

Outage Clustering: From Leaves to Trees

John Heidemann

joint work with Yuri Pradkin, Aqib Nisar
USC/ISI

CAIDA Active Internet Measurements (AIMS) / 15 March 2018

This research is sponsored by the Department of Homeland Security (DHS) Science and Technology Directorate, HSARPA, Cyber Security Division, BAA 11-01-RIKA and Air Force Research Laboratory, Information Directorate under agreement number FA8750-12-2-0344, and contract number D08PC73599. The U.S. Gov't is authorized to reproduce and distribute reprints for Gov't purposes notwithstanding any copyright notation thereon. The views herein are those of the authors and do not necessarily represent those of DHS or the U.S. Gov't.



Copyright © 2017 by John Heidemann
Release terms: CC-BY-NC 4.0 international

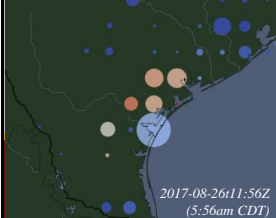


So Much Internet Outage Data...

Trinocular
24x7 since Nov. 2013

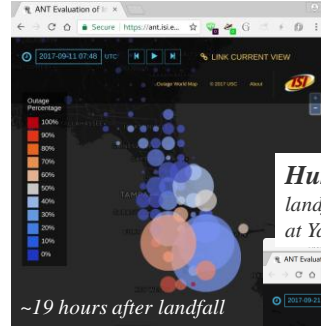


Hurricane Harvey
landfall 2017-08-26t03:10Z
at Port O'Conner, Texas



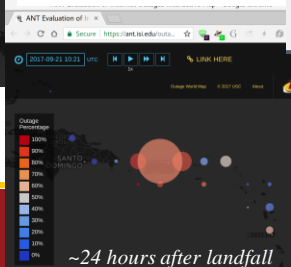
2017-08-26t11:56Z
(5:56am CDT)

Hurricane Irma
landfall 2017-09-10t13:10Z
at Cudjoe Key, Florida



~19 hours after landfall

Hurricane Maria
landfall: 2017-09-20t10:15Z
at Yaboucoa, P.R.



~24 hours after landfall

IODA, UCSD/CAIDA.

IODA Signals for Iraq

Alert Feed for Iraq

Time	Type	Auto
Oct 6th 7:20am	Active	
Oct 6th 7:20am	Probing	
Oct 6th 2016	BGP	7
Oct 6th 5:50am		

Hurricane Sandy: Residential Internet Outages and Recovery - YouTube - Google Chrome

pingin in the rain

ThunderPing, U.MD.

Too Much Long-Term Data?

- USC/ISI's Trinocular: outages, 24x7, since Nov. 2013
- about 40TB (!)
- about 20k observations x 4M blocks:
80G datapoints (!!)
- how to make sense of it?
 - from leaves (edge networks)
 - to trees (events)
 - on the way to understanding the “forest” of Internet reliability

Making Sense of Too Much Data

- **geographic visualization**
interactively explore the world
 - **non-geographic visualization**
begin to reveal patterns
 - **clustering by similarity**
discover underlying dependencies
- with too much data
(40TB and
80G observations)

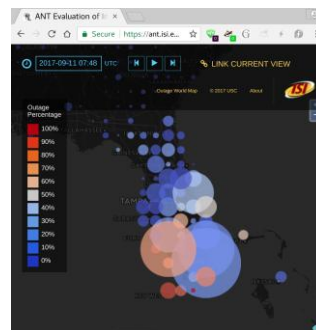
Making Sense of Too Much Data

- **geographic visualization**
interactively explore the world
- **non-geographic visualization**
begin to reveal patterns
with too much data
(40TB and
80G observations)
- **clustering by similarity**
discover underlying dependencies

Geographic Visualization

- on the web: <https://ant.isi.edu/outage/world/>
- key features
 - circle size: *number* of blocks out
 - color: *percent* of blocks out
 - time selection
 - geographic zoom and pan
 - **geography: easy to relate to
(what operators ask for!)**

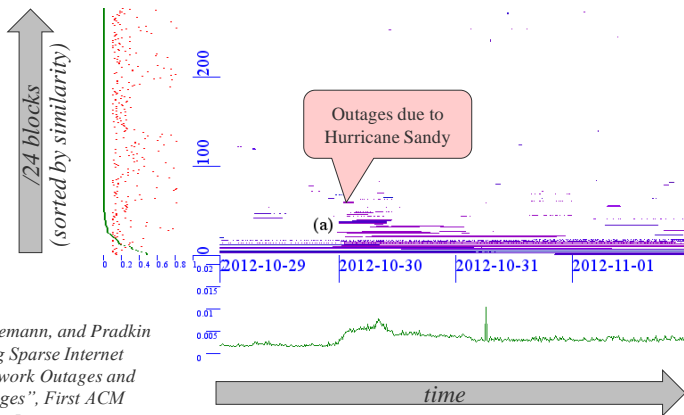
*Florida, ~19 hours after landfall
of Hurricane Irma*



Making Sense of Too Much Data

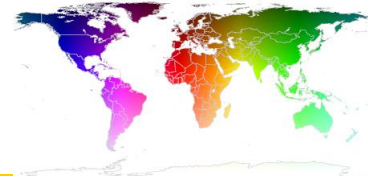
- **geographic visualization**
interactively explore the world
- **non-geographic visualization**
begin to reveal patterns
with too much data (40TB and 80G observations)
- **clustering by similarity**
discover underlying dependencies

Non-Geographic Visualizations: the *Network* in Outages



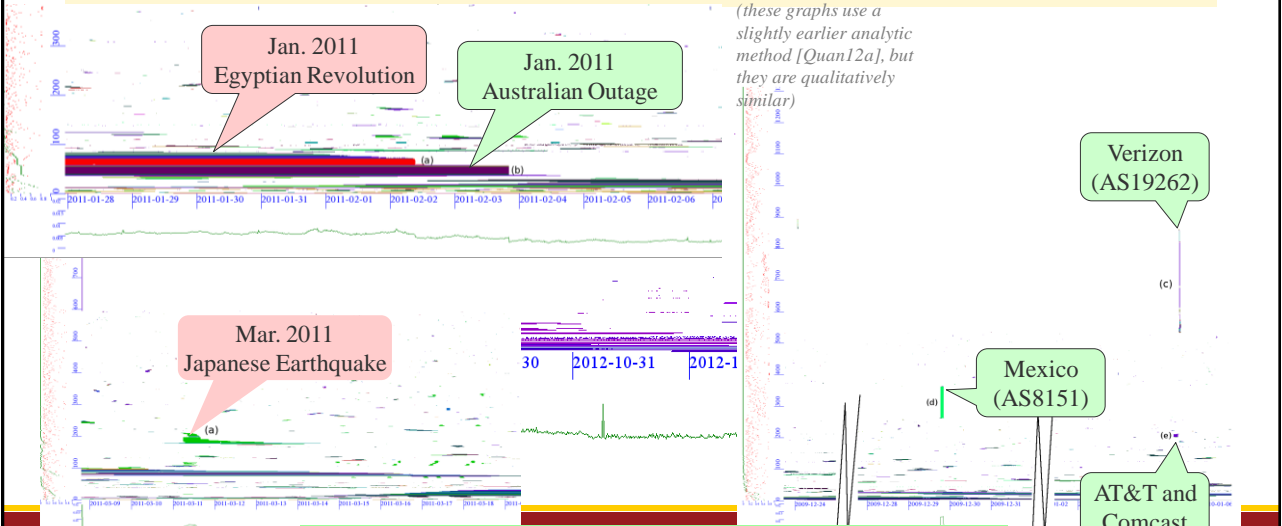
goal: reveal patterns
find dependencies
among networks

(colored areas are outages,
color shows location)



Quan, Heidemann, and Pradkin
"Visualizing Sparse Internet
Events: Network Outages and
Route Changes", First ACM
Workshop on Internet
Visualization, Nov. 2012

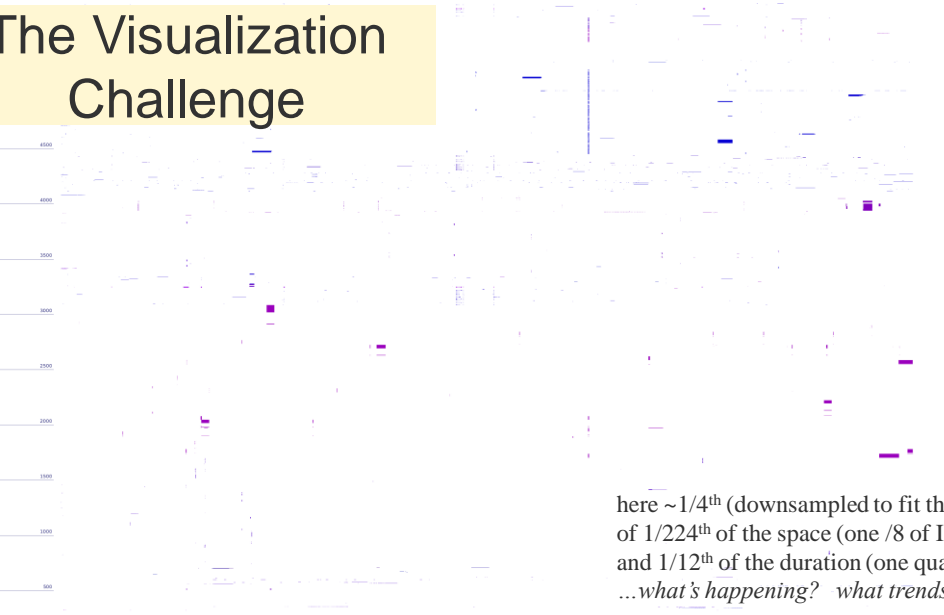
Global Network Outages: **Prominent** and *Unknown*



our goal: understand small *and* big

The Visualization Challenge

↑ 24 blocks (sorted by block IP address)



here ~1/4th (downsampled to fit the screen) of 1/224th of the space (one /8 of IPv4) and 1/12th of the duration (one quarter of ~3 years) ...what's happening? what trends? what's new?

Efficient Visualization

- **visualization with linear ordering algorithm**

- runtime: $O(n \log n \log m)$
- for n blocks and m duration timesteps

- approach:

- map clustering to sorting: $O(n \log n)$ in time
- sort on *multi-timescale bitmap*: $O(\log m)$ in space

Presented at AIMS 2017 (last year!)

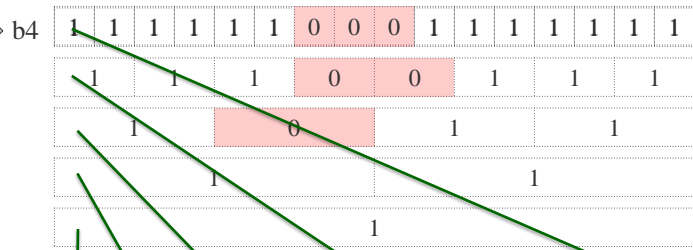
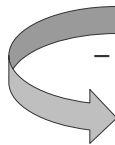
Details in “Back Out: End-to-end Inference of Common Points-of-Failure in the Internet (extended)”. ISI-TR-724, Feb., 2018.

www.isi.edu/~johnh/PAPERS/Heidemann18b.pdf

Multi-Timescale for Similarity

- input: outage timeseries from 5 /24 blocks

- b1 1111 1110 1111 1111
 - b2 1111 1111 1111 1110
 - b3 1111 1100 1111 1111
 - b4 1111 1100 0111 1111
 - b5 1111 1110 1111 1111
- goal: cluster by “similarity”



downsample with mean (keep fractions internally)

concatenate: 1 - 11 - 1011 - 1110 0111 - 1111 1100 0111 1111

Multi-Timescale Finds Similarity

- input: outage timeseries from 5 /24 blocks

```

- b1 1111 1110 1111 1111
- b2 1111 1111 1111 1110
- b3 1111 1100 1111 1111
- b4 1111 1100 0111 1111
- b5 1111 1110 1111 1111
  
```

goal: cluster by “similarity”

- apply to all blocks...

```

- b1 1 - 11 - 1111 - 1110 1111 - 1111 1110 1111 1111
- b2 1 - 11 - 1111 - 1111 1110 - 1111 1111 1111 1110
- b3 1 - 11 - 1011 - 1110 1111 - 1111 1100 1111 1111
- b4 1 - 11 - 1011 - 1110 0111 - 1111 1100 0111 1111
- b5 1 - 11 - 1111 - 1110 1111 - 1111 1110 1111 1111
  
```

Multi-Timescale Finds Similarity

- input: outage timeseries from 5 /24 blocks

```

- b1 1111 1110 1111 1111
- b2 1111 1111 1111 1110
- b3 1111 1100 1111 1111
- b4 1111 1100 0111 1111
- b5 1111 1110 1111 1111
  
```

goal: cluster by “similarity”

define similar as adjacent in multi-timescale vectors

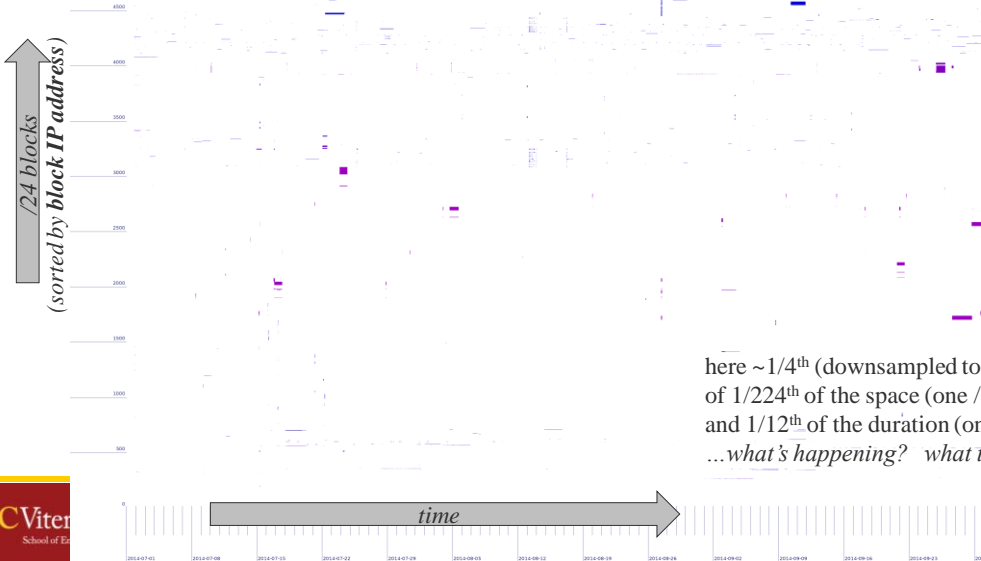
- apply to all blocks and **sort**

```

- b2 1 - 11 - 1111 - 1111 1110 - 1111 1111 1111 1110
- b1 1 - 11 - 1111 - 1110 1111 - 1111 1110 1111 1111
- b5 1 - 11 - 1111 - 1110 1111 - 1111 1110 1111 1111
- b3 1 - 11 - 1011 - 1110 1111 - 1111 1100 1111 1111
- b4 1 - 11 - 1011 - 1110 0111 - 1111 1100 0111 1111
  
```

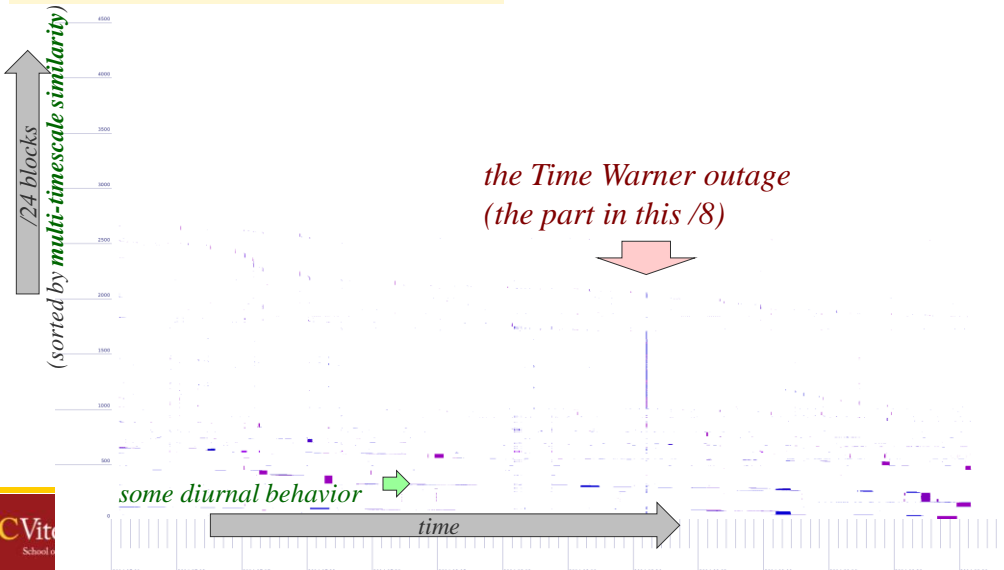
result: better clusters
(Hamming distance from 8 to 4)

The Visualization Challenge



here ~1/4th (downsampled to fit the screen) of 1/224th of the space (one /8 of IPv4) and 1/12th of the duration (one quarter of ~3 years) ...*what's happening? what trends? what's new?*

One Visualization Result



here ~1/4th (downsampled to fit the screen) of 1/224th of the space (one /8 of IPv4) and 1/12th of the duration (one quarter of ~3 years)

Making Sense of Too Much Data

- **geographic visualization**
interactively explore the world
 - **non-geographic visualization**
begin to reveal patterns
 - **clustering by similarity**
discover underlying dependencies
- with too much data
(40TB and
80G observations)

Clustering to Discovery Dependencies

- visualization is nice, but humans can't look at everything
- new clustering algorithms can *discover dependencies*
 - common failure patterns
 - implying common root causes
 - (unconstrained by 2-D visualization)

Clustering Insight

- things fail and recover together => possible dependency
- when *consistently, multiple times* => *probable* dependency

(Details: John Heidemann, Yuri Pradkin, and Aqib Nisar. *Back Out: End-to-end Inference of Common Points-of-Failure in the Internet (extended)*. ISI-TR-724, February, 2018.
<https://www.isi.edu/%7ejohnh/PAPERS/Heidemann18b.html>.)

Clustering Approach

start with timeseries

```

b1 1 - 11 - 1111 - 1110 1111 - 1111 1110 1111 1111
b2 1 - 11 - 1111 - 1111 1110 - 1111 1111 1111 1110
b3 1 - 11 - 1011 - 1110 1111 - 1111 1100 1111 1111
b4 1 - 11 - 1011 - 1110 0111 - 1111 1100 0111 1111
b5 1 - 11 - 1111 - 1110 1111 - 1111 1110 1111 1111
    
```

pick timescale

```

(b1,1,0) (b1,0,3) (b1,1,4) (b1,u,8)
(b2,1,0) (b2,0,3) (b2,u,8)
(b3,1,0) (b3,0,3) (b3,1,4) (b3,u,8)
(b4,1,0) (b4,0,3) (b4,1,5) (b4,u,8)
(b5,1,0) (b5,0,3) (b5,1,4) (b5,u,8)
    
```

find transitions [Heidemann18b, Figure 5]

identify clusters:

b1, b3, b5 are a cluster,
 because
 (b1,b3) (b1,b5) (b3,b5) are all strong edges
 because $C_{b1,b2} = 1$

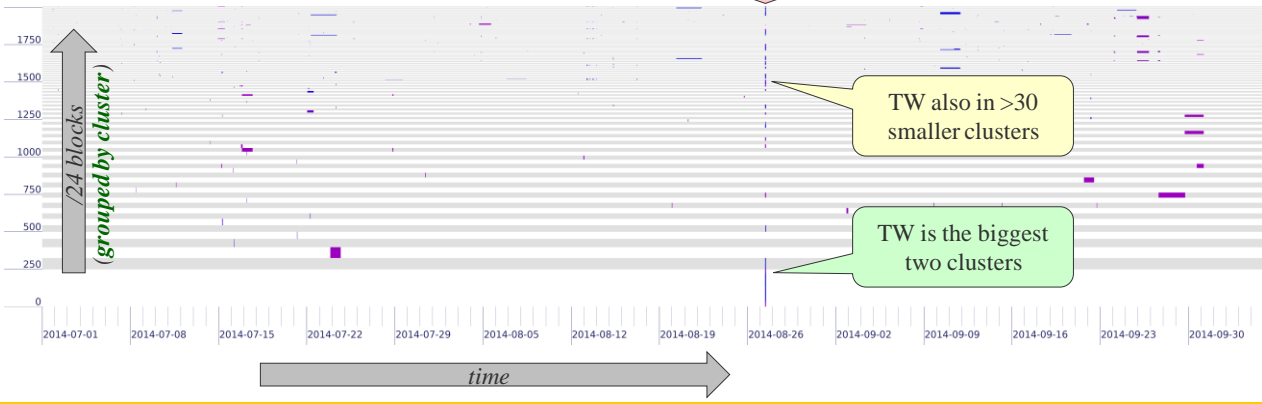
time bin	events...
3	(b1,b3,0) (b1,b4,0) (b1,b5,0) (b3,b4,0) (b3,b5,0) (b4,b5,0)
4	(b1,b3,1) (b1,b5,1) (b3,b5,1)
$N_{b1,b2}$	2 1 2 1 2 1
$C_{b1,b2}$	1 0.5 1 0.5 1 0.5

look for consistent transitions [Heidemann18b, Figure 6]

One Clustering Result

1/224th of the space (one /8 of IPv4)
and 1/12th of the duration (one quarter of ~3 years)

*the Time Warner outage
(the part in this /8)*

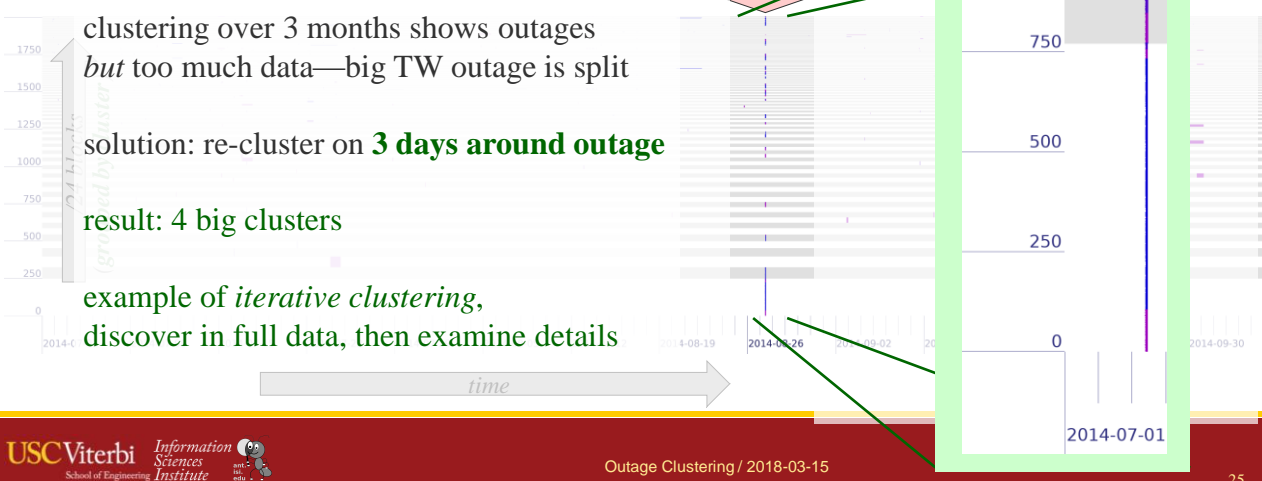


Iterative Clustering

1/224th of the space (one /8 of IPv4)
and 1/12th of the duration (one quarter of ~3 years)
now just **3 days** of time

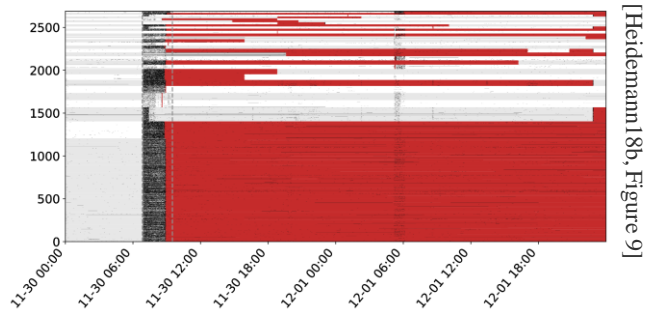
*the Time Warner outage
(the part in this /8)*

clustering over 3 months shows outages
but too much data—big TW outage is split
solution: re-cluster on **3 days around outage**
result: 4 big clusters
example of *iterative clustering*,
discover in full data, then examine details



Clustering Beyond Outages

- cluster generalizes:
detects *temporal correlations*
in *big timeseries*
- we've applied it to
 - outages
 - anycast catchments
 - routing updates
- for anycast:
 - map initial vs. new cluster to binary value
 - skip outages (from DDoS)



anycast catchments for J-Root around the 2015-11-30 DDoS attack

[Heidemann 18b, Figure 9]

Outage Clustering from Here

- just released clustering technical report
- opens many new questions...
 - relating to other information? (like power outages)
 - what is “normal”?
 - can we evaluate policy \Leftrightarrow reliability?
- datasets at <https://ant.isi.edu/datasets/outage> and <https://imactcybertrust.org>
- code available on request

