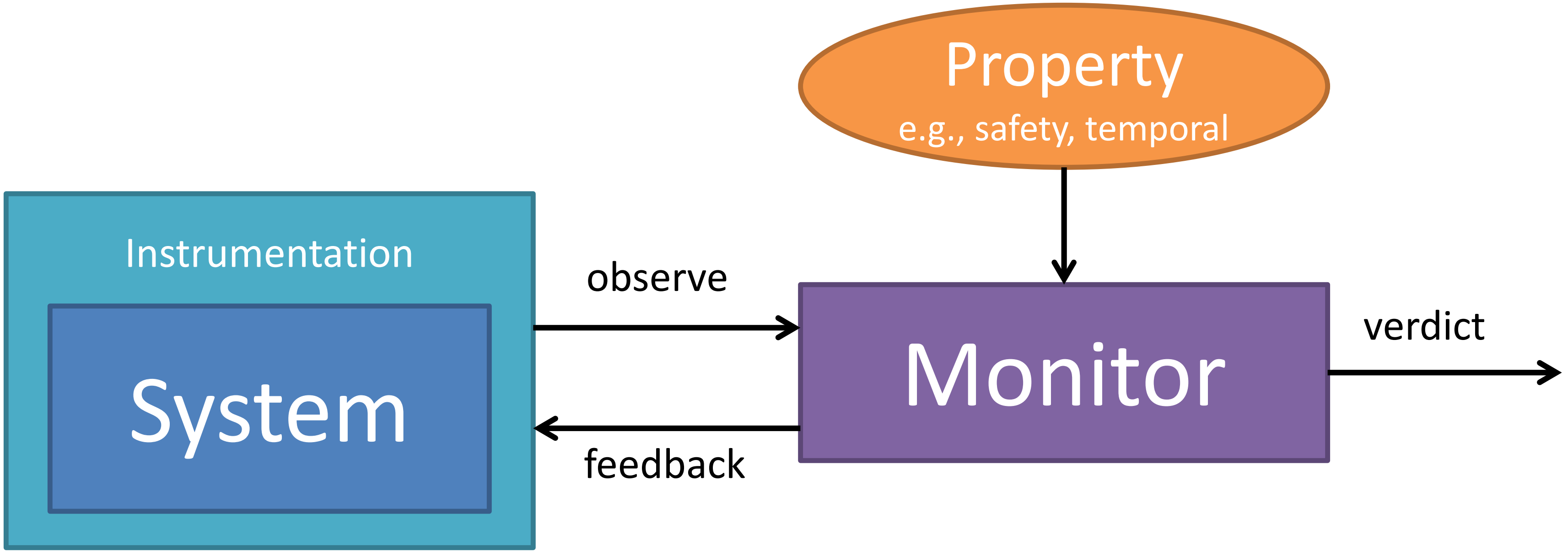
Monitoring Event Frequencies

Bernhard Kragl
IST Austria

Thomas Ferrère Thomas A. Henzinger

Monitoring

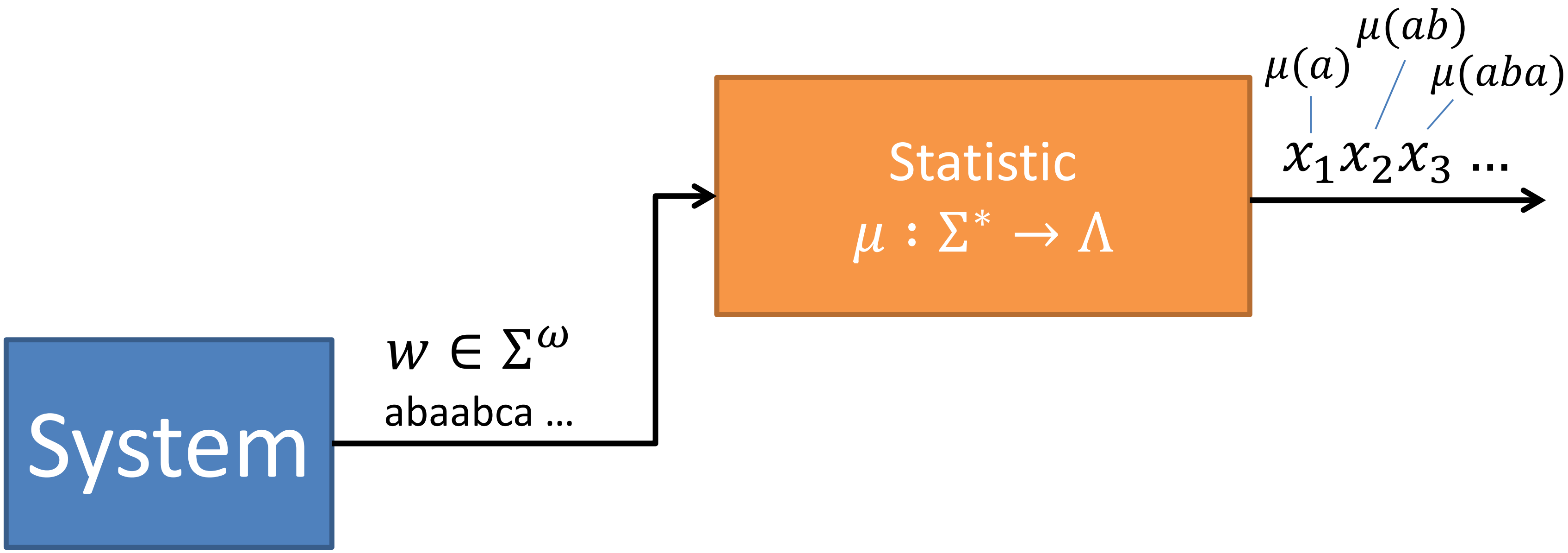
aka Runtime Verification



Performance Monitoring

Quantitative system properties
"Soft" performance indicators

Problem Setting



Example Statistics

- Mode (most frequent value)

$$\mathrm{mode} : \Sigma^* \to \Sigma \cup \{\bot\}$$

$$\mathrm{mode}(w) = a \quad \text{if} \quad |w|_a > |w|_\sigma \quad \text{for all } \sigma \neq a$$

# of occurrences of letter σ in word w

- Median ("middle" value)

$$\mathrm{median} : \Sigma^* \to \Sigma \cup \{\bot\}$$

$$\mathrm{median}(w) = a \quad \text{if}$$

$$\sum_{\sigma > a} |w|_\sigma < \sum_{\sigma \leq a} |w|_\sigma$$

$$\sum_{\sigma < a} |w|_\sigma < \sum_{\sigma \geq a} |w|_\sigma$$

Example Statistics

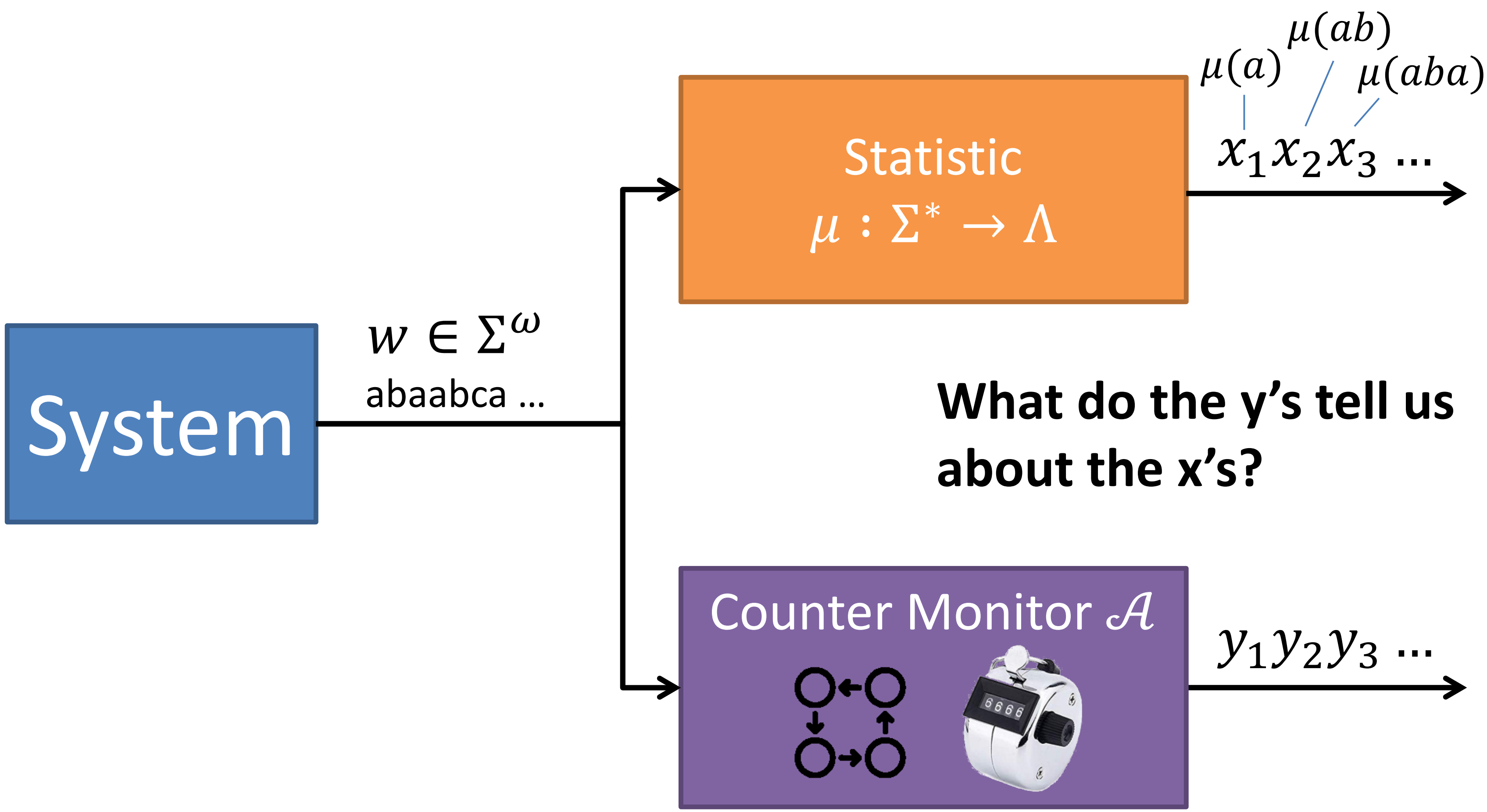
- Mode (most frequent value)

$$\text{mode}(abca) = a$$

- Median ("middle" value)

$$\text{median}(abbcdef) = c$$

Problem Setting



The Monitoring Problem

- Real-Time Monitoring (the past)

  Monitor $\mathcal{A}$ is a *real-time monitor* of statistic μ:

$$[\![\mathcal{A}]\!] = \mu$$

- Limit Monitoring (the future)

  Statistic μ *converges* to value v over random process $\mathcal{P}$ ($\mu(\mathcal{P}) = v$):

$$\mathbb{P}_{w\sim\mathcal{P}}(\lim_{n\to\infty} \mu(w_{1..n}) = v) = 1$$

  Monitor $\mathcal{A}$ is a *limit monitor* of statistic μ over random processes $\mathcal{P}$:

$$[\![\mathcal{A}]\!](\mathcal{P}) = v \quad \Leftrightarrow \quad \mu(\mathcal{P}) = v$$

Precise Real-Time Monitoring

- Mode

  Naïve algorithm: one counter c_σ for every letter σ

  IPv4 protocol: 4,294,967,296 addresses

  UTF-8 encoding: 1,112,064 code points

- Median
  - Offline: selection, median of means (approximate), Mitzenmacher & Upfal (randomized)
  - Online: two heaps for lower and higher halve of values
  - Real-time: no known algorithm

Precise Real-Time Monitoring is Expensive

- Equivalence relation $\equiv_\mu$ over words

$$w \equiv_\mu w' \quad \text{if} \quad \mu(wu) = \mu(w'u) \text{ for all words } u$$

- *Σ-counting* statistic μ

$$w \equiv_\mu w' \quad \Rightarrow \quad \exists n \in \mathbb{Z}\, \forall \sigma \in \Sigma : |w|_\sigma = |w'|_\sigma + n$$

$aabc \equiv_{\mathrm{mode}} a$	over $\Sigma = \{a, b, c\}$	distances: 1
$aabc \not\equiv_{\mathrm{mode}} a$	over $\Sigma = \{a, b, c, d\}$	distances: 1, 0

Precise Real-Time Monitoring is Expensive

- Equivalence relation $\equiv_\mu$ over words

$$w \equiv_\mu w' \quad \text{if} \quad \mu(wu) = \mu(w'u) \text{ for all words } u$$

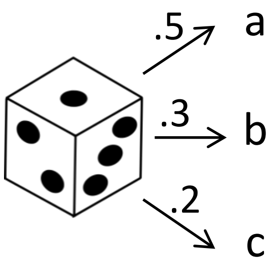
- *Σ-counting* statistic μ

$$w \equiv_\mu w' \quad \Rightarrow \quad \exists n \in \mathbb{Z}\, \forall \sigma \in \Sigma : |w|_\sigma = |w'|_\sigma + n$$

Theorem. Both mode and median are Σ-counting.

Theorem. Any real-time monitor of a Σ-counting statistic requires at least $|\Sigma|$ counters.

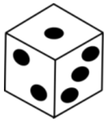
Mode over i.i.d. words



w	c b b a b a c a a b c a c a a a $\cdots$
mode	c - b b b b b - a - - a a a a a $\cdots$

Theorem. w has mode a iff a is unique max. of .

$$\delta_{w_i}(\sigma) = \begin{cases} 1 \text{ if } w_i = \sigma \\ 0 \text{ otherwise} \end{cases} \qquad \mathbb{E}(\delta_{w_i}) =$$ 

$$\underbrace{\frac{\sum_{i=1}^{n} \delta_{w_i}}{n}}_{\text{Empirical distribution}} \xrightarrow{a.s.}$$ 

Strong law of large numbers

Efficient Mode Monitor

- Partition word into chunks of increasing size
- In each chunk, count two letters:
  - Mode candidate x
  - Mode contender y
- After each chunk, keep winner in x and replace y

n	1	2	3	4	5	6 $\cdots$
i	1	1 2	1 2 3	1 2 3 4	1 2 3 4 5	1 $\cdots$
σ	c	b b	a b a	c a a b	c a c a a	a $\cdots$
x	c	c	b	a	a	a $\cdots$
y	c	b	a	c	c	a $\cdots$
c_x	1	0 0	0 1 1	0 1 2 2	0 1 1 2 3	1 $\cdots$
c_y	1	1 2	1 1 2	1 1 1 1	1 1 2 2 2	1 $\cdots$

Only four counters: c_x, c_y, n, i

Monitor correctness over i.i.d. words

Theorem. Our algorithm limit-monitors the mode.

Let a be the mode of an i.i.d. random ω-word.

$$
\begin{aligned}
&\mathbb{P}(\text{eventually always } x = a) \\
\geq\ &\mathbb{P}\big(\text{eventually (always } c_\sigma < c_a \wedge \text{eventually } y = a)\big) \\
\geq\ &\mathbb{P}(\text{always}_{\geq n}\ c_\sigma < c_a \wedge \text{eventually}_{\geq n}\ y = a)
\end{aligned}
$$

$$
\begin{aligned}
&\mathbb{P}(\text{always}_{\geq n}\ c_\sigma < c_a \wedge \text{eventually}_{\geq n}\ y = a) \\
\geq\ &\mathbb{P}(\text{always}_{\geq n}\ c_\sigma < c_a) \\
=\ &1 \quad \text{as} \quad n \to \infty
\end{aligned}
$$

Generalized strong law of large numbers

What we actually do in the paper

- More general setting

  Connected Markov chains

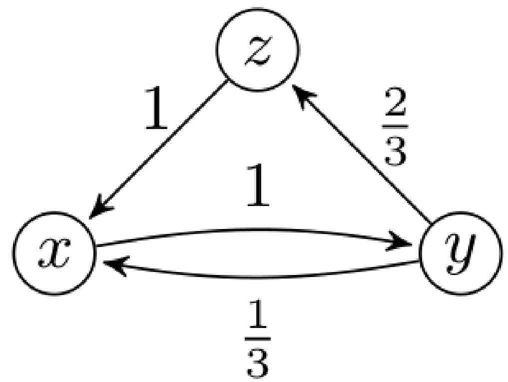
  Key technical result

  Generalized ergodic theorem

- Monitoring general frequency properties

  Frequency formula → Monitor

An Ergodic Theorem over Infixes



ω-word	x	y	z	x	y	z	x	y	x	y	z	x	y	z	x	y	...
Prefixes	0	.5	.33	.25	.4	.33	.29	.38	.33	.4	.36	.33	.38	.36	.33	.38	$\xrightarrow{\text{a.s.}} \frac{3}{8}$
Infixes	0		.5			.33				.5					.2		$\xrightarrow{\text{a.s.}} \frac{3}{8}$

Theorem. The frequency of visiting a state s converges a.s. to the inverse of the expected return time of s, over arbitrary infixes of increasing length.

General Frequency Properties

- Frequency formula ϕ

  Boolean combination of atomic formulas of the form

$$\sum_{\sigma \in \Sigma} f_\sigma \cdot \alpha_\sigma > \alpha$$

frequency term   integer coefficients

"No event occurs 100 times more than any other event."

$$\bigwedge_{\substack{a,b \in \Sigma \\ a \neq b}} f_a < 100 \cdot f_b$$

- Monitoring algorithm

  Evaluate different atomic subformulas of ϕ over different infixes and partially evaluate ϕ in finite state part.

$$\cdots \underbrace{\underbrace{w_{n,1}}_{\phi_1} \underbrace{w_{n,2}}_{\phi_2} \cdots \underbrace{w_{n,k}}_{\phi_k}}_{\phi} \cdots$$

Summary

- Real-time monitoring of important frequency properties can be expensive

- Limit monitoring can be much more efficient

$|\Sigma|$ vs 4 counters for the mode