Model-Based Clustering for Online Crisis Identification in Distributed Computing

D. B. Woodard and M. Goldszmidt
Cornell University and Microsoft Research

Abstract

Distributed computing systems can suffer from occasional catastrophic violation of performance goals; due to the complexity of these systems, manual diagnosis of the cause of the crisis is prohibitive. Recognizing the recurrence of a problem automatically can lead to cause diagnosis and / or informed intervention. We frame this as an online clustering problem, where the labels (causes) of some of the previous crises may be known. We give an effective solution using model-based clustering based on a Dirichlet process mixture; the evolution of each crisis is modeled as a multivariate time series.

We perform fully Bayesian inference on clusters, giving a method for efficient online computation. Such inferences allow for online expected-cost-minimizing decision making in the distributed computing context. We apply our methods to Microsoft's Exchange Hosted Services.

1 Introduction

Distributed systems perform large computing tasks by farming out sub-tasks to a set of servers. These servers may be spread out across geographical and corporate boundaries in data centers containing tens of thousands of machines. Microsoft's Exchange Hosted Services, for instance, provides email processing in which incoming messages are routed to servers that apply a set of filters and other processing steps before directing the emails to the user, a spam repository, or other destination.

Such systems have performance goals such as maximum processing times; when these goals are not met, attempts are made to diagnose the problem and intervene. Such problems can occur, for instance, when demand is high and servers become overloaded, or due to human misconfigurations (e.g., during software updates) or performance problems in lower-level data centers on which the servers rely (e.g., for performing authentication services). Often the violation of the performance goals is severe and occurs for nearly all of the servers, and such periods are simply called system crises.

When a crisis occurs, it is desirable to identify any previous crises that exhibited similar behavior, and consider the set of interventions, if any, that were successful in those cases. Due to the large scale, the interdependence and the distributed nature of the systems, problems tend to recur and, since human diagnosis is expensive, one must recognize the recurrence of a problem in an automated fashion. A set of status measurements for the servers, such as

CPU utilization and queue length and throughput for various tasks, are available for this purpose; there can be hundreds of these measurements per server (Bodik et al., 2009).

We consider the problem of matching a new crisis to previous crises of mixed known and unknown causes; this is an online clustering problem with partial labeling, and is complicated by the incompleteness of the data for the new crisis. Previous work in crisis / failure identification (Cohen et al., 2005; Yuan et al., 2006; Duan and Babu, 2008; Bodik et al., 2009) uses multi-stage approaches in which statistical, machine learning, or ad-hoc methods are combined in a heuristic fashion. While giving practical solutions and valuable insights into the structure and properties of the data, they do not provide a consistent model for the process of interest. They also restrict to either completely labeled or completely unlabeled data, and do not address the incomplete nature of the new crisis data.

We provide a solution using model-based clustering, where the evolution of each crisis is modeled as a multivariate time series. Fully Bayesian inference is performed to estimate the cluster assignments of the set of previous crises; then identification of the type of a new crisis is simply a prediction problem using the incomplete data for this crisis. The fully Bayesian inference on clusters can be performed during the typically lengthy periods between crises, and when a new crisis begins one can then perform rapid prediction of its type by applying a natural approximation.

A Dirichlet process mixture model (Escobar, 1994; Ishwaran and Zarepour, 2002) is used for the cluster assignments; this allows the number of crisis types to increase as the number of crises increases, while maintaining exchangeability between crises. Since the likelihood function can be highly multimodal, we make the inference on clusters as efficient as possible by combining parallel tempering (Geyer, 1991) with a collapsed-space split-merge Markov chain method (Jain and Neal, 2004). We show that this combination can be superior to use of the split-merge method alone.

We describe how to use our identification of a new crisis to perform optimal decision making, i.e., to choose an intervention that minimizes expected cost. This fully accounts for uncertainty in the crisis type assignments and the parameters of those types, only possible using fully Bayesian inference.

To our knowledge this is the first instance of fully Bayesian online clustering in any context. Dirichlet processes have been applied to online clustering of documents by Zhang et al. (2004), obtaining a single cluster assignment based on the posterior, but in order to perform optimal decision-making one must instead integrate over the entire posterior distribution.

We demonstrate the accuracy of our crisis identification method using simulated data and comparing with a distance-based clustering algorithm. Then we apply our method to the Exchange Hosted Services. Priors for the parameters are obtained by a combining information from experts with information in the data, and reflect the fact that the status measurements are chosen with the goal of being indicative of crisis type, i.e. any particular measurement has a non-trivial probability of behaving similarly across crises of a particular type. We show that this careful prior choice improves clustering performance relative to a "default" prior specification.

Performance of our method (which is quite general) applied to the EHS data without any crisis labels is superior to that of the multi-stage data mining methods of Bodik et al. (2009) that were developed using the same data set; accuracy is measured by comparing to

the known causes of some of the crises.

The rest of the article is organized as follows. In Section 2 we describe the data that are typically available for distributed computing centers. Our model for the crisis evolution and crisis types is given in Section 3. Posterior computation for this model is described in Section 4, and methods for online prediction and optimal decision-making are given in Section 5. The simulation study is presented in Section 6, while results for the Exchange Hosted Services are given in Section 7. In Section 8 we draw conclusions.

2 Measuring Performance in Distributed Computing

In distributed computing a common set of measurements from each server capture its current activity and state. These are typically aggregated over time intervals, which in the case of the Exchange Hosted Services (EHS) are 15 minutes long. EHS handles email traffic, applying a sequence of spam filters and other checks for validity, so that some of the measurements are the number of emails that pass each filter, and the number blocked by each filter, during the 15-minute period.

Distributed computing systems have a set of performance goals, often stated as bounds on the acceptable value for one or more of the server measurements (called Key Performance Indicators or KPIs). An extended period of violation of these performance goals is considered to be a system crisis. In EHS the system is considered to be in violation if at least 25% of the servers are above a threshold for a particular KPI (Bodik et al., 2009). Two consecutive violation periods are considered to define the beginning of a crisis in EHS, and the crisis is considered to continue until there are four consecutive periods of non-violation.

Traces of several KPIs and several non-KPI measurements ("metrics") for EHS are shown in Figure 1 for a ten-day period. The KPI traces show the percentage of servers exceeding the threshold for that KPI, and the other traces are the medians of each of the metrics over the set of servers. There are six crises shown here, namely the periods when one of the KPI traces is above the dashed line. The first two crises are known to have particular causes "A" and "B", while the last four crises are believed to have the same cause "C". It is clear that the third metric is elevated during crises of type C, but not during crises of type A or B. The second metric is elevated during crises of type C, but diminished during crises of type A and B. The first metric appears to be elevated during crises of type C, possibly diminished during crises of type B, and not strongly affected by crises of type A.

This plot suggests that the medians of the metrics over the servers are very informative as to the crisis type. Furthermore, the median of any particular metric appears to be consistently either low, normal, or high during crises of a particular type. This is supported by the opinion of EHS experts, so we fit our models on the median values of the metrics, discretizing according to thresholds that define "low", "normal", or "high" values.

We define the normal range of (the median value of) a metric to be the 2nd and 98th quantile of that metric during non-crisis periods. Applying these quantiles to the EHS data, "high" or "low" values of many of the metrics correspond closely with crisis periods. We expect similar dimension reduction and discretization to be effective in other distributed computing systems.

3 Clustering of System Crises

3.1 Crisis Modeling

We use a time series model for crisis evolution. Denote the vector of metrics for the ith crisis in the lth time period after the start of the crisis by $Y_{il\cdot} = (Y_{il1}, \ldots, Y_{ilJ})$; for crises of type k, we assume that the initial state vector $Y_{i1\cdot}$ is sampled from a discrete distribution, and that the state vector $Y_{il\cdot}$ subsequently evolves according to a Markov chain of order q.

Estimation of the full joint distribution of $Y_{i1\cdot}$ and the full transition matrix is infeasible when the number of crises I is small and the number of metrics J is moderate or large, as is typical (for EHS $I = 27$ and $J = 18$). For such small sample sizes, extremely parsimonious conditional independence structures have been found both empirically and theoretically to provide the best accuracy in estimation of a class variable (Friedman et al., 1997; Domingos and Pazzani, 1997; Hand and Yu, 2001). In particular, naive Bayes models, which assume conditional independence of all attributes conditional on the class, and augmented naive Bayes models that assume only pairwise dependencies conditional on the class, have been found to have the best accuracy.

We therefore assume independence of all metrics conditional on the crisis type (dependencies between pairs of metrics can easily be accommodated by replacing each pair of metrics, having three states each, with a single metric that has nine states). Conditional on k, Y_{i1j} then has a discrete distribution with probability vector $\beta^{(jk)} = (\beta_1^{(jk)}, \beta_2^{(jk)}, \beta_3^{(jk)})$, and Y_{ilj} evolves according to a Markov chain of order q over the three states (1:low, 2:normal, 3:high). For parsimony we take $q = 1$; the elements of the row-stochastic Markov transition matrix are denoted by $T_{st}^{(jk)}$ where the subscripts $s, t \in \{1, 2, 3\}$ indicate the states. This gives $8J$ free parameters for each crisis type, where J is the number of metrics. The resulting complete-data likelihood function is the following, where we condition on the unknown type indicators Z_i of each crisis $i = 1, \ldots, I$ and where the values n_{ijst} are the number of transitions of the jth metric from state s to state t during crisis i:

$$\pi\left(\mathcal{D} \mid \{Z_i\}_{i=1}^{I}, \{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}\right) = \prod_{i,j,t}\left[\left(\beta_t^{(jZ_i)}\right)^{\mathbf{1}(Y_{i1j}=t)} \prod_{s}\left(T_{st}^{(jZ_i)}\right)^{n_{ijst}}\right].$$

3.2 Cluster Modeling

The Dirichlet process mixture (DPM) model provides natural prior specification for online clustering, allowing the number of clusters to increase as the number of crises increases. The DPM can be obtained as the limit of a finite mixture model with Dirichlet prior distribution on the mixing proportions (Escobar and West, 1995; Neal, 2000). In our context the DPM is parameterized by a scalar α controlling the expected number of crisis types occurring in a set of crises, and by a prior distribution $G_0(d\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_j)$ for the set of all parameters associated with each crisis type k, where G_0 does not depend on k.

We take G_0 to be the product over j of independent Dirichlet distributions for $\beta^{(jk)}$ (with parameter vectors $a^{(j)} = (a_1^{(j)}, a_2^{(j)}, a_3^{(j)})$), times the product over j and s of independent Dirichlet distributions for the transition matrix rows $T_{s\cdot}^{(jk)}$ (with parameter vectors $b_s^{(j)} = (b_{s1}^{(j)}, b_{s2}^{(j)}, b_{s3}^{(j)})$). The use of such a product Dirichlet prior distribution for the rows of an

unconstrained transition matrix is standard practice (e.g., Carlin et al. (1992); Diaconis and Rolles (2006)).

The DPM model for the crisis types $\{Z_i\}_{i=1}^I$ and crisis parameters $\beta^{(jk)}$, $T_{\cdot\cdot}^{(jk)}$ can be written as follows, in the case where the causes of the crises are all unknown:

$$\begin{aligned}\pi\left(\{Z_i\}_{i=1}^I\right) &= \pi(Z_1)\prod_{i=2}^{I}\pi\left(Z_i|\{Z_{i'}\}_{i'<i}\right) \\ &= \prod_{i=1}^{I}\left[\frac{\alpha}{\alpha+i-1}\mathbf{1}\left(Z_i = m_{i-1}+1\right) + \frac{1}{\alpha+i-1}\sum_{i'<i}\mathbf{1}\left(Z_i = Z_{i'}\right)\right] \qquad (1)\end{aligned}$$

where $m_i = \max\{Z_{i'} : i' \leq i\}$ for $i > 0$ and $m_0 = 0$, and

$$\pi\left(d\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}|\{Z_i\}_{i=1}^I\right) = \prod_{k=1}^{m_I} G_0\left(d\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_j\right). \qquad (2)$$

Here we have integrated over the Dirichlet process, obtaining a generalized Polya urn scheme (Blackwell and MacQueen, 1973). This form is also called the "Chinese Restaurant Process"; each observation i is conceptually a guest who, upon entering a restaurant, either sits at a table that is already occupied, with probability proportional to the number of guests at that table, or sits at an empty table.

When the causes of some of the crises are known (i.e., the partially labeled case), this information can be captured by indicator functions $\mathbf{1}(Z_i = Z_{i'})$ for pairs of crises i, i' that are known to have the same type (denoted by $i \sim i'$). In this case the prior $\pi(\{Z_i\}_{i=1}^I)$ is proportional to the expression in (1) multiplied by $\prod_{i\sim i'}\mathbf{1}(Z_i = Z_{i'})$, while the prior $\pi(d\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}|\{Z_i\}_{i=1}^I)$ is unchanged.

3.3 Choice of prior constants

We select the prior hyperparameters α, $a^{(j)}$, and $b_s^{(j)}$ by combining information elicited from domain experts with information in the data. The former is formal Bayes, while the latter is empirical Bayes (Carlin and Louis, 2009).

According to the Dirichlet process mixture model, the probability that two randomly chosen crises are of the same type is $1/(\alpha+1)$. The EHS experts estimate this to be 0.1, yielding $\alpha = 9$. This implies that the expected number of crisis types in the EHS data is 12.9 (for 27 crises), which the experts agree is reasonable.

Our choices of $a^{(j)}$ and $b_s^{(j)}$ reflect the fact that the metrics are selected to be indicative of crisis type, i.e. any metric has non-trivial probability of behaving similarly across crises of a particular type. EHS experts believe that there is a substantial prior probability for any j and k that one of the values $\beta_t^{(jk)}$ is "close" to one. This is formalized by specifying a 50% prior probability that $\beta_1^{(jk)} > .85$, $\beta_2^{(jk)} > .95$, or $\beta_3^{(jk)} > .85$ (the threshold is higher for $\beta_2^{(jk)}$ since the value "normal" is more common than the values "low" or "high"). $a^{(j)}$ is then uniquely determined by the prior mean for $\beta^{(jk)}$, which we take to be the empirical

distribution $\beta^* = (\beta_1^*, \beta_2^*, \beta_3^*)$ of the first value of all metrics in all crises. Sensitivity to these choice is examined in Section 7.1.

Selection of $b_s^{(j)}$ is analogous. We use the data for all crises and all metrics to find the empirical transition probabilities $T_{s\cdot}^*$ from each starting state s, and set the prior mean of $T_{s\cdot}^{(jk)}$ equal to $T_{s\cdot}^*$. Then we consider the limiting distribution $r^{(jk)} = (r_1^{(jk)}, r_2^{(jk)}, r_3^{(jk)})$ of a Markov chain with transition kernel $T_{\cdot\cdot}^{(jk)}$. Since the metrics tend to behave consistently across crises of a particular type, there is non-trivial probability that one of the values $r_t^{(jk)}$ is "close" to one. This is formalized via a 50% prior probability that $r_1^{(jk)} > .85$, $r_2^{(jk)} > .95$, or $r_3^{(jk)} > .85$. By specifying the prior weight of evidence, meaning the sum of $b_s^{(j)}$, to be equal for each row s of the transition matrix, the values of $b_s^{(j)}$ are then uniquely determined.

Default prior specification: One might alternatively consider the "default" choice of $a_t^{(j)} = b_{st}^{(j)} = 1$ for all j, s, and t; this gives a generalized uniform prior for each of the vectors $\beta^{(jk)}$ and each of the rows $T_{s\cdot}^{(jk)}$ of the transition matrices, and is the most common choice when performing Bayesian inference on discrete distributions or transition matrices (Carlin et al., 1992; Diaconis and Rolles, 2006). However, this choice conflicts with expert opinion and the data, implying a very small prior probability that a particular metric behaves similarly across crises of a particular type, in the sense above. This discrepancy negatively impacts clustering performance (Section 7.1).

4 Posterior Computation

For a fixed set of crises, Markov chain methods can be used to obtain samples from the posterior distribution $\pi(\{Z_i\}_{i=1}^I, \{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k} | \mathcal{D})$ of the clustering model given in Section 3. We use a collapsed-space Markov chain method (Jain and Neal, 2004), modified with parallel tempering (Geyer, 1991). The collapsed-space sampler simulates a Markov chain with target distribution $\pi(\{Z_i\}_{i=1}^I | \mathcal{D})$ on the reduced space $\{Z_i\}_{i=1}^I$; this is possible by marginalizing out the cluster-specific parameters, in our case $\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}$. Posterior samples for the cluster parameters can then be obtained by sampling from their conditional posterior distribution; details are given in Appendix A. Collapsed-space samplers have been found both empirically and theoretically to be more efficient than their full-space counterparts (Liu, 1994).

Collapsed-space sampler: The basic collapsed-space sampler is composed of Gibbs updates of each Z_i in turn. In order to address the potential multimodality of the posterior distribution $\pi(\{Z_i\}_{i=1}^I | \mathcal{D})$, Jain and Neal (2004) add a Metropolis-Hastings move that merges two clusters into one or splits a cluster into two. The authors give empirical evidence showing that the addition of this move speeds convergence.

In the distributed computing context, the number of metrics can be large, and the resulting likelihood can have extremely narrow and well-separated modes corresponding to distinct cluster assignments. Here even the collapsed-space sampler with split-merge moves can have difficulty mixing between the modes. Additionally, convergence diagnostics can be difficult to apply; the parameters are the cluster indicators Z_i of the individual crises, which

for a particular crisis may take only one or two values for the entire simulation even when the mixing is good.

Parallel tempering: In order to further improve the efficiency of the Markov chain, and to facilitate the use of convergence diagnostics, we modify the Markov chain by applying parallel tempering. This technique has been proven to dramatically improve Markov chain efficiency for many multimodal distributions (Woodard et al., 2009). It simulates parallel Markov chains indexed by $l = 1, \ldots, L$ using identical updating strategies but distinct target distributions ϕ_l; we take $\phi_l(\{Z_i\}_{i=1}^I) \propto \pi(\{Z_i\}_{i=1}^I)\pi(\mathcal{D}|\{Z_i\}_{i=1}^I)^{\beta_l}$ where $\pi(\mathcal{D}|\{Z_i\}_{i=1}^I)$ is the marginal likelihood (see Appendix A) and $0 \leq \beta_1 \leq \ldots \leq \beta_L = 1$. The first distribution ϕ_1 is close to the prior if $\beta_1 \approx 0$, so that chain 1 efficiently explores the state space, and the other target distributions ϕ_l interpolate between ϕ_1 and the posterior $\phi_L = \pi(\{Z_i\}_{i=1}^I|\mathcal{D})$. The chains share samples in the sense that swaps are proposed between the states of adjacent chains; these swaps are constructed to guarantee convergence of the joint process to the product distribution $\prod_{l=1}^L \phi_l$. The samples from chain L converge marginally to the posterior ϕ_L, and can be used for Monte Carlo inference.

The "inverse temperatures" β_l are chosen as follows. We take $\beta_1 = 0$, and select the smallest set of inverse temperatures β_l that gives swap acceptance rates of at least 20%; theoretical and empirical results to support this choice of spacing are given in Atchadé et al. (2009). This full set of inverse temperatures are used in Section 7.1 for the EHS data, although for the purposes of the simulation experiments in Section 6 we simplify by using the five largest inverse temperatures ($L = 5$, still chosen to have swap acceptance $\sim 20\%$) instead of the full set; this simplification appears to have little practical impact and saves time for these experiments.

Convergence diagnosis: We apply standard convergence diagnostics (Cowles and Carlin, 1996) to assess convergence of the parallel tempering process. Even if for a particular crisis i the indicator Z_i takes only a single value at the lowest temperature, Z_i takes many values at the higher temperatures, allowing convergence diagnosis.

To detect any lack of convergence due to multimodality, we simulate the parallel tempering process multiple times and apply the convergence diagnostic by Gelman and Rubin (1992). This requires sampling the initial parameter vectors from a distribution that is "overdispersed" relative to the posterior distribution, so we draw these from the prior $\pi(\{Z_i\}_{i=1}^I)$.

5 Online Prediction and Decision-Making

We wish to identify a new crisis in real time, given the data $\mathcal{D}$ from previous crises and the data $\mathcal{D}_{new}$ so far for the new crisis. This consists of estimating $\Pr(Z_{new} = Z_i|\mathcal{D}, \mathcal{D}_{new})$ for each previous crisis $i = 1, \ldots, I$ and $\Pr(Z_{new} \neq Z_i \ \forall i|\mathcal{D}, \mathcal{D}_{new})$ where Z_{new} is an indicator of the type of the new crisis.

5.1 Exact Prediction

To perform inference for Z_{new} we can apply the Markov chain method from Section 4 to the data from past crises plus the data available so far for the new crisis, i.e., clustering the $I+1$ crises. This can be done with as little as a single time period of data for the new crisis, since the time series model given in Section 3.1 still applies. We then have Monte Carlo estimates for the desired probabilities:

$$\hat{\Pr}(Z_{new} = Z_i | \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}(Z_{new}^{(l)} = Z_i^{(l)})$$

$$\hat{\Pr}(Z_{new} \neq Z_i \ \forall i | \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}(Z_{new}^{(l)} \neq Z_i^{(l)} \ \forall i)$$

where $\left(\{Z_i^{(l)}\}_{i=1}^{I}, Z_{new}^{(l)} \right)$ for $l = 1, \ldots, L$ are the posterior sample vectors from the Markov chain.

This is practical when the number of past crises is small (for $I = 15$, $J = 15$ the Markov chain Monte Carlo computation takes less than 5 minutes on a standard processor), but after many crises this is unacceptably slow for a context requiring rapid decision-making.

5.2 Approximate Prediction

We provide an efficient alternative for prediction, based on the approximation:

$$\begin{aligned} \Pr(Z_{new} = Z_i | \mathcal{D}, \mathcal{D}_{new}) &= \sum_{\{Z_i\}_{i=1}^{I}} \Pr(Z_{new} = Z_i | \{Z_i\}_{i=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \Pr(\{Z_i\}_{i=1}^{I} | \mathcal{D}, \mathcal{D}_{new}) \\ &\approx \sum_{\{Z_i\}_{i=1}^{I}} \Pr(Z_{new} = Z_i | \{Z_i\}_{i=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \Pr(\{Z_i\}_{i=1}^{I} | \mathcal{D}) \end{aligned} \tag{3}$$

and the analogous approximation for $\Pr(Z_{new} \neq Z_i \ \forall i | \mathcal{D}, \mathcal{D}_{new})$. These assume that the data from the new crisis do not tell us very much about the past crisis types $\{Z_i\}_{i=1}^{I}$; this is quite accurate in practice, as demonstrated in Section 6.2.

The conditional distribution $\Pr(Z_{new} | \{Z_i\}_{i=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$ of Z_{new} is obtained by:

$$\begin{aligned} \Pr(Z_{new} | \{Z_i\}_{i=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) &\propto \Pr(Z_{new}, \{Z_i\}_{i=1}^{I} | \mathcal{D}, \mathcal{D}_{new}) \\ &\propto \Pr(Z_{new} | \{Z_i\}_{i=1}^{I}) \Pr(\mathcal{D}, \mathcal{D}_{new} | Z_{new}, \{Z_i\}_{i=1}^{I}) \end{aligned}$$

where $\Pr(\mathcal{D}, \mathcal{D}_{new} | Z_{new}, \{Z_i\}_{i=1}^{I})$ is available in closed form as shown in the Appendix, and where (from the Dirichlet process mixture model in (1))

$$\Pr(Z_{new} | \{Z_i\}_{i=1}^{I}) \propto \alpha \mathbf{1}\left(Z_{new} = m_I + 1\right) + \sum_{i'=1}^{I} \mathbf{1}\left(Z_{new} = Z_{i'}\right).$$

Given these facts, we propose the following two-step method:

Method for Approximate Prediction

1. After the end of each crisis, refit the clustering model by simulating the Markov chain described in Section 4. This yields sample vectors $\{Z_i^{(l)}\}_{i=1}^I$ from the posterior distribution $\pi(\{Z_i\}_{i=1}^I|\mathcal{D})$.

2. When a new crisis begins, use its data $\mathcal{D}_{new}$ to calculate the Monte Carlo estimates:

$$\hat{\Pr}(Z_{new} = Z_i|\mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L}\sum_{l=1}^{L} \Pr(Z_{new} = Z_i^{(l)}|\{Z_{i'}^{(l)}\}_{i'=1}^I, \mathcal{D}, \mathcal{D}_{new})$$

$$\hat{\Pr}(Z_{new} \neq Z_i\ \forall i|\mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L}\sum_{l=1}^{L} \Pr(Z_{new} \neq Z_i^{(l)}\ \forall i|\{Z_{i'}^{(l)}\}_{i'=1}^I, \mathcal{D}, \mathcal{D}_{new}).$$

Part 1 is the slower part of the computation, but takes much less than the hours or days that typically pass between crises. The computation in part 2 above is $\mathrm{O}(LIJ)$, very manageable in real time.

5.3 Expected-Cost-Minimizing Decision Making

Given an appropriate cost function, we can use our inferences for a new crisis to perform expected-cost-minimizing decision making. In fact, performing optimal decision making, i.e., while conditioning only on the data and not on particular estimates of the parameters, can only be done via fully Bayesian inference such as we have described (Robert, 2001).

A cost function specifies the total cost of a crisis as a function of the true crisis type and the action taken. Taking an action that quickly resolves a crisis gives low total cost, while taking an action that prolongs a crisis leads to high total cost. The costs of a crisis include, for instance, payouts to clients for violation of service agreements as well as client dissatisfaction.

More precisely, the total cost of the new crisis is a function $C\left[\phi, \left(\{Z_i^*\}_{i=1}^I, Z_{new}^*\right)\right]$ of the action ϕ and the entire vector of true crisis types $\left(\{Z_i^*\}_{i=1}^I, Z_{new}^*\right)$, due to the fact that Z_{new}^* is only meaningful in the context of $\{Z_i^*\}_{i=1}^I$. If we knew C, and given posterior sample vectors $\left(\{Z_i^{(l)}\}_{i=1}^I, Z_{new}^{(l)}\right)$ as in Section 5.1, the expected cost of taking ϕ during the new crisis could be estimated consistently as

$$\mathbf{E}(C) \approx \frac{1}{L}\sum_{l=1}^{L} C\left[\phi, (\{Z_i^{(l)}\}_{i=1}^I, Z_{new}^{(l)})\right].$$

A similar expression is obtained when using the approximation given in Section 5.2.

The expected-cost-minimizing action is the value of ϕ that minimizes $\mathbf{E}(C)$. Although the cost function C is not known in practice, for actions ϕ that have been taken during previous crises the realized costs can be used to estimate C, and for other actions expert knowledge can be used to estimate C.

We will evaluate the accuracy of our crisis identification method while keeping in mind the ultimate goal, namely choosing the optimal action. For this reason we will avoid choosing a

particular estimate of the crisis types $\left(\{Z_i\}_{i=1}^{I}, Z_{new}\right)$, and instead will consider the accuracy of the soft identification, i.e., the posterior distribution over $\left(\{Z_i\}_{i=1}^{I}, Z_{new}\right)$ as given in Sections 5.1 and 5.2.

6 A Simulation Study

We demonstrate the accuracy of our methods on simulated data. We first address the offline setting, i.e., applying the fully Bayesian clustering algorithm described in Sections 3 and 4 to a fixed set of crises. Then we consider accuracy of online prediction.

6.1 Offline Accuracy

We examine offline accuracy, varying the number of crises and metrics and comparing with a distance-based clustering algorithm. We sample I crises of equal length M by first drawing the vector of crisis type indicators and the parameters $\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}$ of the crisis types according to the Dirichlet process mixture model described in Section 3.2, then simulating the metrics for each crisis from the model described in Section 3.1. The crisis lengths must be equal in order to allow for comparison with distance-based clustering.

Inferences for the model-based clustering method are obtained as described in Section 4. We compare this method with K-means, a common distance-based clustering algorithm (Hartigan and Wong, 1979). We apply K-means using Euclidean distance between the observation vectors, which consist of the values (1, 2, or 3) of all of the metrics during all of the time periods of each crisis. We find that in our context standard K-means criteria for estimating the number of crisis types perform very poorly (typically estimating only a few clusters). In order to obtain best-case accuracy measures for K-means, we therefore apply it with either the true number of clusters ("K-Means 1"), or with half the true number of clusters ("K-Means 2"), since using fewer than the true number of clusters can improve the performance of K-means (Booth et al., 2008). These scenarios are unrealistically optimistic since the number of clusters is unknown, but as we will see the performance of K-means is still dramatically lower than that of our model-based clustering method.

Twenty data sets are simulated for each of several combinations of I and J, and the following measures of accuracy are obtained for model-based clustering (MBC) and for K-means 1 and 2:

1. **Pairwise Sensitivity:** Of the pairs of crises that are of the same type, the percentage that are assigned to the same cluster (for MBC, that have posterior probability greater than 0.5 of being in the same cluster).

2. **Pairwise Specificity:** Of the pairs of crises that are not of the same type, the percentage that are not assigned to the same cluster (for MBC, that have posterior probability no more than 0.5 of being in the same cluster).

3. **Error of No. Crisis Types:** The absolute error of the estimated number of crisis types occurring in the data, divided by the true number of crisis types. This is only relevant for MBC, and in this case the posterior mean is used to estimate the number of types.

The first two measures have been used, e.g., in Booth et al. (2008). Some variant of the last measure is used in almost all analyses of clustering performance.

In order to simulate data that are as realistic as possible, we take the crisis length to be that of most of the crises in the EHS data (8 time periods, due to the truncation described in Section 7.1) and take the hyperparameter values $a^{(j)}, b_s^{(j)}$ to be those obtained for the EHS data as described in Section 3.3. We use a smaller value of α ($\alpha = 4$) than that obtained for EHS, in order to estimate the pairwise sensitivity accurately when the number of crises is small.

Values of the accuracy measures are reported in Table 1, averaged over the simulated data sets and along with their standard errors. The performance of both K-means 1 and 2 is dramatically lower than that of MBC in terms of the pairwise accuracy measures. K-means 2 has higher pairwise sensitivity and lower pairwise specificity than K-means 1, since it uses fewer clusters.

Increasing the number of metrics results in improved performance of MBC by all measures. This is expected, since more metrics means more evidence available to estimate the clusters. By contrast, the number of crises in the data does not appear to have a strong effect on any of the accuracy measures. This helps explain the excellent accuracy that we find in the online context (Section 6.2), where the number of crises starts small and gradually increases.

6.2 Online Accuracy

We examine the accuracy of our method in the online context; given a set of simulated crises in a particular order, we predict the type of each crisis based on the data from the previous crises and partial data for the new crisis. Here it is not possible to apply distance-based clustering methods since the length of the available data for the new crisis can be different from that for the previous crises. We instead compare predictive accuracy of the approximate method given in Section 5.2 ("MBC") to that of the exact method in Section 5.1 ("MBC-EX"), in order to justify the approximation. For data simulated as in Section 6.1, we evaluate several measures of accuracy for MBC and MBC-EX:

1. **Full-data misclassification rate:** The percentage of crises whose predicted type is incorrect, using all of the data for the new crisis. Here "correct" predicted type means that $\hat{\Pr}(Z_{new} \neq Z_i \ \forall i | \mathcal{D}, \mathcal{D}_{new}) > 0.5$ if $Z_{new} \neq Z_i \ \forall i$ according to the gold standard (here, the truth), and otherwise that $\hat{\Pr}(Z_{new} = Z_i | \mathcal{D}, \mathcal{D}_{new}) > 0.5$ for some $i \leq I$ such that $Z_{new} = Z_i$ according to the gold standard.

2. **p-period misclassification rate:** The percentage of crises whose predicted type is incorrect, using the first p time periods of data for the new crisis.

3. **Average time to correct identification:** The average number of time periods required to obtain the correct identification, for crises that are correctly identified when using the full data for the new crisis.

We do not evaluate the average time to correct identification for MBC-EX, since this is extremely computationally intensive. The average values of the above accuracy measures over five simulated data sets are shown in Table 2 for several combinations of I and J.

The accuracy is high for both MBC and MBC-EX, correctly classifying over 80% of crises in every setting we considered. The performance of MBC is not significantly worse than that of MBC-EX, showing the accuracy of the approximation given in Section 5.2. Using more metrics shortens the average time to identification, and there is some evidence that it also reduces the misclassification rates.

The accuracy of both methods degrades when using the data from only the first three time periods of the new crisis, but still over 80% of crises are correctly classified in all cases. The average time to correct identification for MBC is between one and two time periods–this means that on average crises are correctly identified before the key performance indicators exceed their thresholds. Such early identification of a crisis is extremely helpful in choosing an appropriate intervention.

7 Application to Exchange Hosted Services

In the first four months of 2008, 27 crises occurred in EHS. The causes of some of these crises have been diagnosed by EHS experts, and are listed in Table 3.

Preprocessing as in Bodik et al. (2009): We choose a subset of the available metrics by applying their feature selection procedure. In cases of pairs of metrics with correlation greater than 0.95 we remove one, leaving 18 metrics.

To facilitate early crisis identification, it is helpful to include the data from the half-hour just before the start of each crisis in fitting the model and estimating the type of a new crisis. Additionally, we do not use data after the first hour and a half of each crisis, since the metrics are not believed to be informative as to the crisis type after this time.

7.1 Offline Application

To test the accuracy of the offline crisis identification method given in Section 4, we apply it to the whole set of EHS crises without the known crisis labels, and compare our results to those labels.

Markov chain trace plots are shown in Figure 2, illustrating the convergence of the chains. The samples of Z_{22} for several values of the inverse temperature β are shown. The chain with $\beta = 1$, which samples from the posterior distribution $\pi(\{Z_i\}_{i=1}^I|\mathcal{D})$, primarily visits the single value $Z_{22} = 2$, while the chains with smaller values of β visit progressively more values of Z_{22}. This facilitates convergence diagnosis and exploration of the space. Using 10^6 iterations, the smallest Geweke diagnostic p-value for the Markov chains is 0.44 after Bonferroni correction, detecting no lack of convergence. Here univariate tests are done for each parameter $Z_i : i \neq 1$ at each inverse temperature. Similarly, we obtain a maximum Gelman-Rubin scale factor of 1.01, again evidence of good convergence. This maximum is taken over $Z_i : i \neq 1$ for inverse temperatures β less than 0.5 (for $\beta \geq 0.5$ there are numerical difficulties, since some Z_i take a single value for almost all iterations).

The sizes of the clusters from the posterior mode cluster assignment are shown in Table 3. This cluster assignment has 58% posterior probability, and along with the second-highest probability assignment accounts for a total of 93.8% of the posterior probability. This second

assignment has only a single difference with the first, namely a change in the labeling of one crisis, increasing the count of type B to 15 and decreasing the count of type K to 5. We will summarize the accuracy of the posterior mode clustering assignment relative to the known causes, but this summary applies equally well to the second assignment since the crisis for which they differ has unknown cause.

Comparison to known causes: The posterior mode crisis labels for the most part match the known causes, with the exception of four uncommon crisis types that are incorrectly clustered with more common types. The largest cluster obtained by MBC corresponds to the cause "overloaded back-end"; all eight of the crises known to be of this type are correctly clustered together, along with six other crises (most of which have unknown cause). The "overloaded back-end" problem occurs due to poor performance of another data center, one on which the servers depend. The EHS technicians cannot intervene to improve the performance of that separate data center, explaining why this is the most common type of crisis.

The two crises of known cause "overloaded front-end" are also correctly clustered together. Similarly, the "database configuration error", "workload spike", and "request routing error" clusters are correctly identified.

Four uncommon crisis types are incorrectly clustered with more common types. For instance, the "configuration error" crisis is clustered with the "overloaded front-end" crises. This type of mistake occurs partly due to the fact that crises having different causes can have the same patterns in their metrics. In the most extreme case, the metrics appear to be indistinguishable between the two crisis types.

However, in the other cases while the large majority of metrics are indistinguishable between the crisis types, a few metrics show distinct behavior between the types. Since we have assumed the parameters of distinct crisis types to be independent a priori, the presence of distinct crisis types with similar patterns for most metrics is very improbable under the prior. Such crisis types are therefore clustered together. This issue could potentially be fixed by creating an appropriate dependence structure between the crisis types in the prior distribution.

7.1.1 Sensitivity to Prior Specification

When applying our method for crisis identification, use of the generalized uniform prior (having $a_t^{(j)} = b_{st}^{(j)} = 1$ for all j, s, and t) can lead to nonsensical results, for reasons described in Section 3.3. Most or all of the crises are assigned to the same cluster; for EHS, all crises except one are assigned to a single cluster.

As long as the prior is not dramatically inconsistent with the data, results are not sensitive to the prior specification. Changing the prior described in Section 3.3 to obtain a 50% prior probability that $\beta_1^{(jk)} > .95$, $\beta_2^{(jk)} > .97$, or $\beta_3^{(jk)} > .95$ and a 50% prior probability that $r_1^{(jk)} > .95$, $r_2^{(jk)} > .97$, or $r_3^{(jk)} > .95$, for instance, does not change the posterior mode clustering assignment.

7.2 Online Application

We evaluate the accuracy of online clustering for the EHS data, relative to the offline clustering assignment. Permuting the order of the EHS crises, we apply the online crisis identification method given in Section 5.1 and evaluate the accuracy measures described in Section 6.2, treating the posterior mode cluster assignments from the offline context (Section 7.1) as the gold standard.

For the original crisis ordering, we obtain a full-data misclassification rate of 7.4%, a 3-period misclassification rate of 14.8%, and an average time to correct identification of 1.81 time periods. This means that on average the crises are identified correctly before the key performance indicators exceed their thresholds. Two-thirds of the crises are identified correctly in the first time period.

Taking the average over five random permutations of the crises, we obtain a full-data misclassification rate of 5.9% (SE=3.4%), a 3-period misclassification rate of 11.8% (SE=3.2%), and an average time to correct identification of 1.56 (SE=0.07). The excellent identification performance of MBC clearly does not depend on the particular ordering of the crises. Our full-data misclassification rate on permuted data is superior to that reported in Bodik et al. (2009), which is over 20% using the same data and permutation procedure.

8 Conclusions

We have given a method for fully Bayesian online crisis identification in distributed computing, and have described how to use this to perform expected-cost-minimizing crisis mitigation. Accuracy has been demonstrated on both simulated data and data from the Exchange Hosted Services; our method outperforms a distance-based classifier in the offline setting, and the multi-stage data mining methods of Bodik et al. (2009) in the online setting, in both cases by a large margin.

Importantly, our method provides natural solutions to several related problems; these are explored in Goldszmidt and Woodard (2009). First, during a crisis one can forecast its evolution. Second, the model-based approach allows for interpretation of the crisis types, which can aid identification of the causes. For instance, one can distinguish the system status metrics that are most strongly associated with crises of a particular type. This question alone has received considerable attention (Cohen et al., 2004; Zhang et al., 2005), and is resolved naturally in the context of our time series model. Finally, one could potentially model not just the evolution of crises of a particular type, but also how this evolution depends on the intervention taken.

References

Atchadé, Y., Roberts, G. O., and Rosenthal, J. S. (2009), Optimal scaling of Metropolis-coupled Markov chain Monte Carlo,. Unpublished manuscript.

Blackwell, D., and MacQueen, J. B. (1973), "Ferguson distributions via Polya schemes," *Annals of Statistics*, 1, 353–355.

Bodik, P., Goldszmidt, M., Fox, A., and Andersen, H. (2009), Fingerprinting the datacenter: Automated classification of performance crises,, Technical report, Microsoft Research.

Booth, J. G., Casella, G., and Hobert, J. P. (2008), "Clustering using objective functions and stochastic search," *Journal of the Royal Statistical Society, Series B*, 70, 119–139.

Carlin, B. P., Gelfand, A. E., and Smith, A. F. M. (1992), "Hierarchical Bayesian analysis of changepoint problems," *Journal of the Royal Statistical Society, Series C*, 41, 389–405.

Carlin, B. P., and Louis, T. A. (2009), *Bayesian Methods for Data Analysis*, 3rd edn, Boca Raton, FL: Chapman and Hall.

Cohen, I., Goldszmidt, M., Kelly, T., Symons, J., and Chase, J. S. (2004), Correlating instrumentation data to system states: A building block for automated diagnosis and control,, in *6th Symposium on Operating Systems Design and Implementation*.

Cohen, I., Zhang, S., Goldszmidt, M., Symons, J., Kelly, T., and Fox, A. (2005), Capturing, indexing, clustering, and retrieving system history,, in *Symposium on Operating System Principles*.

Cowles, M. K., and Carlin, B. P. (1996), "Markov chain Monte Carlo convergence diagnostics: a comparative review," *Journal of the American Statistical Association*, 91, 883–904.

Diaconis, P., and Rolles, S. W. W. (2006), "Bayesian analysis for reversible Markov chains," *Annals of Statistics*, 34, 1270–1292.

Domingos, P., and Pazzani, M. (1997), "On the optimality of the simple Bayesian classifier under zero-one loss," *Machine Learning*, 29, 103–130.

Duan, S., and Babu, S. (2008), Guided problem diagnosis through active learning,, in *Proc. of the International Conference on Autonomic Computing*, pp. 45–54.

Escobar, M. D. (1994), "Estimating normal means with a Dirichlet process prior," *Journal of the American Statistical Association*, 89, 268–277.

Escobar, M. D., and West, M. (1995), "Bayesian density estimation and inference using mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Friedman, N., Geiger, D., and Goldszmidt, M. (1997), "Bayesian Network Classifiers," *Machine Learning*, 29, 131–163.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004), *Bayesian Data Analysis*, 2nd edn, Boca Raton, FL: Chapman & Hall.

Gelman, A., and Rubin, D. B. (1992), "Inference from iterative simulation using multiple sequences," *Statistical Science*, 7, 457–472.

Geyer, C. J. (1991), Markov chain Monte Carlo maximum likelihood,, in *Computing Science and Statistics, Volume 23: Proceedings of the 23rd Symposium on the Interface*, ed. E. Keramidas, Interface Foundation of North America, Fairfax Station, VA, pp. 156–163.

Goldszmidt, M., and Woodard, D. B. (2009), Bayesian Inference for Crisis Characterization in Distributed Computing,. Unpublished.

Hand, D. J., and Yu, K. (2001), "Idiot's Bayes: Not so stupid after all?," *International Statistical Review*, 69, 385–398.

Hartigan, J. A., and Wong, M. A. (1979), "Algorithm AS-136: A K-means clustering algorithm," *Applied Statistics*, 28, 100–108.

Ishwaran, H., and Zarepour, M. (2002), "Dirichlet prior sieves in finite normal mixtures," *Statistica Sinica*, 12, 941–963.

Jain, S., and Neal, R. M. (2004), "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model," *Journal of Computational and Graphical Statistics*, 13, 158–182.

Liu, J. S. (1994), "The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem," *Journal of the American Statistical Association*, 89, 958–966.

Neal, R. M. (2000), "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, 9, 249–265.

Robert, C. P. (2001), *The Bayesian Choice*, 2nd edn, New York: Springer-Verlag.

Woodard, D. B., Schmidler, S. C., and Huber, M. (2009), "Conditions for rapid mixing of parallel and simulated tempering on multimodal distributions," *Annals of Applied Probability*, 19, 617–640.

Yuan, C., Lao, N., Wen, J., Li, J., Zhang, Z., Wang, Y., and Ma, W. (2006), Automated known problem diagnosis with event traces,, in *Eurosys 2006*, Leuven, Belgium.

Zhang, J., Ghahramani, Z., and Yang, Y. (2004), A probabilistic model for online document clustering with application to novelty detection,, in *Advances in Neural Information Processing Systems*.

Zhang, S., Cohen, I., Goldszmidt, M., Symons, J., and Fox, A. (2005), Ensembles of models for automated diagnosis of system performance problems,, in *Dependable Systems and Networks*.

A Markov Chain Monte Carlo Computations

The marginal likelihood given $\{Z_i\}_{i=1}^I$ is:

$$\begin{aligned}\pi(\mathcal{D}|\{Z_i\}_{i=1}^I) &= \int \pi\left(\mathcal{D} \mid \{Z_i\}_{i=1}^I, \beta^{(jk)}, \{T_{\cdot\cdot}^{(jk)}\}_{j,k}\right) \;\; \pi\left(d\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k} \mid \{Z_i\}_{i=1}^I\right) \\ &= \int \left[\prod_{i=1}^{I}\prod_{j,t}\left(\beta_t^{(jZ_i)}\right)^{\mathbf{1}(Y_{i1j}=t)}\right]\prod_{k=1}^{m_I}\prod_{j}\text{Dirichlet}(d(\beta_1^{(jk)},\beta_2^{(jk)},\beta_3^{(jk)});a^{(j)})\times \\ &\quad \left[\prod_{i=1}^{I}\prod_{j,s,t}\left(T_{st}^{(jZ_i)}\right)^{n_{ijst}}\right]\prod_{k=1}^{m_I}\prod_{j,s}\text{Dirichlet}(d(T_{s1}^{(jk)},T_{s2}^{(jk)},T_{s3}^{(jk)});b_s^{(j)})\end{aligned}$$

where $\text{Dirichlet}(d(T_{s1}^{(jk)},T_{s2}^{(jk)},T_{s3}^{(jk)});b_s^{(j)})$ is the finite-dimensional Dirichlet distribution for $(T_{s1}^{(jk)},T_{s2}^{(jk)},T_{s3}^{(jk)})$ with parameter vector $b_s^{(j)}$. By multinomial-Dirichlet conjugacy (Gelman et al., 2004),

$$\pi(\mathcal{D}|\{Z_i\}_{i=1}^I) = \prod_{k=1}^{m_I}\prod_{j}\left[\frac{\Gamma\left(\sum_{t\in\mathbb{Z}_3}a_t^{(j)}\right)\prod_{t\in\mathbb{Z}_3}\Gamma\left(a_t^{(j)}+\sum_{i:Z_i=k}\mathbf{1}(Y_{i1j}=t)\right)}{\Gamma\left(\sum_{t\in\mathbb{Z}_3}\left[a_t^{(j)}+\sum_{i:Z_i=k}\mathbf{1}(Y_{i1j}=t)\right]\right)\prod_{t\in\mathbb{Z}_3}\Gamma\left(a_t^{(j)}\right)}\right]\times \tag{4}$$

$$\prod_{k=1}^{m_I}\prod_{j,s}\left[\frac{\Gamma\left(\sum_{t\in\mathbb{Z}_3}b_{st}^{(j)}\right)\prod_{t\in\mathbb{Z}_3}\Gamma\left(b_{st}^{(j)}+\sum_{i:Z_i=k}n_{ijst}\right)}{\Gamma\left(\sum_{t\in\mathbb{Z}_3}\left[b_{st}^{(j)}+\sum_{i:Z_i=k}n_{ijst}\right]\right)\prod_{t\in\mathbb{Z}_3}\Gamma\left(b_{st}^{(j)}\right)}\right]. \tag{5}$$

The posterior distribution of $\{Z_i\}_{i=1}^I$ is proportional to the product of $\pi(\{Z_i\}_{i=1}^I)$ and $\pi(\mathcal{D}|\{Z_i\}_{i=1}^I)$, given in (1) and (5), respectively. A Markov chain can then be constructed to sample on this reduced space. For instance, a Gibbs sampler for $\{Z_i\}$ updates each Z_i conditional on $Z_{[-i]} = \{Z_{i'}\}_{i'\neq i}$. The posterior distribution of Z_i conditional on $Z_{[-i]}$ is proportional to $\pi(\{Z_i\}_{i=1}^I|\mathcal{D})$; computation consists of enumerating over the possible values of Z_i and normalizing to obtain the conditional distribution. The possible options are that Z_i is equal to one of the values in $Z_{[-i]}$, or that it is not equal to any of the values in $Z_{[-i]}$. Notice that any of these possibilities may require relabeling of the crisis types, to ensure that the first occurrences of the types are correctly ordered.

Once we have obtained a set of posterior samples of $\{Z_i\}_{i=1}^I$ by simulating such a Markov chain, we can also obtain posterior samples of $\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}$, by noticing that

$$\begin{aligned}\pi(\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k}|\{Z_i\}_{i=1}^I,\mathcal{D}) &= \prod_{k=1}^{m_I}\prod_{j}\text{Dirichlet}(d(\beta_1^{(jk)},\beta_2^{(jk)},\beta_3^{(jk)});\hat{a}^{(j)})\times \\ &\quad \prod_{k=1}^{m_I}\prod_{j,s}\text{Dirichlet}(d(T_{s1}^{(jk)},T_{s2}^{(jk)},T_{s3}^{(jk)});\hat{b}_s^{(j)})\end{aligned} \tag{6}$$

where $\hat{a}_t^{(j)} = a_t^{(j)} + \sum_{i:Z_i=k} \mathbf{1}(Y_{i1j} = t)$ and $\hat{b}_{st}^{(j)} = b_{st}^{(j)} + \sum_{i:Z_i=k} n_{ijst}$ for $t = 1, 2, 3$. For each posterior sample of $\{Z_i\}_{i=1}^I$, generate one sample from $\pi(\{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k} | \{Z_i\}_{i=1}^I, \mathcal{D})$; this gives joint posterior samples of the full set of parameters $(\{Z_i\}_{i=1}^I, \{\beta^{(jk)}, T_{\cdot\cdot}^{(jk)}\}_{j,k})$.

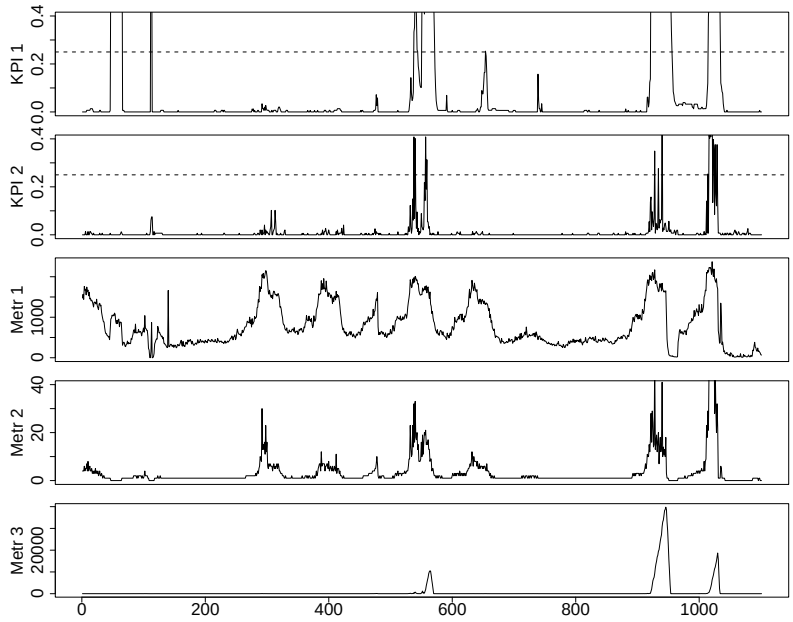

Figure 1: Traces of several KPIs and metrics for EHS over a period of ten days.

No. Crises	No. Metrics	Method	Pairwise Sensitivity	Pairwise Specificity	% Error No. Types
15	**10**	**MBC**	**94.6 (2.08)**	**99.0 (0.50)**	**9.3 (1.87)**
		K-Means 1	47.8 (4.26)	95.3 (0.57)	–
		K-Means 2	74.8 (5.39)	77.9 (1.73)	–
15	**15**	**MBC**	**99.0 (1.00)**	**99.4 (0.41)**	**3.7 (0.95)**
		K-Means 1	69.6 (4.76)	97.0 (0.54)	–
		K-Means 2	88.3 (4.01)	78.2 (2.13)	–
25	**10**	**MBC**	**91.9 (1.88)**	**98.8 (0.40)**	**7.4 (1.58)**
		K-Means 1	57.7 (3.19)	95.5 (0.54)	–
		K-Means 2	76.0 (4.01)	82.9 (1.16)	–
25	**15**	**MBC**	**99.6 (0.23)**	**99.9 (0.05)**	**3.5 (1.13)**
		K-Means 1	56.5 (3.76)	95.8 (0.57)	–
		K-Means 2	82.4 (4.76)	83.0 (1.83)	–
35	**10**	**MBC**	**97.6 (0.65)**	**99.8 (0.08)**	**6.4 (1.81)**
		K-Means 1	56.5 (3.43)	95.9 (0.48)	–
		K-Means 2	74.0 (3.93)	83.9 (1.15)	–
35	**15**	**MBC**	**99.5 (0.24)**	**99.9 (0.03)**	**3.4 (0.67)**
		K-Means 1	59.3 (4.07)	97.8 (0.27)	–
		K-Means 2	81.1 (4.74)	86.7 (1.48)	–

Table 1: Offline accuracy of MBC and K-Means for simulated data. Standard errors are shown in parentheses.

No. Crises	No. Metrics	Method	Full-data Misclassification	3-period Misclassification	Avg. Time to Identification
15	**10**	**MBC**	**6.7 (3.0)**	**10.7 (4.5)**	**1.31 (0.11)**
		MBC-EX	8 (2.5)	10.7 (4.5)	
15	**15**	**MBC**	**6.7 (5.2)**	**9.3 (6.2)**	**1.13 (0.08)**
		MBC-EX	5.3 (3.9)	8.0 (4.9)	
25	**10**	**MBC**	**13.6 (2.7)**	**15.2 (2.7)**	**1.33 (0.13)**
		MBC-EX	9.6 (2.0)	15.2 (3.4)	
25	**15**	**MBC**	**2.4 (1.6)**	**4.0 (1.8)**	**1.15 (0.06)**
		MBC-EX	3.2 (1.5)	3.2 (1.5)	

Table 2: Online accuracy of MBC and MBC-EX for simulated data. Standard errors are shown in parentheses.

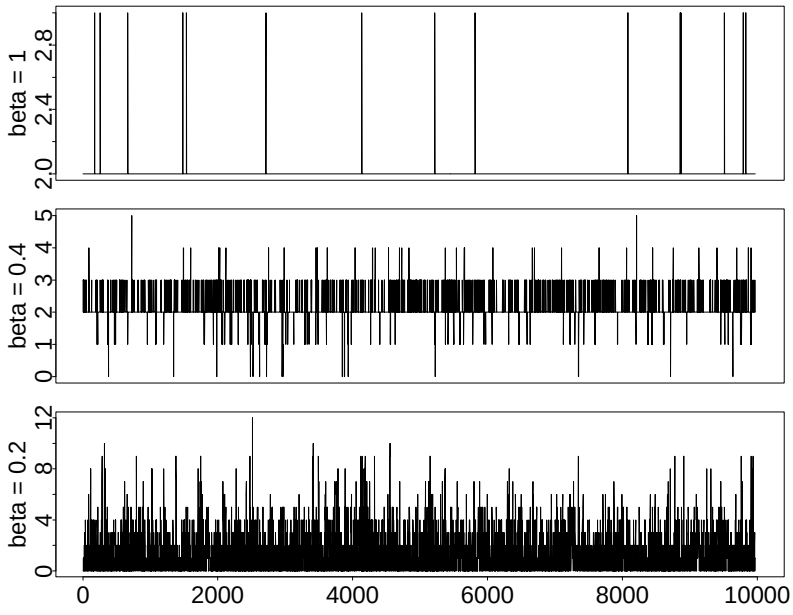

Figure 2: Trace plots of the parallel tempering Markov chain samples of Z_{22}. Three inverse temperatures β are shown; x-axes correspond to the iterations of the Markov chain.

ID	Cause	No. of known crises	No. identified by MBC	No. MBC crises matching known
A	overloaded front-end	2	3	2
B	overloaded back-end	8	14	8
C	database configuration error	1	2	1
D	configuration error	1	0	0 (labeled as A)
F	performance issue	1	0	0 (labeled as B)
G	middle-tier issue	1	0	0 (labeled as K)
I	whole DC turned off and on	1	0	0 (labeled as B)
J	workload spike	1	1	1
K	request routing error	1	6	1

Table 3: EHS crises types. The number of crises known to be of each type is given in column 3. The number of crises identified by MBC as being of this type is given in column 4, and the number of these that correspond to the crises of known type is given in column 5.