Making Sense of BGP

Tina Wong, Van Jacobson, Cengiz Alaettinoglu

Packet Design Inc.

NANOG 30

Miami, FL

February 9, 2004

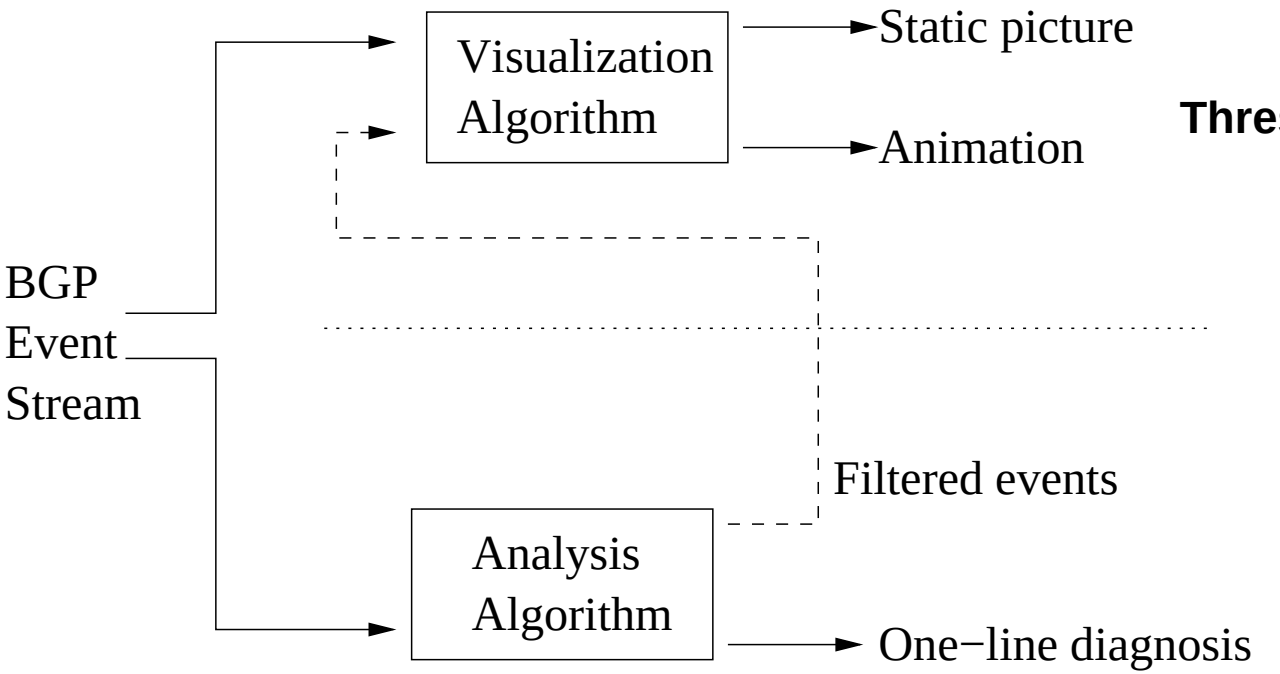
Overview

We have developed two new algorithms to help diagnose BGP problems:

- A visualization technique we call *TAMP* that shows the large-scale structure of some set of BGP routes.
- An analysis technique we call *Stemming* helps to do root-cause analysis of BGP event streams.

Both algorithms are driven by raw BGP event streams. They have no built-in models, statistical or otherwise. Their goal is to show the world *as the routers see it*.

Architecture



Threshold and Merge Prefixes (TAMP)

Stemming

The analysis and visualization modules can be used together or separately.

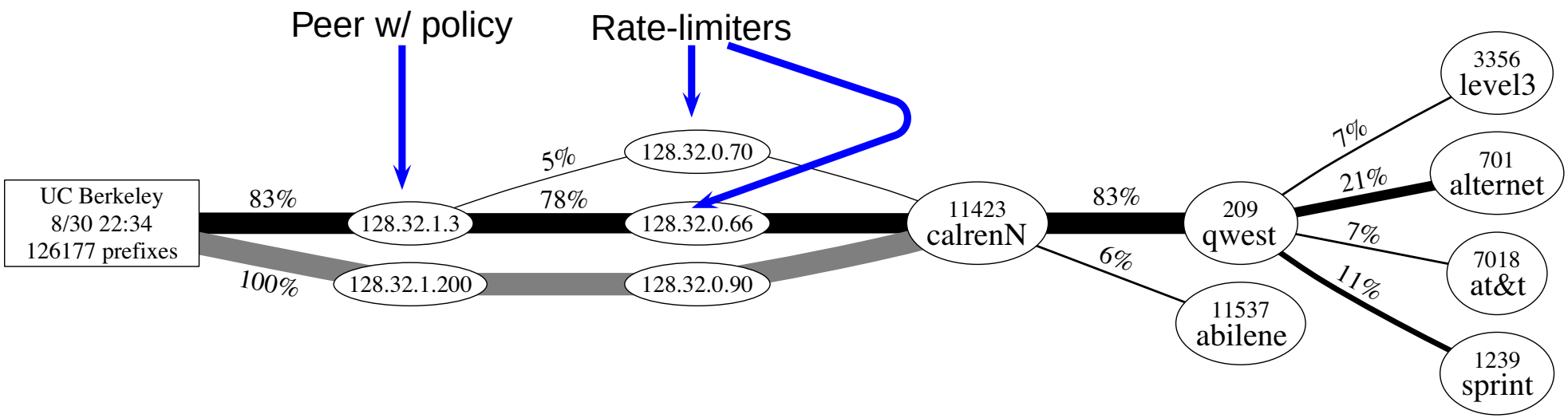
We will first go over static part of TAMP.

TAMP (Threshold and Merge Prefixes) BGP routing visualization

- For a particular BGP peer & time, that peer's routes form a tree rooted at the peer.
  - Take the number of prefixes using an edge as its weight (interested in edges that are heavily used).
  - Prune all edges with weight below a specified threshold (result is subtree of original tree and has the same root).
- Construct this subtree for each peer and/or time then merge trees by combining common prefixes.

TAMP shows the large-scale structure of a set of routes. By appropriately choosing the set different problems can be diagnosed.

Static TAMP visualization
Berkeley's BGP topology

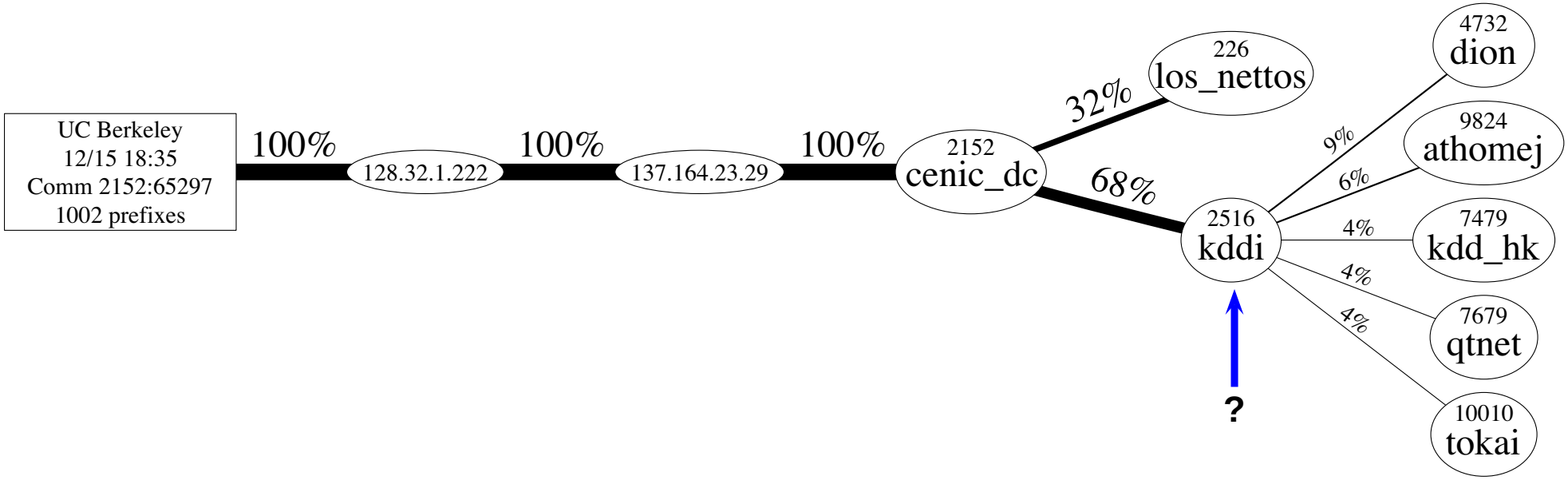


- edge thickness proportional to number of prefi xes

Mostly see what is expected, except...

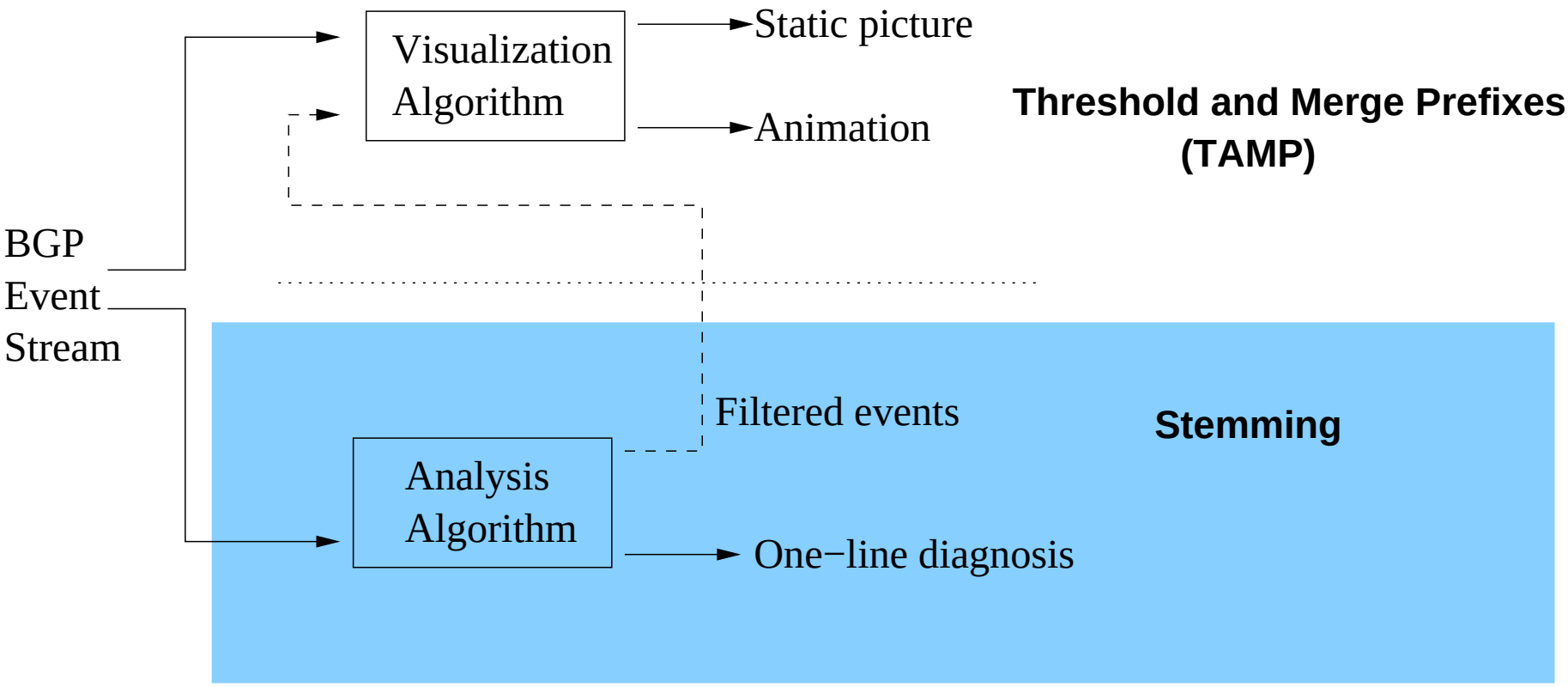
128.32.1.3 configured to carry commodity Internet traffic, and split prefixes onto 2 rate-limiting BGP Nexthops for load balancing. Division is less than optimal: one carried 78%, the other only 5%. Unintended.

Static TAMP visualization
CENIC Los Nettos routes (2152:65297) as seen by Berkeley



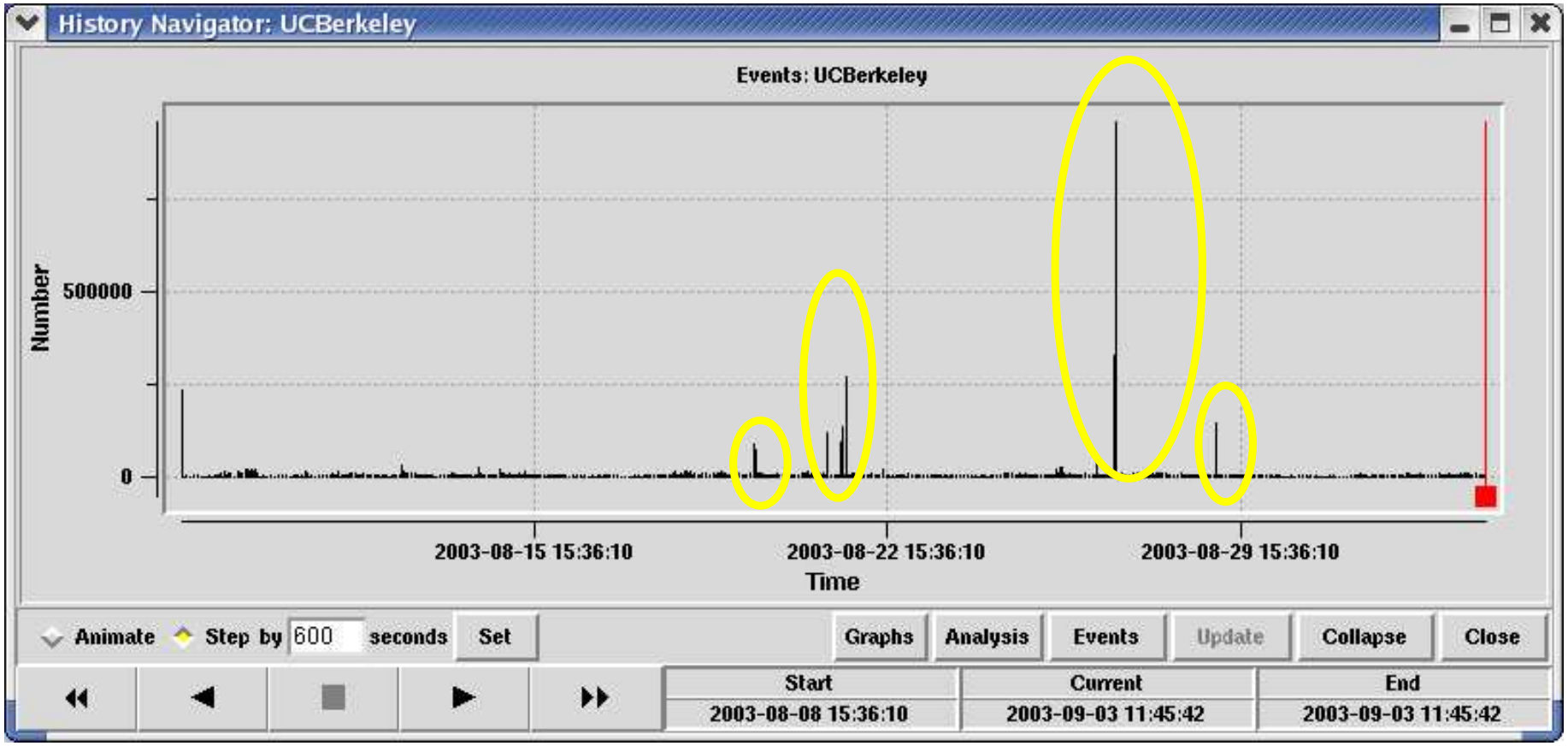
The topological meaning of CENIC community 2152:25297 is supposed to be routes coming from Los Nettos. Why are routes from KDDI tagged?

Architecture



A static picture is often not enough because BGP topology is fluid. To understand BGP events, conduct analysis over the events and pick out routes to map onto visualization. This is a hard problem...

BGP event rate at Berkeley during August 2003



BGP is extremely chatty.

A million events: What happened? Where is it happening? Is it ignorable? Is action necessary?

BGP talks at the prefix-level

```
1061481377.471241 W 128.32.1.3 160.124.0.0/16 ORIGIN: IGP ASPATH: 11423 209 701 1299 5713 6083
  NEXT_HOP: 128.32.0.70 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 209:888 11423:65350 11423:65352
1061481377.474808 W 128.32.1.3 207.191.1.0/24 ORIGIN: INCOMPLETE ASPATH: 11423 11422 209 4519
  NEXT_HOP: 128.32.0.66 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 11422:65350 11422:65352
1061481379.147718 A 128.32.1.200 192.103.144.0/24 ORIGIN: IGP ASPATH: 11423 209 7018 12282
  NEXT_HOP: 128.32.0.90 MED: 10 LOCAL_PREF: 70 COMMUNITIES: 209:888 209:889 11423:65350 11423:65352
1061481379.149378 W 128.32.1.200 192.96.25.0/24 ORIGIN: IGP ASPATH: 11423 209 701 1299 1299 1299 5713 608
  NEXT_HOP: 128.32.0.90 MED: 10 LOCAL_PREF: 70 COMMUNITIES: 209:888 11423:65350 11423:65352
1061481379.153554 W 128.32.1.200 212.22.132.0/23 ORIGIN: IGP ASPATH: 11423 209 1239 3228 21408
  NEXT_HOP: 128.32.0.90 MED: 10 LOCAL_PREF: 70 COMMUNITIES: 209:888 11423:65350 11423:65352
1061481379.153554 A 128.32.1.200 212.22.132.0/23 ORIGIN: IGP ASPATH: 11423 209 3356 8968 21408
  NEXT_HOP: 128.32.0.90 MED: 10 LOCAL_PREF: 70 COMMUNITIES: 209:888 11423:65350 11423:65352
1061481379.154543 A 128.32.1.200 207.77.60.0/24 ORIGIN: IGP ASPATH: 11423 209 701 705
  NEXT_HOP: 128.32.0.90 MED: 10 LOCAL_PREF: 70 COMMUNITIES: 209:888 11423:65350 11423:65352
1061481408.097538 W 128.32.1.3 207.77.60.0/24 ORIGIN: IGP ASPATH: 11423 209 701 705
  NEXT_HOP: 128.32.0.66 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 209:888 11423:65350 11423:65352
1061481408.100072 W 128.32.1.3 199.0.17.0/24 ORIGIN: IGP ASPATH: 11423 11422 209 1239 3602 7456
  NEXT_HOP: 128.32.0.66 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 11422:65350 11422:65352
1061481423.653095 W 128.32.1.3 12.2.41.0/24 ORIGIN: IGP ASPATH: 11423 209 7018 13606
  NEXT_HOP: 128.32.0.66 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 209:888 209:889 11423:65350 11423:65352
1061481423.653095 W 128.32.1.3 216.206.24.0/24 ORIGIN: IGP ASPATH: 11423 209 7018 13606
  NEXT_HOP: 128.32.0.66 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 209:888 209:889 11423:65350 11423:65352
1061481423.666821 W 128.32.1.3 195.80.160.0/19 ORIGIN: IGP ASPATH: 11423 209 1239 5400 15410 8778
  NEXT_HOP: 128.32.0.66 MED: 5 LOCAL_PREF: 80 COMMUNITIES: 209:888 11423:65350 11423:65352
```

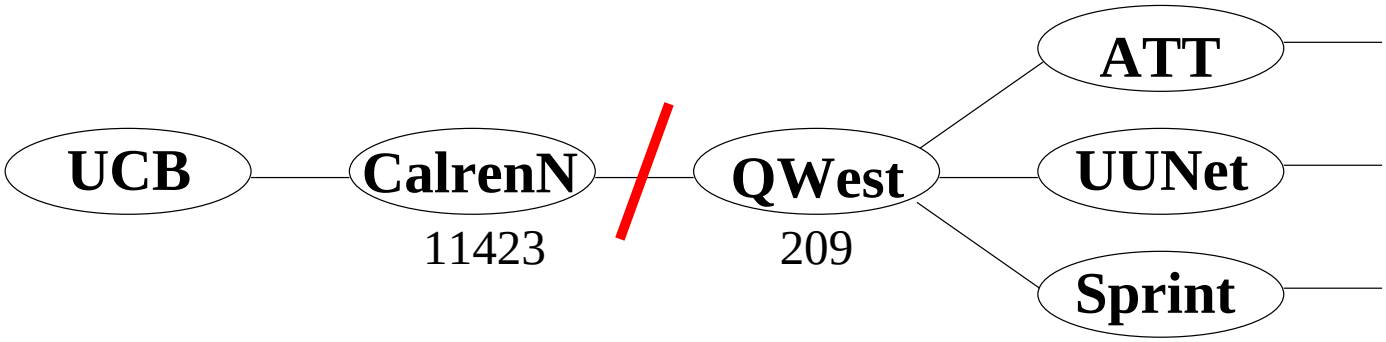
A million events does not mean a million different things happened. BGP cannot say the actual incident, e.g. peering down or flaky router or prefix loss. It can only tell you indirectly via prefix withdrawals & announcements.

Extracting correlation with Stemming analysis

```
W 128.32.1.3   NEXT_HOP: 128.32.0.70 ASPATH: 11423 209 701 1299 5713
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 11422 209 4519
W 128.32.1.200 NEXT_HOP: 128.32.0.90 ASPATH: 11423 209 701 1299 5713
W 128.32.1.200 NEXT_HOP: 128.32.0.90 ASPATH: 11423 209 1239 3228 21408
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 209 701 705
...
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 11422 209 1239 3602
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 209 7018 13606
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 209 7018 13606
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 209 1239 5400 15410
W 128.32.1.3   NEXT_HOP: 128.32.0.66 ASPATH: 11423 209 1239 5400 15410
```

Most of the withdrawals have 11423-209 in their AS paths. Why?

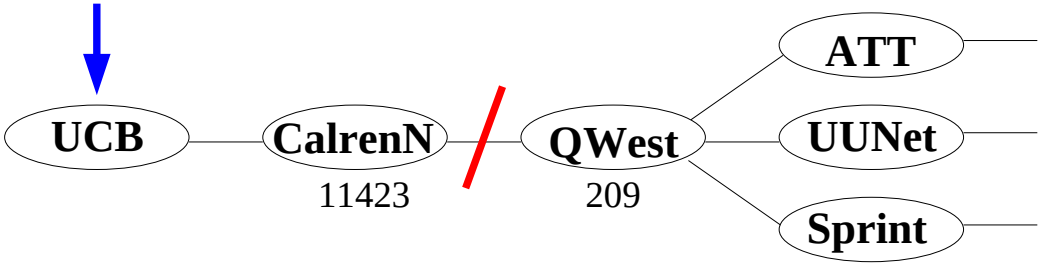
Extracting correlation with Stemming analysis (cont'd)



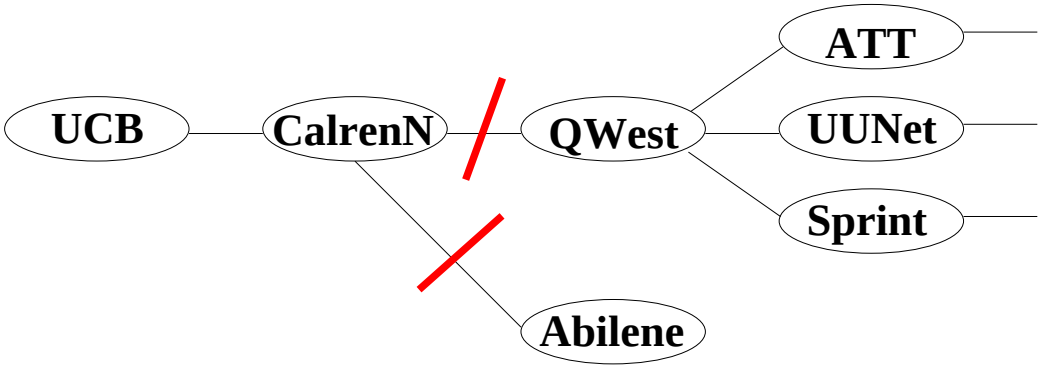
Because a branch of the original routing tree is cut. The common portion tells you where the cut is, which is at the last edge. E.g. CalrenN-QWest.

We call the common portion a *stem*.

Stemming challenges



Finding the stem is not trivial. It is not a simple correlation problem as events are all correlated at the root. E.g. UCB.



Must also distinguish among multiple simultaneous failures: need to find all of them and pull them apart. E.g. failure on CalrenN-QWest vs failure on CalrenN-Abilene.

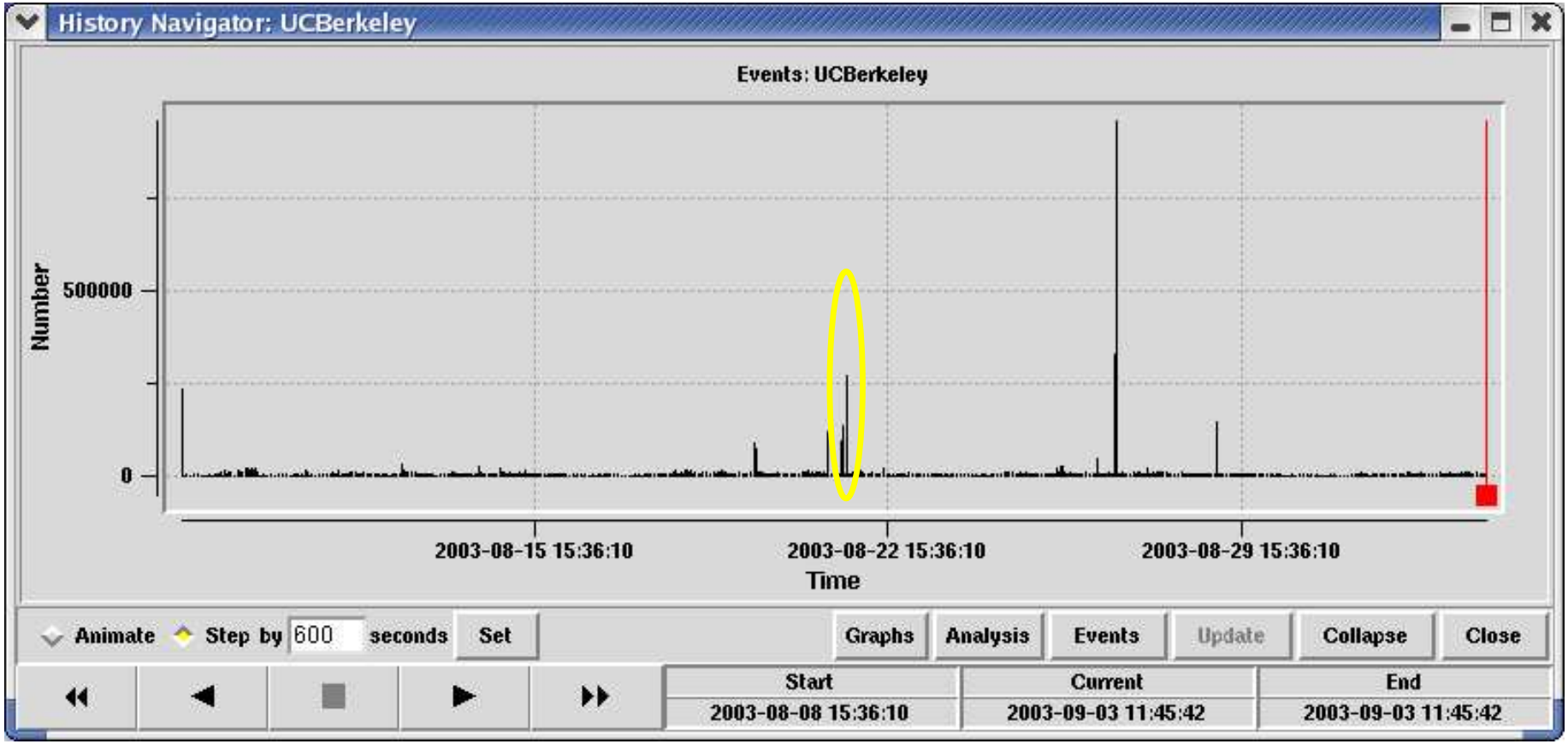
Stemming challenges (cont'd)

Common problem in statistics, usually solved using hierarchical clustering, PCA or MDS, but computationally expensive – $O(N^2)$ or $O(N \log N)$, where N is number of possible path stems.

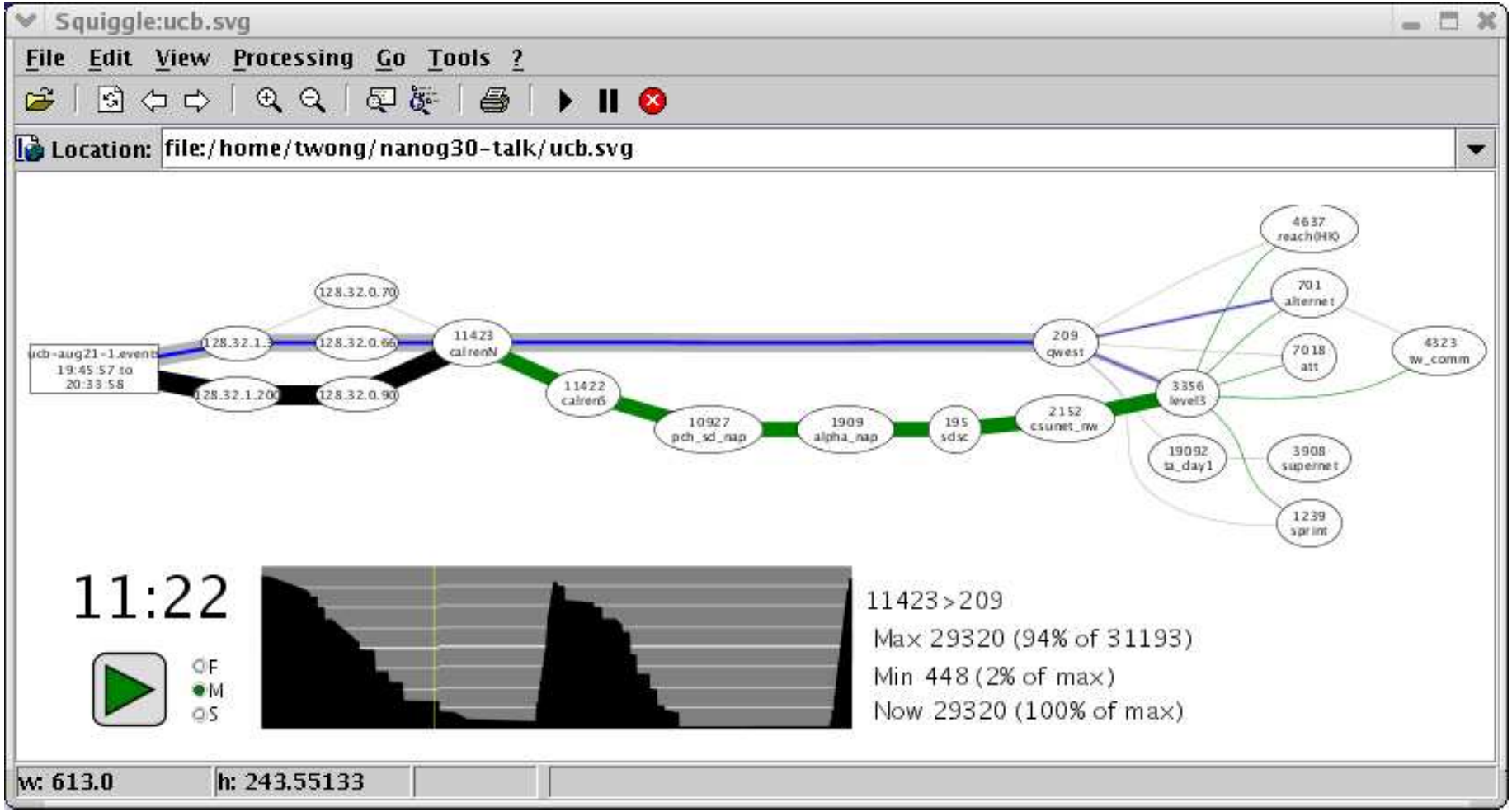
We have developed a linear time algorithm call *Stemming* that exploits AS path structures. Runs real-time on a modern processor.

Works on tree-like topologies such as Berkeley's, and complex, forest-like ones of Tier-1 providers.

BGP event rate at Berkeley during August 2003
What happened there?



Very long backup path and community mishap at Berkeley



Note: this is an animation

Go to: http://www.packetdesign.com/technology/presentations/nanog-30/index.htm

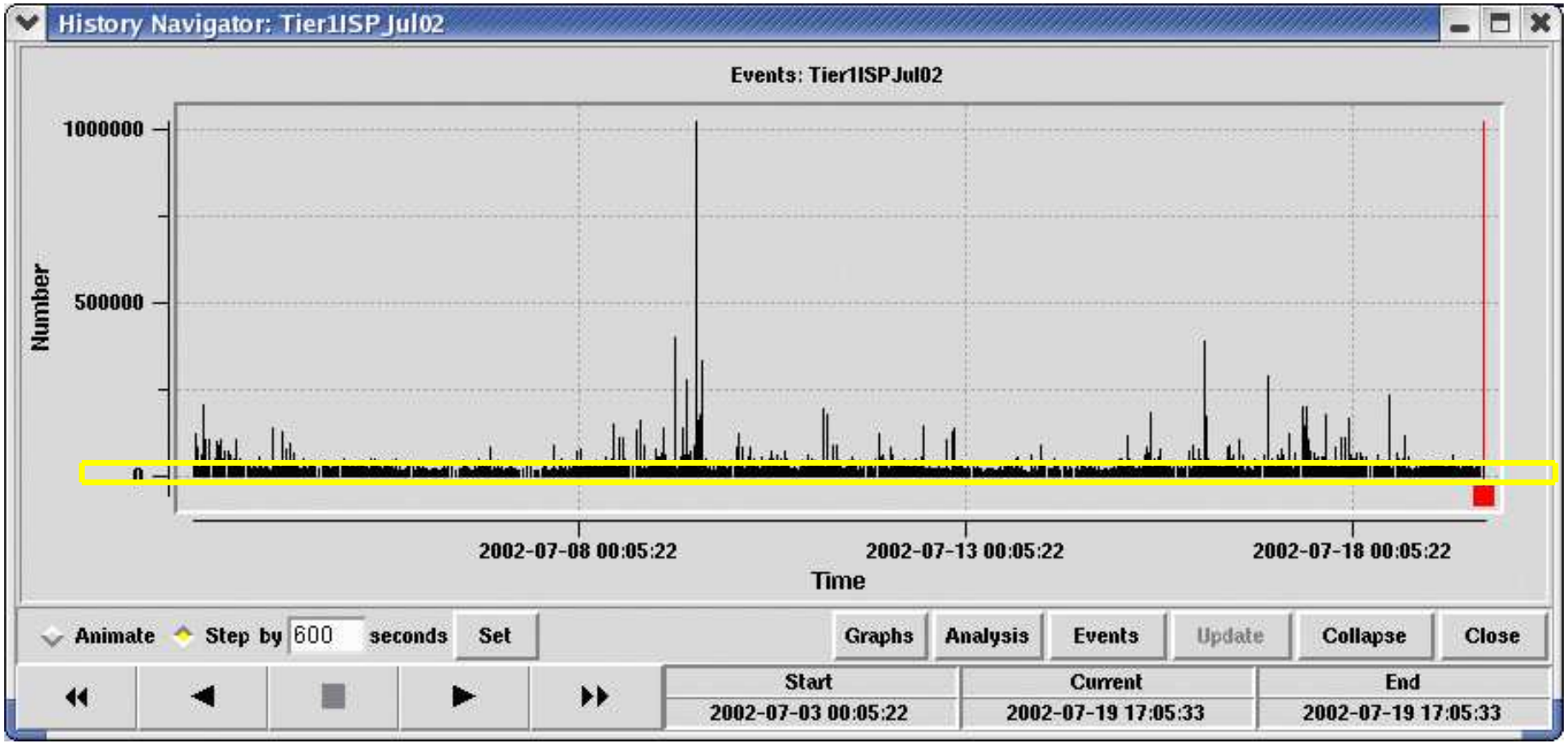
Temporal independence in Stemming

Many serious problems often do not result in distinctive event spikes. For example, persistent route oscillations. Can Stemming help?

Yes. Stemming is not just about explaining event spikes.

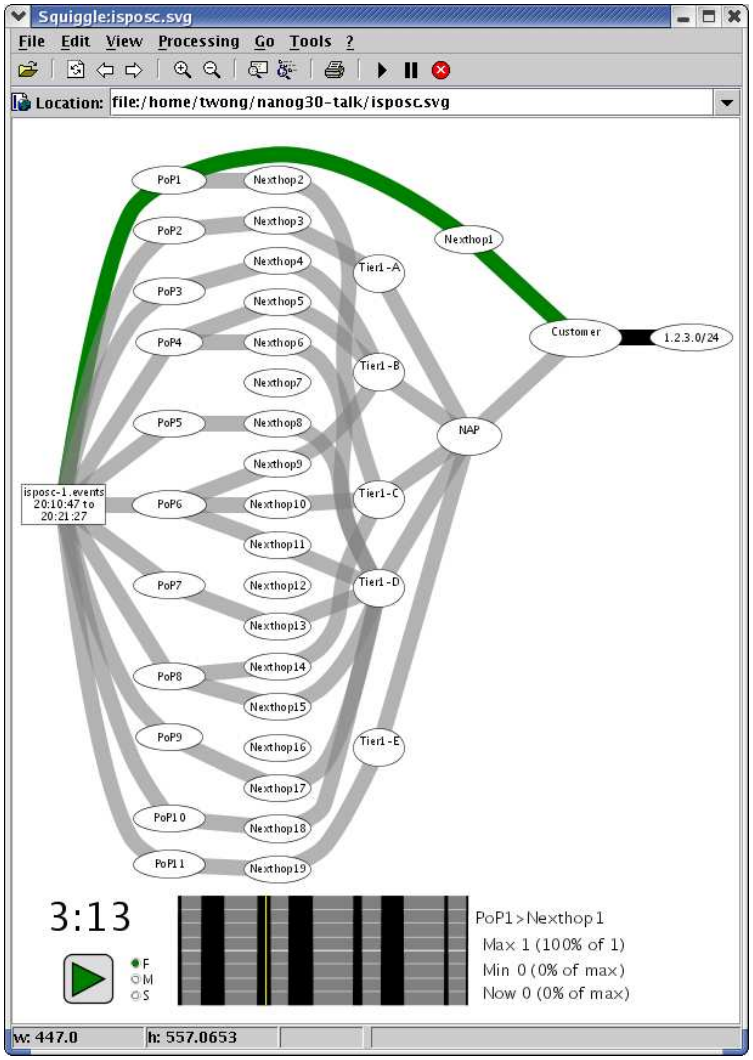
One of Stemming's important characteristics is temporal independence. Correlation is a well-defined property at any time-scale. By looking at a long enough time period, a continuous flapping on even a single prefix would overwhelm other correlations.

BGP event rate at a US Tier-1 ISP during July 2002



Persistent route flapping once a minute *looks like* BGP noise here (yellow box). This lasted for at least 1.5 months, and resulted in a very unhappy customer.

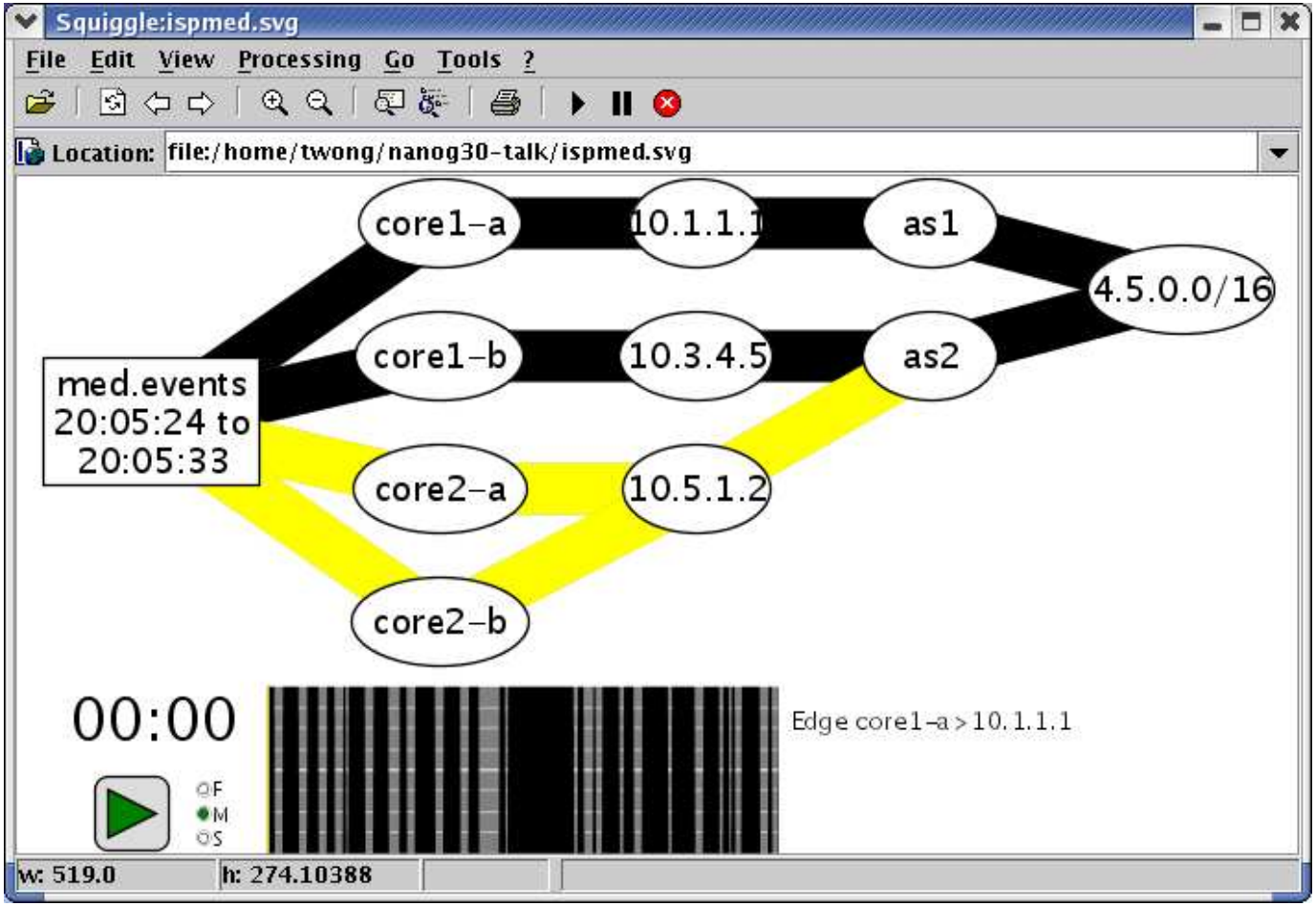
Persistent customer route flapping at Tier1 ISP



Note: this is an animation

Go to: http://www.packetdesign.com/technology/presentations/nanog-30/index.htm

Persistent MED oscillation at Tier1 ISP



Note: this is an animation

Go to: http://www.packetdesign.com/technology/presentations/nanog-30/index.htm

Output of root-cause analysis

```
aug21 19:45:57 BGP 128.32.1.3 lost 28,293 prefixes:
calrenN>qwest peering failed over to
calrenN>calrenS>pch_sd_nap>alpha_nap>sdsc>cenic_dc>level3

aug21 20:05:25 BGP 128.32.1.3 gained 28,293 prefixes:
calrenN>qwest peering restored
```

Acknowledgments

Thank you, Ken Lindahl and UC Berkeley NetOp.

Thank you, anonymous Tier1 ISP.

Questions?

http://www.packetdesign.com/technology/presentations/nanog-30/index.htm