There's something about MRAI

Timing diversity can exponentially worsen BGP convergence

Alex Fabrikant

Google Research
(Work done while at Princeton University)

Joint work with Umar Syed (Penn) and Jennifer Rexford (Princeton)

BGP makes the world go ’round

- For those just walking in: BGP keeps the Internet glued together. Pretty important, right?

BGP makes the world go ’round

- For those just walking in: BGP keeps the Internet glued together. Pretty important, right?
- Lots of progress on understanding the formal "question that BGP computes an answer to"

BGP makes the world go ’round

- For those just walking in: BGP keeps the Internet glued together. Pretty important, right?
- Lots of progress on understanding the formal "question that BGP computes an answer to"
- What of the process of computation?
  - A number of measurement studies of convergence
  - We set out for a detailed theory of worst-case rate of convergence

Why worst-case?

- The internet is scaling and evolving
- Makes for solid best-practices recommendations

Why worst-case?

- The internet is scaling and evolving
- Makes for solid best-practices recommendations
  - Network ops know:
    BGP trouble + pager = sleep deprivation

Why worst-case?

- The internet is scaling and evolving
- Makes for solid best-practices recommendations
  - Network ops know:
    BGP trouble + pager = sleep deprivation

    Measurement studies vs worst-case analysis

    =

    sleeping pill vs a cure for insomnia
  - (both are important!)

Why worst-case?

- The internet is scaling and evolving
- Makes for solid best-practices recommendations
  - Network ops know:
    BGP trouble + pager = sleep deprivation

    Measurement studies vs worst-case analysis

    =

    sleeping pill vs a cure for insomnia
  - (both are important!)
- We explore limiting our models to get more realistic bounds

BGP model - a quick sketch

- Atomic autonomous systems, a graph of edges
- AS's route preferences: at *least* tractable (more later)
- We focus on single-destination
- SPVP-based model of dynamics [GSW'02]

BGP convergence

- [Obradovic'02] and [Sami,Schapira,Zohar'09] say linear convergence, if no dispute wheels
  - Linear in the depth of the customer-provider hierarchy [O'02;SSZ'09]
  - Linear in #ASes even without G-R constraints, given no dispute wheels [SSZ'09]
  - (What's a dispute wheel? [Gao,Rexford'01]: don't worry about it, this condition holds* assuming the Internet is based on economically-sensible interactions)

BGP convergence

- [Obradovic'02] and [Sami,Schapira,Zohar'09] say linear convergence, if no dispute wheels
- [Karloff'04]: exponentially slow convergence!
  ([Labovitz'01] got harsher results, but in an odd model)

BGP convergence

- [Obradovic'02] and [Sami,Schapira,Zohar'09] say linear convergence, if no dispute wheels
- [Karloff'04]: exponentially slow convergence!
  ([Labovitz'01] got harsher results, but in an odd model)
- What happened?

BGP convergence

- [Obradovic'02] and [Sami,Schapira,Zohar'09] say linear convergence, if no dispute wheels
- [Karloff'04]: exponentially slow convergence!
  ([Labovitz'01] got harsher results, but in an odd model)
- What happened?
- The key is units of time:
  - fair phases [O'02;SSZ'09]
  - events [K'04]

Real BGP timing: MRAI

- Min Route Advertisement Interval (MRAI): how frequently should I send updates to my neighbor?
  - MRAI = the Internet's "clock"
- Originally: 30 seconds (in the 1995 RFC)
- Recently:
  - "Can you hear me now?!: it must be BGP" [Kushman, et al'07]
  - Vendors and ISPs are dropping MRAI timers
  - An Internet Draft [Jakma '08–'10] calls for removing the recommended value

Real BGP timing: MRAI

- Min Route Advertisement Interval (MRAI): how frequently should I send updates to my neighbor?
  - MRAI = the Internet's "clock"
- Originally: 30 seconds (in the 1995 RFC)
- Recently:
  - "Can you hear me now?!: it must be BGP" [Kushman, et al'07]
  - Vendors and ISPs are dropping MRAI timers
  - An Internet Draft [Jakma '08–'10] calls for removing the recommended value
- Our results: a gallery combinatorial worst-case scenarios where incremental deployment of these changes risks worsening convergence!

Talk outline

1. What to model?
2. Control-plane convergence
3. What is “Realistic”?
4. Data-plane consequences
5. Implications & open problems

Talk outline

What changes with the new MRAI proposals?

- Incremental deployment is a given!
- Deployment measured by:
  - Timing disparity: the ratio between slowest and fastest MRAI in use, r
  - Timing diversity: the number of different distinct values (*species*) of MRAI, s

What changes with the new MRAI proposals?

- Incremental deployment is a given!
- MRAI recommendation changed?
- Deployment measured by:
  - Timing disparity: the ratio between slowest and fastest MRAI in use, r grows
  - Timing diversity: the number of different distinct values (*species*) of MRAI, $s \rightarrow s + 1$

What changes with the new MRAI proposals?

- Incremental deployment is a given!
- MRAI recommendation removed?
- Deployment measured by:
  - Timing disparity: the ratio between slowest and fastest MRAI in use, r grows
  - Timing diversity: the number of different distinct values (*species*) of MRAI, $s \rightarrow ???$

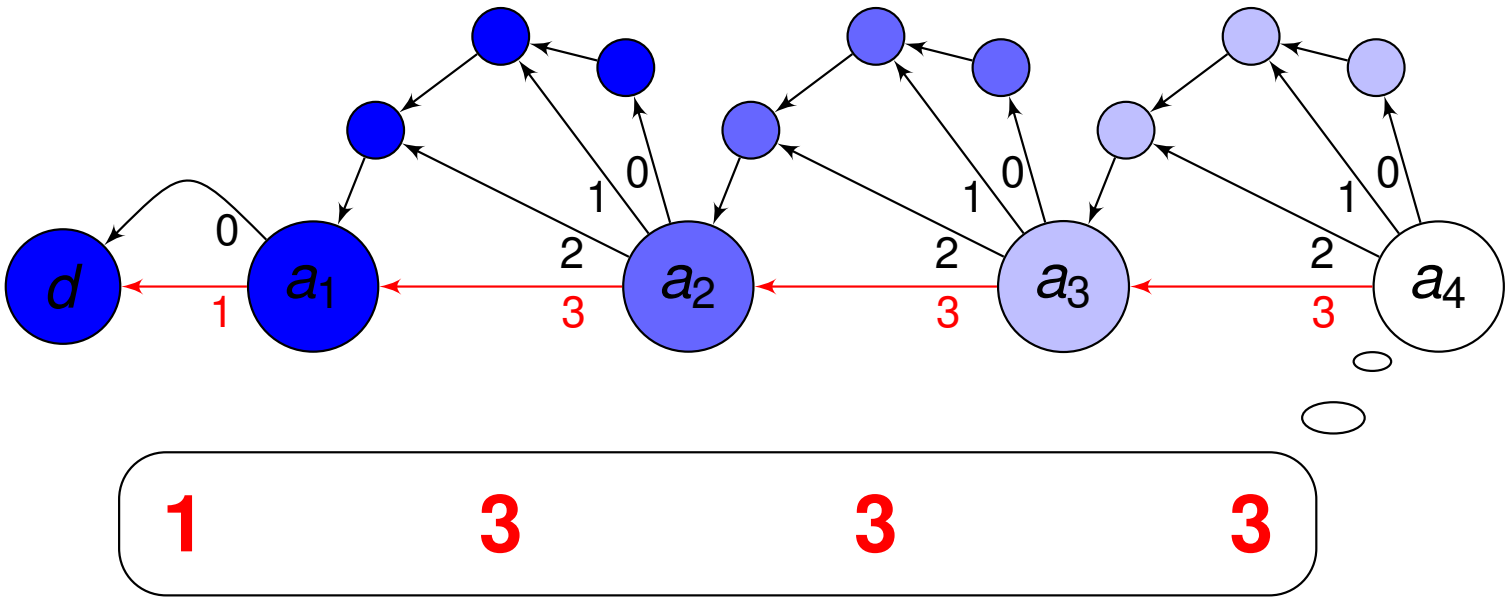
What makes for "good convergence"?

- Of course, time until convergence (in seconds)
- But also:
  1. # max BGP messages sent per link
  2. # BGP messages sent system-wide
  3. # max FIB updates per node
- We consider the dependence on both:
  1. the number of ASes (n)
  2. the customer-provider “depth” of the Internet (α)

Talk outline

MRAI Diversity: messages sent

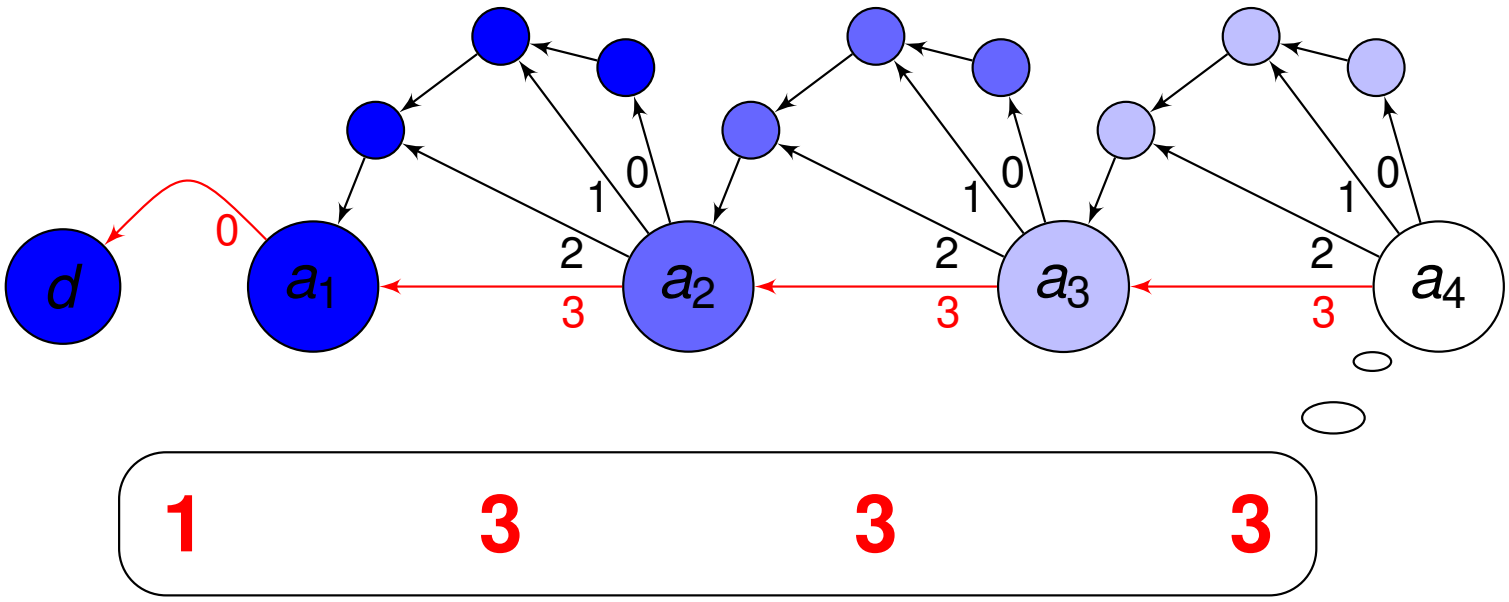
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

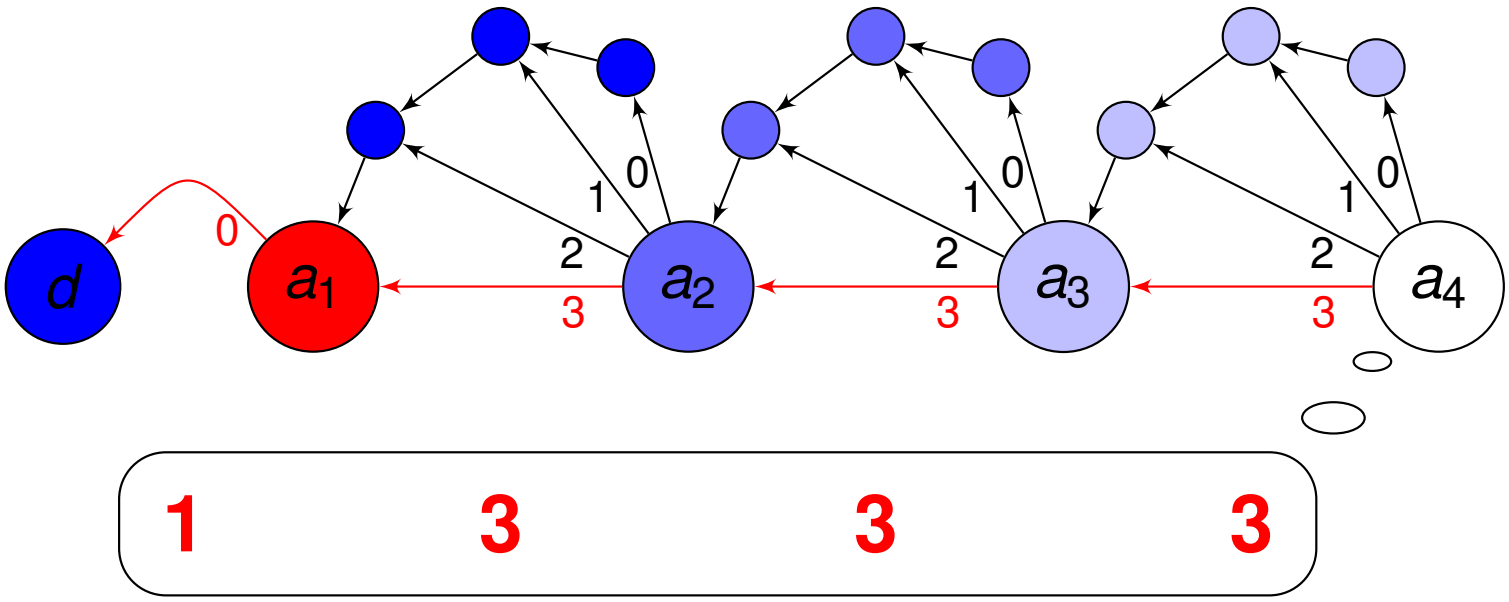
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

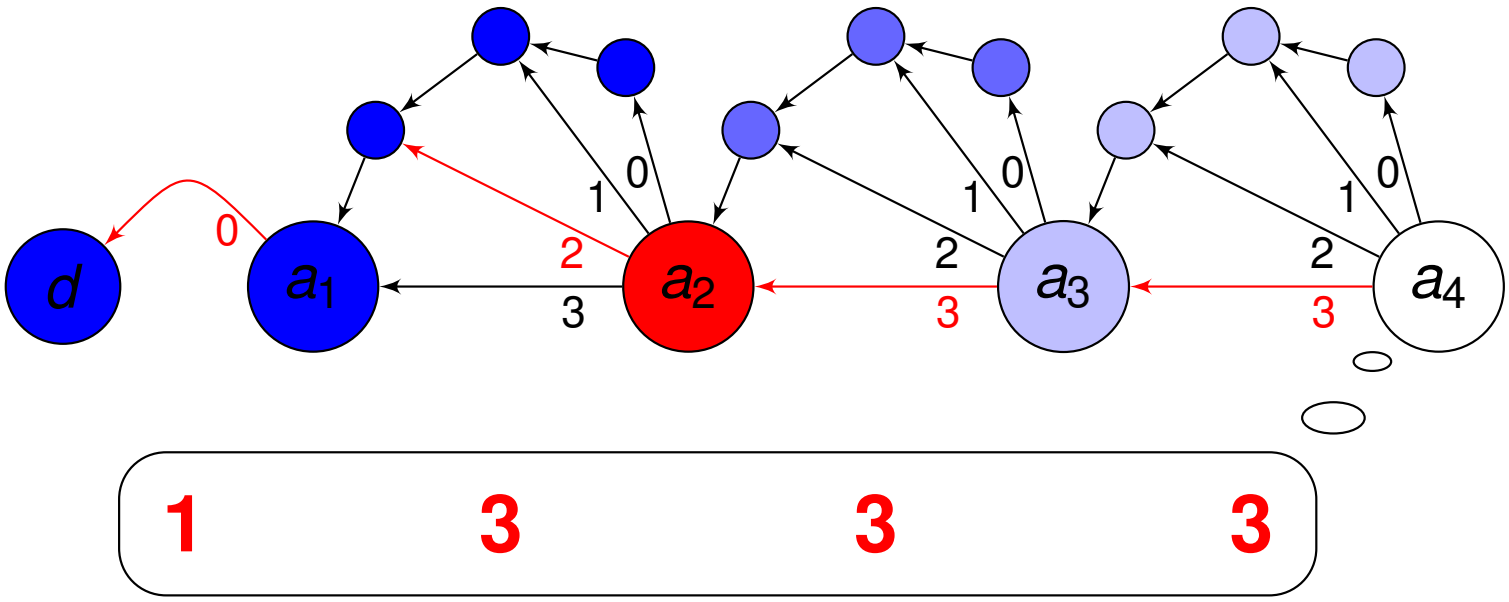
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

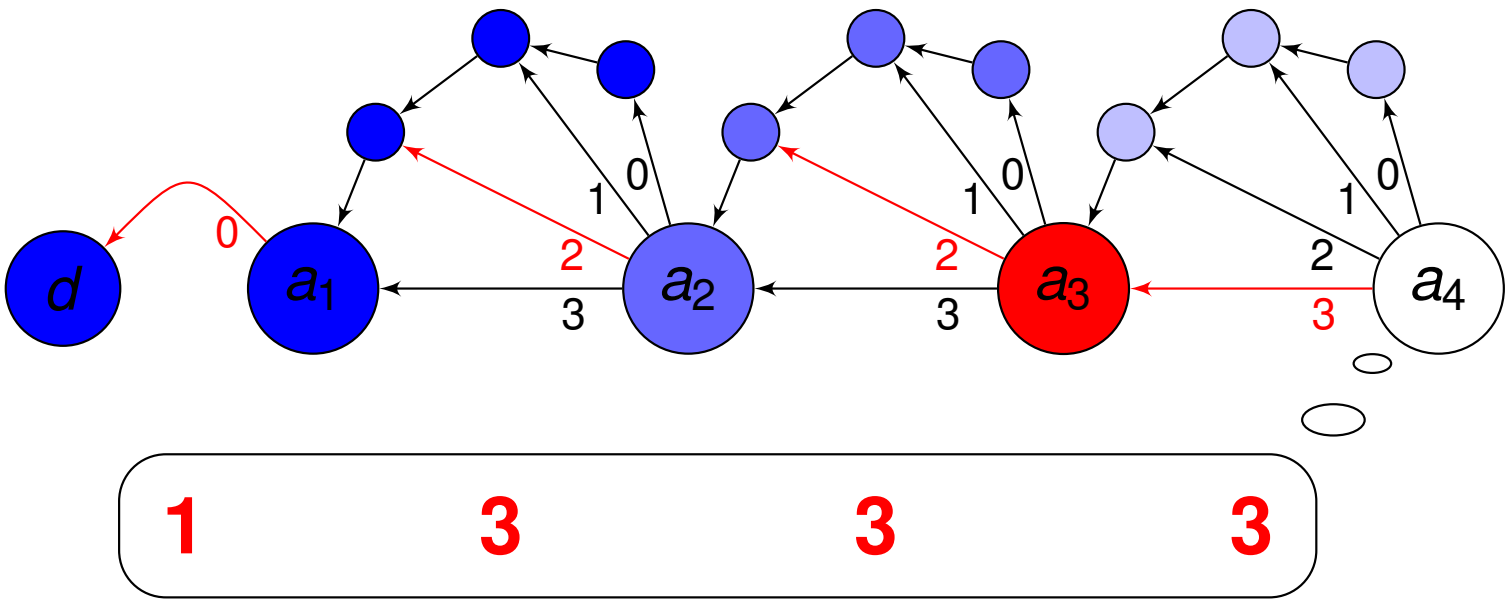
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

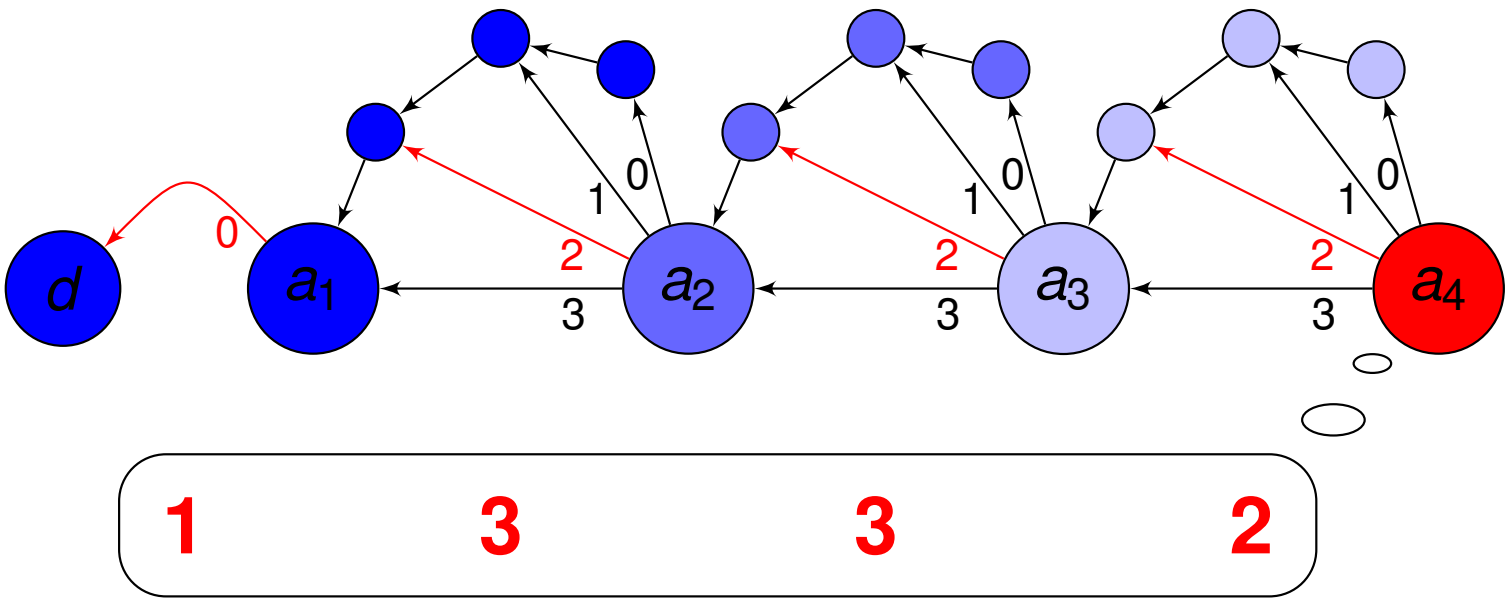
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

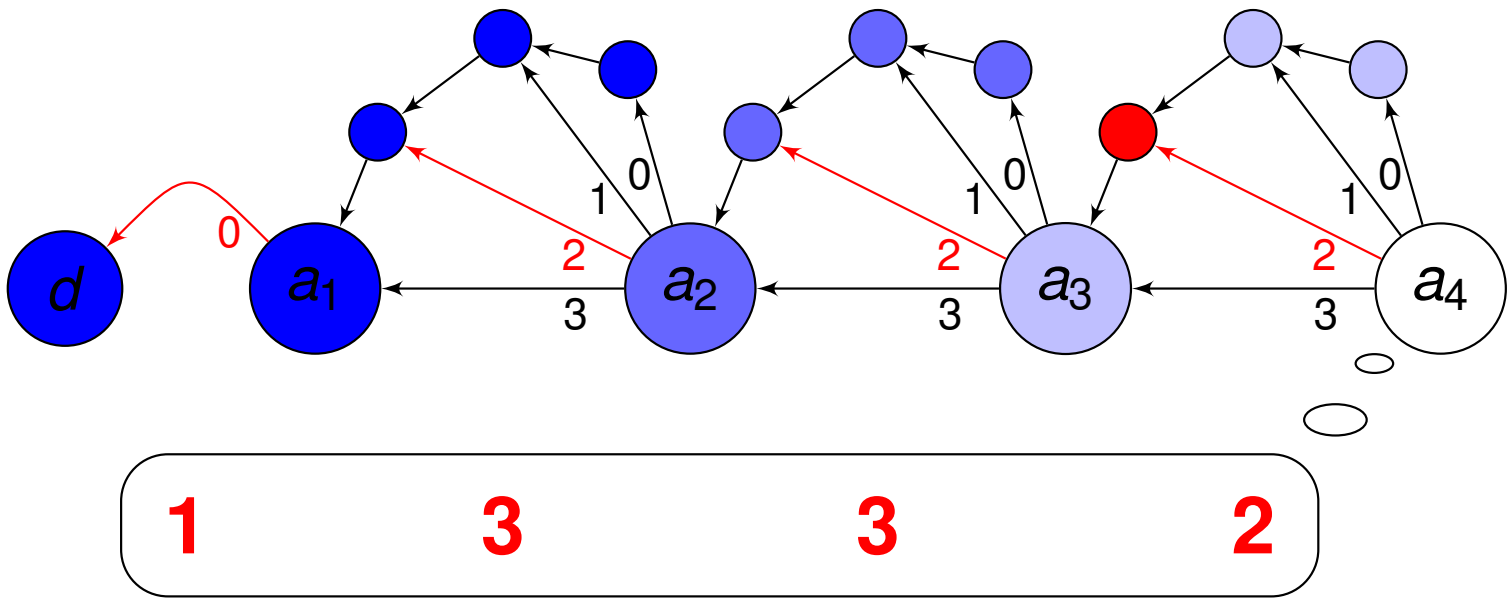
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

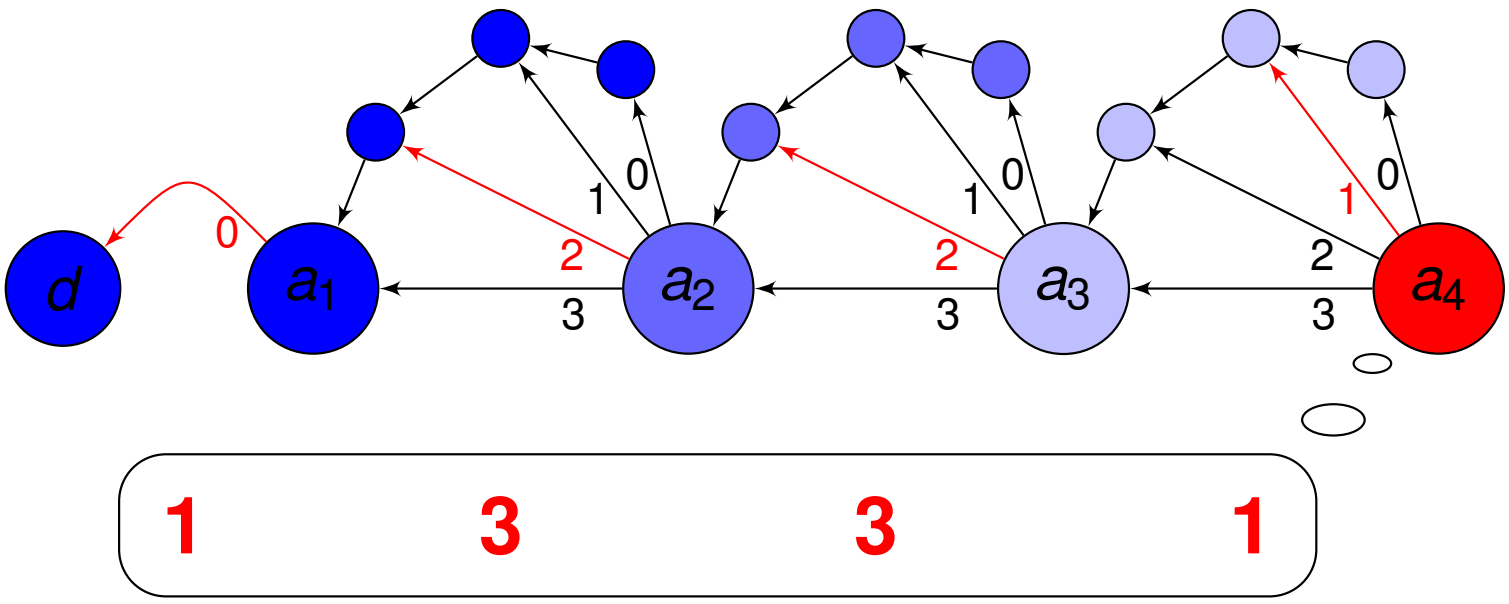
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

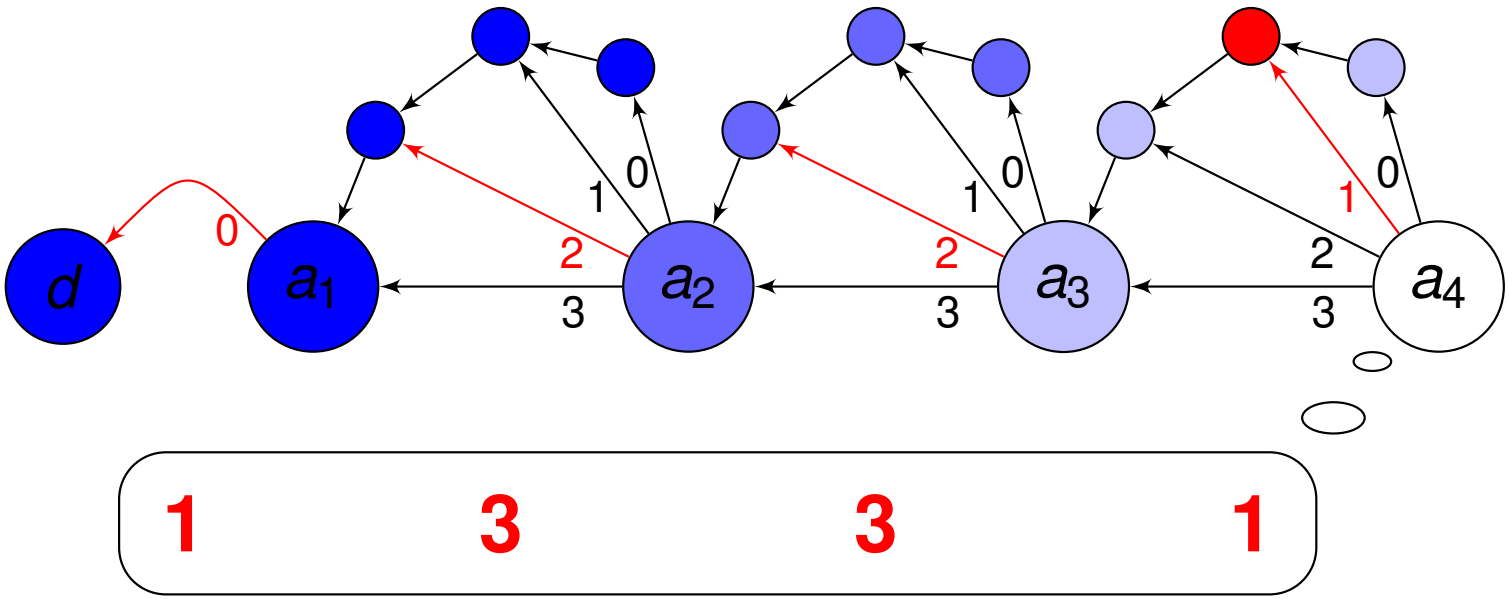
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

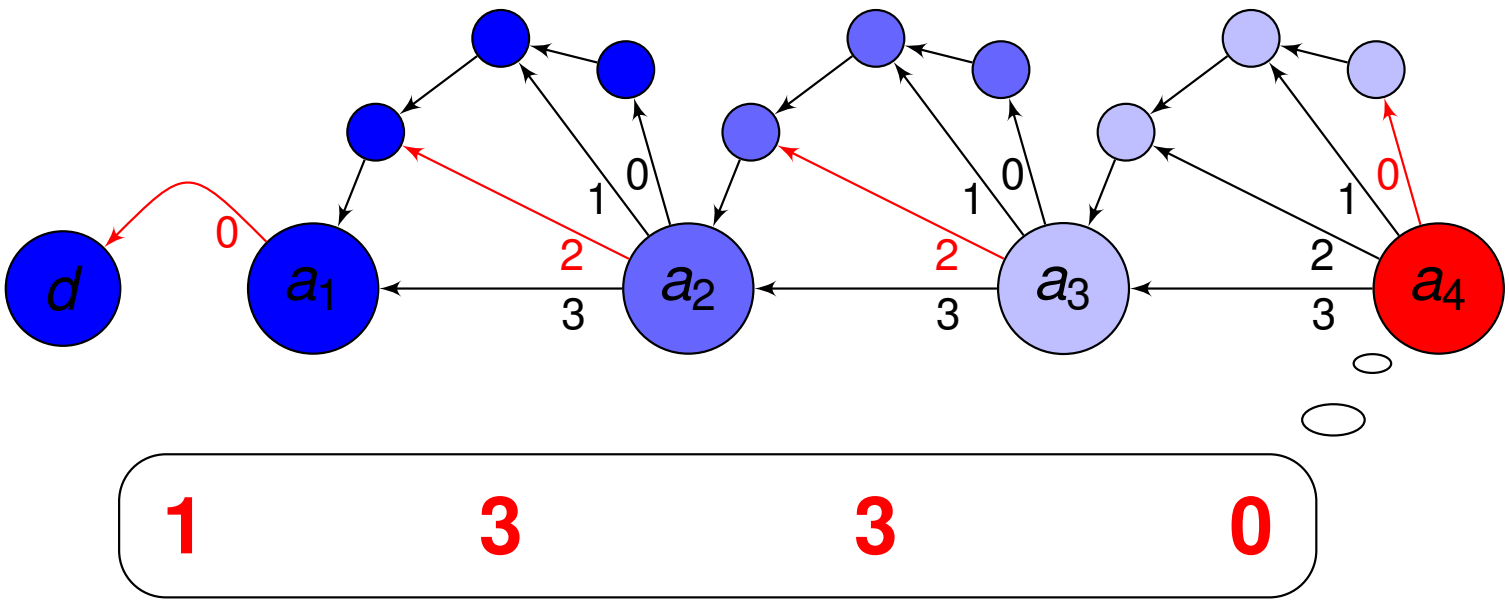
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

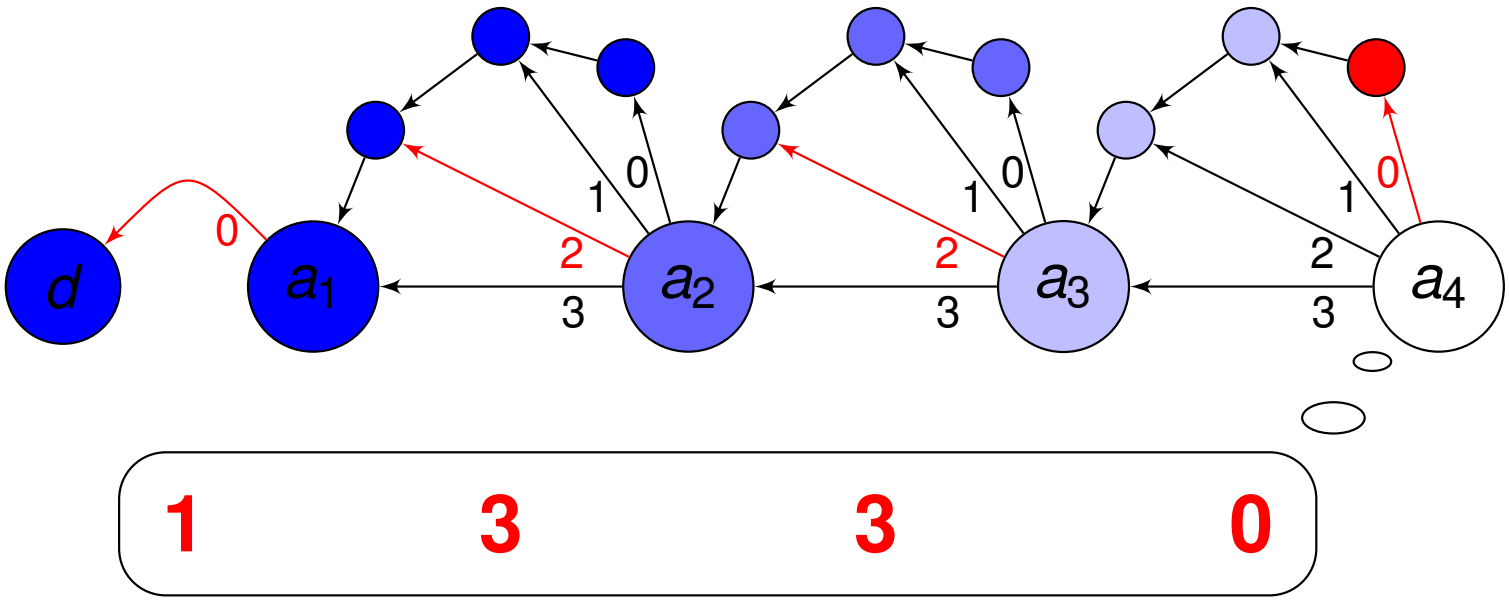
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

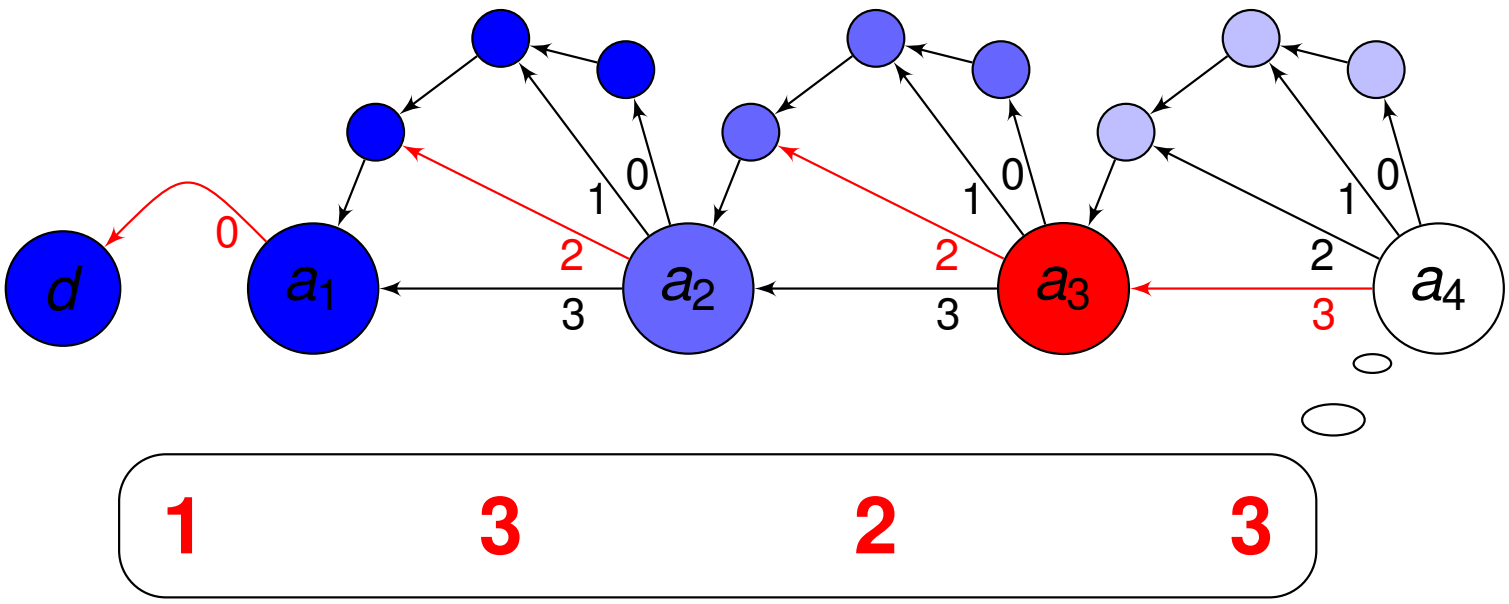
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

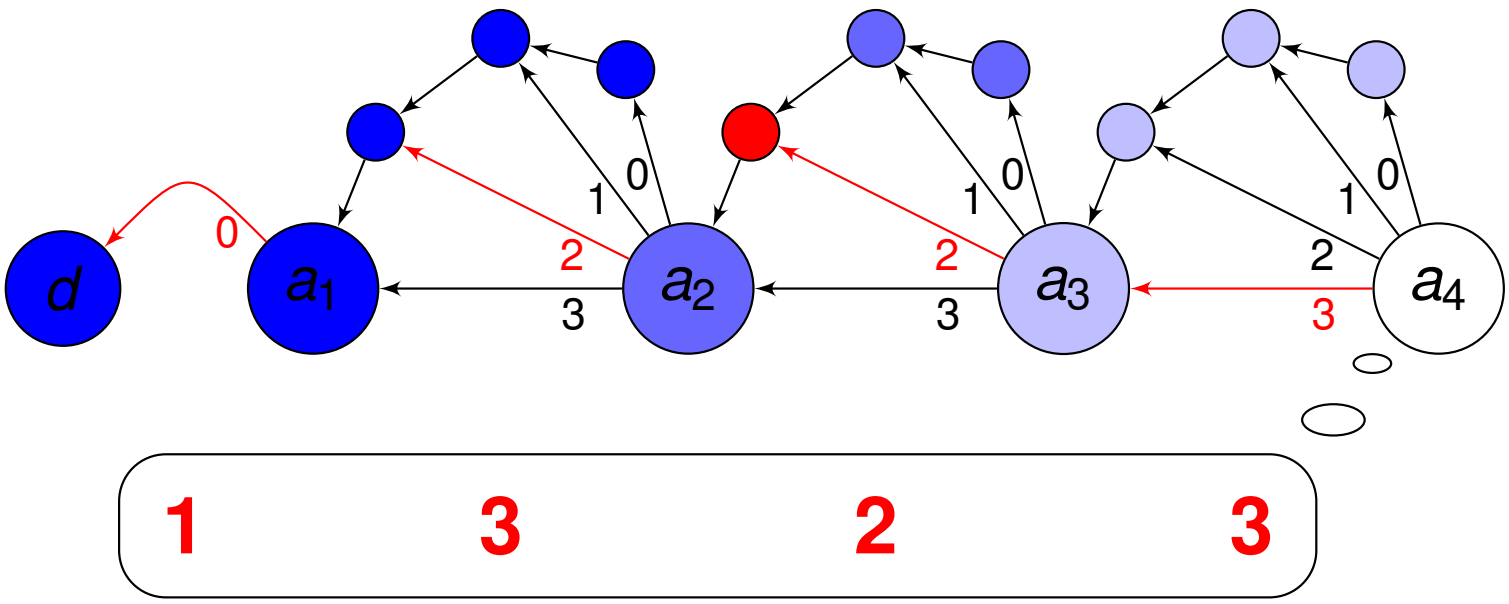
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

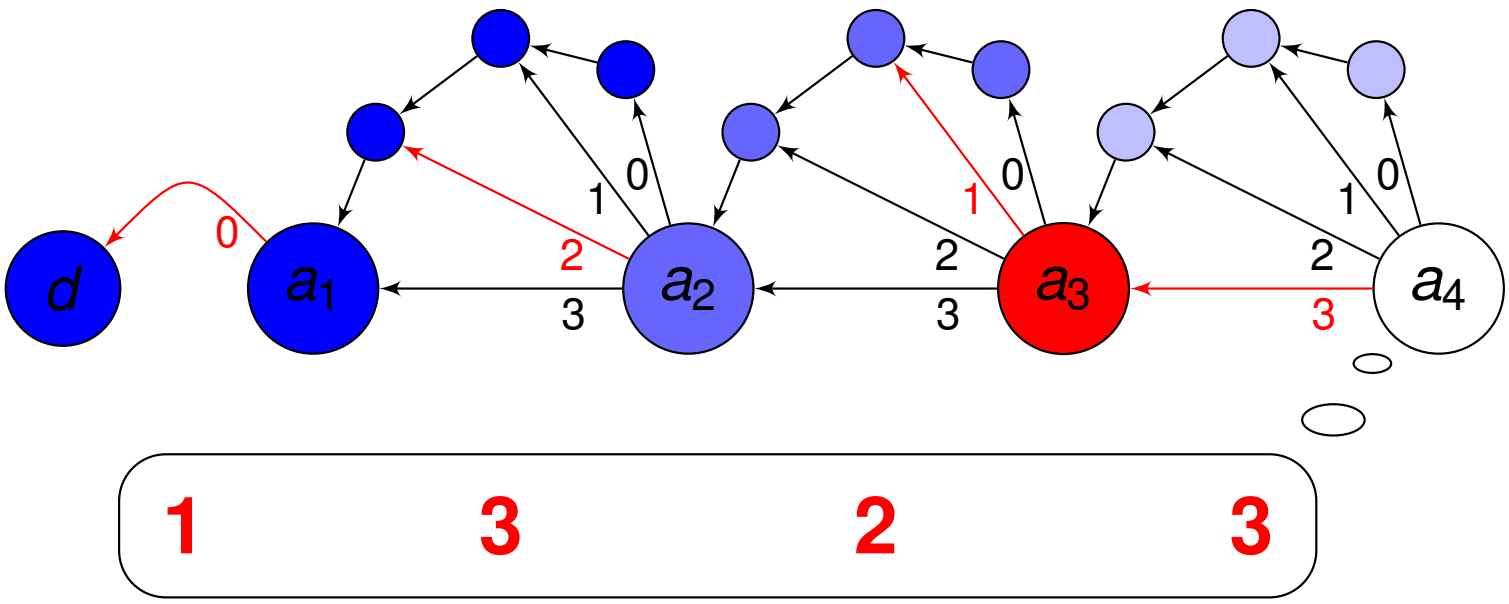
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

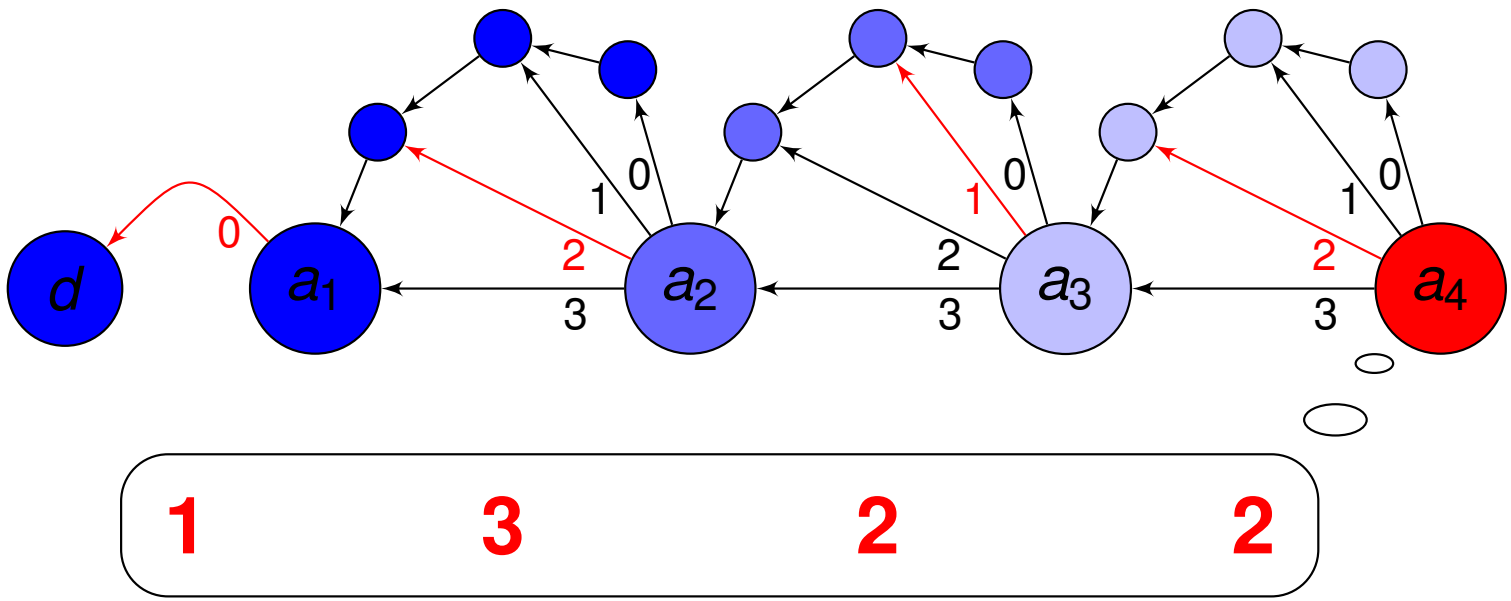
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

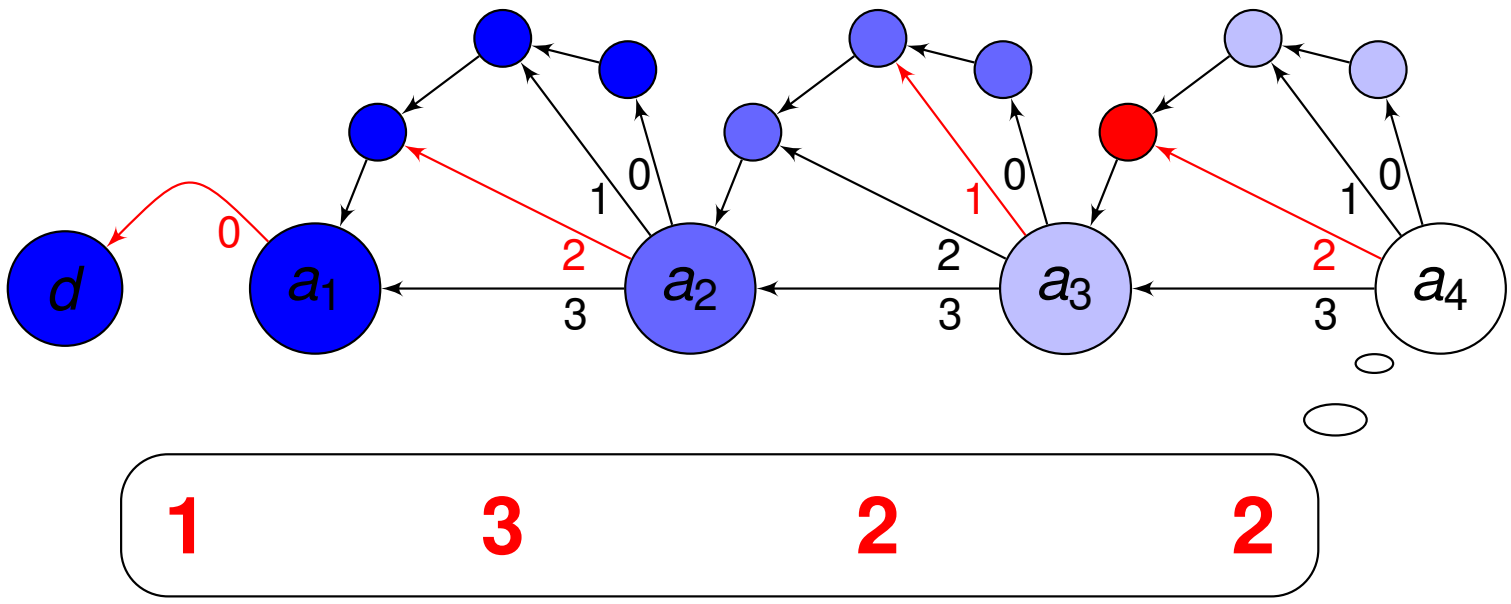
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

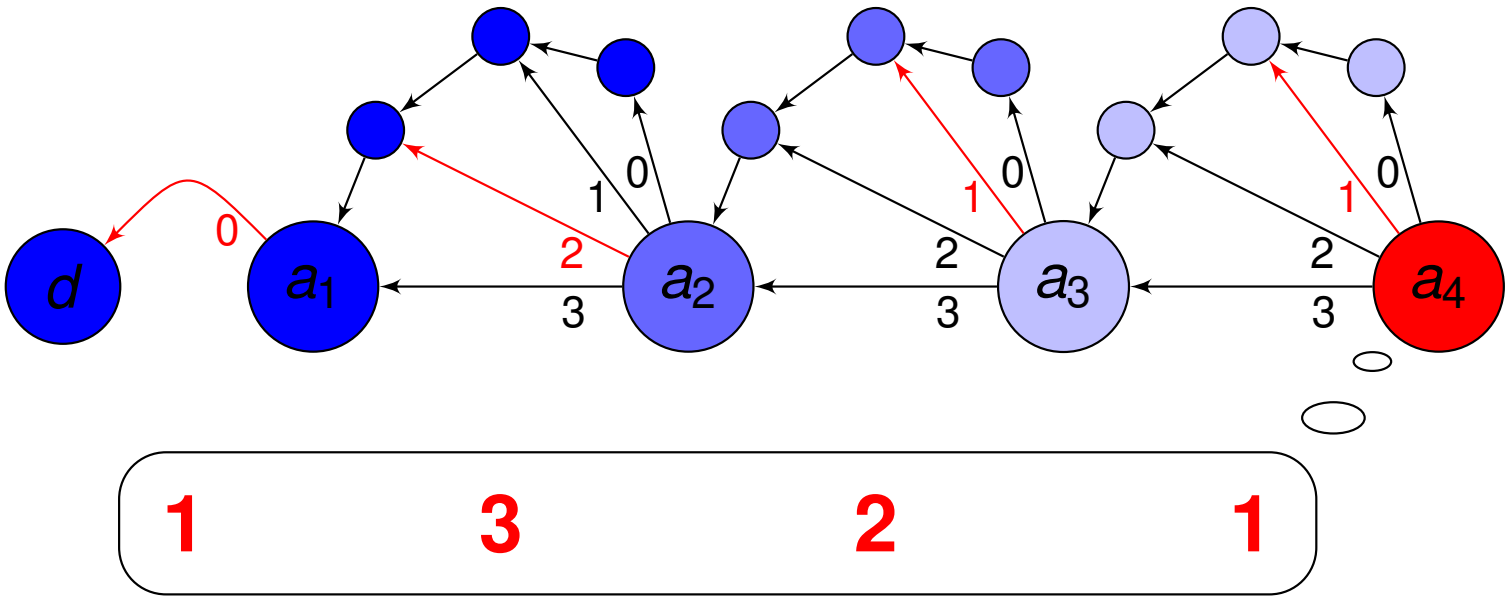
- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- Darker = slower MRAI
- a_4 counts down in base-4 from 1333 to 0333

MRAI Diversity: messages sent

- Many time scales $\Rightarrow$ old information can traverse combinatorially many paths before disappearing:



- a_4 sends $\Theta\left(\left(\frac{n}{s}\right)^s\right) = \Theta(\alpha^s)$ messages
- Total messages and forwarding updates: same bound

MRAI Diversity: upper bound on messages

- "Good" news: doesn't get much worse
- We prove: a worst-case timing gives $O(n^{2s})$ convergence
- Roughly: "Steinerize" known bounds (Sami et al), and carefully look at worst-case sequences of time offsets

Convergence time and MRAI diversity vs disparity

- Time: grows linearly with the slowest MRAI (each activation = fair phase), but is that really a consolation?
- Linear is better than exponential, but think about the numbers!
- Incremental deployment of 30 sec $\rightarrow$ $\sim$100 ms might cause:
  - no time improvements
  - 300-fold increase in #messages?!
- To measure convergence *time*, need to consider MRAI disparity

MRAI Disparity: upper bounds

- Let $r = \frac{\max \text{MRAI}}{\min \text{MRAI}}$
- [SSZ'09] bounds the convergence to n phases:
  Slowest AS: n activations
  Fastest AS: nr activations
- Upper bounds are easy consequences of [SSZ'09]:

MRAI Disparity: upper bounds

- Let $r = \frac{\max \text{MRAI}}{\min \text{MRAI}}$
- [SSZ'09] bounds the convergence to n phases:
  Slowest AS: n activations
  Fastest AS: nr activations
- Upper bounds are easy consequences of [SSZ'09]:
  1. Time: $O(nr)$ (units: fastest MRAI)

MRAI Disparity: upper bounds

- Let $r = \frac{\max \text{MRAI}}{\min \text{MRAI}}$
- [SSZ'09] bounds the convergence to n phases:
  Slowest AS: n activations
  Fastest AS: nr activations
- Upper bounds are easy consequences of [SSZ'09]:
  1. Time: $O(nr)$ (units: fastest MRAI)
  2. Max routing updates per edge: $O(nr)$ (1 message/activation)

MRAI Disparity: upper bounds

- Let $r = \frac{\max \text{MRAI}}{\min \text{MRAI}}$
- [SSZ'09] bounds the convergence to n phases:
  Slowest AS: n activations
  Fastest AS: nr activations
- Upper bounds are easy consequences of [SSZ'09]:
  1. Time: $O(nr)$ (units: fastest MRAI)
  2. Max routing updates per edge: $O(nr)$ (1 message/activation)
  3. Total routing updates: $O(nrE)$

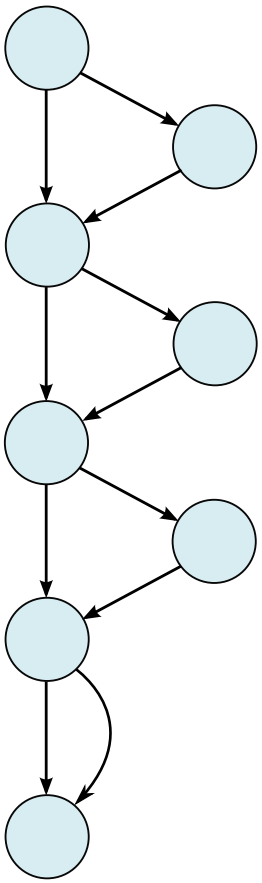
MRAI Disparity: upper bounds

- Let $r = \frac{\max \text{MRAI}}{\min \text{MRAI}}$
- [SSZ'09] bounds the convergence to n phases:
  Slowest AS: n activations
  Fastest AS: nr activations
- Upper bounds are easy consequences of [SSZ'09]:
  1. Time: $O(nr)$ (units: fastest MRAI)
  2. Max routing updates per edge: $O(nr)$ (1 message/activation)
  3. Total routing updates: $O(nrE)$
  4. Total forwarding updates: $O(n^2 r)$

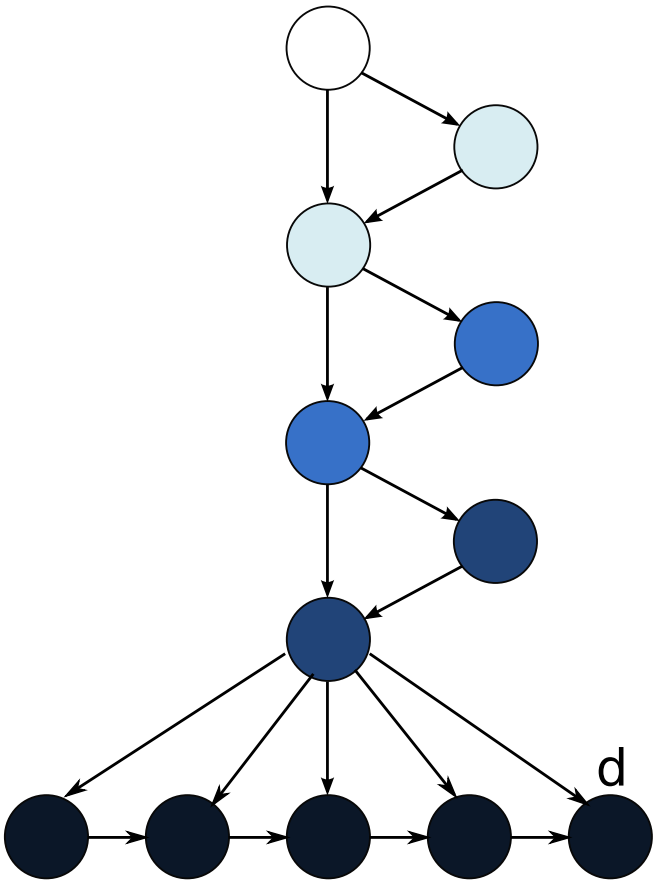
MRAI Disparity: upper bounds

- Let $r = \frac{\max \text{MRAI}}{\min \text{MRAI}}$
- [SSZ'09] bounds the convergence to α phases:
  Slowest AS: α activations
  Fastest AS: αr activations
- Upper bounds are easy consequences of [SSZ'09]:
  1. Time: $O(\alpha r)$ (units: fastest MRAI)
  2. Max routing updates per edge: $O(\alpha r)$ (1 message/activation)
  3. Total routing updates: $O(\alpha r E)$
  4. Total forwarding updates: $O(\alpha^2 r)$

MRAI Diversity: tight lower bounds

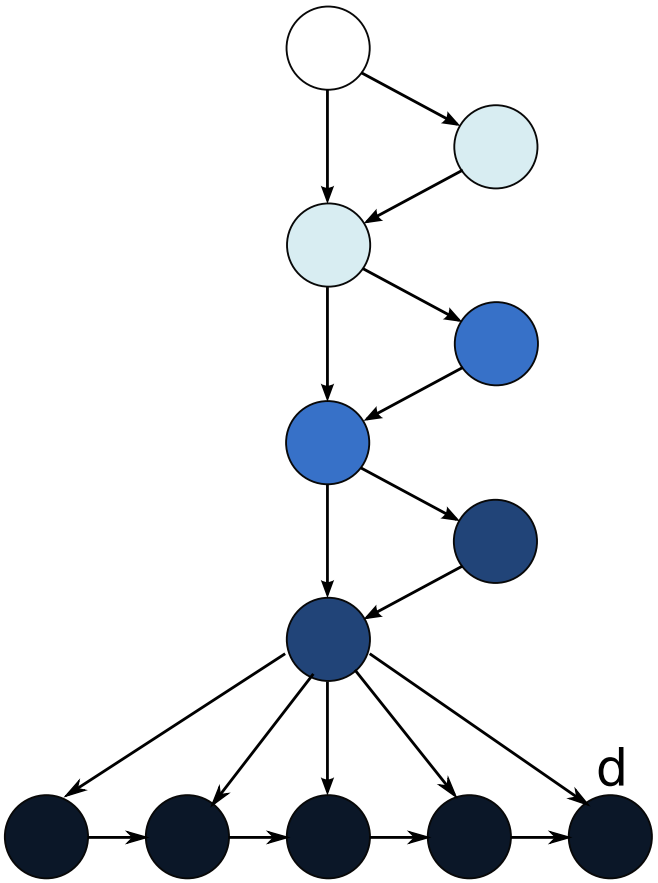


MRAI Diversity: tight lower bounds



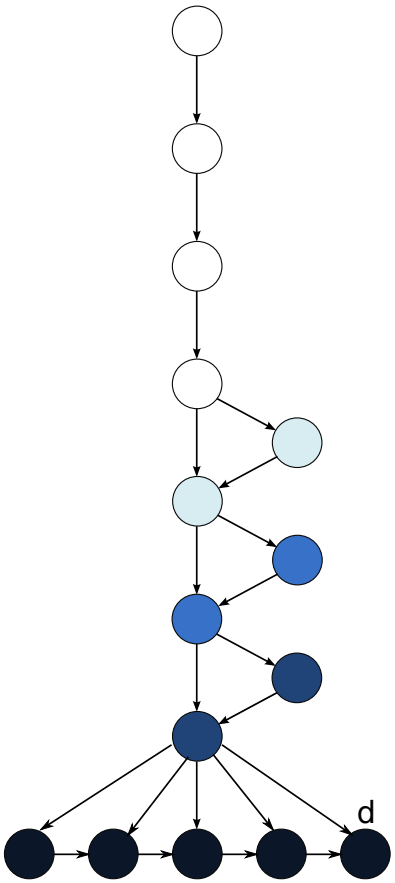
1. Time: $O(nr)$

MRAI Diversity: tight lower bounds



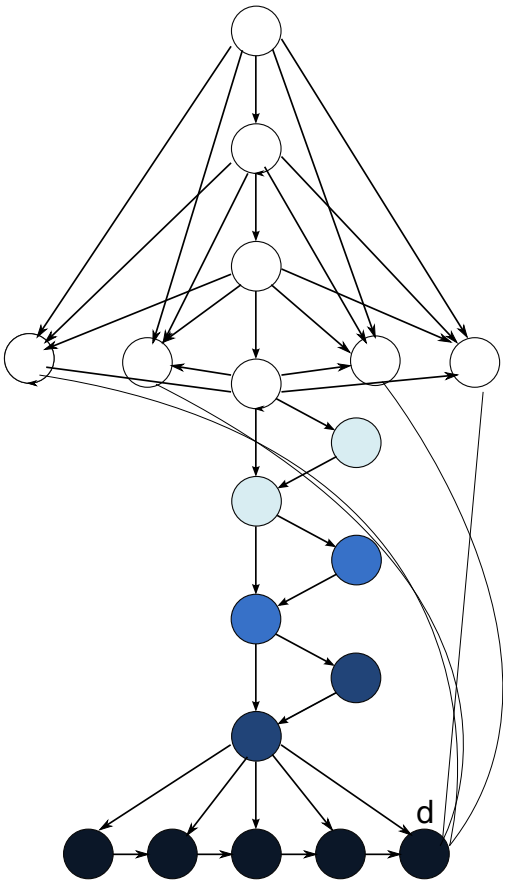
1. Time: $O(nr)$
2. Route updates/edge: $O(nr)$ (1 message/activation)

MRAI Diversity: tight lower bounds



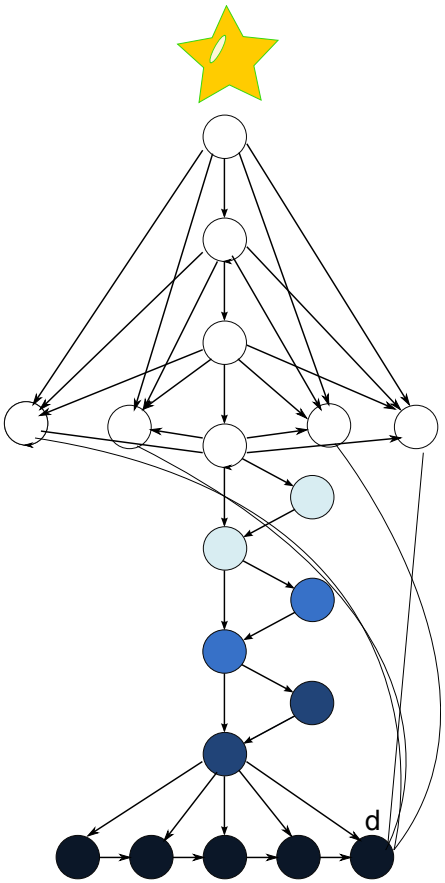
1. Time: $O(nr)$
2. Route updates/edge: $O(nr)$ (1 message/activation)
3. Total route updates: $O(nrE)$ (E: number of edges)

MRAI Diversity: tight lower bounds



1. Time: $O(nr)$
2. Route updates/edge: $O(nr)$ (1 message/activation)
3. Total route updates: $O(nrE)$ (E: number of edges)
4. Total forwarding updates: $O(n^2 r)$ (each node can update $\leq nr$ times)

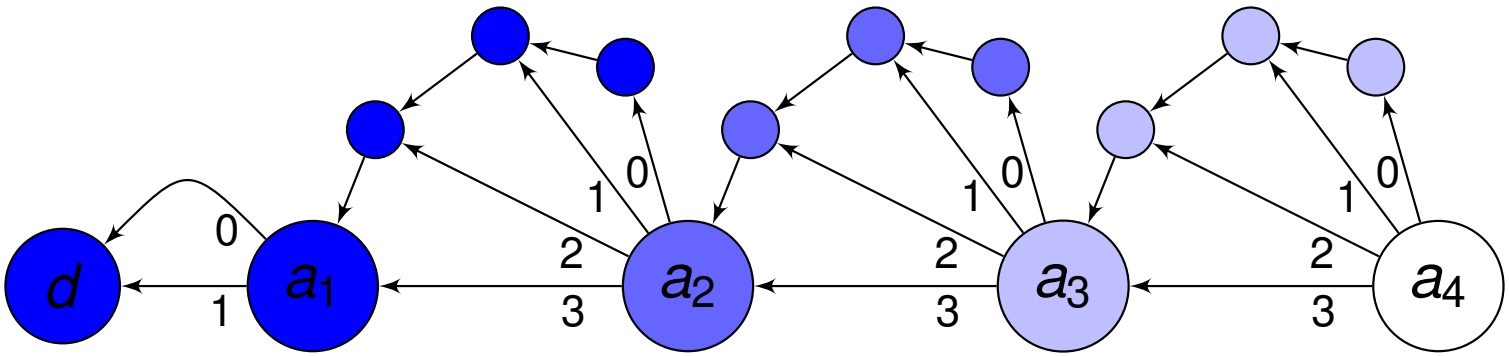
MRAI Diversity: tight lower bounds



1. Time: $O(\alpha r)$
2. Route updates/edge: $O(\alpha r)$ (1 message/activation)
3. Total route updates: $O(\alpha r E)$ (E: number of edges)
4. Total forwarding updates: $O(\alpha^2 r)$ (each node can update $\leq \alpha r$ times)

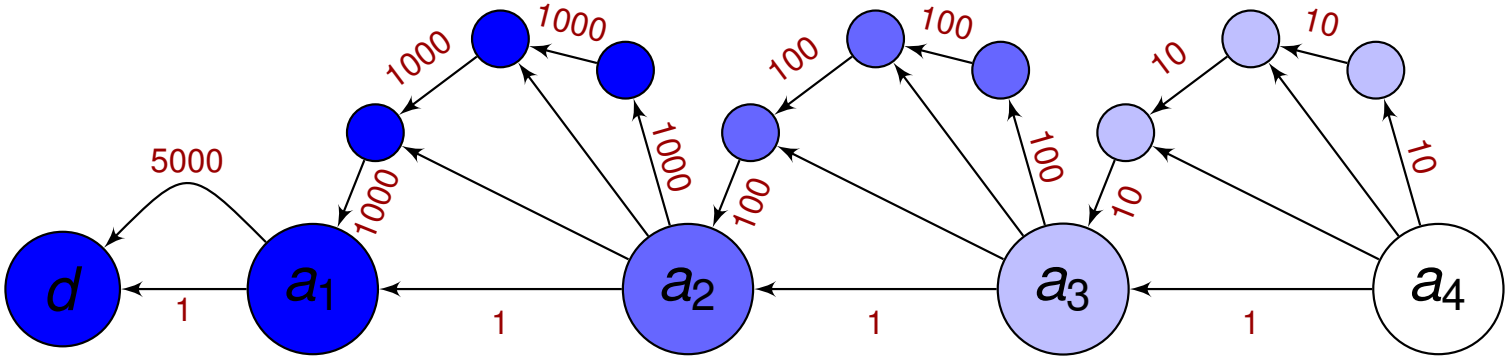
Talk outline

Structures too weird to be interesting?



- "Count in base k": sounds weird?

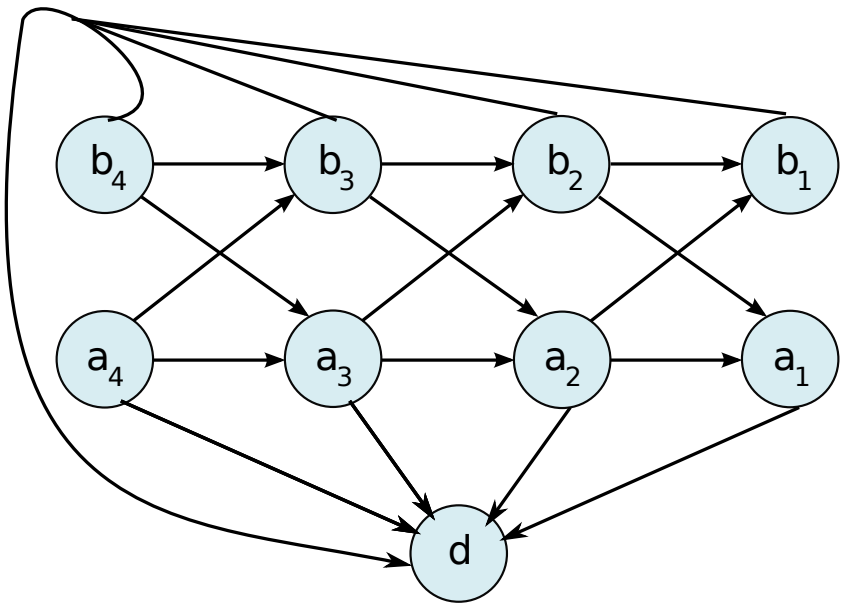
Structures too weird to be interesting?



- "Count in base k": sounds weird?
  - Isomorphic to “optimize latency”!
  - ...or net packet loss (or any semiring)
  - Business relationship constraints (Gao-Rexford) do hold

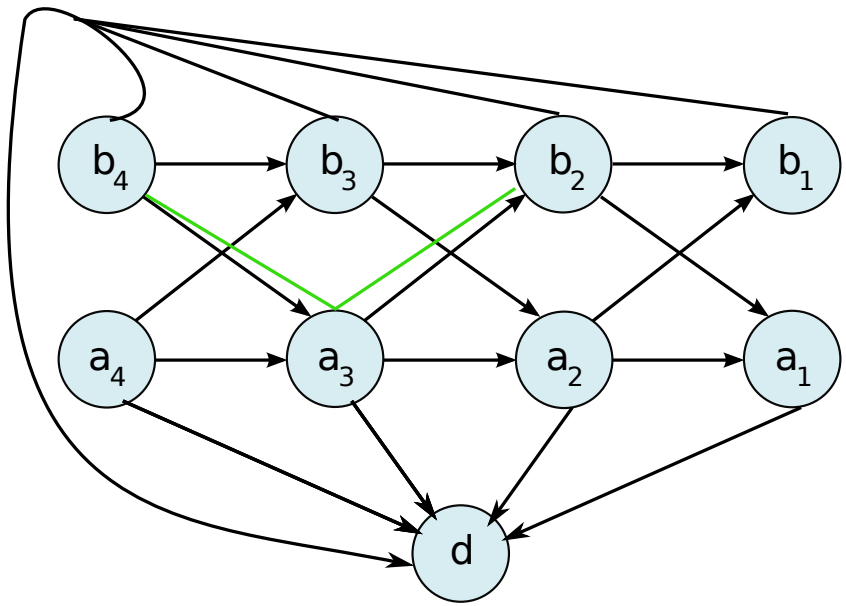
Paths too long? Too many options?

Paths too long? Too many options?



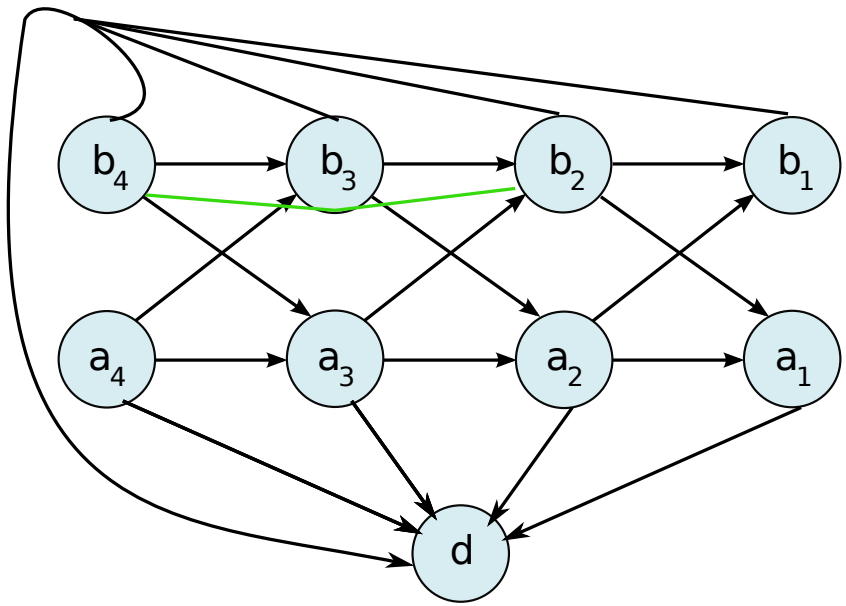
- 4 paths per node, each of length 3, still exponential

Paths too long? Too many options?



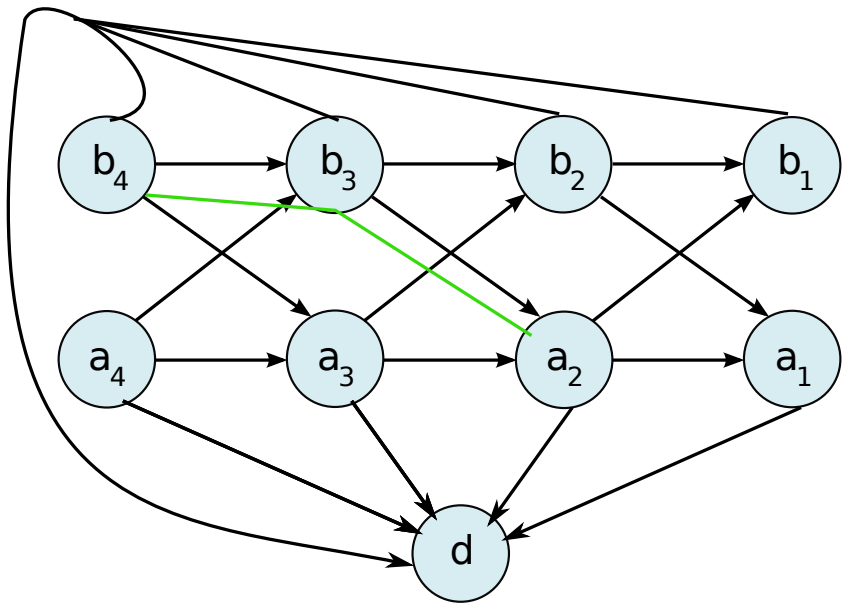
- 4 paths per node, each of length 3, still exponential

Paths too long? Too many options?



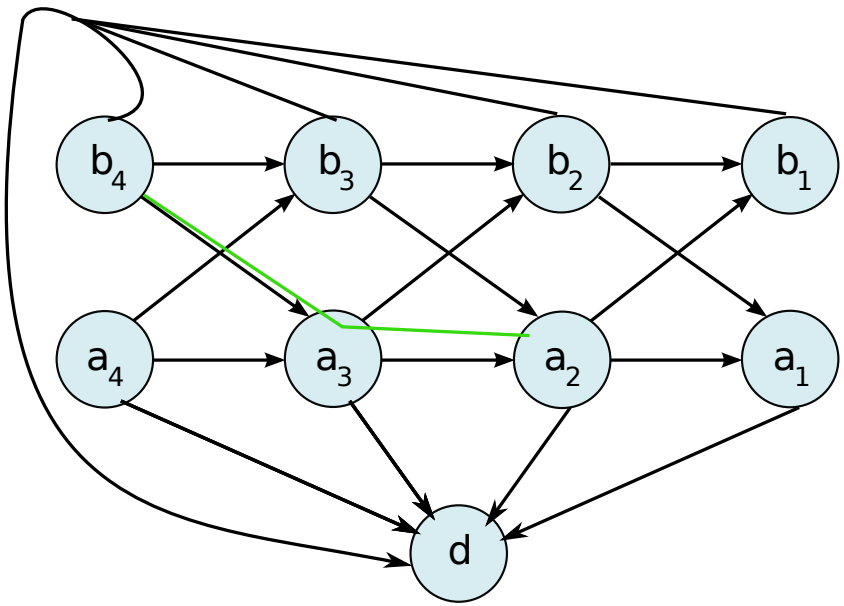
- 4 paths per node, each of length 3, still exponential

Paths too long? Too many options?



- 4 paths per node, each of length 3, still exponential

Paths too long? Too many options?



- 4 paths per node, each of length 3, still exponential

But MRAI has jitter!

- The BGP-4 RFC requires that routers add $\pm 25\%$ jitter to MRAIs to avoid weird resonance behaviors.
- The above bounds rely on tight timings between different MRAIs, but only in one direction:
  - If jitter broadens the gap between two MRAIs, all the above bounds work
- E.g., jitterless exponential bounds $+$ jitter $=$ exponential expectation bounds but with the exponent halved

Talk outline

Data-plane consequences: time

- Data-plane black holes can last as long as control-plane oscillation
- ...even under the worst-case control-plane gadgets in particular

Data-plane consequences: forwarding updates

- The number of forwarding changes is comparable to the number of routing changes ($\Omega(\#\text{ routing changes}/n^2)$)
- Proof: between any two forwarding changes, the routing changes just propagate once up the "routing forest" (can be non-tree in mid-oscillation)

Talk outline

A big ~~open~~ mostly closed problem: MRAI=0

- A MRAI-less world
  - The old model of BGP no longer usable: plentiful transient announcements uncover a lot of wild possibilities [Suchara,F.,Rexford, INFOCOM'11]
  - Lots of implications for BGP theory and engineering (well beyond just convergence behavior)

A big ~~open~~ mostly closed problem: MRAI=0

- A MRAI-less world
  - The old model of BGP no longer usable: plentiful transient announcements uncover a lot of wild possibilities [Suchara,F.,Rexford, ~~INFOCOM'11~~ 20 minutes ago]
  - Lots of implications for BGP theory and engineering (well beyond just convergence behavior)

Somewhat open problem: good ranking functions

- What kinds of local preference functions are sufficient to eliminate the problem?
- We prove: next-hop-only (and consistent-export) local preferences yield polynomial convergence in all senses:
  1. FIB updates: at most once per neighbor
  2. Routing changes: $M = O(n^2 \cdot \#\text{ fwding updates})$
- Open: More realistic good families of local preference functions?

Open problems

- We can't measure others' preferences. Look at real occurence of *risky graph structures*?
- MRAI is usually per-session, not per-prefix: what if multiple destinations are in flux?
- Find simple, practical heuristics extensions of the MRAI mechanism that resolve this:
  - Conjecture: there exists an MRAI-like scheme that slows down announcements about distant destinations, and avoids all the convergence problems above.

Our bottom line

- There is are benefits to globally-uniform MRAI timers
- *Worst-case scenario* will exponentially worsen with proposed changes, so...

Our bottom line

- There is are benefits to globally-uniform MRAI timers
- *Worst-case scenario* will exponentially worsen with proposed changes, so...
  1. Stop.
  2. Collaborate.
  3. And listen.

(with apologies to Vanilla Ice)

Our bottom line

- There is are benefits to globally-uniform MRAI timers
- *Worst-case scenario* will exponentially worsen with proposed changes, so...
  1. The sky isn't falling, but we will benefit from a better-understood deployment of such changes.
  2. Coordinated timing changes are much better – encourage uniformity (unless we have solid evidence for mitigating circumstances
  3. Use an understanding of worst-case patterns to look for them in measurement data to evaluate the risks.