DIAGNOSIS OF STOCHASTIC DISCRETE EVENT SYSTEMS BASED ON N-GRAM MODELS WITH WILDCARD CHARACTERS

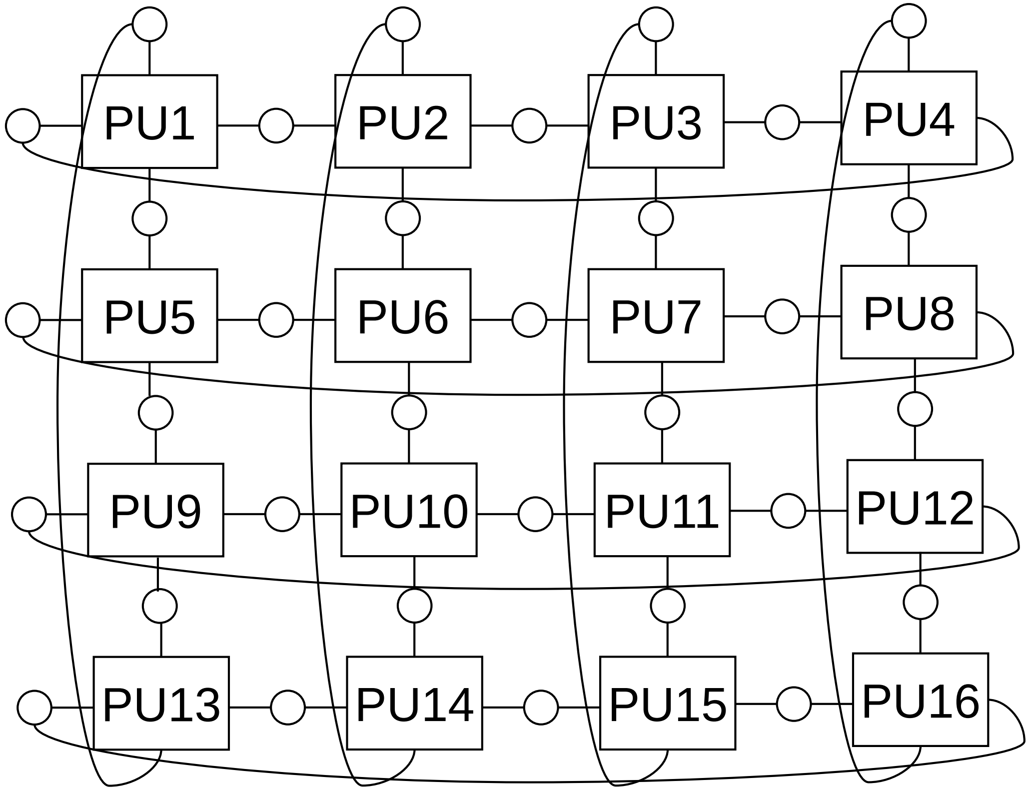
Kunihiko HIRAISHI, Miwa YOSHIMOTO and Koichi KOBAYASHI
School of Information Science, JAIST

Who am I?

- My Current Research Interests include
  - Theoretical Computer Science / System Science
    - Formal Modeling of Systems / Formal Verification,
    - Discrete Event Systems / Hybrid Systems,
    - Systems Biology,
    - Optimization and Algorithms
  - Service Science including Human Activities (New)
    - Smart Voice Messaging System in Nursing/Caregiving Services

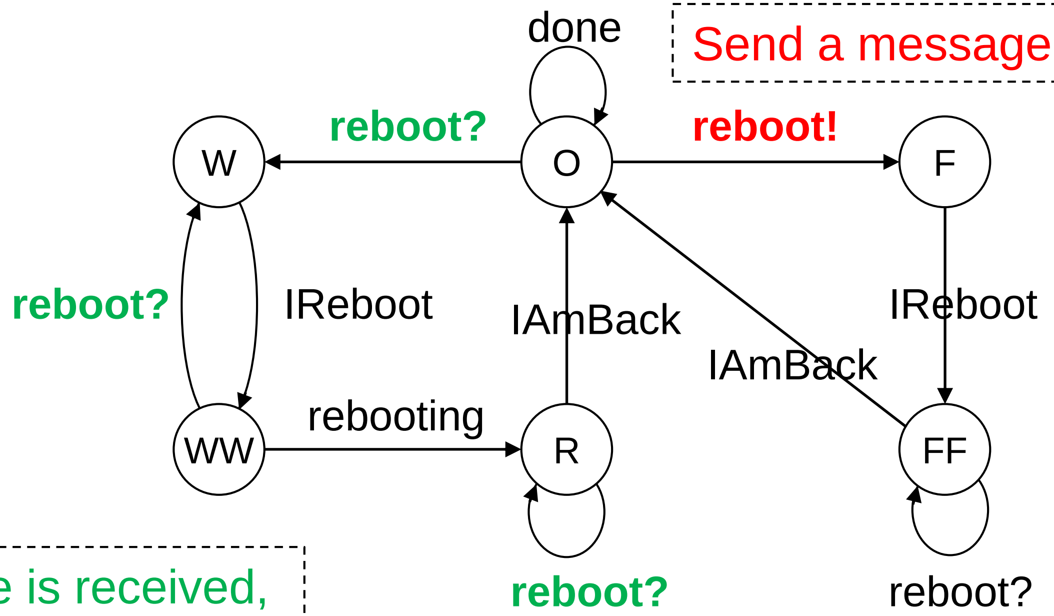
Example: Diagnosis of a Multi-Processor System

A. Grastien, Diagnosis of Discrete-Event Systems Using Satisfiability Algorithms, AAAI2007, pp.305-310 (2007).



Example: Diagnosis of a Multi-Processor System

Send a message to all neighbors



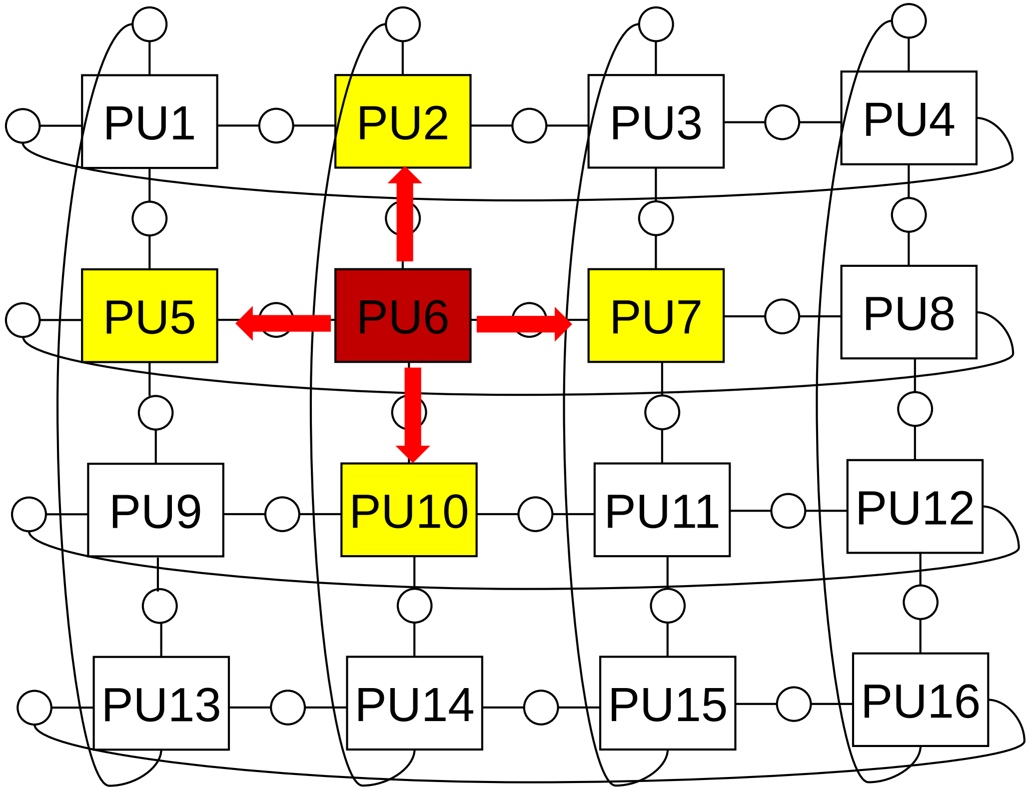
If a message is received, then the PU also reboots.

Processing Unit（PU)

There are at most 6^{16} states in total.

Example: Diagnosis of a Multi-Processor System

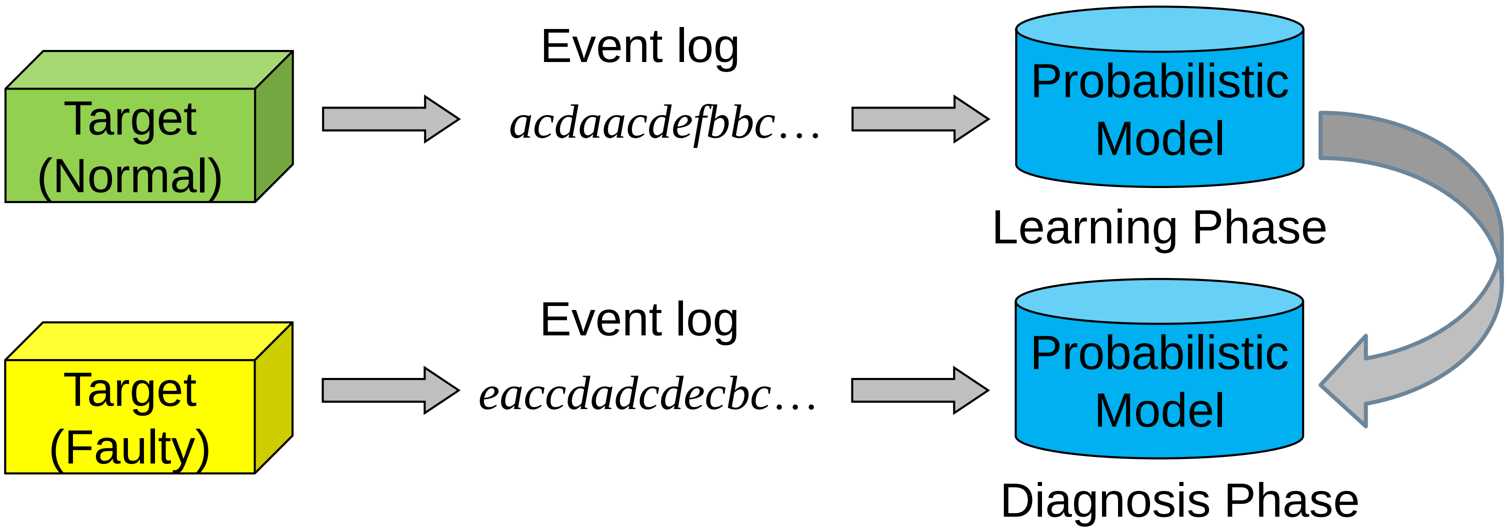
Message “Reboot!” Is sent to all neighbors.



Proposal of Model-less Diagnosis (MLD)

- Diagnosis of Discrete Event Systems using event-logs only.
- It aims to identify
  - whether some faults have occurred or not,
  - which type of faults has occurred,
  - the time faults have occurred.

Proposal of Model-less Diagnosis (MLD)



Compare the event log with the model

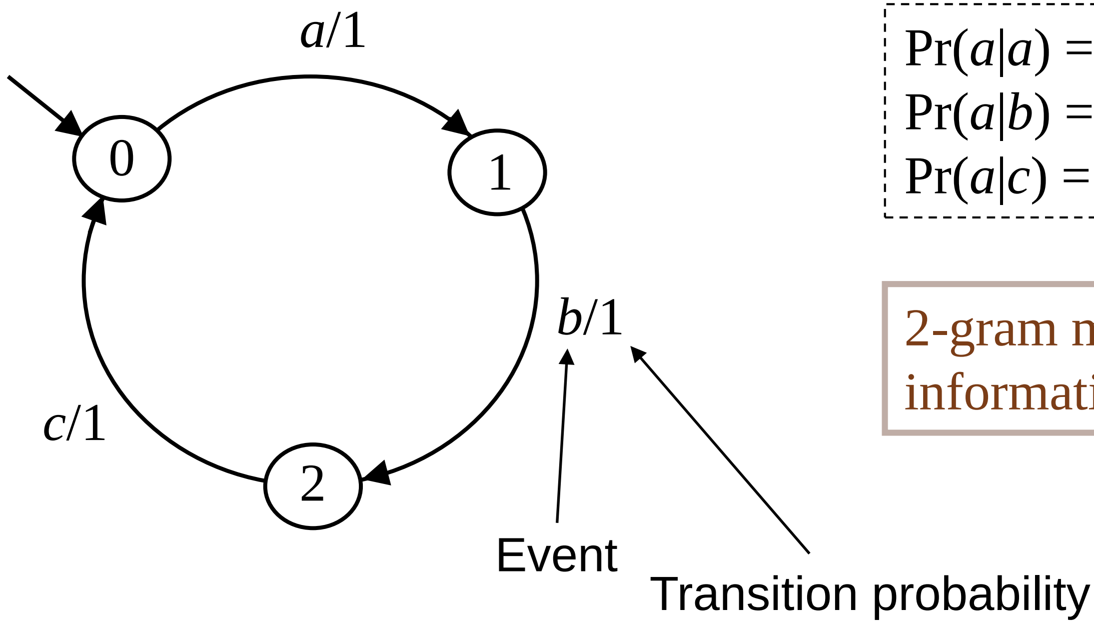
Related Work

- Model-based diagnosis (MBD): The exact system model is required. Computational complexity is very high.
  MBD cannot handle 6^{16} states!
- Rule-based diagnosis: Empirical knowledge on the system is required. Many results in AI.
  A priori knowledge is not required in MLD.
- Process mining: It uses event-logs, but obtaining complete process models is the goal.
  Simpler models sufficient for the diagnosis are used in MLD.

Probabilistic Model: *N*-Gram Model

- *N*-gram: A string of length N.
- Let e_1, e_2, e_3, …, e_i, … be a sequence of event symbols generated by the target system.
- Suppose that the probability that e_i occurs depends only on the N-1 gram just before e_i. Then a collection of conditional probabilities $\Pr(e_i \mid e_{i-N+1} \ldots e_{i-1})$ approximately represent the system behavior.
- The idea was shown by C. E. Shannon.

N-Gram Model can approximate Markov chains



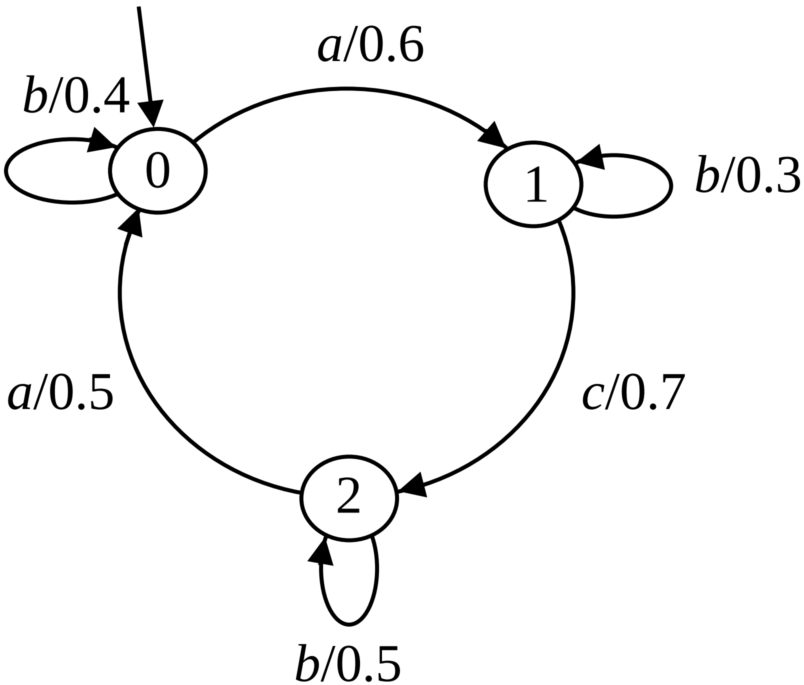
Discrete-time Markov chain

2-gram model

$\Pr(a|a) = 0$, $\Pr(b|a) = 1$, $\Pr(c|a) = 0$
$\Pr(a|b) = 0$, $\Pr(b|b) = 0$, $\Pr(c|b) = 1$
$\Pr(a|c) = 1$, $\Pr(b|c) = 0$, $\Pr(c|c) = 0$

2-gram model has the complete information of the Markov chain.

N-Gram Model can approximate Markov chains



Discrete-time Markov chain

No *N*-gram model can represent the Markov chain since both state 0 and state 1 are reached by unbounded sequence b^*a.

N-Gram Model can approximate Markov chains

3-Gram Model

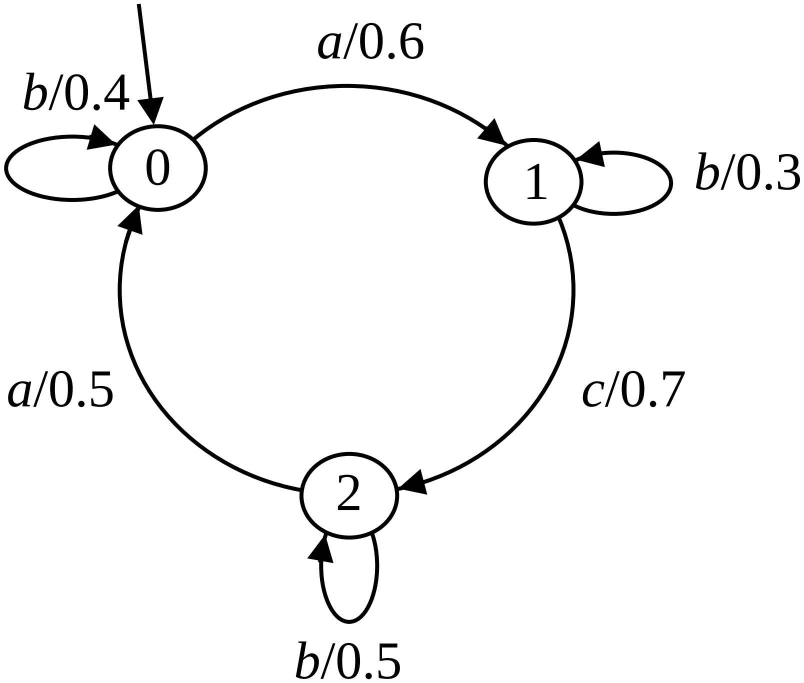
	Possible states	*a*	*b*	*c*
aa	1	0	1/2	1/2
ab	0, 1			
ac	2	1/2	1/2	0
ba	0, 1			
bb	0, 1, 2			
bc	2	1/2	1/2	0
ca	0	3/5	2/5	0
cb	2	1/2	1/2	0
cc	–			

N-Gram Model can approximate Markov chains

4-Gram Model

	Possible states	*a*	*b*	*c*
aaa	–			
aab	0	3/5	2/5	0
aac	–			
aba	1	0	1/2	1/2
abb	0	3/5	2/5	0
abc	–			
aca	0	3/5	2/5	0
acb	2	1/2	1/2	0
…				
bba	0, 1			

N-Gram Model for Ergodic Markov Chains



Steady state probability

$$P = \begin{bmatrix} 0.4 & 0.6 & 0 \\ 0 & 0.3 & 0.7 \\ 0.5 & 0 & 0.5 \end{bmatrix},$$

$$\pi = [\pi_0, \pi_1, \pi_2]$$

$$\pi = \pi P, \pi_0 + \pi_1 + \pi_2 = 1$$

$$\pi_0 = 35/107, \pi_1 = 30/107, \pi_2 = 42/107$$

N-Gram Model for Ergodic Markov Chains

The set of possible states after sequence *ab* is { 0, 1 }.

$$\Pr(a \mid ab) = 0.6 \cdot \frac{\pi_0}{\pi_0 + \pi_1} = 21/65$$

relative probability

$$\Pr(b \mid ab) = 0.4 \cdot \frac{\pi_0}{\pi_0 + \pi_1} + 0.3 \cdot \frac{\pi_1}{\pi_0 + \pi_1} = 23/65$$

$$\Pr(c \mid ab) = 0.7 \cdot \frac{\pi_1}{\pi_0 + \pi_1} = 21/65$$

N-Gram Model for Ergodic Markov Chains

The set of possible states after sequence *bb* is { 0, 1, 2 }.

$$\Pr(a \mid bb) = 0.6 \cdot \frac{\pi_0}{\pi_0 + \pi_1 + \pi_2} + 0.5 \cdot \frac{\pi_2}{\pi_0 + \pi_1 + \pi_2} = 42/107$$

$$\Pr(b \mid bb) = 0.4 \cdot \frac{\pi_0}{\pi_0 + \pi_1 + \pi_2} + 0.3 \cdot \frac{\pi_1}{\pi_0 + \pi_1 + \pi_2} + 0.5 \cdot \frac{\pi_2}{\pi_0 + \pi_1 + \pi_2} = 44/107$$

$$\Pr(c \mid bb) = 0.7 \cdot \frac{\pi_1}{\pi_0 + \pi_1 + \pi_2} = 21/107$$

N-Gram Model for Ergodic Markov Chains

3-Gram Model based on steady state probability

	Possible states	*a*	*b*	*c*
aa	1	0	1/2	1/2
ab	0, 1	21/65	23/65	21/65
ac	2	1/2	1/2	0
ba	0, 1	21/65	23/65	21/65
bb	0, 1, 2	42/107	44/107	21/107
bc	2	1/2	1/2	0
ca	0	3/5	2/5	0
cb	2	1/2	1/2	0
cc	–			

Derivation of *N*-Gram Models from Event Logs

- If a Markov chain is ergodic, there exists an *N*-gram model that precisely represent its behavior at the steady state.
- Such an *N*-gram is obtained from sufficiently long event sequence w generated by the Markov chain:

$$\Pr(\sigma|y) = \frac{O_{y\sigma}(w)}{\sum_{\sigma' \in \Sigma} O_{y\sigma'}(w)}$$

where

y is an (*N* -1)-gram, and

$O_x(w)$ denote the number of occurrences of x in w.

Sequence Profiling

0. Give w_{ref} (an event sequence in normal situation) and
   w_k $(k = 0, \ldots, m)$ (event sequences in various faulty situations).
1. *Learning phase*: Learn an *N*-gram model M_{ref} from w_{ref}.
2. *Diagnosis Phase*: Given the observed sequence w_{test} and *r* (> *N*):

① Based on M_{ref}, compute the expected number of times E_j that each *r*-gram s_j occurs in w_{test}.

② Count the number of times O_j that each *r*-gram s_j occurs in w_{test}.

③ Compute logratio $d^{r,N}(s_j) = \log(O_j / E_j)$ for each *r*-gram s_j. We call it *specificity* of s_j. Let $D^{r,N}(w_{test})$ be the vector each component of which corresponds to an r-gram s_j and has the value $d^{r,N}(s_j)$. We call $D^{r,N}(w_{test})$ *profile* of w_{test}.

④ Compute correlation coefficients between $D^{r,N}(w_{test})$ and $D^{r,N}(w_k)$ $(k = 0, \ldots, m)$. Output k that gives the highest correlation.

Sequence Profiling

Test sequence w_{test}

acdaacdefbbc…

N-gram model for normal situation

Actual value

acd : 20
cda : 25
daa : 11
aac : 10
…

Expected value

acd : 18.3
cda : 22.1
daa : 15.6
aac : 8.9
…

$r = 3$
$(r \geq N)$

Profile of w_{test}

acd : log(20/18.3)
cda : log(25/22.1)
daa : log(11/15.6)
aac : log(10/8.9)
…

All components are almost 0 in normal situation.

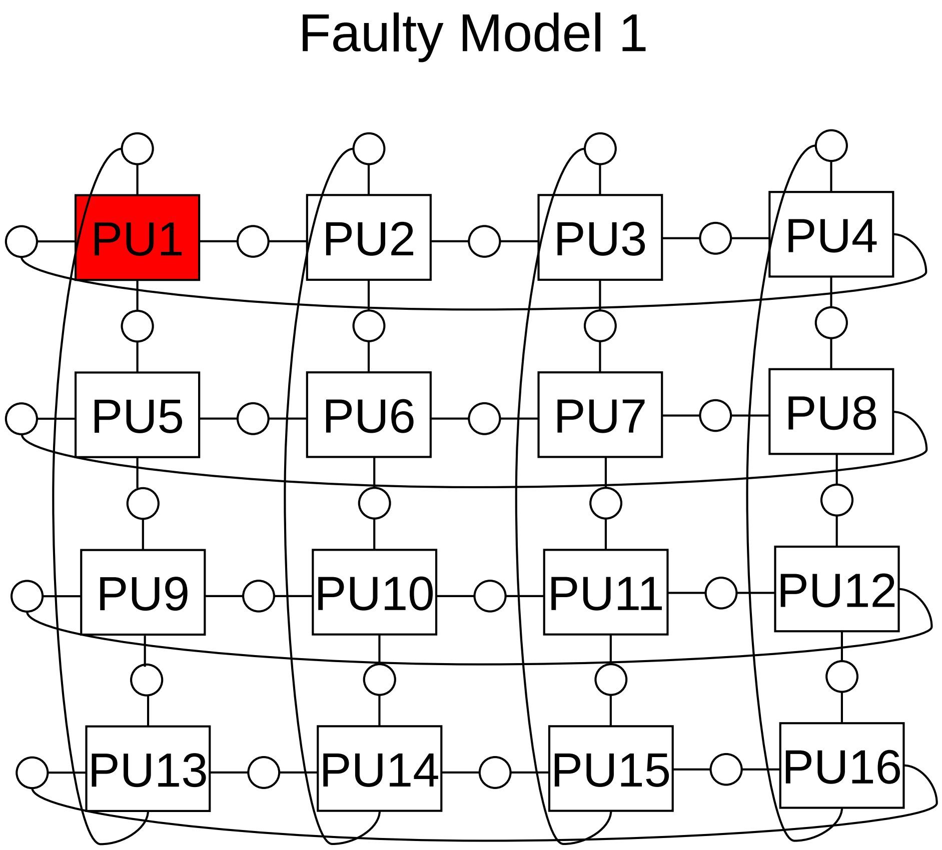
Profile is a dimensionless quantity.

Introducing Wildcard Characters into Patterns

- In distributed processing environment such as cloud systems, event-logs from subsystems are interleaved.
- To eliminate effect by subsystems that may not be related to the faults, we consider *masking* of sequence patterns by wildcard characters.
- Let "n" be the wildcard character and $\Sigma = \{ a, b \}$. "n" can be substituted for any symbol of Σ, e.g., $anbn \Rightarrow aaba, aabb, abba, abbb$. We call such a pattern *a motif*.
- This idea is inspired by Genom Sequence Analysis.

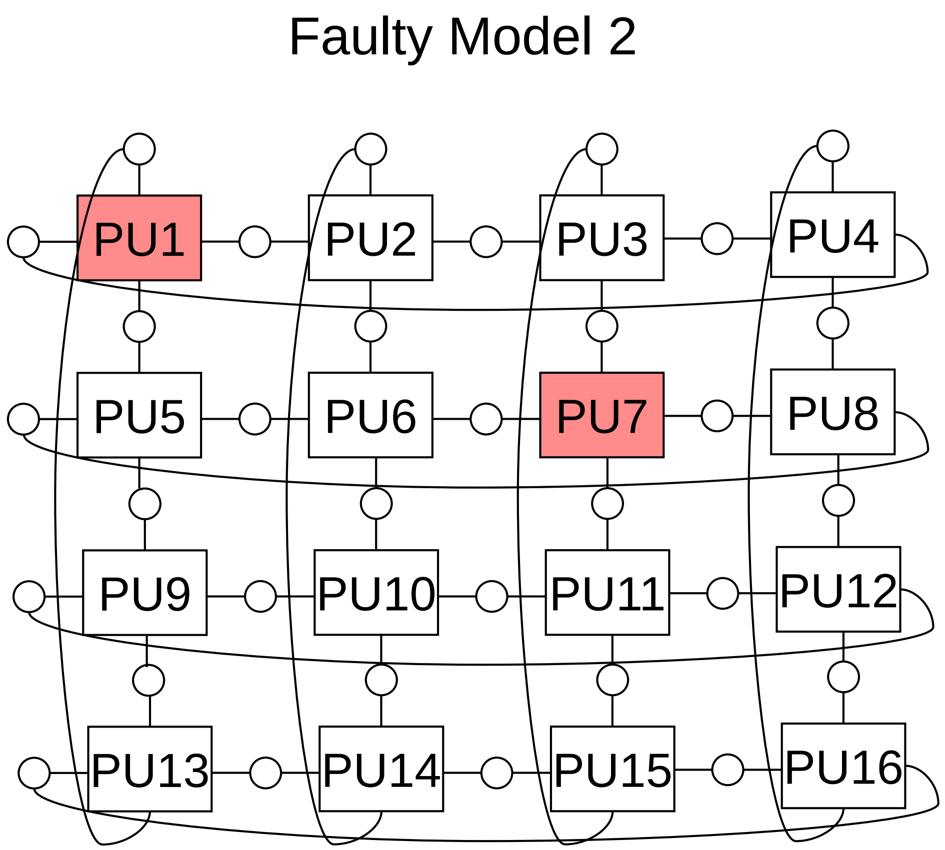
Faulty Model 1

Frequency of
reboot x 5

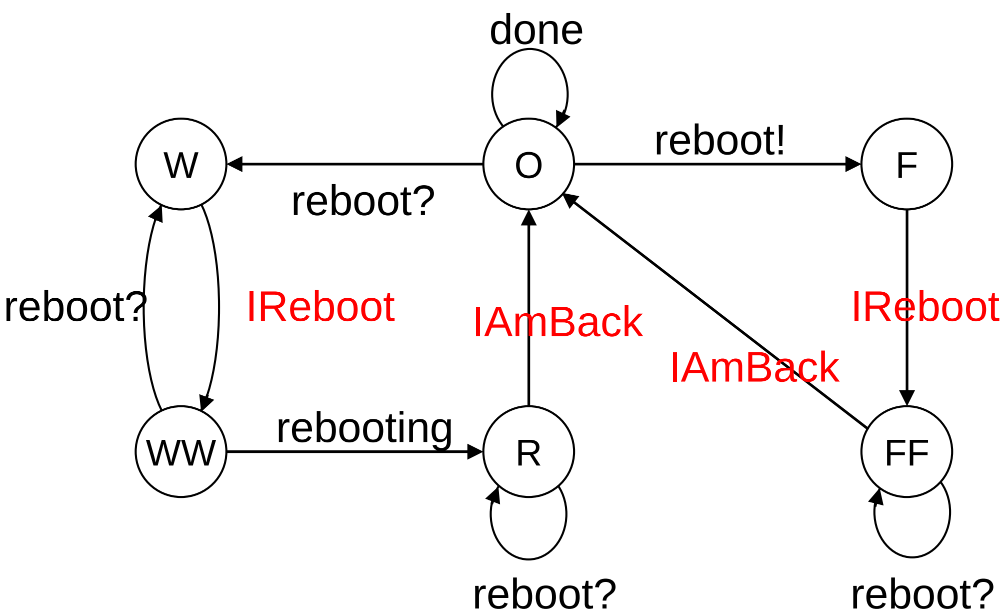


Faulty Model 2

Frequency of reboot x 2.5

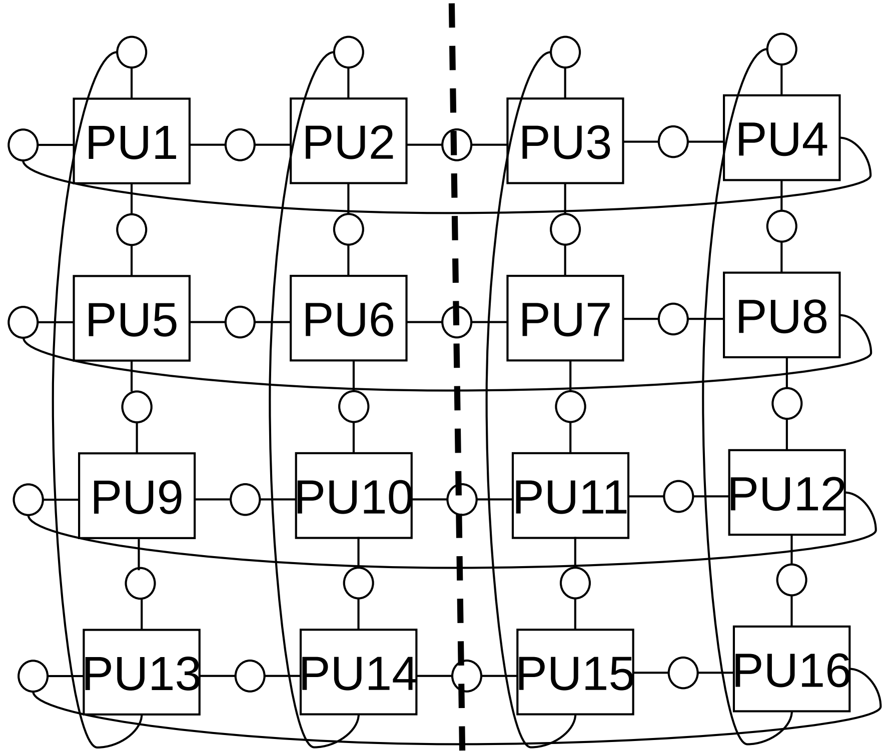


Event Observation and Abstraction



IReboot and IAmBack are observable, and other events are unobservable

Event Observation and Abstraction



Left

Right

IReboot: a
IAmBack: x

IReboot: b
IAmBack: y

Experiment

- Event sequences
  - w_{ref} : Normal model
  - w_0 : Normal model
  - w_1 : Faulty model 1
  - w_2 : Faulty model 2
  - w_{test} : Faulty model 2 $\Rightarrow$ correct estimation
- We use a simulator based on Stochastic Petri Nets to obtain event-logs.

Results

TABLE I

STATISTICS ON THE NUMBER OF EVENT OCCURRENCES.

	w_0	w_1	w_2	w_{test}
a	12,980	14,634	13,468	13,381
x	8,562	8,506	8,619	8,663
y	8,581	8,177	8,623	8,656
b	13,237	12,918	13,493	13,506

Results

r = 3
wildcard 0
N = 0, …, 3

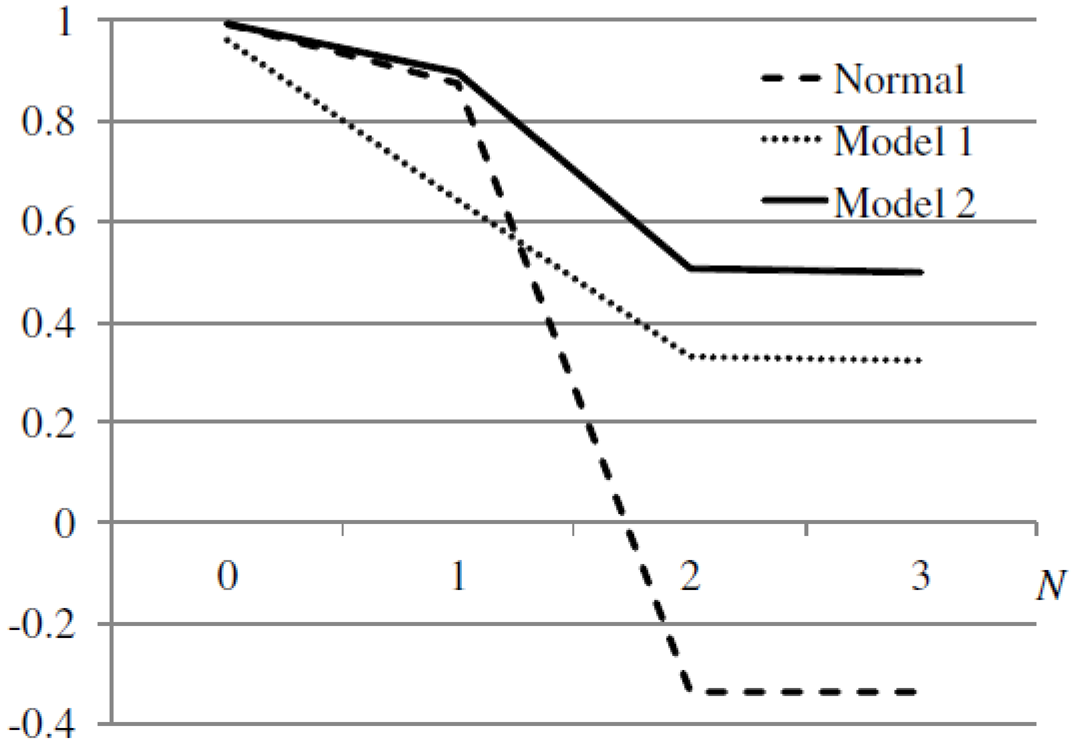


Fig. 6. Correlation coefficients between $\mathcal{D}^{3,0}(w_{test})$ and other $\mathcal{D}^{3,0,N}(w_i)$'s for $N = 0, \cdots, 3$.

Results

r = 6
wildcard 0
N = 0, …, 5

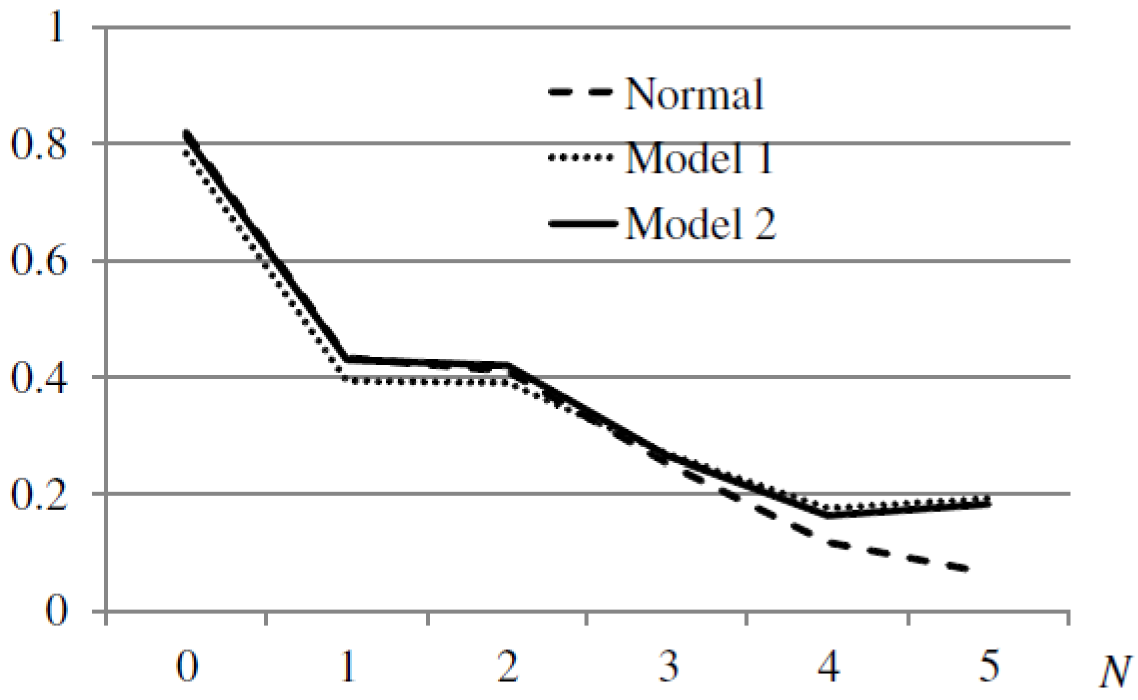


Fig. 7. Correlation coefficients between $\mathcal{D}^{6,0}(w_{test})$ and other $\mathcal{D}^{6,0,N}(w_i)$'s for $N = 0, \cdots, 5$.

Results

$r = 6$
wildcard 3
$N = 0, \ldots, 5$

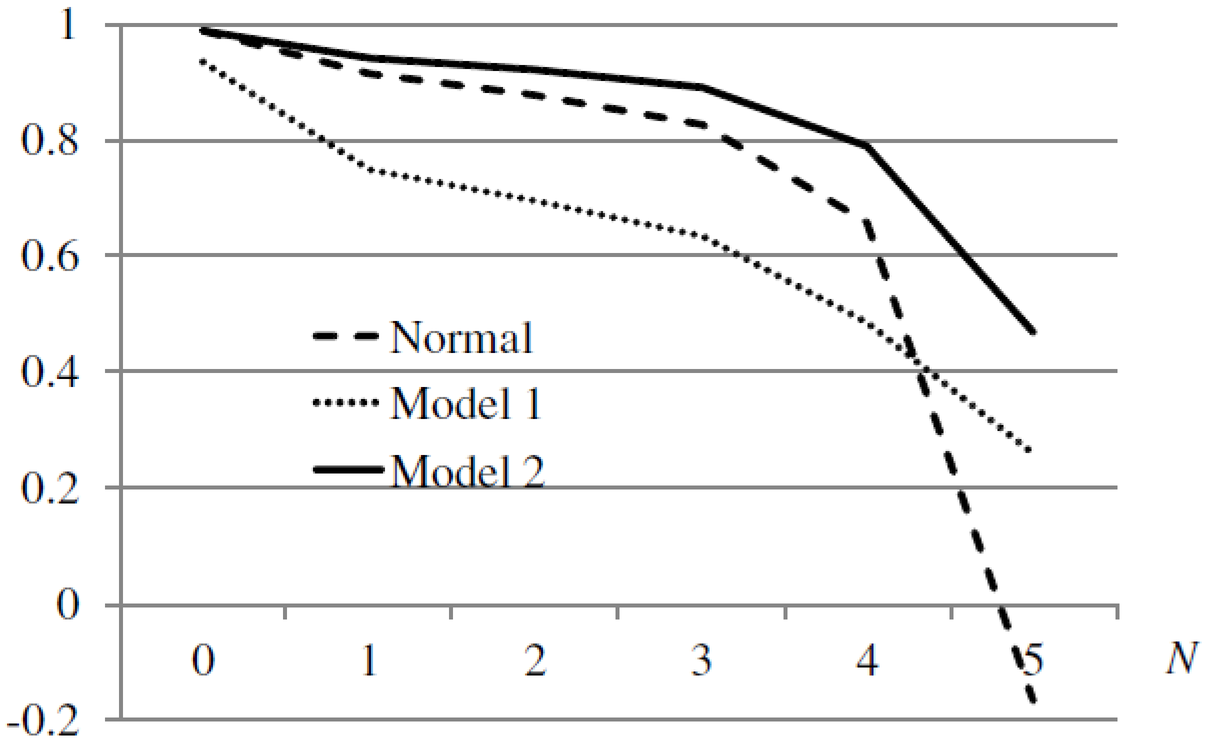


Fig. 8. Correlation coefficients between $\mathcal{D}^{6,3}(w_{test})$ and other $\mathcal{D}^{6,3,N}(w_i)$'s for $N = 0, \cdots, 5$.

Results

r = 6
wildcard 0,…,5
N = 5

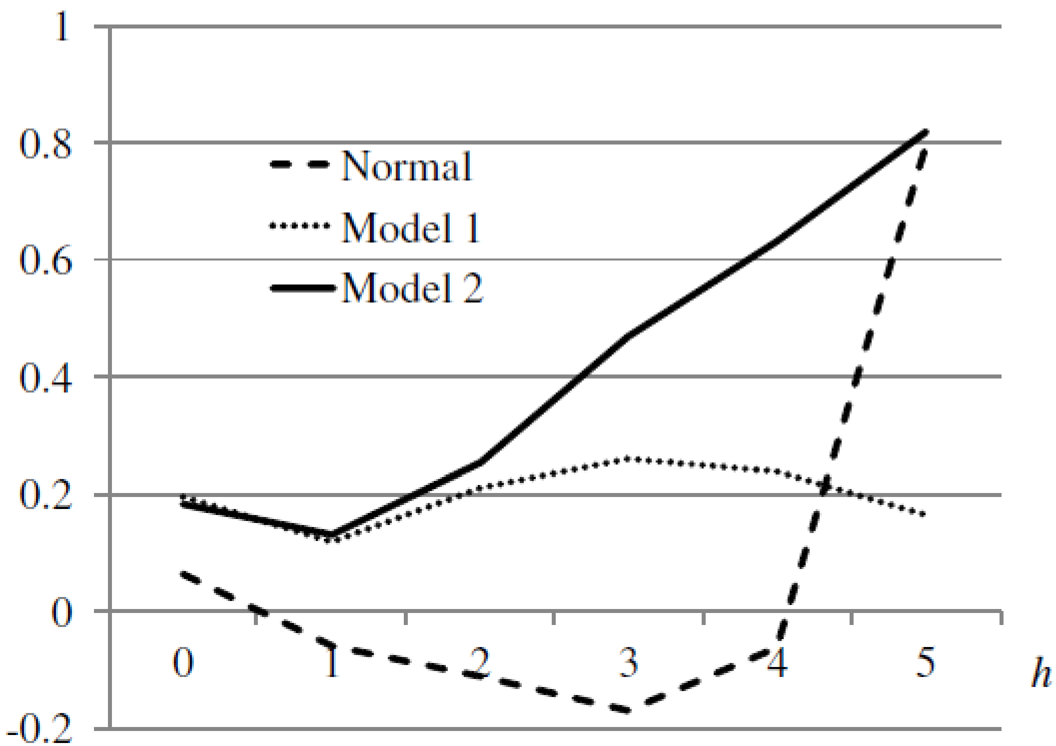


Fig. 9. Correlation coefficients between $\mathcal{D}^{6,h,5}(w_{test})$ and other $\mathcal{D}^{6,h,5}(w_i)$'s for $h = 0, \cdots, 5$.

Results

TABLE III

TOP 10 SEQUENCES IN FAULTY MODEL 1.

Sequence	Specificity
$yaxnnn$	0.446642379
$xnaxnn$	0.385756871
$yaynnn$	0.3695851
$ayannn$	0.363505804
$xaynnn$	0.360871346
$ananna$	0.330605425
$axannn$	0.321329947
$xnaynn$	0.31243332
$axnann$	0.308730344
$yanynn$	0.304838413

TABLE IV

TOP 10 SEQUENCES IN FAULTY MODEL 2.

Sequence	Specificity
$xbxnnn$	0.239204665
$yaxnnn$	0.178624349
$annnya$	0.167122029
$yaynnn$	0.155783797
$xbnynn$	0.144907917
$ayannn$	0.142148264
$yanxnn$	0.141006404
$ynaxnn$	0.132237506
$axnnny$	0.12804503
$anybnn$	0.127656013

Sequence patterns with high (low) singularity are used for the cause of the fault.

Future Work

- Wildcard characters with variable length
- Considering duration of inter-event time
- Online diagnosis
- Using more complex probabilistic models, e.g., context-sensitive models