Monitoring Algorithmic Fairness

under Partial Observations.

Thomas A. Henzinger | Konstantin Kueffner | Kaushik Mallik

Institute of Science and Technology Austria

TLDL.

Too Long Didn’t Listen.

$$f : \Sigma^n \to \mathbb{R}$$

some function

$$\overrightarrow{X} := \left(X_t\right)_{t>0}$$

a stochastic process.

$$\mathbb{E}(f(X_{t:t+n}))$$

Want to compute.

$$\mathbb{P}$$

Unknown.

$$\vec{x}_t := x_1, \ldots, x_t$$

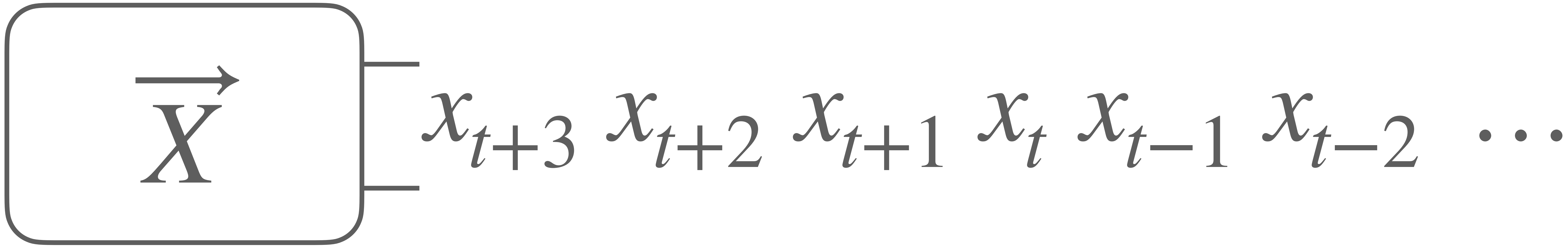
Realisation.

$$\mathbb{P} \in \mathscr{P}$$

Assumptions.

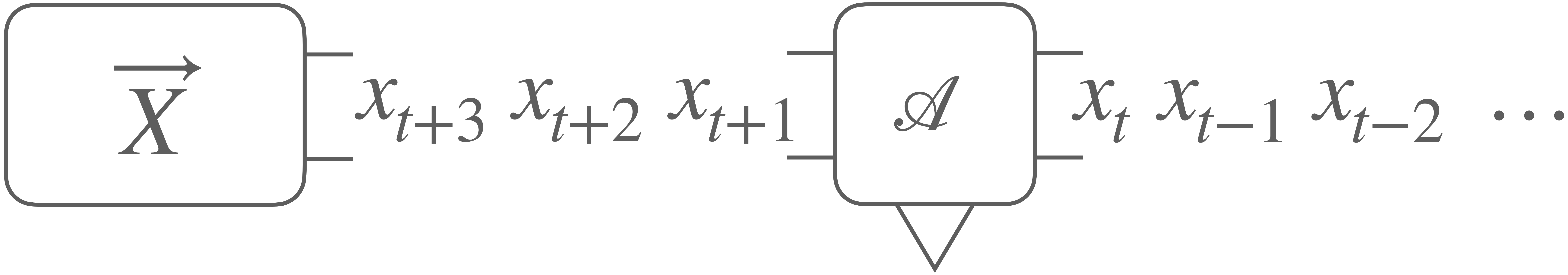
$$\mathbb{E}(f(X_{t:t+n})) \in [\hat{l}(\vec{x}_t), \hat{u}(\vec{x}_t)]$$

Estimate.

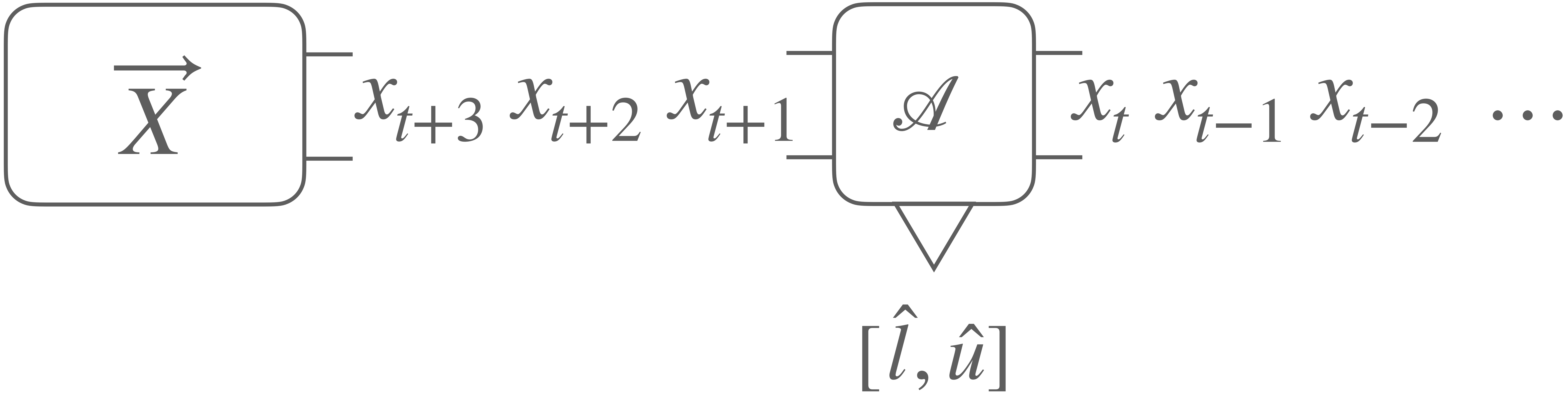


$\overrightarrow{X}$

$x_{t+3}\ x_{t+2}\ x_{t+1}\ x_t\ x_{t-1}\ x_{t-2}\ \cdots$


$\overrightarrow{X}$
x_{t+3} x_{t+2} x_{t+1}
$\mathscr{A}$
x_t x_{t-1} x_{t-2} $\cdots$

$\mathbb{E}(f(\overrightarrow{X}_{t:t+n})) \in \mathscr{A}(\overrightarrow{x_t})$ with probability $1-\delta$


$\overrightarrow{X}$
x_{t+3} x_{t+2} x_{t+1}
$\mathscr{A}$
x_t x_{t-1} x_{t-2} $\cdots$
$[\hat{l}, \hat{u}]$

What we did:

Proposed a monitor for estimating such properties over a restricted class of Hidden Markov Models.

Monitoring Algorithmic Fairness

under Partial Observations.

Extended Cut

Thomas A. Henzinger | Konstantin Kueffner | Kaushik Mallik

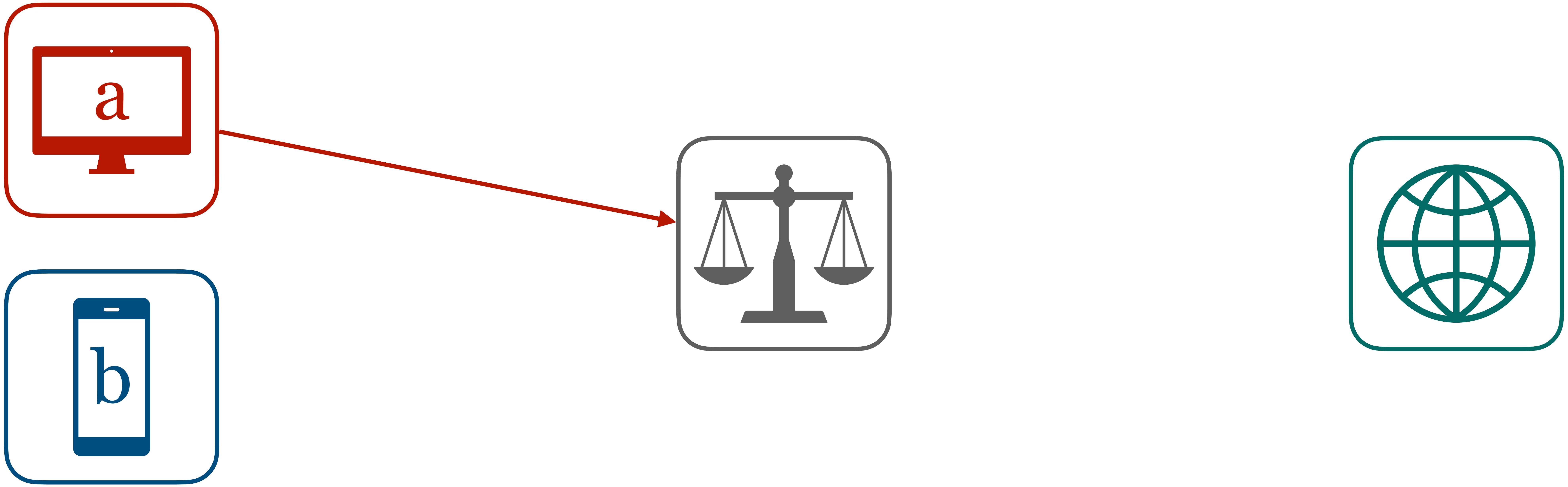
Institute of Science and Technology Austria

Example.

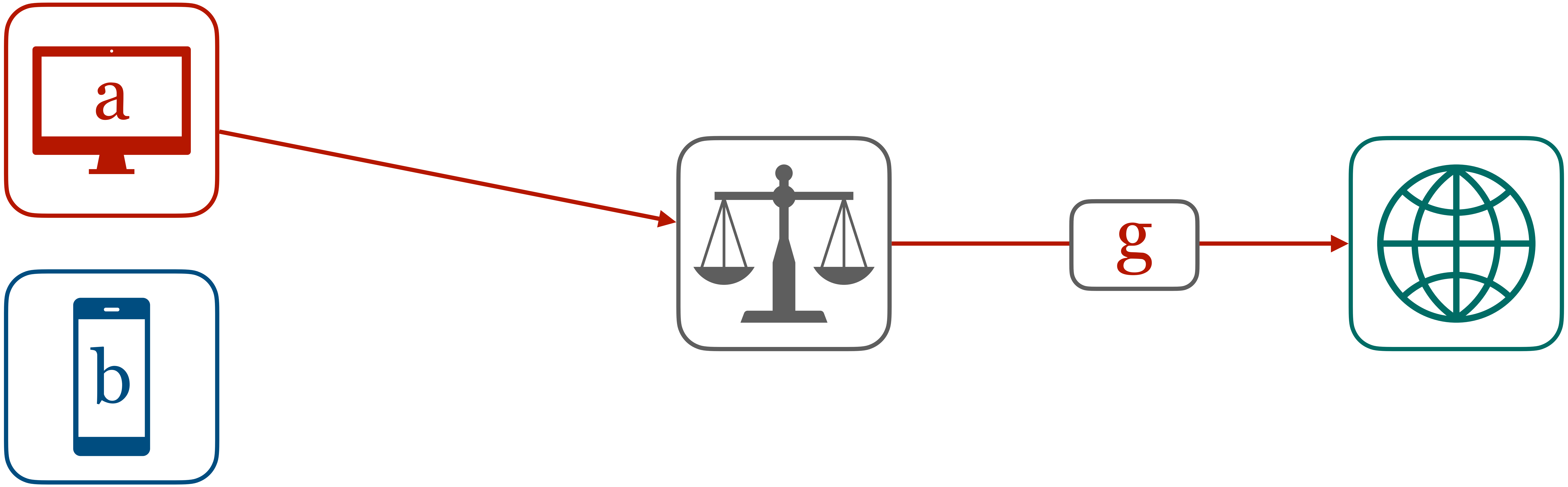
A simple resource allocation problem.

a

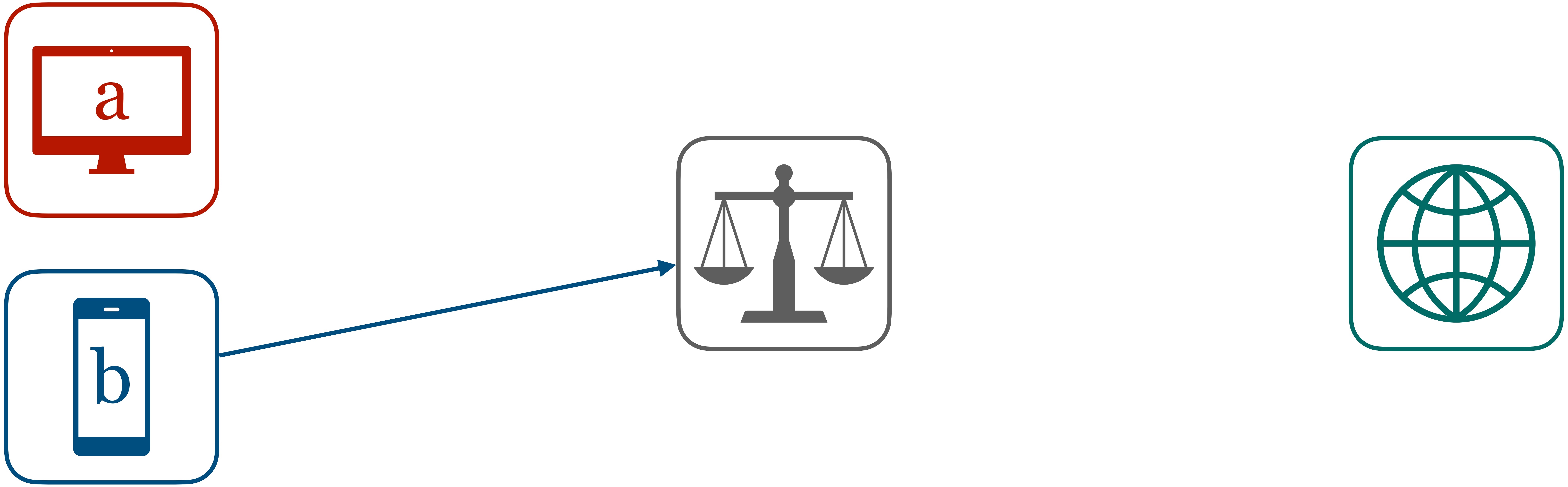
b



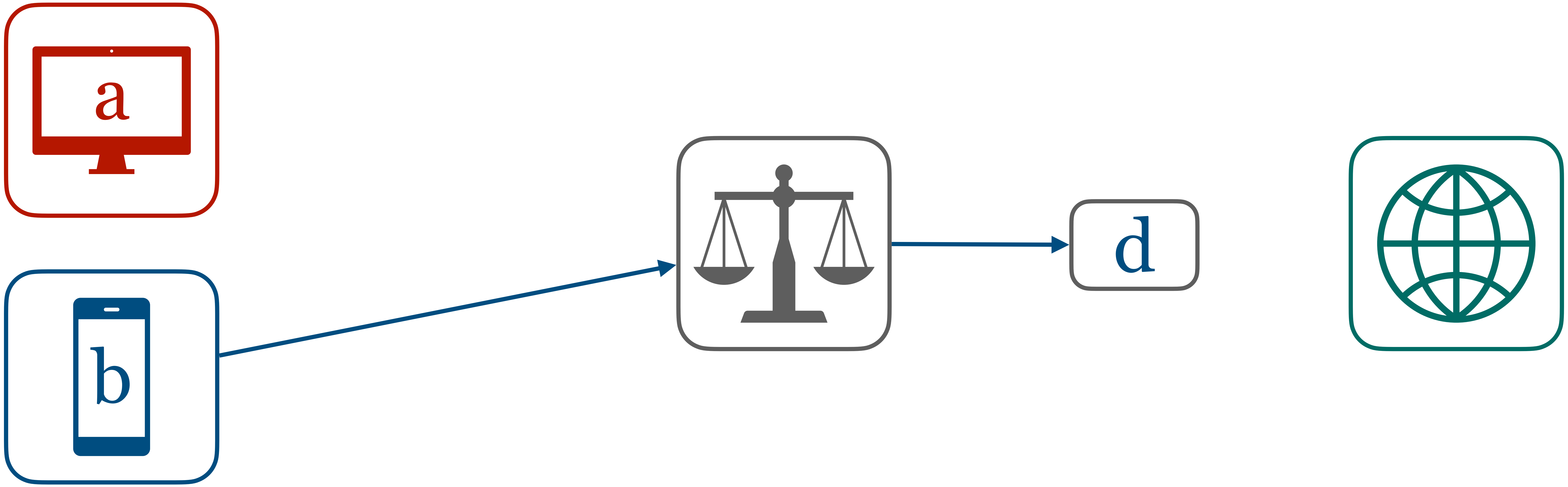
a



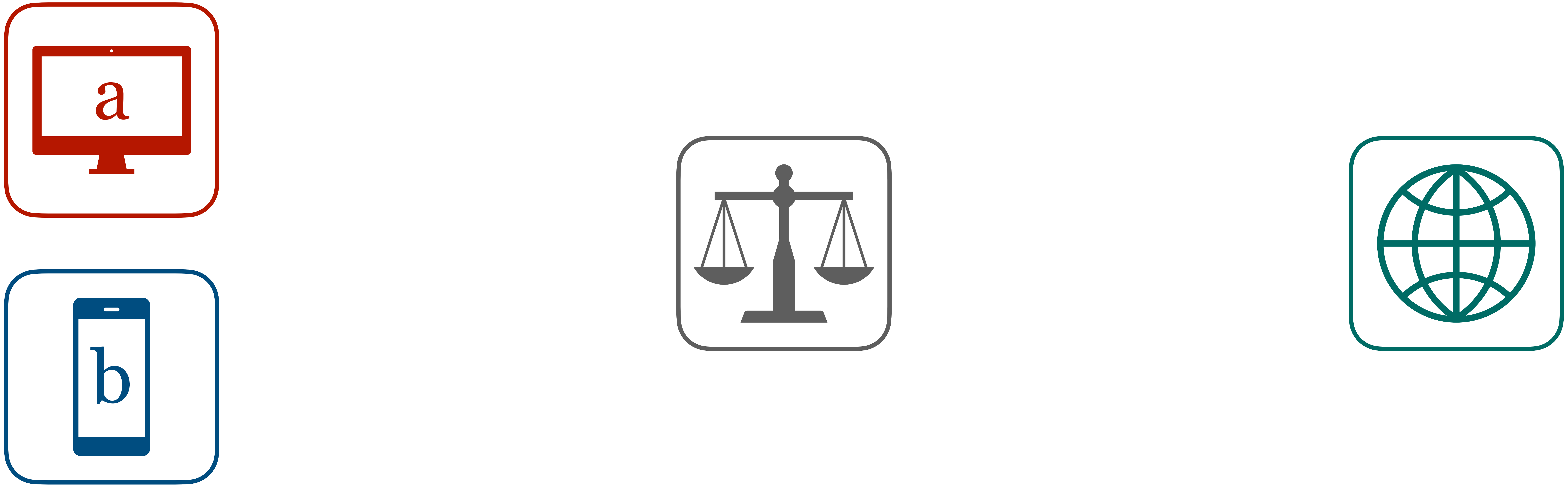
a g



a g b



a g b d



a g b d b d a g b g . . . . . .

Is the arbiter “fair”.

Many possible interpretations.

$$(\Box\Diamond a \rightarrow \Box\Diamond g) \wedge (\Box\Diamond b \rightarrow \Box\Diamond g)$$

~~$(\Box\Diamond a \to \Box\Diamond g) \land (\Box\Diamond b \to \Box\Diamond g)$~~

$$\mathbb{P}(\mathbf{g} \mid a) - \mathbb{P}(\mathbf{g} \mid b)$$

$$\mathbb{P}(\mathbf{g}) - \mathbb{P}(\mathbf{g})$$

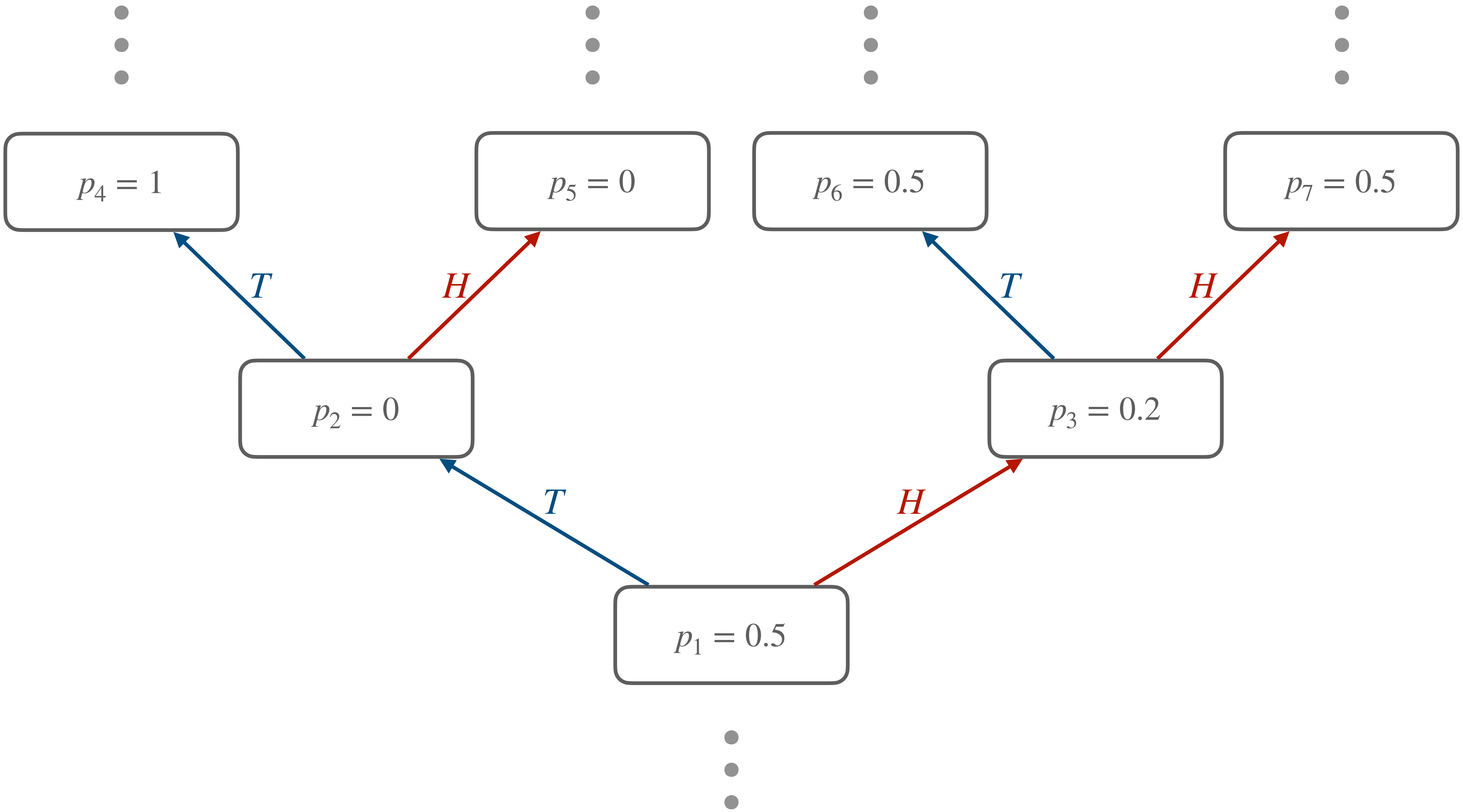
Ok. Let’s try to …

… give meaning to those probabilities.

$$\mathbb{P}(\mathrm{g}) \Rightarrow \mathbb{P}(\mathrm{H})$$

$$\mathbb{P}(\mathrm{d}) \Rightarrow \mathbb{P}(\mathrm{T})$$

$$\ldots, X_{t-3}, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, X_{t+3}, \ldots$$

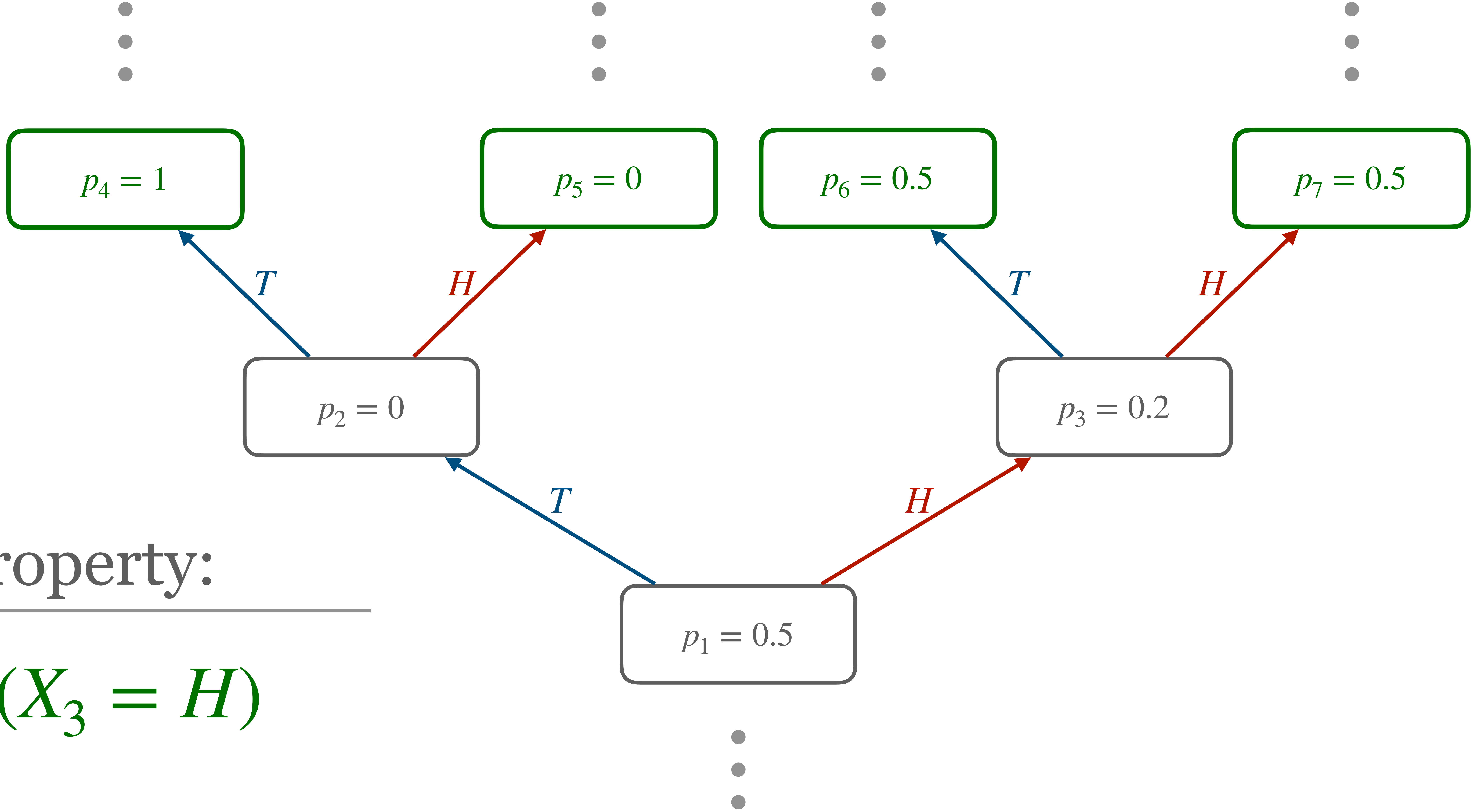

$p_4 = 1$
$p_5 = 0$
$p_6 = 0.5$
$p_7 = 0.5$
T
H
T
H
$p_2 = 0$
$p_3 = 0.2$
T
H
$p_1 = 0.5$

How fair is it…

at time t?

$$\ldots, X_{t-3}, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, X_{t+3}, \ldots$$

$$\mathbb{P}(X_t = \mathbf{H})$$



Property:

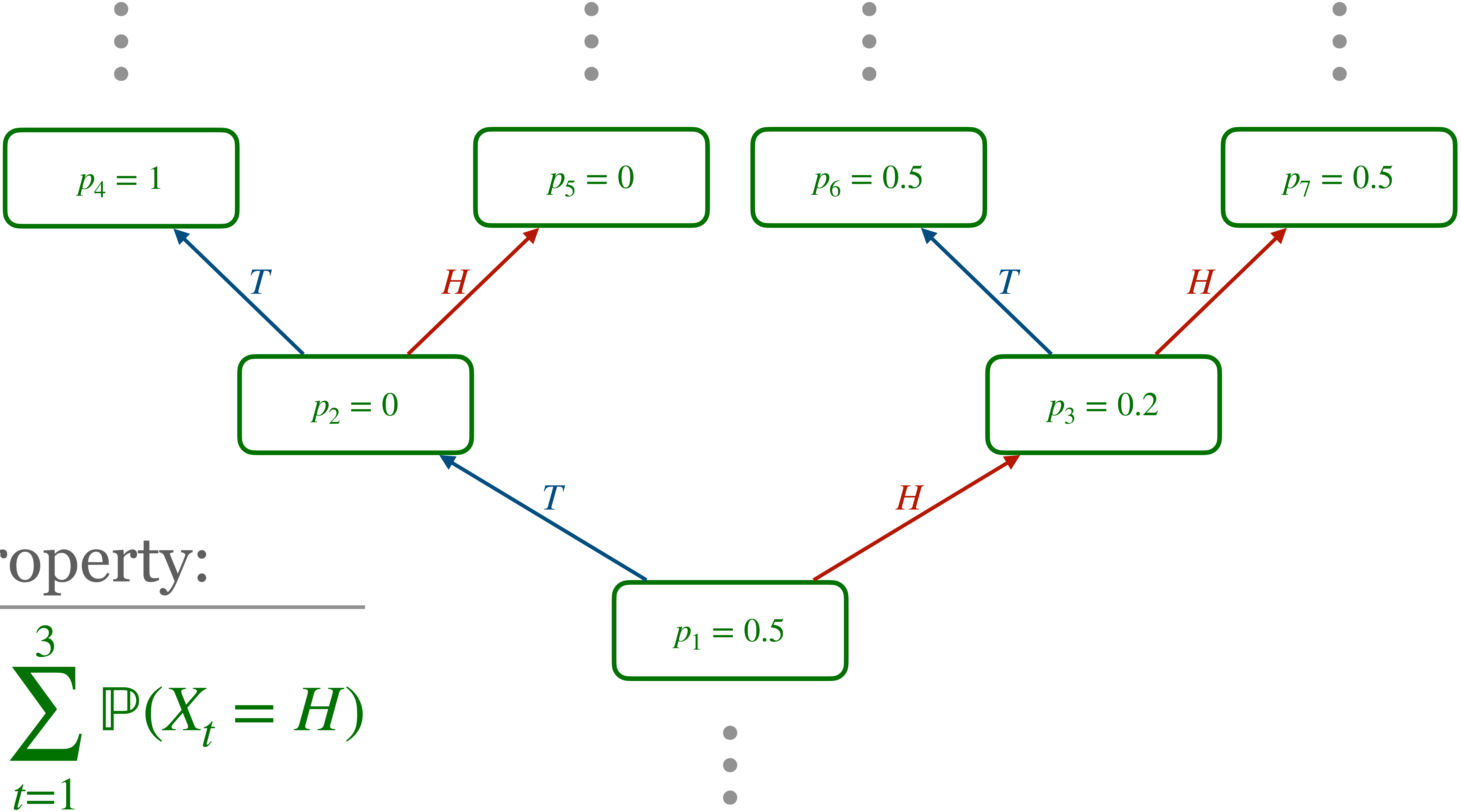
$\mathbb{P}(X_3 = H)$

How fair is it…

on average?
(up to time t)

$$\ldots, \underline{X_{t-3}, X_{t-2}, X_{t-1}, X_t,} \; X_{t+1}, X_{t+2}, X_{t+3}, \ldots$$

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{P}(X_k = \mathrm{H})$$



Property:

$$\frac{1}{3}\sum_{t=1}^{3}\mathbb{P}(X_t = H)$$

How fair is it…

on average?
(in the limit.)

$$\ldots, X_{t-3}, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, X_{t+3}, \ldots$$

$$\lim_{t\to\infty} \frac{1}{t}\sum_{k=1}^{t} \mathbb{P}(X_k = \mathrm{H})$$

Static Properties

The “classic” perspective.

Dynamic Properties

The runtime perspective.

$$x_3 = T$$

$$x_2 = H$$

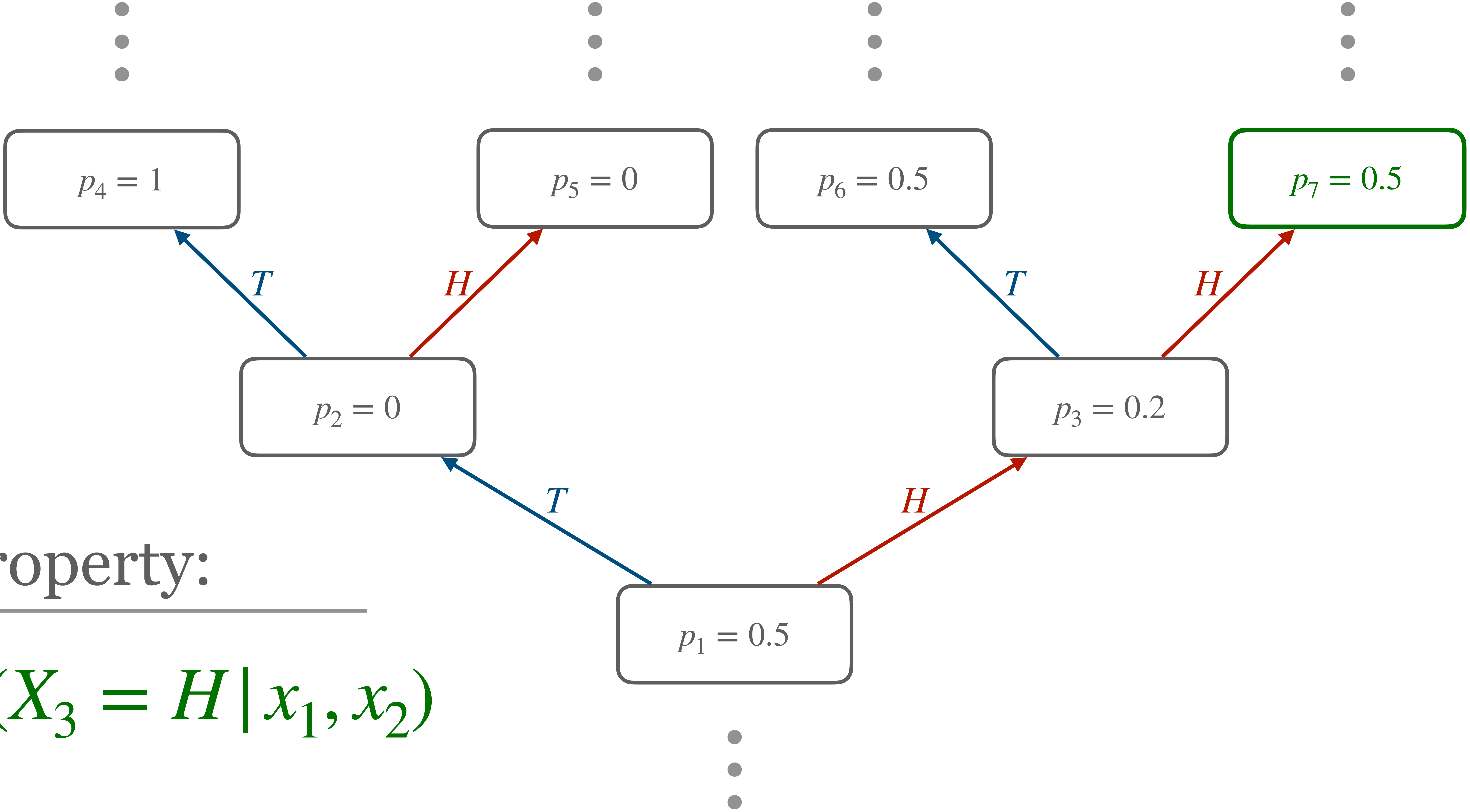
$$x_1 = H$$

How fair is it…

at this very moment?

$$\ldots, X_{t-3}, X_{t-2}, X_{t-1}, X_t, X_{t+1}, X_{t+2}, X_{t+3}, \ldots$$

$$\mathbb{P}(X_t = \mathbf{H} \mid \vec{x}_{t-1})$$



Property:

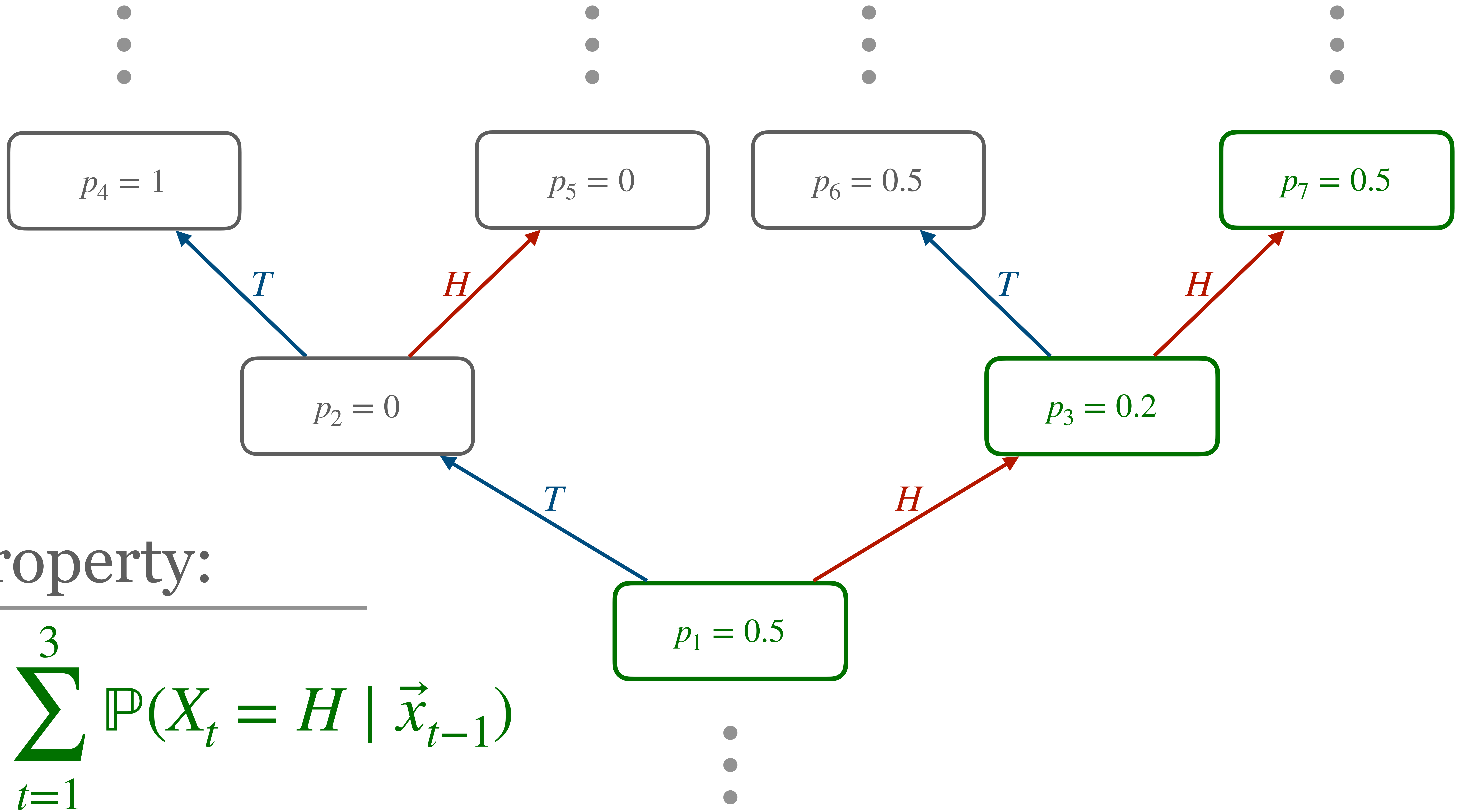
$$\mathbb{P}(X_3 = H \,|\, x_1, x_2)$$

How fair was it…

in the past on average?

$$\dots, \underline{X_{t-3}, X_{t-2}, X_{t-1}, X_t,}\ X_{t+1}, X_{t+2}, X_{t+3}, \dots$$

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{P}(X_k = \mathrm{H} \mid \vec{x}_{k-1})$$



Property:

$$\frac{1}{3}\sum_{t=1}^{3}\mathbb{P}(X_t = H \mid \vec{x}_{t-1})$$

Properties.

Let's be a little bit more general.

$$f : \Sigma^* \to \mathbb{R}$$

Atomic Function

$$\overrightarrow{X} = \left(X_t\right)_{t>0}$$

Stochastic Process

Arithmetic Expressions over:

$\mathbb{E}(f(\overrightarrow{X}_t))$ $\ldots$ $\mathbb{E}(f(\overrightarrow{X}_t) \mid \overrightarrow{X}_t)$

Static *Dynamic*

Arithmetic Expressions over:

$$\mathbb{E}(f(\overrightarrow{X}_t)) \qquad \ldots \qquad \mathbb{E}(f(\overrightarrow{X}_t) \mid \overrightarrow{X}_t)$$

Static *Dynamic*

Arithmetic Expressions over:

$$\boxed{\mathbb{E}(f(\overrightarrow{X}_t))} \quad \ldots \quad \mathbb{E}(f(\overrightarrow{X}_{t:t+n}) \mid \overrightarrow{X}_t)$$

Static

Dynamic

I know….a bit ironic.

For more on dynamic properties see:
Runtime Monitoring of Dynamic Fairness Properties (FAccT23)

System.

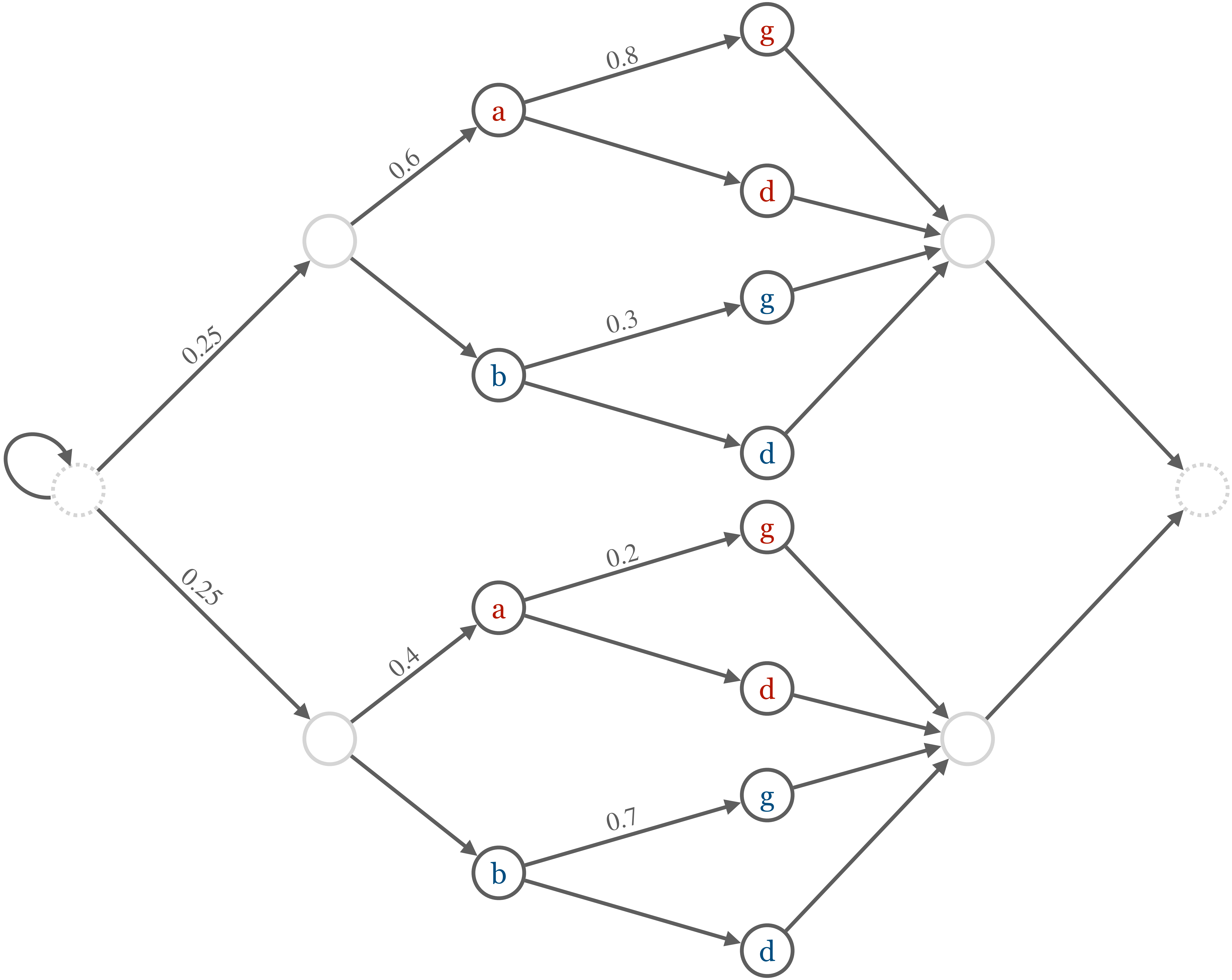
What assumptions do we make?

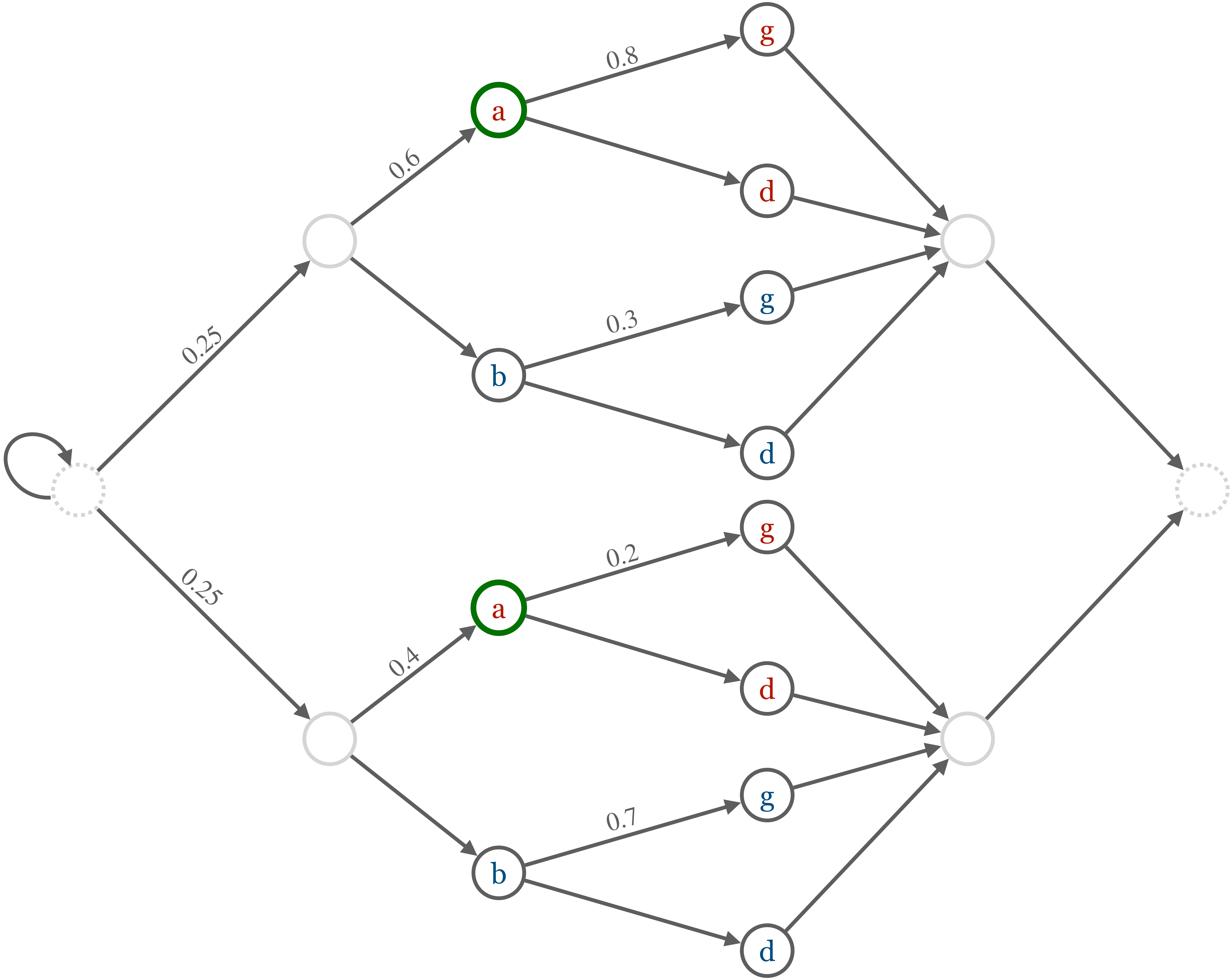
We assume…

… the system is a stationary, aperiodic, irreducible, labelled Markov chain with known mixing time τ_{mix}.

Labelled Markov chain.

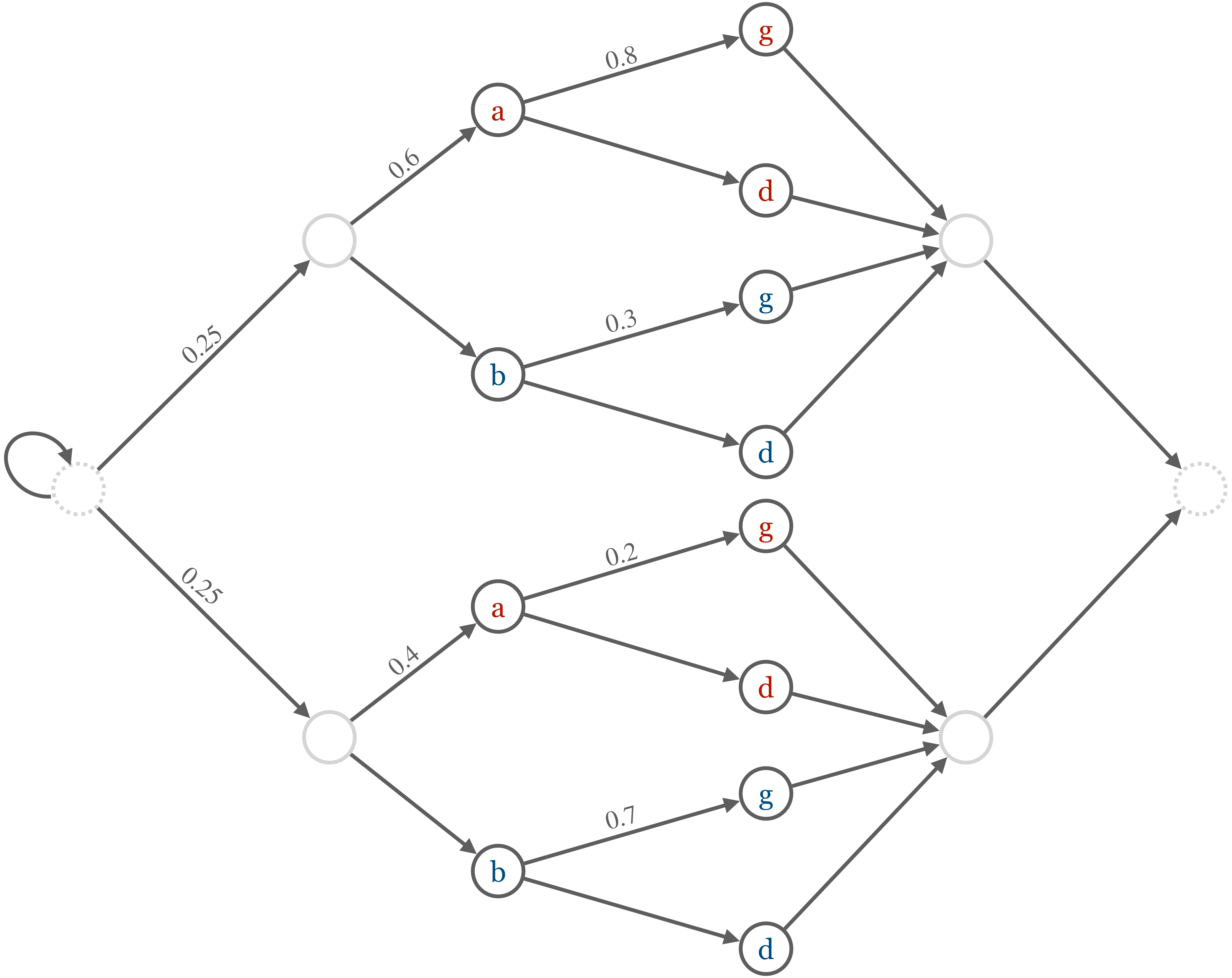
Markov chain where states deterministically map to observables.

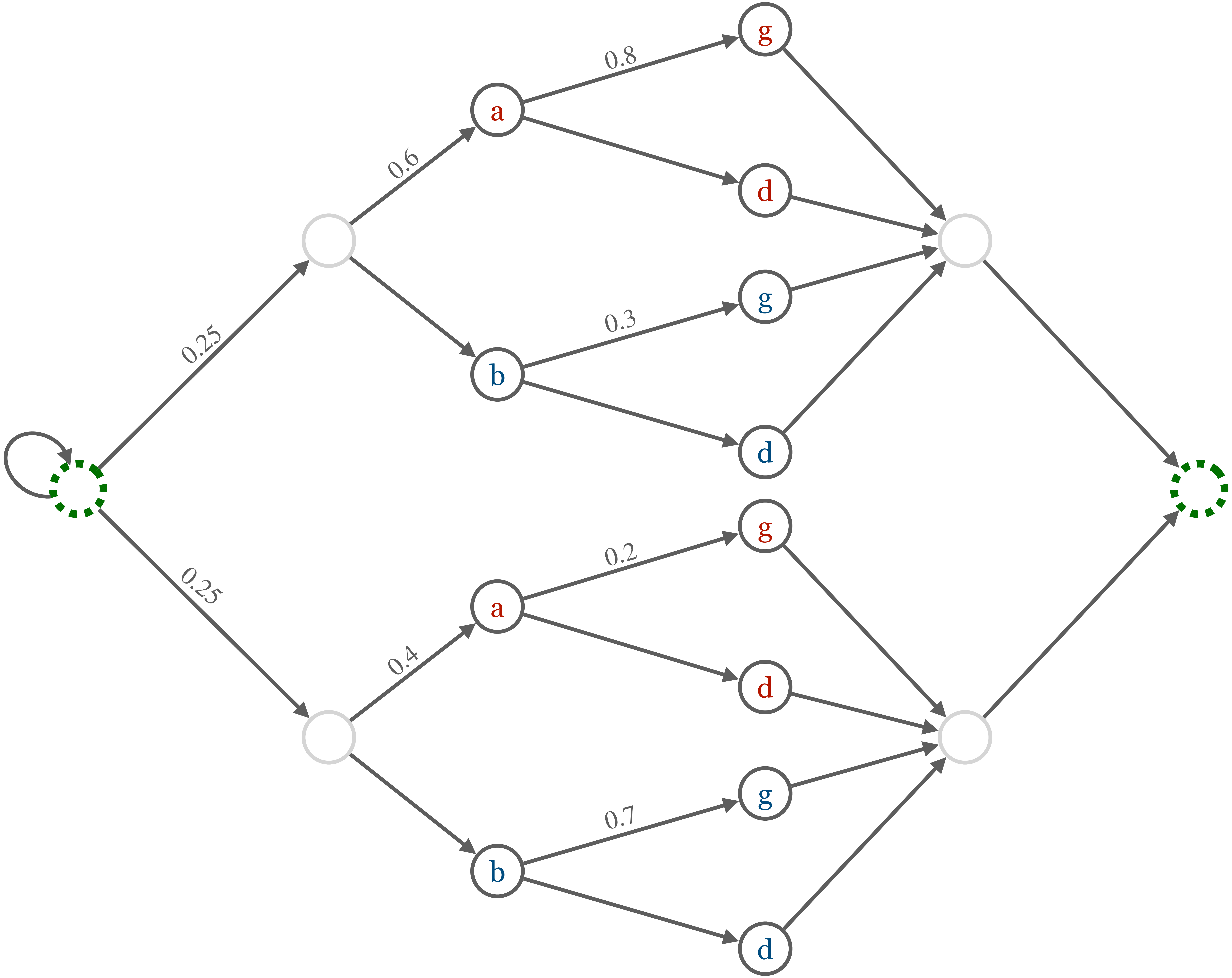

0.25
0.25
0.6
0.4
a
b
a
b
0.8
0.3
0.2
0.7
g
d
g
d
g
d
g
d


0.8
g
a
0.6
d
0.25
g
0.3
b
d
g
0.2
0.25
a
0.4
d
g
0.7
b
d

Irreducible.

The underlying graph is a strongly connected component.


0.25
0.25
0.6
0.4
a
b
a
b
0.8
0.3
0.2
0.7
g
d
g
d
g
d
g
d

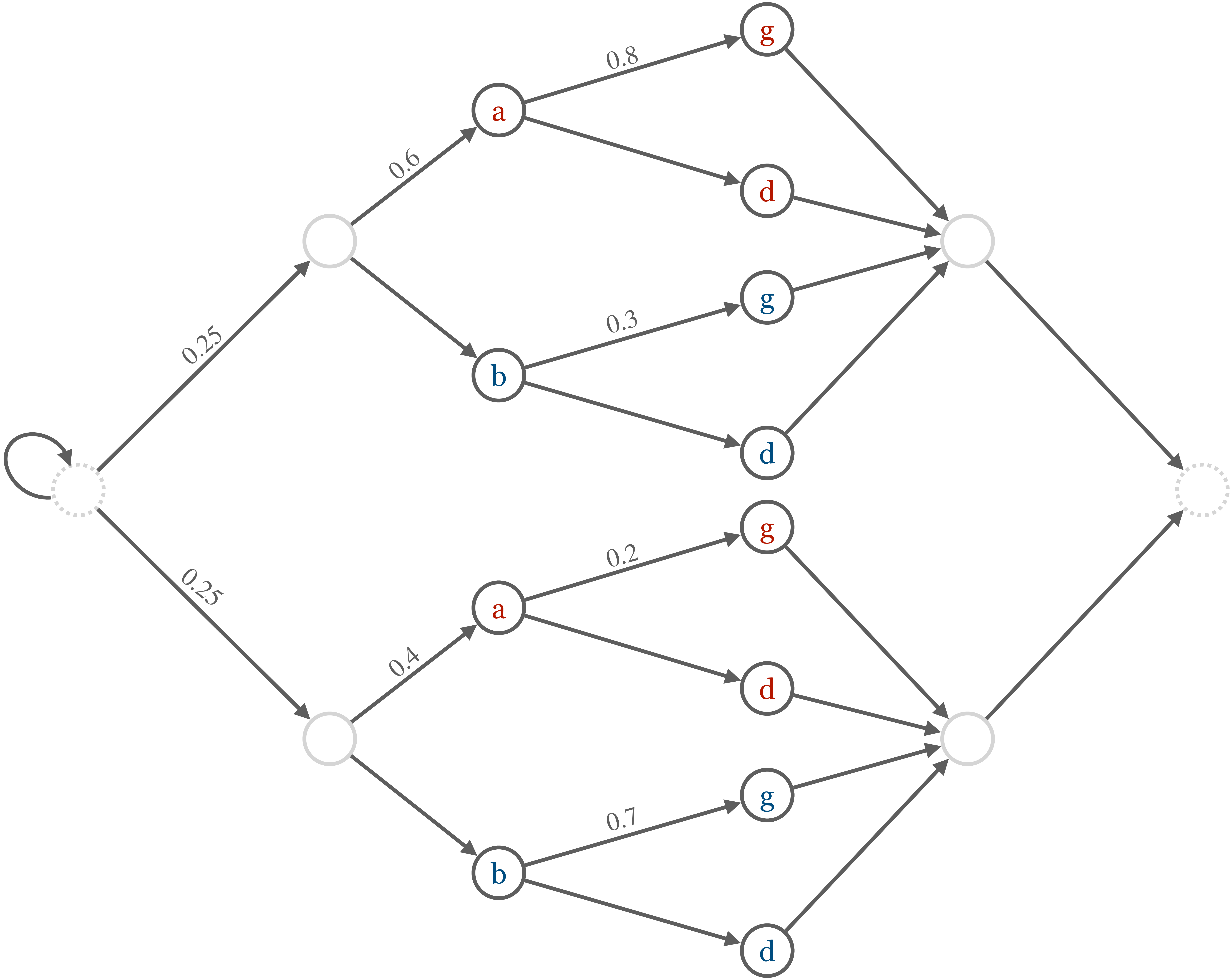

0.8
g
a
0.6
d
0.25
g
0.3
b
d
0.2
g
0.25
a
0.4
d
g
0.7
b
d

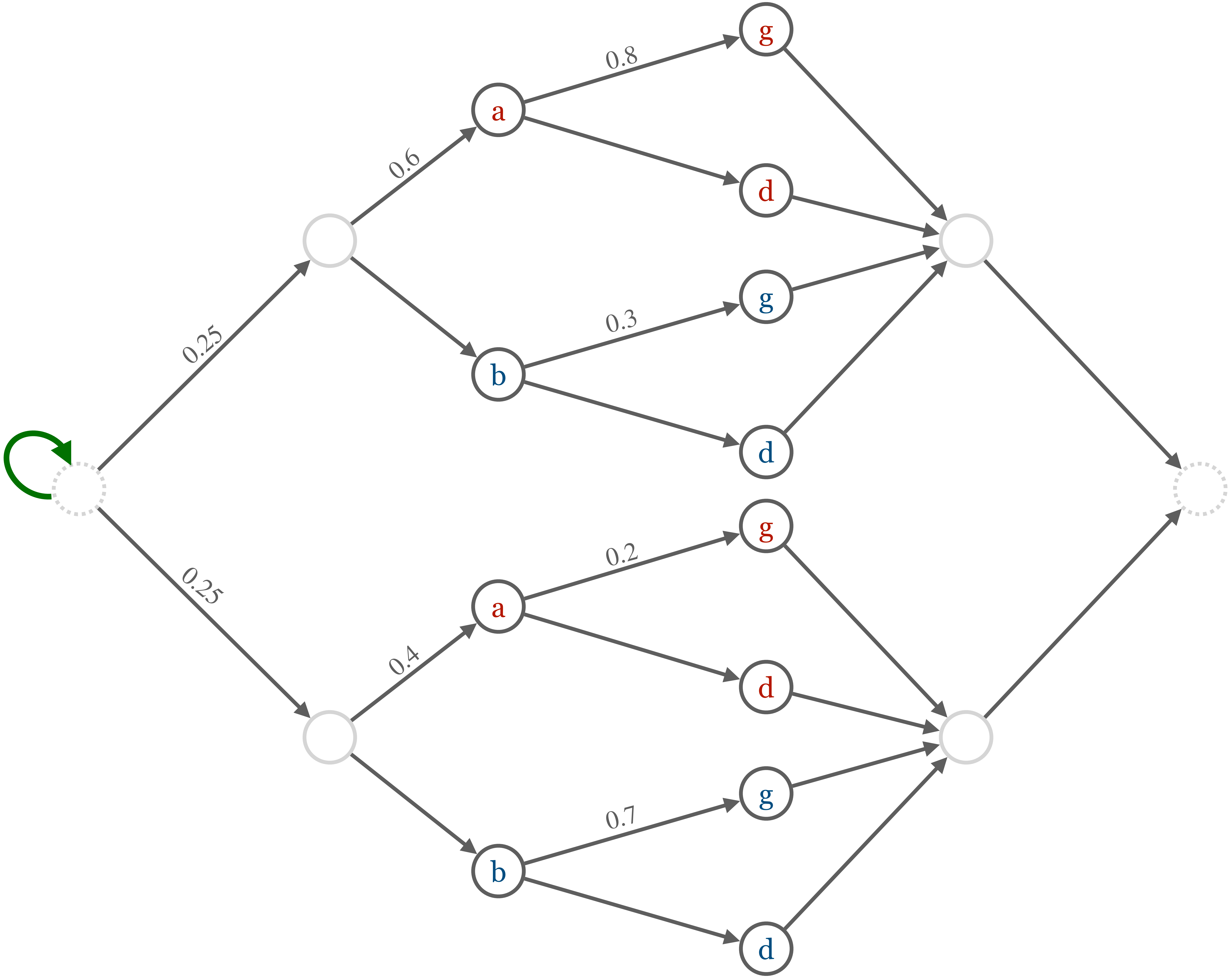
Aperiodic.

You can return to the same state in (almost) arbitrary number of steps.

$\exists q \in Q :$

$$\gcd\{n \in \mathbb{N} \mid \mathbb{P}(X_n = q \mid X_1 = q) > 0\} = 1$$

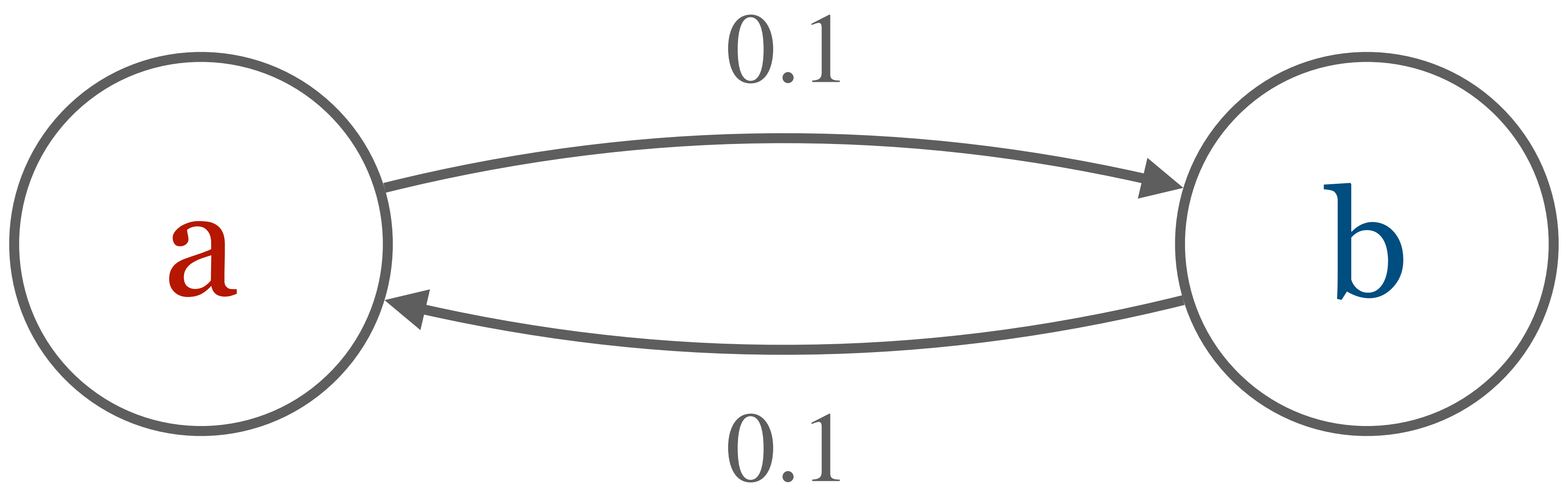

g
0.8
a
d
0.6
0.25
b
0.3
g
d
g
0.2
0.25
a
d
0.4
g
0.7
b
d


0.25
0.25
0.6
0.4
a
b
a
b
0.8
0.3
0.2
0.7
g
d
g
d
g
d
g
d

Stationarity.

…the distribution over states does not change.

$$\pi = \pi \cdot P$$

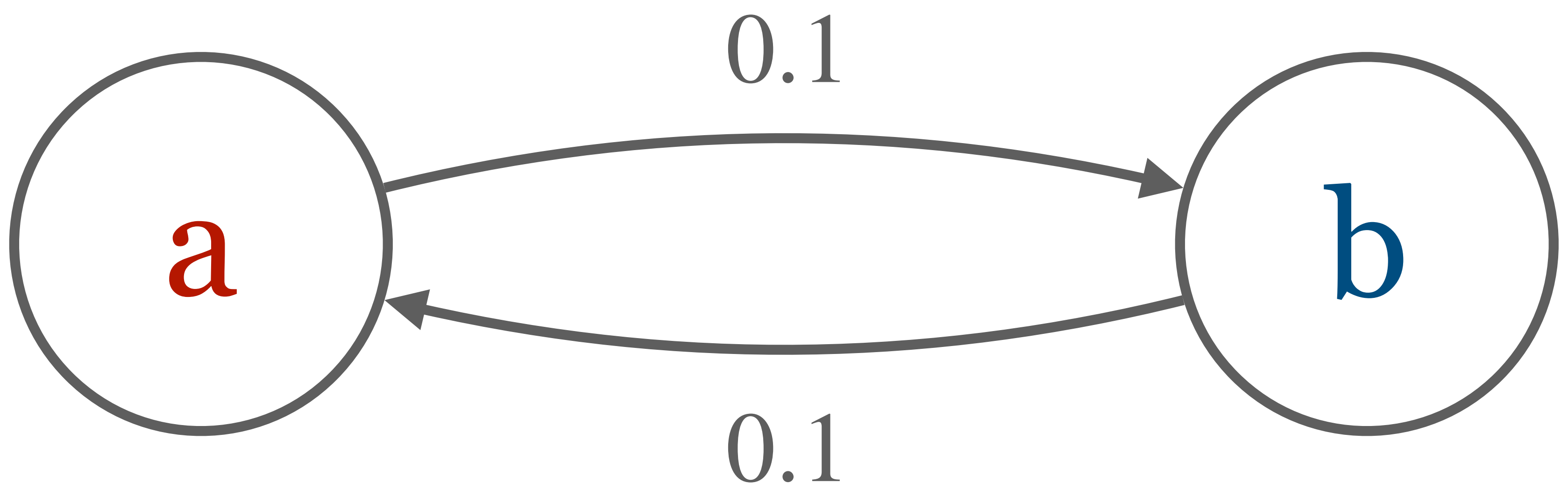


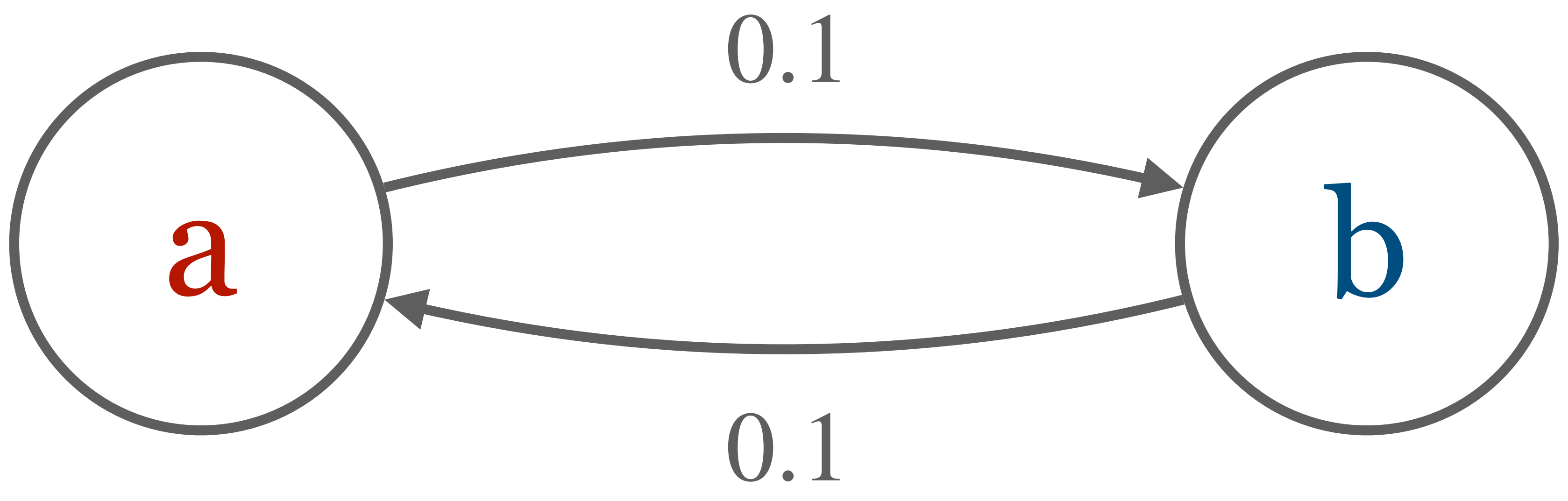
$$\begin{pmatrix}0.5\\0.5\end{pmatrix}\cdot\begin{pmatrix}0.9 & 0.1\\0.1 & 0.9\end{pmatrix}=\begin{pmatrix}0.5\\0.5\end{pmatrix}$$

Mixing Time.

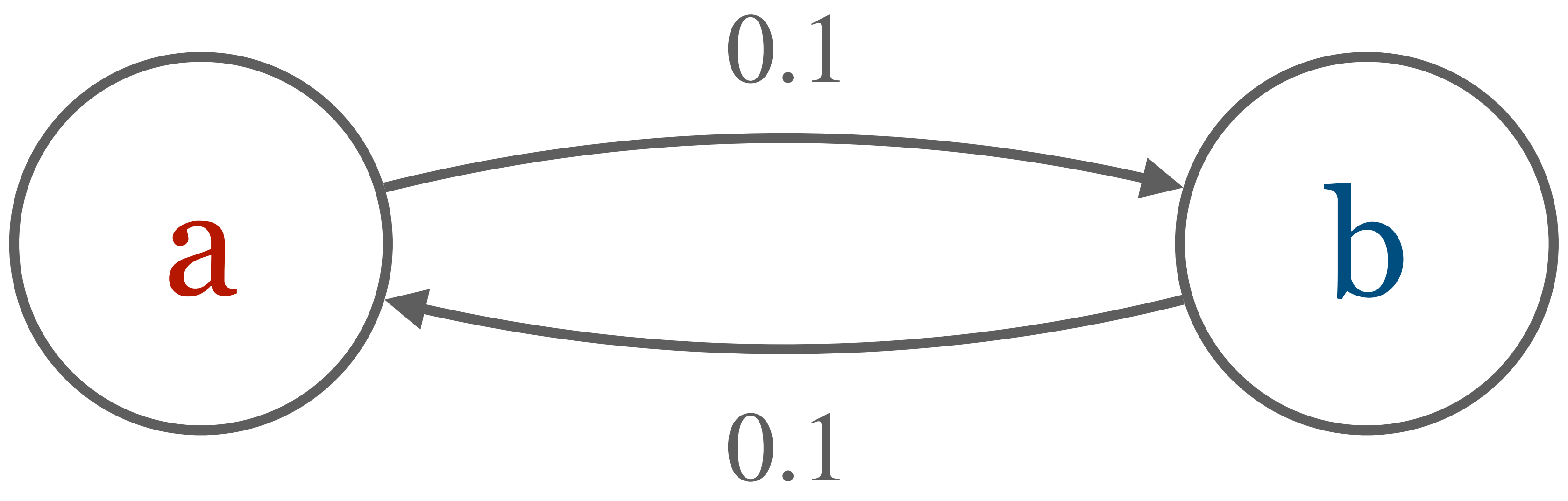
Any irreducible, aperiodic Markov chain eventually reaches its stationary distribution. This time is the mixing time.

$$\tau_{mix}(\varepsilon) = \min_{t} \left\{ \sup_{\mu} \|\mu \cdot P^t - \pi\|_{TV} \leq \varepsilon \right\}$$


0.1
a
b
0.1



$$\begin{pmatrix}1\\0\end{pmatrix}\cdot\begin{pmatrix}0.9 & 0.1\\0.1 & 0.9\end{pmatrix}=\begin{pmatrix}0.9\\0.1\end{pmatrix}$$



$$\begin{pmatrix}1\\0\end{pmatrix}\cdot\begin{pmatrix}0.9 & 0.1\\0.1 & 0.9\end{pmatrix}=\begin{pmatrix}0.9\\0.1\end{pmatrix}$$

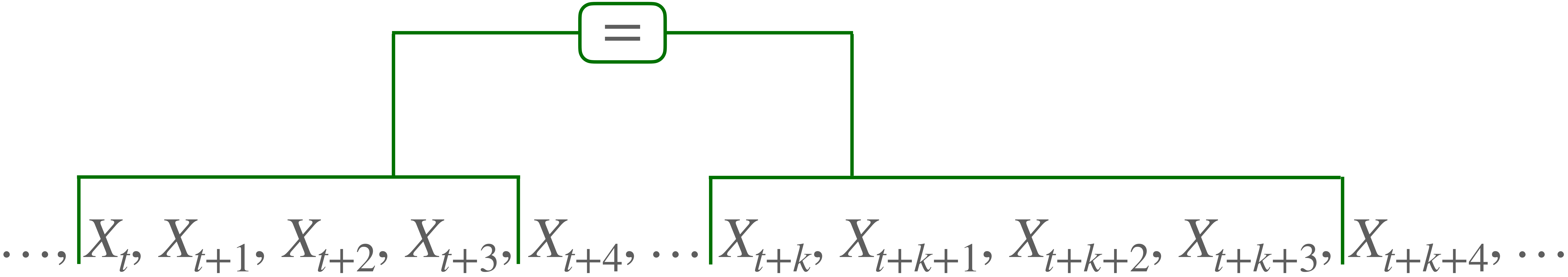
$$\vdots$$

$$\begin{pmatrix}1\\0\end{pmatrix}\cdot\begin{pmatrix}0.9 & 0.1\\0.1 & 0.9\end{pmatrix}^{20}=\begin{pmatrix}0.506\\0.494\end{pmatrix}$$

Stationarity?

What do we gain?

$$\ldots, \overbrace{X_t, X_{t+1}, X_{t+2}, X_{t+3},} X_{t+4}, \ldots \overbrace{X_{t+k}, X_{t+k+1}, X_{t+k+2}, X_{t+k+3},} X_{t+k+4}, \ldots$$



$$\ldots, X_t, X_{t+1}, X_{t+2}, X_{t+3}, X_{t+4}, \ldots X_{t+k}, X_{t+k+1}, X_{t+k+2}, X_{t+k+3}, X_{t+k+4}, \ldots$$

$$\mathbb{E}(f(\overrightarrow{X}_{t:t+n}))$$

$$\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}(f(\overrightarrow{X}_{t:t+n}))$$

$$\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}(f(\overrightarrow{X}_{t:t+n}))$$

$$\mathbb{E}(f(\overrightarrow{X}_{1:n+1}))$$

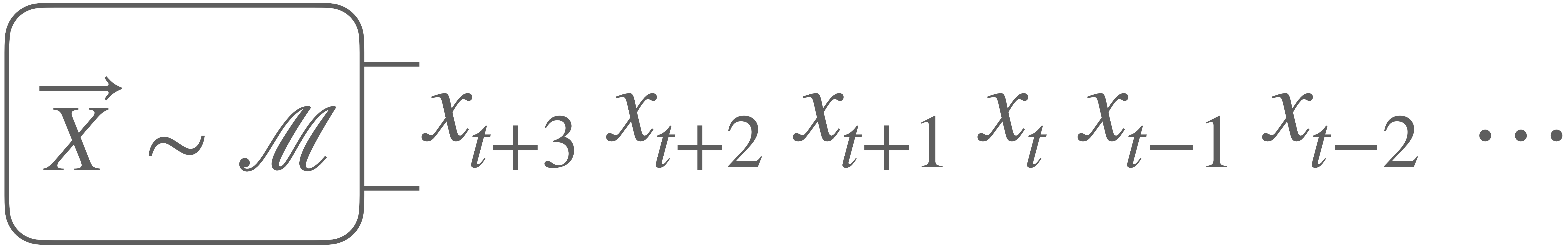
$$\left.\begin{array}{r}\mathbb{E}(f(\overrightarrow{X}_{t:t+n})) \\[2em] \frac{1}{t}\sum_{k=1}^{t}\mathbb{E}(f(\overrightarrow{X}_{t:t+n})) \\[2em] \lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^{t}\mathbb{E}(f(\overrightarrow{X}_{t:t+n}))\end{array}\right| \longrightarrow \; f(\mathscr{M})$$

Mixing Time?

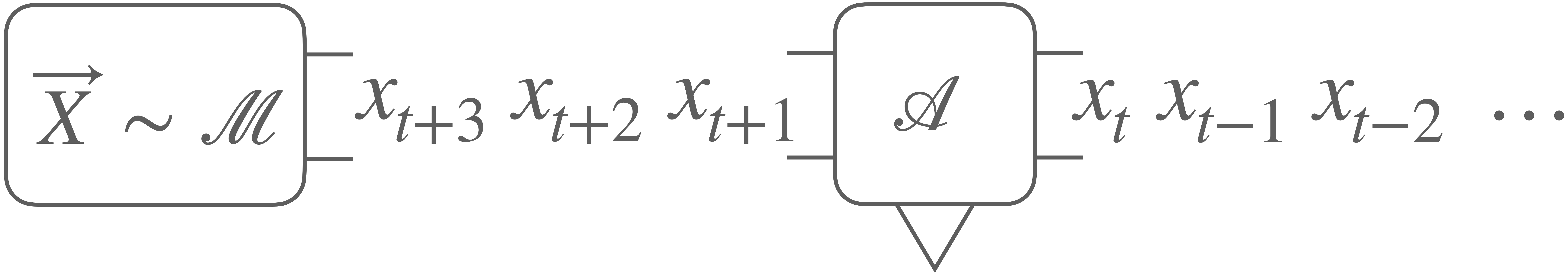
What do we gain?
To be continued.

Problem.

What do we trying to do?

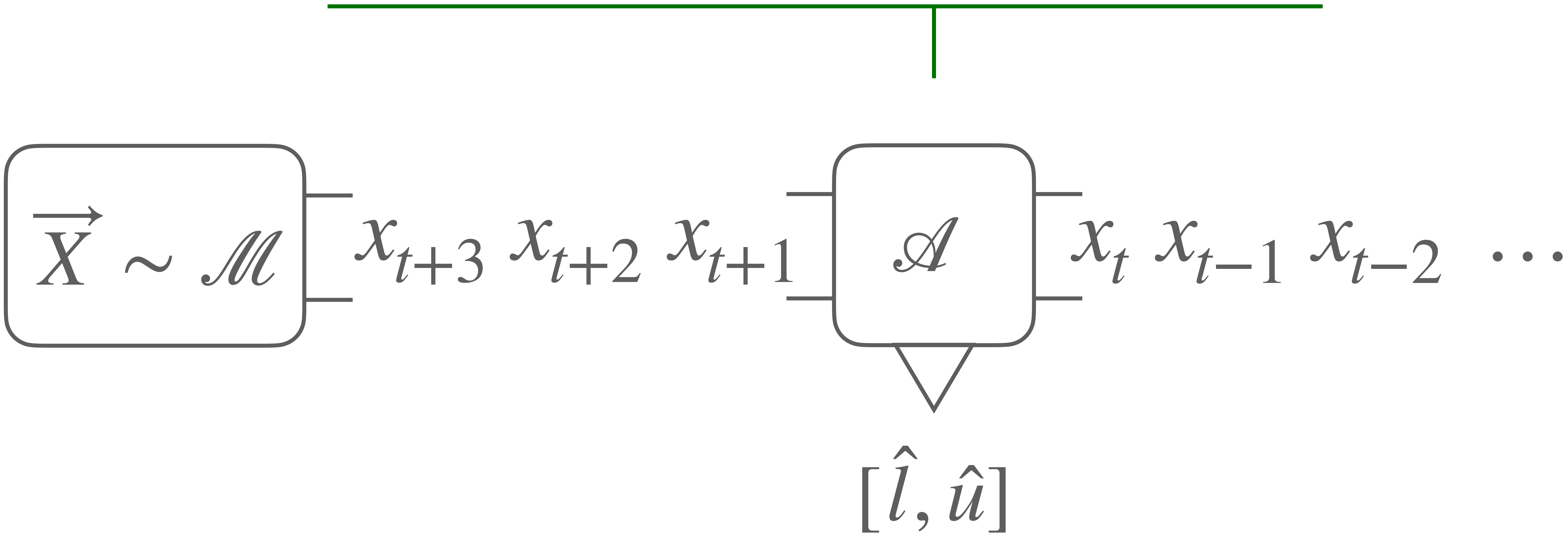


$$\boxed{\overrightarrow{X} \sim \mathscr{M}} \quad x_{t+3}\ x_{t+2}\ x_{t+1}\ x_t\ x_{t-1}\ x_{t-2}\ \cdots$$



$\overrightarrow{X} \sim \mathscr{M}$ x_{t+3} x_{t+2} x_{t+1} $\mathscr{A}$ x_t x_{t-1} x_{t-2} $\cdots$

$f(\mathscr{M}) \in \mathscr{A}(\overrightarrow{x_t})$ with probability $1 - \delta$


$\overrightarrow{X} \sim \mathscr{M}$
$x_{t+3}\ x_{t+2}\ x_{t+1}$
$\mathscr{A}$
$x_t\ x_{t-1}\ x_{t-2}\ \cdots$
$[\hat{l}, \hat{u}]$

$$\mathbb{P}\left(f(\mathscr{M}) \in \mathscr{A}(\overrightarrow{X}_t)\right) \geq 1 - \delta$$

Algorithm.

A sketch.

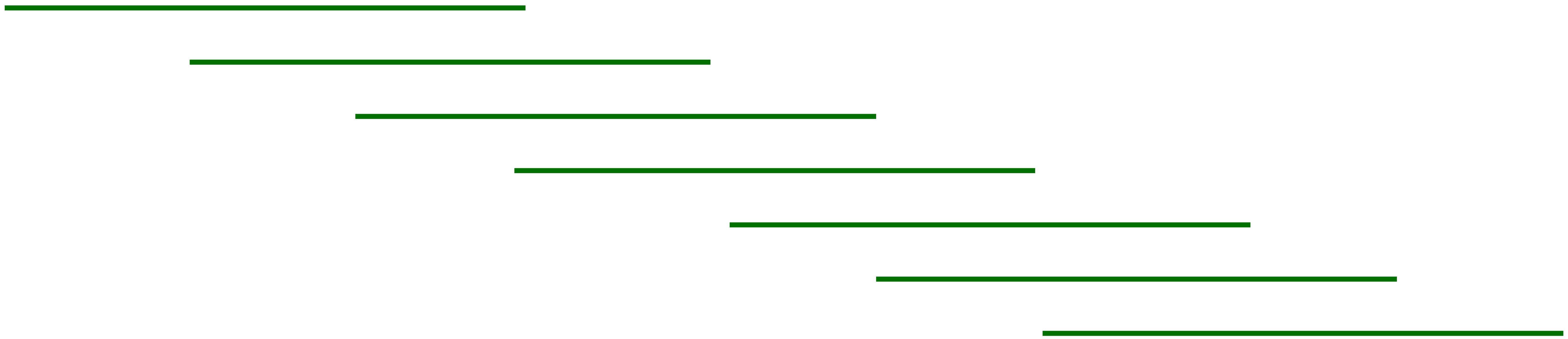
$$\mathbb{E}\left(f(\overrightarrow{X}_{t:t+n})\right) = \mathbb{E}\left(f(\overrightarrow{X}_{t+k:t+k+n})\right)$$

From stationarity

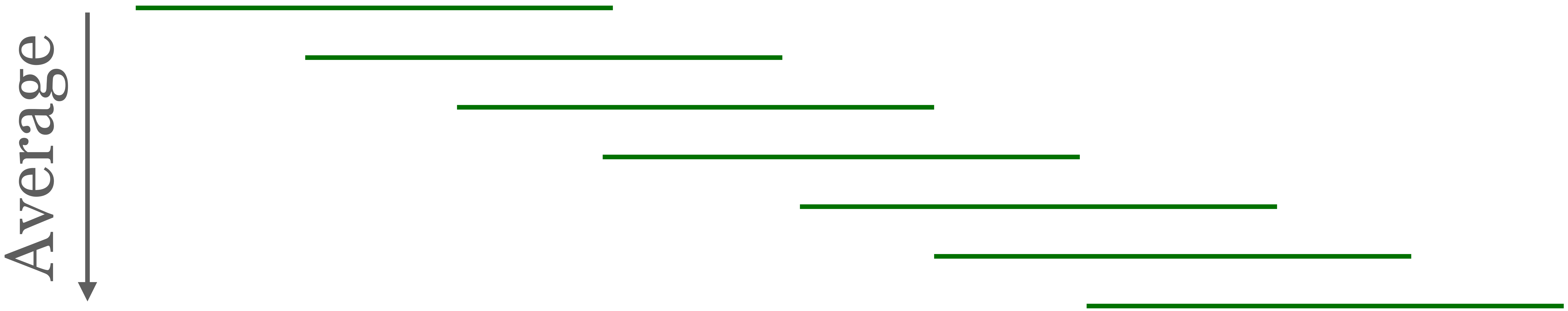
$$\underset{f(x_1, x_2, x_3).}{\underline{x_1, x_2, x_3,}} \; x_4, x_5, x_6, x_7, x_8, x_9$$

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$$

$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$$



$$x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8, x_9$$



$$\hat{f}(\vec{x}_t) := \frac{1}{t-n+1}\sum_{i=1}^{t-n+1} f(\vec{x}_{i:i+n+1})$$

Estimator

$$\mathbb{E}(\hat{f}(\overrightarrow{X}_t)) = f(\mathscr{M})$$

Unbiased

$$\overbrace{|\hat{f}(\vec{x}_t) - \hat{f}(\vec{x}'_t)|}^{\vec{x}_t \text{ and } \vec{x}'_t \text{ differ only in position } i} \leq c_i(t)$$

Lipschitz continuous

$$\mathbb{P}\left(\,|f(\mathscr{M}) - \hat{f}(\overrightarrow{X}_t)| \geq \varepsilon\,\right) \leq \gamma(\varepsilon, \tau_{mix}, \{c_i(t)\}_i)$$

McDiarmid's inequality for MCs

$$\mathbb{P}\left(\, |f(\mathscr{M}) - f(\overrightarrow{X}_t)| \geq \varepsilon \right) \leq 2 \cdot \exp\left(-\frac{2\varepsilon^2}{\sqrt{\Sigma_{i=1}^{n} c_i^2}^{\,2} \cdot 9 \cdot \tau_{mix}} \right)$$

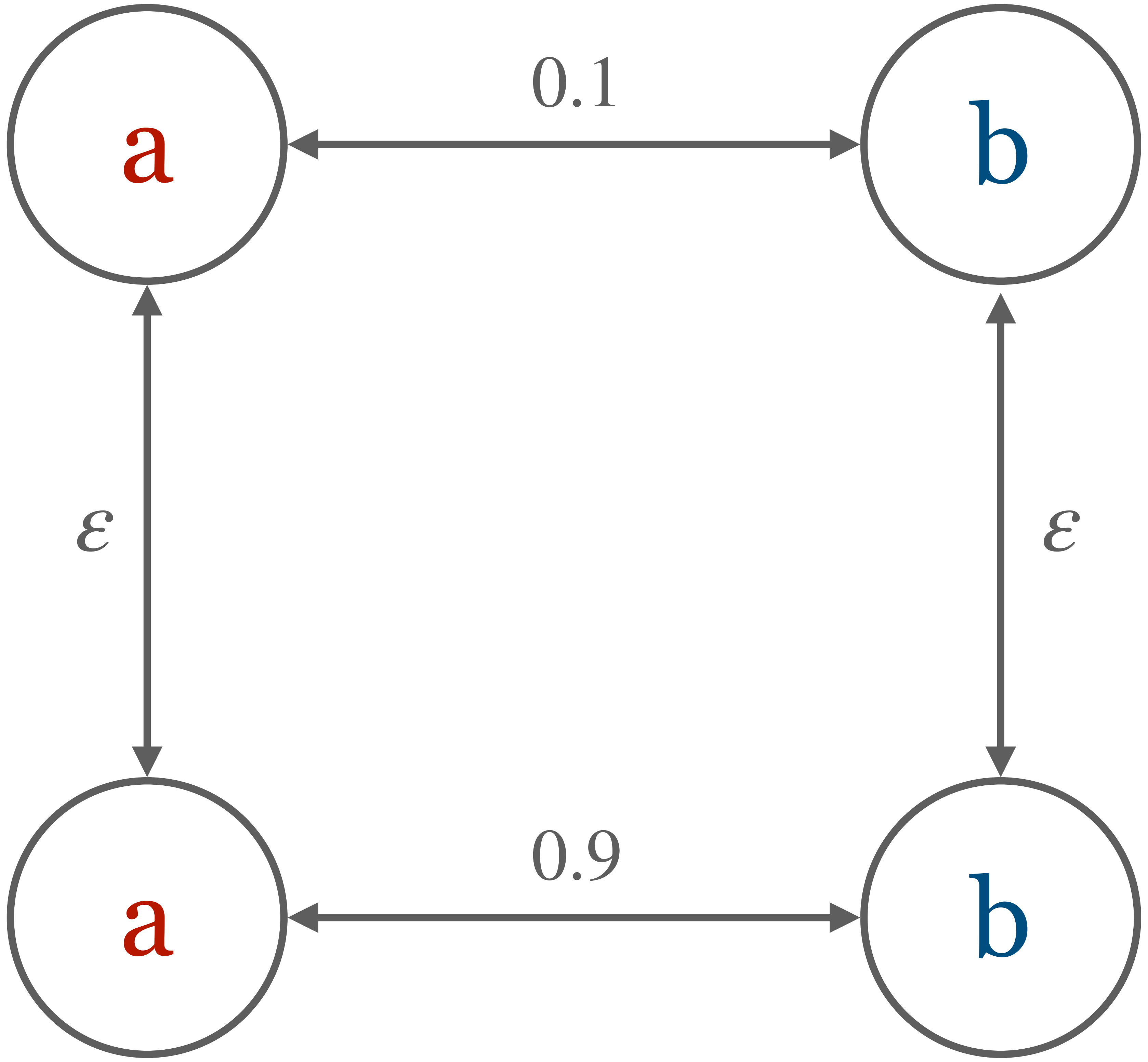
McDiarmid's inequality for MCs

Result can easily be extended to…

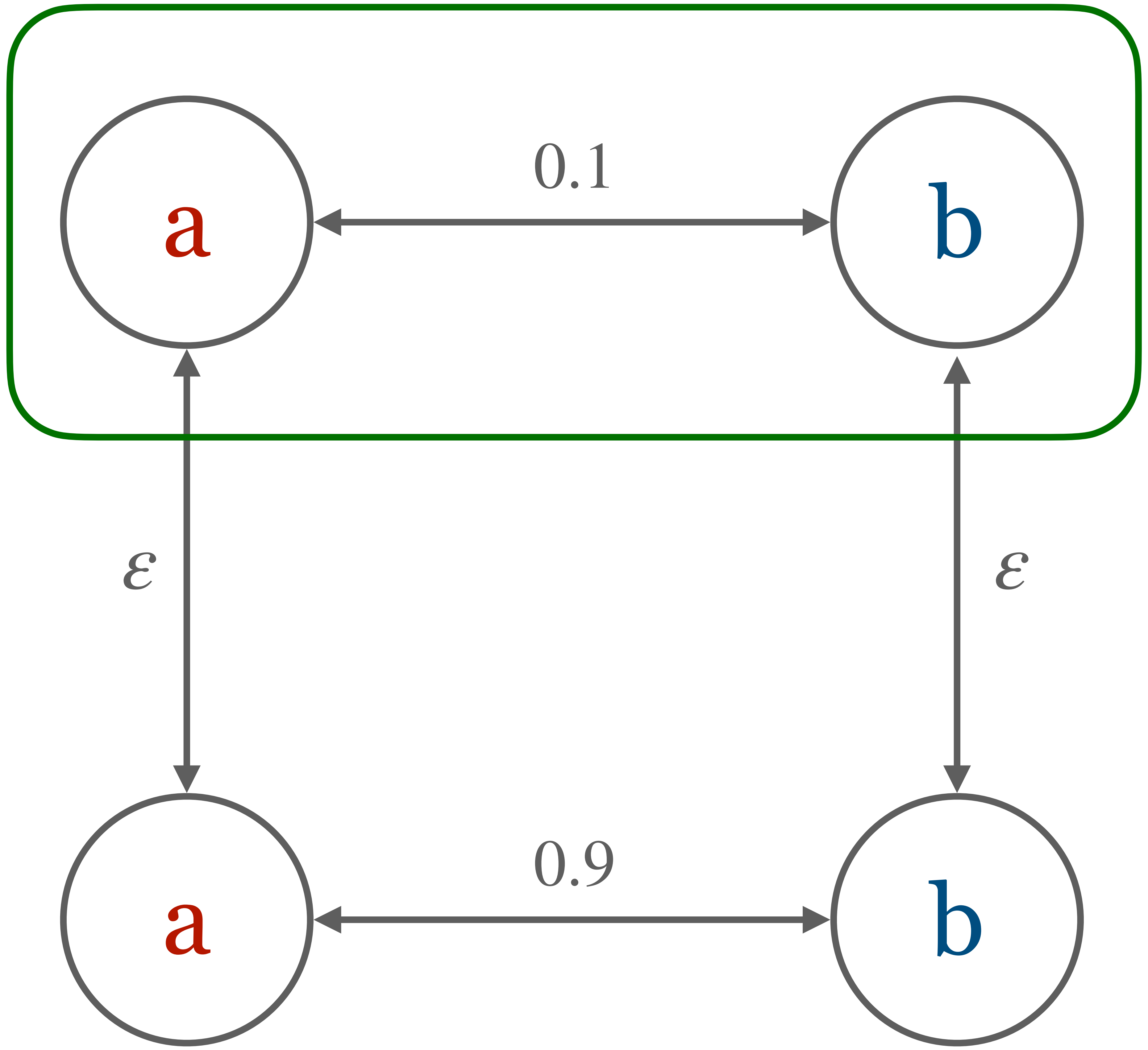
… arithmetic expressions over
expected values of atomic functions
using union bound and interval arithmetic.

Mixing Time?

Because of Dependency.

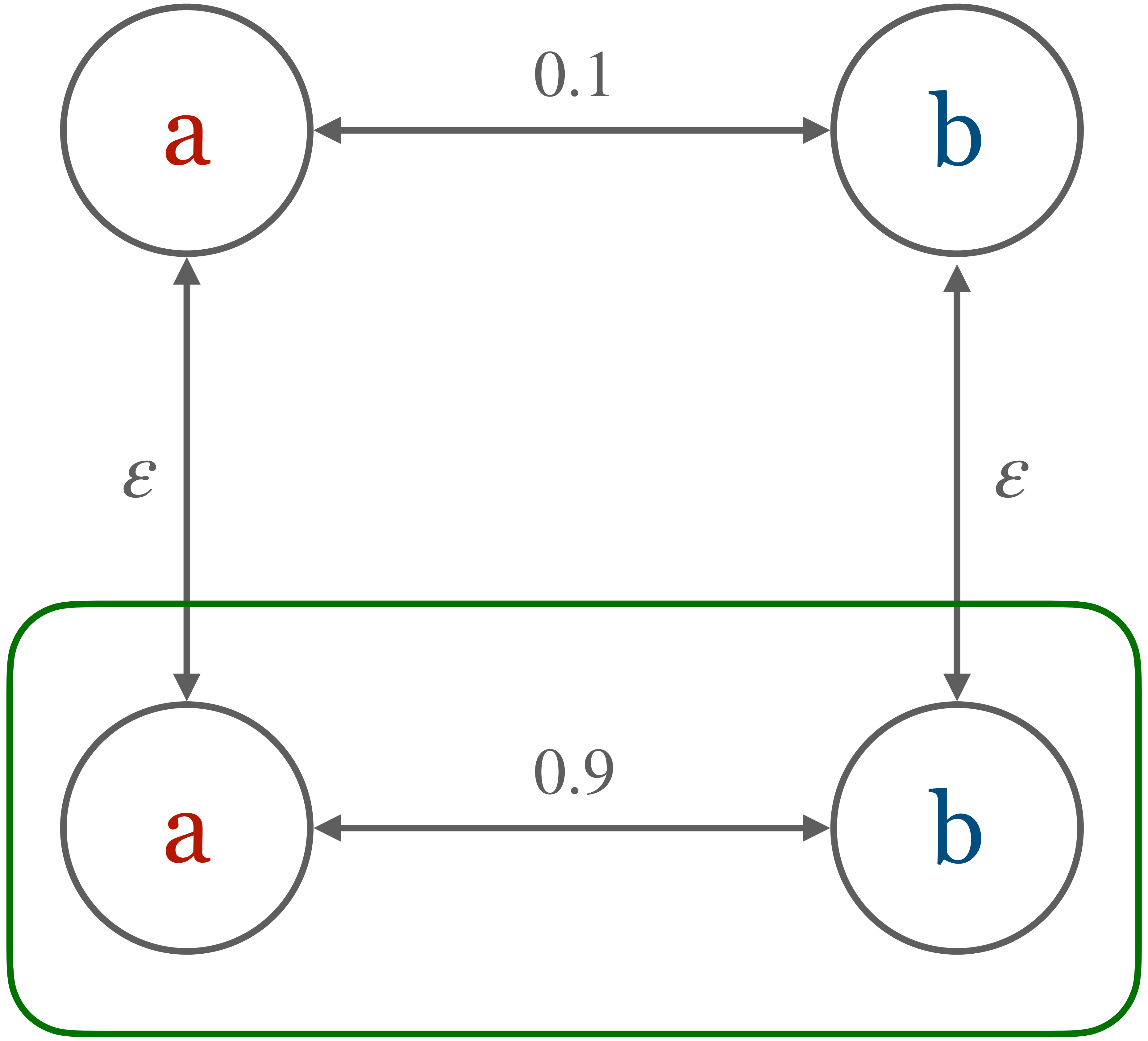


$$\mathbb{P}(b \mid a) = 0.5$$



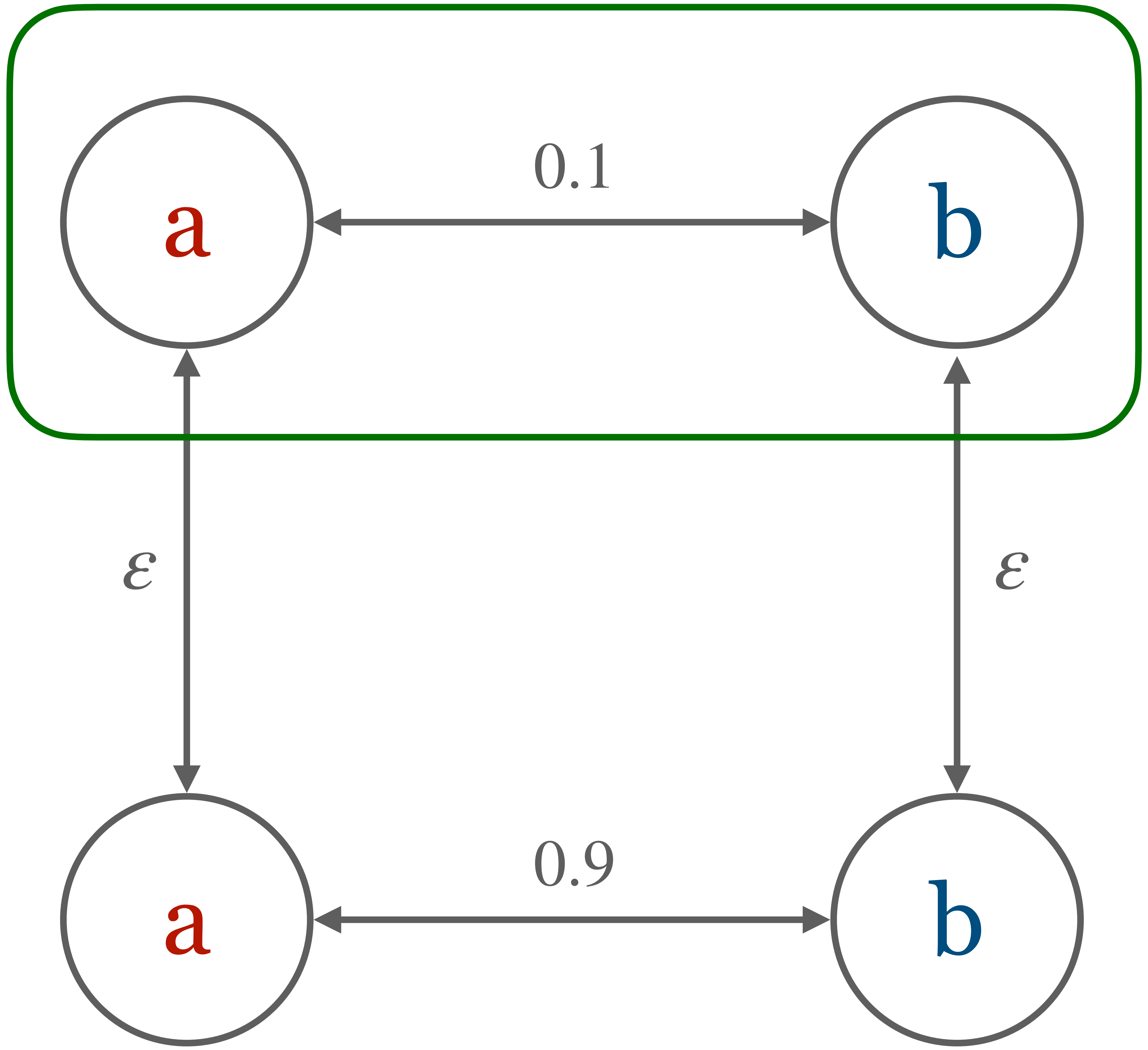
$$\mathbb{P}(b \mid a) = 0.5$$

$$\hat{f}(\vec{x}_t) \approx 0.1$$



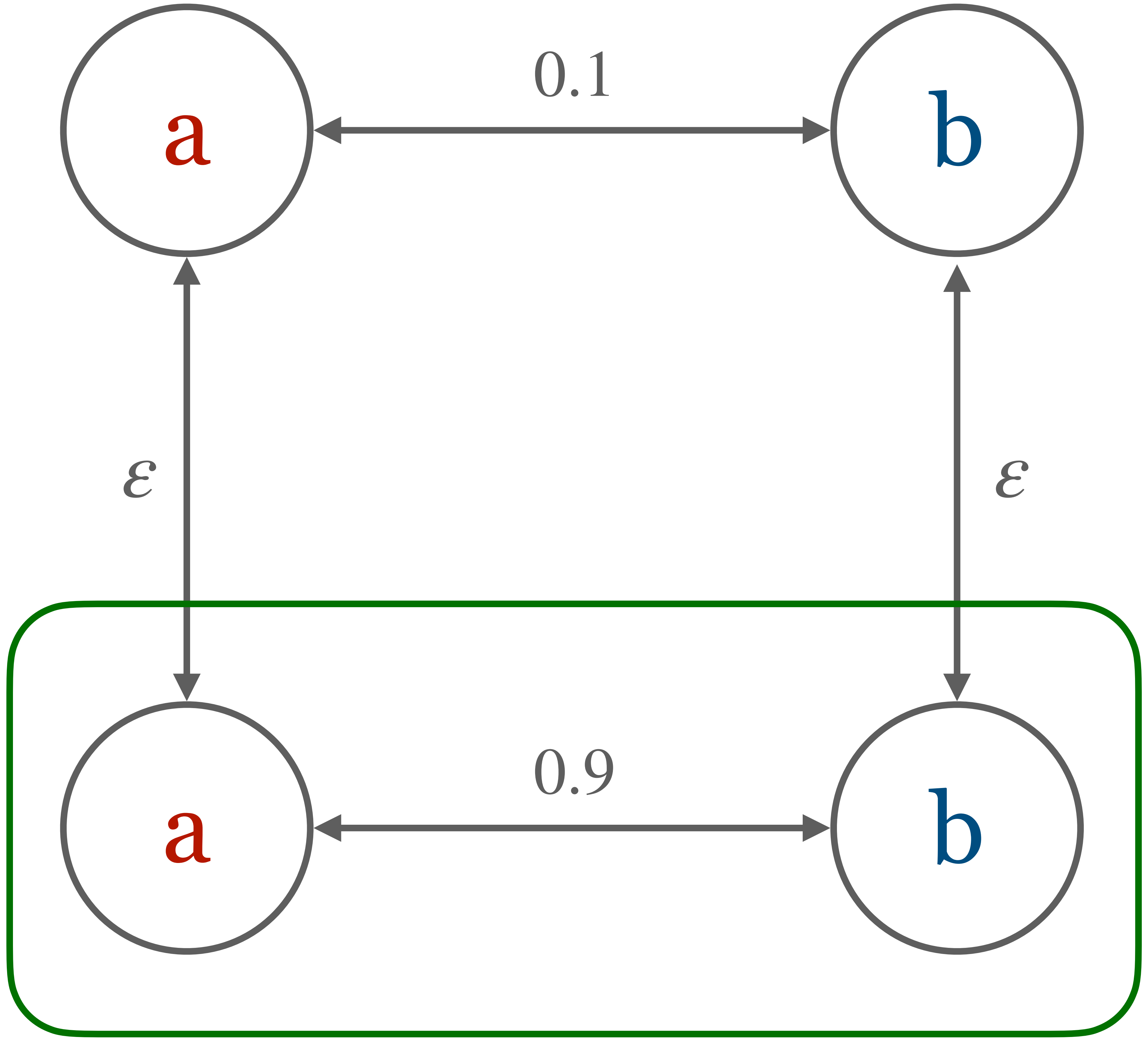
$$\mathbb{P}(b \mid a) = 0.5$$

$$\hat{f}(\vec{x}_t) \approx 0.1$$



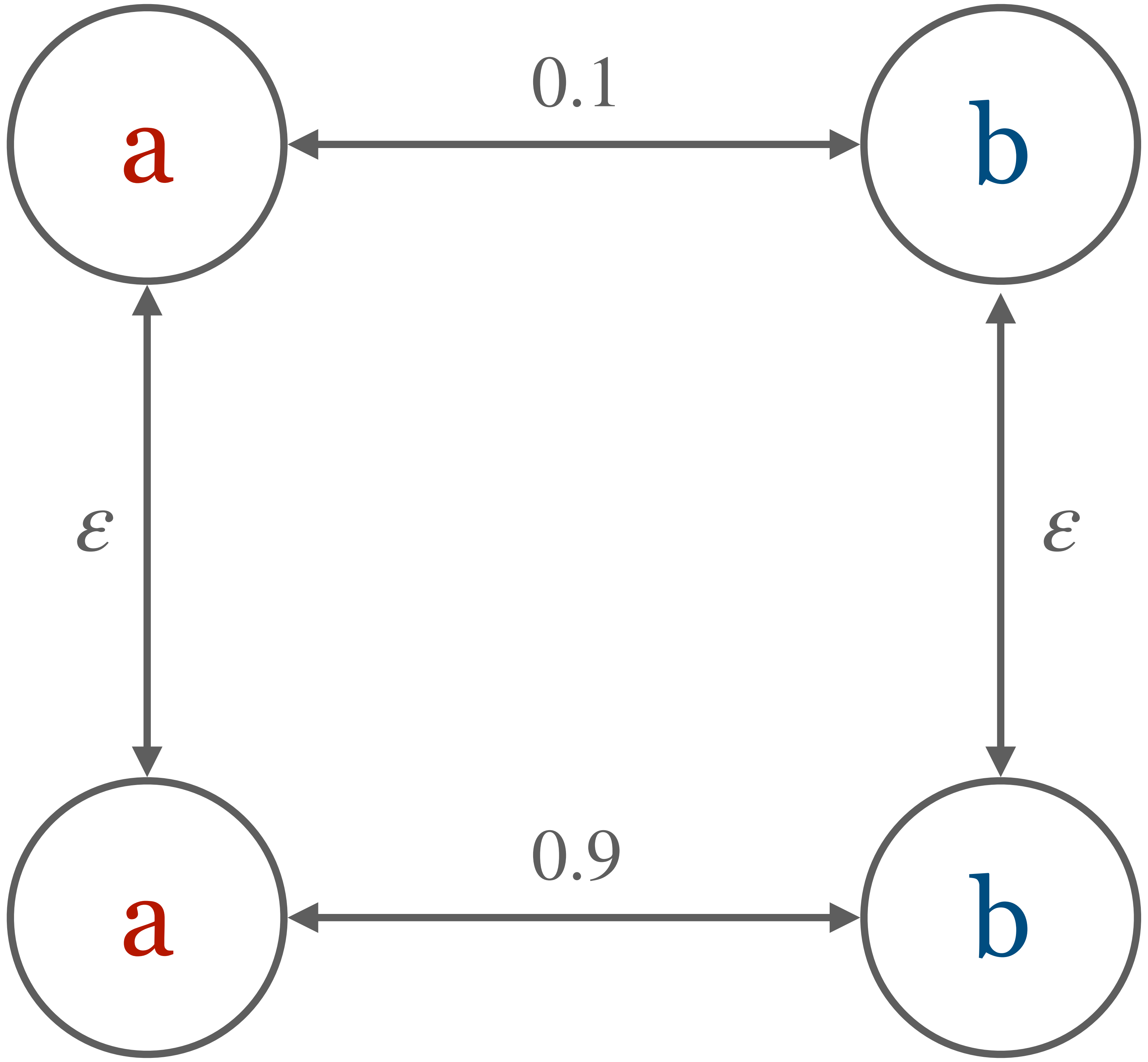
$$\mathbb{P}(b \mid a) = 0.5$$

$$\hat{f}(\vec{x}_t) \approx 0.1$$



$$\mathbb{P}(b \mid a) = 0.5$$

$$\hat{f}(\vec{x}_t) \approx 0.1$$



$$\mathbb{P}(b \mid a) = 0.5$$

$$\hat{f}(\vec{x}_t) \rightarrow 0.5$$

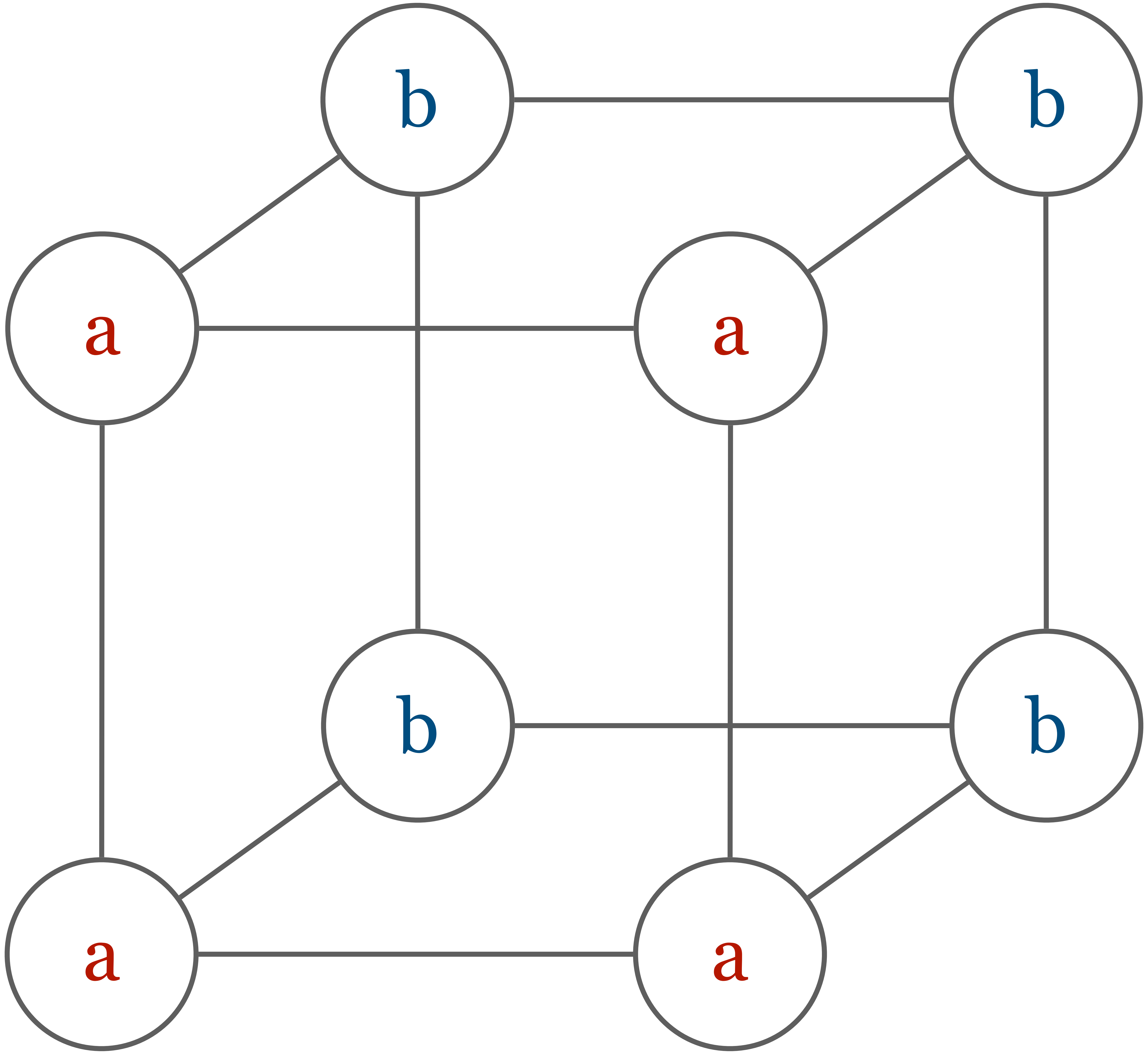
You don’t know…

…when you have seen all behaviour you need to see.

Experiments.

3D-Hypercube (i.e. a cube).

D’Amour et al. 2020. Fairness is not static: deeper understanding of long term fairness via simulation studies


b
b
a
a
b
b
a
a

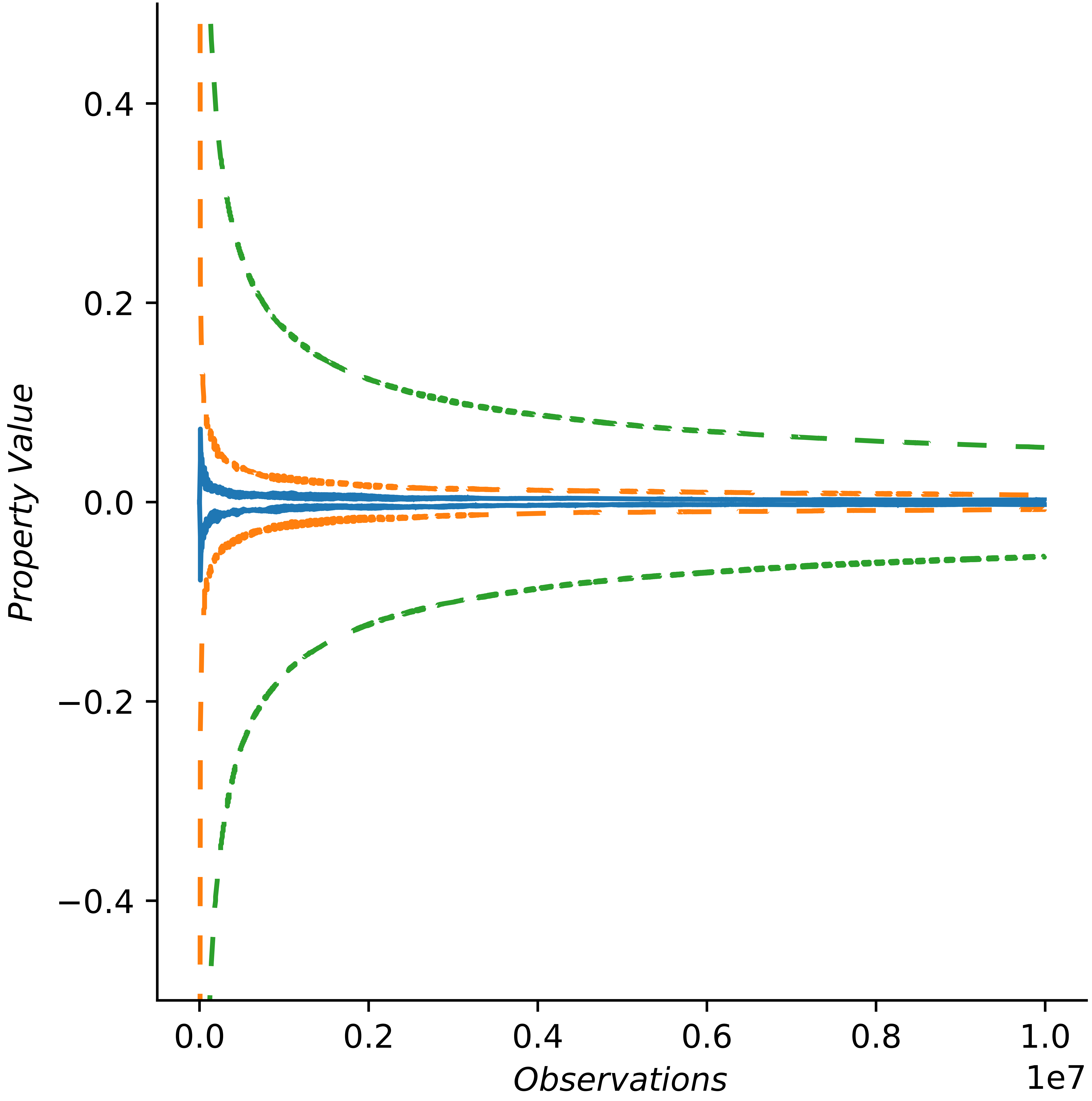
$$\mathbb{P}(aa) - \mathbb{P}(bb)$$

Property.

$$\mathbb{1}[x = a, y = a]$$

$$f_{aa}(x, y) - f_{bb}(x, y)$$

Function.


Property Value
0.4
0.2
0.0
−0.2
−0.4
0.0
0.2
0.4
0.6
0.8
1.0
1e7
Observations

Related Work.

What has been done so far?

Static verification of algorithmic fairness

Albarghouthi, et al. "Fairsquare: probabilistic verification of program fairness.” OOPSLA 2017.
Bastani et al. "Probabilistic verification of fairness properties via concentration." OOPSLA 2019.
Ghosh et al. “Justicia: A stochastic sat approach to formally verify fairness.” AAAI 2021.
Sun, et al. "Probabilistic verification of neural networks against group fairness." FM 2021.
Ghosh, et. al. "Algorithmic fairness verification with graphical models." AAAI 2022.

Monitoring algorithmic fairness

Albarghouthi and Vinitsky. "Fairness-aware programming." FAccT 2019.
Henzinger et al. “Monitoring Algorithmic Fairness.” CAV 2023.
Henzinger et al. "Runtime Monitoring of Dynamic Fairness Properties." FAccT 2023.

Summary.

Main points.

Interested in monitoring “distributional” properties, e.g. conditional expectation, of stochastic processes.

Proposed a monitor for estimating such properties over a restricted class of Hidden Markov Models.

Leverage tools from non-asymptotic statistics to provide valid guarantees for each time step.

We focused on monitoring Algorithmic Fairness, but those techniques have wide applicability.

ISTA

Institute of
Science and
Technology
Austria