This is a repository copy of Accurate Analysis of Quality Properties of Software with Observation-Based Markov Chain Refinement.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/113451/

Version: Accepted Version

Proceedings Paper:

Reuse
Items deposited in White Rose Research Online are protected by copyright, with all rights reserved unless indicated otherwise. They may be downloaded and/or printed for private study, or other acts as permitted by national copyright laws. The publisher or other rights holders may allow further reproduction and re-use of the full text version. This is indicated by the licence information on the White Rose Research Online record for the item.

Takedown
If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.
Accurate Analysis of Quality Properties of Software with Observation-Based Markov Chain Refinement

Colin Paterson and Radu Calinescu
Department of Computer Science, University of York, UK

Abstract—We introduce a tool-supported method for the automated refinement of continuous-time Markov chains (CTMCs) used to assess quality properties of component-based software. Existing research focuses on improving the efficiency of CTMC analysis and on identifying new applications for this analysis. As such, ensuring that the analysis is accurate by using CTMCs that closely model the behaviour of the analysed software has received relatively little attention. Our new method addresses this gap by refining the high-level CTMC model of a component-based software system based on observations of the execution times of its components. Our refinement method reduced analysis errors by 77–90.3% for a service-based system implemented using six public web services from three different providers, improving the accuracy of the analysis and significantly reducing the risk of invalid software engineering decisions.

I. INTRODUCTION

Software performance and reliability engineering [1] uses mathematical models to predict performance, reliability and other quality properties of software [2]. These models include Petri Nets [3], queuing networks [4] and Markov chains [5], [6], [7], and may be built manually or extracted automatically from more general software models such as UML-MARTE [8] or Palladio [9].

Despite major advances in this research area, it is still very challenging to ensure that such mathematical models are sufficiently accurate to support the design and verification of real systems. Our paper addresses this challenge for continuous-time Markov chains (CTMCs), a type of stochastic state-transition models used to analyse quality software properties both at design time [5], [10], [11] and at runtime [12], [13], [14], [15]. We introduce an observation-based Markov chain refinement (OMNI) method that significantly improves the accuracy with which quality software properties can be analysed. OMNI starts from a high-level abstract CTMC whose states correspond to operations performed by different components of the analysed system, and uses observed execution times for the components to refine this CTMC. The high-level CTMC can be generated (e.g., from annotated UML activity diagrams as in [10]) or can be provided by the software engineers. The execution time observations can be obtained by unit testing components individually before the software system is built or, for existing systems, from their logs.

OMNI comprises two stages. The first stage makes the CTMC more realistic through the addition of states and transitions that model the fact that software components have non-zero minimum execution times. We use additional states and transitions corresponding to Erlang distributions [16] for this purpose. The CTMC is then further refined in the second OMNI stage by using phase-type distributions [17] to model the variations in the execution time of the components. The refined CTMC supports the accurate analysis of a wide range of quality software properties expressed in continuous stochastic logic [18], preventing many invalid design decisions and verification conclusions associated with traditional CTMC analysis. Moreover, the refined CTMCs can be analysed using existing probabilistic model checkers such as PRISM [19].

To support the use of OMNI, we describe rigorous techniques for synthesising its Erlang and phase-type distributions. These distributions model the execution of software components with far greater accuracy than the exponential distributions from existing CTMC modelling, which match only the first moment of the unknown distributions of the execution time observations. In addition, we provide a tool that implements the OMNI method, producing refined CTMC models that can be analysed with the probabilistic model checker PRISM [19]. Finally, we present a case study that shows how software engineers can use OMNI to avoid multiple invalid design decisions suggested by traditional CTMC analysis. To ensure the reproducibility of our results, we provide the models, code and data from this work on our project webpage www-users.cs.york.ac.uk/cap/OMNI/.

The remainder of the paper is organised as follows. Sections II and III introduce the theoretical background for our work and a running example that we use to motivate and illustrate the refinement method, respectively. We then present the OMNI method in Section IV and the tool that we implemented to automate its use in Section V. Section VI evaluates the effectiveness of our method through a case study that uses OMNI to analyse the service-based system from the running example, and Section VII discusses related work. We analyse threats to validity in Section VIII, and present our conclusions and future work directions in Section IX.

II. PRELIMINARIES

Continuous Time Markov Chains—Markov chains are mathematical models for stochastic processes evolving in time [20]. A continuous-time stochastic process \( \{X(t) : t \geq 0\} \) with countable state space \( S \) is a continuous-time Markov chain (CTMC) if it has the Markov property

\[
\Pr(X(t_{n+1}) = s_{n+1} \mid X(t_n) = s_n, X(t_{n-1}) = s_{n-1}, \ldots) = \Pr(X(t') = s' \mid X(t) = s),
\]

\( 1 \)
CSL formulae are interpreted over a CTMC whose states are quality properties analysed over CTMCs [18], [21]. Model checkers used to analyse CTMCs in software performance engineering (e.g., PRISM [19], OMNI) uses continuous stochastic logic (CSL) extended with rewards to specify the quality properties analysed over CTMCs [18], [21].

**Definition 1:** A continuous-time Markov chain is a tuple
\[(S, \pi, R),\] (2)
where \(S\) is a finite set of states, \(\pi : S \rightarrow [0, 1]\) is an initial-state probability vector such that the probability that the CTMC is initially in state \(s_i\) is given by \(\pi(s_i)\), and \(R : S \times S \rightarrow \mathbb{R}\) is a transition rate matrix such that, for any states \(s_i \neq s_j\) from \(S\), \(R(s_i, s_j) \geq 0\) specifies the rate with which the CTMC transitions from state \(s_i\) to state \(s_j\), and \(R(s_i, s_i) = -\sum_{s_j \in S \setminus \{s_i\}} R(s_i, s_j)\).

We use the notation \(\text{CTMC}(S, \pi, R)\) for the CTMC (2). The probability that this CTMC will transition from state \(s_i\) to another state within \(t\) time units is \(1 - e^{-t \sum_{s_k \in S \setminus \{s_i\}} R(s_i, s_k)}\), and the probability that the new state is \(s_j \in S \setminus \{s_i\}\) is
\[p_{ij} = R(s_i, s_j) / \sum_{s_k \in S \setminus \{s_i\}} R(s_i, s_k).\] (3)

**Continuous Stochastic Logic—**Similar to the probabilistic model checkers used to analyse CTMCs in software performance engineering (e.g., PRISM [19], OMNI) uses continuous stochastic logic (CSL) extended with rewards to specify the quality properties analysed over CTMCs [18], [21].

**Definition 2:** Let \(AP\) be a set of atomic propositions, \(a \in AP\), \(p \in [0, 1]\), \(I\) an interval in \(\mathbb{R}\) and \(\bowtie \in \{\geq, >, <, \leq\}\). Then a state formula \(\Phi\) and a path formula \(\Psi\) in CSL are defined by the following grammar:
\[
\Phi ::= true \mid a \mid \Phi \land \Phi \mid \neg \Phi \mid P_{\bowtie p}[\Psi] \mid S_{\bowtie p}[\Psi]. \tag{4}
\]

CSL formulae are interpreted over a CTMC whose states are labelled with atomic propositions from \(AP\) by a function \(L : S \rightarrow 2^{AP}\). Path formulae only occur inside the probabilistic operator \(P\) and steady-state operator \(S\), which define bounds on the probability of system evolution, e.g., a state \(s\) satisfies a formula \(P_{\bowtie p}[\Psi]\) if the probability of the future evolution of the system meets the bound \(\bowtie p\). For a path, the “next” formula \(X\Phi\) holds if \(\Phi\) is satisfied in the next state; the “bounded until” formula \(\Phi_1 U^T \Phi_2\) holds if before \(\Phi_2\) becomes true at time \(t \in I\), \(\Phi_1\) is satisfied continuously in the interval \([0, t)\).

If \(I = [0, \infty)\), the formula is termed “unbounded until”. The notation \(F^T \Phi \equiv true U^T \Phi\) is used when the first part of an until formula is true, and \(s \models \Phi\) and \(M \models \Phi\) indicate that \(\Phi\)

<table>
<thead>
<tr>
<th>Label</th>
<th>Third-party service</th>
<th>URL</th>
<th>rate ((s^{-1}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>location</td>
<td>Bing location service</td>
<td><a href="http://dev.virtualearth.net/REST/v1/Locations/">http://dev.virtualearth.net/REST/v1/Locations/</a></td>
<td>9.62</td>
</tr>
<tr>
<td>arrivals</td>
<td>Thales rail arrival board</td>
<td><a href="http://www.livedepartureboards.co.uk/ldbws/">http://www.livedepartureboards.co.uk/ldbws/</a></td>
<td>19.88</td>
</tr>
<tr>
<td>departures</td>
<td>Thales rail departures board</td>
<td><a href="http://www.livedepartureboards.co.uk/ldbws/">http://www.livedepartureboards.co.uk/ldbws/</a></td>
<td>19.46</td>
</tr>
<tr>
<td>search</td>
<td>Bing web search</td>
<td><a href="https://api.datamarket.azure.com/Bing/Search/v1/">https://api.datamarket.azure.com/Bing/Search/v1/</a></td>
<td>1.85</td>
</tr>
<tr>
<td>weather</td>
<td>WebserviceX.net weather service</td>
<td><a href="http://www.webservicex.net/globalweather.asmx?op=GetWeather">http://www.webservicex.net/globalweather.asmx?op=GetWeather</a></td>
<td>1.11</td>
</tr>
<tr>
<td>traffic</td>
<td>Bing traffic service</td>
<td><a href="http://dev.virtualearth.net/REST/v1/Traffic/">http://dev.virtualearth.net/REST/v1/Traffic/</a></td>
<td>2.51</td>
</tr>
</tbody>
</table>

where \(s_1, \ldots, s_{n-1}, s_n, s_{n+1} \in S\) and \(0 \leq t_1 \leq \ldots \leq t_{n-1} \leq t_n \leq t_{n+1}, n \geq 1\), is any sequence of \(n + 1\) times. For the work presented in this paper, we use the following CTMC definition adapted from [21].

Motivate our work, we consider a travel web application that needs to use the third-party services from Table I in order to handle two types of requests:

1. Requests from users who plan to meet and entertain a visitor arriving by train. These requests are expected to occur with probability \(p_1\).
2. Requests from users looking for a possible destination for a day trip by train. These requests are expected to occur with probability \(1 - p_1\).

The high-level abstract CTMC from Fig. 1 (derived from an activity diagram) models the handling of a request by the application. The initial state \(s_1\) corresponds to finding the location of the train station. For the first request type, this is followed by finding a possible destination \((s_2)\), obtaining suitable restaurants in the area \((s_3)\), obtaining a traffic report for the route from the user’s location to the station \((s_4)\), and returning the response to the user \((s_7)\). For the second request type, state \(s_1\) is followed by finding a possible destination \((s_3)\), obtaining a weather forecast for this destination \((s_5)\), and returning the response to the user \((s_7)\). Once a suitable destination is selected, the traffic report is obtained \((s_6)\) and the response is returned to the user \((s_7)\). The operation execution rates \(\lambda_1\) to \(\lambda_6\) depend on the components used for these operations, and the engineers want to decide if the real services from Table I are suitable for building the application. If they are, the engineers need: (i) to select appropriate request-handling times to be promised in the
application service-level agreement (SLA); and (ii) to choose a pricing scheme for the application. Therefore, the engineers are interested to assess the following quality properties of the application variant built using these services:

**P1** The fraction of user requests handled in under $T$ seconds, for $0 < T \leq 4$.

**P2** The fraction of “day trip” requests handled in under $T$ seconds, for $0 < T \leq 4$.

**P3** The expected profit per request handled, assuming that 1 cent is charged for requests handled within $T$ seconds for $0 < T \leq 3$.

Service response times are assumed exponentially distributed in CTMC modelling, so the engineers use observed response times $t_{i1}, \ldots, t_{in}$ for service $i$ (taken from existing logs or obtained through testing) to estimate the service rate $\lambda_i$ as

$$\lambda_i = \left( \frac{t_{i1} + t_{i2} + \cdots + t_{in}}{n} \right)^{-1}.$$  \hspace{1cm} (5)

Finally, they use the CTMC to analyse the CSL-formalised properties **P1**–**P3**:

\[
\begin{align*}
\textbf{P1} & : P_{\pi} = \mathbb{P}[^{[0, T]}_\text{complete}] \\
\textbf{P2} & : P_{\pi} = \mathbb{P}[^{[0, T]}_\text{complete}] - \mathbb{P}^{(3, \infty)}_\text{complete} \\
\textbf{P3} & : P_{\pi} = \mathbb{P}^{(0, T)}_\text{complete} - 2 \cdot P_{\pi} \cdot \mathbb{P}^{(3, \infty)}_\text{complete}
\end{align*}
\]

where $0 < T \leq 4$ for **P1** and **P2**, and $0 < T \leq 3$ for **P3**. To replicate this process, we built a prototype version of the application and used it to handle 270 randomly generated requests for $p_1 = 0.3$ and $p_2 = 0.1$. We obtained sample execution times for each service (between 81 for arrivals and search and 270 for location and traffic), and used (5) to calculate the estimate service rates in Table I. We then used the model checker PRISM [19] to analyse the CTMC for these rates, and thus to predict the values of properties (6). To assess the accuracy of the predictions, we calculated the actual values of these properties using detailed timing information logged by our application. The predictions obtained through CTMC analysis and the actual property values are compared in Fig. 2. The errors reported in the figure give the area difference between the actual and predicted property values:

$$\text{error} = \int_0^{T_{\text{max}}} |\text{actual}(T) - \text{predicted}(T)| \, dT,$$  \hspace{1cm} (7)

where $T_{\text{max}} = 4$ for properties **P1** and **P2**, and $T_{\text{max}} = 3$ for **P3**. We will later use these errors to measure the accuracy improvements due to our OMNI model refinement. For now, recall that the engineers must make decisions based only on the predicted property values; two such decisions could be:

- Implement the application with the services from Table I, with an SLA promising that 40% of the requests will be handled within 1s (property **P1**), 35% of the “day trip” requests will be handled also within 1s (property **P2**), and charge 1 cent for requests handled within 1s (property **P3**). This decision would be wrong, as both promises would be violated by a wide margin, and the actual profit would be under a third of the predicted profit (cf. Fig. 2).

- Look for alternative services for the application, because not even 80% of the requests or “day trip” requests are handled within 2s, and/or because the profit is below 0.7 cents per request when charging 1 cent for each request handled within 2s. This decision would also be wrong, since all the constraints it is based on would actually be satisfied by the application (Fig. 2).

Clearly, we chose the times and bounds above so as to show that the current practice of using idealised CTMC models may lead to blatantly invalid decisions. Using other times and bounds will yield valid decisions. However, we argue that engineering decisions must be consistently valid, and not down to chance.

IV. CTMC REFINEMENT METHOD

Let CTMC($S, \pi, R$) be a high-level CTMC model of a system such that:

- each state $s_i \in S$ corresponds to operation $i$ of the system and $\pi(s_i)$ is the probability that this is the initial operation;
- for any $s_i \neq s_j$ from $S$, $R(s_i, s_j) = p_{ij} \lambda_i$, where $p_{ij}$ is the (known or estimated as in [22], [23]) probability (3) that operation $i$ is followed by operation $j$, and $\lambda_i$ is calculated as in (5), using $n > 0$ observed execution times $t_{i1}, t_{i2}, \ldots, t_{in}$ of operation $i$.

This CTMC model makes the typical assumption that operation execution times are exponentially distributed. However, this assumption is almost always invalid for two reasons. First, each software operation $i$ has a minimum execution time $t_i > 0$ such that its probability of completion within $t_i$ time units is zero. Second, even the “holding times”
Fig. 3: Accurate analysis of quality properties with OMNI

t_{i1} - \xi, t_{i2} - \xi, \ldots, t_{in} - \xi of an operation i are rarely exponentially distributed. As we showed in Section III, these issues can lead to erroneous analyses of the quality properties of the modelled system.

OMNI overcomes these issues by producing a refined CTMC whose analysis with existing model checkers supports the accurate evaluation of quality properties of software (Fig. 3). This is achieved through the use of phase-type distributions (PHDs) that accurately model the variations in execution times for each component. The PHD fitting of distributions with deterministic delays is known to require extremely large numbers of states, and such delays are best fitted by Erlang distributions [24], as we confirm experimentally in Section VI. As such, the OMNI refinement comprises two stages. In the first stage, the high-level CTMC is refined to better model the minimum execution times of software operations. We term this stage delay extraction. In the second stage, the CTMC is further refined through the use of PHD fitting. We call this stage holding-time modelling. The two OMNI stages are described next.

A. Stage I: Delay Extraction

In this OMNI stage, the CTMC is extended with additional states and transitions that model the minimum execution times (i.e. delays) of the software operations by means of Erlang distributions, i.e., sums of several independent exponential distributions with the same rate [16]. With the notation above, state \( s_i \) and its \( m_i \geq 1 \) outgoing transitions with rates \( p_{1i}\lambda_i, p_{2i}\lambda_i, \ldots, p_{mi}\lambda_i \) are replaced (Fig. 4) with a sequence of delay-modelling states \( s_{i1}, s_{i2}, \ldots, s_{ik} \) that encode an Erlang-\( k_i \) distribution with rate \( \lambda_{Ei}^i \), and a state \( s_i' \) with outgoing transitions of rates \( p_{1i}\lambda_i', p_{2i}\lambda_i', \ldots, p_{mi}\lambda_i' \) to the same next states as \( s_i \). However, delays are not modelled perfectly by Erlang distributions: for any error \( \epsilon \in (0, 1) \), there is a (small) probability \( p \) that the refined CTMC leaves state \( s_{ik} \) within \( L_i(1-\epsilon) \) time units of entering \( s_{i1} \). Given specific values for \( \lambda_i \) and \( p \), the theorem below supports the calculation of the parameters \( k_i, \lambda_{Ei}^i \) and \( \lambda_i' \) for our delay extraction.

**Theorem 1:** Given an error bound \( \epsilon \in (0, 1) \), if the delay-extraction refinement parameters \( k_i, \lambda_{Ei}^i \) and \( \lambda_i' \) satisfy

\[
1 - \sum_{l=0}^{k_i-1} \frac{(k_i(1-\epsilon))^l}{l!} e^{-k_i(1-\epsilon)} = p, \tag{8}
\]

\[
\lambda_{Ei}^i \frac{k_i}{L_i} = \lambda_i \text{ and } \lambda_i' = \frac{\lambda_i}{1 - \lambda_i L_i} \tag{9}
\]

for some value \( p \in (0, 1) \) then the following properties hold for the refined CTMC:

(i) The probability that the CTMC leaves state \( s_{ik} \) within \( L_i(1-\epsilon) \) time units from entering state \( s_{i1} \) is \( p \);

(ii) The expected time for the CTMC to leave \( s_{i1} \) after entering state \( s_{ik} \) within \( L_i(1-\epsilon) \) time units is \( p \).

**Proof:** To prove (i), we note that the cumulative distribution function of an Erlang-\( k \) distribution with rate \( \lambda \) is

\[
F(k, \lambda, x) = 1 - \sum_{l=0}^{k-1} \frac{(\lambda x)^l}{l!},\]

so (8) can be rewritten as

\[
F(k_i, \lambda_{Ei}^i, L_i(1-\epsilon)) = p \quad \text{since} \quad k_i = \lambda_{Ei}^i L_i \quad \text{according to (9)}.
\]

Therefore, the probability that the Erlang delay model from Fig. 4 will transition from entering state \( s_{i1} \) to exiting state \( s_{ik} \) within \( L_i(1-\epsilon) \) time units is \( p \). For part (ii), the expected time for the CTMC to leave state \( s_i' \) after entering state \( s_{i1} \) is the sum of the mean of the Erlang-\( k_i \) distribution with rate \( \lambda_{Ei}^i \) and the mean of the exponential distribution with rate \( \lambda_i' \), i.e.,

\[\frac{\lambda_{Ei}^i}{\lambda_i} + \frac{1}{\lambda_i'} = L_i + \frac{1}{\lambda_i'} - L_i = \lambda_i^{-1} \quad \square \]

Thus, we can calculate the delay model parameters for operation \( i \) as follows:

1. Approximate the minimum execution time for operation \( i \) as \( L_i = \min_{j=1}^{m_i} t_{ij} \);
2. Choose a small error \( \epsilon \in (0, 1) \) and a small probability \( p \) (e.g., \( \epsilon = 0.1 \) and \( p = 0.05 \)), and solve (8) for \( k_i \), e.g., by using a numeric solver and rounding the result up to an integer value or—since \( k_i \) only depends on \( \epsilon \) and \( p \), and is independent of \( L_i \)—by using a table of precomputed \( k_i \) values as in Table II;
3. Calculate \( \lambda_{Ei}^i \) and \( \lambda_i' \) using (9).

The theorem below gives the format of the refined CTMC after the delay extraction stage. For convenience, we consider that the delay extraction procedure was applied to all states of the initial model CTMC(S,S,ż,R), which involves setting \( k_i = 0 \) and \( \lambda_i' = \lambda_i \) in the Erlang delay model from Fig. 4(b) for states \( s_i \in S \) that do not require delay extraction (e.g. state \( s_7 \) from our CTMC in Fig. 1).

**Theorem 2:** Applying the OMNI delay extraction procedure to a high-level model CTMC(S,S,ż,R) yields the refined model CTMC(S',S',ż,R'), where:

\[S' = S_z \cup S_{s_{i1}, s_{i2}, \ldots, s_{ik}}, \quad \pi'(s_i') = \pi(s_i) \quad \text{and} \quad \pi'(s_{i1}) = \cdots = \pi'(s_{ik}) = 0, \forall s_i \in S; \]

\[R'(s_{ik}, s_{ik+1}) = R'(s_{ik}, s_i) = \lambda_{Ei}^i, 1 \leq k < k_i, \forall s_i \in S; \]

\[R'(s_i', s_j) = p_{ij}\lambda_i' \quad \forall s_i \in S \quad \forall s_j \in S; \]

\[R'(s_i', s_j') = 0 \quad \forall s_j \in S \setminus \{s_i\}, 1 \leq k < k_i, 1 \leq l \leq k; \]

\[R'(s_i', s') = -\sum_{s_i' \in S \setminus \{s_i'\}} R'(s_i', s'), \forall s' \in S'. \]

**Proof:** The proof is by construction, cf. Fig. 4. \quad \square
by synthesising a PHD that models the “holding times” sample as the distribution of the time to absorption in a CTMC continuous distribution with a strictly positive density evaluation [17] and can approximate arbitrarily close any phase-type distributions.

Stage II: Holding-Time Modelling

Errors. Nevertheless, there are still time points three properties, reducing the margin for engineering decision error (7) decreased significantly (by between 49%–70%) for all predicted actual and predicted property values differ noticeably, e.g., the predicted P1 and P2 values for T = 1.4s are 0.565 (instead of an actual value of 0.723) and 0.462 (instead of 0.674), respectively. The next OMNI stage addresses this difference.

Example 1 Consider again the web application from our motivating example, and its three quality properties from (6). We applied the OMNI delay extraction to each state associated with a service in the CTMC from Fig. 1. We set ε = 0.1 and p = 0.05, so state s_i, 1 ≤ i ≤ 6, from the original CTMC was replaced with an Erlang delay model as in Fig. 4(b), where k_i = 259 (cf. Table II) and the values for λ'^{i}_{s} and λ'^{j}_{s} are given in Table III. Accordingly, the refined CTMC had 6 × 259 more states and transitions than the original CTMC. Fig. 5 compares the actual values of properties P1–P3 with the values predicted through the analysis of the refined CTMC. As anticipated, the error (7) decreased significantly (by between 49%–70%) for all three properties, reducing the margin for engineering decision errors. Nevertheless, there are still time points T where the actual and predicted property values differ noticeably, e.g., the predicted P1 and P2 values for T = 1.4s are 0.565 (instead of an actual value of 0.723) and 0.462 (instead of 0.674), respectively. The next OMNI stage addresses this difference.

B. Stage II: Holding-Time Modelling

We further refine the CTMC from the first OMNI stage by using phase-type distributions (PHDs) to model the “holding times” of the system operations. A PHD [17] is defined as the distribution of the time to absorption in a CTMC with one absorbing state, i.e., the time to reach the only state s_i ∈ S of a CTMC (2) for which R(s_i, s_j) = 0 for all s_j ∈ S. PHDs support efficient numerical and analytical evaluation [17] and can approximate arbitrarily close any continuous distribution with a strictly positive density in (0, ∞) [25]. OMNI exploits these advantages of PHDs by synthesising a PHD that models the “holding times” sample

\[ \text{sample}_i = (t'_{i1} = t_{i1} - t_{i1}, t'_{i2} = t_{i2} - t_{i1}, \ldots, t'_{in} = t_{in} - t_{i1}) \] (10)

for each operation i, and replacing state s'_i of the Erlang delay model from Fig. 4(b) with the CTMC associated with this PHD. Before describing the OMNI PHD synthesis algorithm and replacement procedure for state s'_i, we need to introduce several basic concepts about PHDs. First, consider the transition rate matrix R of a CTMC (S, π, R) with one absorbing state and N ≥ 1 transient (i.e., non-absorbing) states called the phases of the PHD. Without loss of generality, we will assume that the last row of R corresponds to the absorbing state, i.e.

\[ R = \begin{bmatrix} D_0 & d_1 \\ 0 & 0 \end{bmatrix} \] (11)

where the N × N sub-matrix D_0 specifies only transition rates between transient states, 0 is a 1 × N row vector with zero elements, and d_1 is an N × 1 vector whose elements specify the transition rates from the transient states to the absorbing state. The elements from each row of R add up to zero (cf. Definition 1), so we additionally have D_01 + d_1 = 0, where 1 and 0 are column vectors of N ones and N zeros, respectively. Thus, d_1 = −D_01 and the PHD associated with this CTMC is fully defined by the sub-matrix D_0 and the row vector π_0 containing the first N elements of the initial probability vector π (as in most practical applications, we are only interested in PHDs that are acyclic and that cannot start in the absorbing state). We use the notation PHD(π_0, D_0) for this PHD.

The fitting of phase-type distributions to empirical data received considerable attention [17], [26], with effective PHD fitting algorithms developed based on techniques such as moment matching [27], expectation maximisation [28], [29] and Bayes estimation [30], [31]. Recently, these algorithms have been used within PHD fitting approaches that: (a) partition the dataset into segments [29] or clusters [32] of “similar” data points; (b) employ an established algorithm to fit a PHD with a simple structure to each data segment or cluster; and (c) use these simple PHDs as branches of a PHD that fits the whole dataset. These approaches achieve better trade-offs between the size, accuracy and complexity of the final PHD than direct algorithms applied to the entire dataset.

The function HTMODEL from Algorithm 1 performs the holding-time modelling for state s'_i of the CTMC(S', π', R') model from Theorem 2. HTMODEL applies Reinecke et al.’s cluster-based PHD fitting approach [32], [33], [34] to the holding times sample_{i}(10) lines 2–16), uses this PHD to derive the parameters of the refined CTMC(S'', π'', R'') (lines 17–20), and returns this CTMC in line 21.

The PHD fitting is carried out by the while loop in lines 5–16, which iteratively assesses the suitability of PHDs obtained when partitioning sample_{i} into c = MinC, MinC +...
TABLE III: Erlang delay model parameters for the states of the CTMC from Fig. 1

<table>
<thead>
<tr>
<th>$s_i$</th>
<th>label</th>
<th>$t_i$ (ms)</th>
<th>$\lambda_i^E$ (s$^{-1}$)</th>
<th>$\lambda_i^r$ (s$^{-1}$)</th>
<th>$s_i$</th>
<th>label</th>
<th>$t_i$ (ms)</th>
<th>$\lambda_i^E$ (s$^{-1}$)</th>
<th>$\lambda_i^r$ (s$^{-1}$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$s_1$</td>
<td>location</td>
<td>49</td>
<td>5285</td>
<td>18.21</td>
<td>$s_4$</td>
<td>search</td>
<td>209</td>
<td>1239</td>
<td>3.01</td>
</tr>
<tr>
<td>$s_2$</td>
<td>arrivals</td>
<td>45</td>
<td>5756</td>
<td>188.81</td>
<td>$s_5$</td>
<td>weather</td>
<td>706</td>
<td>367</td>
<td>5.14</td>
</tr>
<tr>
<td>$s_3$</td>
<td>departures</td>
<td>45</td>
<td>5756</td>
<td>156.76</td>
<td>$s_6$</td>
<td>traffic</td>
<td>179</td>
<td>1447</td>
<td>4.57</td>
</tr>
</tbody>
</table>

Algorithm 1: Holding-time modelling with parameters:
- $MinC$ (minimum number of PHD clusters)
- $MaxC$ (maximum number of PHD clusters)
- $MaxP$ (maximum number of cluster phases)
- $FittingAlg$ (basic PHD fitting algorithm)
- $MaxSteps$ (maximum steps without improvement)

1: function HTMODEL($S', \pi', R'$, $s_i', \text{sample}_i, \alpha$)
2: \[ \text{PHD}(\pi_0, D_0) \leftarrow \text{null}, \minErr = \infty \]
3: \[ \text{improvement} \leftarrow 0, \text{steps} \leftarrow 0 \]
4: \[ c \leftarrow \text{MinC} \]
5: while \( c \leq \text{MaxC} \land \text{steps} \leq \text{MaxSteps} \) do
6: \[ \text{phd} \leftarrow \text{CBFITTING(sample, c, FittingAlg, MaxP)} \]
7: \[ \text{err} \leftarrow \Delta \text{CDF(sample, phd)} \]
8: if \( \text{err} < \minErr \) then
9: \[ \text{improvement} \leftarrow \text{improvement} + (\minErr - \text{err}) \]
10: \[ \minErr \leftarrow \text{err} \]
11: \[ \text{PHD}(\pi_0, D_0) \leftarrow \text{phd} \]
12: if \( \text{improvement} \geq \alpha \) then
13: \[ \text{improvement} \leftarrow 0, \text{steps} \leftarrow 0 \]
14: else
15: \[ \text{steps} \leftarrow \text{steps} + 1 \]
16: \[ c \leftarrow c + 1 \]
17: \[ N_i \leftarrow \text{SIZEOF}(D_0) \]
18: \[ S'' \leftarrow (S' \setminus \{s_i\}) \cup \{s_{i1}, s_{i2}, \ldots, s_{iN_i}\} \]
19: \[ \pi'' = \left[ \frac{\lambda_i^E \pi_0}{\lambda_i^r + 1} \right] \]
20: \[ R'' \leftarrow \begin{bmatrix} \pi''(S' \setminus \{s_i\}, S'' \setminus \{s_i\}) & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ D_0 & \cdot & \cdot & \cdot \end{bmatrix} \]
21: return $\text{CTMC}(S'', \pi'', R'')$

1, \ldots, \text{MaxC} \text{ clusters. Line 6 obtains a PHD with c branches (corresponding to partitioning sample}_i \text{ into c clusters) and up to MaxP phases by using the function CBFITTING, which implements the cluster-based PHD fitting from [32]. The FittingAlg argument of CBFITTING specifies the basic PHD fitting algorithm applied to each cluster as explained above, and can be any of the standard moment matching, expectation maximisation or Bayes estimation PHD fitting algorithms. The quality of the c-branch PHD is established in line 7 by using the CDF-difference metric [32] to compute the difference err between sample}_i \text{ and the PHD. If the statement in lines 8–11 identifies the PHD with the lowest err value so far, making a record of it in line 10. Any reductions in err (i.e., "improvements") are cumulated in improvement (line 9), and the while loop terminates early if the iteration counter steps exceeds MaxSteps before improvement reaches the threshold $\alpha \geq 0$ supplied as an argument to HTMODEL and the steps counter is reset (line 13).

On exit from the while loop, the elements $D_0, d_1 = -D_0$, and $\pi_0$ of the best-fit phase-type distribution PHD($\pi_0, D_0$) recorded in line 11 are used to calculate the parameters $S'', R''$ and $\pi''$ of the refined CTMC. These calculations are shown in lines 17–20, under the simplifying assumption that $R'$ maintains the transition rates for the outgoing and incoming transitions of states $s_{i_k}$ and $s'_i$ in the last two rows and columns, respectively. This assumption is easily satisfiable by reordering the rows and columns of the matrix, so that $R'$ has the structure

$$R' = \begin{bmatrix} R'(S' \setminus \{s_i\}, S'' \setminus \{s_i\}) & 0 \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ D_0 & \cdot & \cdot & \cdot \end{bmatrix}$$

where $R'(S' \setminus \{s_i\}, S'' \setminus \{s_i\})$ denotes matrix $R'$ without the last row and column. The following justifies the construction of the refined CTMC.

Theorem 3: The tuple $(S'', \pi'', R'')$ synthesised by Algorithm 1 defines a valid CTMC in which the probability that $s_{j1}$ is the first state from $S'' \cap S'$ reached after state $s_{i_k}$, is $p_{i_k}$, and the mean time of reaching any state in $S'' \cap S'$ from $s_{i_k}$, is given by the mean time to reach the absorbing state of PHD($\pi_0, D_0$).

Proof: Since $\pi'$ is the initial probability vector of CTMC($S', \pi', R'$), its elements sum to 1.0 and do so the elements of $\pi''$, which is therefore a valid probability vector. It is immediate to show that the rows of $R''$ contain non-negative elements outside the main diagonal and that the elements on each row add up to zero, so $R''$ is a valid transition rate matrix. Accordingly, $(S'', \pi'', R'')$ satisfies Definition 1 and is a valid CTMC. For the second part of the theorem, assume that the $N_i$ transient states of the (acyclic) PHD($\pi_0, D_0$) are reached with probabilities $x_1, x_2, \ldots, x_{N_i}$. We have $[x_1 x_2 \cdots x_{N_i}] \cdot d_1 = 1$. In the CTMC returned by Algorithm 1, we note from $R''$ that the next state after $s_{i_k}$ is necessarily one of $s'_{i1}, s'_{i2}, \ldots, s'_{iN_i}$, and that the probabilities of reaching these “PHD-based” states are given by the elements of $\pi_0$. Thus, the probabilities of reaching the states $s'_{i1}, s'_{i2}, \ldots, s'_{iN_i}$ from $s_{i_k}$ are $x_{i1}, x_{i2}, \ldots, x_{iN_i}$ and so the probability of reaching $s_{j1}$ from $s_{i_k}$ is $[x_1 x_2 \cdots x_{N_i}] \cdot p_{i_k} \cdot d_1 = p_{i_k} \cdot [x_1 x_2 \cdots x_{N_i}] \cdot d_1 = p_{i_j}$. Finally, since the next state after $s_{i_k}$ corresponds to a transient state of PHD($\pi_0, D_0$), the mean time to reach any other state (i.e., a state from $s_{i1}, s_{i2}, \ldots$) is the same as the mean time to reach the absorbing state of the PHD.

Example 2: We used our OMNI implementation to perform Stage II of the refinement on the CTMC from the motivating example. Algorithm 1 was executed for each of our six services, with $\alpha = 0.1$ and with the configuration parameters
We evaluated OMNI experimentally to answer the following research questions (RQs). RQ1 (Accuracy): How effective are OMNI models at predicting quality property values for other system runs than the one used to collect the datasets used for the refinement? For any stochastic process, a (sufficiently large) training dataset should be sufficient to capture the behaviour of the system. Models based on this dataset should be robust when evaluated against datasets from other system runs (i.e., no overfitting). To assess if OMNI models have this property, we used three four-hour runs of the application from Section IV. The differences between actual and predicted values for each additional dataset are presented in Fig. 7a. The models used in this analysis are the initial CTMC from Fig. 1 (labelled “exponential” in the diagrams) and the refined CTMCs obtained after each OMNI stage for the training dataset and the OMNI parameters from Examples 1 and 2. A visual inspection of the results shows that each OMNI stage significantly improves the accuracy of the analysis. The reduction in the cumulated prediction error (7) across the three testing datasets was in the ranges 82.5–88.6% for property P1, 77–90.3% for P2 and 83.3–89.8% for P3.

To eliminate the risk that these improvements were due to an advantageous partition of the data into training and testing sets, we carried out 30 more experiments in which the observations were randomly partitioned into four new datasets, one used for training and the others for assessing fitting errors. The errors from these experiments, summarised as box plots in Fig. 7b, show that OMNI consistently outperforms traditional CTMC modelling irrespective of the choice of training data, confirming the accuracy and robustness of our method.

RQ2 (Refinement Granularity): What is the effect of varying the refinement granularity on the model accuracy, size and verification time? We ran experiments to evaluate the accuracy of the OMNI predictions when varying: (i) the size $k_i$ of the Erlang delay model (cf. Table II); and (ii) the threshold $\alpha$ from Algorithm 1 (with the other OMNI parameters as in Examples 1 and 2). The values of $k_i$ and $\alpha$ determine the granularity of the refinement from the two OMNI stages, with larger $k_i$ and smaller $\alpha$ values corresponding to finer-grained refinement (and larger refined CTMCs). The experimental results (Table IV) show that when only Stage I of OMNI is used, increasing $k_i$ initially reduces the analysis error for all properties, with significant improvements even for small $k_i$. However, increasing $k_i$ beyond a certain value (approx. 100 for our system) has little effect on the model accuracy. In Stage II, the analysis error decreases when smaller $\alpha$ values are used. However, this decrease is much less significant when changing from $\alpha = 0.1$ to $\alpha = 0.05$ than when changing from $\alpha = 0.2$ to $\alpha = 0.1$. Both results show the presence of a point of “diminishing returns” in continuing the refinement beyond a certain level of granularity (which is necessarily application and training dataset dependent). As a further remark, a coupling effect between $\alpha$ and $k_i$ is suggested by the results. This increases the limit at which $k_i$ still provides improvements in model accuracy, with Stage I+II model accuracy showing

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{fig6.png}
\caption{Properties P1–P3 predicted after holding-time modelling (dashed line) vs. actual (continuous line); $\text{error}_0$, $\text{error}_i$ and $\text{error}_II$ are the prediction errors before OMNI and after each OMNI stage, respectively.}
\end{figure}
improvements as $k_i$ is increased from 100 to 259.

Table IV also reports the number of CTMC states at each level of granularity, and the associated verification times—comprising the time to build the model and to verify each property at a single time point using PRISM on a MacBook Pro with 2.9 GHz Intel i5 processor and 16Gb of memory. As expected, the model size and verification time increase rapidly with the refinement granularity, up to over 2000 states and 12s for several models. Importantly, the results indicate that this increase can be limited by not refining the model beyond the granularity at which the predicted property values stabilise, which can be determined e.g. by applying OMNI iteratively at increasing levels of refinement granularity. Another positive insight is that when computational resources are at a premium, small OMNI refined models still provide considerable accuracy improvements.

The component-based system from our case study is realistic, but relatively simple. Therefore, it is worth noting that the increase in model size after the OMNI refinement is only linear in the number of system components. Moreover, as OMNI uses acyclic PHDs, the number of transitions also increases only linearly with refinement. As modern model checkers can analyse CTMCs with $10^5$–$10^6$ states [36], we expect OMNI to scale well to much larger system sizes. We confirmed these hypotheses (Fig. 8) by running additional OMNI refinement experiments for models of systems with 12, 24, 48 and 96 components. These larger models (available on our project webpage) were obtained by combining 2, 4, 8 and 16 instances of our six-component high-level CTMC from Fig. 1, and OMNI was used with observation sets drawn randomly from those of our six web services, $k = 259$ and $\alpha = 0.1$. The verified property was only P1 from (6) for $T = 20s$ (as P2 and P3 cannot be meaningfully extrapolated to the larger systems).

**RQ3 (Training dataset size): What is the effect of the training dataset size on the refined model accuracy?** We ran experiments to assess the accuracy of OMNI predictions when the refined CTMC was derived using subsets comprising only 20%, 40%, 60% and 80% of the elements from our 270-entry training dataset. For each subset size we used $k_i = 259$ and $\alpha = 0.1$, and ran 30 experiments with subset elements drawn randomly from the complete training dataset. As expected, the results (depicted in Table V) show that the OMNI prediction accuracy
TABLE V: Training dataset size effect on prediction accuracy, shown as average error and standard deviation over 30 runs

<table>
<thead>
<tr>
<th>Dataset††</th>
<th>P1 Error</th>
<th>P2 Error</th>
<th>P3 Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>100%††</td>
<td>0.037 sd N/A*</td>
<td>0.039 sd N/A*</td>
<td>0.063 sd N/A*</td>
</tr>
<tr>
<td>80%</td>
<td>0.054 sd 0.019</td>
<td>0.054 sd 0.015</td>
<td>0.109 sd 0.053</td>
</tr>
<tr>
<td>60%</td>
<td>0.063 sd 0.023</td>
<td>0.066 sd 0.024</td>
<td>0.130 sd 0.066</td>
</tr>
<tr>
<td>40%</td>
<td>0.065 sd 0.021</td>
<td>0.063 sd 0.020</td>
<td>0.149 sd 0.068</td>
</tr>
<tr>
<td>20%</td>
<td>0.074 sd 0.029</td>
<td>0.075 sd 0.028</td>
<td>0.157 sd 0.068</td>
</tr>
</tbody>
</table>

Initial CTMC 0.325 sd N/A* 0.402 sd N/A* 0.377 sd N/A*

†Percentage of complete 270-element training dataset ††Single run using entire data set

RQ4 (Comparison to PHD): How does OMNI compare to a single-stage PHD fitting approach? As PHDs can fit any positive continuous distribution, we assessed the benefit of having a delay extraction stage in OMNI. To this end, we switched off OMNI’s delay extraction and produced PHD-only refined CTMCs for $\alpha \in \{0.1, 0.05, 0.01\}$. Table VI compares these models to OMNI-refined CTMCs of equivalent size (i.e. small, medium or large) that we constructed by using $\alpha = 0.1$, with enough delay-modelling states to not exceed the PHD-only model sizes. The experimental results show that for all three model sizes, the OMNI models yield more accurate results (with similar or better verification times, and much smaller refinement times) than the PHD models. This confirms existing theoretical results which show that PHD fitting of deterministic delays requires an excessive number of states, and that an Erlang distribution (as used in the first OMNI stage) is the best fit for a fixed delay [24].

Finally, the cluster-based PHD fitting method from [32] was used to implement the efficient PHD-fitting tool HyperStar [33]. However, the PHD fitting method and tool from [32], [33] generate a PHD only for a single dataset. OMNI uses this method and tool in its second stage, to refine the high-level CTMC model of a component-based software system.

Non-PHD-based approaches to combining Markov models with real data range from e.g. using Monte Carlo simulation to analyse properties of discrete-time Markov chains with uncertain parameters [39] to using semi-Markov chains to model holding times governed by general distributions [40]. However, none of these approaches can offer the guarantees and tool support provided by OMNI thanks to its exploitation of established CTMC model checking techniques.

VIII. THREATS TO VALIDITY

Construct validity threats may arise due to the assumptions made when implementing our prototype web application and collecting the datasets used for the model refinement experiments. To mitigate them, we implemented the web application using standard Java technologies, and we collected the datasets from six real web services from three different providers, at different times of day and on two different days. We also used different datasets for training and testing, and we analysed typical performance and cost properties of the application.

Internal validity threats can be due to the stochastic nature of the analysed component-based system or bias in interpreting the experimental results. To address these threats, we provided formal proofs for our method, and reported results from multiple independent experiments that use different values for the OMNI parameters, and analyse three system properties at several levels of refinement granularity.

External validity threats may exist if the stochastic features of the analysed system are not indicative of the features of other software systems. To reduce this threat, trace data was obtained from a range of real web services, such that the characteristics of the distributions used exhibited features which we would expect to see in a many real-world applications, e.g., regions of zero density, multi-modal response times and long tails. Additionally external threats exists if the refined CTMC model cannot be verified within the resources of a traditional computer. The OMNI approach allows for the refinement to be carried out at different levels of granularity and our evaluation suggests that significant improvements in prediction accuracy can be achieved with modest enlargement of the models.

IX. CONCLUSION

We introduced OMNI, an observation-based CTMC refinement method and tool that significantly improve the accuracy with which quality properties of software systems can be analysed using CTMC models. We evaluated OMNI using data obtained from real web services, demonstrating its robustness.

In future work, we will assess the sensitivity of OMNI to the size of the training datasets, and its effectiveness at estimating a wider range of execution time distributions for the system components. In addition, we plan to augment OMNI with
the ability to intelligently vary the level of refinement across components, so that fine-grained refinement is only used for components that need it.

ACKNOWLEDGEMENTS

This paper presents research sponsored by the UK MOD. The information contained in it should not be interpreted as representing the views of the UK MOD, nor should it be assumed it reflects any current or future UK MOD policy.

REFERENCES