SWINBURNE UNIVERSITY OF TECHNOLOGY

BGP Path Exploration Damping (PED)

Mattia Rossi

mrossi@swin.edu.au

Centre for Advanced Internet Architectures (CAIA)
Swinburne University of Technology

Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
- Reduction of update load
- MRAI and PED convergence time compared

Conclusions and future work

Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
- Reduction of update load
- MRAI and PED convergence time compared

Conclusions and future work

About this talk

- Selected for the Applied Networking Research Prize (ANRP) based on a peer-reviewed paper:
- Paper published in the IEEE Journal on Selected Areas in Communications (JSAC), October 2010 [1]
  - A Technique for Reducing BGP Update Announcements through Path Exploration Damping.
  - Geoff Huston, Mattia Rossi, Grenville Armitage
- Project sponsored by the Cisco University Research Program Fund

[1] http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5586440

What is Path Exploration Damping?

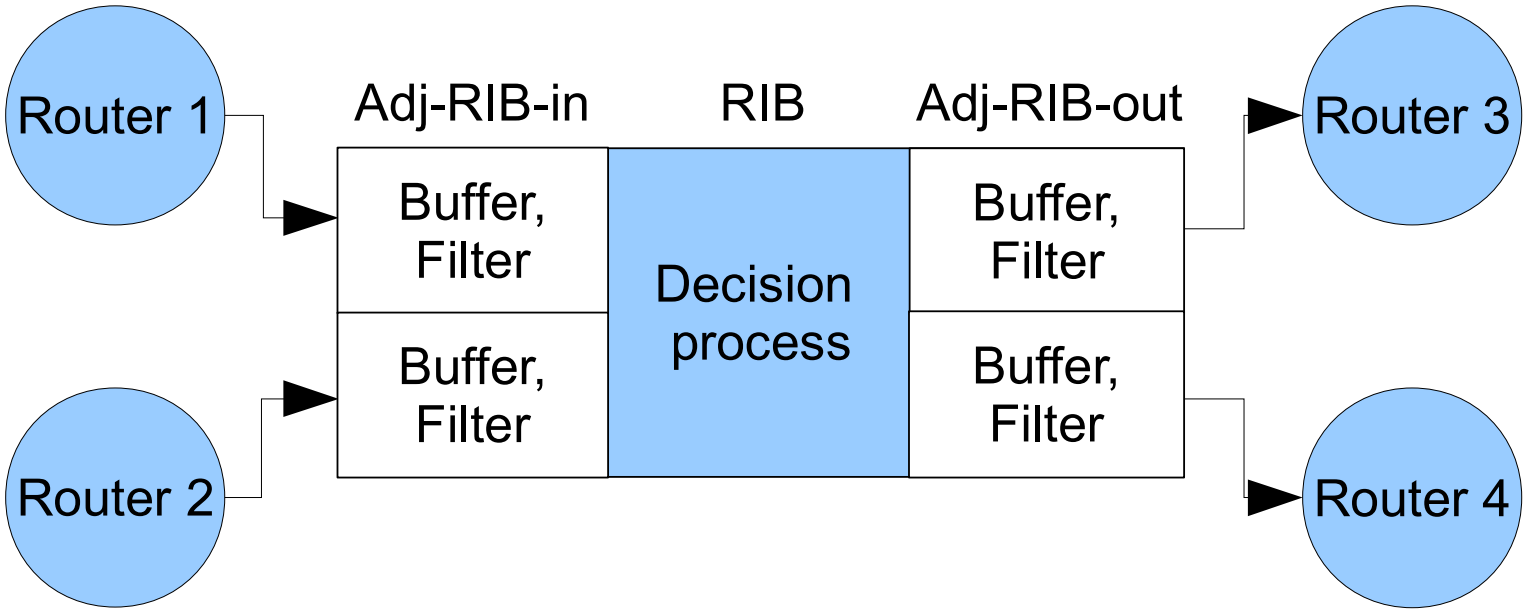
- Path Exploration Damping $\rightarrow$ PED
- Algorithm intended to replace MRAI and RFD
- Methods to reduce update churn and convergence time in BGP
- BGP is the de-facto standard for inter-domain routing
  - BGP – Border Gateway Protocol
  - MRAI – Minimum Route Advertisment Interval
  - RFD – Route Flap Damping

PED implementations

- CAIA
  - Implemented in Quagga
  - Patch available for Quagga 0.99.13
  - Download at http://caia.swin.edu.au/urp/bgp/tools.html
- Cisco (by Mohammed Mirza)
  - Implemented in CISCO IOS-XE Experimental Version 15.1
  - Running on Cisco ASR1002 (2RU)
  - Currently tested at APNIC Pty. Ltd., Australia

Outline

Simplified BGP speaker design

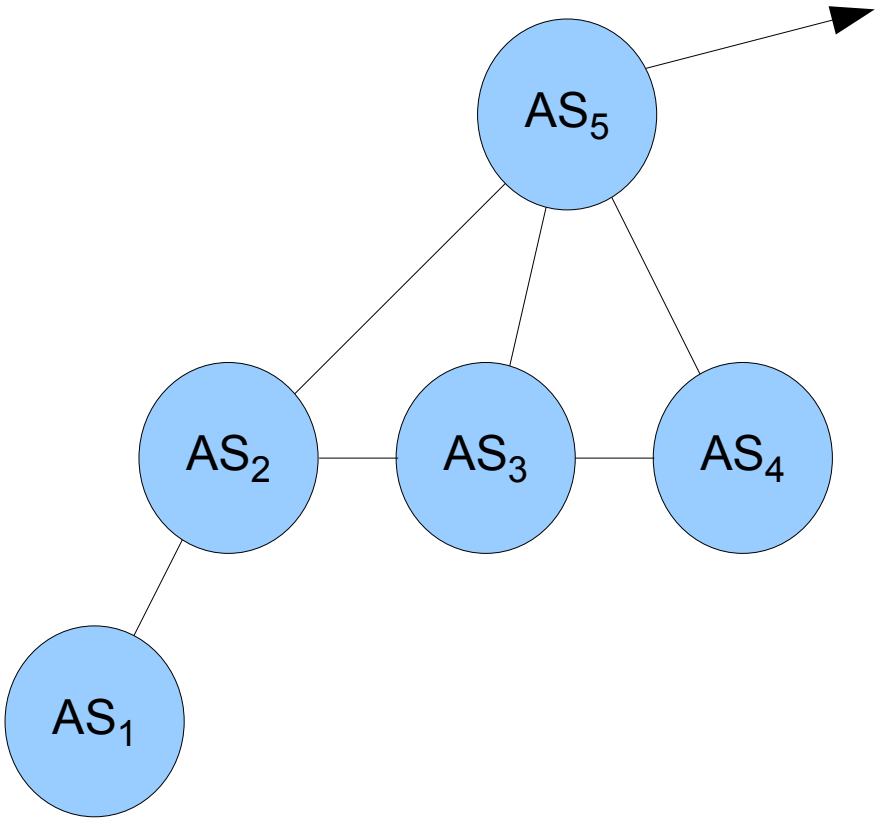


RIB: Routing Information Base – Routing Table

Adj-RIB-(in,out): Adjacency RIB (in,out)

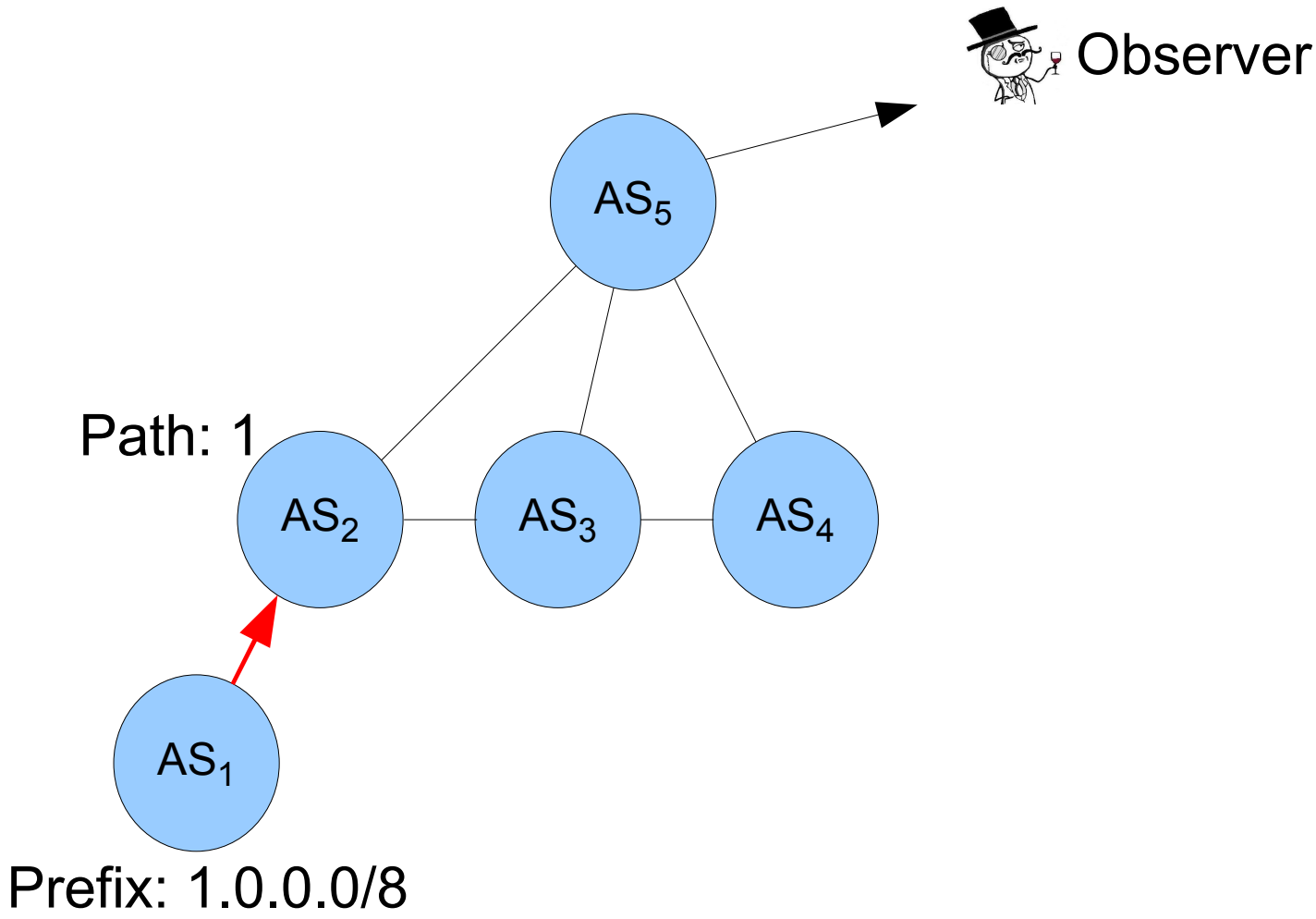
Basic BGP dynamics

Observer

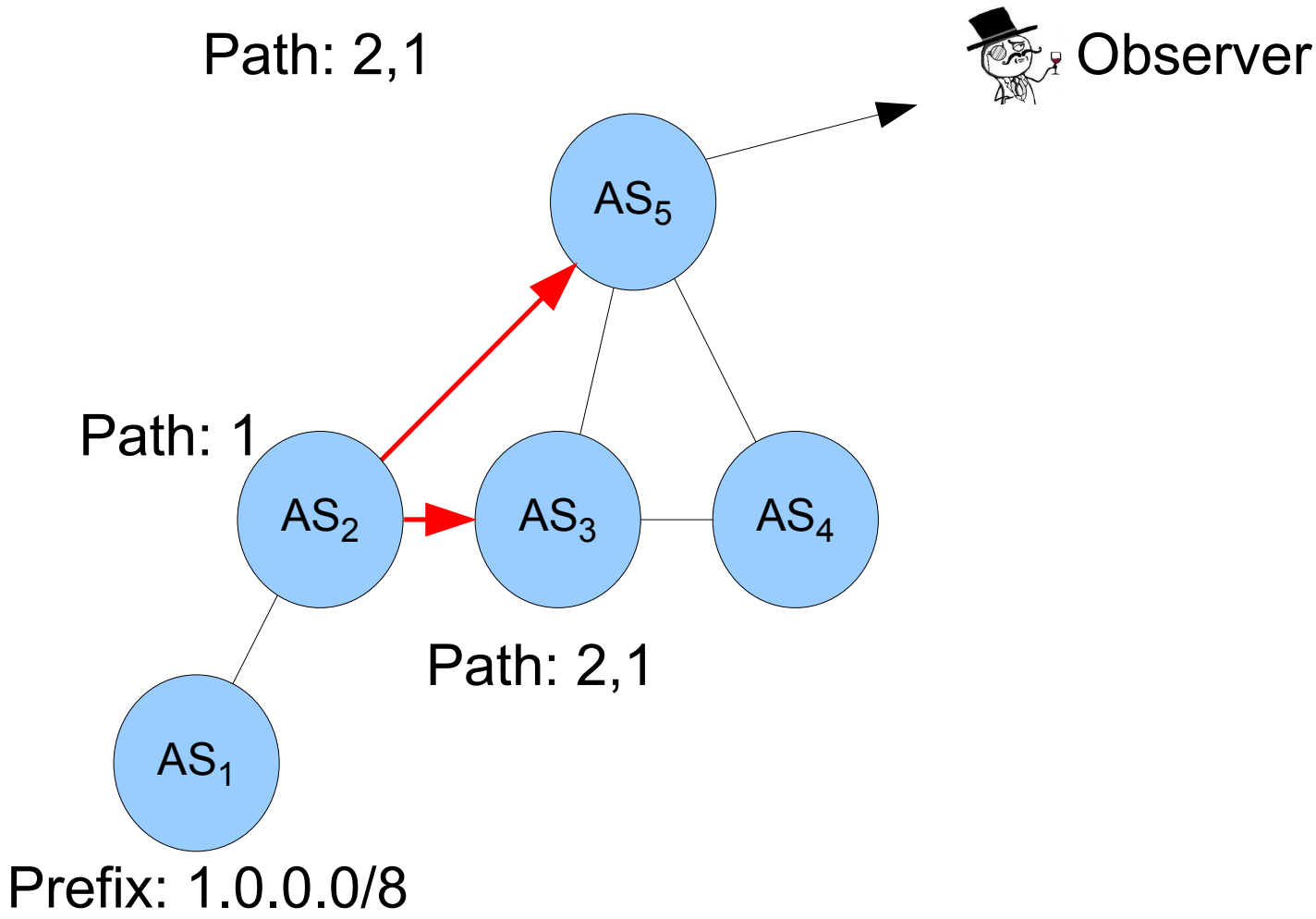


Prefix: 1.0.0.0/8

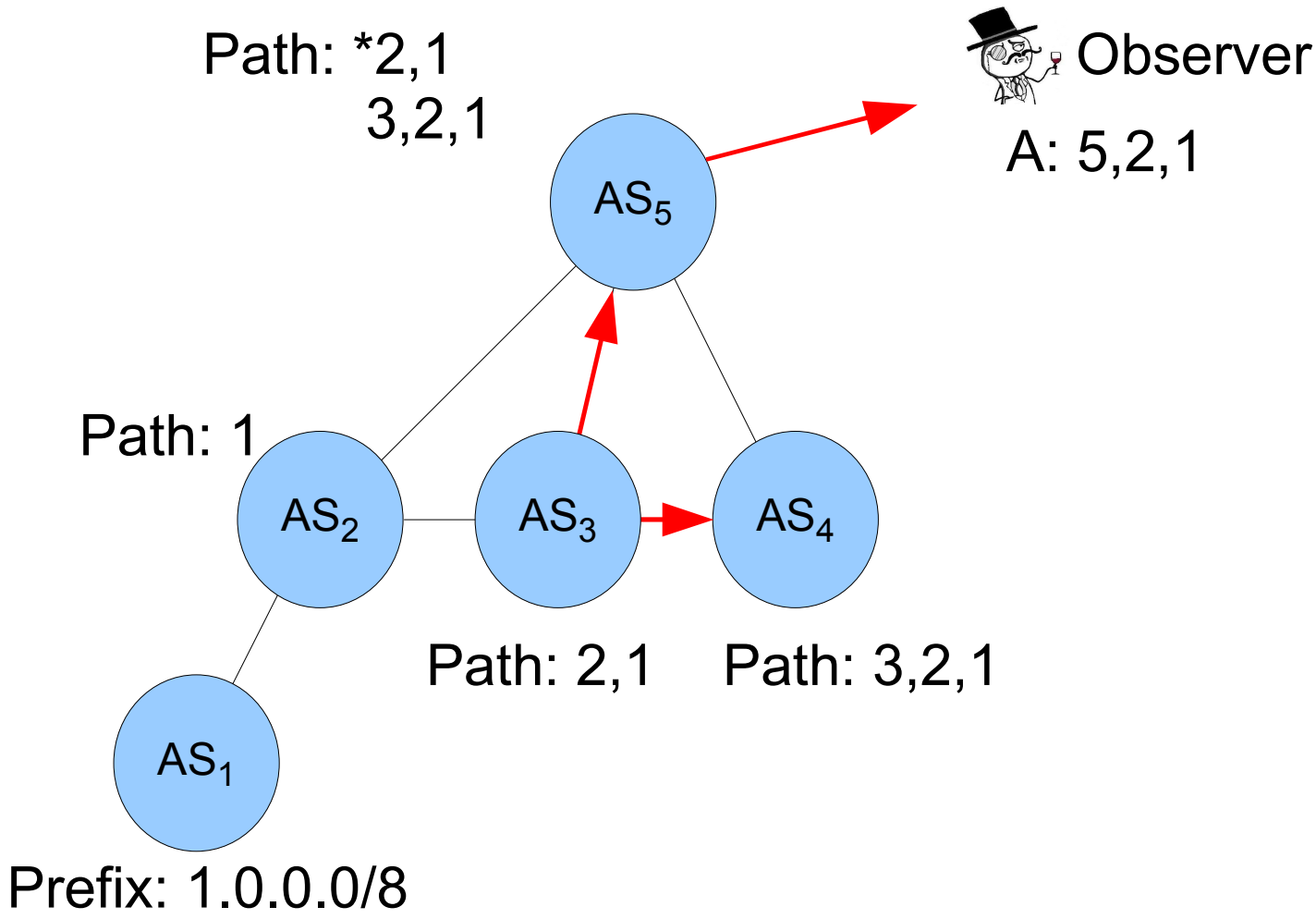
Basic BGP dynamics



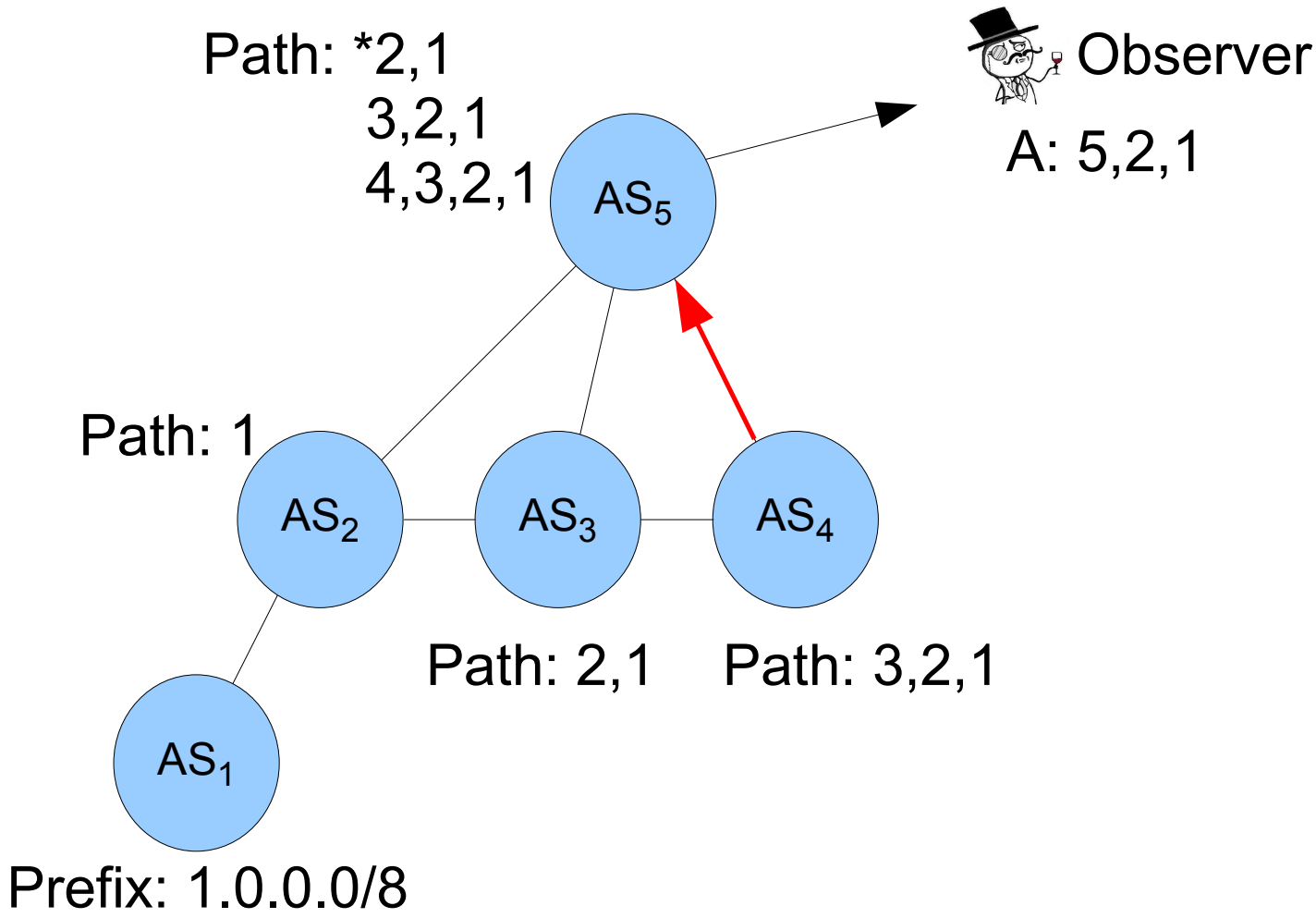
Basic BGP dynamics



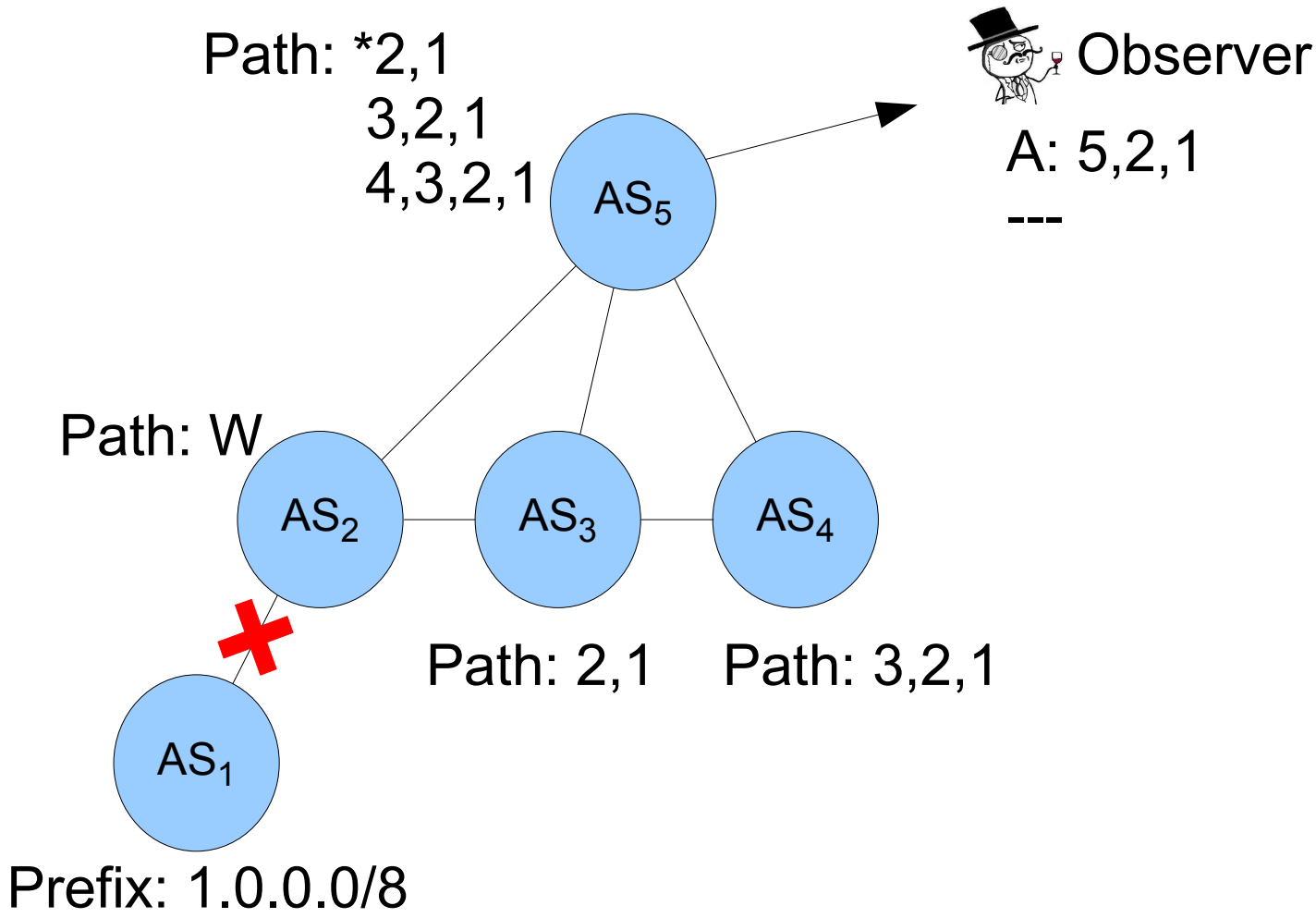
Basic BGP dynamics



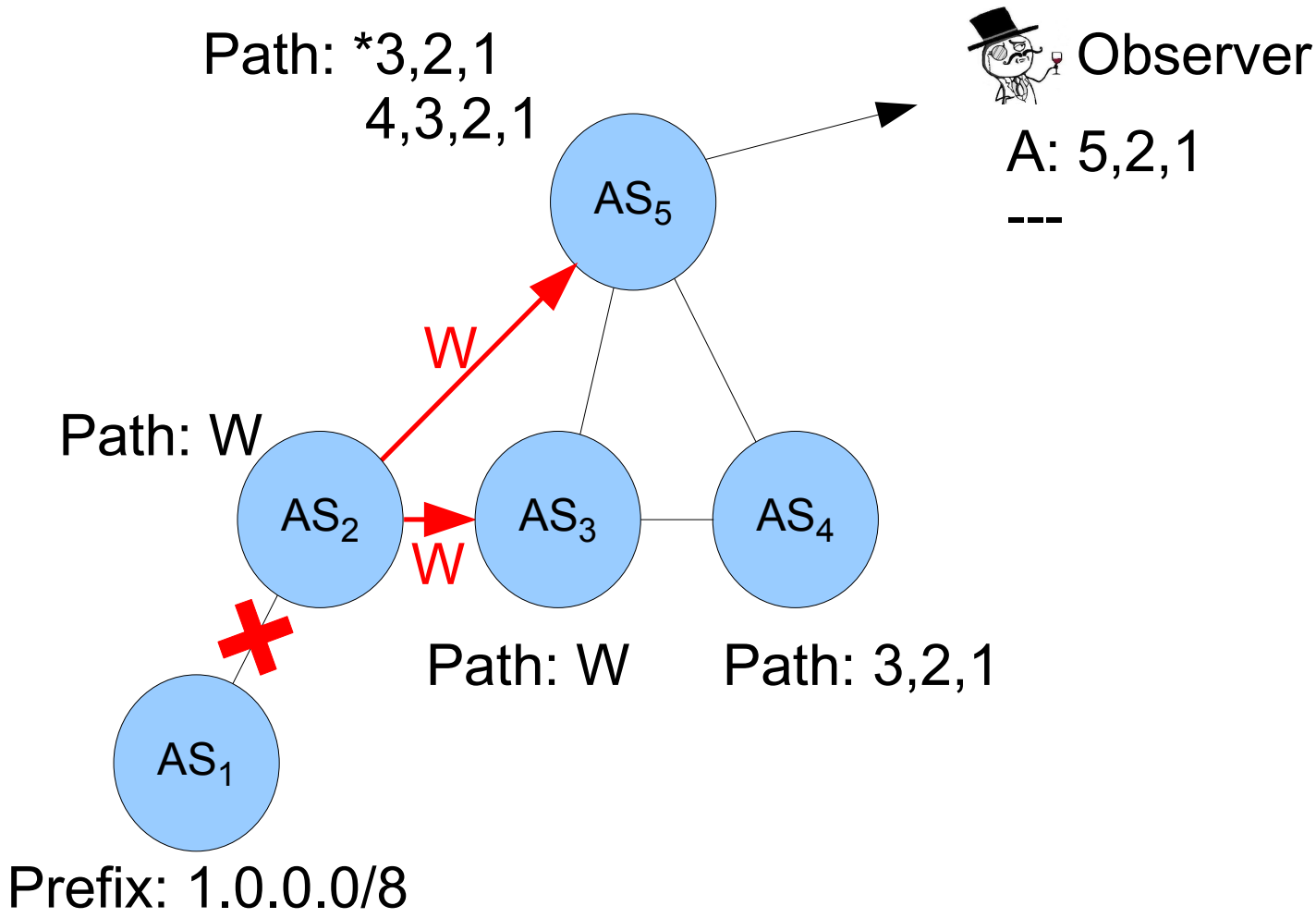
Basic BGP dynamics



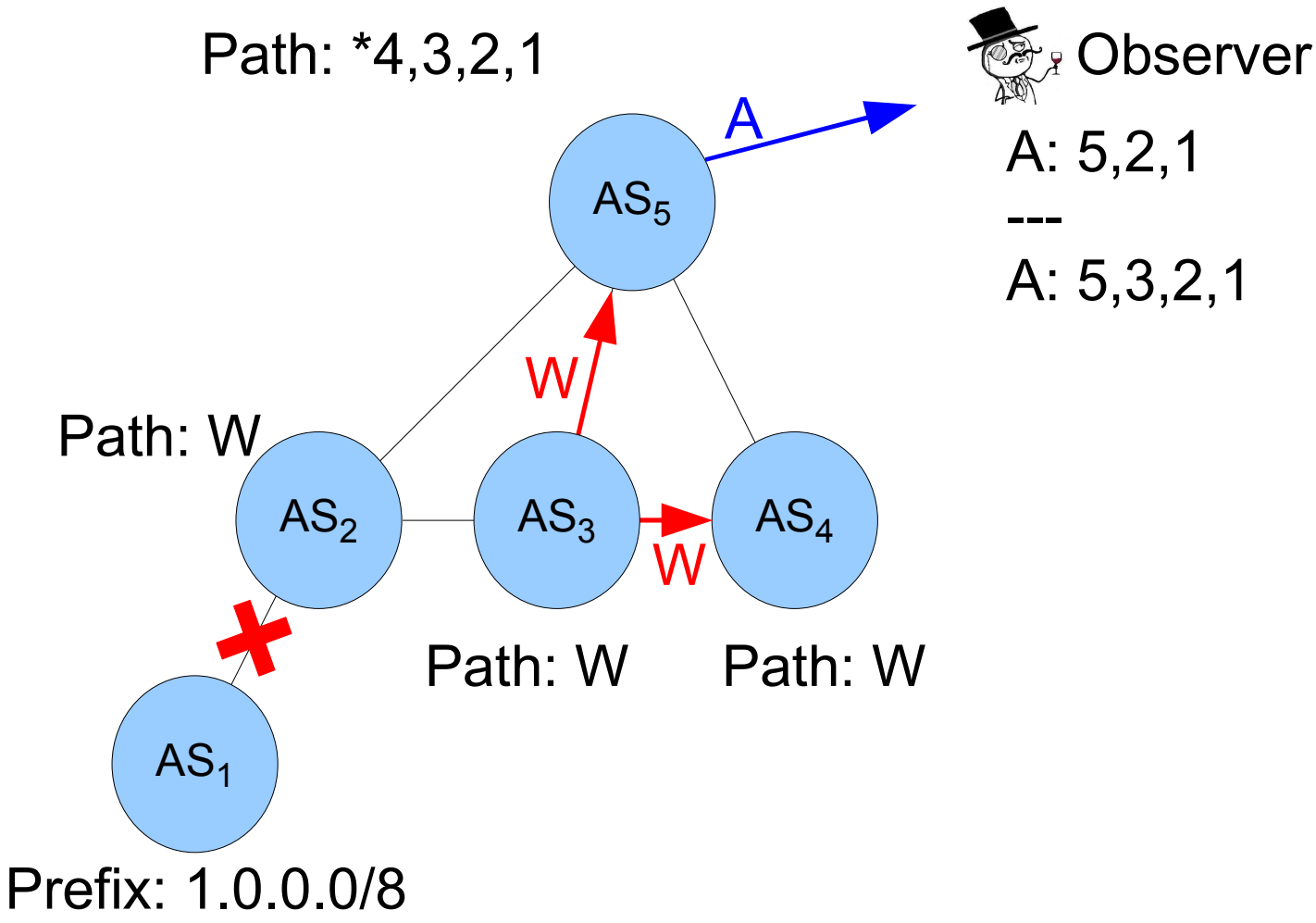
Path Exploration



Path Exploration



Path Exploration

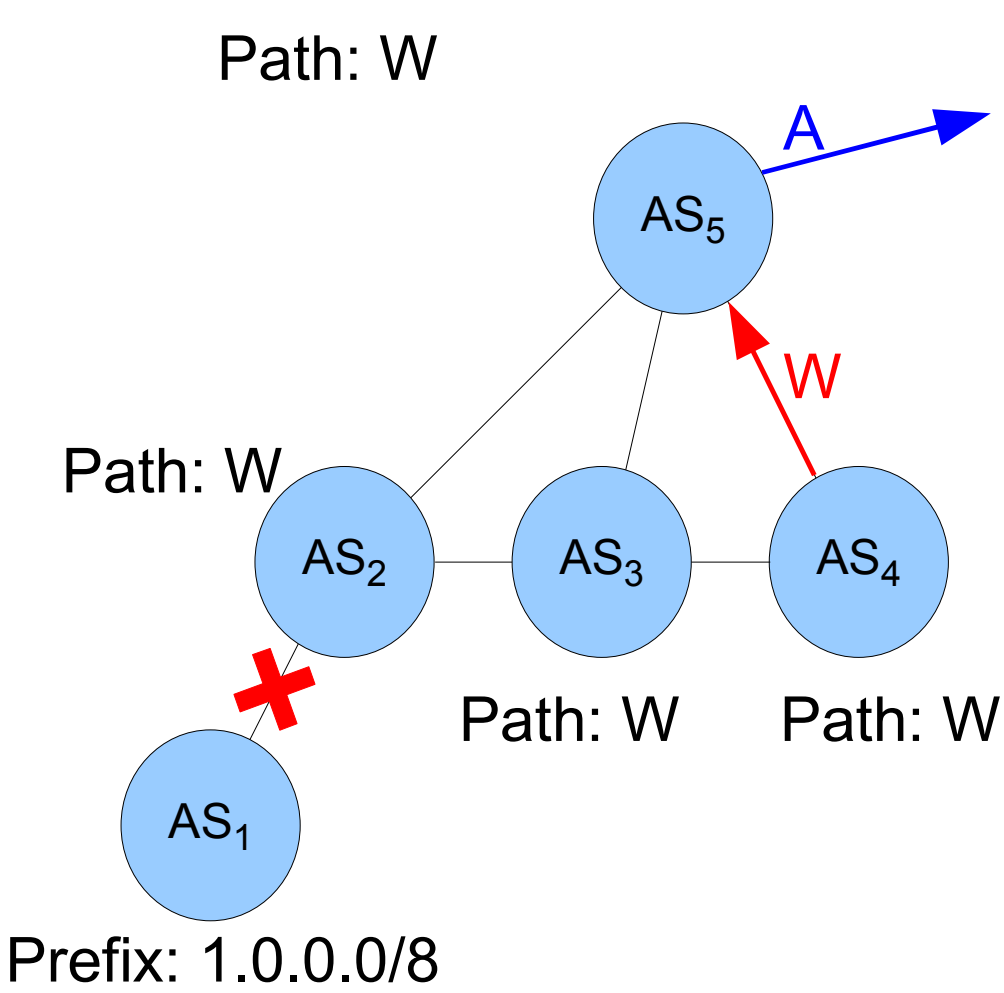


Observer

A: 5,2,1

A: 5,3,2,1

Path Exploration



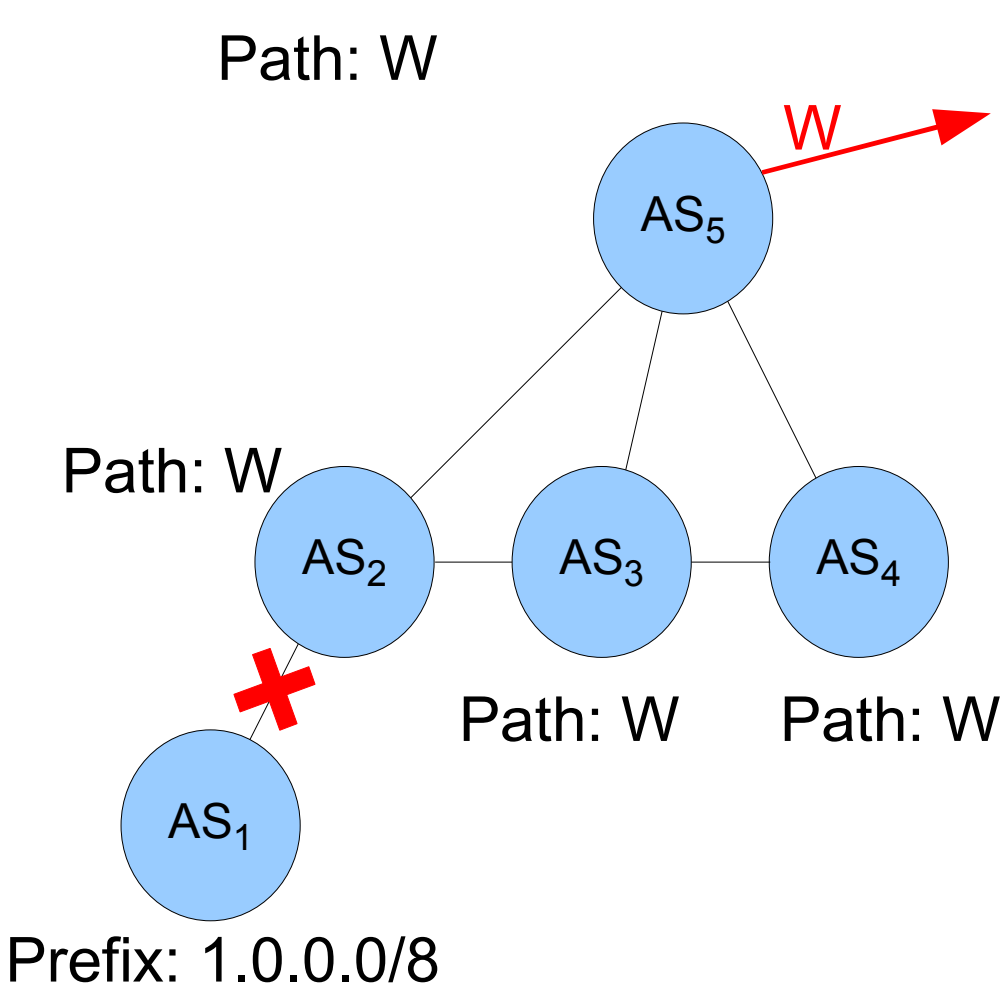
Observer

A: 5,2,1

A: 5,3,2,1

A: 5,4,3,2,1

Path Exploration



Observer

A: 5,2,1

A: 5,3,2,1

A: 5,4,3,2,1

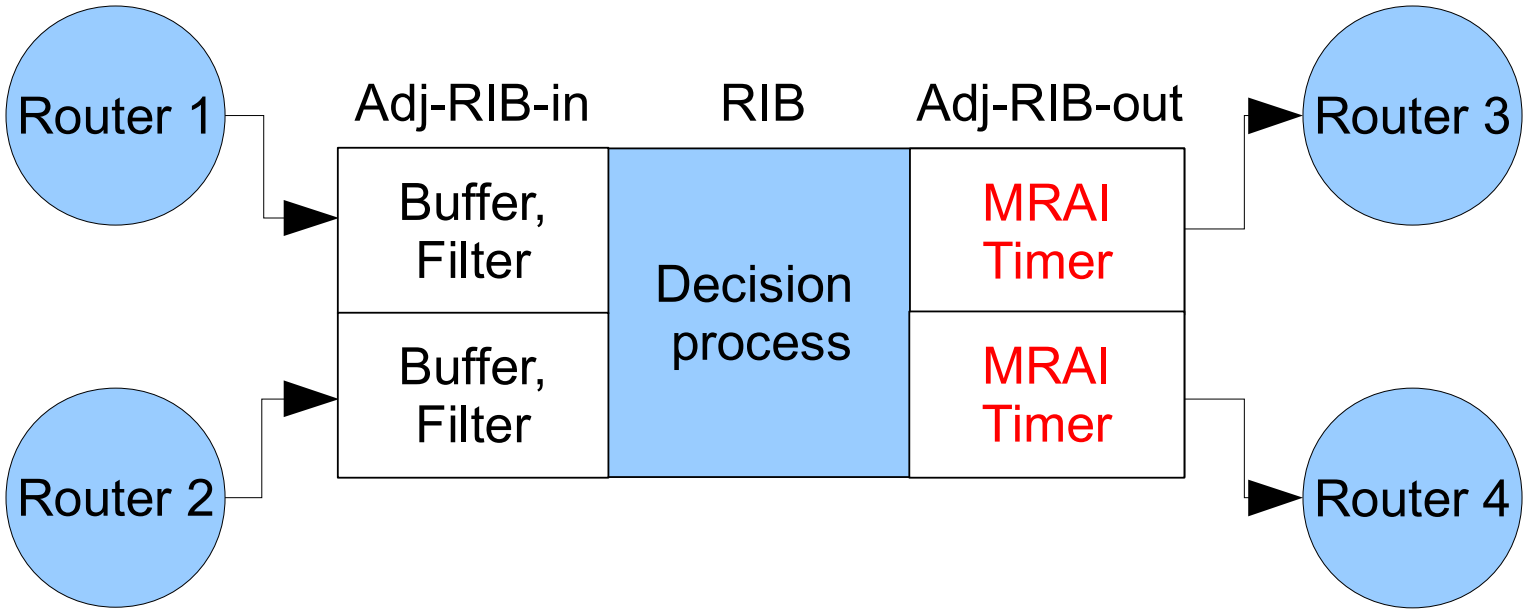
W

What is Path Exploration?

- An update sequence lengthening the AS-path gradually until stability is reached

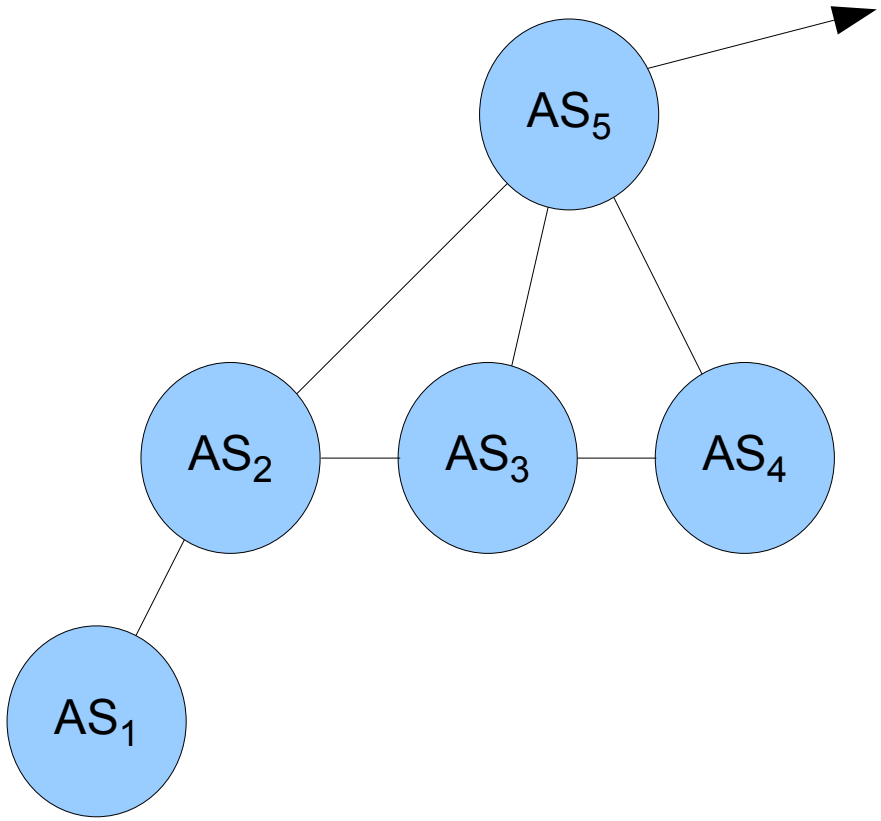
MRAI and RFD

- Need to decrease BGP chattiness and Path Exploration
- Minimum Route Advertisment Interval – MRAI (RFC 1771)
  - Apply 30s (default) delay on announcements



Minimum Route Advertisment Interval – MRAI

Observer

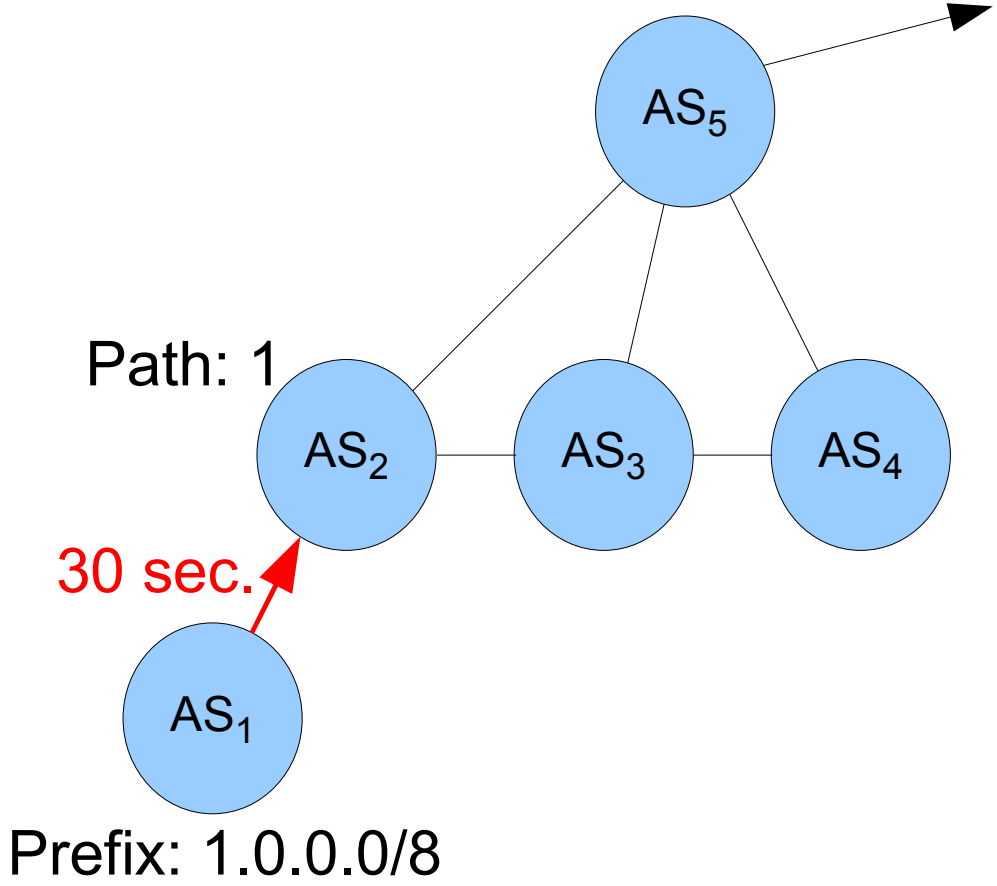


Prefix: 1.0.0.0/8

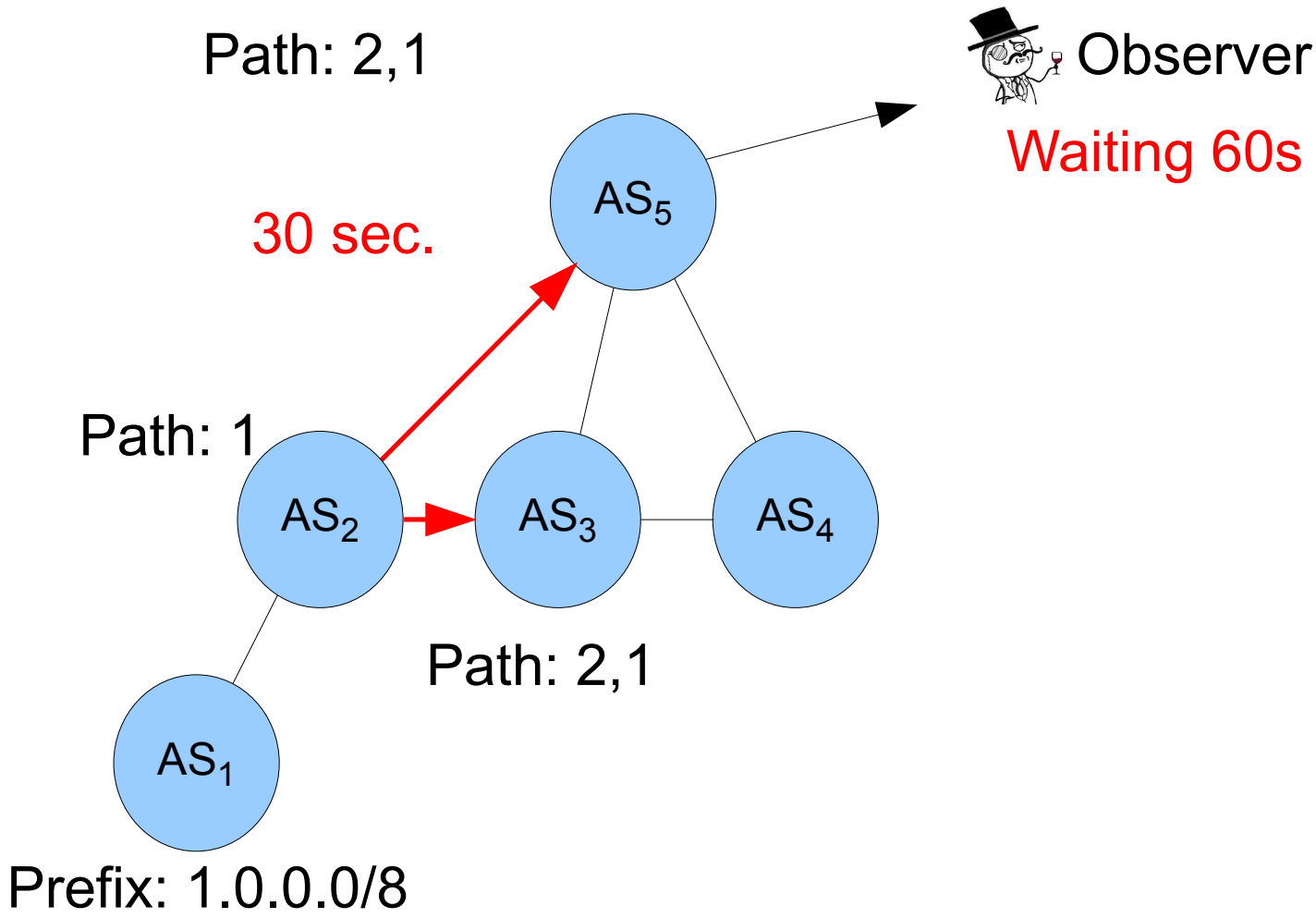
Minimum Route Advertisment Interval – MRAI

Observer

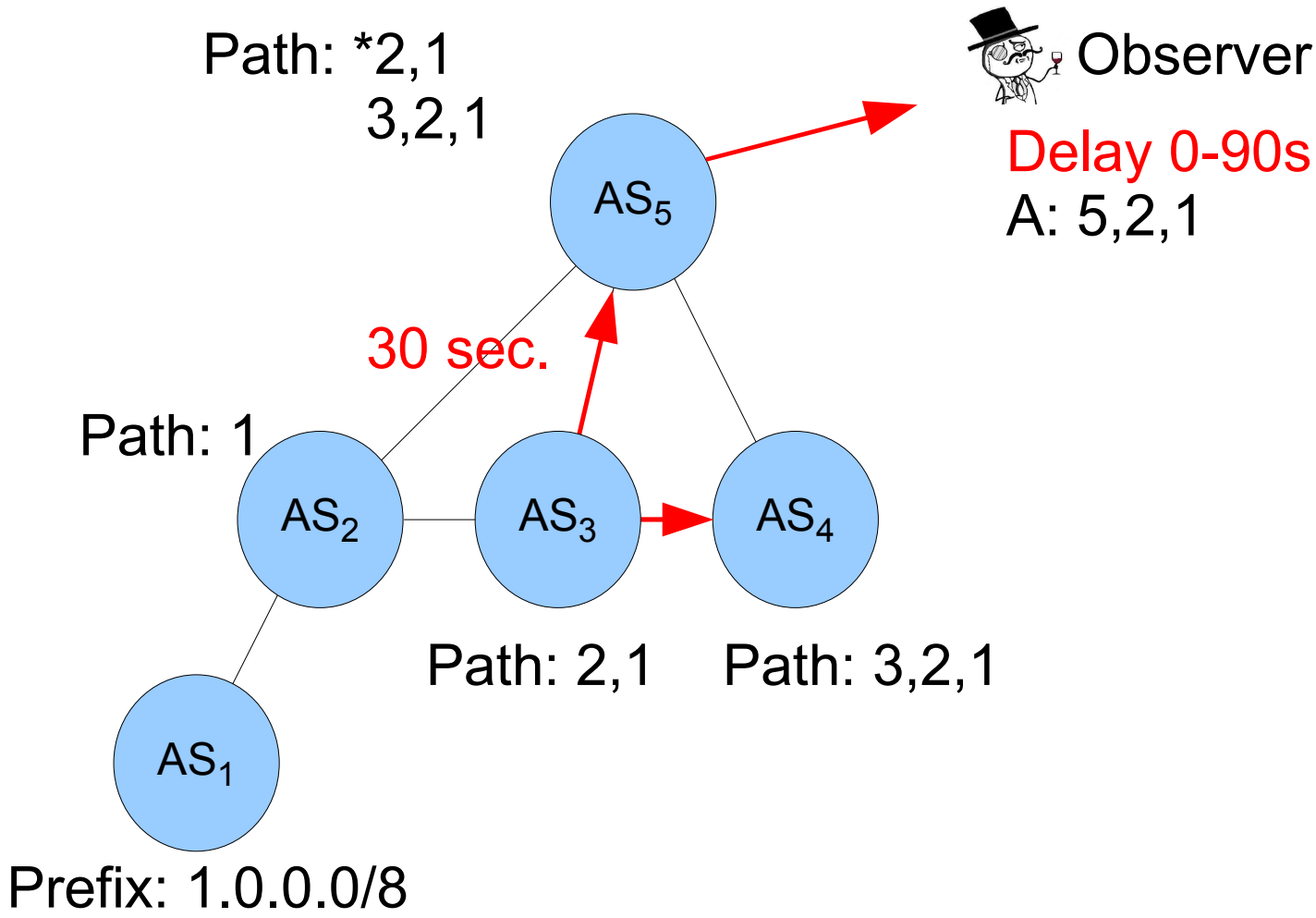
Waiting 30s



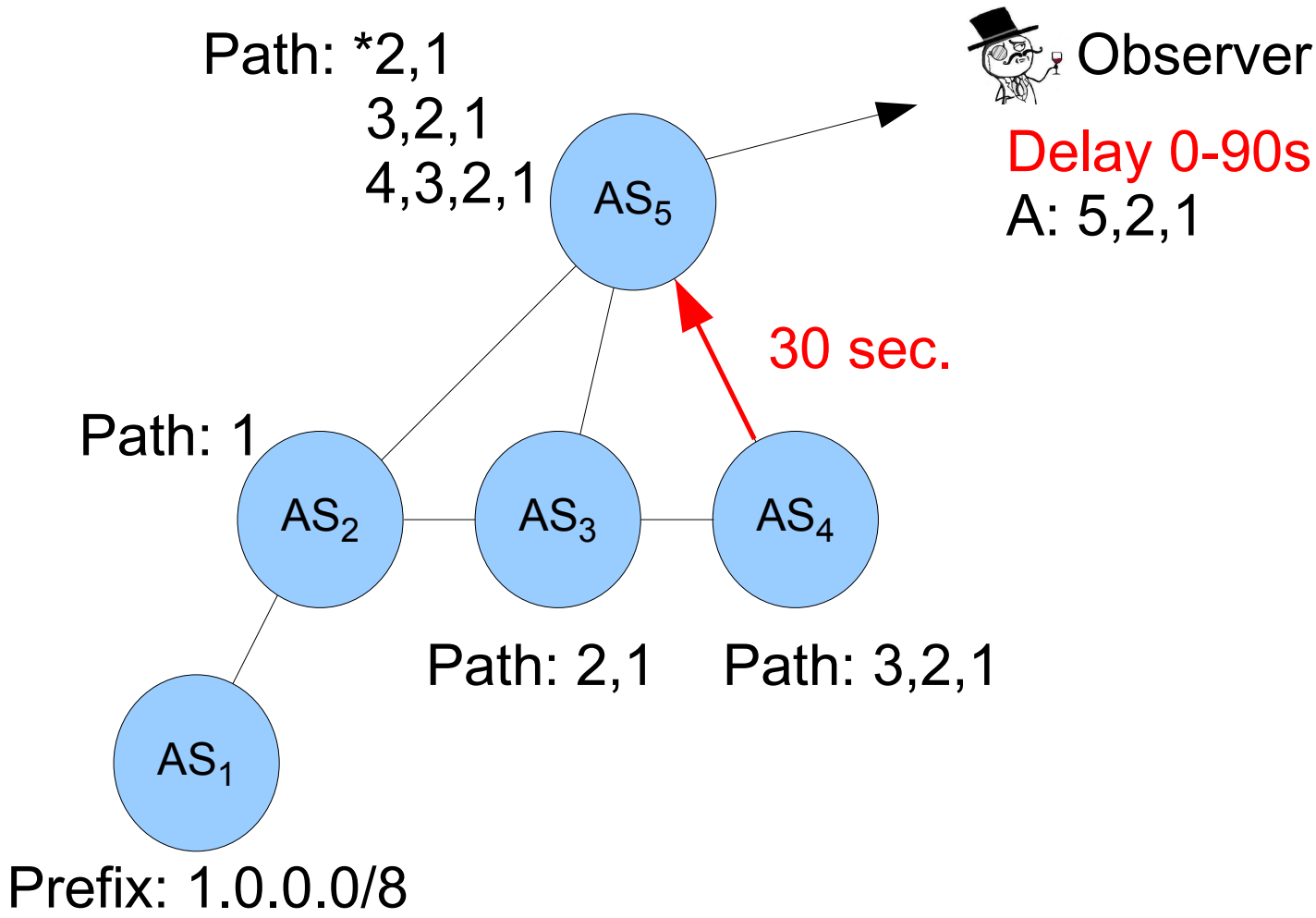
Minimum Route Advertisment Interval – MRAI



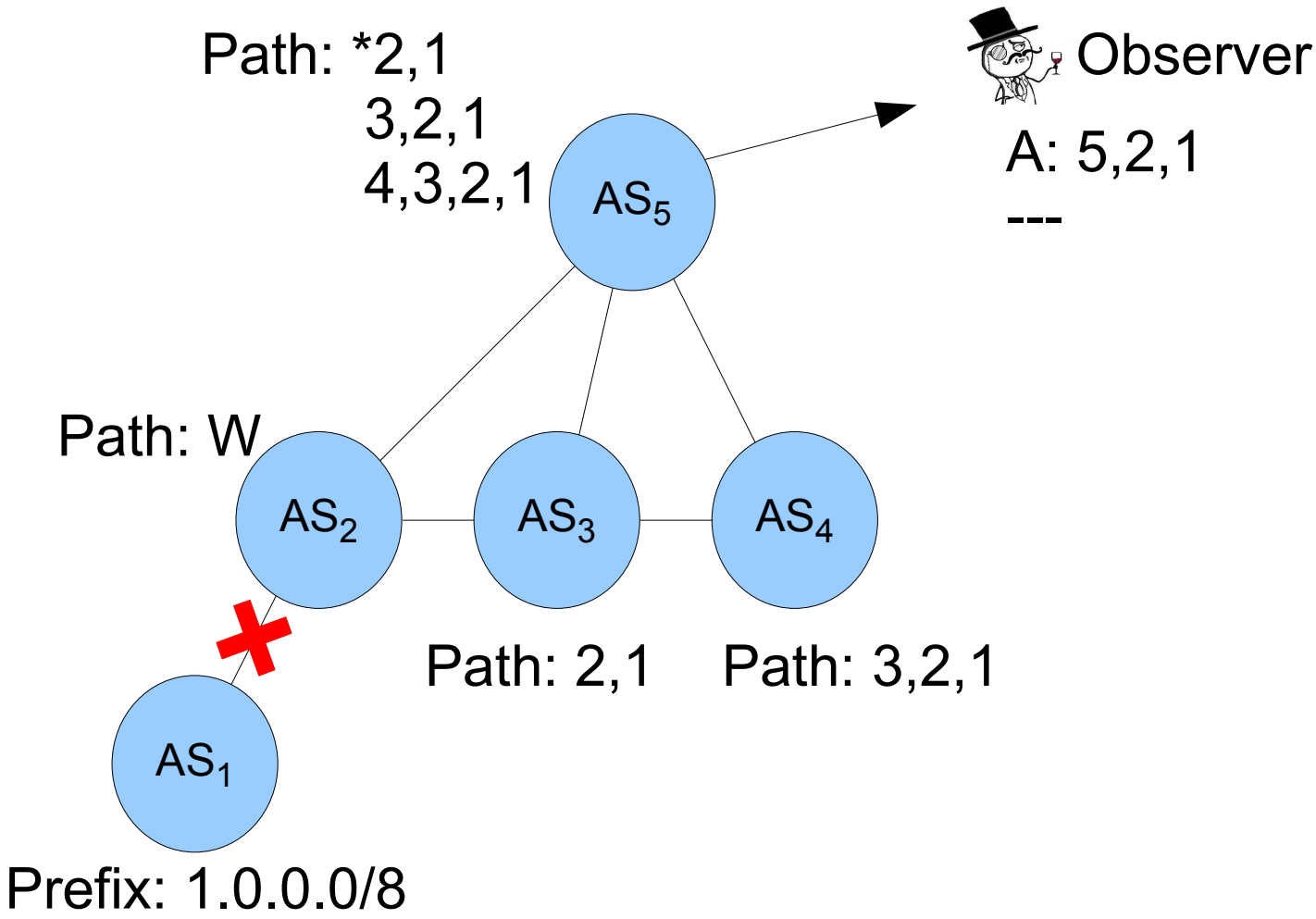
Minimum Route Advertisment Interval – MRAI



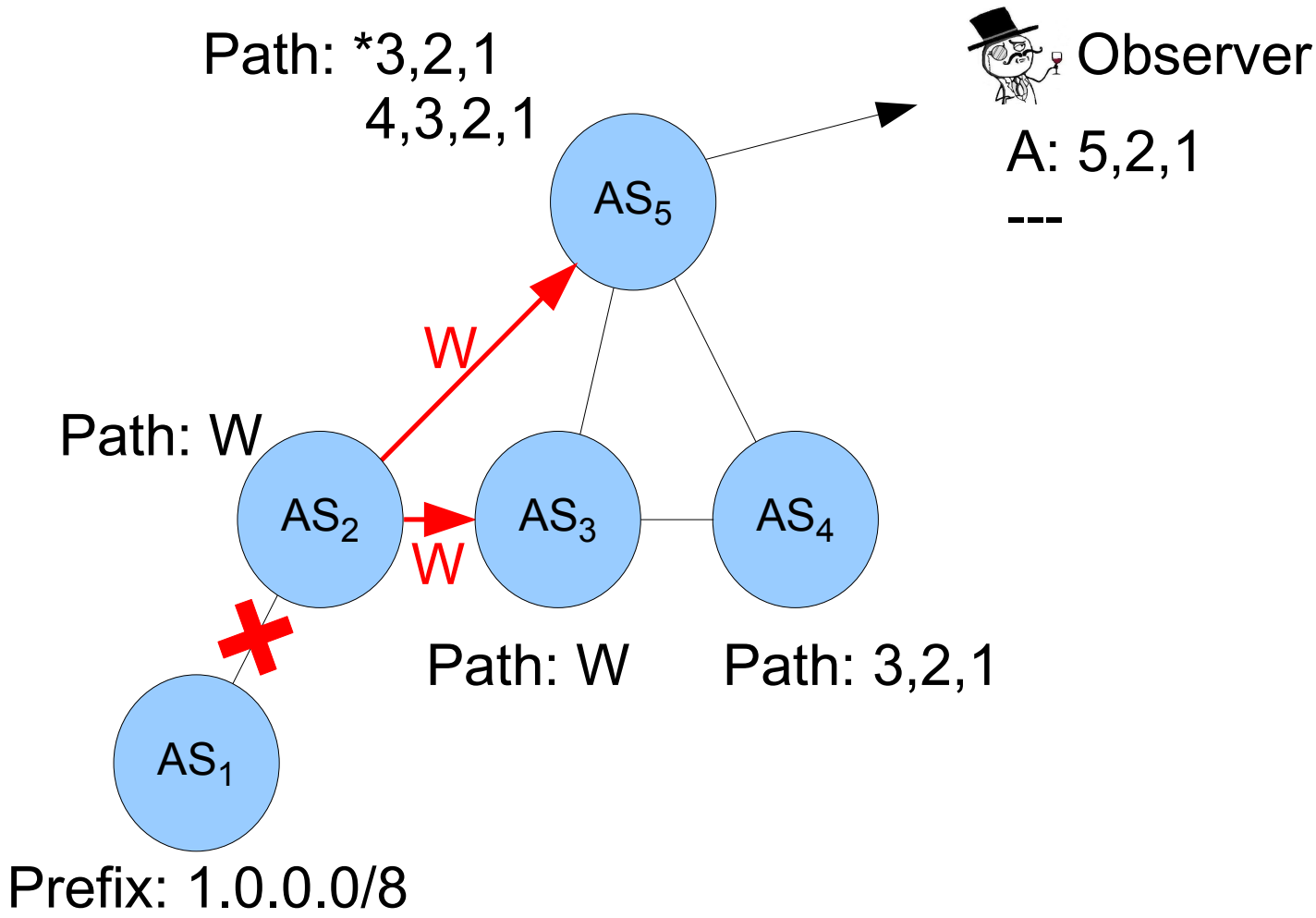
Minimum Route Advertisment Interval – MRAI



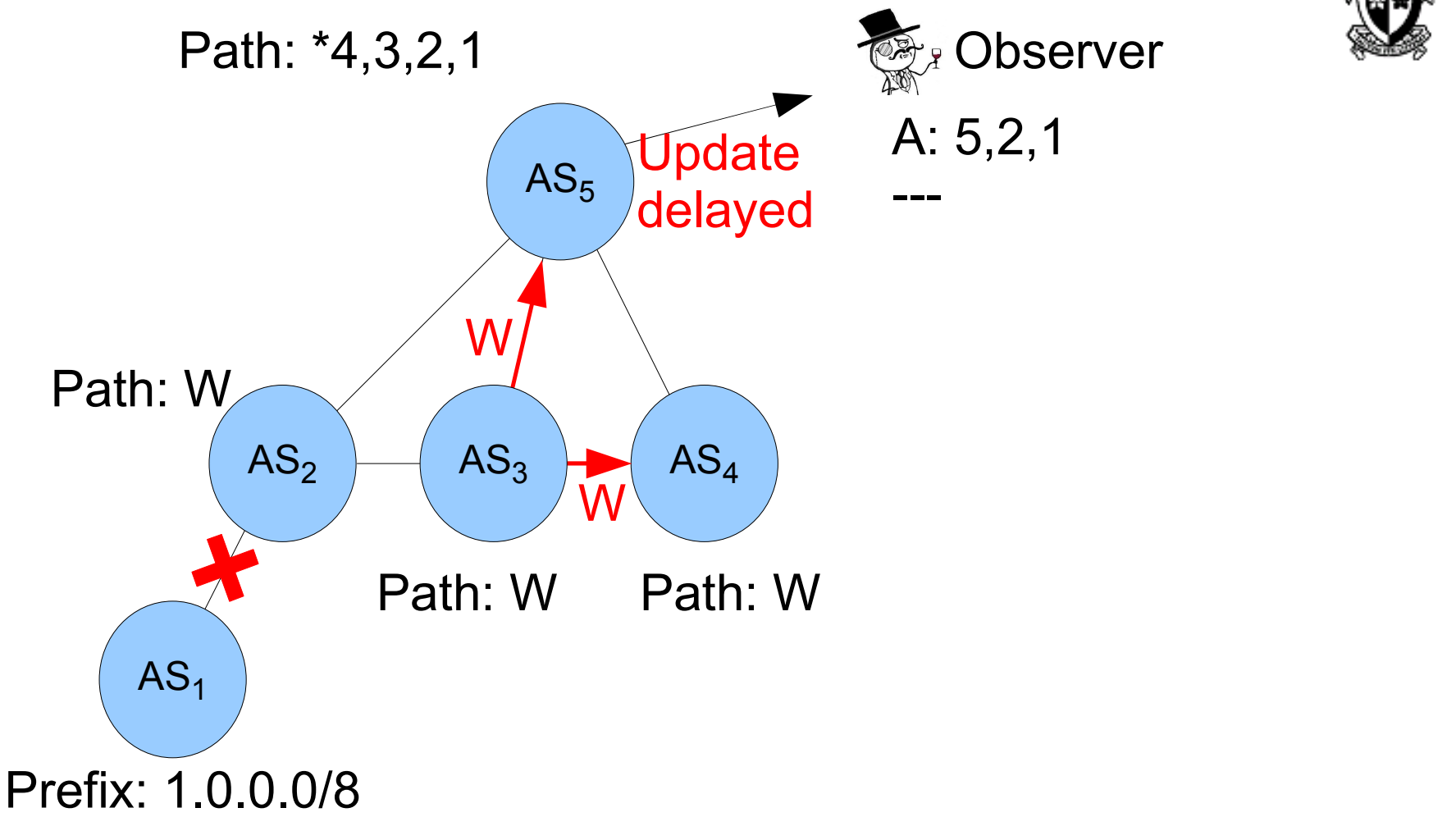
Minimum Route Advertisment Interval – MRAI



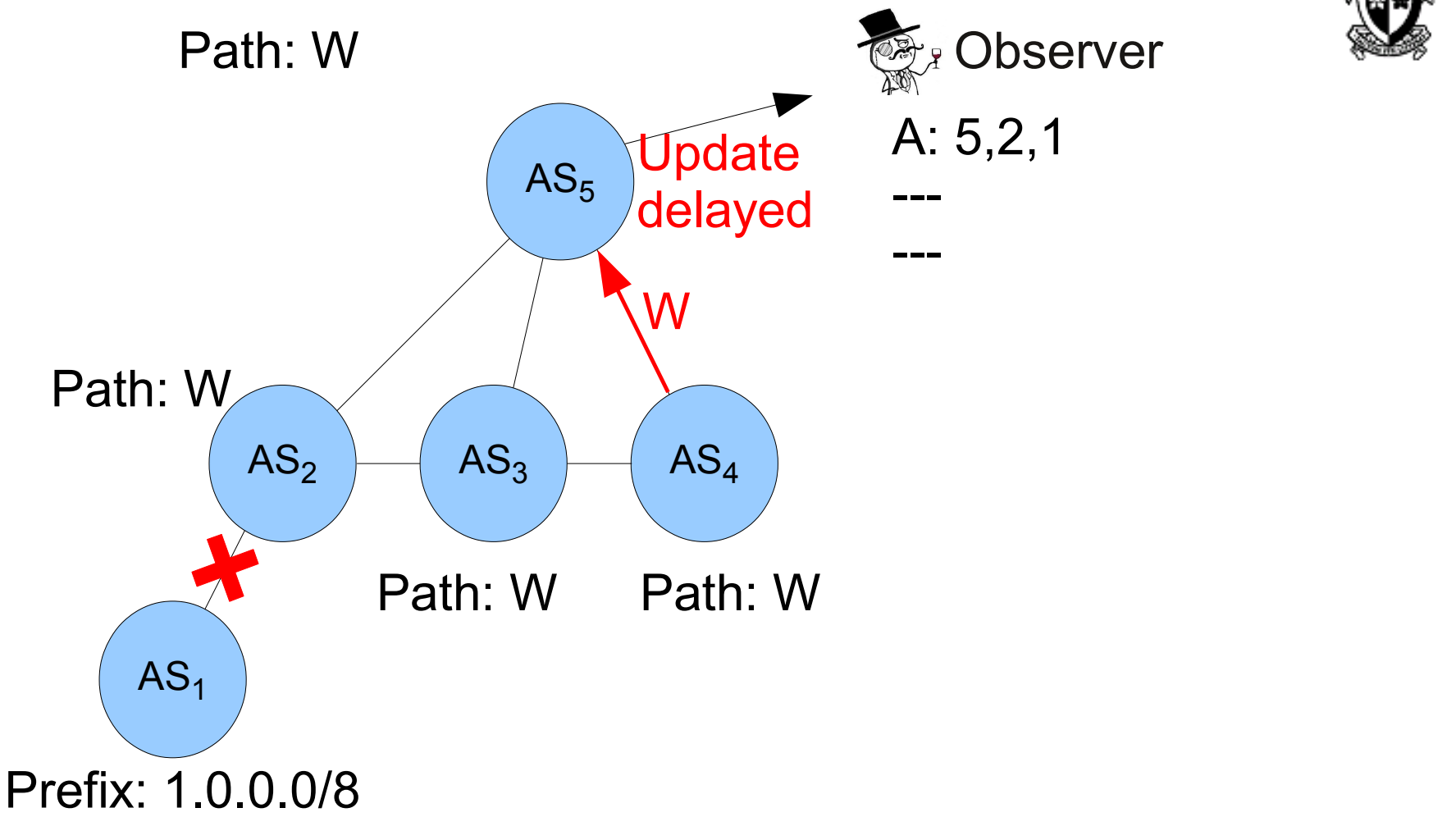
Minimum Route Advertisment Interval – MRAI



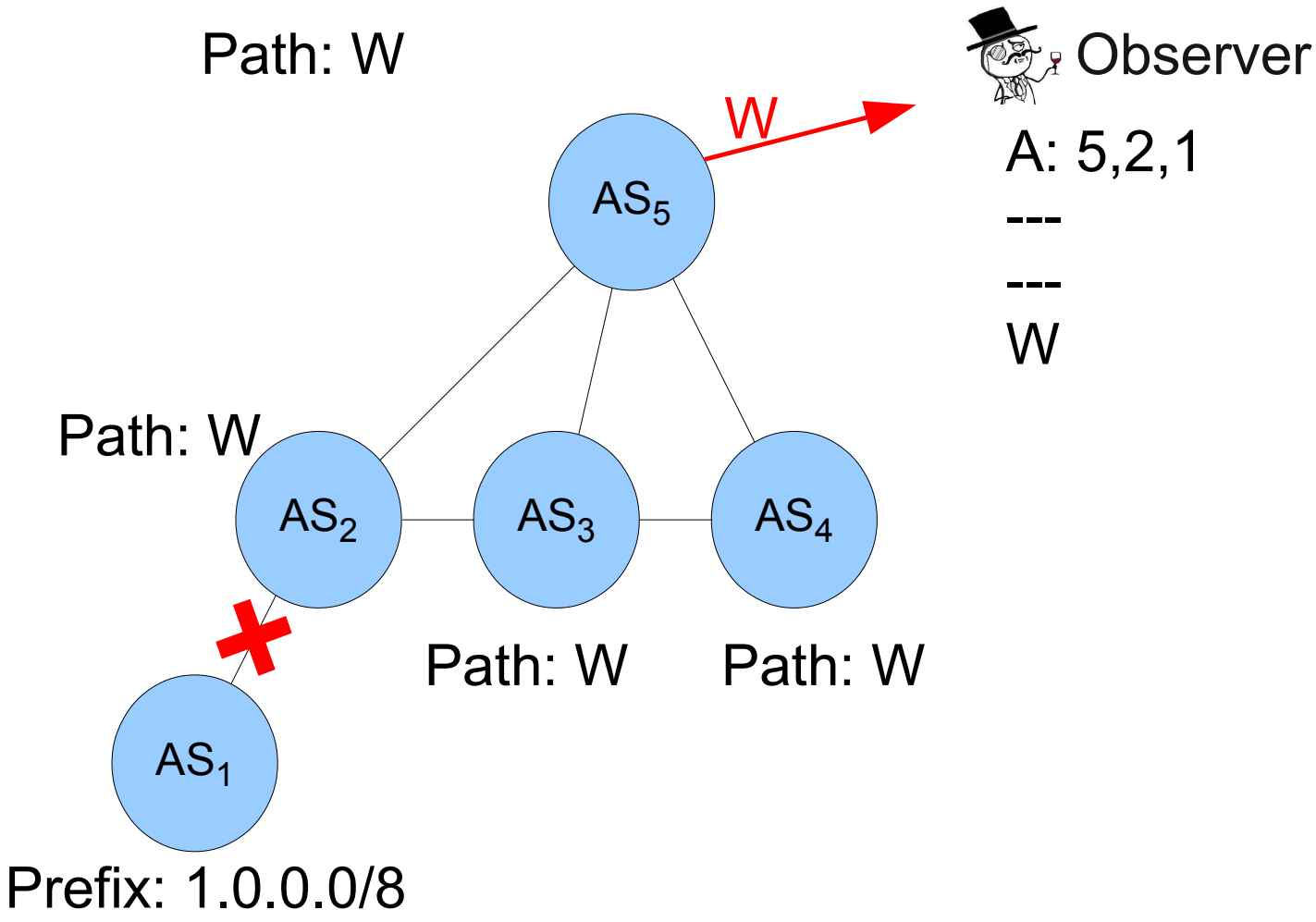
Minimum Route Advertisment Interval – MRAI



Minimum Route Advertisment Interval – MRAI



Minimum Route Advertisment Interval – MRAI

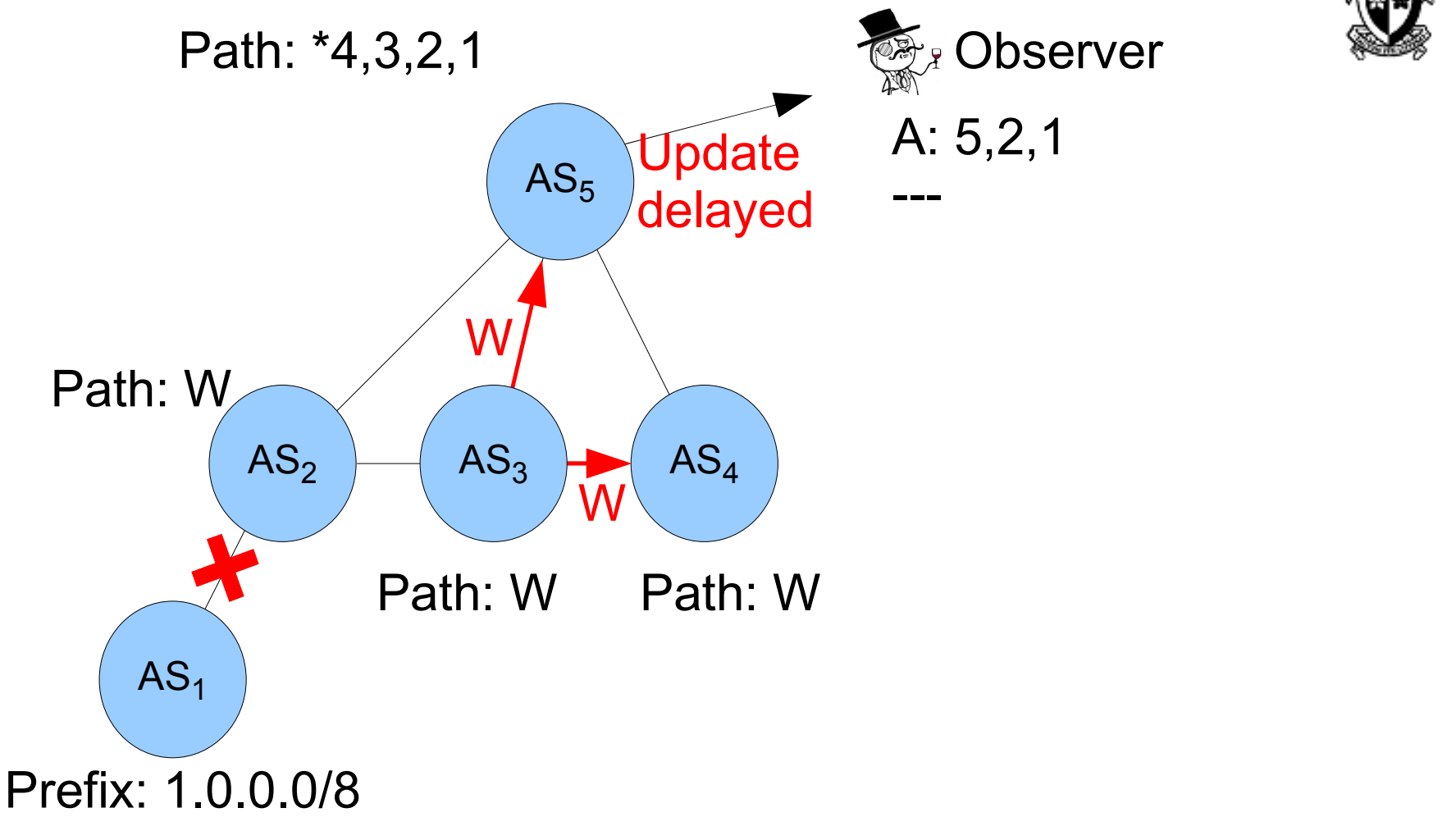


Observer

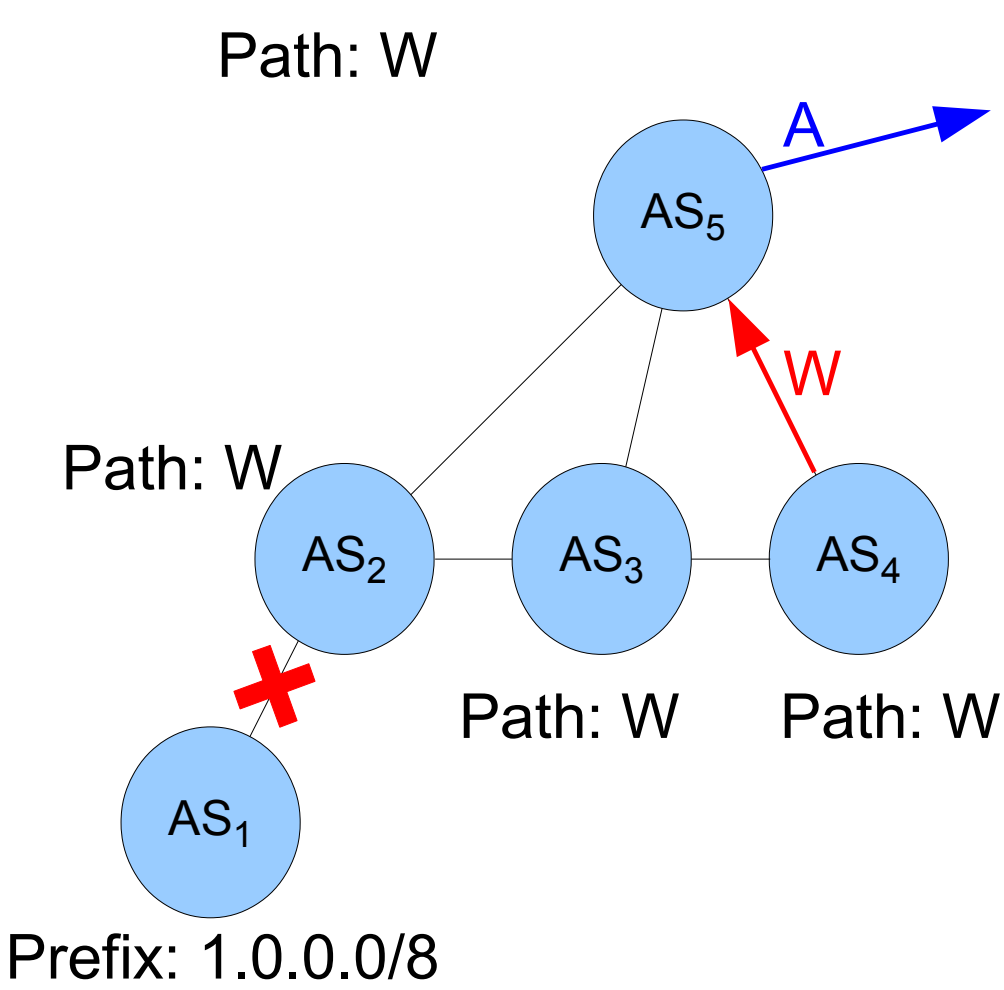
A: 5,2,1

W

Minimum Route Advertisment Interval – MRAI



Minimum Route Advertisment Interval – MRAI

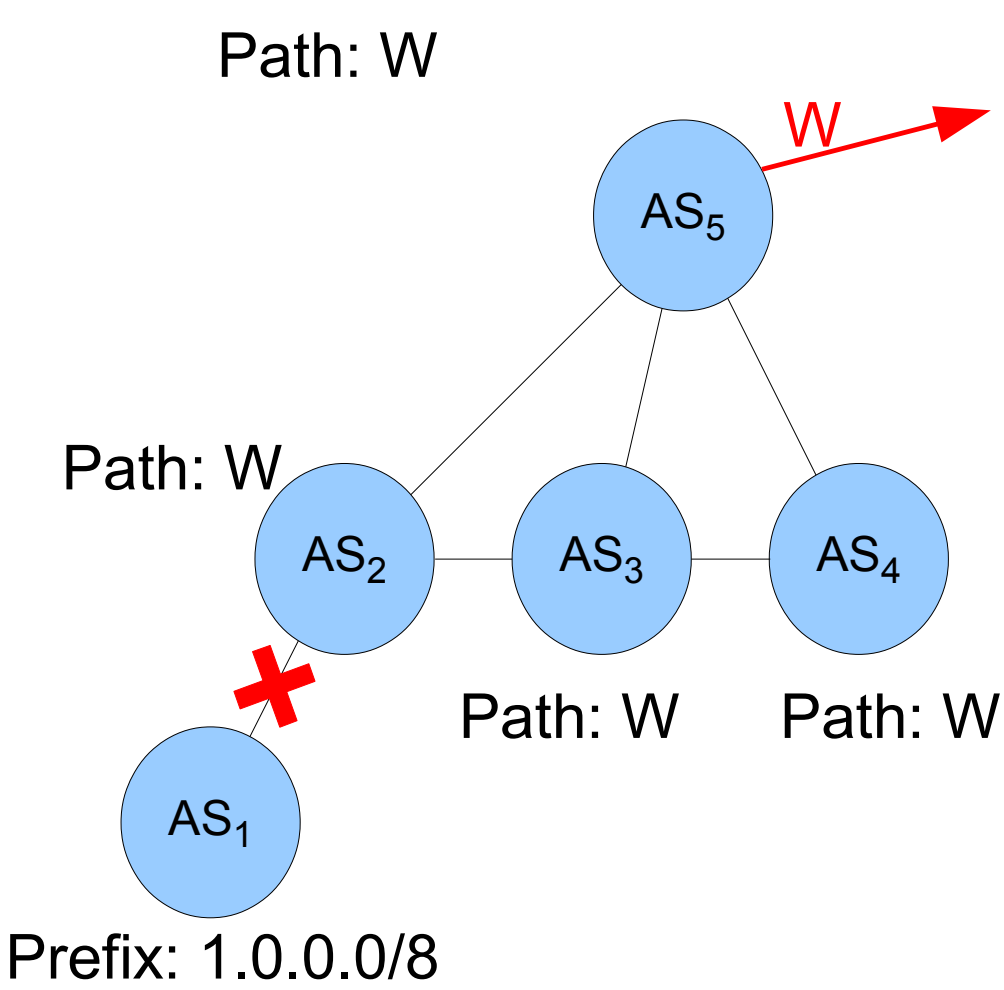


Observer

A: 5,2,1

A: 5,4,3,2,1

Minimum Route Advertisment Interval – MRAI



Observer

A: 5,2,1

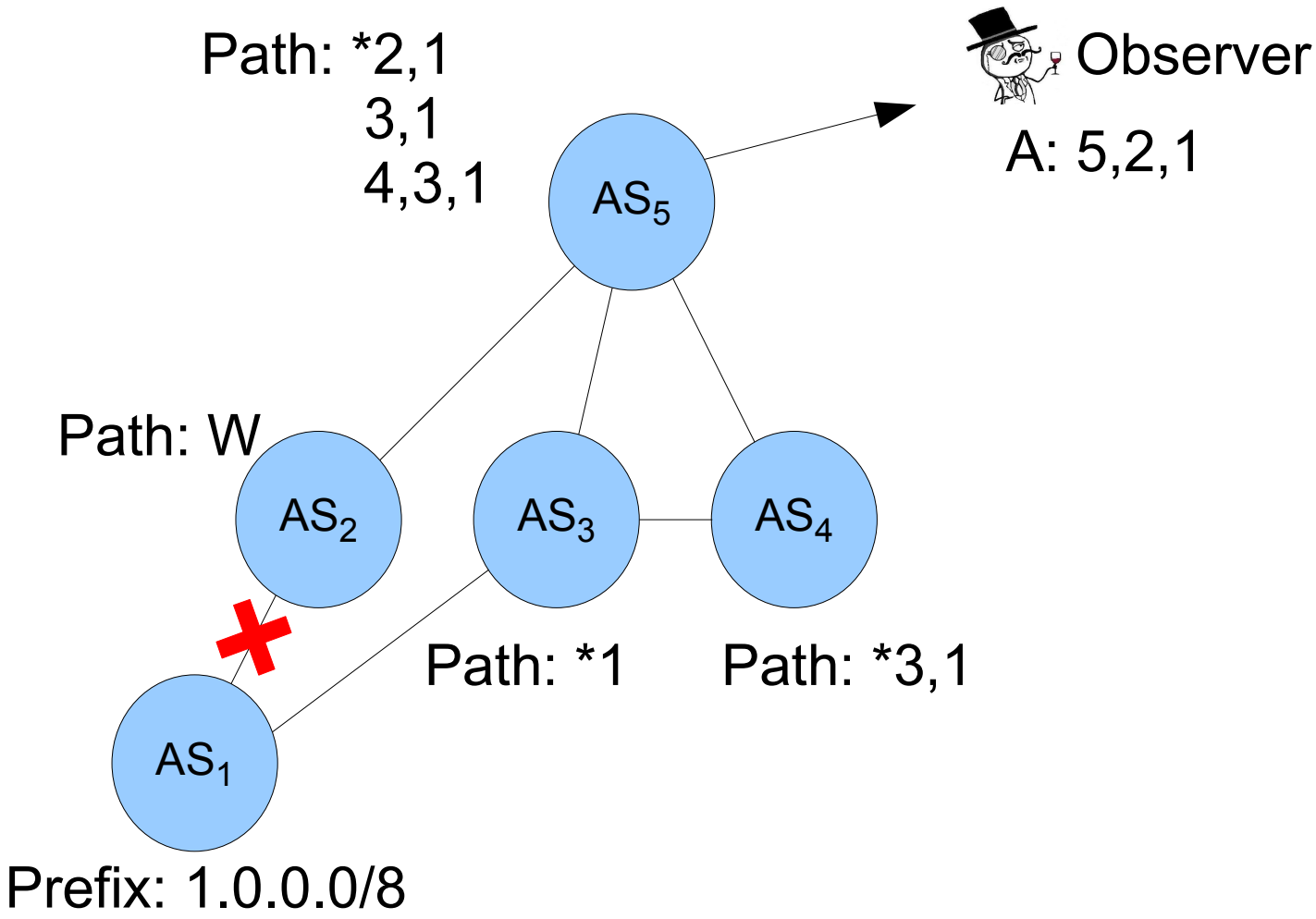
A: 5,4,3,2,1

W

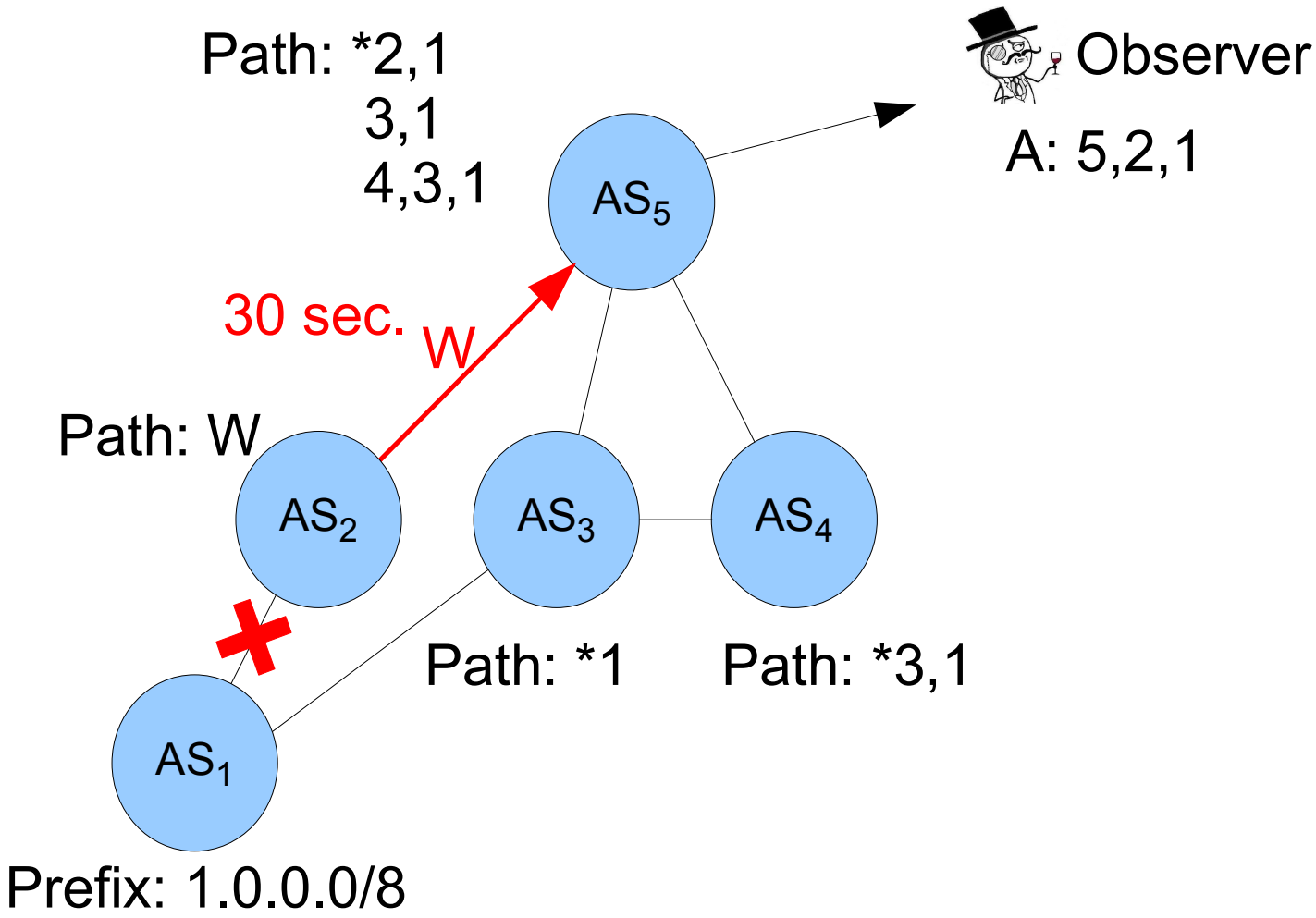
MRAI and RFD

- Need to decrease BGP chattiness and Path Exploration
- Minimum Route Advertisment Interval – MRAI (RFC 1771)
  - Apply 30s (default) delay on announcements
  - MRAI on withdrawals (WRATE) allowed per RFC 4271

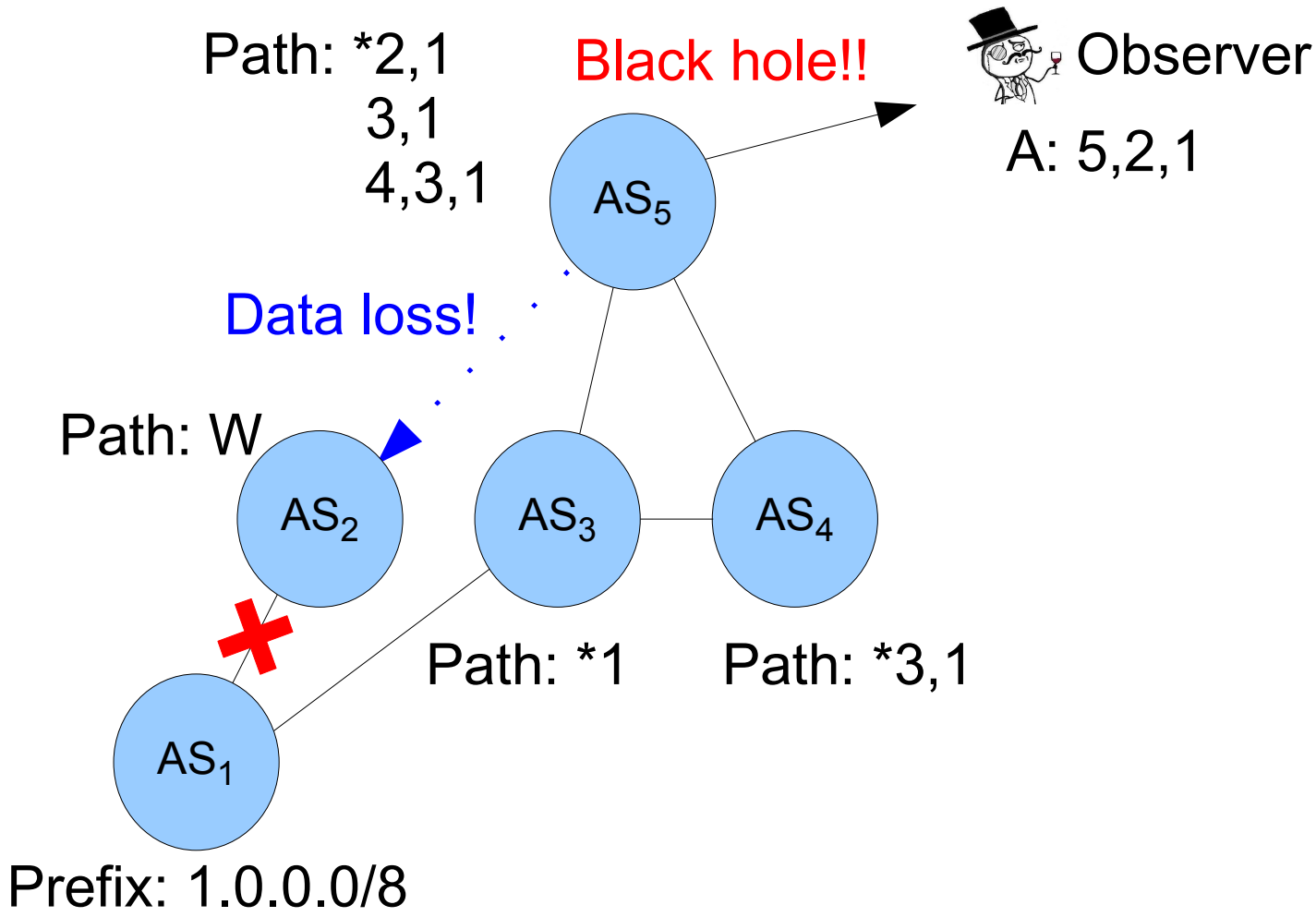
MRAI on withdrawals – WRATE



MRAI on withdrawals – WRATE



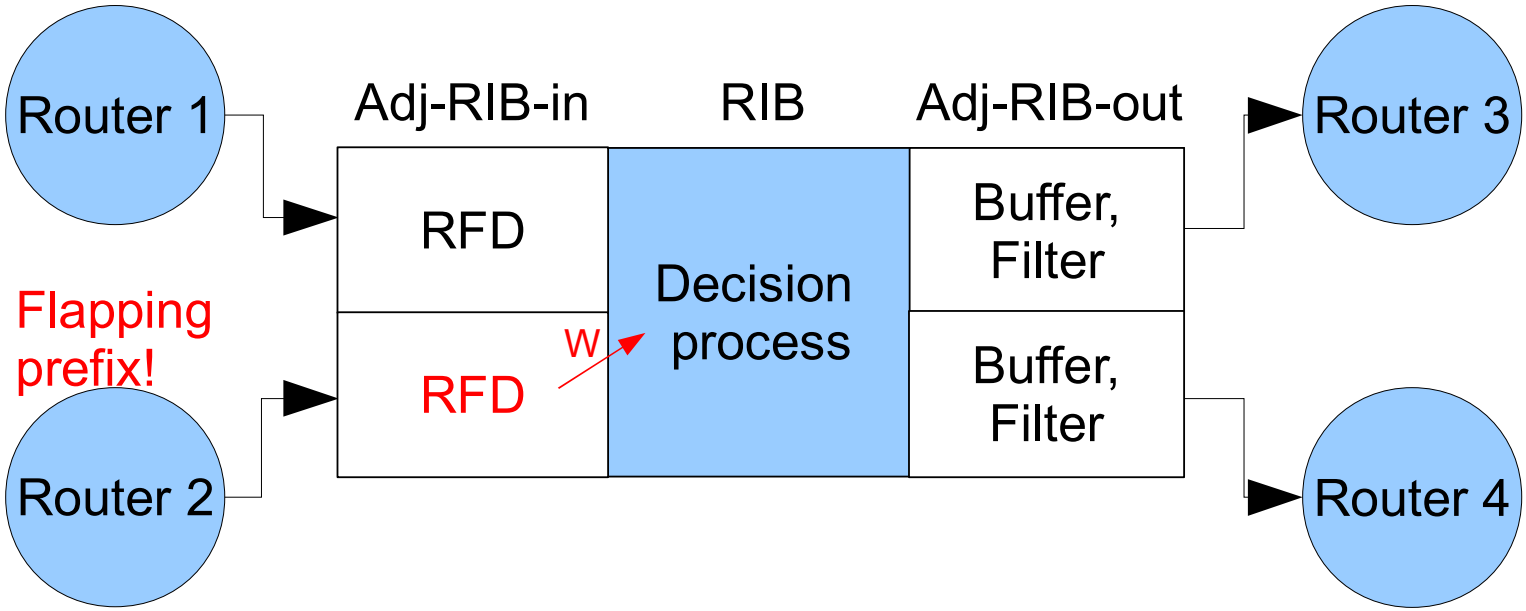
MRAI on withdrawals – WRATE



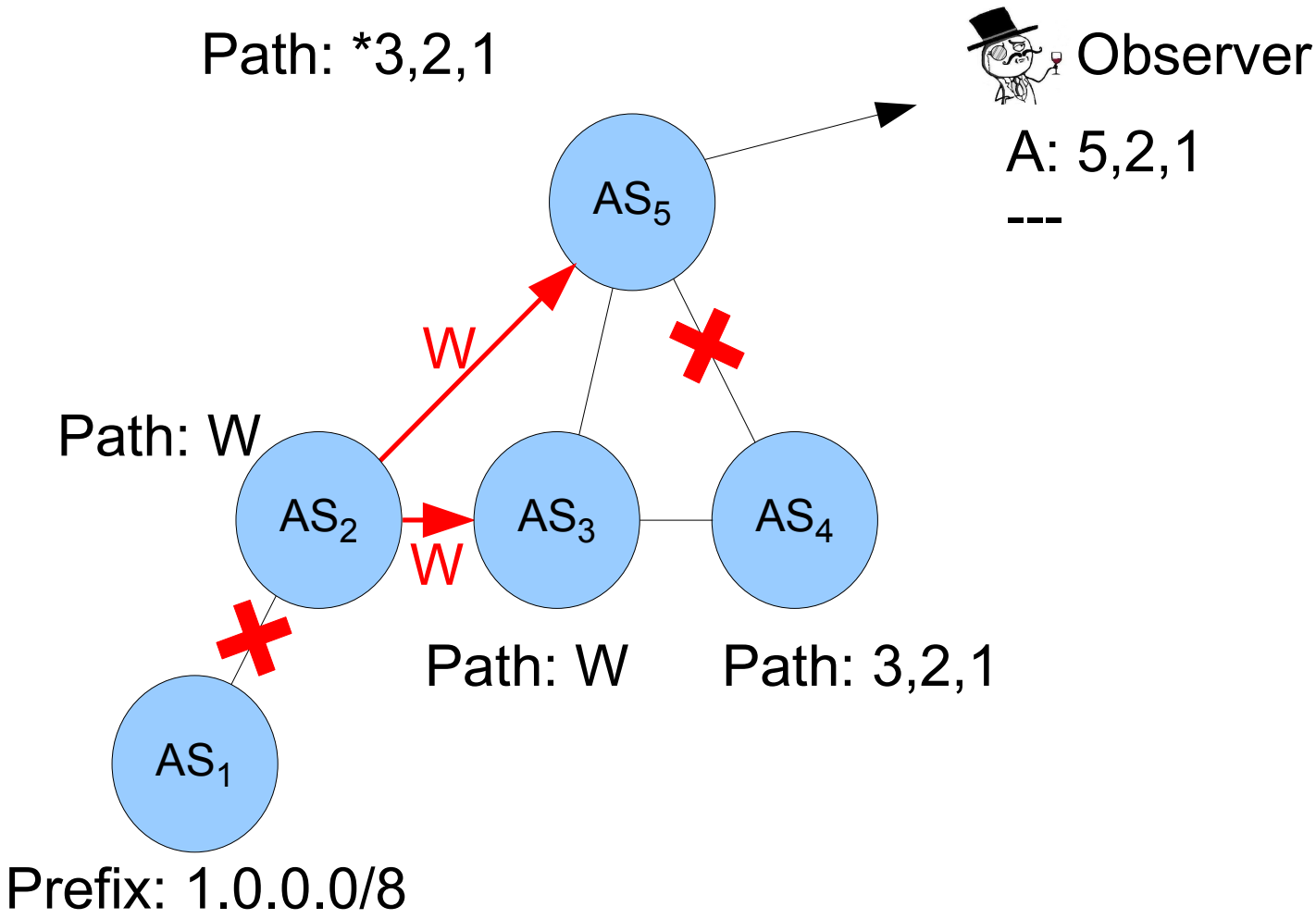
MRAI and RFD

- Need to decrease BGP chattiness and Path Exploration
- Minimum Route Advertisment Interval – MRAI (RFC 1771)
  - Apply 30s (default) delay on announcements
  - MRAI on withdrawals (WRATE) allowed per RFC 4271
- Route Flap Damping – RFD (RFC 2439)
  - Flapping = sequence of announcements and withdrawals
  - Suppress flapping prefixes for 1 hour (or more)

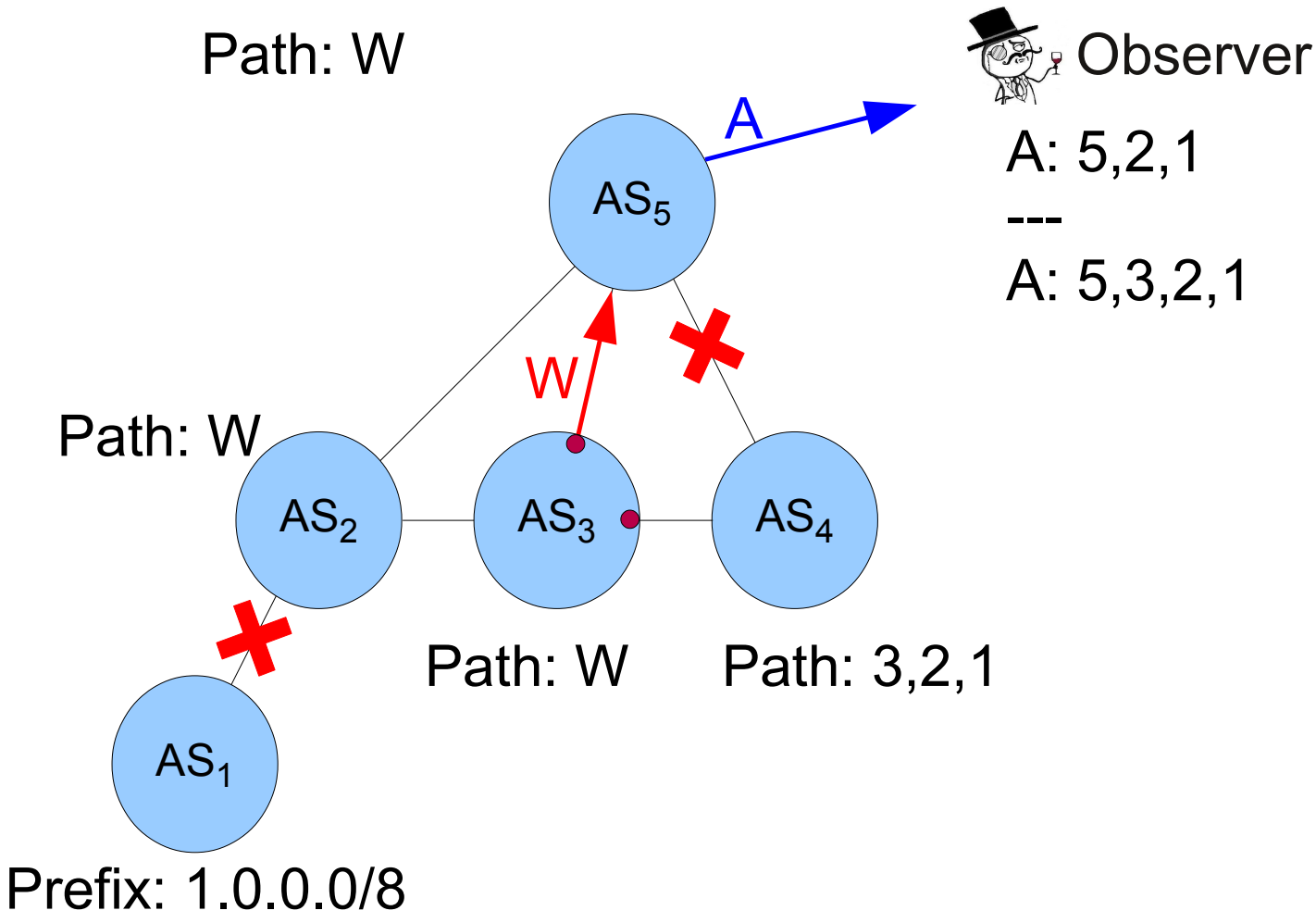
Route Flap Damping – RFD



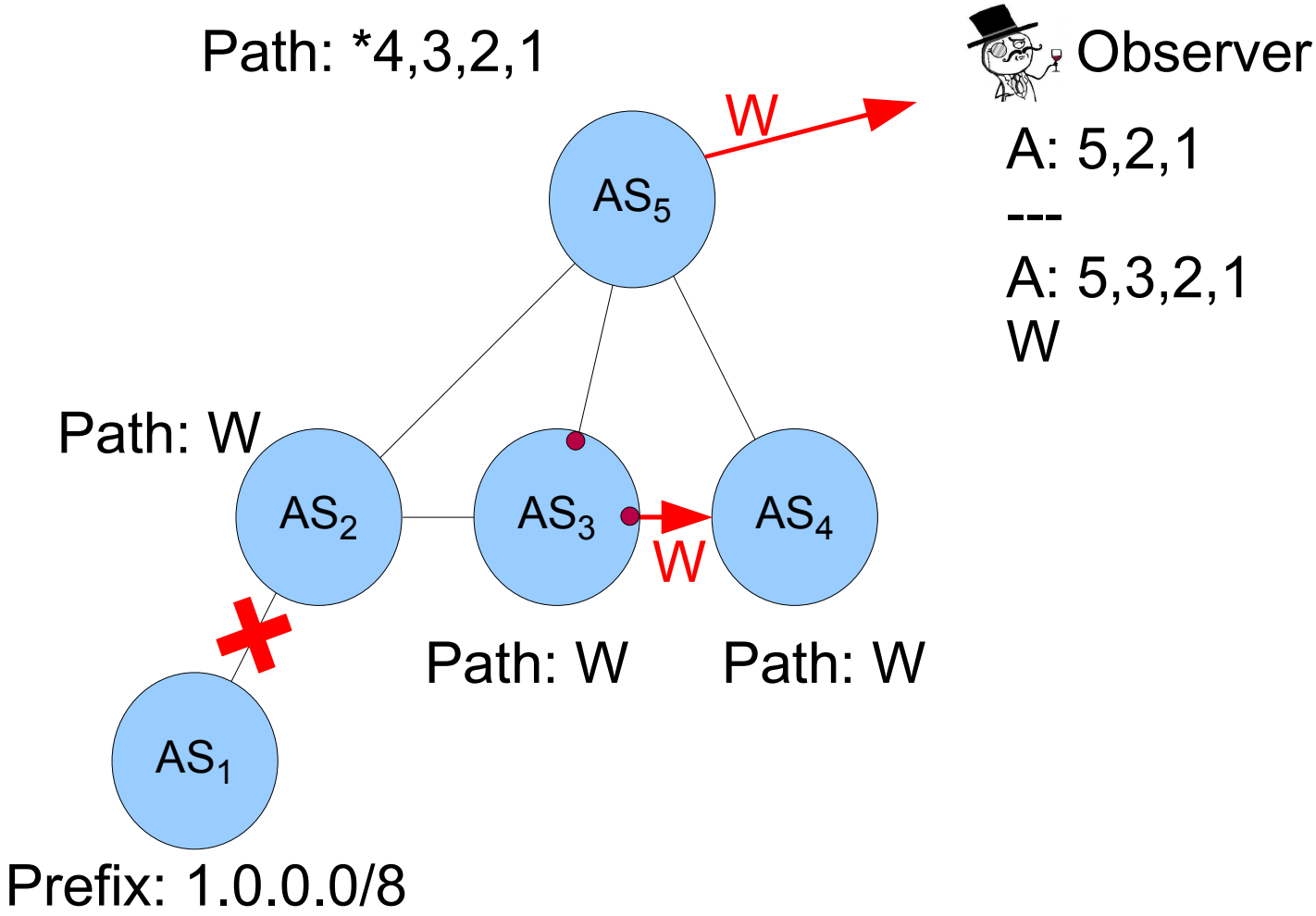
Route Flap Damping – RFD



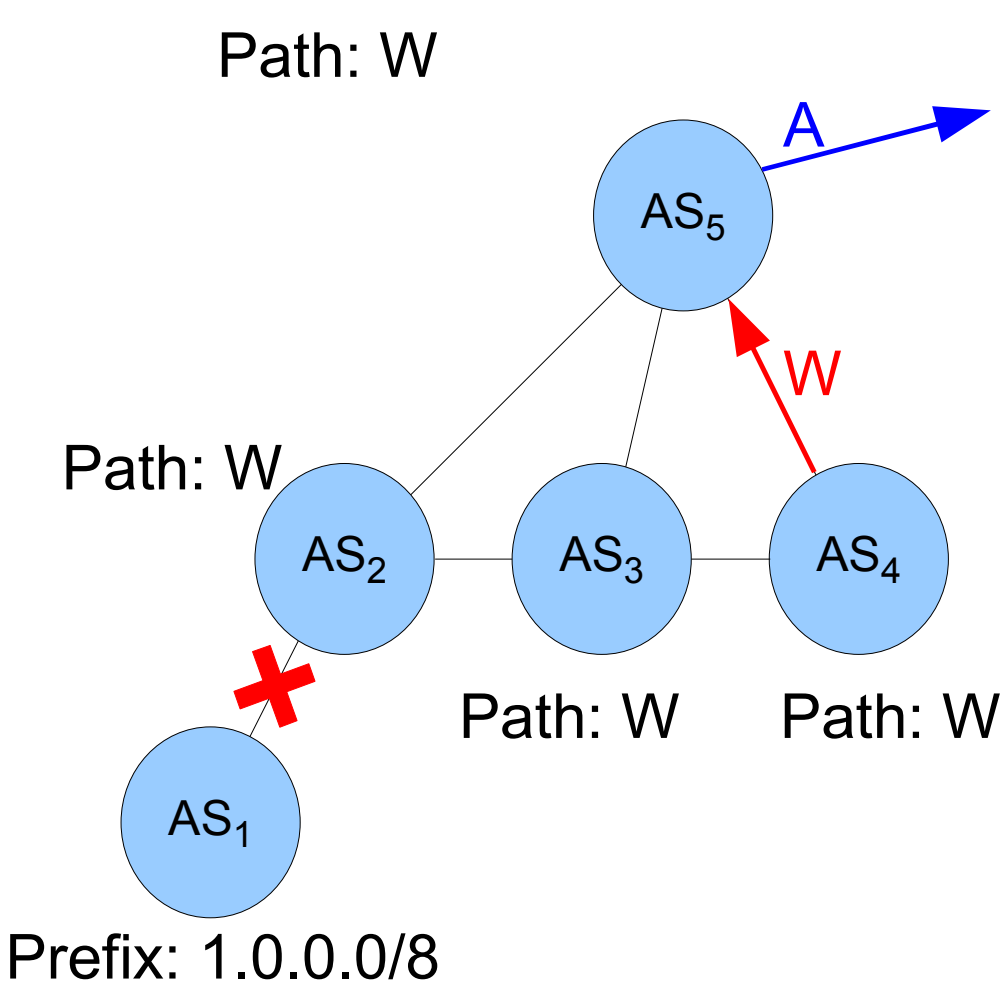
Route Flap Damping – RFD



Route Flap Damping – RFD



Route Flap Damping – RFD



Observer

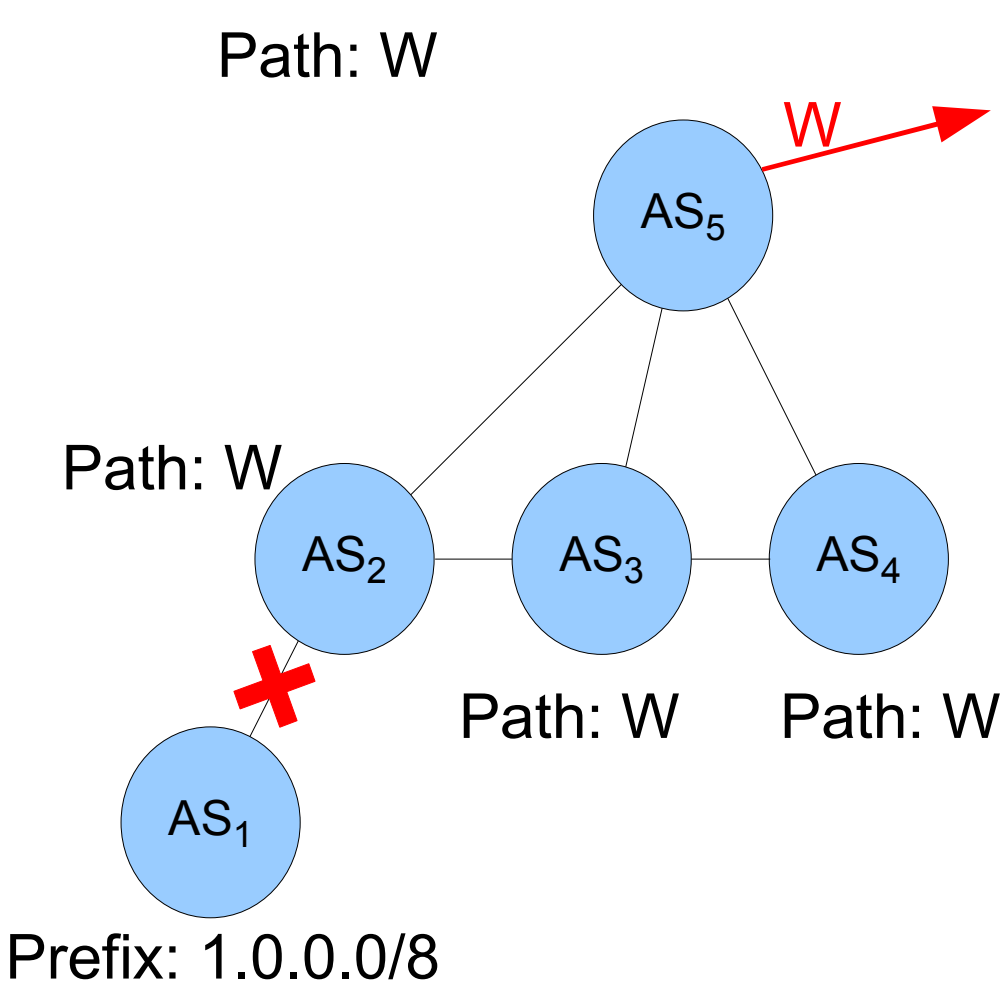
A: 5,2,1

A: 5,3,2,1

W

A: 5,4,3,2,1

Route Flap Damping – RFD



Observer

A: 5,2,1

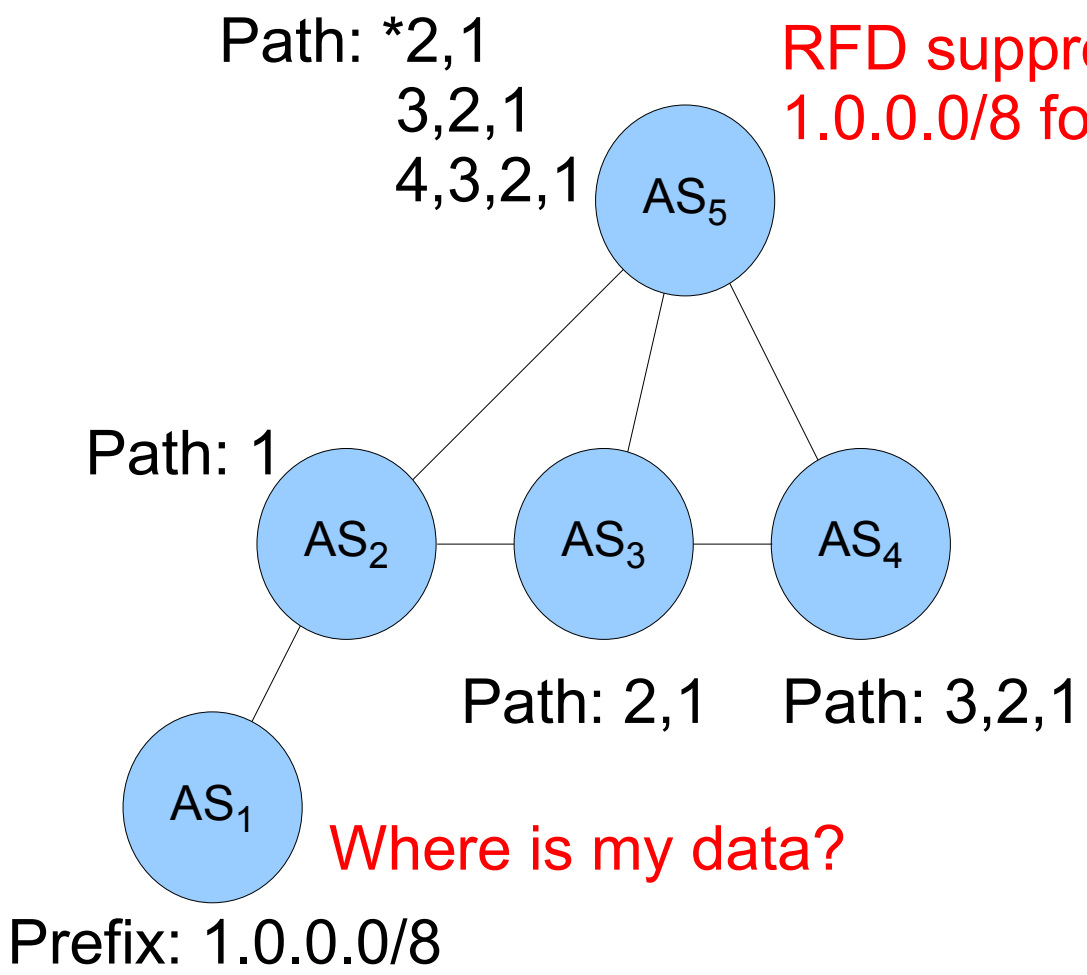
A: 5,3,2,1

W

A: 5,4,3,2,1

W

Route Flap Damping – RFD



What are the problems?

- MRAI exhibits unpredictable behavior
- WRATE creates black holes
- RFD penalizes prefix owners even if the misbehavior happened further upstream
- We want something better!

Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results

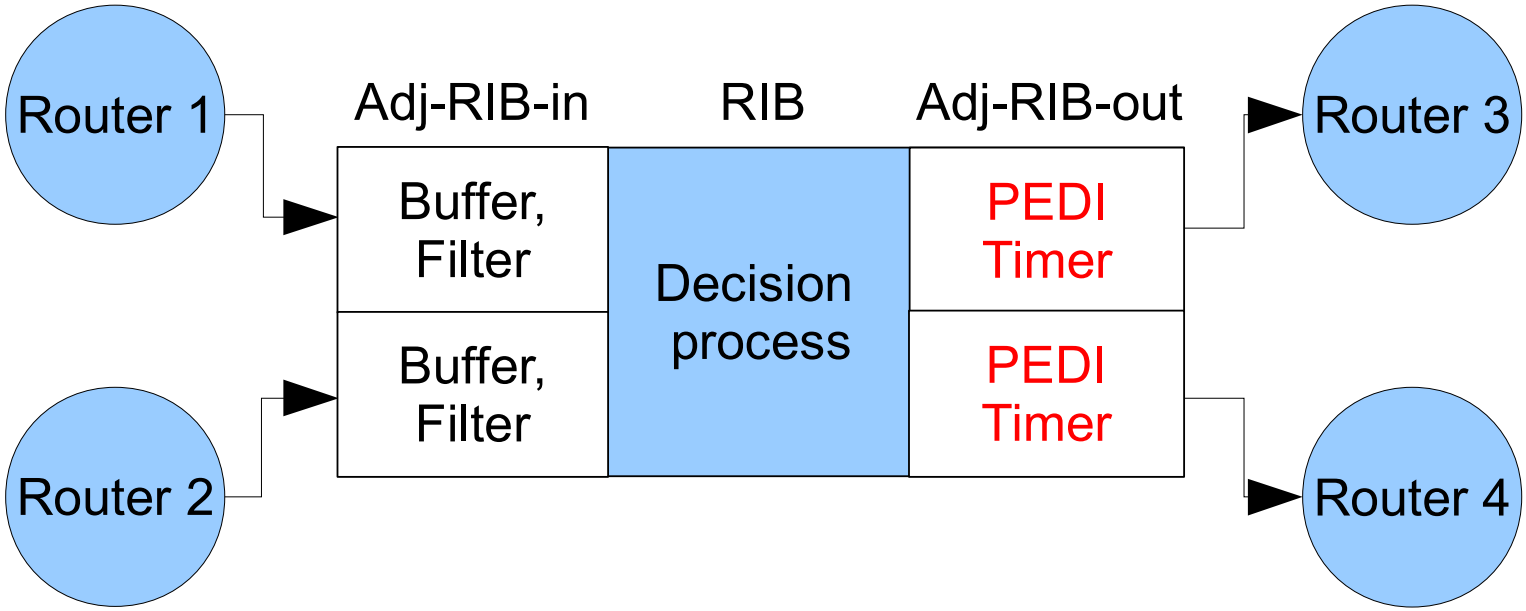
- Reduction of update load
- MRAI and PED convergence time compared

Conclusions and future work

PED algorithm

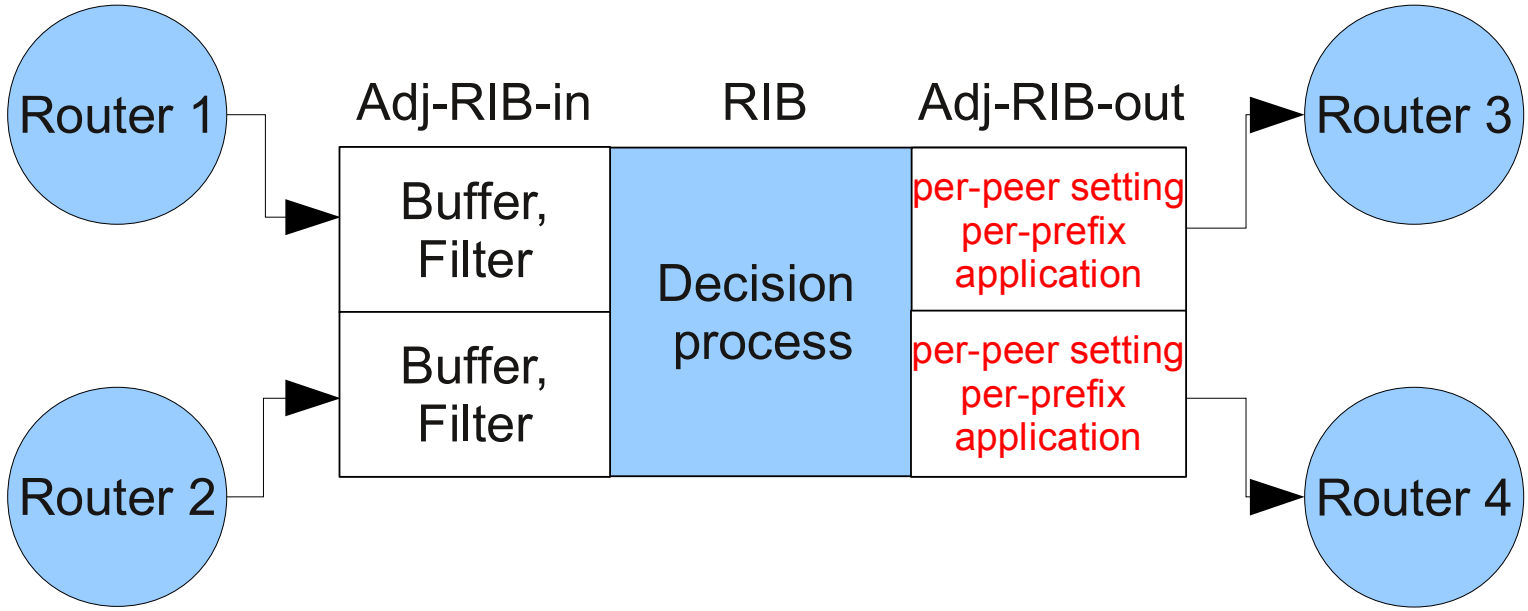
- Delay BGP announcements if the announced AS path is longer than the previously known AS path

PED algorithm illustrated

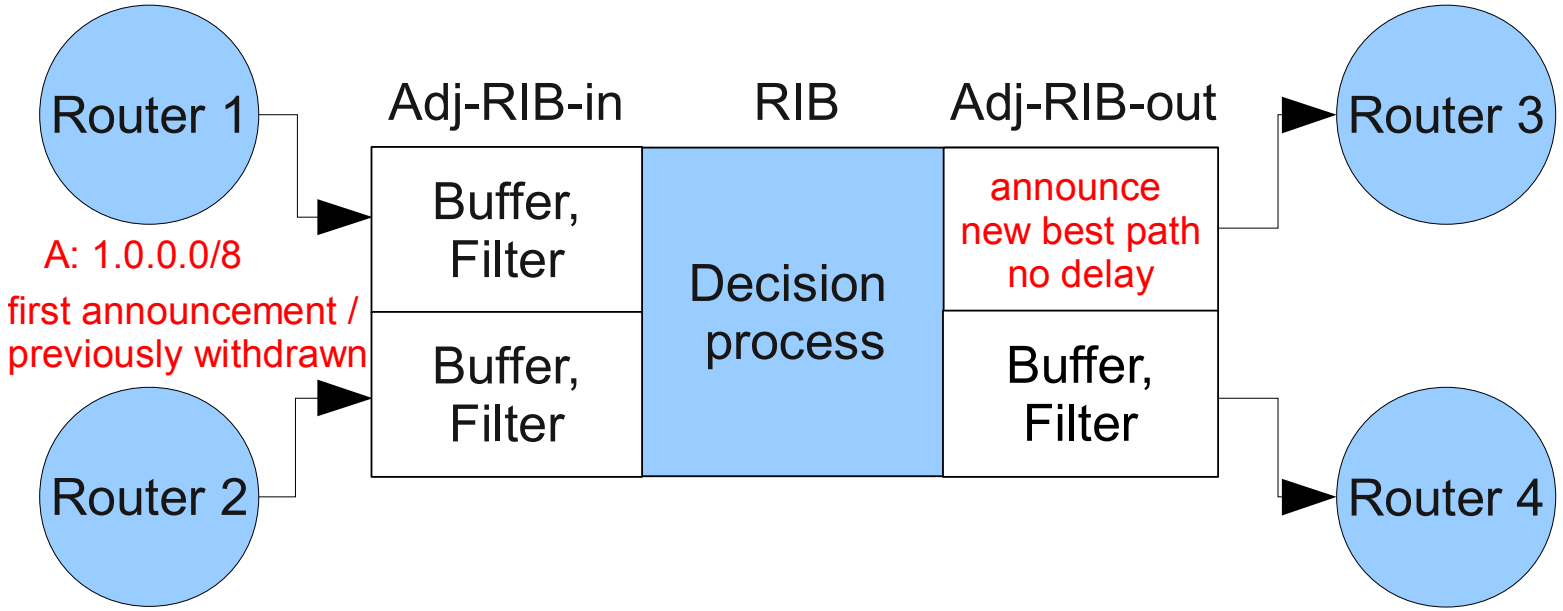


PEDI: Path Exploration Damping Interval

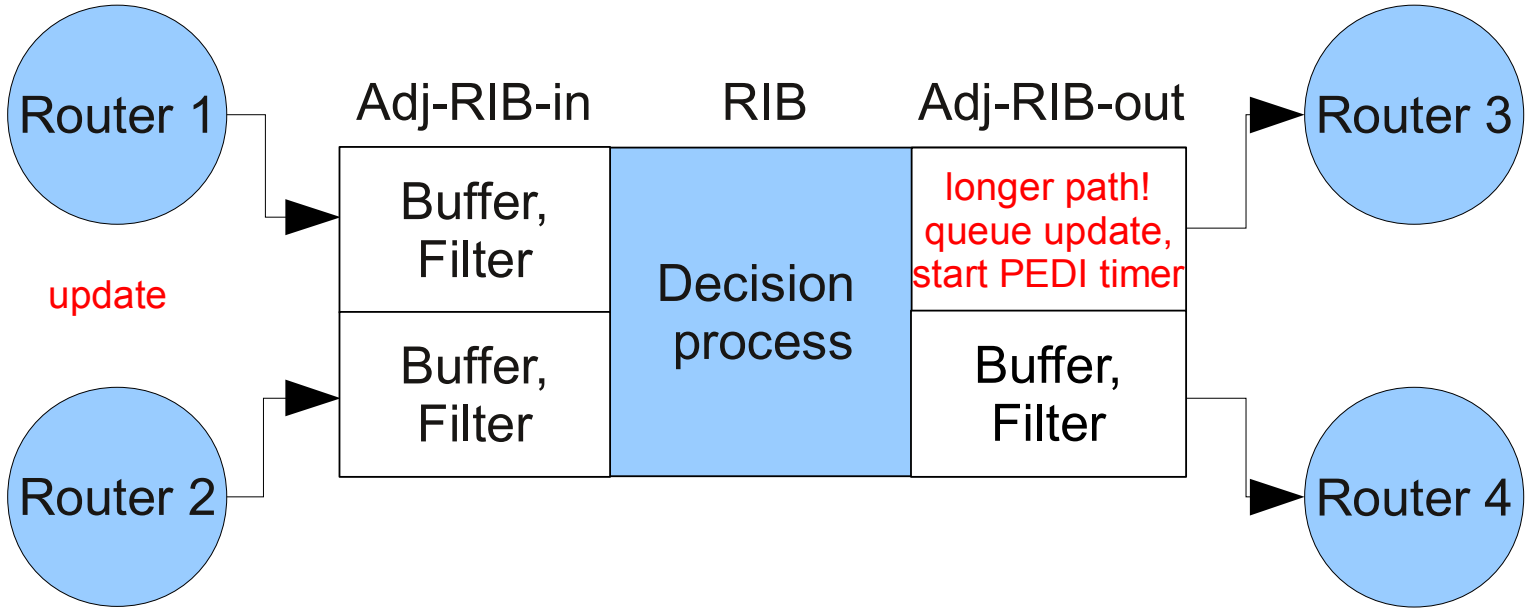
PED algorithm illustrated



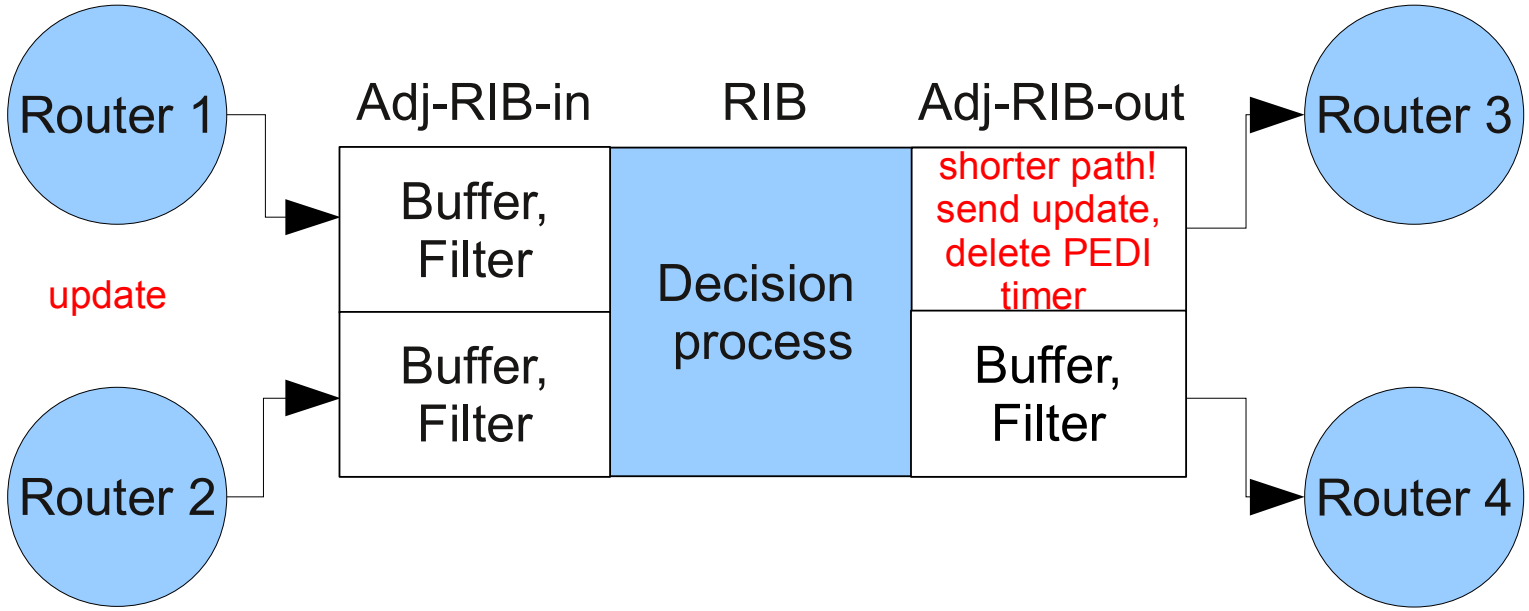
PED algorithm illustrated



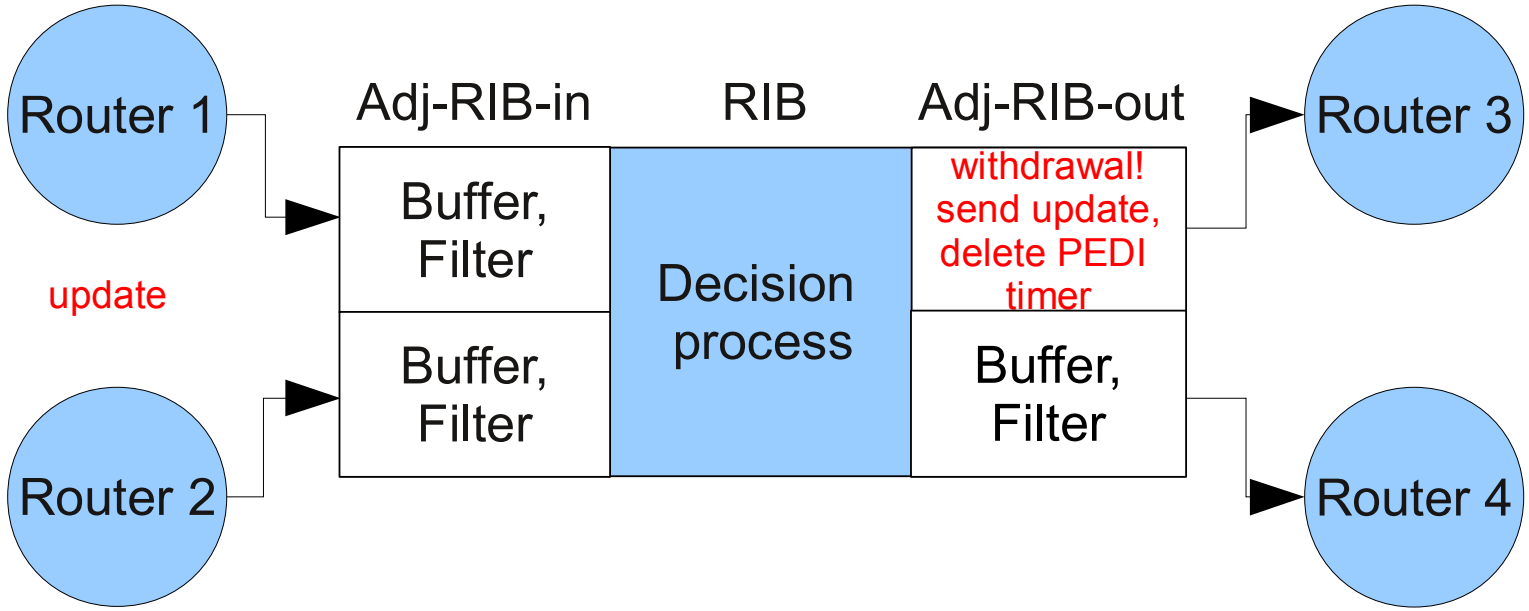
PED algorithm illustrated



PED algorithm illustrated



PED algorithm illustrated



Outline

Introduction

Motivation

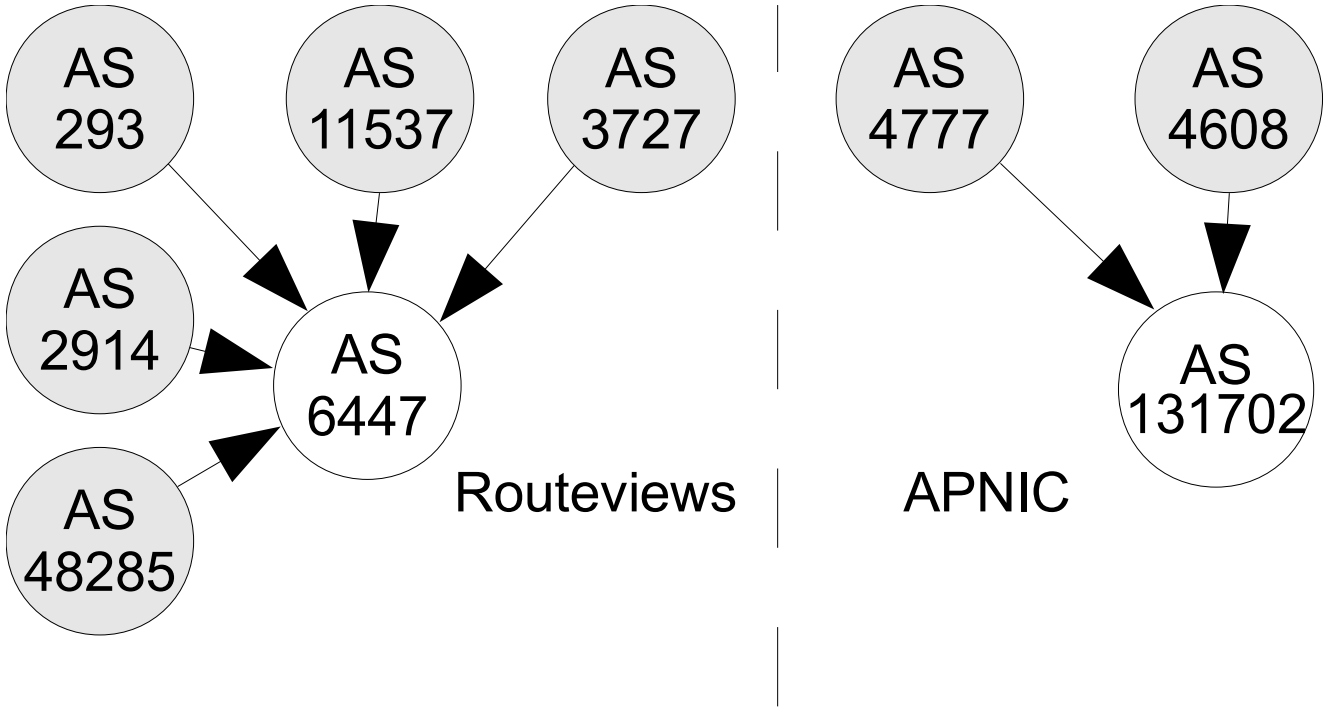
Path Exploration Damping - PED

Experimental results
- Reduction of update load
- MRAI and PED convergence time compared

Conclusions and future work

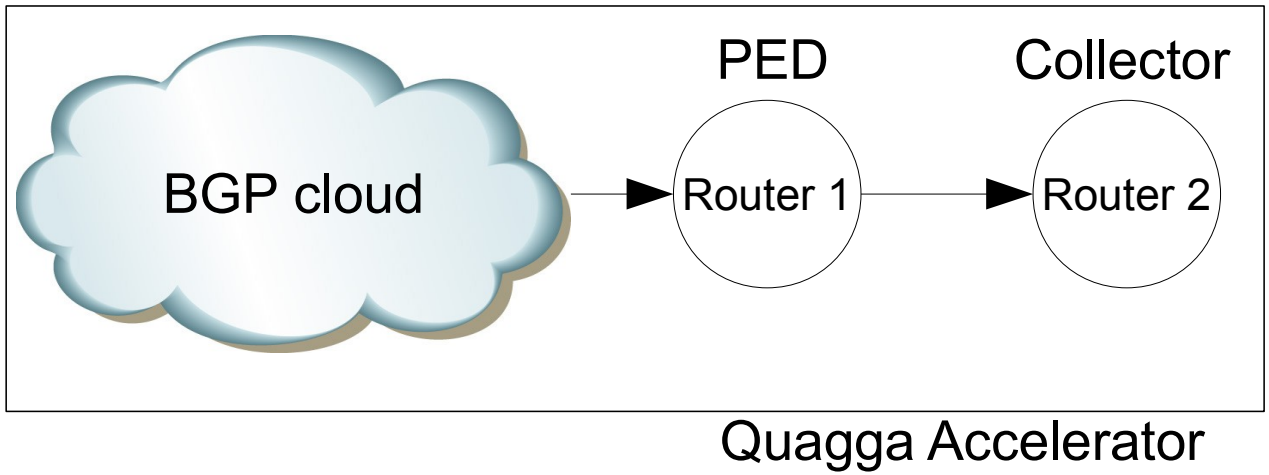
PED – Data analysis

- Experiments using 24 hours of real BGP updates
- Two datasets:
  1. APNIC Pty. Ltd., Australia (2 peers)
  2. University of Oregon – Routeviews, US (5 peers)



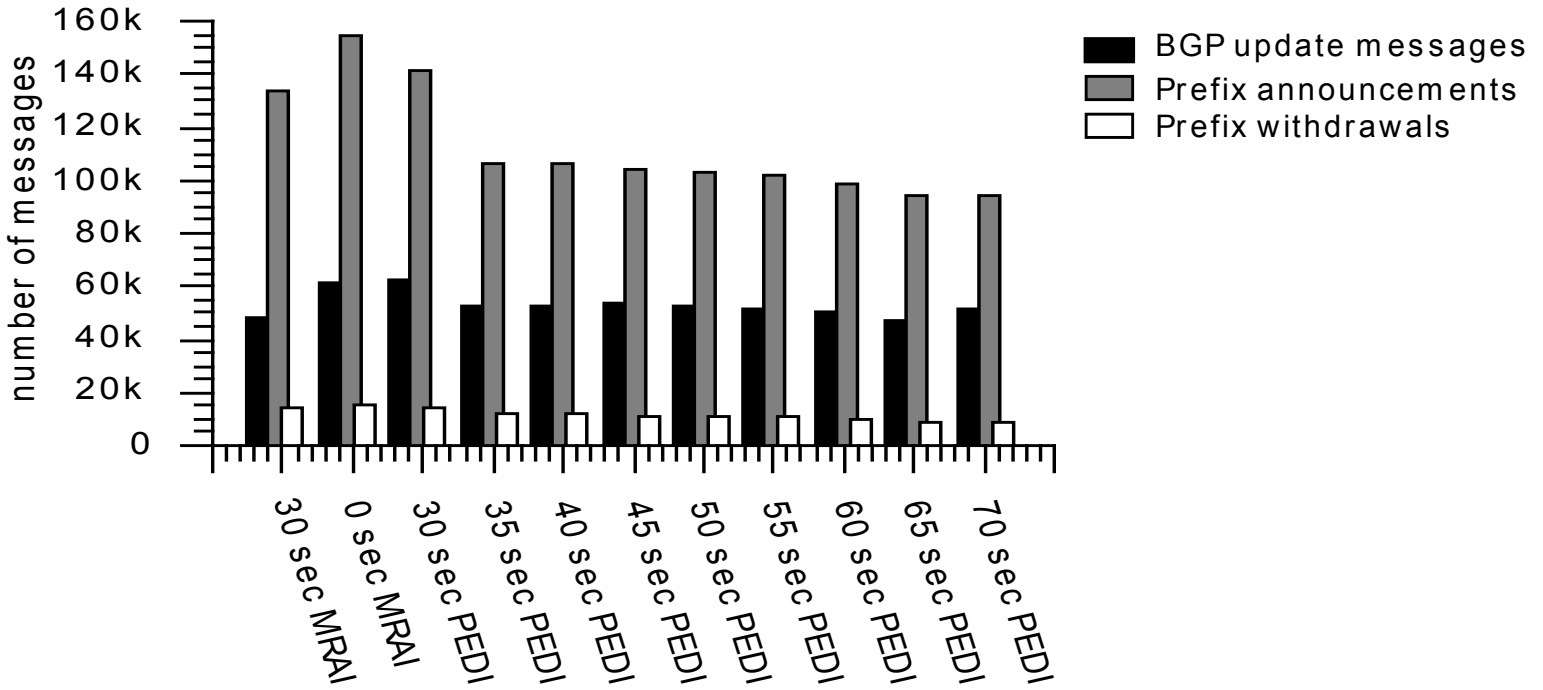
PED – Data analysis

- Replayed using the *Quagga-Accelerator*

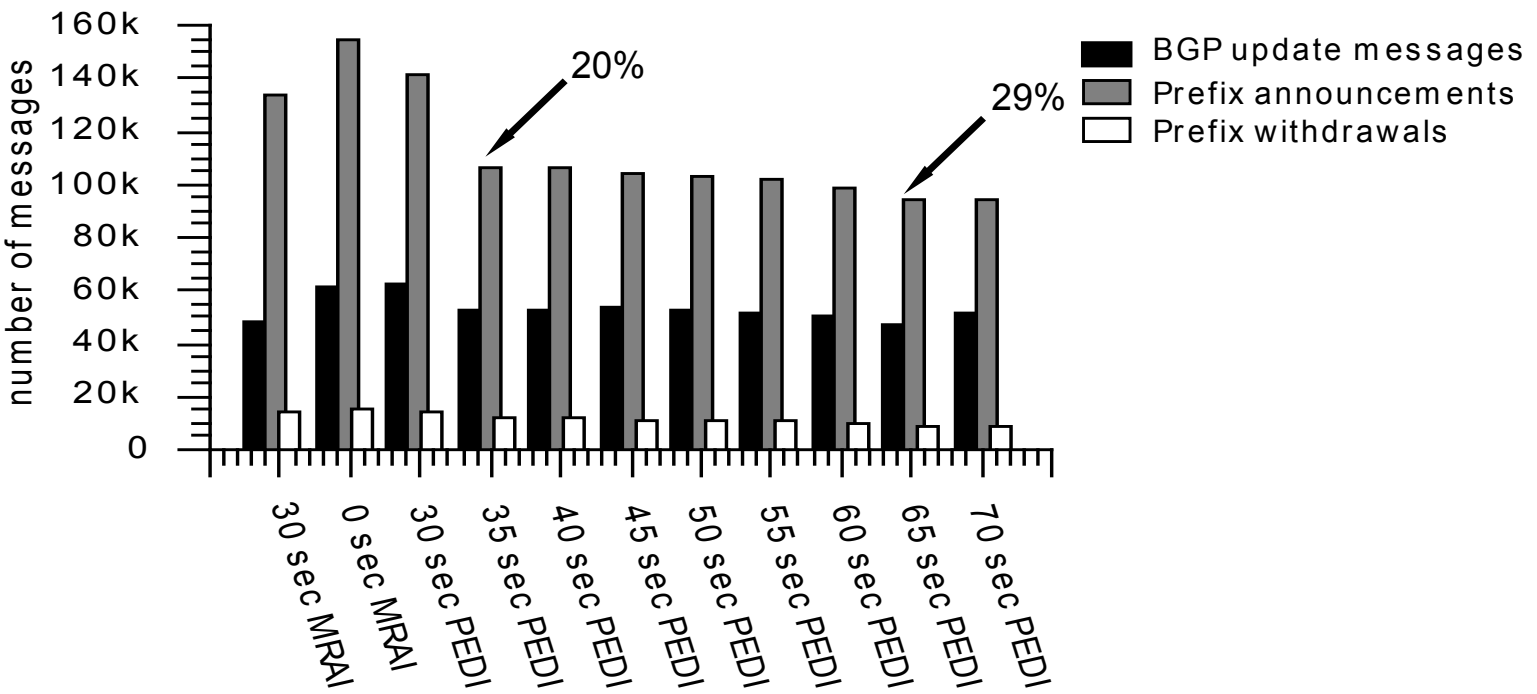


Outline

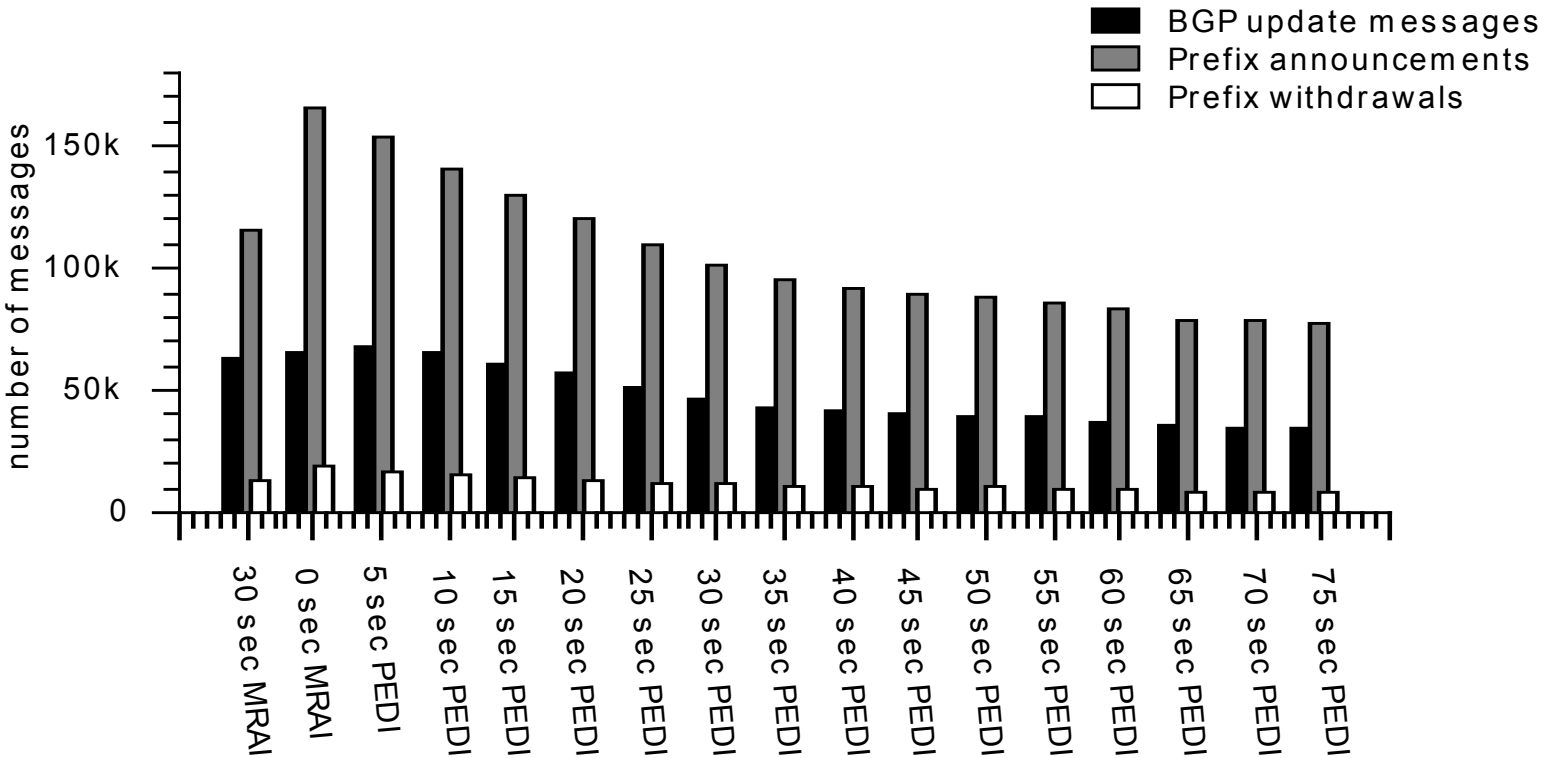
Reduction of update load – APNIC



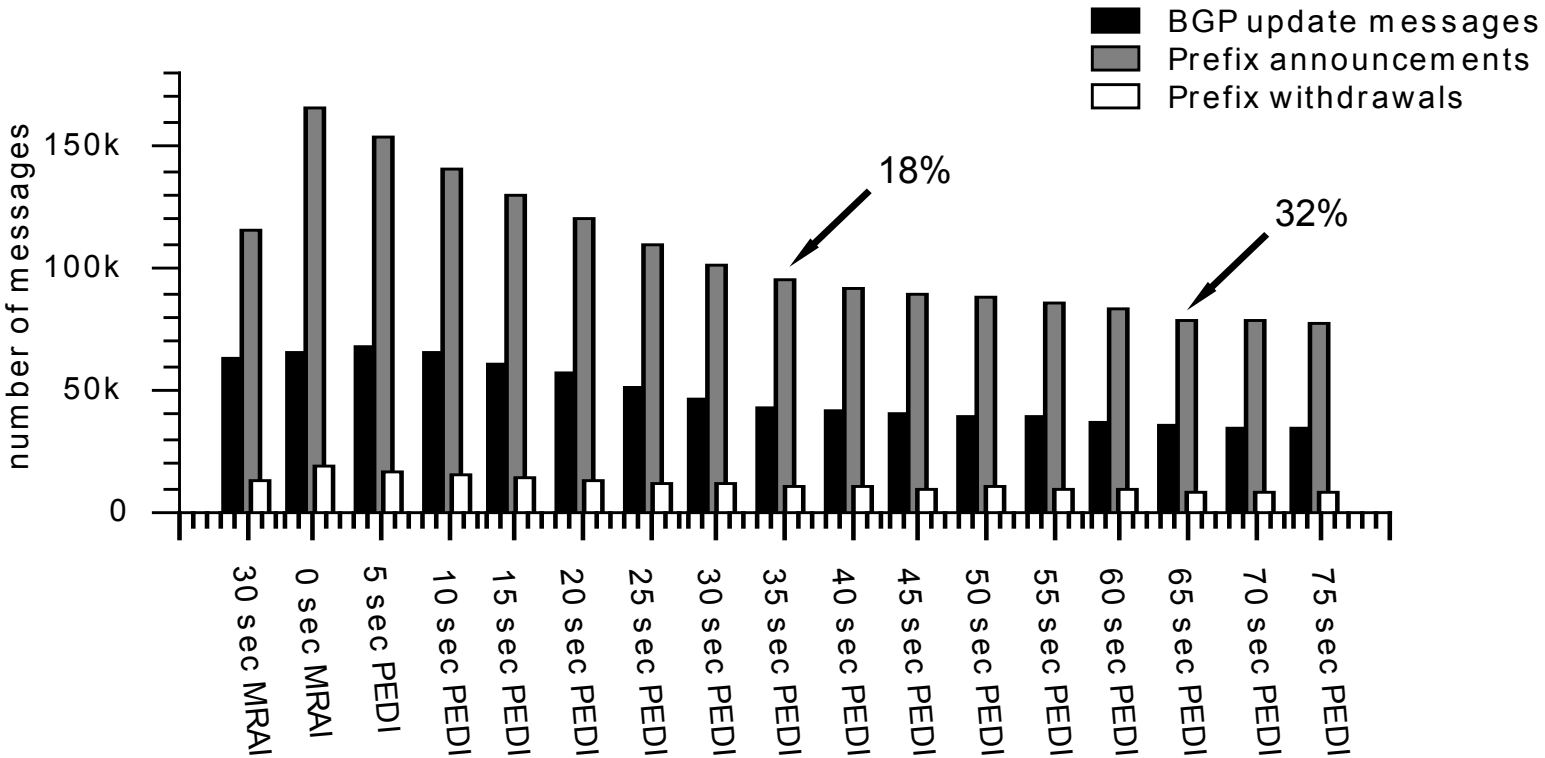
Reduction of update load – APNIC



Reduction of update load – Routeviews



Reduction of update load – Routeviews

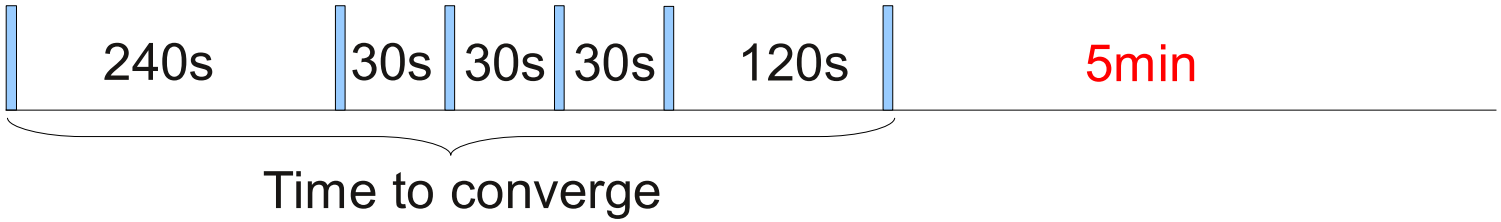


Outline

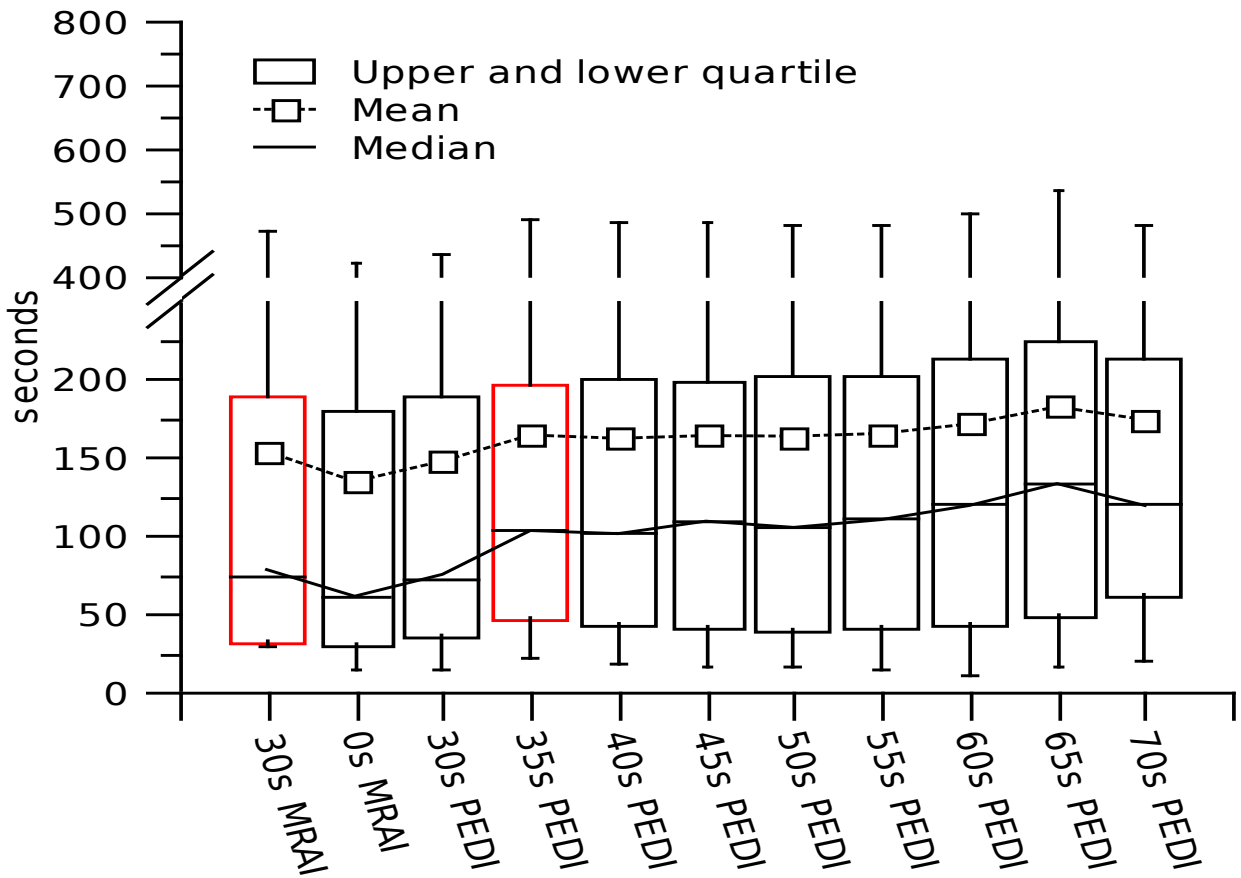
Convergence time approximation

- Control Plane convergence (Optimality)
- Data plane convergence – Forwarding path (Reachability)

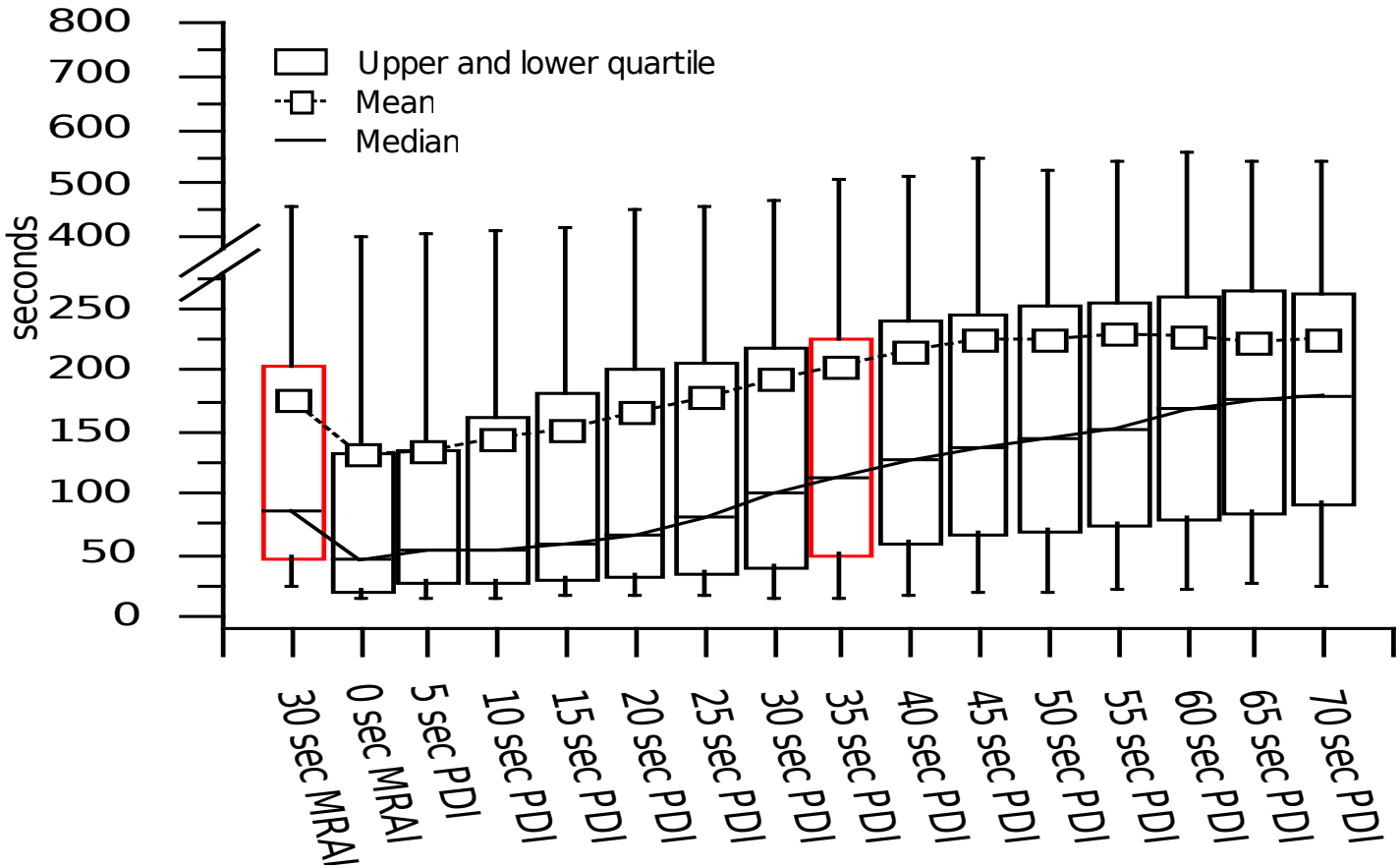
- Convergence measured from a single point of view (Optimality)
  - Router 2
- Convergence defined as a route being stable for 5 minutes



Optimality approximation – APNIC data



Optimality approximation – Routeviews data

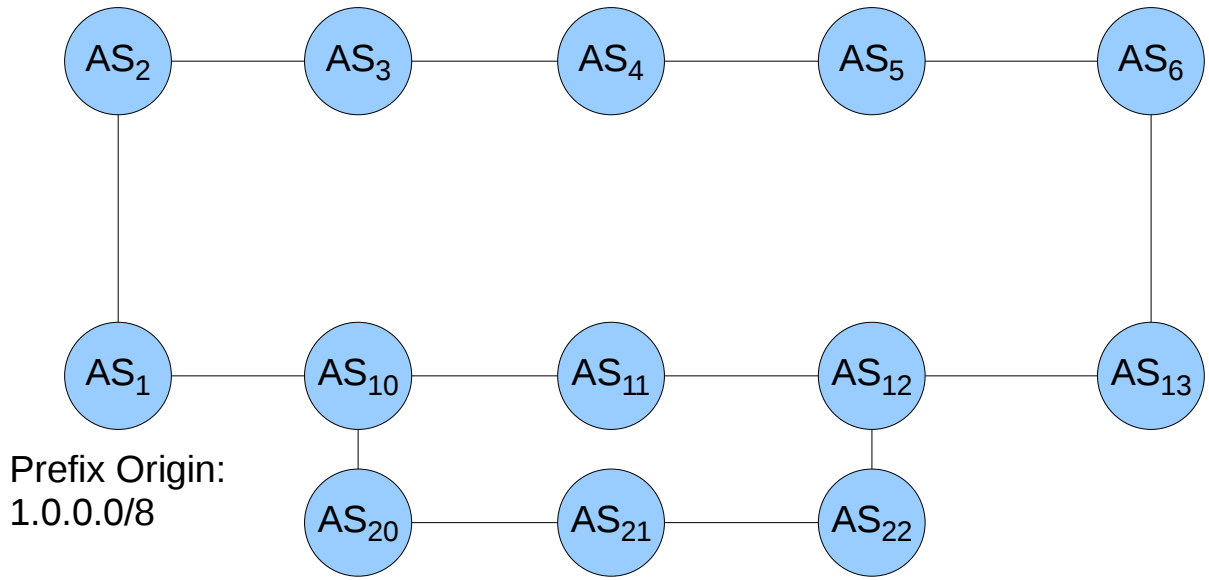


Detailed convergence time analysis

- Investigate Reachability vs. Optimality
- Analysis of a whole BGP system
- Impact on convergence of 4 events causing instability:
  - Link failure along the path – alternative path exists
  - Link recovery
  - Prefix withdrawal
  - Prefix announcement

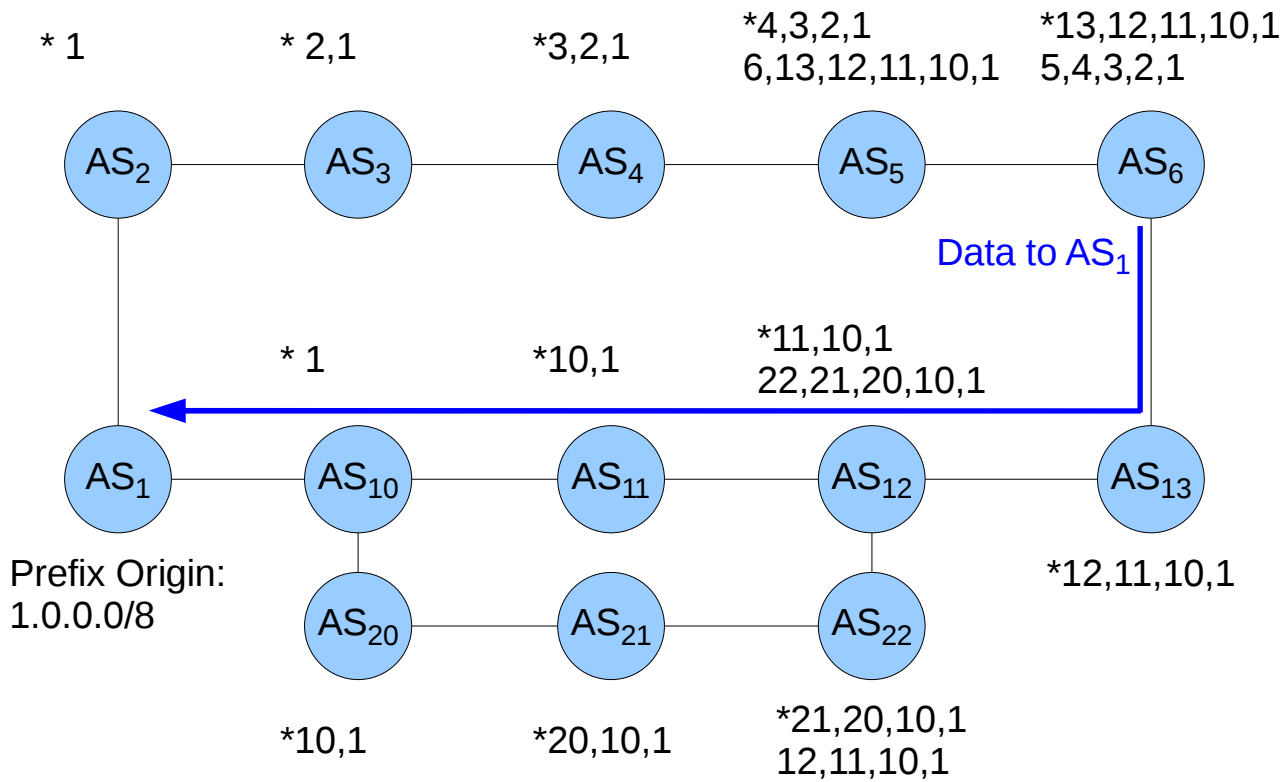
Detailed convergence time analysis

- Experimental analysis over 20 testruns
- Simple example topology
  - ASes are single BGP speakers
- Example: 30s MRAI on all ASes and 35s PEDI on all ASes



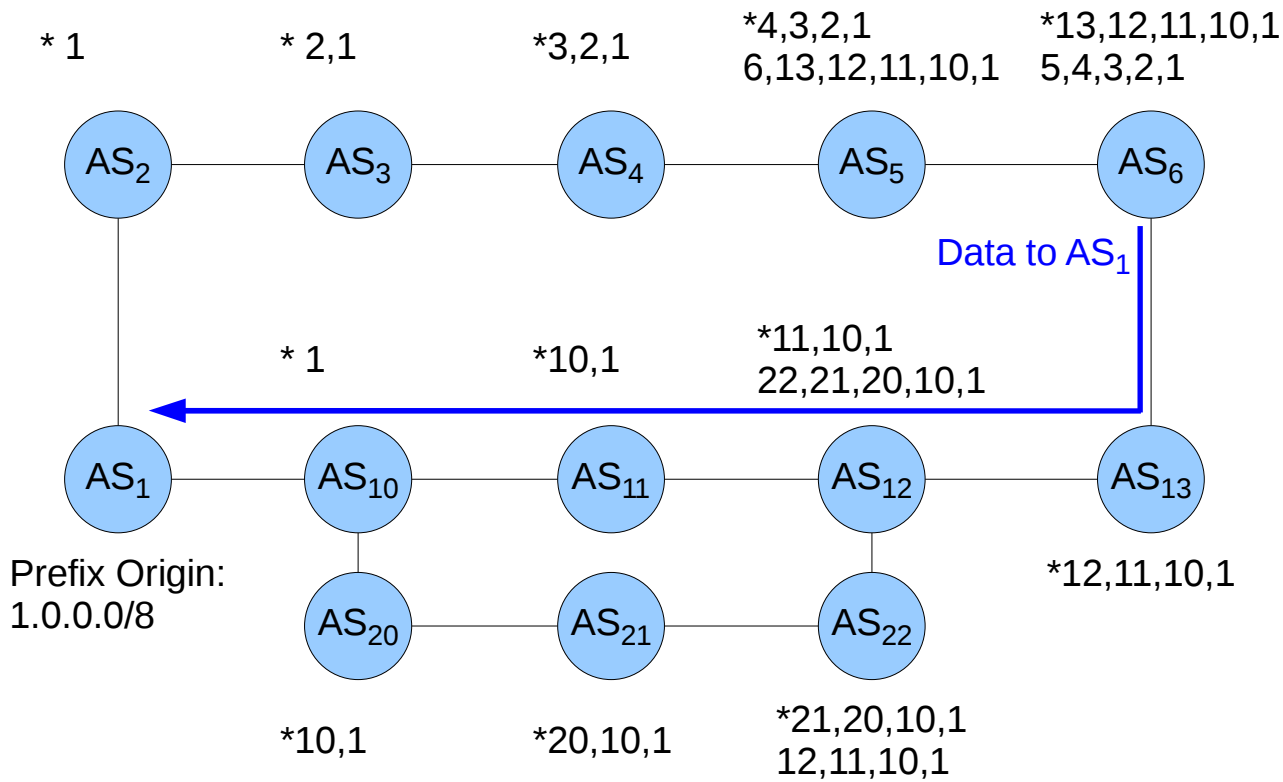
Optimality – Announcement of initial route

Stable System:



Optimality – Announcement of initial route

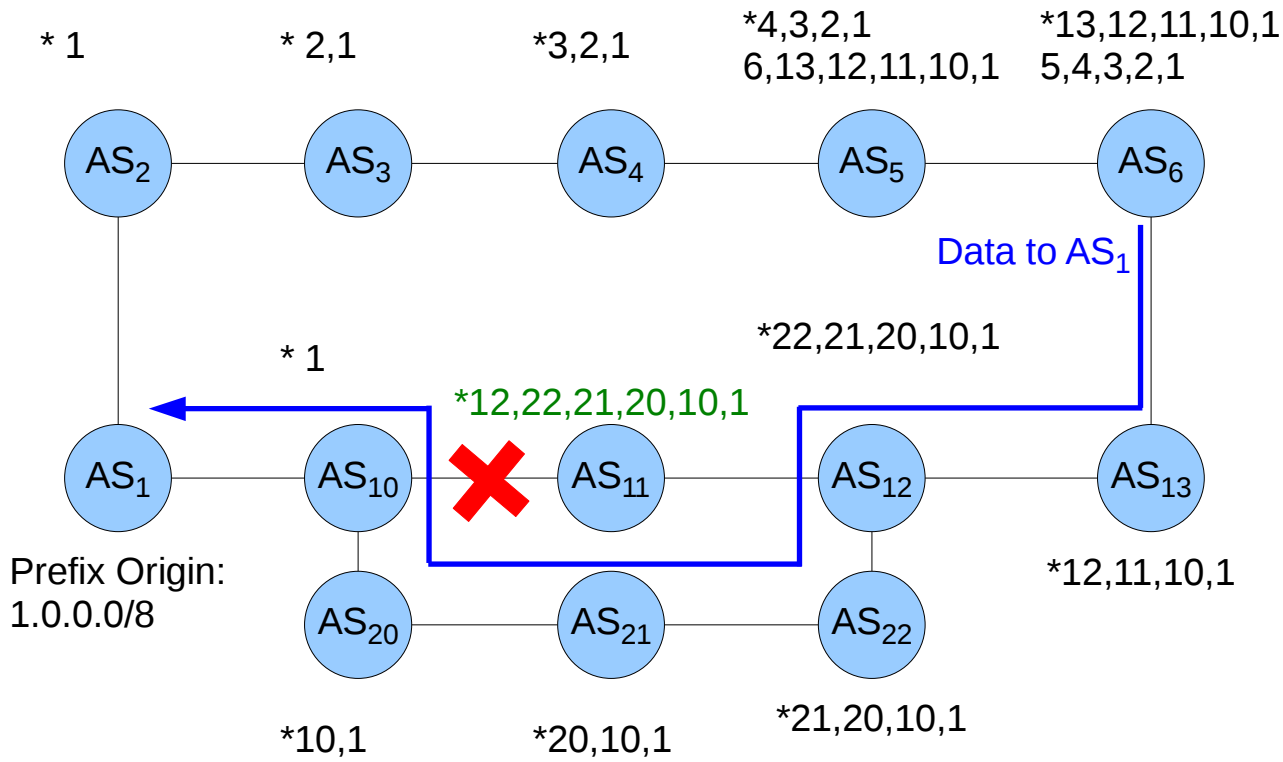
Stable System:



- Announcement of initial route at AS_6:
  - PED: 0 seconds
  - MRAI: 60-120 seconds

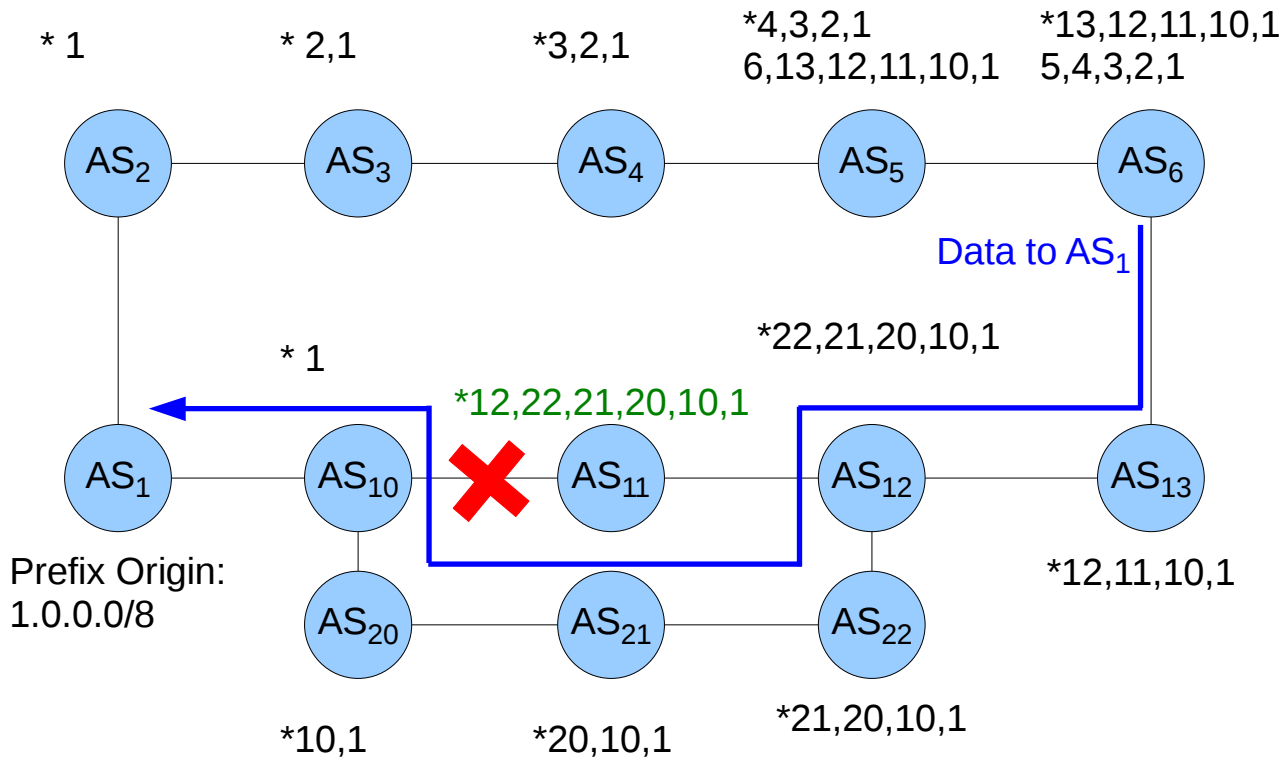
Reachability – Link failure

Link failure between AS_{10} and AS_{11}:



Reachability – Link failure

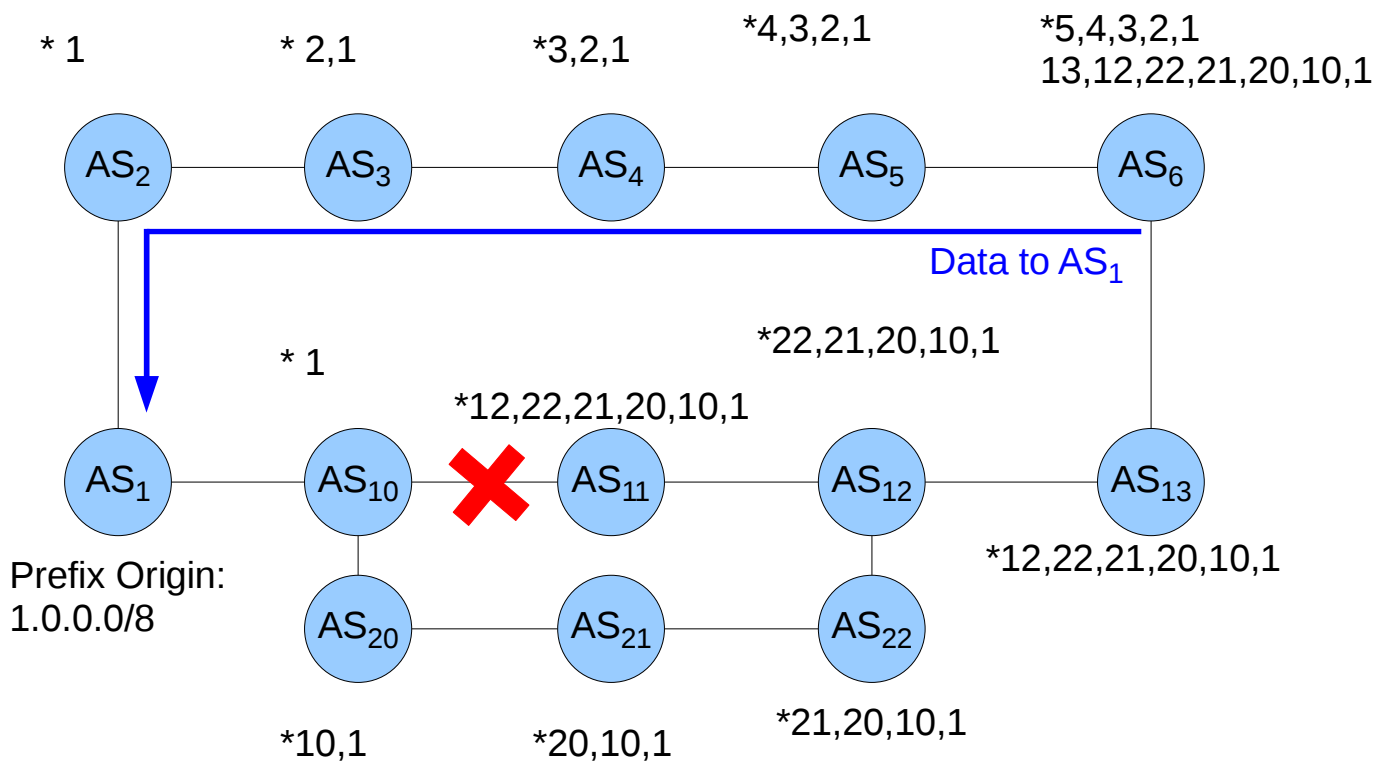
Link failure between AS_{10} and AS_{11}:



- Reachability achieved (AS_{11})
  - PED: 0 seconds
  - MRAI: 0-4 or 29-30 seconds

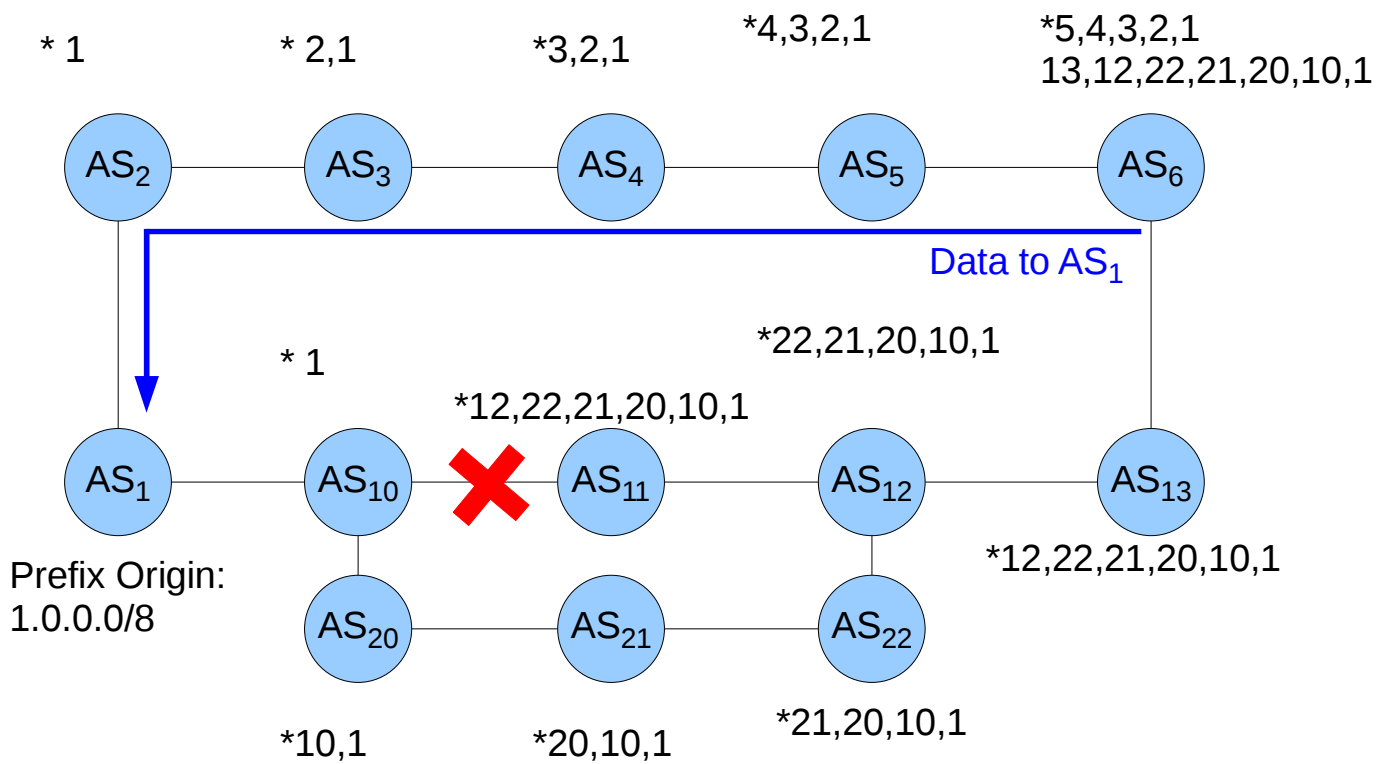
Optimality – Link failure

Link failure between AS_{10} and AS_{11}:



Optimality – Link failure

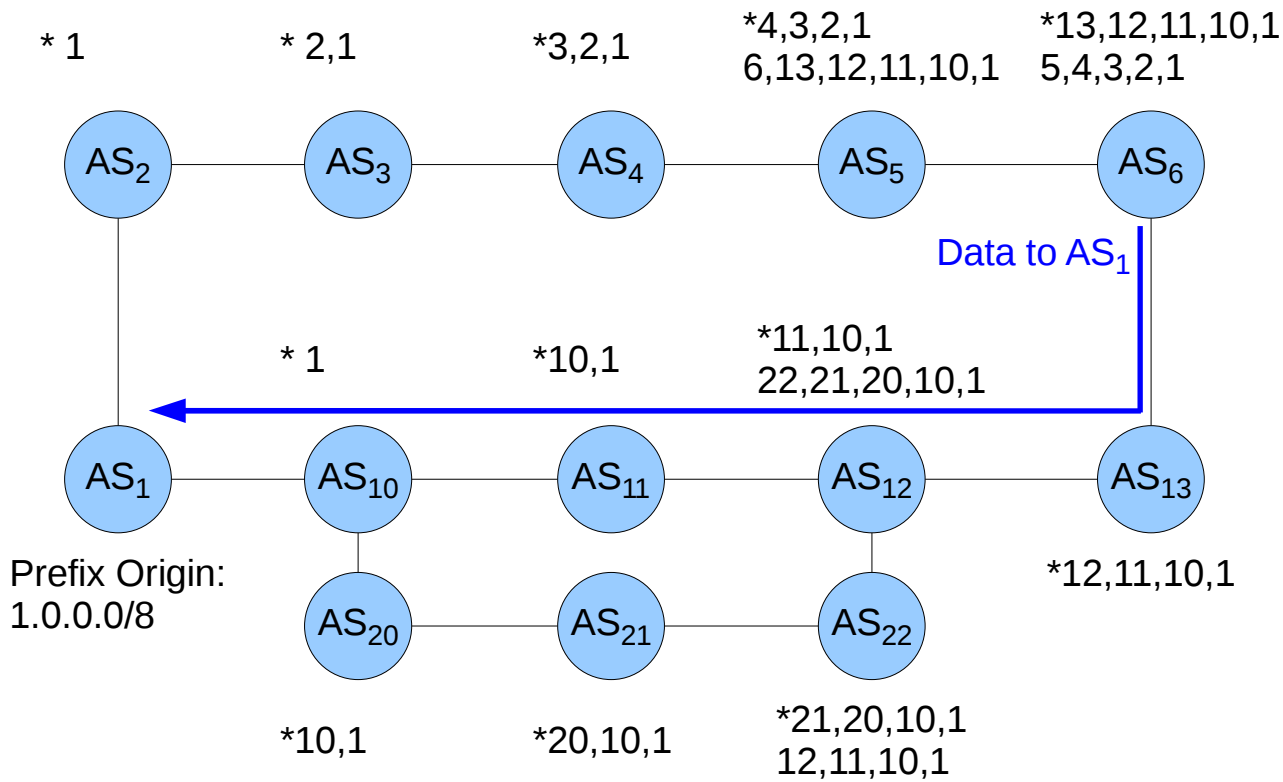
Link failure between AS_{10} and AS_{11}:



- Optimality achieved:
  - PED: 66 seconds (+-jitter)
  - MRAI: 2-58 seconds

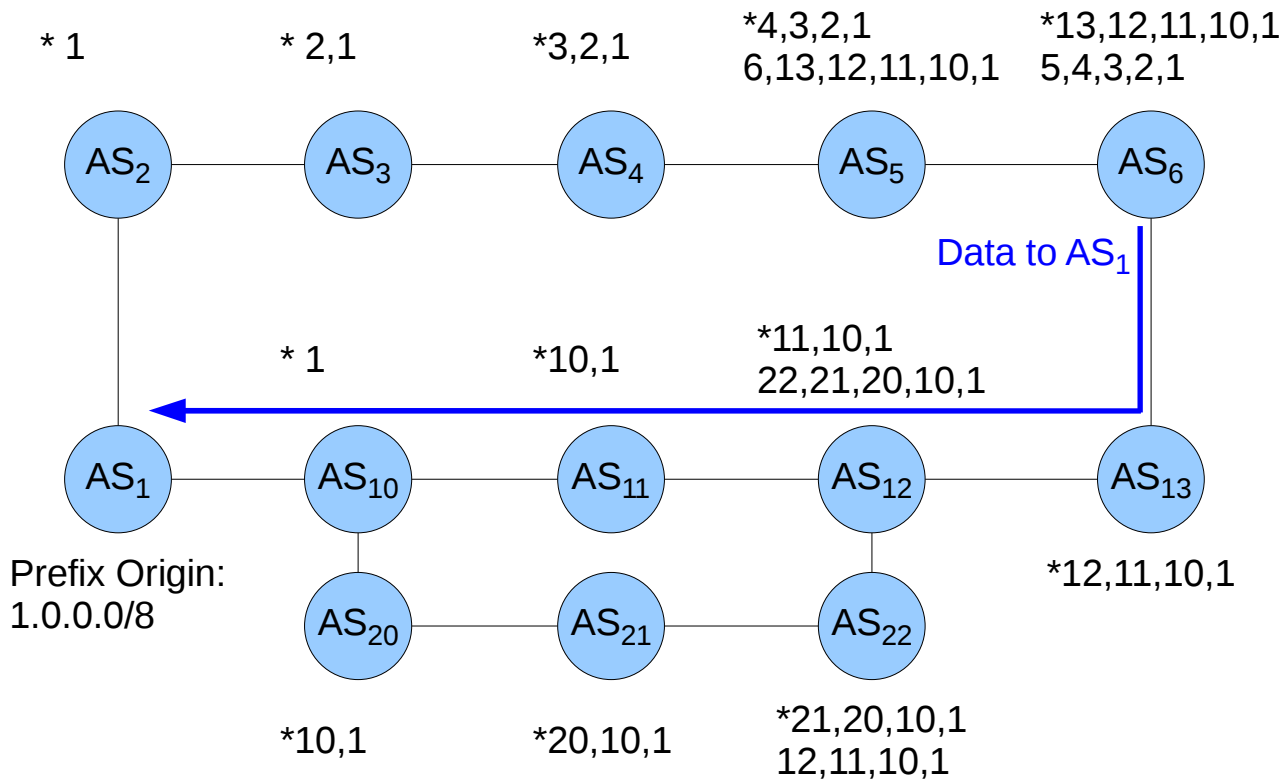
Optimality – Link recovery

Link recovery between AS_{10} and AS_{11}:



Optimality – Link recovery

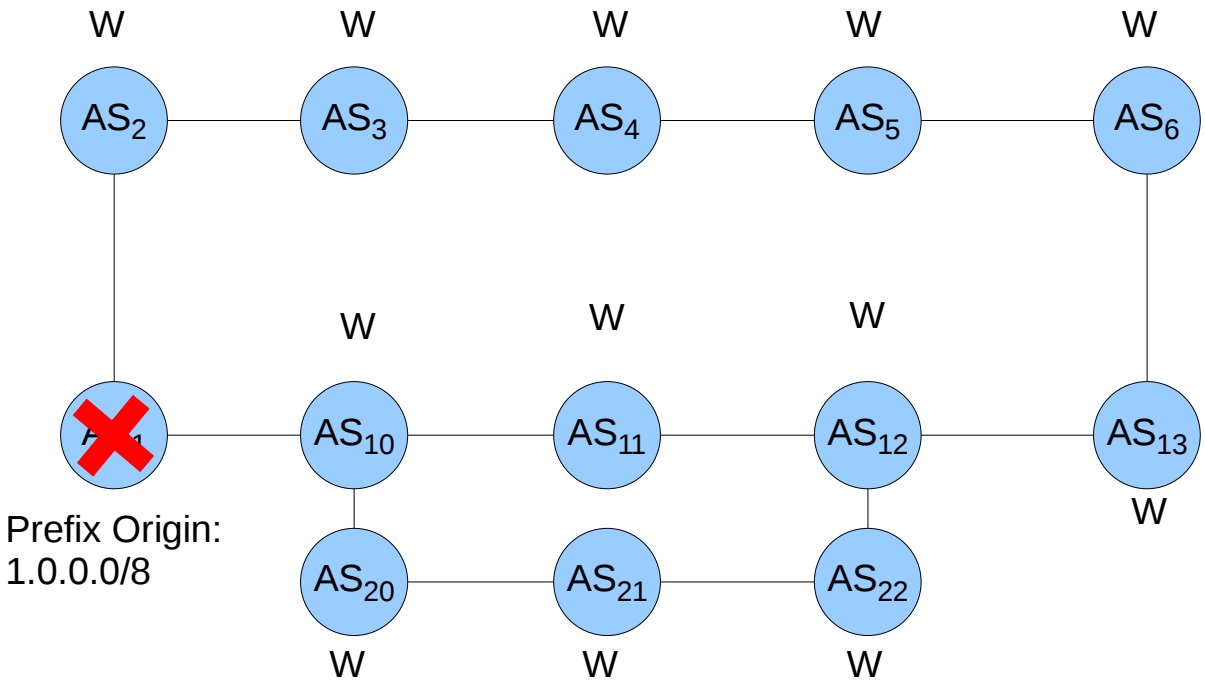
Link recovery between AS_{10} and AS_{11}:



- Optimality achieved:
  - PED: 0 seconds
  - MRAI: 31-33 and 55-60 seconds

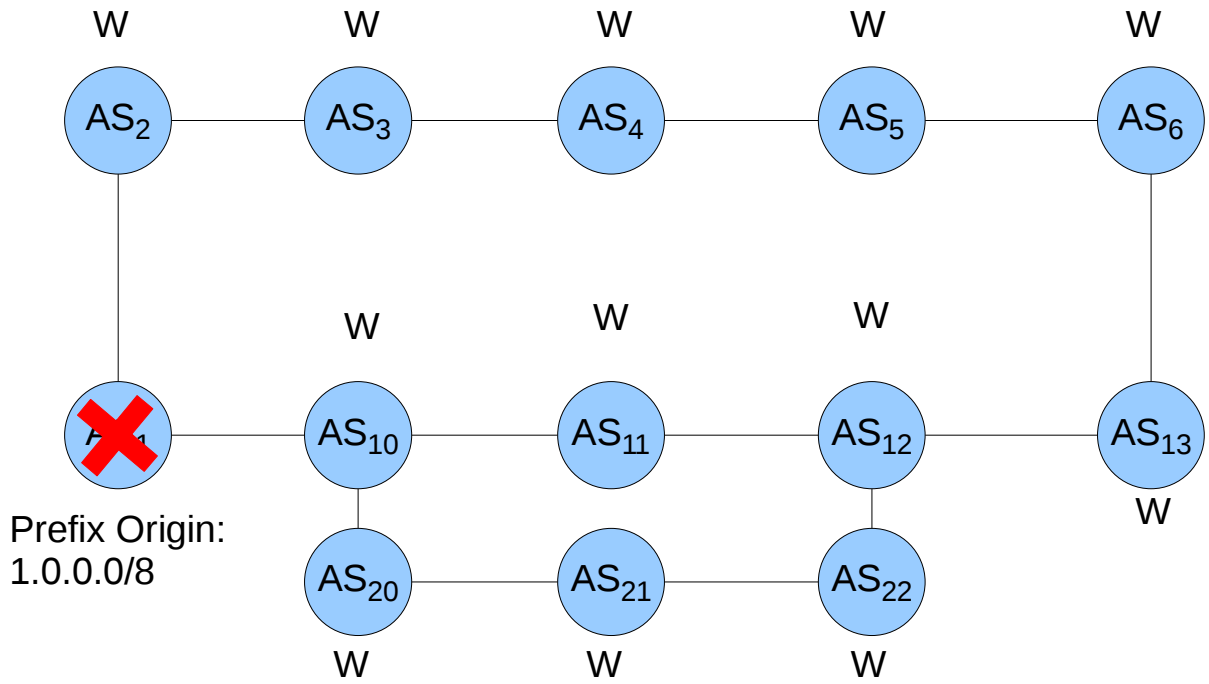
Optimality/Reachability – Prefix withdrawn

Prefix withdrawal at AS_1



Optimality/Reachability – Prefix withdrawn

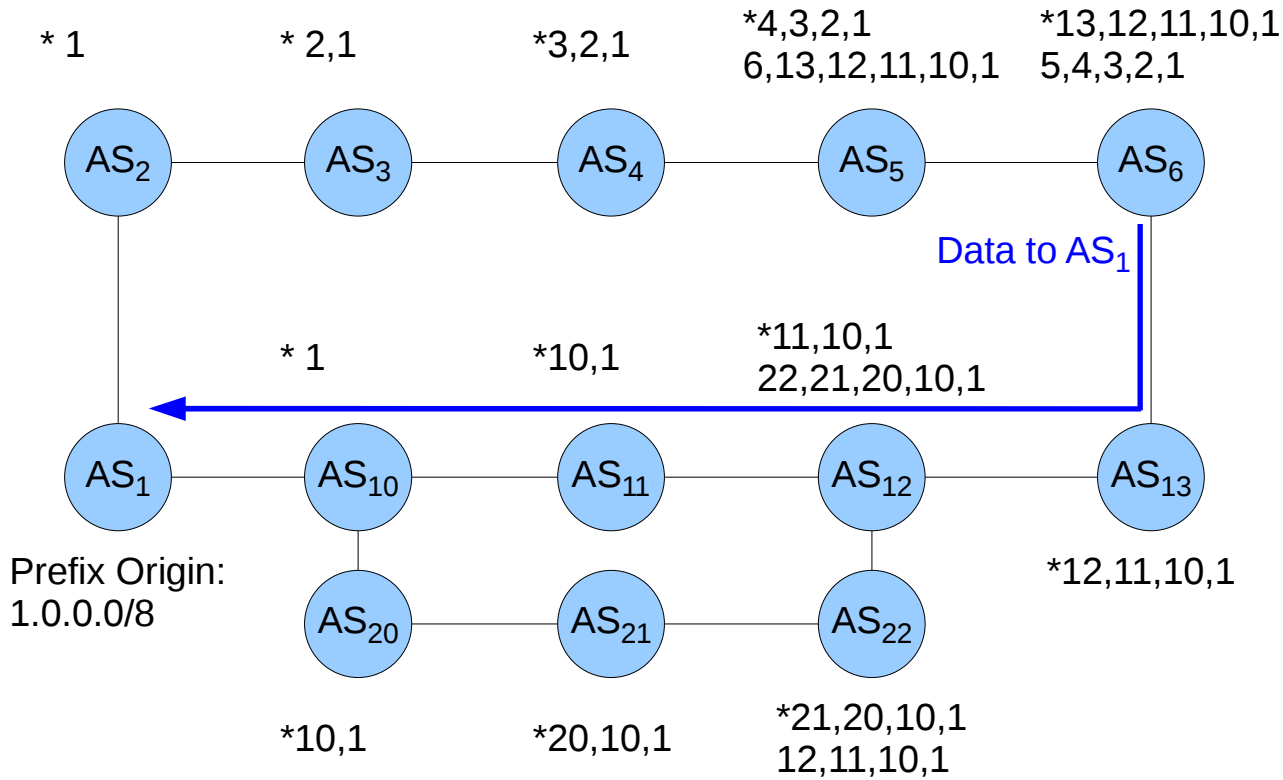
Prefix withdrawal at AS_1



- Optimality achieved (route withdrawn on every AS):
  - PED: 0 seconds
  - MRAI: 0 seconds

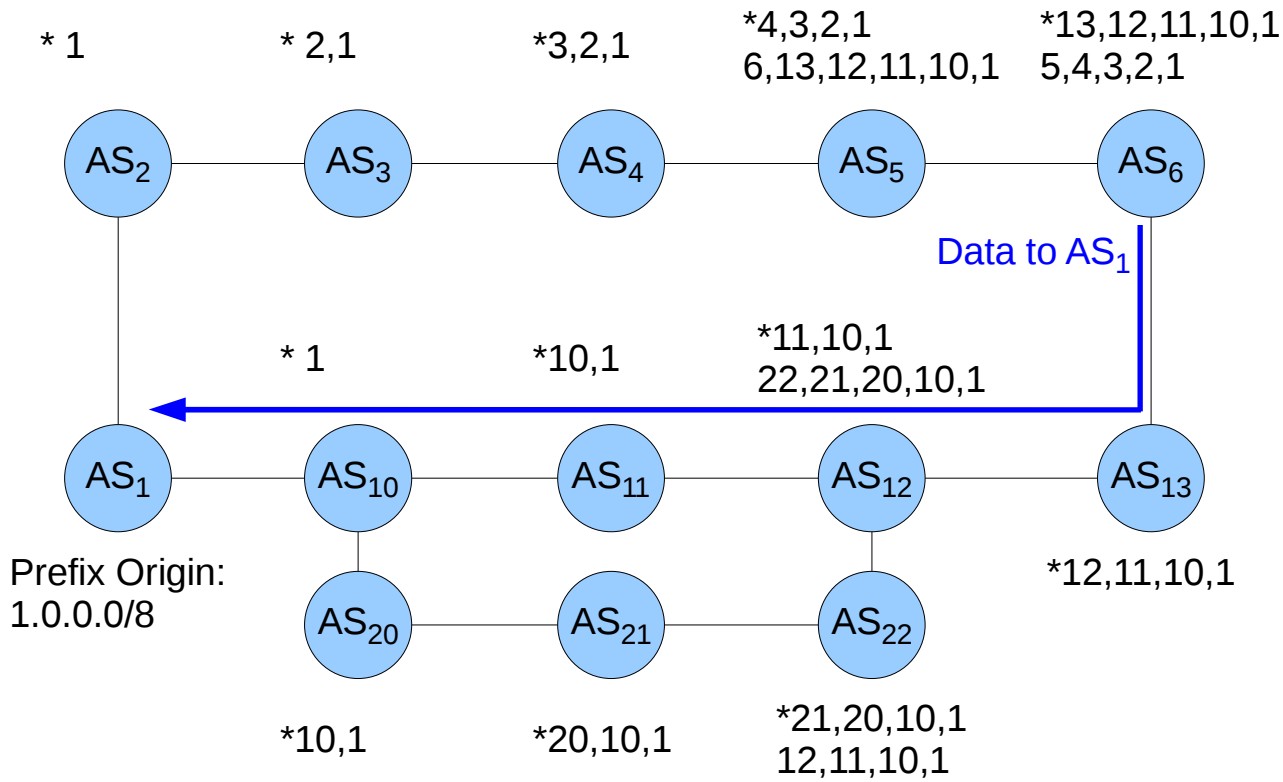
Optimality/Reachability – Prefix (re)-announced

Prefix announcement at AS_1



Optimality/Reachability – Prefix (re)-announced

Prefix announcement at AS_1



- Optimality achieved (same as initial announcement):
  - PED: 0 seconds
  - MRAI: 32-34, 58-60, 76-90 seconds

Outline

Introduction

Motivation

Path Exploration Damping - PED

Experimental results
- Reduction of update load
- MRAI and PED convergence time compared

Conclusions and future work

Conclusions

- In this talk:
  - PED decreases update load
  - PED converges to Reachability as fast or faster than MRAI
  - PED converges to Optimality slower than MRAI in one case
- In the paper:
  - PED interacts well with MRAI
  - PED can be deployed incrementally
  - A single PED speaker is beneficial to the BGP system
  - 35s PEDI is a safe defaulte value in the MRAI dominated Internet
- In the future:
  - Dynamic PEDI per prefix
  - More heuristics