Online Model-Based Clustering for Crisis Identification in Distributed Computing

Dawn B. Woodard*
Cornell University
Ithaca, NY 14850

Moises Goldszmidt†
Microsoft Research
Mountain View, CA 94043

Abstract

Large-scale distributed computing systems can suffer from occasional severe violation of performance goals; due to the complexity of these systems, manual diagnosis of the cause of the crisis is too slow to inform interventions taken during the crisis. Rapid automatic recognition of the recurrence of a problem can lead to cause diagnosis and 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 a fast and accurate solution using model-based clustering based on a Dirichlet process mixture; the evolution of each crisis is modeled as a multivariate time series.

In the periods between crises we perform full Bayesian inference for the past crises, and as a new crisis occurs we apply fast approximate Bayesian updating. These inferences allow real-time expected-cost-minimizing decision making that fully accounts for uncertainty in the crisis labels and other parameters. We apply and validate our methods using simulated data and data from a production computing center with hundreds of servers running a 24/7 email-related application.

Keywords: Dirichlet process, mixture model, time series, Bayesian, Monte Carlo, optimal decision.

*Dawn B. Woodard is Assistant Professor in Cornell University's School of Operations Research and Information Engineering, 206 Rhodes Hall, Ithaca, NY 14850 (email: woodard@cornell.edu)

† Moises Goldszmidt is Principal Researcher, Microsoft Research, 1065 La Avenida, Mountain View, CA, 94043 (email: moises@microsoft.com)

1. INTRODUCTION

Commercial distributed computing providers offer remotely hosted processing services. The providers accomplish this computing by farming out to servers that may be spread across geographical and corporate boundaries in centers containing tens of thousands of machines. For instance, Microsoft offers email processing via an Email Hosted Service (EHS), in which incoming messages are routed to servers that apply a set of spam filters before directing remaining emails to the user.

Such systems have performance requirements such as limits on processing times, set in agreements with clients; violation of these limits (a "crisis") leads to cash penalties and potential loss of contracts, so rapid diagnosis and intervention is critical when a violation occurs. Such problems can happen, 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 computing centers on which the servers rely (e.g., for performing authentication services).

When a crisis occurs, we wish to rapidly identify any previous crises of the same type, and take the intervention that has been most effective in the previous occurrences. Due to the large scale, the interdependence and the distributed nature of the systems, problems tend to recur and human diagnosis is very slow, so 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.

We consider the problem of matching a currently occurring (and thus incompletely observed) crisis to previous crises of mixed known and unknown causes. This is an online clustering problem with partial labeling that is complicated by the incompleteness of the data for the new crisis. By online clustering we mean the task of grouping (in real time) observations that arrive in a temporal sequence. Previous work in online crisis/failure identification (Cohen et al. 2005, Yuan et al. 2006, Duan and Babu 2008, Bodik et al. 2009) uses multi-stage approaches combining statistical, machine learning, or ad-hoc methods. While giving practical solutions, they do not provide a complete model for the process of interest. They also restrict to either completely labeled or completely unlabeled data, and do not satisfactorily address the incomplete nature of the new crisis data.

We provide a solution using online model-based clustering, where the evolution of each crisis is modeled as a multivariate time series. In the periods between crises we perform full Bayesian inference for the past crises, and as a new crisis occurs we apply fast approximate Bayesian updating.

A Dirichlet process mixture model (Escobar 1994; Ishwaran and Zarepour 2002) is used for the cluster assignments; this allows us to automatically estimate the number of clusters from the data, and to quantify our uncertainty regarding the number of clusters. Since the posterior distribution 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).

Fully Bayesian inference of this kind is required to perform optimal decision mak-

ing while accounting for uncertainty in the crisis type assignments and the parameters of those types. We describe how to use our Bayesian identification of a new crisis to choose an intervention that minimizes the expected cost of the crisis.

Online clustering based on Dirichlet process or related mixture models has been previously addressed by Sato (2001), Zhang, Ghahramani and Yang (2004), and Gomes, Welling and Perona (2008). These papers focus on clustering very large numbers of observations, motivated by the need to automatically categorize huge volumes of news stories and images; Zhang et al. (2004), for instance, cluster 62,962 documents having 100,000 features. These authors therefore develop fast approximate methods. Zhang et al. (2004) obtain a single "best" cluster assignment for each observation, not updating the cluster assignments of existing observations as new observations arrive, nor quantifying uncertainty regarding the cluster assignment. Sato (2001) and Gomes et al. (2008) use a variational approximation to the posterior distribution, where the approximating distribution is completely factorizable and has a simple parametric form for the marginal distribution of each parameter. Such an approximation can be useful for very large sample sizes, where more precise inference is intractable, but is hard to justify otherwise.

In our context the number of observations is small to moderate, and the focus is on accurate clustering and quantification of uncertainty, including uncertainty regarding the cluster assignments. For this reason we perform fully Bayesian online clustering without resorting to a variational approximation. Our Markov chain method simulates accurately from the posterior distribution, updating the cluster assignments of old observations as more data become available, handling the multimodality of the posterior distribution, and capturing dependencies between parameters.

Our main contribution is to solve an important applied problem by combining Dirichlet process mixture models for time series observations with sophisticated

Markov chain Monte Carlo methods. To our knowledge we are also the first to do fully Bayesian real-time online clustering without resorting to a variational approximation.

We demonstrate the accuracy of our crisis identification method using simulated data and comparing with a state-of-the-art maximum likelihood / maximum a posteriori clustering algorithm; our method is far more accurate in these simulations. Then we apply our method to the Email Hosted Service. Priors for the parameters are obtained by combining information from experts with information in the data, and reflect the fact that the server status measurements are chosen with the goal of being indicative of crisis type.

An alternative to clustering is given by classification methods, which can be applied by using the available labels and creating a category for crises of unknown cause. We do not take this approach, since it is less informative than clustering: it can categorize the new crisis as having unknown type, but cannot identify it as having the same type as several specific unlabeled past crises. In the typical case where few or no labels are available (see Sec. 8), the output of classification approaches is not very meaningful.

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 EHS 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, and are typically aggregated over fixed-length time intervals. EHS handles email traffic, applying a sequence of spam filters, so that some of the measurements are the number of emails that pass each filter, and the number blocked by each filter, during the time interval.

Distributed computing systems have a set of performance goals, defined in agreements with clients. An extended period of violation of these performance goals is considered to be a system crisis. In EHS, for instance, the system is considered to be in violation if at least a predetermined percentage of the servers are above a threshold for a "key performance indicator." 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 server measurements ("metrics") for EHS are shown in Figure 1 for a ten-day period; the median value over the servers is plotted. Crisis periods are highlighted and labeled according to their types, which were diagnosed afterwards. The first two crises are known to have particular causes "A" and "B", while the last four crises are known 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 (and essential) in other distributed computing systems. The number of servers in these systems is typically huge and is increasing at a rapid pace, so it is important to use data summaries that do not grow in dimension with the number of servers.

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 $\mathbf{Y}_{il\cdot} = (\mathbf{Y}_{il1}, \ldots, \mathbf{Y}_{ilJ})$; for crises of type k, we assume that the initial state vector $\mathbf{Y}_{i1\cdot}$ is sampled from a discrete distribution, and that the state vector $\mathbf{Y}_{il\cdot}$ subsequently evolves according to a Markov chain of order q.

Estimation of the full joint distribution of $\mathbf{Y}_{i1\cdot}$ 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 the EHS data $I = 27$ 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 (Domingos and Pazzani 1997; Friedman, Geiger and Goldszmidt 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 the pair of metrics, having three states each, with a single metric that has nine states). Conditional on k, $\mathbf{Y}_{i1j}$ then has a discrete distribution with probability vector $\gamma^{(jk)} = (\gamma_1^{(jk)}, \gamma_2^{(jk)}, \gamma_3^{(jk)})$, and $\mathbf{Y}_{ilj}$ 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 $\mathbf{T}_{st}^{(jk)}$ where the subscripts $s, t \in \{1, 2, 3\}$ indicate the states. The resulting complete-data likelihood function is as follows, where we condition on the unknown type indicators $\mathbf{Z}_i$ 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:

$$\pi\left(\mathcal{D} \mid \{\mathbf{Z}_i\}_{i=1}^{I}, \{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}\right) = \prod_{i,j,t}\left[\left(\gamma_t^{(j\mathbf{Z}_i)}\right)^{\mathbf{1}(\mathbf{Y}_{i1j}=t)} \prod_{s}\left(\mathbf{T}_{st}^{(j\mathbf{Z}_i)}\right)^{n_{ijst}}\right]. \quad (1)$$

For simplicity we will use π to indicate likelihood, prior, and posterior distributions, as distinguished by their arguments.

3.2 Cluster Modeling

The Dirichlet process mixture (DPM) model provides natural prior specification for online clustering, allowing estimation of the number of clusters while maintaining exchangeability between observations (Escobar and West 1995). A DPM can be obtained as the limit of a finite mixture model with Dirichlet prior distribution on the mixing proportions (Neal 2000; Rasmussen 2000). In our context the DPM is parameterized by a scalar α controlling the expected number of types occurring in a set of crises, and by a prior distribution $G_0(\{\gamma^{(j\cdot)}, \mathbf{T}^{(j\cdot)}\}_j)$ for the set of all parameters associated with each crisis type k.

We take G_0 to be the product over j of independent Dirichlet distributions for $\gamma^{(jk)}$ (with parameter vectors $\mathbf{a}^{(j)}$), times the product over j and s of independent Dirichlet distributions for the transition matrix rows $\mathbf{T}_{s\cdot}^{(jk)}$ (with parameter vectors $\mathbf{b}_s^{(j)}$). The use of such a product Dirichlet prior distribution for the rows of an unconstrained transition matrix is standard practice (e.g. Carlin, Gelfand and Smith (1992), Diaconis and Rolles (2006)).

The DPM model for the crisis types $\{\mathbf{Z}_i\}_{i=1}^I$ and crisis parameters $\gamma^{(jk)}$, $\mathbf{T}^{(jk)}$ can be described as follows, in the case where the causes of the crises are all unknown. The first crisis has type 1 ($\mathbf{Z}_i = 1$) by definition. For subsequent crises,

$$\Pr(\mathbf{Z}_i = z|\mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}) = \frac{\#\{i' < i : \mathbf{Z}_{i'} = z\}}{i - 1 + \alpha} \qquad \text{for } z \in \{\mathbf{Z}_{i'}\}_{i'<i}$$
$$\Pr(\mathbf{Z}_i \neq \mathbf{Z}_{i'}\ \forall i' < i|\mathbf{Z}_1, \ldots, \mathbf{Z}_{i-1}) = \frac{\alpha}{i - 1 + \alpha}.$$

This is 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.

Conditional on the cluster assignments $\mathbf{Z}_i$, the parameters of each cluster k are independently distributed according to G_0. Thus the DPM model can be written

$$\begin{aligned}\pi\left(\{\mathbf{Z}_i\}_{i=1}^I\right) &= \pi(\mathbf{Z}_1)\prod_{i=2}^{I}\pi\left(\mathbf{Z}_i|\{\mathbf{Z}_{i'}\}_{i'<i}\right) \\ &= \prod_{i=1}^{I}\left[\frac{\alpha}{\alpha+i-1}\mathbf{1}\left(\mathbf{Z}_i = m_{i-1}+1\right) + \frac{1}{\alpha+i-1}\sum_{i'<i}\mathbf{1}\left(\mathbf{Z}_i = \mathbf{Z}_{i'}\right)\right] \qquad (2)\end{aligned}$$

$$\pi\left(\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}|\{\mathbf{Z}_i\}_{i=1}^I\right) = \prod_{k=1}^{m_I} G_0\left(\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_j\right). \qquad (3)$$

where $m_i = \max\{\mathbf{Z}_{i'} : i' \leq i\}$ for $i > 0$ and $m_0 = 0$. In this description the Dirichlet process has been integrated out, obtaining a generalized Polya urn scheme (Blackwell and MacQueen 1973).

When the causes of some of the crises are known (the partially labeled case), this information can be captured by indicator functions $\mathbf{1}(\mathbf{Z}_i = \mathbf{Z}_{i'})$ for pairs of crises i, i' that are known to have the same type (denoted by $i \sim i'$) and $\mathbf{1}(\mathbf{Z}_i \neq \mathbf{Z}_{i'})$ for pairs of crises i, i' that are known to have different types (denoted by $i \not\sim i'$). In this case the prior $\pi(\{\mathbf{Z}_i\}_{i=1}^I)$ is proportional to the expression (2) multiplied by the restriction

$$\prod_{i \sim i'} \mathbf{1}(\mathbf{Z}_i = \mathbf{Z}_{i'}) \prod_{i \not\sim i'} \mathbf{1}(\mathbf{Z}_i \neq \mathbf{Z}_{i'}) \tag{4}$$

while the prior $\pi(\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k} | \{\mathbf{Z}_i\}_{i=1}^I)$ is unchanged. Our computational methods extend trivially to accommodate the partial labels, by simply disallowing configurations that are incompatible with those labels (see Section 4).

3.3 Choice of prior constants

We select the prior hyperparameters α, $\mathbf{a}^{(j)}$, and $\mathbf{b}_s^{(j)}$ 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). Sensitivity to these choices is examined in Section 7.1.

We elicit α from experts; results are remarkably insensitive to the choice of α (Sec. 6.1 & 7.1). According to the DPM model, the probability that two randomly chosen crises have the same type is $1/(\alpha + 1)$. In the EHS example, domain experts estimate this to be 10%, yielding $\alpha = 9$. This implies that the expected number of types in the EHS data is 12.9 (for 27 crises), which the experts agree is reasonable.

For $\mathbf{a}^{(j)}$ and $\mathbf{b}_s^{(j)}$ one might consider the "default" choice of $\mathbf{a}_t^{(j)} = \mathbf{b}_{st}^{(j)} = 1$ for all j, s, and t; this gives a generalized uniform prior for each of the vectors $\gamma^{(jk)}$ and each of the rows $\mathbf{T}_{s\cdot}^{(jk)}$ of the transition matrices, and is a common choice when performing Bayesian inference on discrete distributions or transition matrices when the number of parameters is fixed (e.g. Carlin et al. 1992). In our context this prior is very diffuse relative to the likelihood (explained below). Unfortunately, diffuse priors can be

problematic when comparing models with parameter spaces of different dimensions, as is the case in the clustering context where we are comparing different values of the number of clusters. In particular, such priors can dramatically favor the model with the smallest number of parameters (cf. Kass and Raftery 1995). We encountered this problem when attempting to use the generalized uniform prior for the EHS data or simulated data that mimicked the EHS data; all of the crises were assigned to a single cluster with very high posterior probability. The generalized uniform prior is not used for document clustering with DPMs, perhaps for this reason; it is replaced by either a data-based prior specification or a symmetric Dirichlet distribution with parameter less than one (Blei, Griffiths, Jordan and Tenenbaum 2004; Zhang et al. 2004).

To explain why the generalized uniform prior is very diffuse relative to the likelihood in the distributed computing context, we note that real data from distributed computing have very specific properties that are simply not compatible with much of the parameter space. For instance, the metrics more commonly take the value 2 ("normal") than the values 1 ("low") or 3 ("high"); also, the metrics change values infrequently in the time series, so that any reasonable transition matrix $\mathbf{T}^{(jk)}$ must have diagonal elements close to one.

To create a sensible prior for $\mathbf{a}^{(j)}$, we first follow Zhang et al. (2004) in taking the prior mean of the vector $\gamma^{(jk)}$ to be the empirical frequencies over the entire dataset, i.e. the empirical distribution $\gamma^* = (\gamma_1^*, \gamma_2^*, \gamma_3^*)$ of the first value of all metrics in all crises observed so far. This implies that $\mathbf{a}^{(j)} = c^{(j)}(\gamma_1^*, \gamma_2^*, \gamma_3^*)$ for some constant $c^{(j)}$ that controls the prior variance of $\gamma^{(jk)}$. Our choice of this constant will 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. For this reason the EHS experts believe that there is a substantial prior probability for any j and k that one of the values $\gamma_t^{(jk)}$ is "close" to one. This is formalized by specifying

a 50% prior probability that $\max_t \gamma_t^{(jk)} > .9$ (results are insensitive to this choice; Sec 7.1). This choice uniquely determines the value of $c^* = c^{(j)}$, which can be found by an iterative numerical procedure that simulates from Dirichlet($c\gamma^*$), checks the proportion of samples for which one of the values is "close" to one as defined above, and adjusts c.

Selection of $\mathbf{b}_s^{(j)}$ is analogous. We use the data for all metrics and all crises observed so far to find the empirical transition probabilities $\mathbf{T}^*_{s\cdot}$ from each starting state s, and set the prior mean of $\mathbf{T}^{(jk)}_{s\cdot}$ equal to $\mathbf{T}^*_{s\cdot}$. By taking the prior weight of evidence (the sum of $\mathbf{b}_s^{(j)}$) to be equal for each row s of the transition matrix, we must have $\mathbf{T}^{(jk)} = d^{(j)}\mathbf{T}^*$ for some constant $d^{(j)}$. To choose $d^{(j)}$, we consider the limiting distribution $r^{(jk)} = (r_1^{(jk)}, r_2^{(jk)}, r_3^{(jk)})$ of a Markov chain with transition kernel $\mathbf{T}^{(jk)}$. 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% probability that $\max_t r_t^{(jk)} > .9$. Once again, $d^* = d^{(j)}$ is found by simulation.

4. POSTERIOR COMPUTATION

For a fixed set of crises, Markov chain methods can be used to obtain samples from the posterior distribution $\pi(\{\mathbf{Z}_i\}_{i=1}^I, \{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}|\mathcal{D})$ of the clustering model given in Sec. 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 $\pi(\{\mathbf{Z}_i\}_{i=1}^I|\mathcal{D})$ on the reduced space $\{\mathbf{Z}_i\}_{i=1}^I$; this is possible by integrating out the cluster-specific parameters, in our case $\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}$. Posterior samples for the cluster parameters can then be obtained by sampling from their conditional posterior distribution; details are in the Appendix.

Theoretical and empirical support for the use of collapsed-space Gibbs samplers

is given in Liu (1994). The author shows that the operator norm of a collapsed-space Gibbs sampler is less than or equal to that of the corresponding full-space Gibbs sampler (Liu 1994, Thm.1). This result applies to our context, giving evidence that a Gibbs sampler over the parameters $\{\mathbf{Z}_i\}_{i=1}^I$ is likely more efficient than a sampler on the full space that alternates Gibbs updates of $\{\mathbf{Z}_i\}_{i=1}^I$ with Gibbs updates of the cluster-specific parameters. Liu (1994) also finds improved empirical performance of a collapsed-space Gibbs sampler over the full-space Gibbs sampler in a context that has similarities to ours.

Collapsed-space sampler. The basic collapsed-space Gibbs sampler for DPMs consists of Gibbs updates of each $\mathbf{Z}_i$ in turn; this method was first used by Neal (1992) and Escobar (1994). In order to address the potential multimodality of the posterior distribution, Jain and Neal (2004) add a Metropolis-Hastings move that merges two clusters into one or splits a cluster into two. They 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 narrow and well-separated modes corresponding to distinct cluster assignments. Here even the collapsed-space sampler with split-merge moves can have difficulty moving between the modes. Additionally, convergence diagnostics can be difficult to apply; the parameters are the cluster indicators $\mathbf{Z}_i$ 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. Standard convergence diagnostics such as Geweke's diagnostic or the Gelman-Rubin diagnostic are degenerate when applied to a constant-valued time series (Cowles and Carlin 1996), so one could only apply convergence diagnostics to non-constant $\mathbf{Z}_i$ or to non-constant summaries of the parameter vector $\{\mathbf{Z}_i\}_{i=1}^I$, such as the log-likelihood.

Parallel tempering. In order to 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, Schmidler and Huber 2009). It simulates parallel Markov chains indexed by $m = 1, \ldots, M$ using identical updating strategies but distinct target distributions ϕ_m; we take $\phi_m(\{\mathbf{Z}_i\}_{i=1}^I) \propto \pi(\{\mathbf{Z}_i\}_{i=1}^I)\pi(\mathcal{D}|\{\mathbf{Z}_i\}_{i=1}^I)^{\beta_m}$ where $\pi(\mathcal{D}|\{\mathbf{Z}_i\}_{i=1}^I)$ is the likelihood of the data conditioned only on the cluster assignments, and where $0 \leq \beta_1 \leq \ldots \leq \beta_M = 1$. 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_m interpolate between ϕ_1 and the posterior $\phi_M = \pi(\{\mathbf{Z}_i\}_{i=1}^I|\mathcal{D})$. 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 $\prod_{m=1}^M \phi_m$. The samples from chain M converge marginally to the posterior ϕ_M, and can be used for Monte Carlo inference.

The "inverse temperatures" β_m are chosen as follows. We take $\beta_1 = 0$, and select the smallest set of inverse temperatures β_m that gives swap acceptance rates of at least 20%; theoretical and empirical results to support this choice are given in Atchadé, Roberts and Rosenthal (2009).

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 $\mathbf{Z}_i$ takes only a single value at the lowest temperature ($\beta_M = 1$), $\mathbf{Z}_i$ takes many values at the higher temperatures (β_m small), 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 $\pi(\{\mathbf{Z}_i\}_{i=1}^I)$.

Handling partial labeling. Extending the above method to the partially labeled case is simple; first, the Markov chain should be initialized in a configuration that satisfies the restrictions given in (4). When simulating the Markov chain, the parameters $\mathbf{Z}_i$ for crises that are known to have the same type are updated as a single parameter, which enforces the restriction that $\mathbf{Z}_i = \mathbf{Z}_{i'}$ for all $i \sim i'$. To enforce the restriction that $\mathbf{Z}_i \neq \mathbf{Z}_{i'}$ for all $i \not\sim i'$, only moves that satisfy this restriction should be considered in the Gibbs and Metropolis-Hastings updates.

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 the probabilities $\pi(\mathbf{Z}_{new} = \mathbf{Z}_i|\mathcal{D}, \mathcal{D}_{new})$ that the new crisis has the same type as each previous crisis $i = 1, \ldots, I$ and the probability $\pi(\mathbf{Z}_{new} \neq \mathbf{Z}_i \ \forall i|\mathcal{D}, \mathcal{D}_{new})$ of this being a new type of crisis ($\mathbf{Z}_{new}$ is an indicator of the type of the new crisis).

5.1 Exact Prediction

To perform inference for $\mathbf{Z}_{new}$ 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. The Markov chain then has parameter vector $(\{\mathbf{Z}_i\}_{i=1}^I, \mathbf{Z}_{new})$ and limiting distribution equal to the posterior distribution $\pi(\{\mathbf{Z}_i\}_{i=1}^I, \mathbf{Z}_{new}|\mathcal{D}, \mathcal{D}_{new})$. Call the iterations of this chain $(\{\mathbf{Z}_i^{(l)}\}_{i=1}^I, \mathbf{Z}_{new}^{(l)})$ for $l = 1, \ldots, L$.

Then we have the following Monte Carlo estimates for $\pi(\mathbf{Z}_{new} = \mathbf{Z}_i|\mathcal{D}, \mathcal{D}_{new})$ and

$\pi(\mathbf{Z}_{new} \neq \mathbf{Z}_i \ \forall i | \mathcal{D}, \mathcal{D}_{new})$:

$$\hat{\pi}(\mathbf{Z}_{new} = \mathbf{Z}_i | \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}(\mathbf{Z}_{new}^{(l)} = \mathbf{Z}_i^{(l)}) \tag{5}$$

$$\hat{\pi}(\mathbf{Z}_{new} \neq \mathbf{Z}_i \ \forall i | \mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L} \sum_{l=1}^{L} \mathbf{1}(\mathbf{Z}_{new}^{(l)} \neq \mathbf{Z}_i^{(l)} \ \forall i).$$

The right-hand side of (5), for instance, is simply the proportion of samples from the Markov chain for which the new crisis is in the same cluster as the ith crisis. Since the Markov chain converges to the posterior distribution, we have that $\hat{\pi}(\mathbf{Z}_{new} = \mathbf{Z}_i | \mathcal{D}, \mathcal{D}_{new})$ converges to $\pi(\mathbf{Z}_{new} = \mathbf{Z}_i | \mathcal{D}, \mathcal{D}_{new})$ as $L \to \infty$ (cf. Tierney, 1994).

The more computationally intensive part of this approach is simulating the Markov chain; having obtained these samples the Monte Carlo computation in (5) is nearly instantaneous. The above-described approach is practical when the number of past crises is small (for $I = 15$, $J = 10$ simulating the Markov chain for 10^5 iterations takes less than 10 minutes on a standard processor), but after many crises 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:

$$\begin{aligned} \pi(\mathbf{Z}_{new} = \mathbf{Z}_i | \mathcal{D}, \mathcal{D}_{new}) &= \sum_{\{\mathbf{Z}_{i'}\}_{i'=1}^{I}} \pi(\mathbf{Z}_{new} = \mathbf{Z}_i | \{\mathbf{Z}_{i'}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \pi(\{\mathbf{Z}_{i'}\}_{i'=1}^{I} | \mathcal{D}, \mathcal{D}_{new}) \\ &\approx \sum_{\{\mathbf{Z}_{i'}\}_{i'=1}^{I}} \pi(\mathbf{Z}_{new} = \mathbf{Z}_i | \{\mathbf{Z}_{i'}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \pi(\{\mathbf{Z}_{i'}\}_{i'=1}^{I} | \mathcal{D}) \end{aligned} \tag{6}$$

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

The conditional distribution $\pi(\mathbf{Z}_{new} | \{\mathbf{Z}_{i'}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$ of $\mathbf{Z}_{new}$ is expressed as:

$$\pi(\mathbf{Z}_{new} | \{\mathbf{Z}_{i'}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}) \propto \pi(\mathbf{Z}_{new} | \{\mathbf{Z}_{i'}\}_{i'=1}^{I}) \pi(\mathcal{D}, \mathcal{D}_{new} | \mathbf{Z}_{new}, \{\mathbf{Z}_{i'}\}_{i'=1}^{I}) \tag{7}$$

where $\pi(\mathcal{D}, \mathcal{D}_{new}|\mathbf{Z}_{new}, \{\mathbf{Z}_{i'}\}_{i'=1}^{I})$ is available in closed form as shown in the Appendix, and where (from the Dirichlet process mixture model in (2))

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

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 $\{\mathbf{Z}_i^{(l)}\}_{i=1}^{I}$ from the posterior distribution $\pi(\{\mathbf{Z}_i\}_{i=1}^{I}|\mathcal{D})$.

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

$$\hat{\pi}(\mathbf{Z}_{new} = \mathbf{Z}_i|\mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L}\sum_{l=1}^{L} \pi(\mathbf{Z}_{new} = \mathbf{Z}_i^{(l)}|\{\mathbf{Z}_{i'}^{(l)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$$

$$\hat{\pi}(\mathbf{Z}_{new} \neq \mathbf{Z}_i\ \forall i|\mathcal{D}, \mathcal{D}_{new}) = \frac{1}{L}\sum_{l=1}^{L} \pi(\mathbf{Z}_{new} \neq \mathbf{Z}_i^{(l)}\ \forall i|\{\mathbf{Z}_{i'}^{(l)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new}).$$

   For each l the discrete distribution $\pi(\mathbf{Z}_{new}|\{\mathbf{Z}_{i'}^{(l)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$ is first computed, by enumerating (7) over the possible values $k = 1, \ldots, m_I + 1$ of $\mathbf{Z}_{new}$ and normalizing. Then $\pi(\mathbf{Z}_{new} = \mathbf{Z}_i^{(l)}|\{\mathbf{Z}_{i'}^{(l)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$ and $\pi(\mathbf{Z}_{new} \neq \mathbf{Z}_i^{(l)}\ \forall i|\{\mathbf{Z}_{i'}^{(l)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$ are found by choosing the appropriate element of the vector $\pi(\mathbf{Z}_{new}|\{\mathbf{Z}_{i'}^{(l)}\}_{i'=1}^{I}, \mathcal{D}, \mathcal{D}_{new})$.

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 $O(LIJ)$, 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 (conditioning only on the data and not on particular parameter estimates), can only be done using fully Bayesian inference (Robert 2001).

A cost function specifies the total cost of a crisis as a function of the true crisis type and the intervention taken. Taking an intervention that quickly resolves a crisis gives low total cost, while taking an intervention 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 $C\left[\phi, \left(\{\mathbf{Z}_i^*\}_{i=1}^I, \mathbf{Z}_{new}^*\right)\right]$ of the intervention ϕ and the entire vector of true crisis types $\left(\{\mathbf{Z}_i^*\}_{i=1}^I, \mathbf{Z}_{new}^*\right)$, due to the fact that $\mathbf{Z}_{new}^*$ is only meaningful in the context of $\{\mathbf{Z}_i^*\}_{i=1}^I$. If we knew C, and given posterior sample vectors $\left(\{\mathbf{Z}_i^{(l)}\}_{i=1}^I, \mathbf{Z}_{new}^{(l)}\right)$ as in Section 5.1, the expected cost of taking ϕ during the new crisis could be estimated consistently as

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

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

The expected-cost-minimizing intervention is the value of ϕ that minimizes $\mathbf{E}(C)$. Although the cost function C is not known in practice, for interventions ϕ that have been taken during previous crises the realized costs can be used to estimate C, and for other interventions 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 intervention. For this reason we will avoid choosing a particular estimate of the crisis types $\left(\{\mathbf{Z}_i\}_{i=1}^I, \mathbf{Z}_{new}\right)$, and instead will consider the accuracy of the soft identification, i.e. the posterior distribution over $\left(\{\mathbf{Z}_i\}_{i=1}^I, \mathbf{Z}_{new}\right)$ as given in Sec. 5.1 & 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 clustering algorithm described in Sec. 3 & 4 to a fixed set of crises. Then we consider accuracy of online clustering.

6.1 Offline Accuracy

We examine offline accuracy, varying the number of crises and metrics and comparing to an alternative model-based clustering algorithm detailed in Fraley and Raftery (2002) and extended and supported in numerous subsequent analyses (Fraley, Raftery and Wehrens 2005; Li, Fraley, Bumgarner, Yeung and Raftery 2005; Raftery and Dean 2006). We also have compared our method with a distance-based clustering method, which fared extremely poorly due to the difficulty of choosing an appropriate distance metric; those results are not reported here.

The clustering method of Fraley and Raftery (2002) fits a finite mixture model via maximum likelihood using the expectation-maximization (EM) algorithm (Dempster, Laird and Rubin 1977). The initial clustering is obtained using model-based hierarchical agglomerative clustering (cf. Banfield and Raftery 1993), and the number of clusters is chosen using the Bayesian Information Criterion (Schwarz 1978). The method of Fraley and Raftery (2002), like ours, yields a probabilistic ("soft") assignment of observations to clusters, but unlike our method chooses a particular number of clusters and particular values for the cluster parameters, rather than obtaining a posterior distribution over these quantities.

We sample I crises of equal length M by first drawing the number of clusters from a uniform distribution on the integers from 1 to I. Conditional on the number of clusters we sample the vector of cluster probabilities from a generalized uniform distribution, then sample the cluster indicators according these probabilities (unrepresented clusters are dropped, reducing the number of clusters). We sample the

cluster parameters $\gamma^{(jk)}, \mathbf{T}^{(jk)}$ from finite Dirichlet / product Dirichlet distributions, then simulate the metrics for each crisis from the time series model given in Sec. 3.1.

In order to simulate data that are as realistic as possible, we take the hyperparameter values $\mathbf{a}^{(j)}, \mathbf{b}_s^{(j)}$ for the finite Dirichlet distributions to be those obtained for the EHS data as described in Section 3.3, and take the crisis length to be the median length in the EHS data (8 time periods).

Twenty datasets are simulated for each of several combinations of I and J, and the following measures of accuracy are obtained for our method ("DPM") and for the method of Fraley and Raftery (2002) ("ML-BIC"):

1. **Pairwise Sensitivity:** Of the pairs of crises that are of the same type, the percentage that have 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 have 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. For DPM, the posterior mean is used to estimate the number of types.

These measures have been used, e.g., in Booth, Casella and Hobert (2008).

Values of the accuracy measures are reported in Table 1, averaged over the simulated datasets and along with their standard errors (when applying DPM we have chosen α so that the prior mean # clusters is $I/4$, although multiplying or dividing α by five gives virtually identical results). The accuracy of DPM is better than ML-BIC by all measures, and dramatically better in terms of both pairwise sensitivity and estimating the number of clusters. The performance of ML-BIC degrades substantially as the number of metrics increases.

The main problem is that EM in this context rarely changes the cluster assignments from their initial values; the final cluster assignments are probabilistic, but typically place nearly all probability mass on the initial cluster assignments. This means that the initial cluster assignment provided by hierarchical clustering corresponds to a local mode of the likelihood function, so that EM stays in that local mode. In our context there can be few observations per cluster and the observations have moderate to high dimensionality; this leads to many such local modes of the likelihood function.

We altered the ML-BIC algorithm in several ways to attempt to fix this problem, without success. First, we smoothed the initialization of EM, placing only 80% of the prior mass for each observation on the cluster assignment from hierarchical clustering, and the remainder on other cluster assignments. Second, we smoothed the surface over which the maximization is performed, by placing a prior on the parameters of the finite mixture model and maximizing the posterior density instead of the likelihood, as suggested by Fraley and Raftery (2002). We used two prior specifications for the cluster-specific parameters $\gamma^{(jk)}$ and $\mathbf{T}^{(jk)}$: the generalized uniform prior ("MAP-UNIF") and the prior described in Section 3.3 ("MAP-INFO"). For the cluster probabilities (conditional on the number of clusters) we used a generalized uniform prior. The resulting algorithms perform better than ML-BIC but still far worse than DPM for all values of I and J. Results for MAP-UNIF are shown in Table 1; the accuracy of MAP-INFO was only slightly better than ML-BIC and is not reported.

While the performance of ML-BIC and its variants degrades as the number of metrics increases, the performance of DPM improves by all measures in this simulation. More metrics means both more noise in the data and more information available to estimate the clusters. While the performance of the EM-based methods suffers due to

overfitting the noise, DPM succeeds in taking advantage of the additional information.

The number of crises in the data does not appear to have a strong effect on the accuracy of DPM. This helps explain the excellent accuracy that we find in the online context (Section 6.2), where the number of crises starts small and increases over time.

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 estimate the type of each crisis based on the data from the previous crises and partial data for the new crisis. Due to the poor performance of the expectation-maximization approaches in the offline context, we do not consider these methods further. We instead compare predictive accuracy of the approximate method given in Section 5.2 ("DPM") to that of the exact method in Section 5.1 ("DPM-EX"), in order to justify the approximation. For data simulated as in Section 6.1, we evaluate several measures of accuracy for DPM and DPM-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 $\hat{\pi}(\mathbf{Z}_{new} \neq \mathbf{Z}_i \ \forall i | \mathcal{D}, \mathcal{D}_{new}) > 0.5$ if $\mathbf{Z}_{new} \neq \mathbf{Z}_i \ \forall i$ according to the gold standard (here, the truth), and otherwise that $\hat{\pi}(\mathbf{Z}_{new} = \mathbf{Z}_i | \mathcal{D}, \mathcal{D}_{new}) > 0.5$ for some $i \leq I$ such that $\mathbf{Z}_{new} = \mathbf{Z}_i$ 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 DPM-EX, since this is extremely computationally intensive. The average values of the above accuracy

measures over five simulated datasets are shown in Table 2 for several combinations of I and J.

The accuracy is high for both DPM and DPM-EX, correctly classifying over 80% of crises in every setting we considered. The performance of DPM is not significantly worse than that of DPM-EX, showing the accuracy of the approximation given in Section 5.2. There is some evidence that using more metrics shortens the average time to identification and 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 DPM is between one and two time periods. Such early identification of a crisis is extremely helpful in choosing an appropriate intervention.

7. APPLICATION TO THE EMAIL HOSTED SERVICE

27 crises occurred in EHS during the first four months of 2008. The causes of some of these crises have been diagnosed by EHS experts, and are listed in the third column of Table 3. For confidentiality reasons, the letters assigned to these crisis types do not correspond with the letters shown in Figure 1. All of the EHS data used here are free from human interventions.

Preprocessing. We choose a subset of the available metrics by applying the feature selection procedure in Bodik, Goldszmidt, Fox and Andersen (2009). 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 by the EHS experts 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 Sections 3 and 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 $\mathbf{Z}_{22}$ for several values of the inverse temperature β are shown. The chain with $\beta = 1$, which is designed to draw from the posterior distribution $\pi(\{\mathbf{Z}_i\}_{i=1}^{I}|\mathcal{D})$, primarily visits the single value $\mathbf{Z}_{22} = 2$, while the chains with smaller values of β visit progressively more values of $\mathbf{Z}_{22}$. 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 $\mathbf{Z}_i : i \neq 1$ at each inverse temperature (omitting cases where $\mathbf{Z}_i$ was constant for the entire chain). Similarly, we obtain a maximum Gelman-Rubin scale factor of 1.01, again evidence of good convergence. This maximum is taken over $\mathbf{Z}_i : i \neq 1$ for inverse temperatures β less than 0.5 (for $\beta \geq 0.5$ there are numerical difficulties, since some $\mathbf{Z}_i$ take a single value for almost all iterations).

The sizes of the clusters from the posterior mode cluster assignment are shown in the fourth column of 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 I 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 our method 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 computing center, one on which the servers depend. The EHS technicians do not have authority to fix the performance of that separate computing center, explaining why this is the most common type of crisis. It is also the most important type of crisis to correctly identify, since although the problem cannot be fixed, the technicians know the best intervention for minimizing the effect of such a 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 (e.g., closely related 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. This result suggests to computing center operators the need for additional metrics to distinguish between these crisis types.

In the other cases of incorrectly merged crisis types, while the large majority of metrics are indistinguishable between the two types a few metrics show distinct behavior. Since we have assumed the parameters of distinct crisis types to be inde-

pendent 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 be fixed by creating a hierarchical dependence structure between crisis types in the prior distribution. This structure would be realistic, since in fact one can define the true crisis "causes" according to a coarse division into a few main causes, or into finely sub-divided causes.

The tendency to merge small clusters with larger clusters could also be due to our use of the Dirichlet process, under which the model for the cluster assignments has a single parameter α. A more flexible approach would be to use the Pitman-Yor process (Pitman and Yor 1997), which is a generalization of the Dirichlet process that has two parameters for the cluster assignment model and a richer ability to represent the distribution of cluster sizes. We expect that the methods given in this paper would generalize naturally to this process, and leave this as future work.

Sensitivity to Prior Specification. Results are very insensitive to the choice of α; multiplying or dividing α by up to a factor of 20 does not change the posterior mode cluster assignment. Results are also not sensitive to the choice of the constants $\mathbf{a}^{(j)}$ and $\mathbf{b}_s^{(j)}$, as long as this choice is not dramatically inconsistent with the data and with the experts' belief that the metrics provide substantial information about the crisis type. For instance, changing the prior median of $\max_t \gamma_t^{(jk)}$ and $\max_t r_t^{(jk)}$ from 0.90 to 0.99 or 0.85 does not change the posterior mode clustering assignment. Reducing this value to 0.80 switches the ranks of the two most probable cluster assignments, but the total posterior probability of these two cluster assignments is still high (89.9%). Reducing it further to 0.70 gives the same results as 0.80 except that the posterior probability of these two most probable cluster assignments declines to 66.7%. Note that even with this large change in the prior, the accuracy results reported in Table 3, column 5 are unchanged.

We examine sensitivity to the prior mean of $\gamma^{(jk)}$ and $\mathbf{T}_{s\cdot}^{(jk)}$ by taking the data-based prior mean described in Section 3.3 and mixing with a uniform distribution. When the mixture proportions are 3/4 on the data-based mean and 1/4 on the uniform distribution, the posterior mode clustering assignment is unchanged. When we reduce the weight on the data-based mean to 1/2, the ranks of the two most probable cluster assignments are switched, but they still account for 90.3% of the posterior probability. Reducing it further to 1/4, clusters A and H become merged. Thus the prior has to change dramatically before results change in a meaningful way.

7.2 Online Application

We evaluate the accuracy of online clustering for the EHS data, relative to the offline clustering assignment. We apply the online crisis identification method given in Sec. 5.2 and evaluate the accuracy measures described in Sec. 6.2, treating the posterior mode cluster assignment from the offline context (Sec. 7.1) as the gold standard.

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 even before the technical start of the crisis. Two-thirds of the crises are identified correctly in the first time period.

To check whether the good identification performance of our method is specific to the particular ordering of the crises, we also permute the crises and evaluate performance. 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). These results are even better than for the original ordering.

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 intervention. Accuracy has been demonstrated on both simulated data and data from a production system (EHS); our method dramatically outperforms a state-of-the-art maximum likelihood / maximum a posteriori clustering method in the offline setting, and sees very little loss of accuracy in the online setting relative to the offline setting.

Importantly, our method provides natural solutions to several related problems; these are explored in Goldszmidt and Woodard (2010). 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 and suggest promising interventions. 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.

The above uses give our approach an advantage over another potential alternative: directly learning a mapping from the metrics to the best intervention. Such a mapping avoids an explicit model for the metrics, and so cannot be used for any of these related purposes (which are essential to the operators of such systems).

Here we have used a parsimonious model for the crises, due to the small sample sizes available for inference. Such small sample sizes are characteristic of crisis identification in the environment of large-scale distributed computing supporting internet services. The system as a whole undergoes frequent updates and occasional configu-

ration changes. This is due in part to the constant addition of features and in part to changes required for long-term crisis resolution. The result is that the learning process has to be restarted often, so that there are almost never more than fifty relevant past crises, and typically labels are available for only a minority. Thus, although one could theoretically allow the model to become more complex as the sample size increases, in practice one almost never reaches sample sizes large enough for this to be useful.

9. ACKNOWLEDGEMENTS

The authors thank the editors and reviewers for their helpful suggestions that led to many improvements in this article. Many thanks to Hans Andersen and David LaFaurie (Microsoft) for their help in understanding the EHS system and the data. Thanks also to Peter Bodick for his help in the initial exploratory data analysis and for many useful conversations on the topic of this article. This research was partially supported by the National Science Foundation (CMMI-0926814).

APPENDIX: MARKOV CHAIN MONTE CARLO COMPUTATIONS

The likelihood of the data conditioned only on $\{\mathbf{Z}_i\}_{i=1}^I$ is:

$$\pi(\mathcal{D}|\{\mathbf{Z}_i\}_{i=1}^I) = \int \pi\left(\mathcal{D} \mid \{\mathbf{Z}_i\}_{i=1}^I, \{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}\right) \pi\left(\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k} \mid \{\mathbf{Z}_i\}_{i=1}^I\right) d\gamma^{(jk)} d\mathbf{T}^{(jk)}$$

Using (1) and (3) and by multinomial-Dirichlet conjugacy (cf. Gelman, Carlin, Stern, and Rubin 2004),

$$\pi(\mathcal{D}|\{\mathbf{Z}_i\}_{i=1}^I) = \prod_{k=1}^{m_I}\prod_j \left[\frac{\Gamma\left(\sum_t \mathbf{a}_t^{(j)}\right)\prod_t \Gamma\left(\mathbf{a}_t^{(j)} + \sum_{i:\mathbf{Z}_i=k} \mathbf{1}(\mathbf{Y}_{i1j}=t)\right)}{\Gamma\left(\sum_t\left[\mathbf{a}_t^{(j)} + \sum_{i:\mathbf{Z}_i=k}\mathbf{1}(\mathbf{Y}_{i1j}=t)\right]\right)\prod_t \Gamma\left(\mathbf{a}_t^{(j)}\right)}\right] \times$$
$$\prod_{k=1}^{m_I}\prod_{j,s}\left[\frac{\Gamma\left(\sum_t \mathbf{b}_{st}^{(j)}\right)\prod_t \Gamma\left(\mathbf{b}_{st}^{(j)} + \sum_{i:\mathbf{Z}_i=k} n_{ijst}\right)}{\Gamma\left(\sum_t\left[\mathbf{b}_{st}^{(j)} + \sum_{i:\mathbf{Z}_i=k} n_{ijst}\right]\right)\prod_t \Gamma\left(\mathbf{b}_{st}^{(j)}\right)}\right]. \quad (8)$$

The posterior distribution of $\{\mathbf{Z}_i\}_{i=1}^I$ is proportional to the product of $\pi(\{\mathbf{Z}_i\}_{i=1}^I)$ and $\pi(\mathcal{D}|\{\mathbf{Z}_i\}_{i=1}^I)$, given in (2) and (8), respectively. A Markov chain can then be constructed to sample on this reduced space. For instance, a Gibbs sampler for $\{\mathbf{Z}_i\}$ updates each $\mathbf{Z}_i$ conditional on $\mathbf{Z}_{[-i]} = \{\mathbf{Z}_{i'}\}_{i'\neq i}$. The posterior distribution of $\mathbf{Z}_i$ conditional on $\mathbf{Z}_{[-i]}$ is proportional to $\pi(\{\mathbf{Z}_i\}_{i=1}^I|\mathcal{D})$; computation consists of enumerating over the possible values of $\mathbf{Z}_i$ and normalizing to obtain the conditional distribution. The possible options are that $\mathbf{Z}_i$ is equal to one of the values in $\mathbf{Z}_{[-i]}$, or that it is not equal to any of the values in $\mathbf{Z}_{[-i]}$. 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 posterior samples of $\{\mathbf{Z}_i\}_{i=1}^I$ by simulating such a Markov chain, we can also obtain posterior samples of $\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}$, by noticing that

$$\pi(\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k}|\{\mathbf{Z}_i\}_{i=1}^I, \mathcal{D}) = \prod_{k=1}^{m_I}\prod_j \text{Dirichlet}((\gamma_1^{(jk)}, \gamma_2^{(jk)}, \gamma_3^{(jk)}); \hat{\mathbf{a}}^{(j)}) \times$$
$$\prod_{k=1}^{m_I}\prod_{j,s} \text{Dirichlet}((\mathbf{T}_{s1}^{(jk)}, \mathbf{T}_{s2}^{(jk)}, \mathbf{T}_{s3}^{(jk)}); \hat{\mathbf{b}}_s^{(j)})$$

where $\hat{\mathbf{a}}_t^{(j)} = \mathbf{a}_t^{(j)} + \sum_{i:\mathbf{Z}_i=k} \mathbf{1}(\mathbf{Y}_{i1j} = t)$, $\hat{\mathbf{b}}_{st}^{(j)} = \mathbf{b}_{st}^{(j)} + \sum_{i:\mathbf{Z}_i=k} n_{ijst}$ for $t = 1, 2, 3$, and where $\text{Dirichlet}((\gamma_1^{(jk)}, \gamma_2^{(jk)}, \gamma_3^{(jk)}); \hat{\mathbf{a}}^{(j)})$ is the finite Dirichlet density for $\gamma^{(jk)}$ with parameter vector $\hat{\mathbf{a}}^{(j)}$. For each posterior sample of $\{\mathbf{Z}_i\}_{i=1}^I$, generate one sample

from $\pi(\{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k} | \{\mathbf{Z}_i\}_{i=1}^I, \mathcal{D})$; this gives joint posterior samples of the full set of parameters $(\{\mathbf{Z}_i\}_{i=1}^I, \{\gamma^{(jk)}, \mathbf{T}^{(jk)}\}_{j,k})$.

REFERENCES

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

Banfield, J. D., and Raftery, A. E. (1993), "Model-based Gaussian and non-Gaussian clustering," *Biometrics*, 49, 803–821.

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

Blei, D. M., Griffiths, T. L., Jordan, M. I., and Tenenbaum, J. B. (2004), "Hierarchical topic models and the nested Chinese restaurant process," in *Advances in Neural Information Processing Systems*, eds. S. Thrun, L. Saul, and B. Schölkopf, MIT Press, Cambridge, MA.

Bodik, P., Goldszmidt, M., Fox, A., Woodard, D. B., and Andersen, H. (2009), "Fingerprinting the datacenter: automated classification of performance crises," in *EuroSys 2010*, ed. G. Muller.

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 *Proc. of the 6th Symposium on Operating Systems Design and Implementation*, eds. E. Brewer and P. Chen, pp. 231–244.

Cohen, I., Zhang, S., Goldszmidt, M., Symons, J., Kelly, T., and Fox, A. (2005), "Capturing, indexing, clustering, and retrieving system history," in *Proc. of the 20th ACM Symposium on Operating System Principles*, eds. A. Herbert and K. P. Birman, pp. 105–118.

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.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977), "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

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*, eds. J. Strassner and S. Dobson, 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.

Fraley, C., and Raftery, A. E. (2002), "Model-based clustering, discriminant analysis, and density estimation," *Journal of the American Statistical Association*, 97, 611–631.

Fraley, C., Raftery, A., and Wehrens, R. (2005), "Incremental model-based clustering for large datasets with small clusters," *Journal of Computational and Graphical Statistics*, 14, 529–546.

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. (2010), "Bayesian inference for crisis characterization in distributed computing," Unpublished manuscript.

Gomes, R., Welling, M., and Perona, P. (2008), "Incremental learning of nonparametric Bayesian mixture models," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, eds. K. Boyer, M. Shah, and T. Syeda-Mahmood.

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

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.

Kass, R. E., and Raftery, A. E. (1995), "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

Li, Q., Fraley, C., Bumgarner, R. E., Yeung, K. Y., and Raftery, A. E. (2005), "Donuts, scratches and blanks: Robust model-based segmentation of microarray images," *Bioinformatics*, 21, 2875–2882.

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. (1992), "Markov chain sampling methods for Dirichlet process mixture models," in *Proc. of the 11th International Workshop on Maximum Entropy and Bayesian Methods of Statistical Analysis*, eds. C. R. Smith, G. J. Erickson, and P. O. Neudorfer, Kluwer Academic Publishers, pp. 197–211.

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

Pitman, J., and Yor, M. (1997), "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, 25, 855–900.

Raftery, A. E., and Dean, N. (2006), "Variable selection for model-based clustering," *Journal of the American Statistical Association*, 101, 168–178.

Rasmussen, C. E. (2000), "The infinite Gaussian mixture model," in *Advances in Neural Information Processing Systems*, eds. S. A. Solla, T. K. Leen, and K. R. Muller, pp. 554–560.

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

Sato, M. (2001), "Online model selection based on the variational Bayes," *Neural Computation*, 13, 1649–1681.

Schwarz, G. (1978), "Estimating the dimension of a model," *Annals of Statistics*, 6, 461–464.

Tierney, L. (1994), "Markov chains for exploring posterior distributions (with discussion)," *Annals of Statistics*, 22, 1701–1762.

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*, ed. W. Zwaenepoel, pp. 375–388.

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*, ed. Y. Weiss.

Zhang, S., Cohen, I., Goldszmidt, M., Symons, J., and Fox, A. (2005), "Ensembles of models for automated diagnosis of system performance problems," in *Proc. of the International Conference on Dependable Systems and Networks*, ed. A. Bondavalli, pp. 644–653.

Table 1: Offline accuracy of DPM and EM-based methods for simulated data.

No. Crises	No. Metrics	Method	Pairwise Sensitivity	Pairwise Specificity	% Error No. Types
15	**10**	**DPM**	**96.6 (1.45)**	**99.5 (0.29)**	**5.3 (1.22)**
		ML-BIC	54.0 (5.21)	98.0 (0.54)	77.4 (27.96)
		MAP-UNIF	58.6 (5.14)	97.8 (0.57)	77.4 (27.96)
15	**15**	**DPM**	**98.5 (0.90)**	**99.9 (0.05)**	**8.9 (3.71)**
		ML-BIC	39.8 (4.81)	99.9 (0.10)	113.0 (32.97)
		MAP-UNIF	49.6 (5.80)	99.5 (0.23)	113.0 (32.97)
25	**10**	**DPM**	**94.6 (2.49)**	**99.8 (0.10)**	**7.6 (1.62)**
		ML-BIC	59.1 (4.78)	98.6 (0.31)	24.2 (6.11)
		MAP-UNIF	67.1 (4.89)	97.1 (0.90)	24.2 (6.11)
25	**15**	**DPM**	**99.7 (0.32)**	**99.7 (0.19)**	**2.7 (0.84)**
		ML-BIC	40.9 (4.11)	99.8 (0.07)	86.0 (15.0)
		MAP-UNIF	57.6 (5.14)	99.8 (0.10)	86.0 (15.0)
35	**10**	**DPM**	**93.1 (1.43)**	**99.6 (0.09)**	**8.2 (1.68)**
		ML-BIC	61.2 (4.04)	98.0 (0.24)	35.0 (9.81)
		MAP-UNIF	68.5 (4.07)	97.8 (0.29)	35.0 (9.81)
35	**15**	**DPM**	**97.9 (0.95)**	**99.9 (0.06)**	**3.0 (0.60)**
		ML-BIC	46.2 (3.56)	99.7 (0.09)	51.8 (9.81)
		MAP-UNIF	52.1 (3.77)	99.5 (0.20)	51.8 (9.81)

NOTE: Accuracies are averaged over 10 datasets, with standard errors shown in parentheses.

Table 2: Online accuracy of DPM and DPM-EX for simulated data.

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

NOTE: Accuracies are averaged over five datasets, with standard errors shown in parentheses.

Table 3: Crises types in data from Microsoft's EHS computing center.

ID	Cause	No. of known crises	No. identified by DPM	No. DPM 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)
E	performance issue	1	0	0 (labeled as B)
F	middle-tier issue	1	0	0 (labeled as I)
G	whole DC turned off and on	1	0	0 (labeled as B)
H	workload spike	1	2	1
I	request routing error	1	6	1

NOTE: The number of crises known to be of each type is given in column 3. The number of crises identified by DPM 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.

Figure 1: Traces of several metrics for Microsoft's EHS computing center over a period of ten days; crisis periods are highlighted and labeled according to known type.

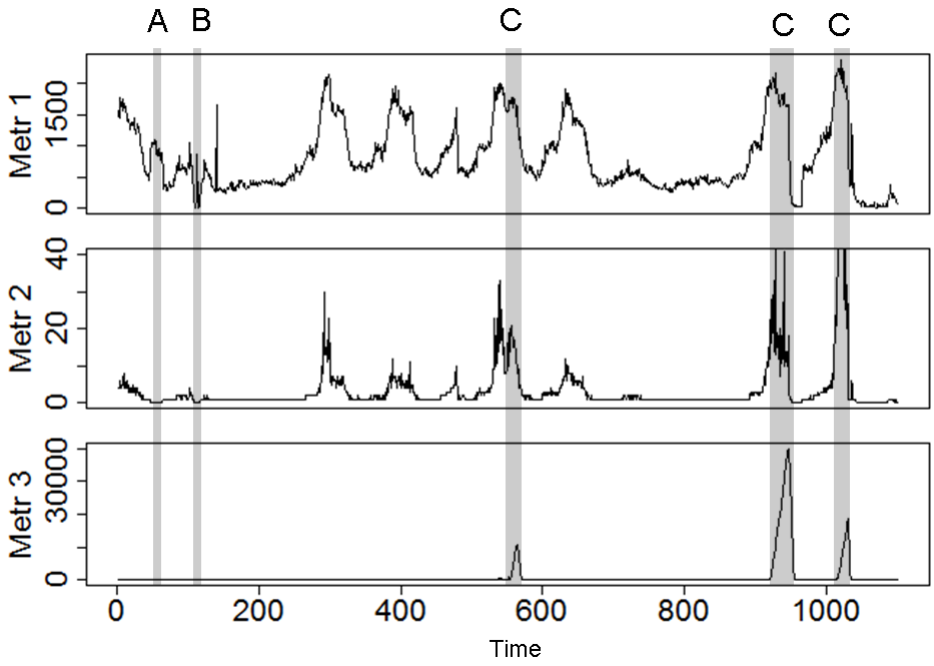

Figure 2: Trace plots of the parallel tempering Markov chain samples of $\mathbf{Z}_{22}$. Three inverse temperatures β are shown; x-axes correspond to the (post-thinning) iterations of the Markov chain.