Is the BGP Sky Falling?

Geoff Huston

APNIC

May 2009

Conventional BGP Wisdom

IAB Workshop on Inter-Domain routing in October 2006 – RFC 4984:

> **“routing scalability is the most important problem facing the Internet today and must be solved”**

BGP measurements

There are a number of ways to "measure" BGP:

1. Assemble a large set of BGP peering sessions and record everything
   - RIPE NCC's RIS service
   - Route Views
2. Perform carefully controlled injections of route information and observe the propagation of information
   - Beacons
   - AS Set manipulation
   - Bogon Detection and Triangulation
3. Take a single BGP perspective and perform continuous recording of a number of BGP metrics over a long baseline

BGP measurements

There are a number of ways to "measure" BGP:

1. Assemble a large set of BGP peering sessions and record everything
   - RIPE NCC's RIS service
   - Route Views
2. Perform carefully controlled injections of route information and observe the propagation of information
   - Beacons
   - AS Set manipulation
   - Bogon Detection and Triangulation
3. Take a single BGP perspective and perform continuous recording of a number of BGP metrics over a long baseline

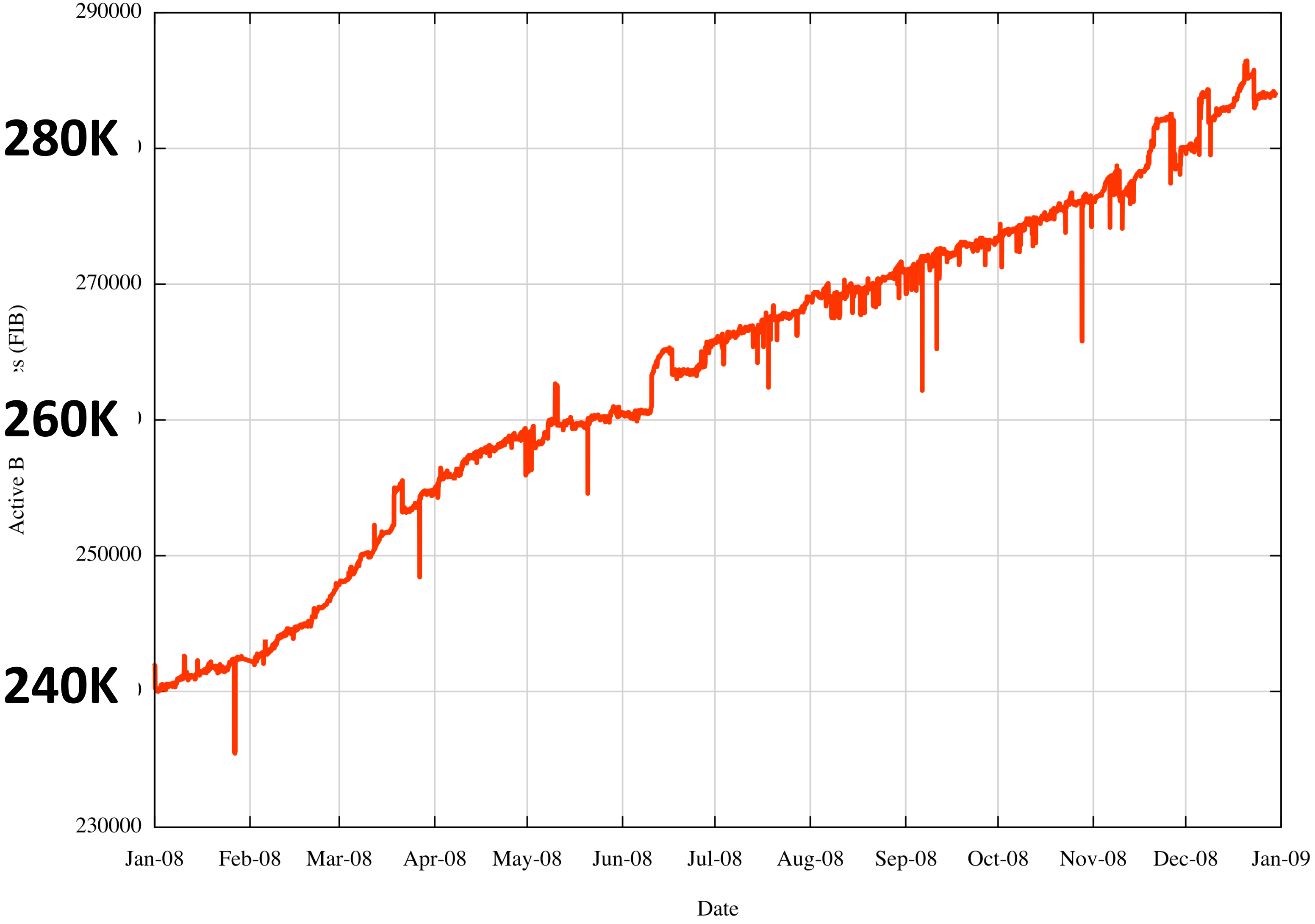
AS131072 (or AS2.0) BGP measurement

- Successor to the AS1221 observation point
- Data collection since 1 July 2007 (since 2000 for AS1221)
- Passive data measurement technique (no advertisements or probes)
- Quagga platform, connected to AS4608 and AS4777 via eBGP
- IPv4 and IPv6 simultaneous
- Archive of all BGP updates and daily RIB dumps
- Data and reports are continuously updated and published: http://bgp.potaroo.net

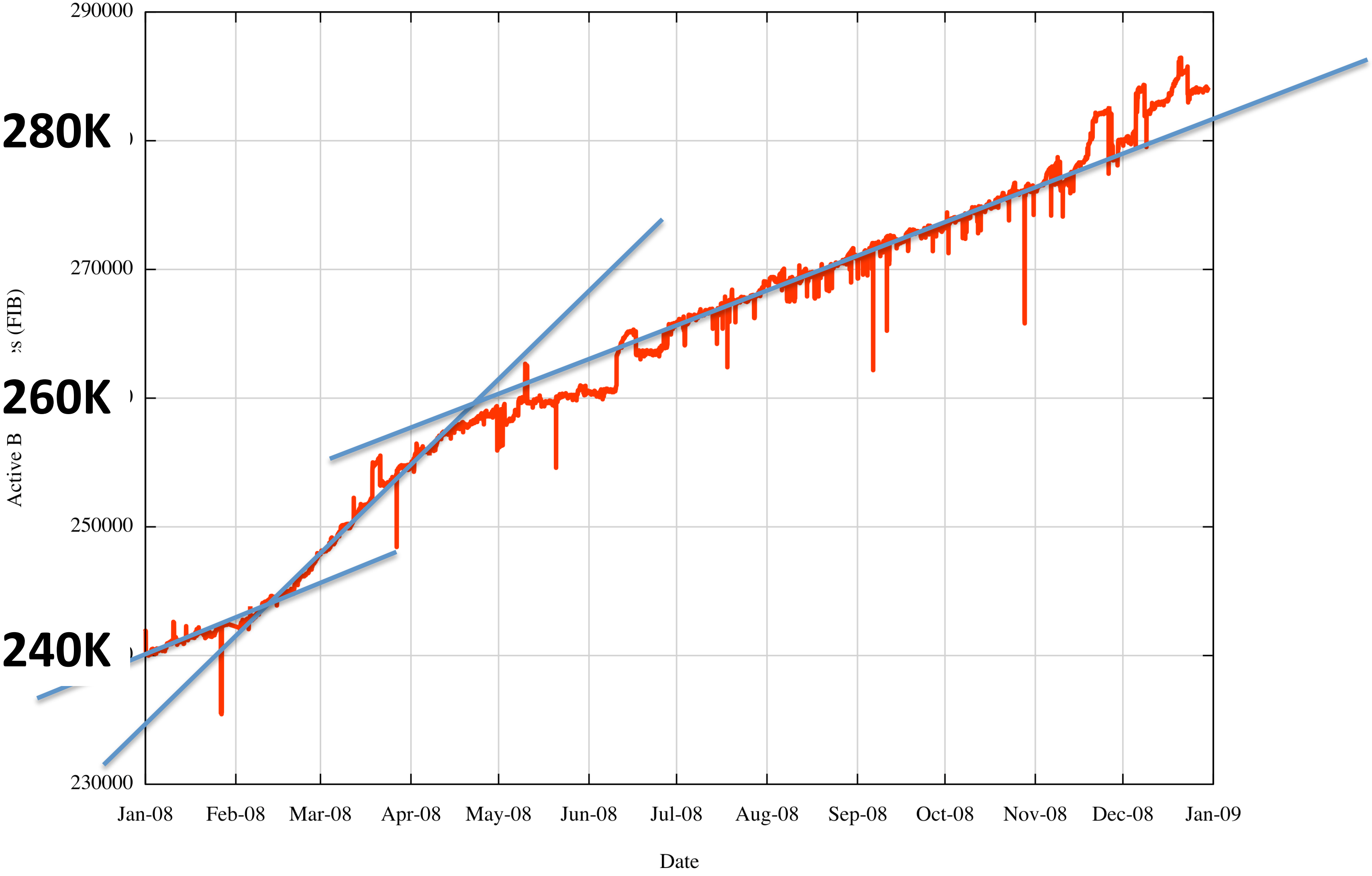
Some Caveats

- This is a measurement at the EDGE, not in the MIDDLE
- It is a single stream measurement, not an aggregated measurement
- This is a measurement of the ‘production network’ used for forwarding traffic
- There is NO iBGP traffic being measured
- This is what an eBGP customer may see in terms of load for a single eBGP feed

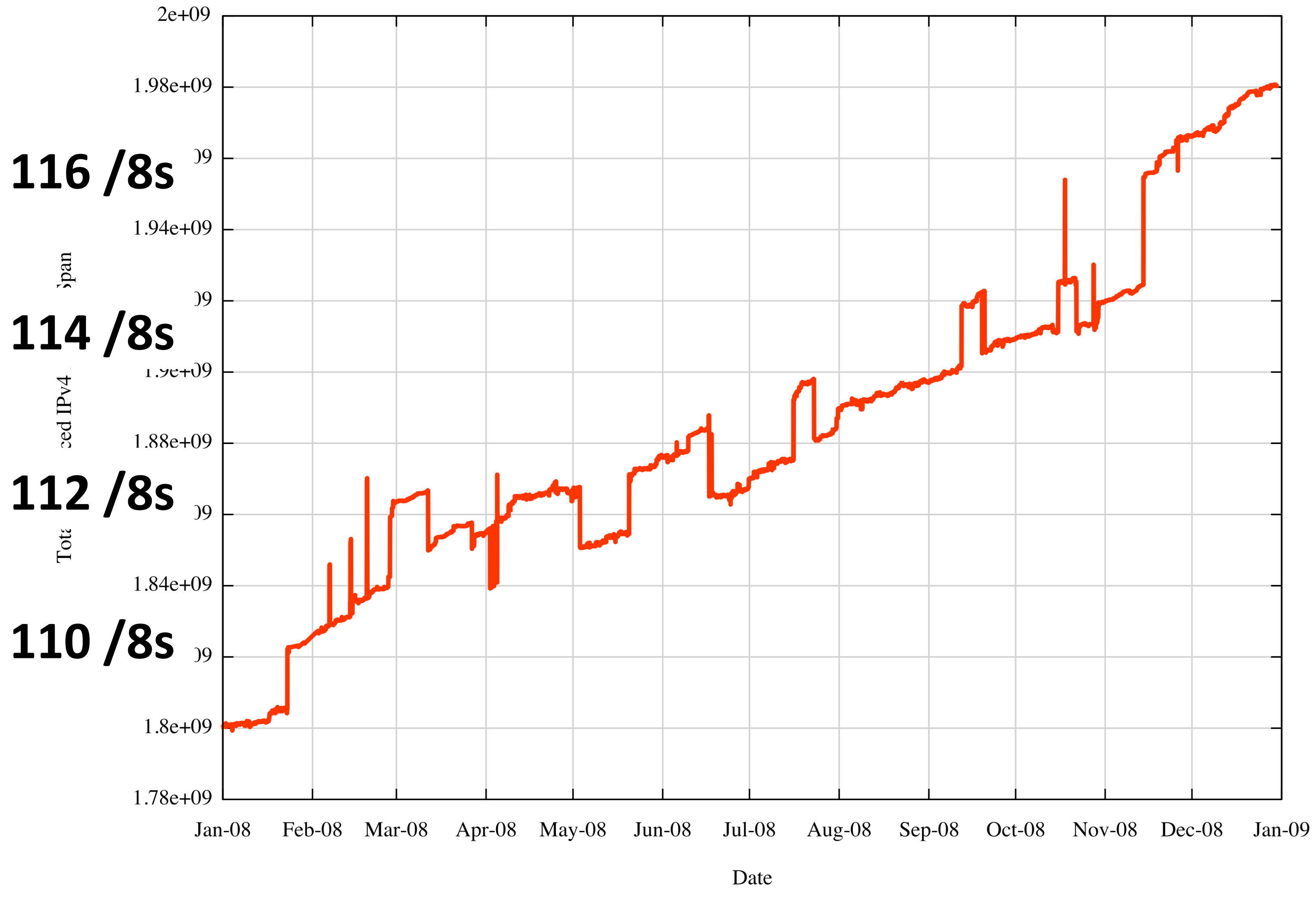
IPv4 BGP Prefix Count



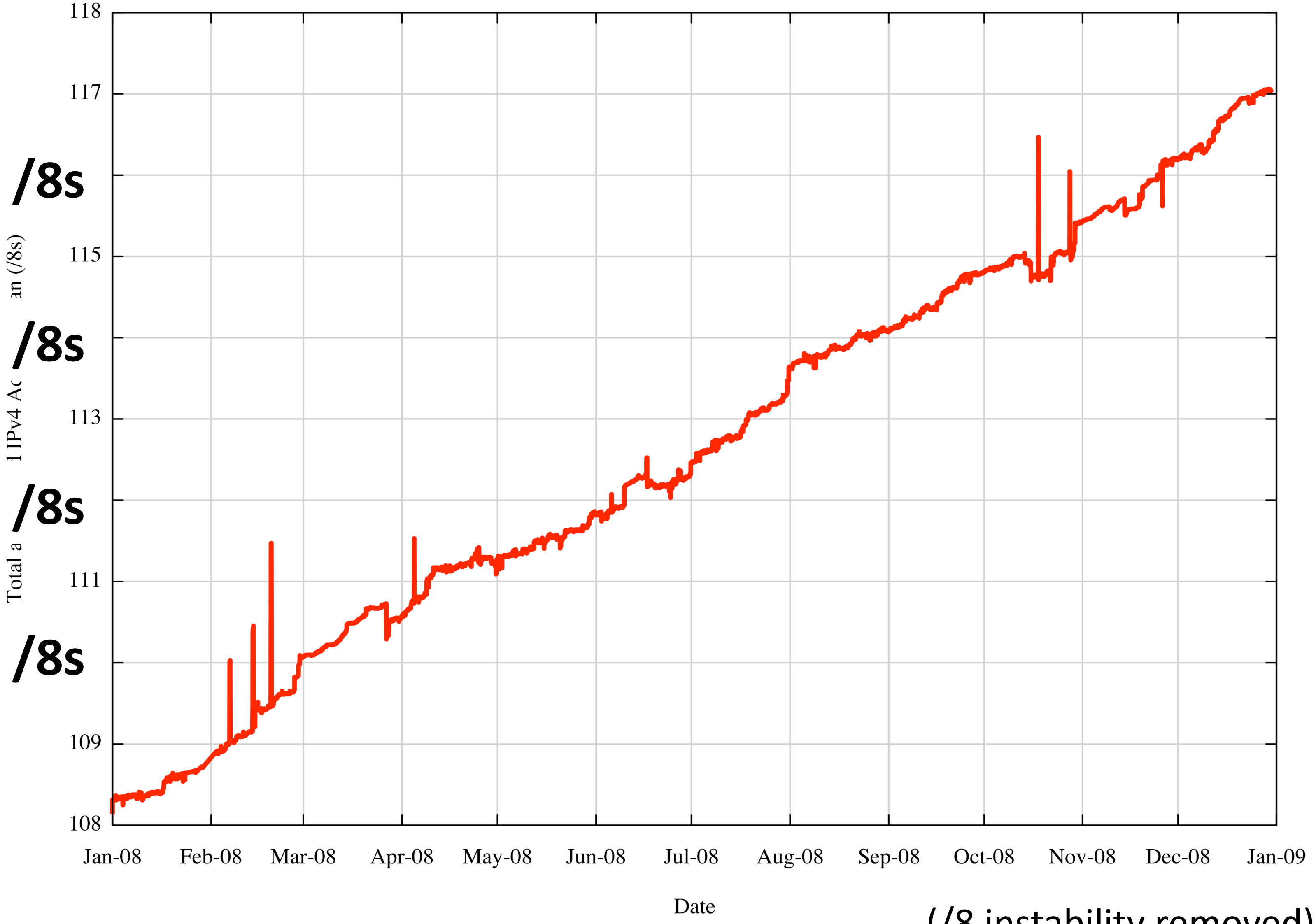
IPv4 BGP Prefix Count



IPv4 Routed Address Span



IPv4 Routed Address Span



116 /8s

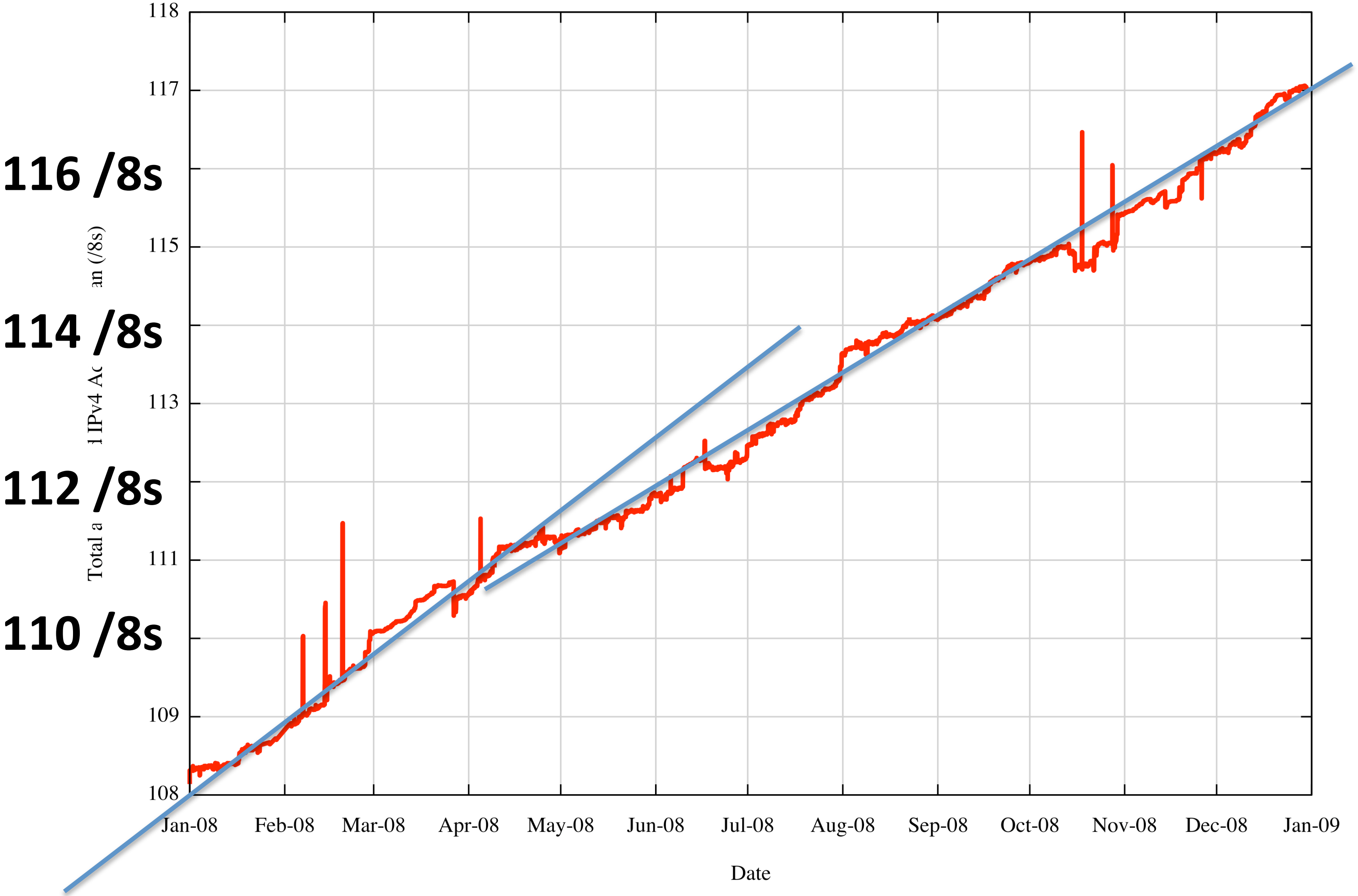
114 /8s

112 /8s

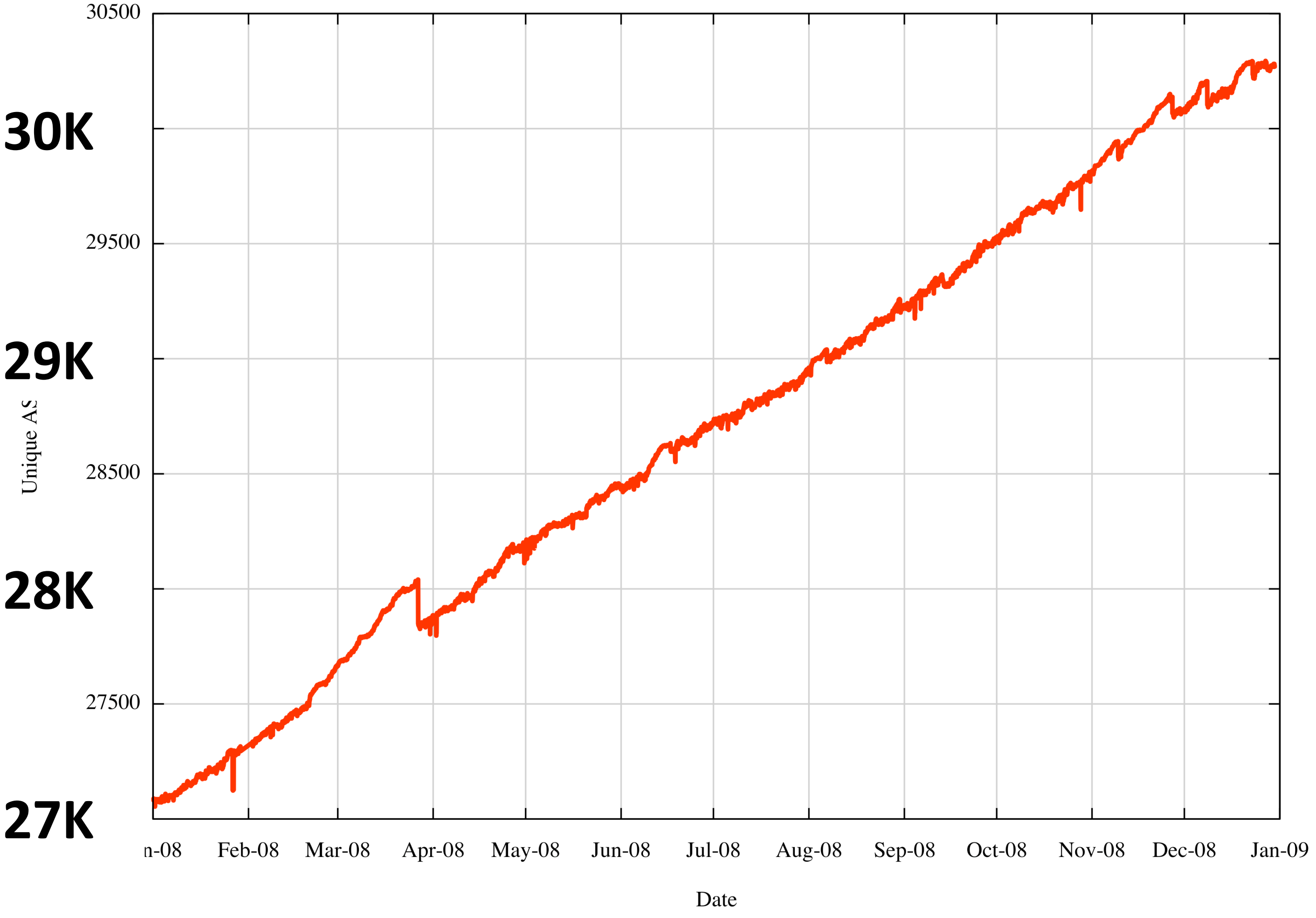
110 /8s

(/8 instability removed)

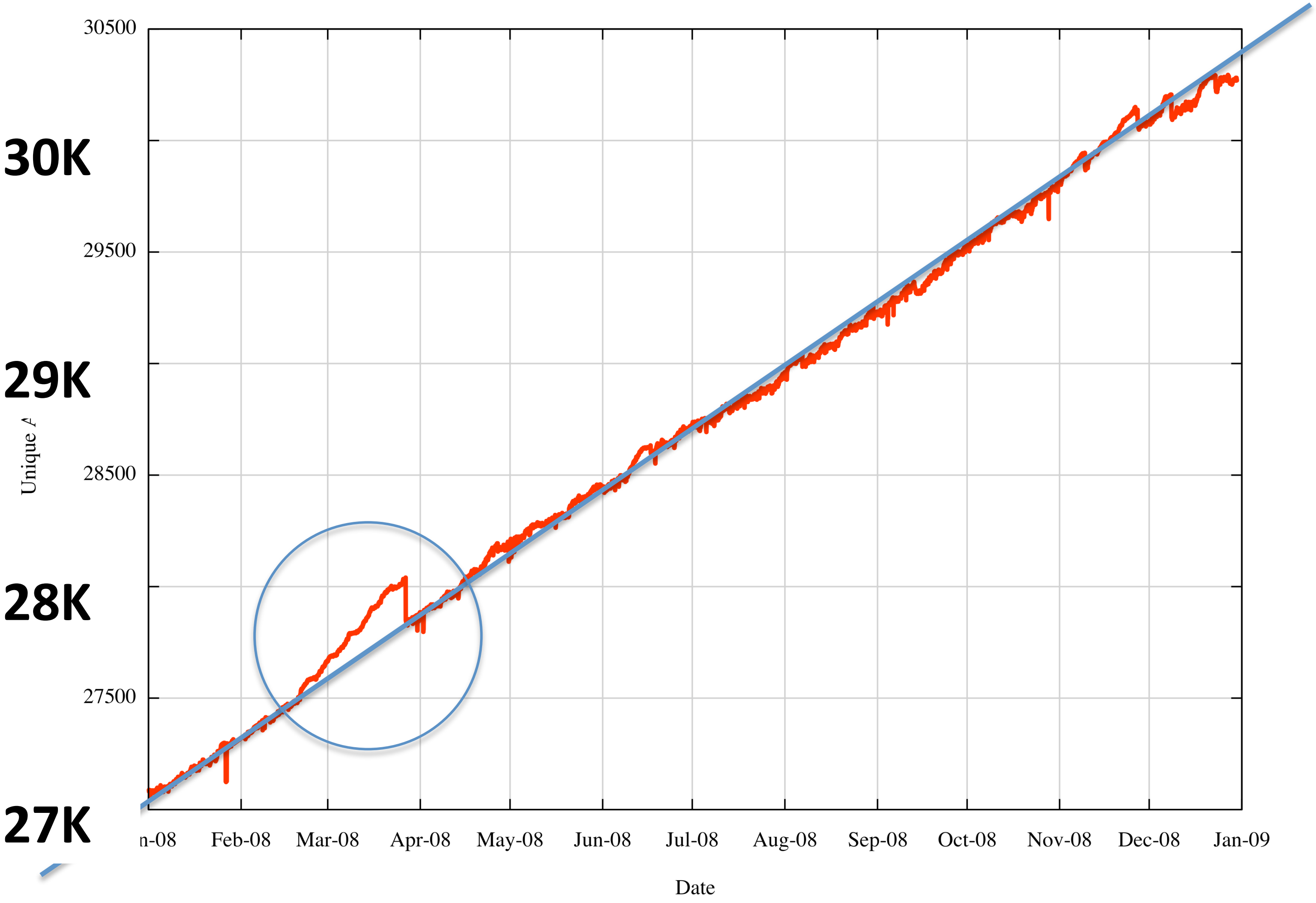
IPv4 Routed Address Span



IPv4 Routed AS Count



IPv4 Routed AS Count



IPv4 Vital Statistics for 2008

	Jan-08	Dec-08	
Prefix Count	245,000	286,000	**+17%**
Roots	118,000	133,000	+13%
More Specifics	127,000	152,000	+20%
Address Span	106.39	118.44	**+11%**
AS Count	27,000	30,300	**+11%**
Transit	3,600	4,100	+14%
Stub	23,400	26,200	+11%

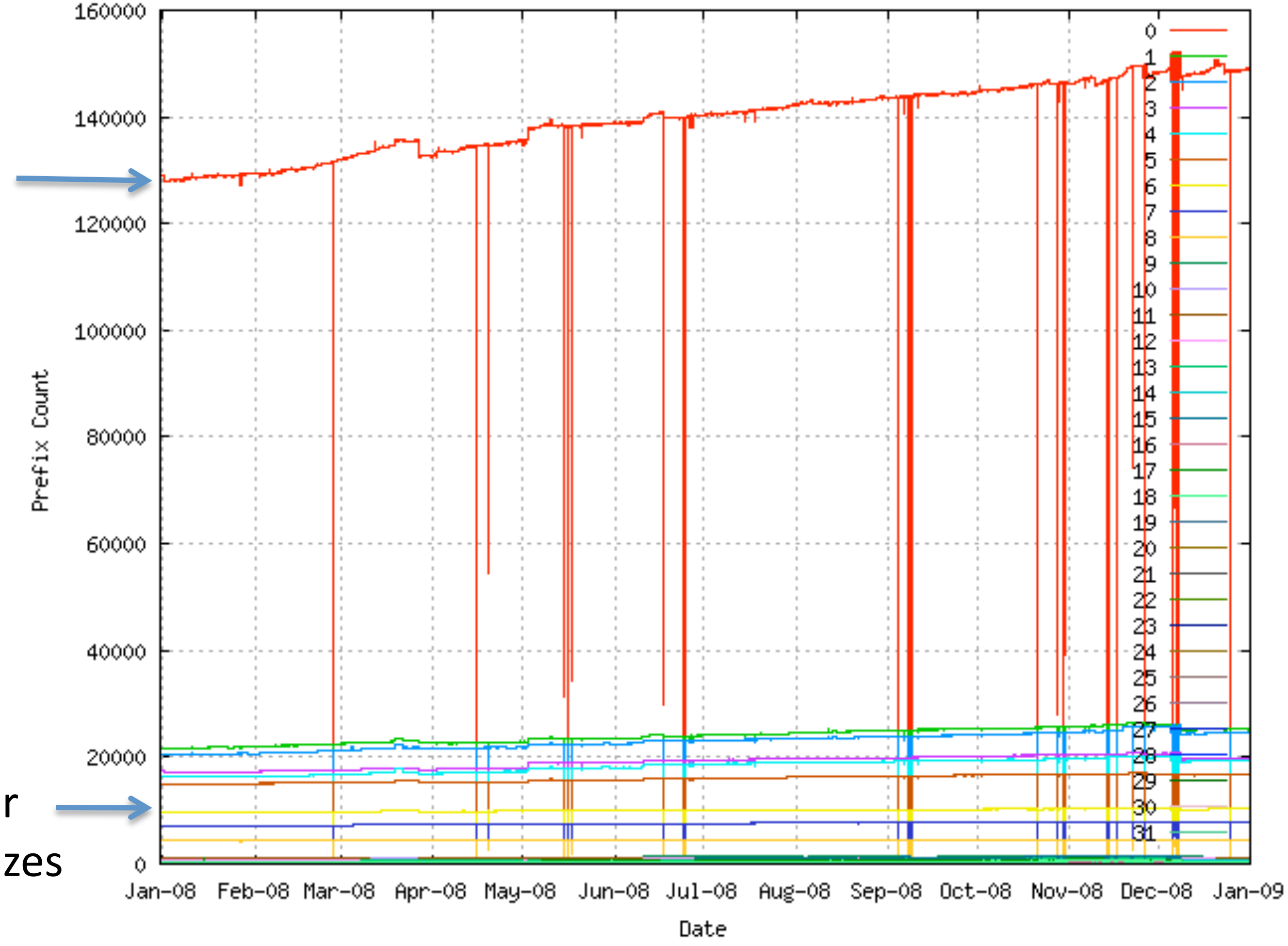
Some Observations

- Growth in IPv4 deployment slowed considerably as of the end of April 2008
  - Is this a possible consequence of the financial crash of 2008?
- Fragmentation of the IPv4 routing space continues to grow at a faster pace than underlying growth of the network itself

IPv4 prefix distribution

- Its all about /24’s

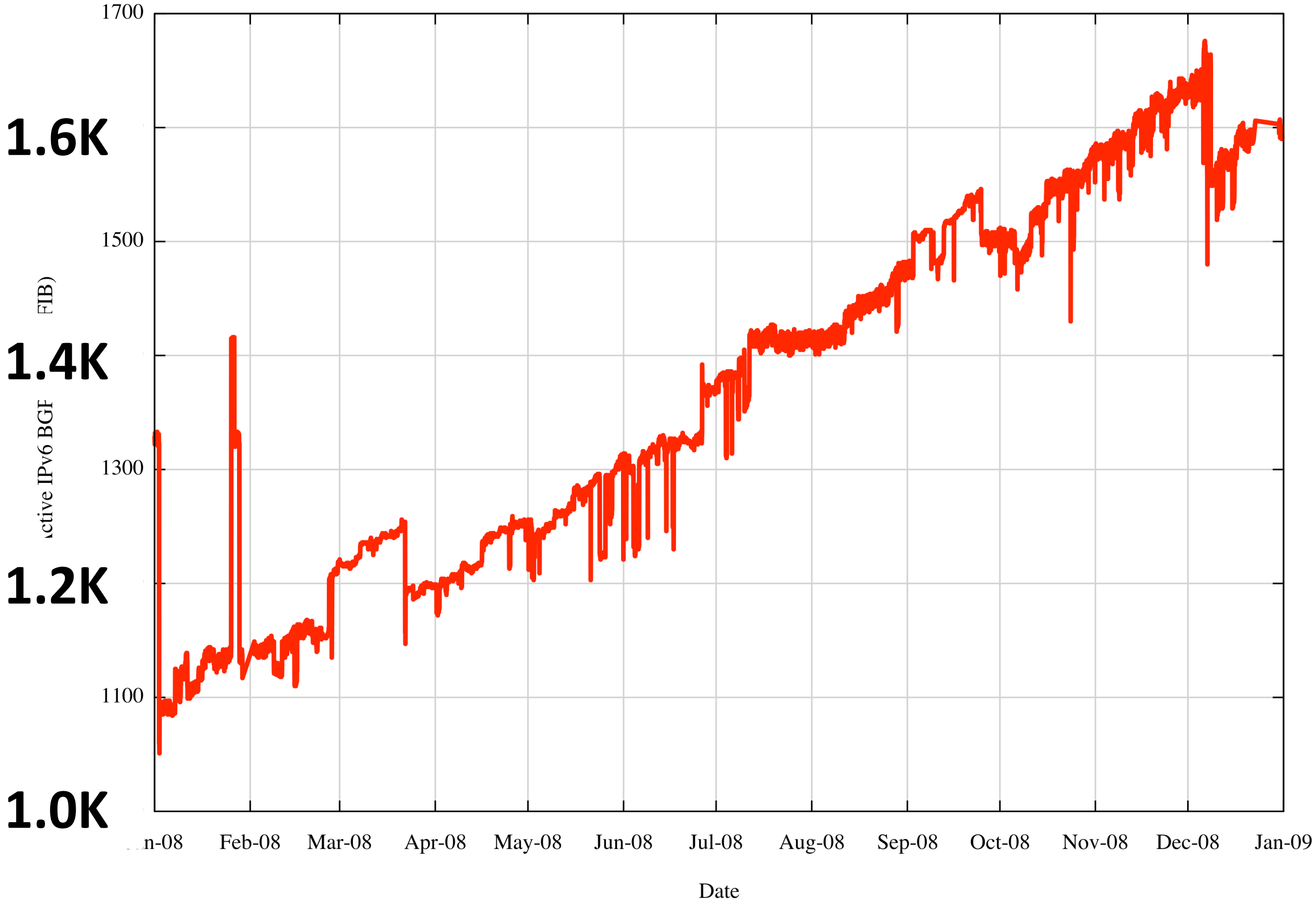
IPv4 prefix distribution



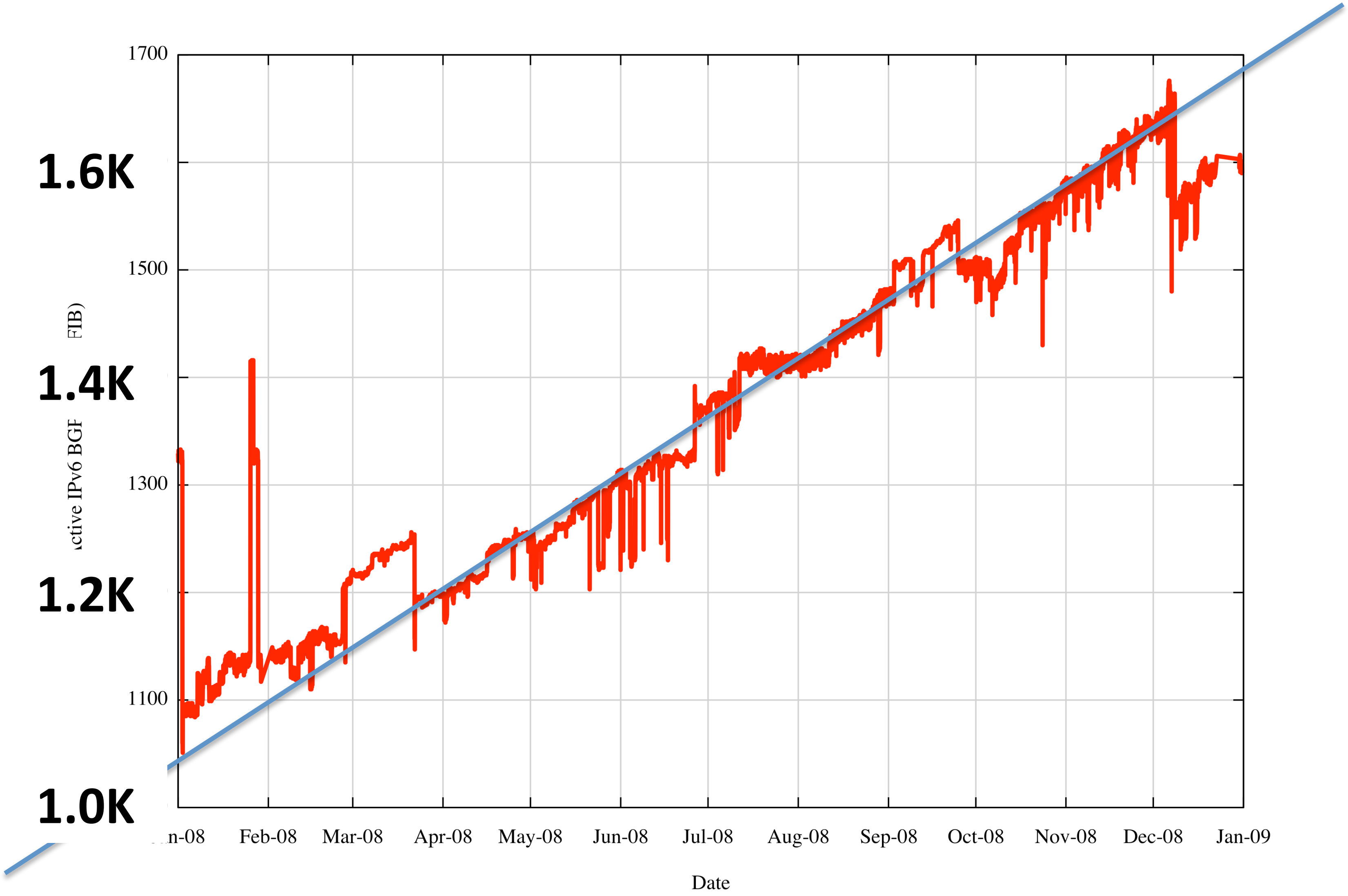
/24s

All other prefix sizes

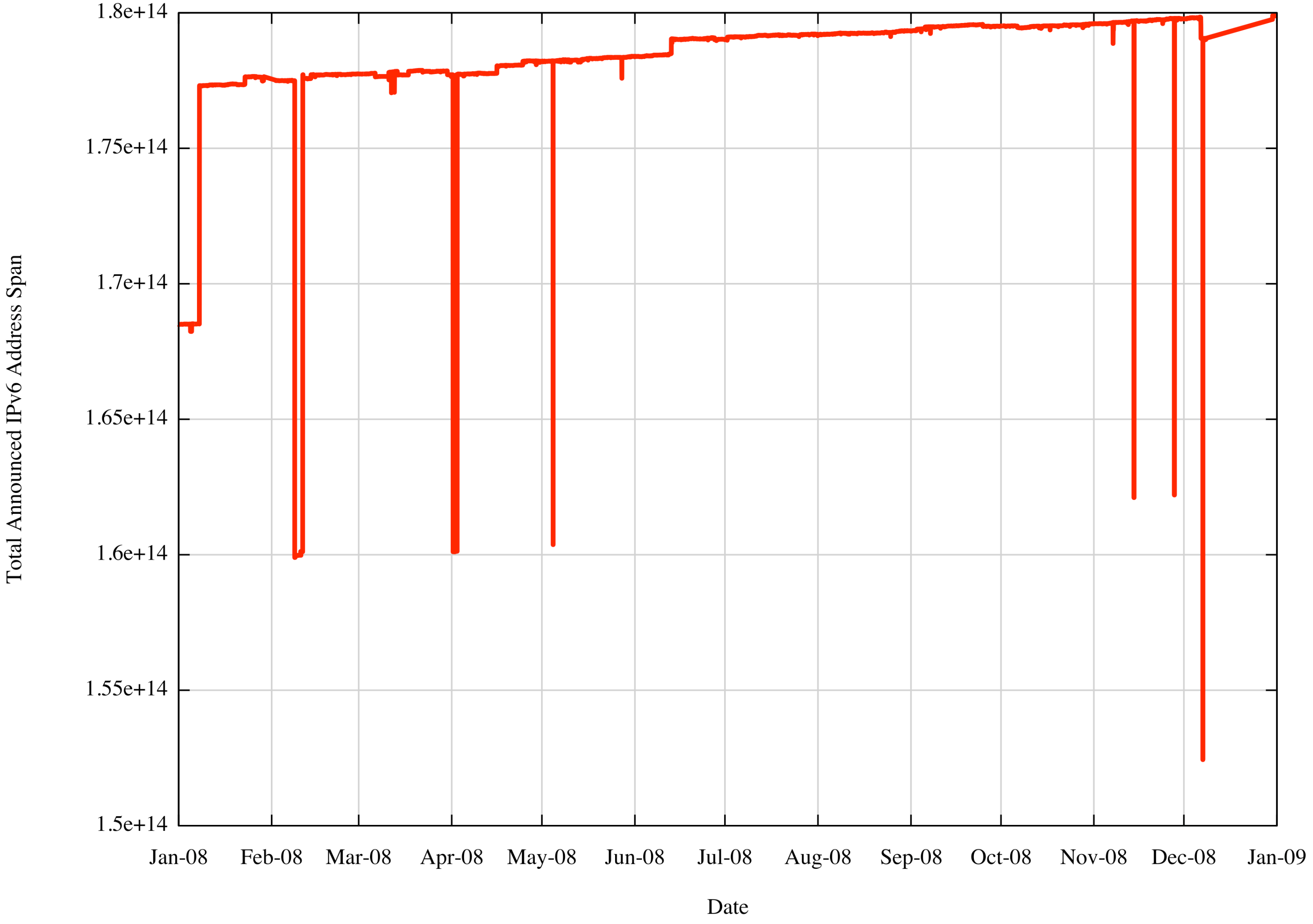
IPv6 BGP Prefix Count



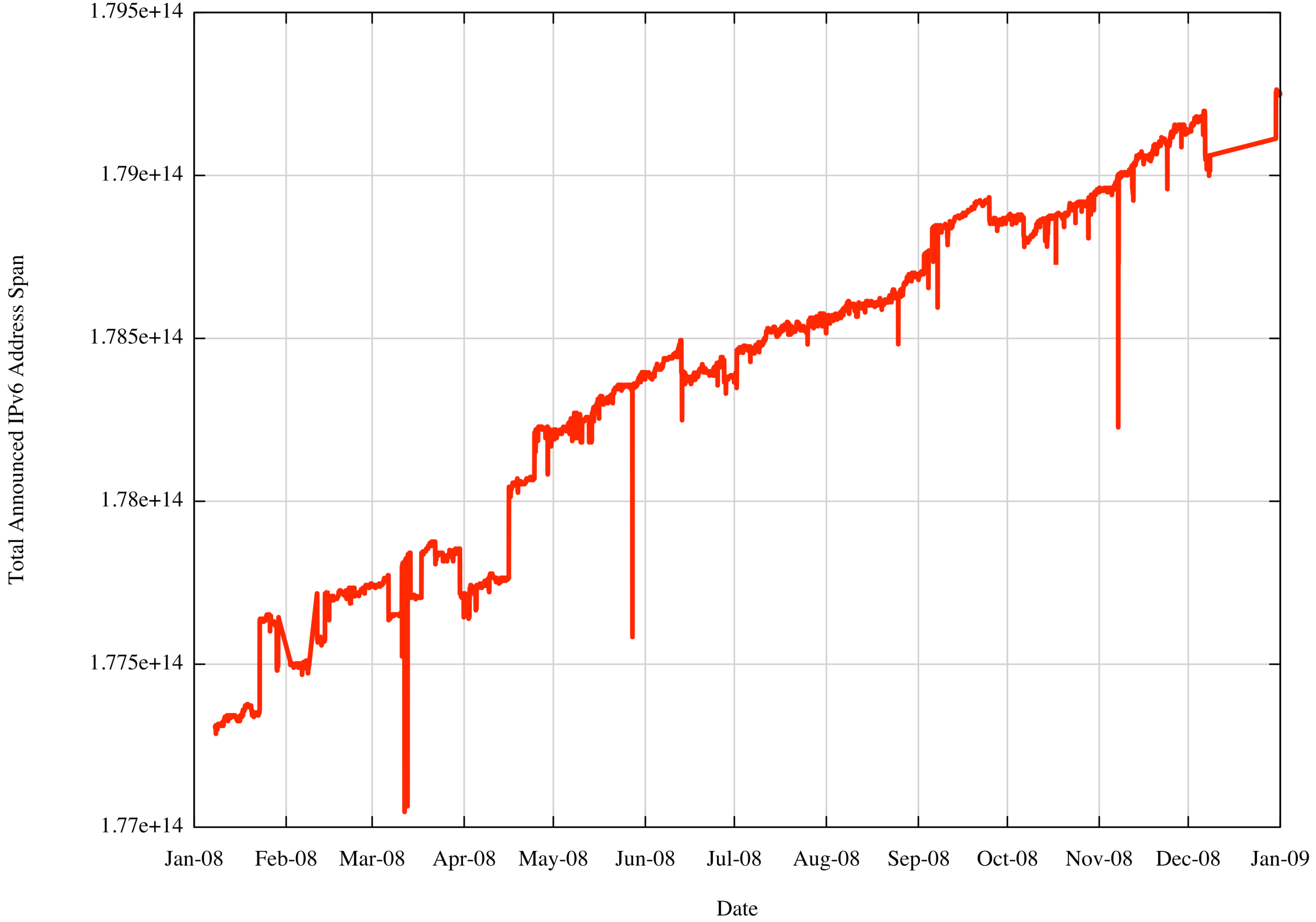
IPv6 BGP Prefix Count



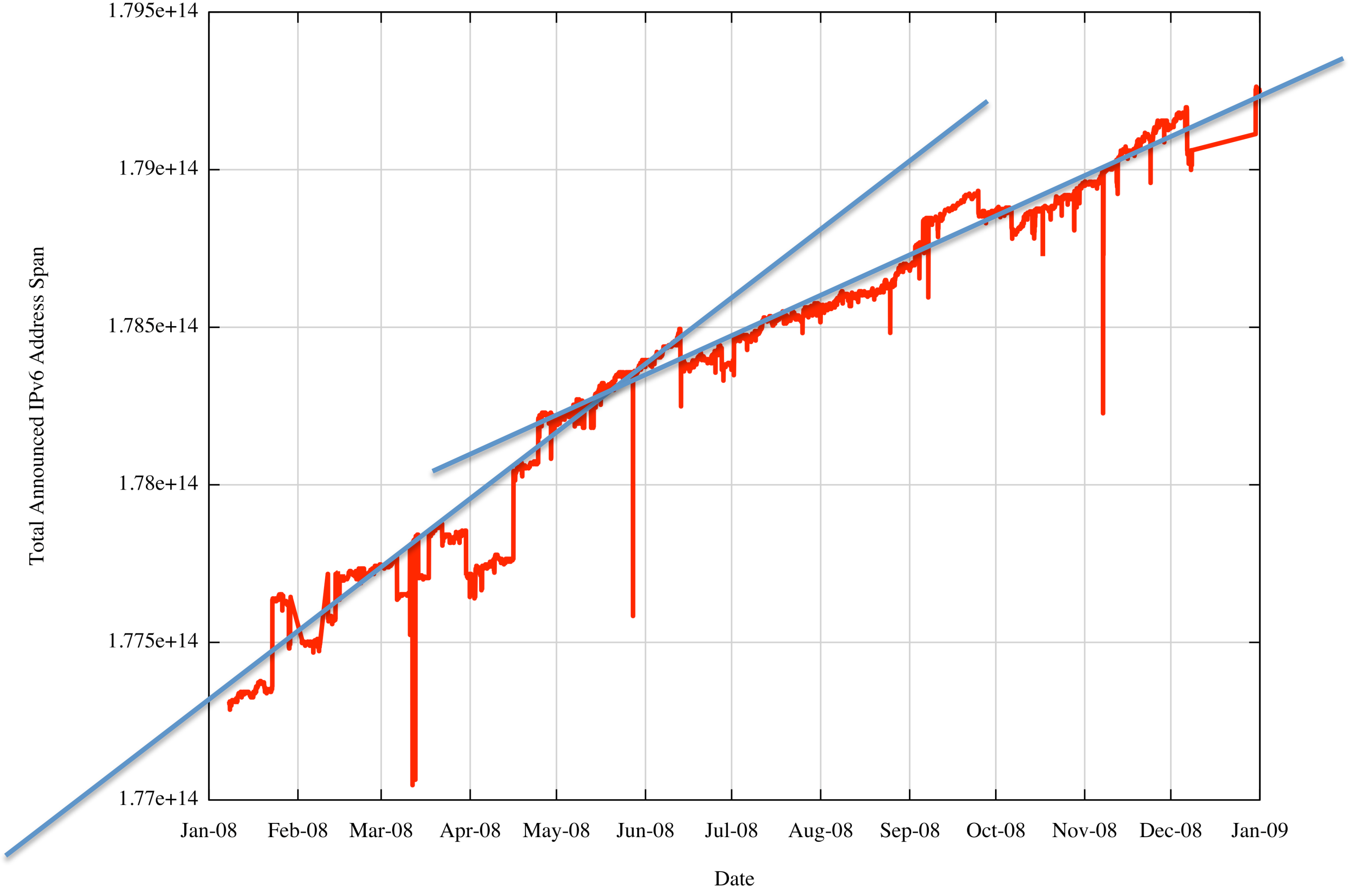
IPv6 Routed Address Span



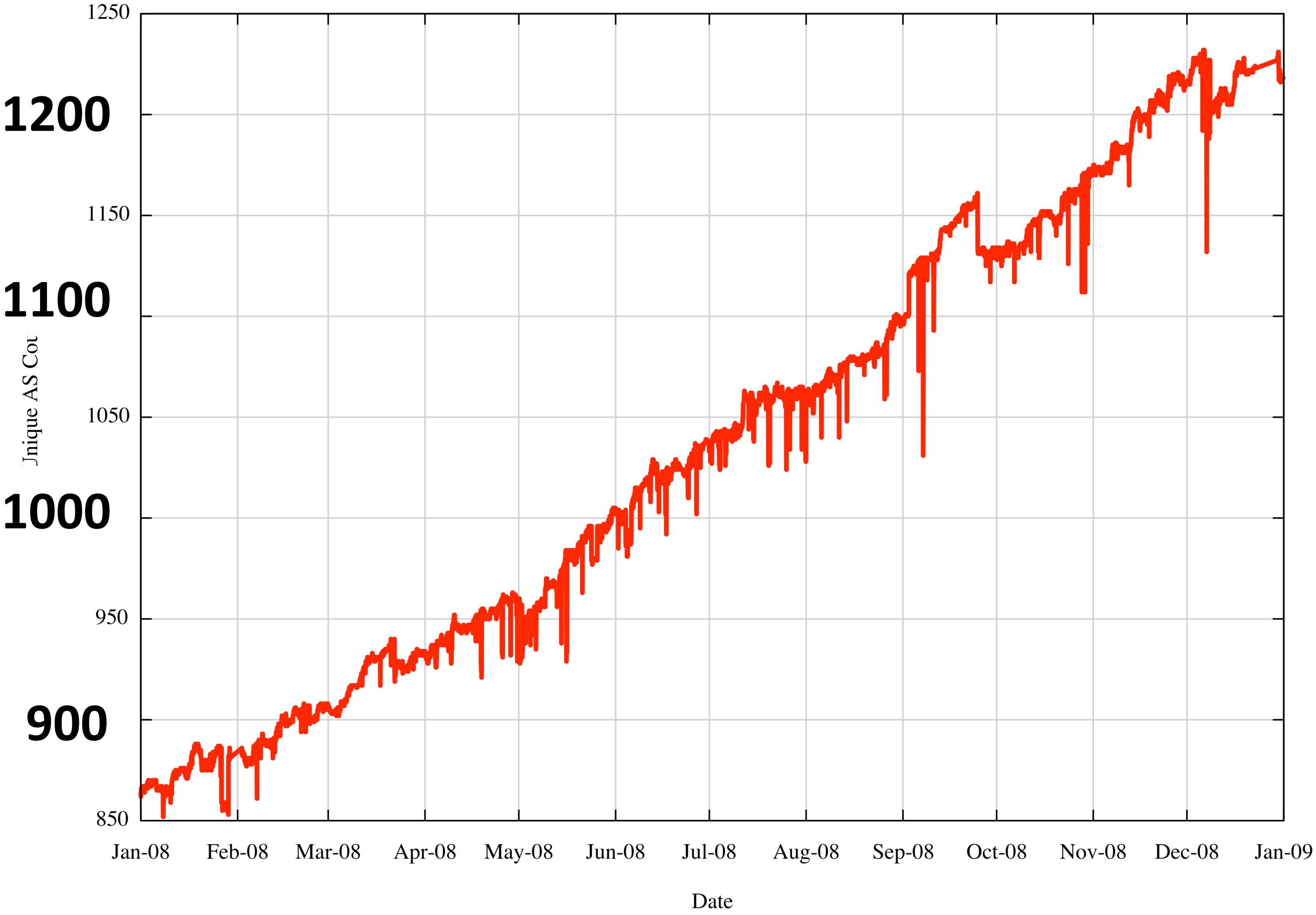
IPv6 Routed Address Span



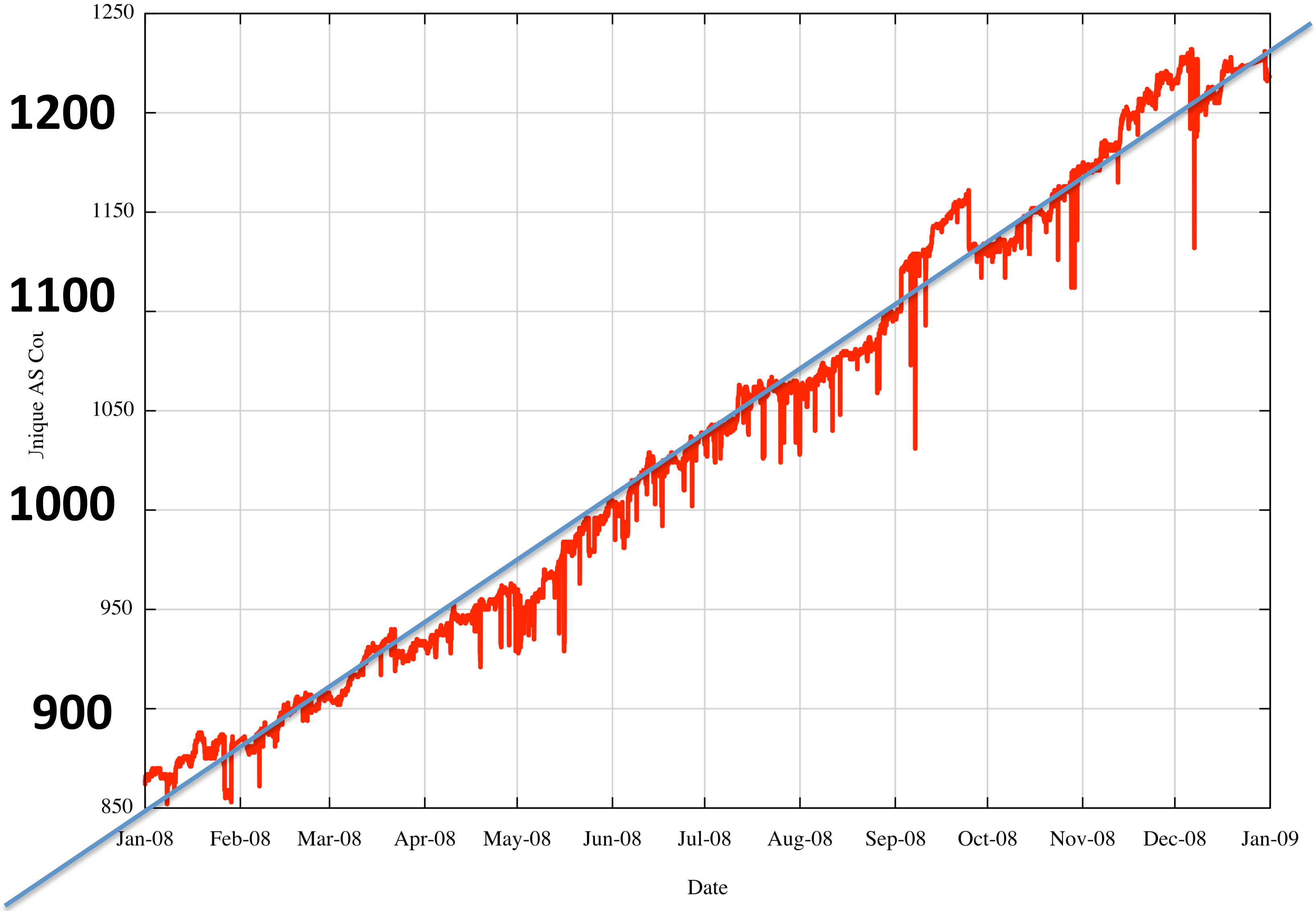
IPv6 Routed Address Span



IPv6 Routed AS Count



IPv6 Routed AS Count



IPv6 Vital Statistics for 2008

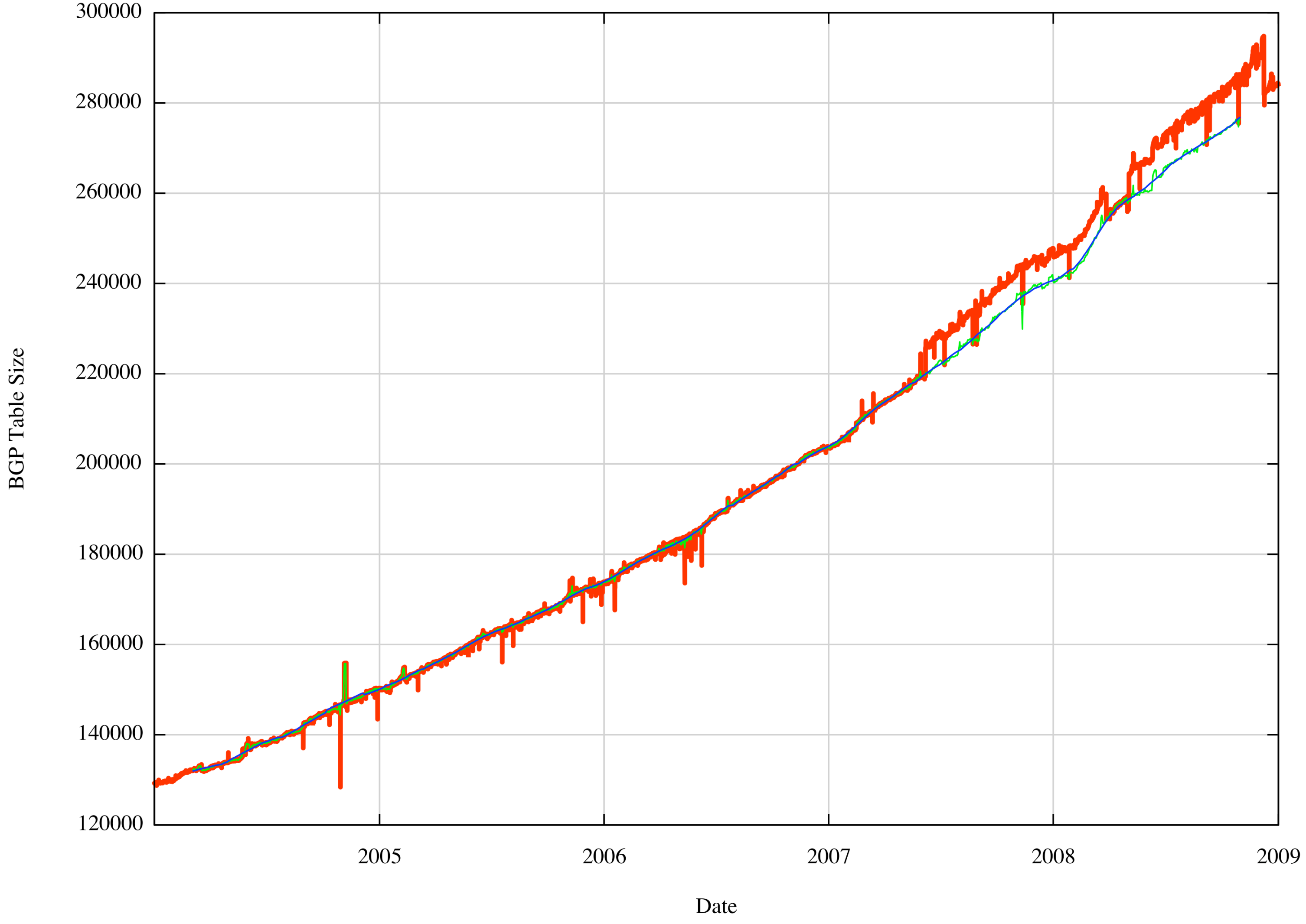
	Jan-08	Dec-08	
Prefix Count	1,050	1,600	**52%**
Roots	840	1,300	55%
More Specifics	210	300	43%
Address Span	/16.67	/16.65	**1%**
AS Count	860	1,230	**43%**
Transit	240	310	29%
Stub	620	920	48%

BGP Projections

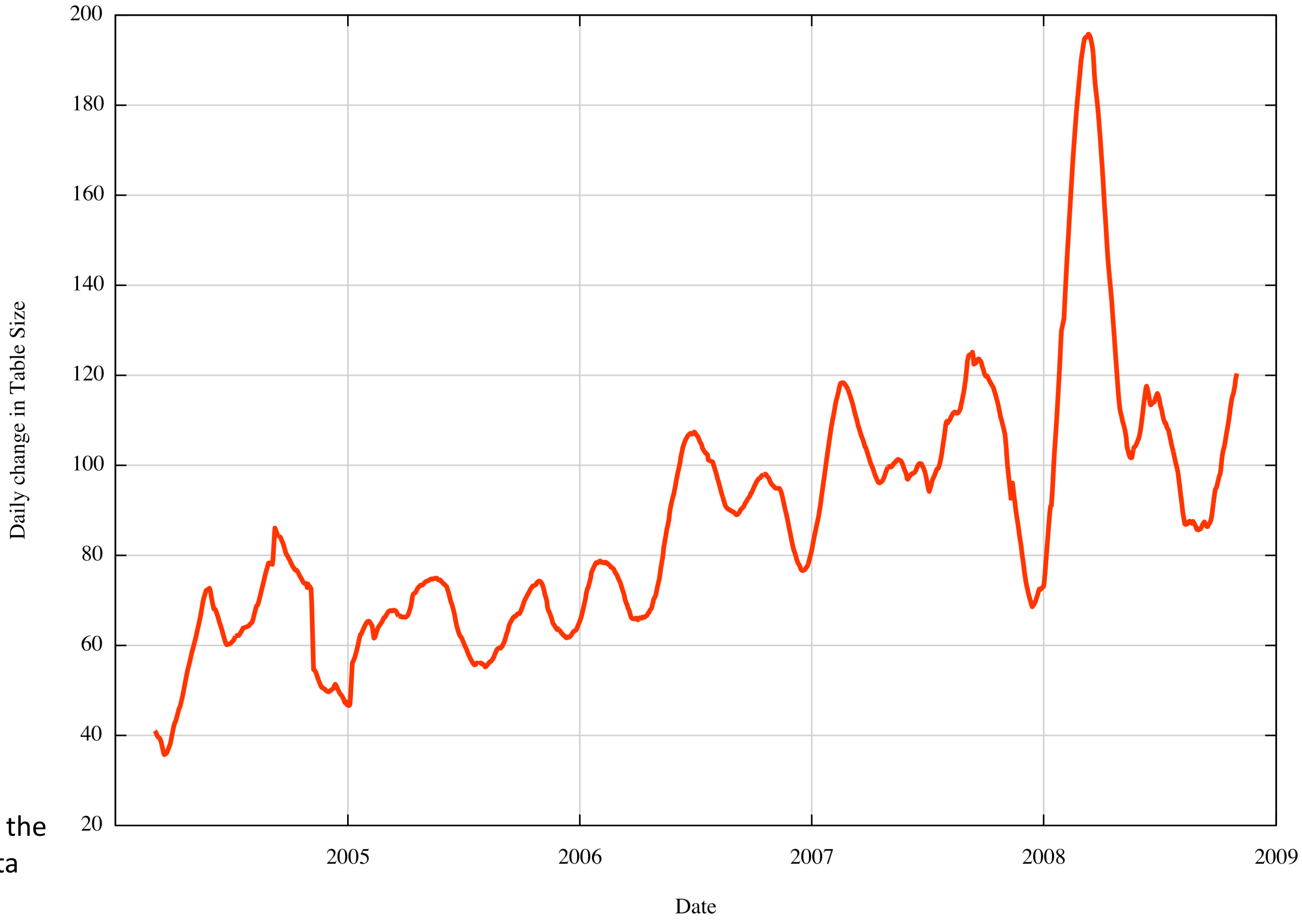
Use IPv4 BGP table size data to generate a 4 year projection of the IPv4 routing table size

- smooth data using a sliding window average
- take first order differential
- generate linear model using least squares best fit
- integrate to produce a quadratic data model

IPv4 Table Size – 60 months data window

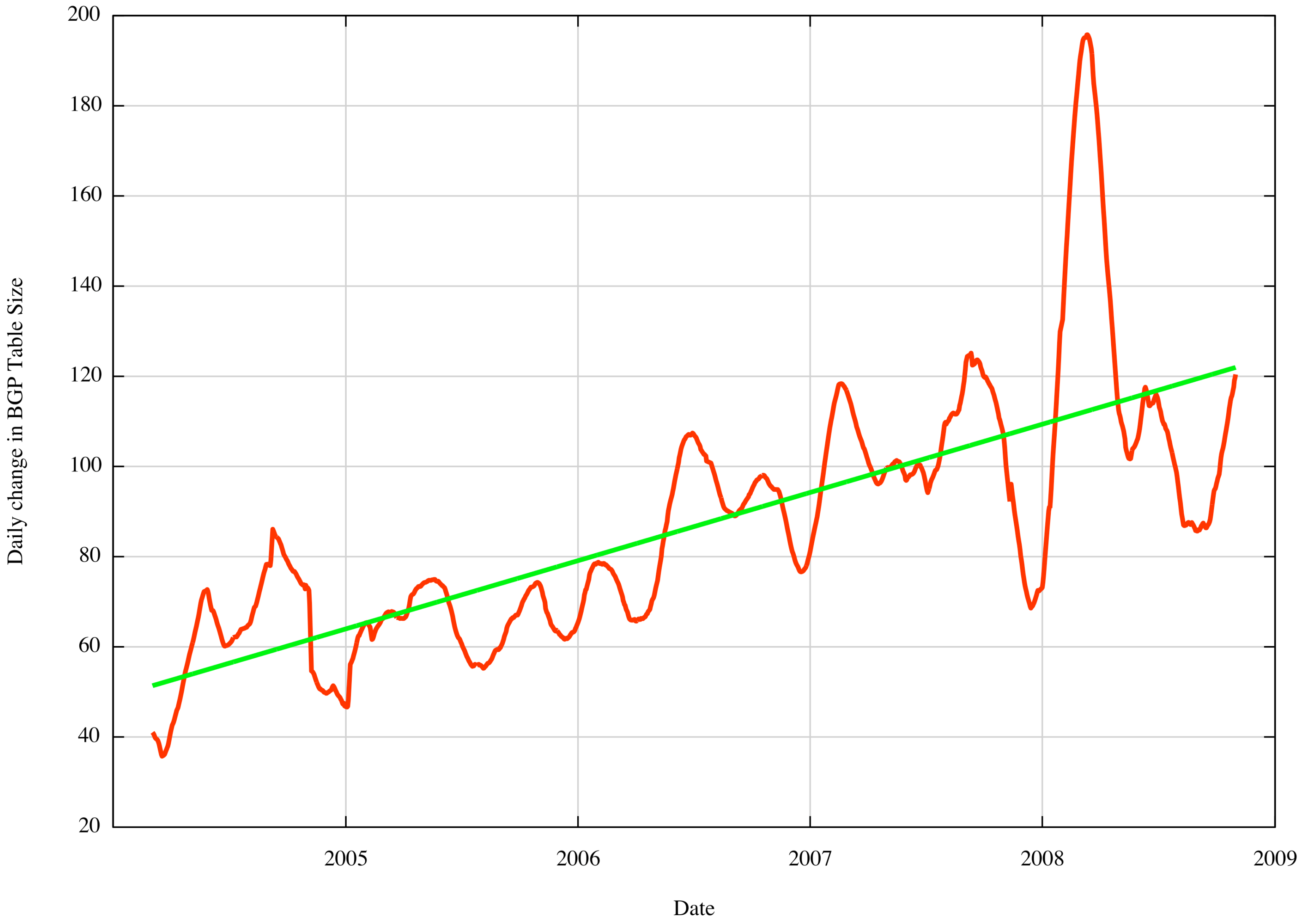


Daily Growth Rates

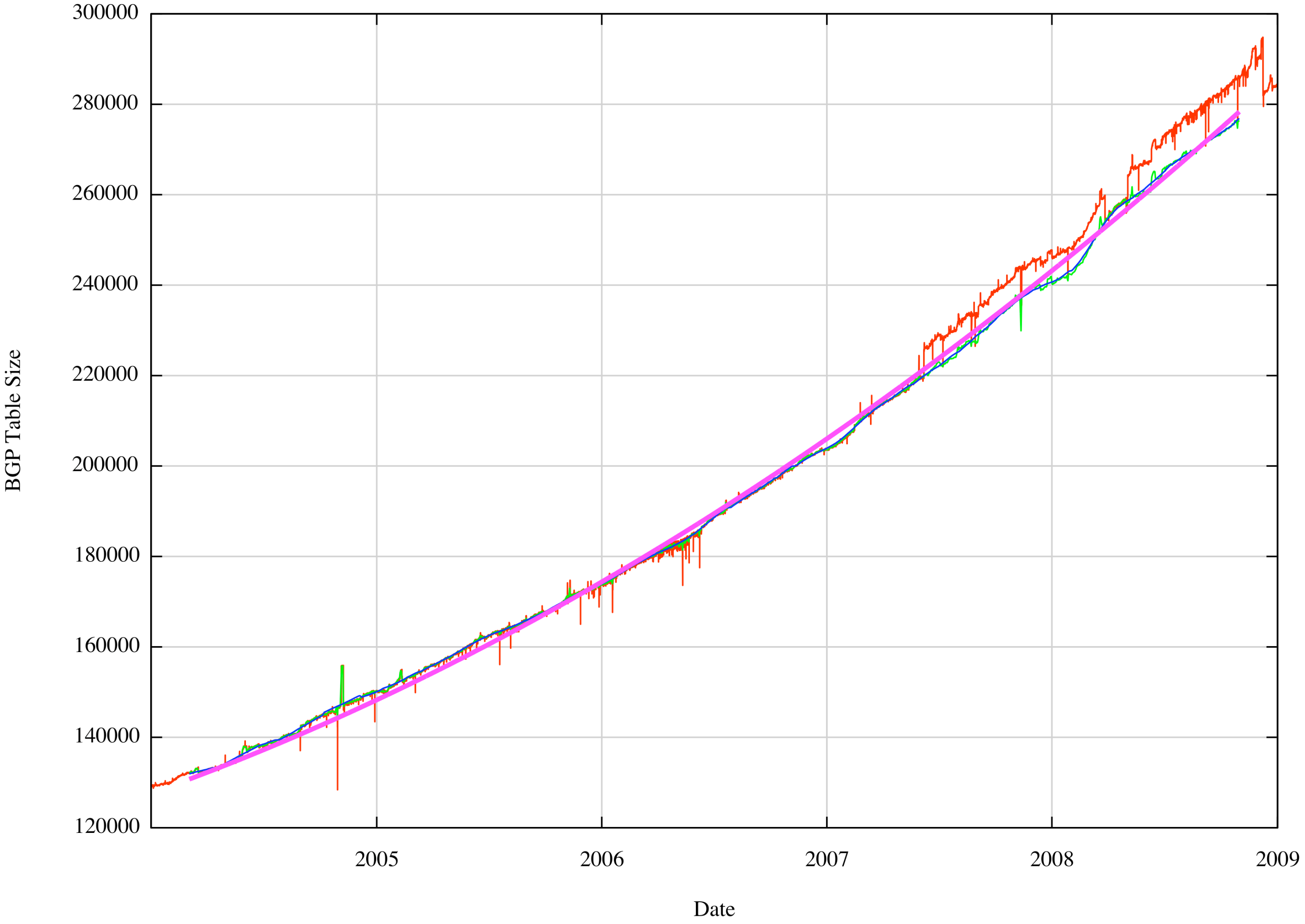


First order differential of the smoothed data

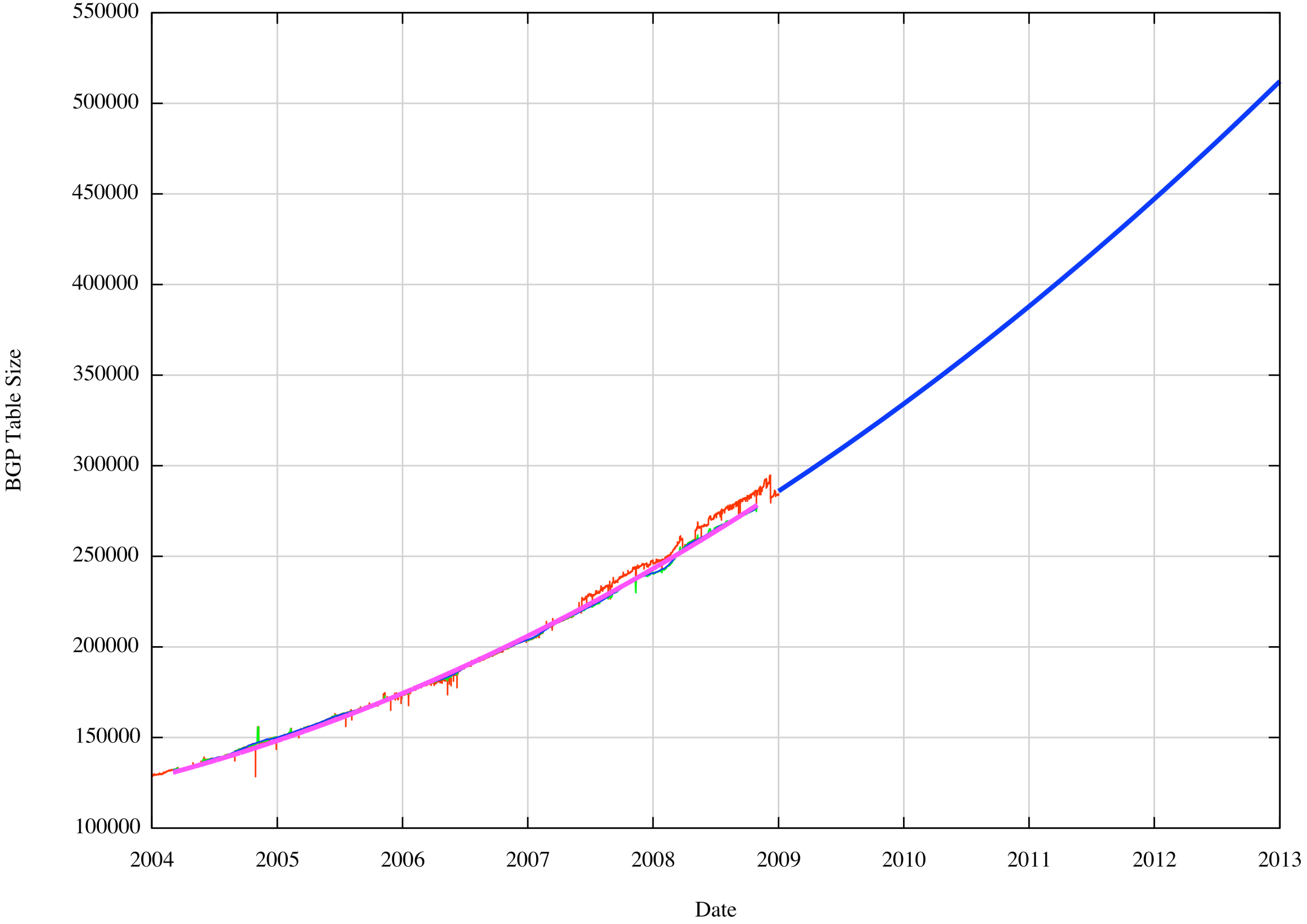
Daily Growth Rates



IPv4 Table Size
Quadratic Growth Model



IPv4 Table Size Quadratic Growth Model - Projection



BGP Table Size Predictions

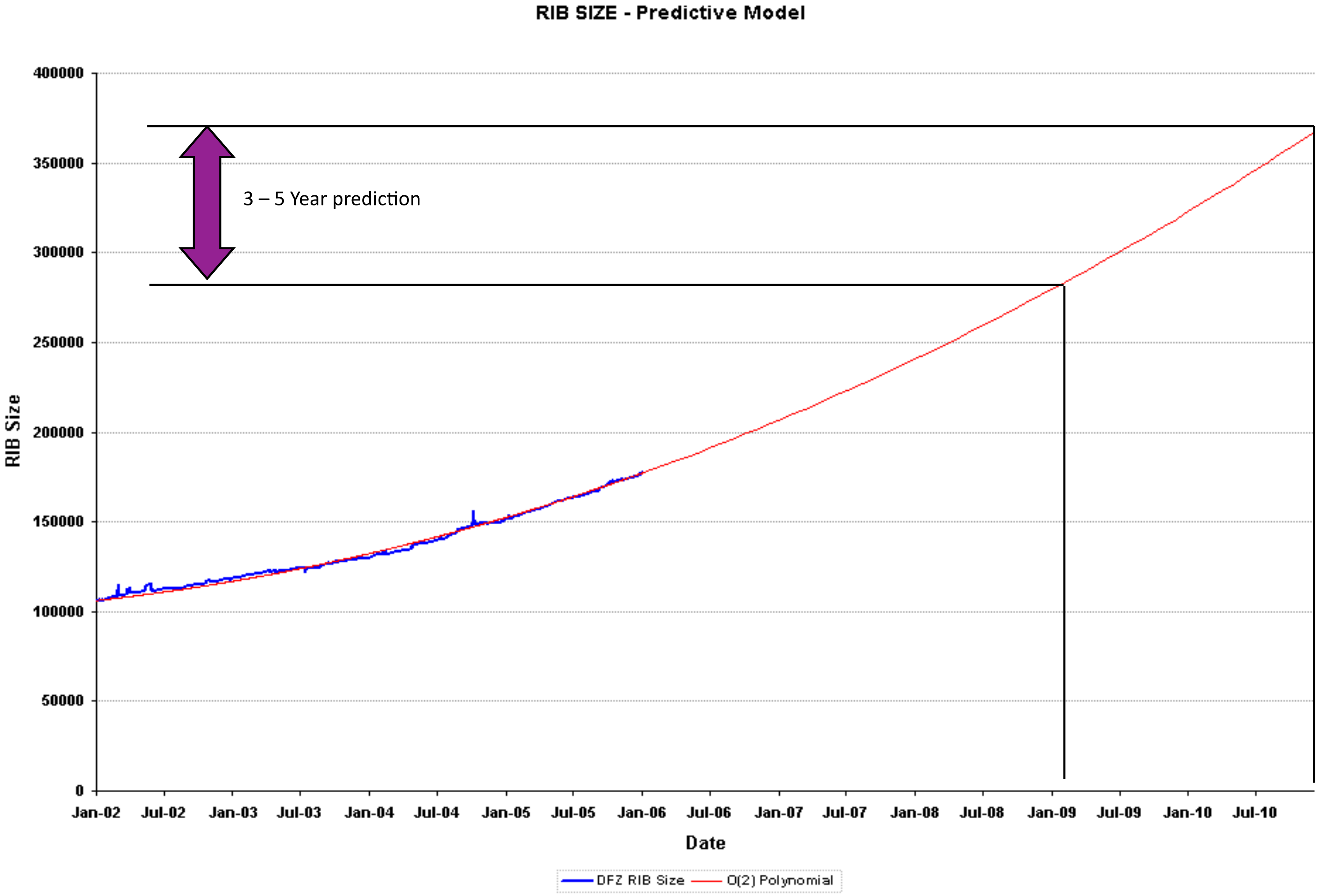
May 2009	285,000 entries
12 months	335,000 entries
24 months	388,000 entries
36 months*	447,000 entries
48 months*	512,000 entries

* These numbers are dubious due to IPv4 address exhaustion pressures. It is possible that the number will be larger than the values predicted by this model.

Back in 2006 ….

- This modeling work on the BGP table size was performed at the end of 2005 to generate a 3 and 5 year projection

2006 prediction



BGP Table Size Predictions

May 2009	285,000 entries (2006: 275,000)
12 months	335,000 entries
24 months	388,000 entries (2006: 370,000)
36 months*	447,000 entries
48 months*	512,000 entries

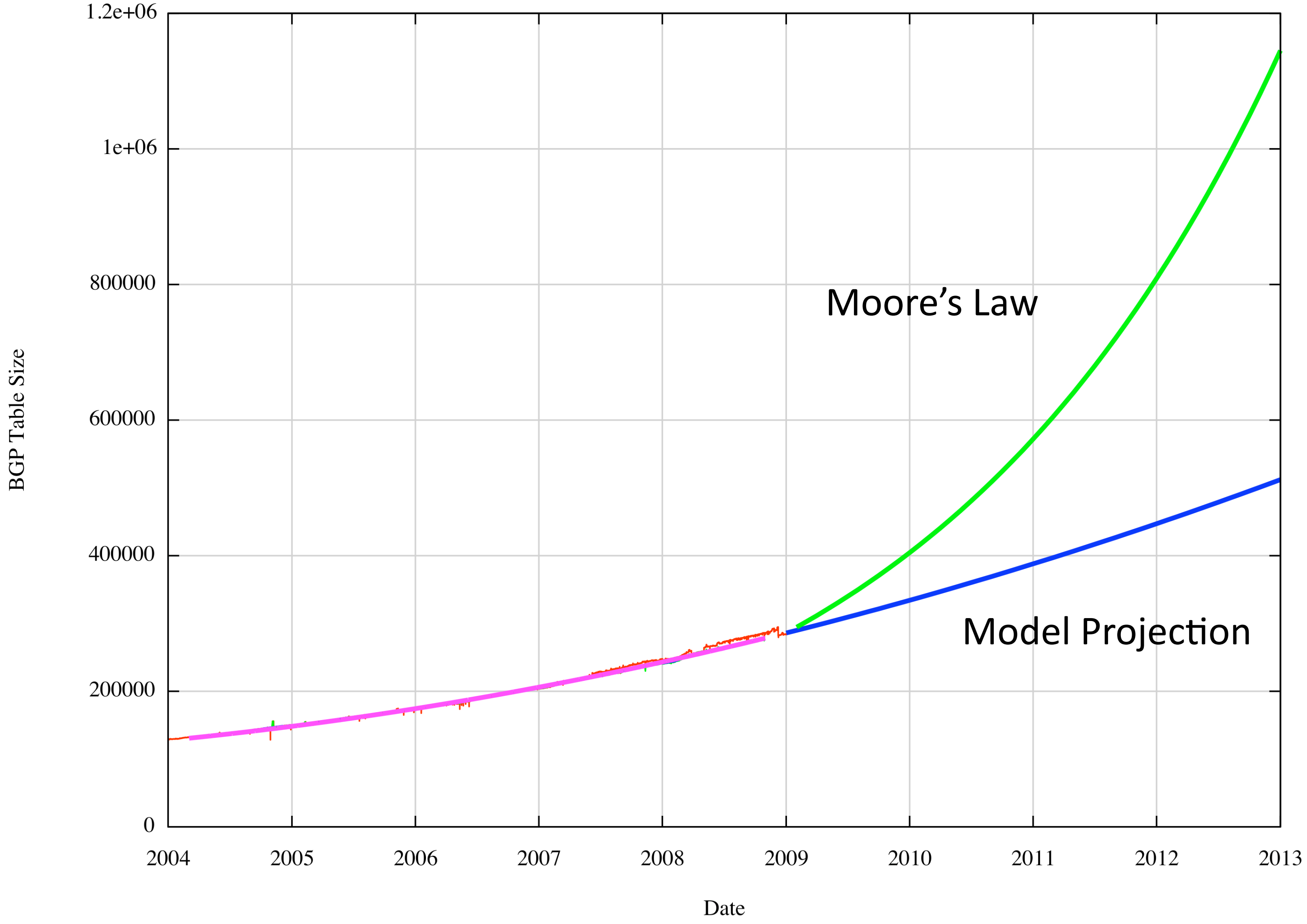
* These numbers are dubious due to IPv4 address exhaustion pressures. It is possible that the number will be larger than the values predicted by this model.

Is this a Problem?

BGP Scaling and Table Size

- As long as growth rates stay within the general parameters of Moore’s Law the unit cost of the routing function should not escalate
  - assuming that Moore’s law continues to hold
  - and assuming that routing table growth is driven by similar factors as in the recent past

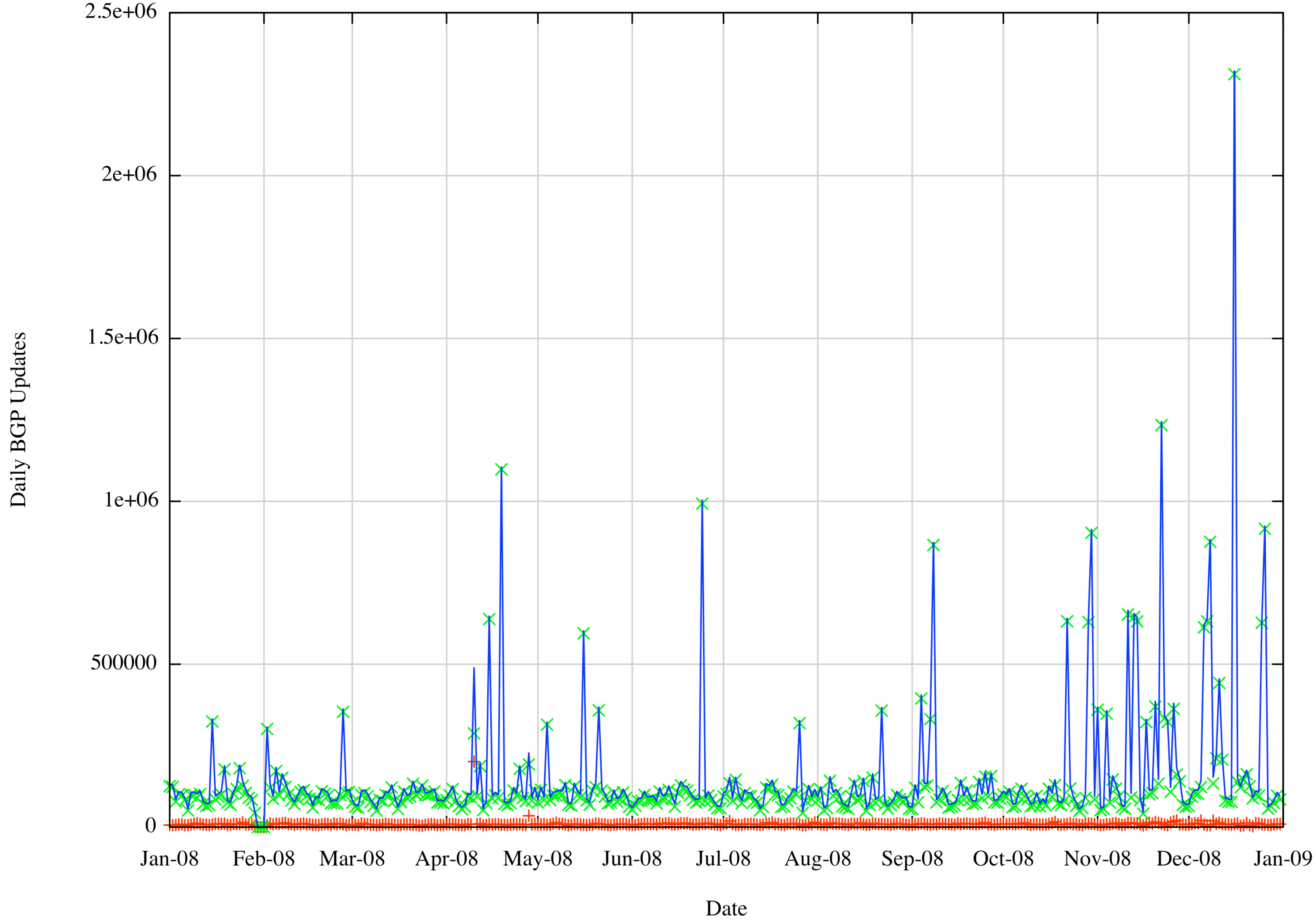
Projections against Moore's Law



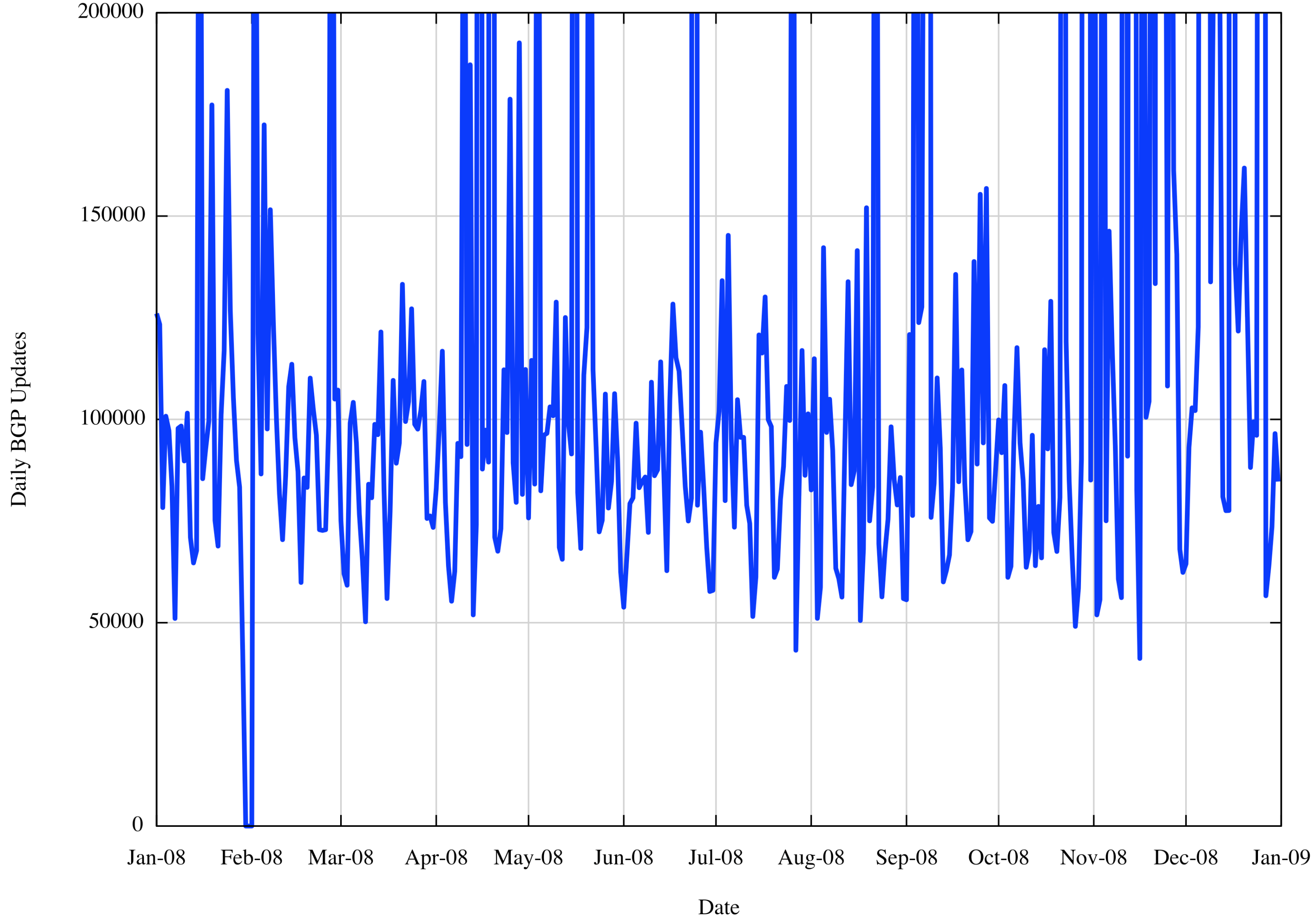
BGP Scaling and Stability

- Is it the size of the RIB or the level of dynamic update and routing stability that is the concern here?
- So lets look at update trends in BGP…

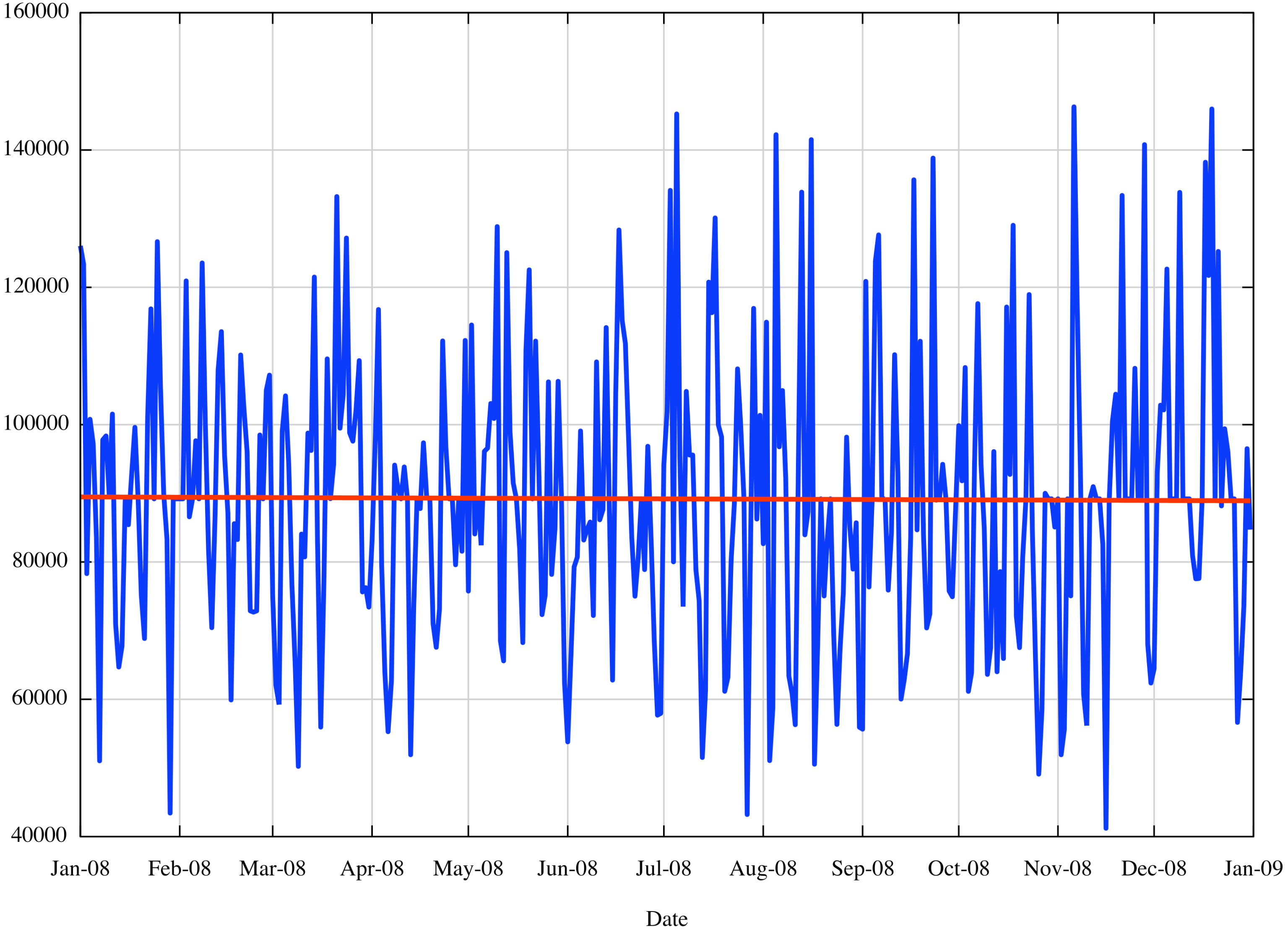
Daily Announce and Withdrawal Rates



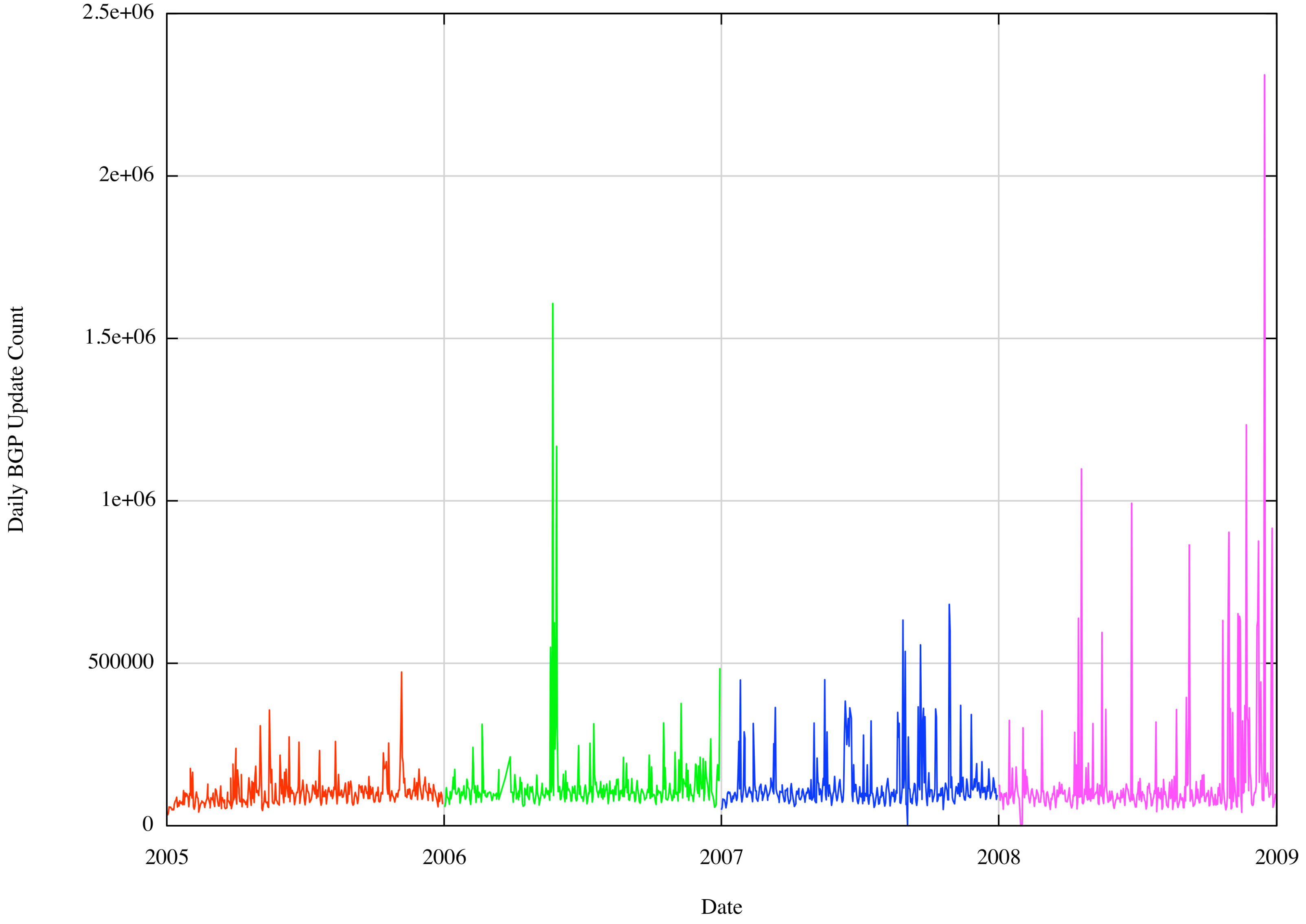
Daily Updates



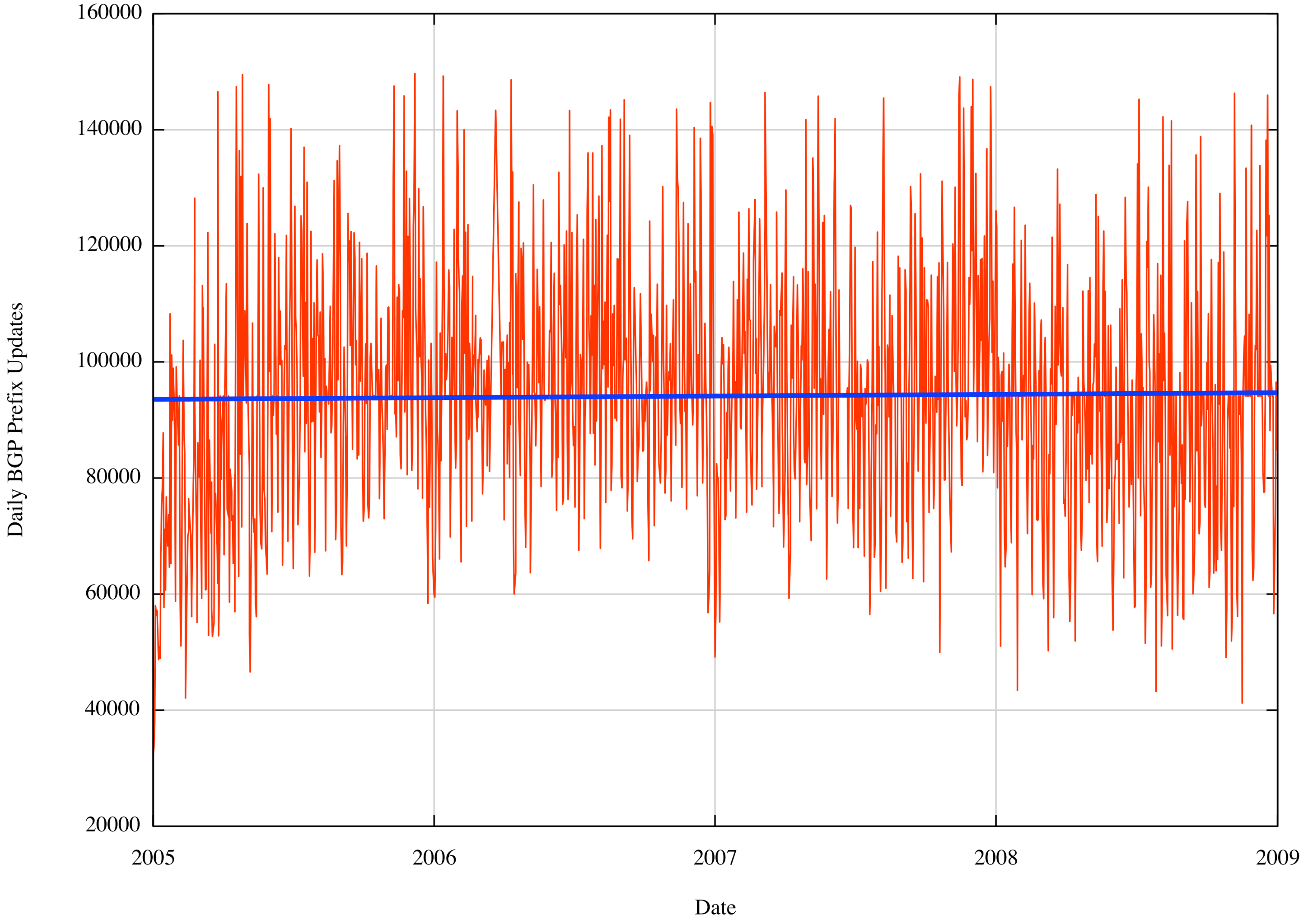
Best Fit to Updates



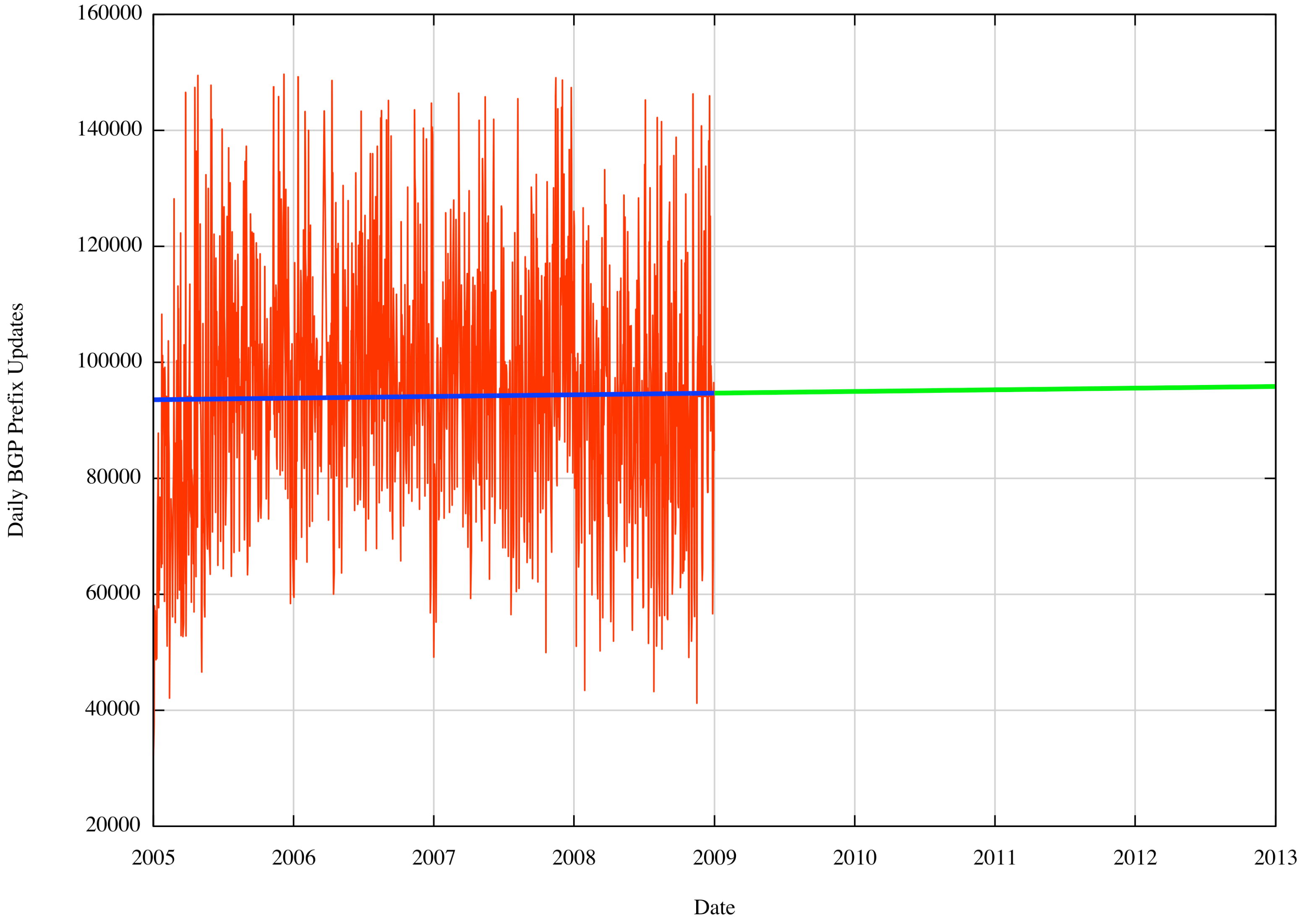
BGP Updates - Extended Data Set



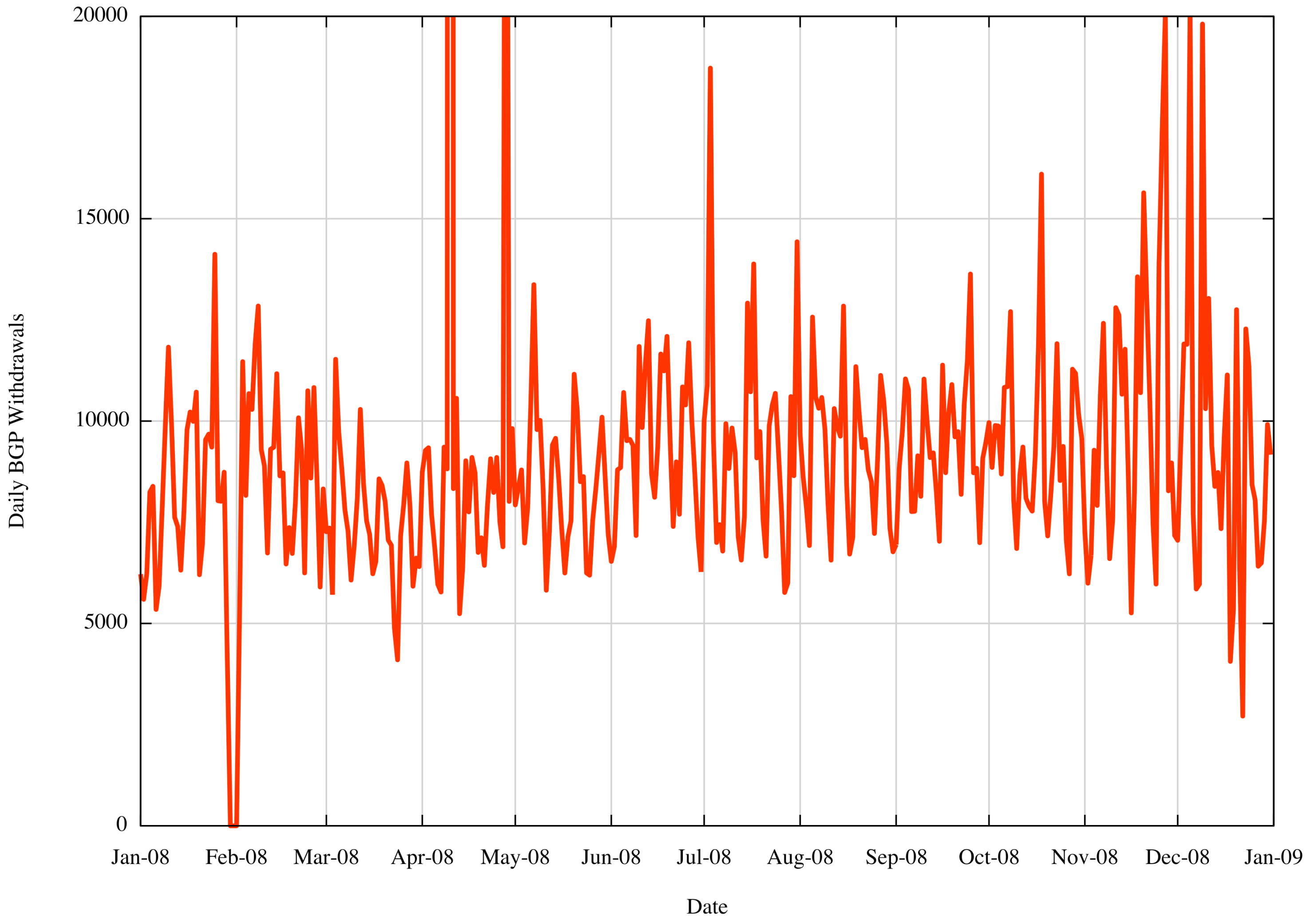
BGP Updates - Extended Data Set



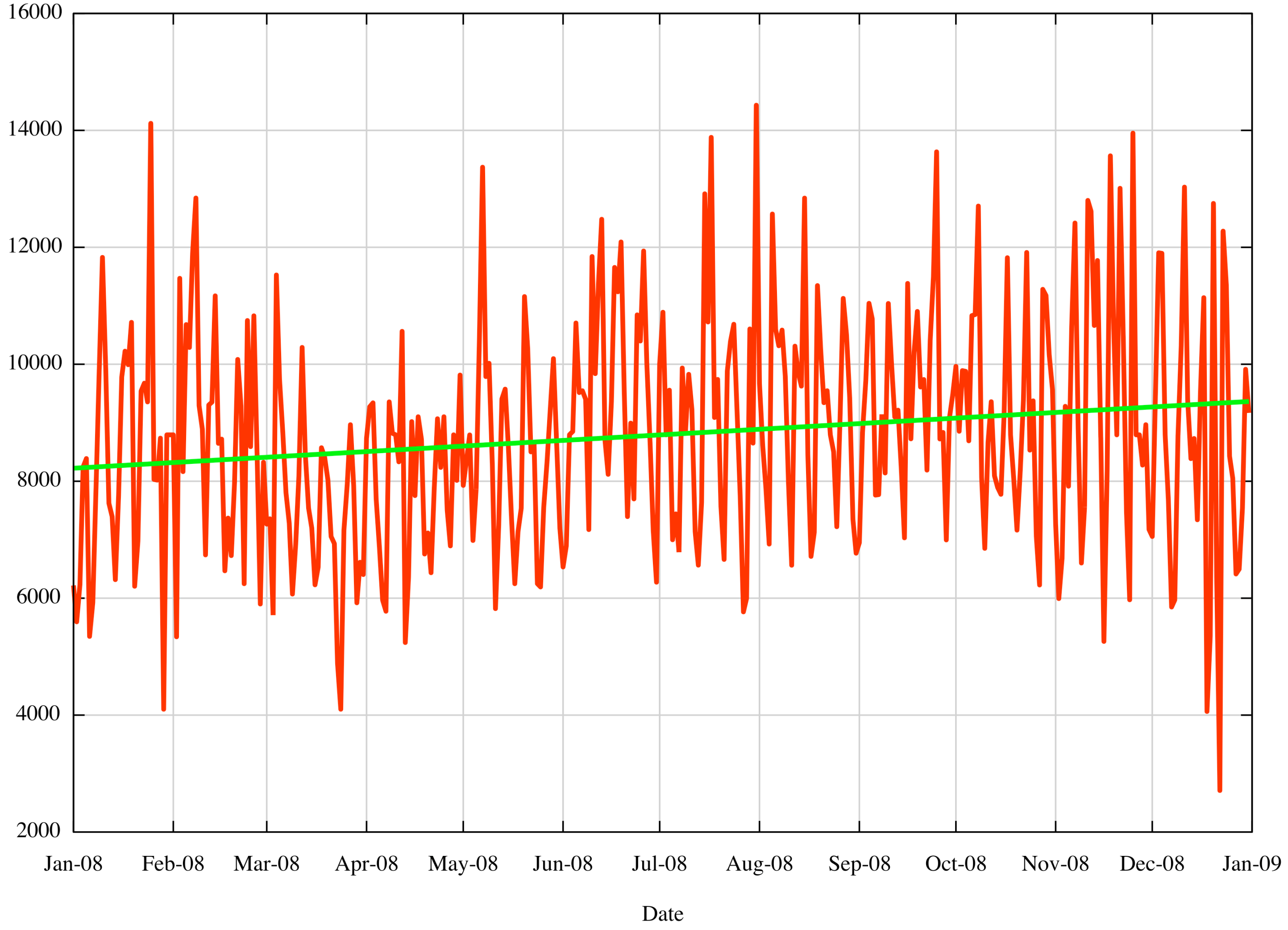
BGP Update Projection



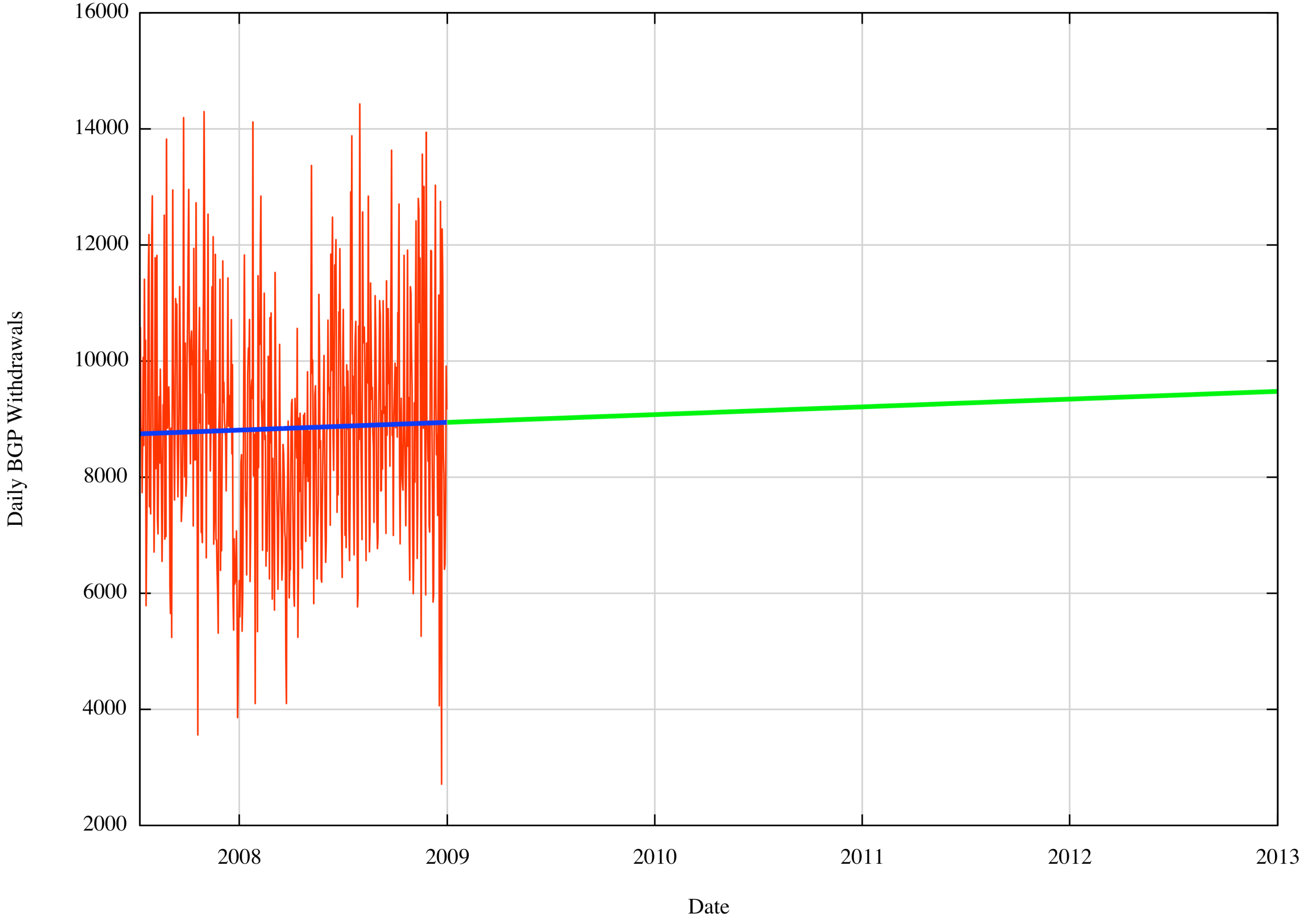
Daily Withdrawals



Best Fit to WDLs



BGP Withdrawal Projection



Why is this?

- Why are the levels of growth in BGP updates not proportional to the size of the routing table?
  - growth rates of BGP updates appear to be far smaller than the growth rate of the routing space itself

Convergence in BGP

- BGP is a distance vector protocol
- This implies that BGP may send a number of updates in a tight “cluster” before converging to the “best” path
- This is clearly evident in withdrawals and convergence to (longer) secondary paths

For Example

Withdrawal at source at 08:00:00 03-Apr of 84.205.77.0/24 at MSK-IX, as observed at AS 2.0

Announced AS Path: <4777 2497 9002 12654>

Received update sequence:

```
08:02:22 03-Apr + <4777 2516 3549 3327 12976 20483 31323 12654>
08:02:51 03-Apr + <4777 2497 3549 3327 12976 20483 39792 8359 12654>
08:03:52 03-Apr + <4777 2516 3549 3327 12976 20483 39792 6939 16150 8359 12654>
08:04:28 03-Apr + <4777 2516 1239 3549 3327 12976 20483 39792 6939 16150 8359 12654>
08:04:52 03-Apr -  <4777 2516 1239 3549 3327 12976 20483 39792 6939 16150 8359 12654>
```

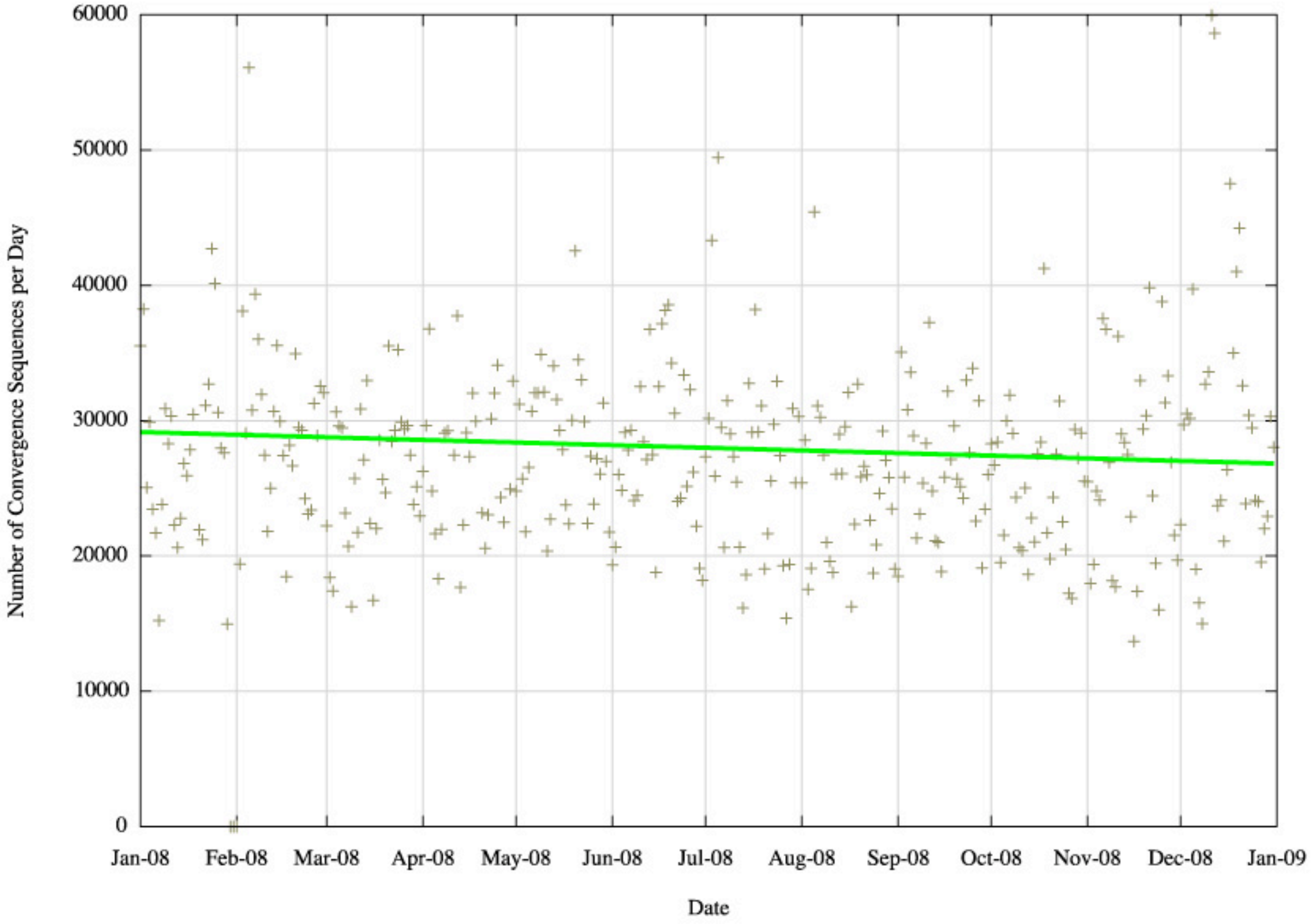
1 withdrawal at source generated a convergence sequence of 5 events, spanning 150 seconds

Measurement Approach for stability behaviour

- Group all updates into “convergence sequences” using a stability timer of 130 seconds
    - A prefix is “stable” if no updates or withdrawals for that prefix are received in a 130 second interval
    - A “convergence sequence” is a series of updates and withdrawals that are spaced within 130 seconds or each other
- Remove all isolated single update events (generally related to local BGP session reset)

Stability Trends

Number of "Convergence Sequences" per day for 2008



Stability Trends

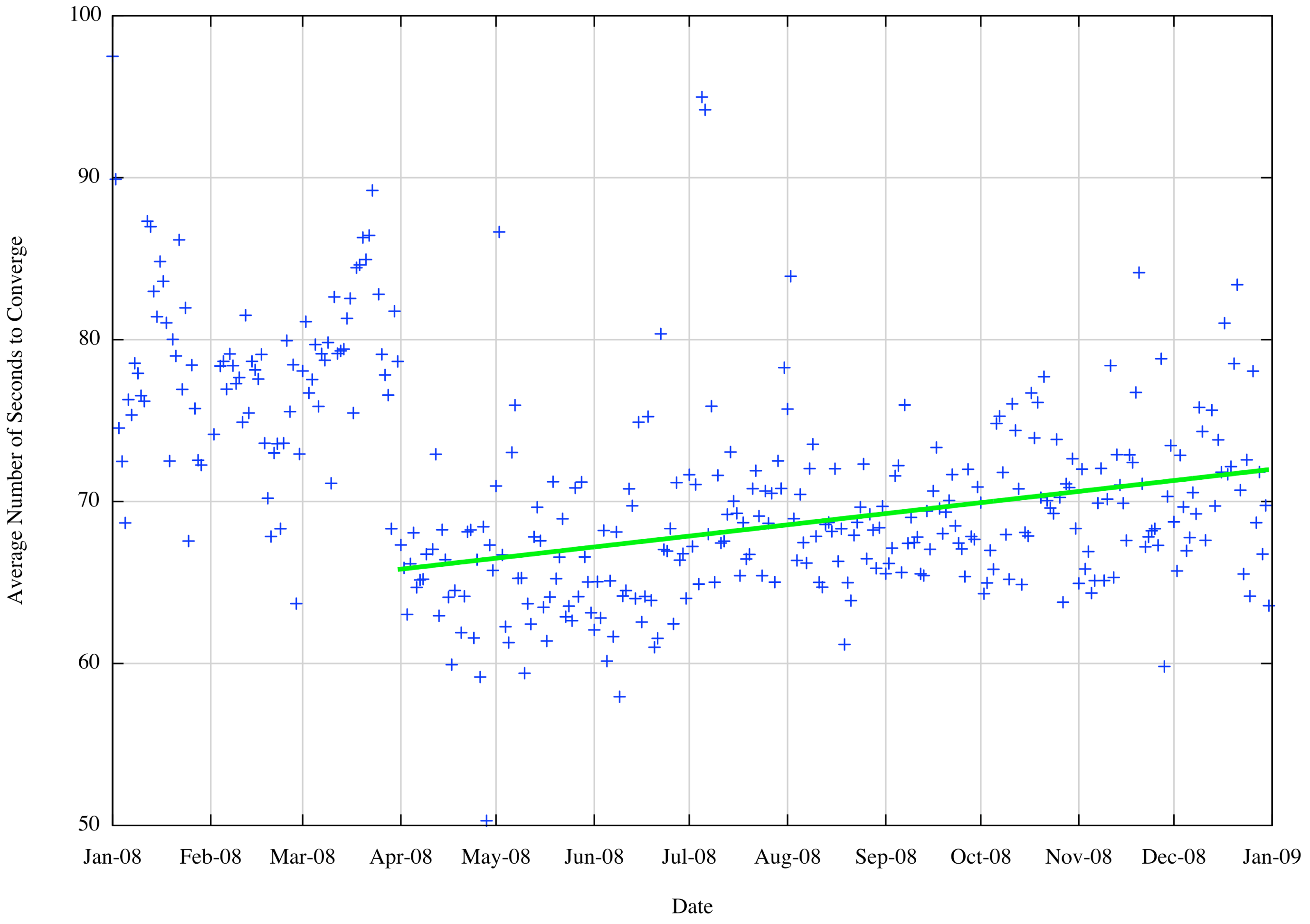
- The trend average number of prefixes that generated convergence sequences dropped from 29,156 to 26,835, or a drop of 8% over the year
- The BGP RIB grew by 17% (245,000 to 286,000)
- The relative occurrence of instability dropped by a 27% over the year (11.9% to 9.3%)

- BGP was trending to greater stability in relative terms over 2008

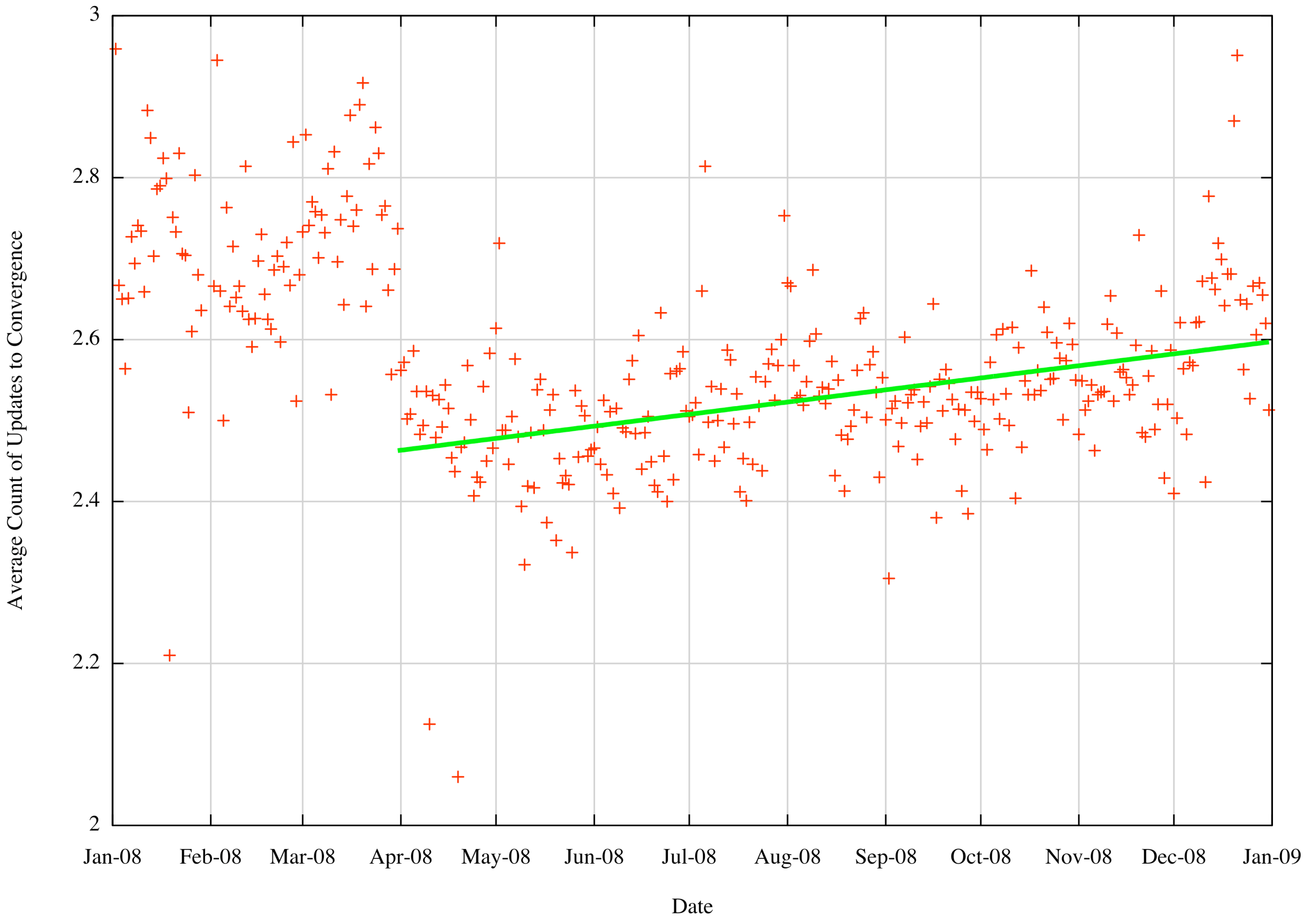
Stability Trends

- Is that’s the case why isn’t the number of BGP updates and withdrawals decreasing over time?

Average Convergence Time



Average Convergence Updates



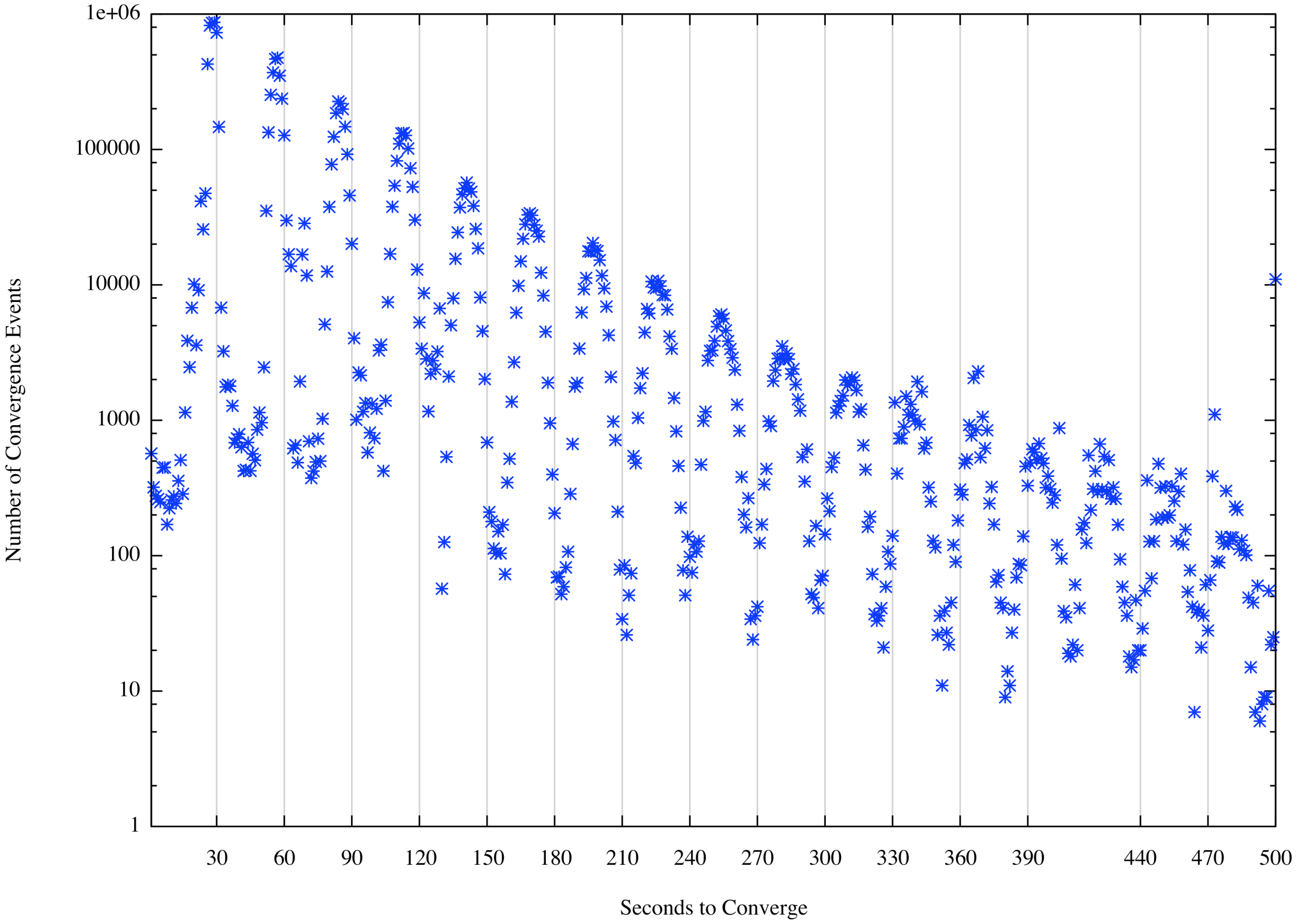
Convergence Trends

- In 9 months the average time to converge increased by 9% (65.8 seconds to 71.2 seconds) or an annual rate of 12%
- The number up BGP updates increased by 5% (2.46 to 2.59 updates) or an annual rate of 6.9%

Convergence Trends

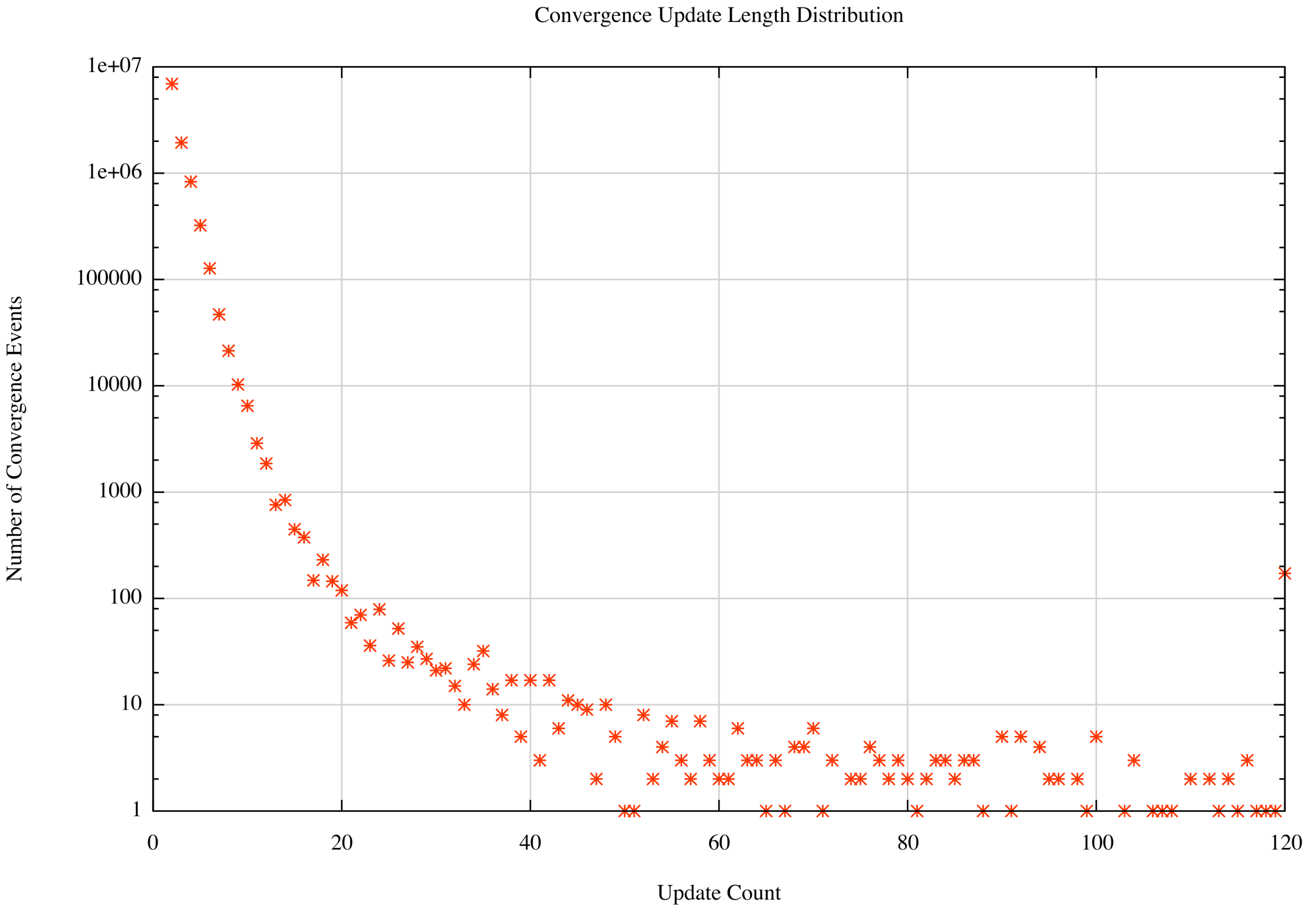
- Fewer instability events, taking slightly longer to converge and slightly more updates to reach convergence
- Is the a general trend, or a case of a skewed distribution driving the average values?

Convergence Distribution



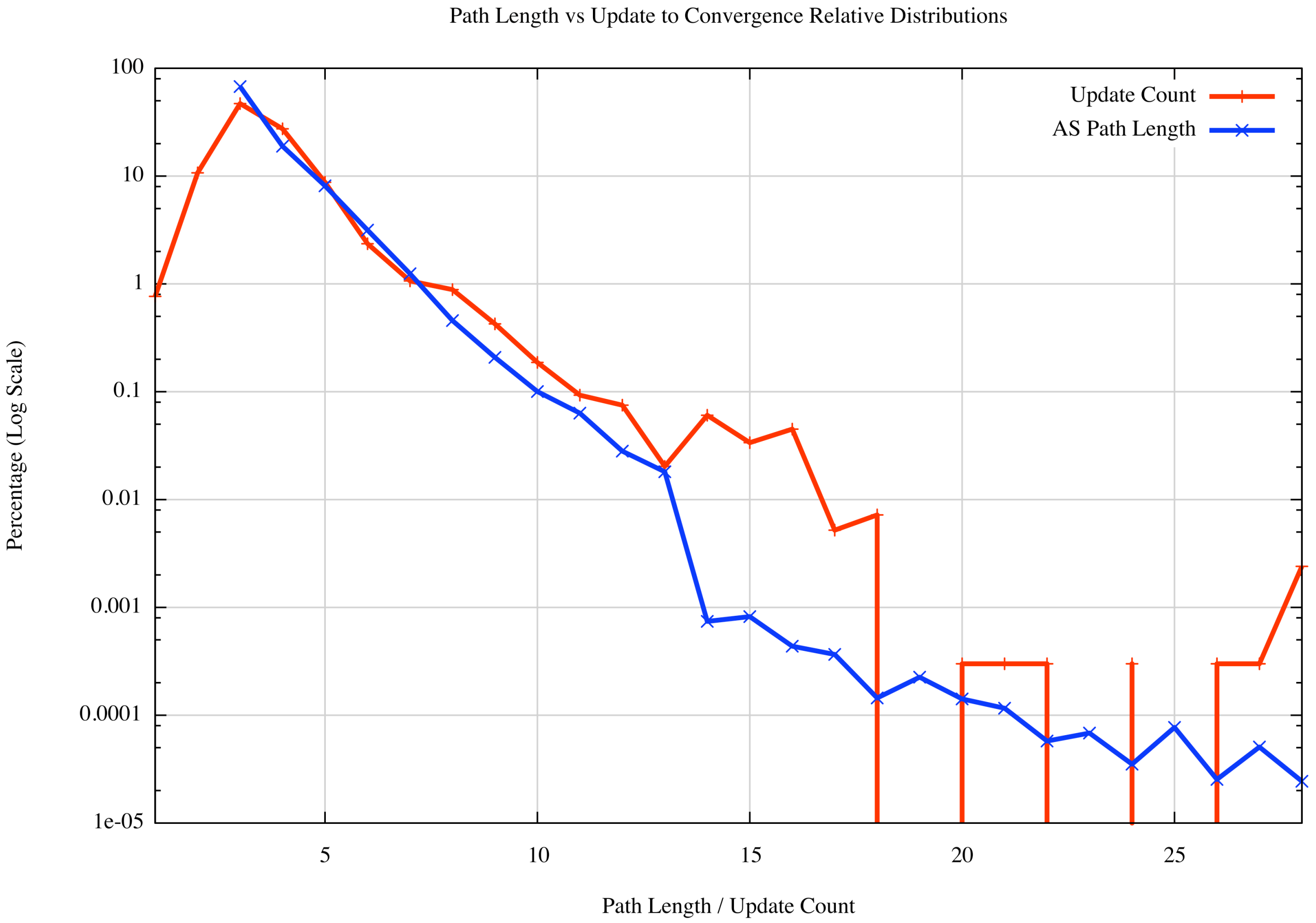
Time to reach converged state has strong 27 second peaks
Default 27 -30 second MRAI timer is the major factor here

Convergence Distribution



Number of updates to reach convergence has exponential decay in the distribution. Does this correlate to the distribution of AS path lengths in the routing table?

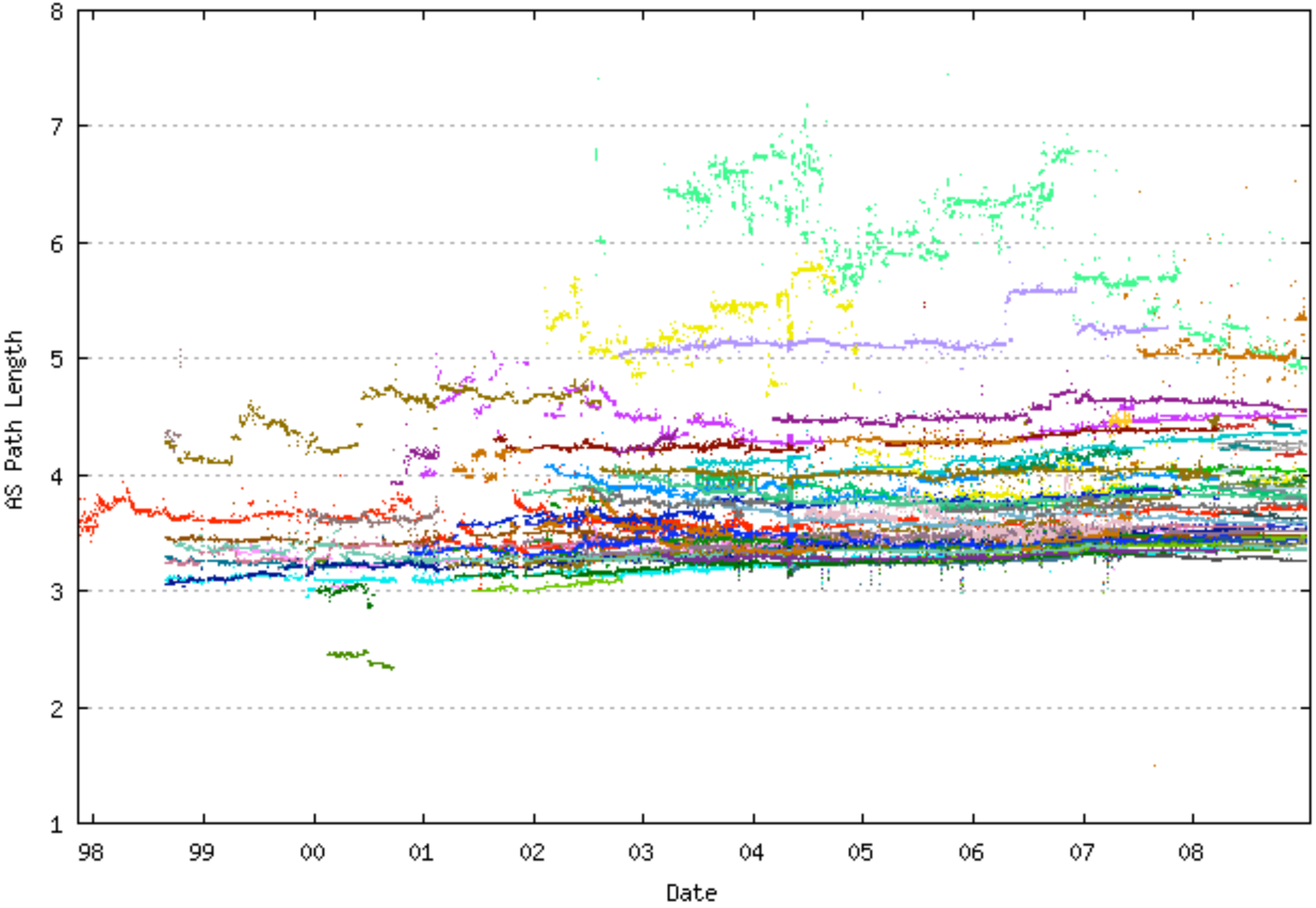
Convergence Distribution



Observations

- There is a plausible correlation between AS Path Length Distribution and Convergence Update Distribution for counts <= 13
- This is a possible indication that the number of updates to reach convergence and the time to reach convergence is related to AS Path Length for most (99.84%) of all instability events
- Other events are related to longer term instability that may have causes beyond conventional protocol behaviour of BGP

Average AS Path Length is long term stable



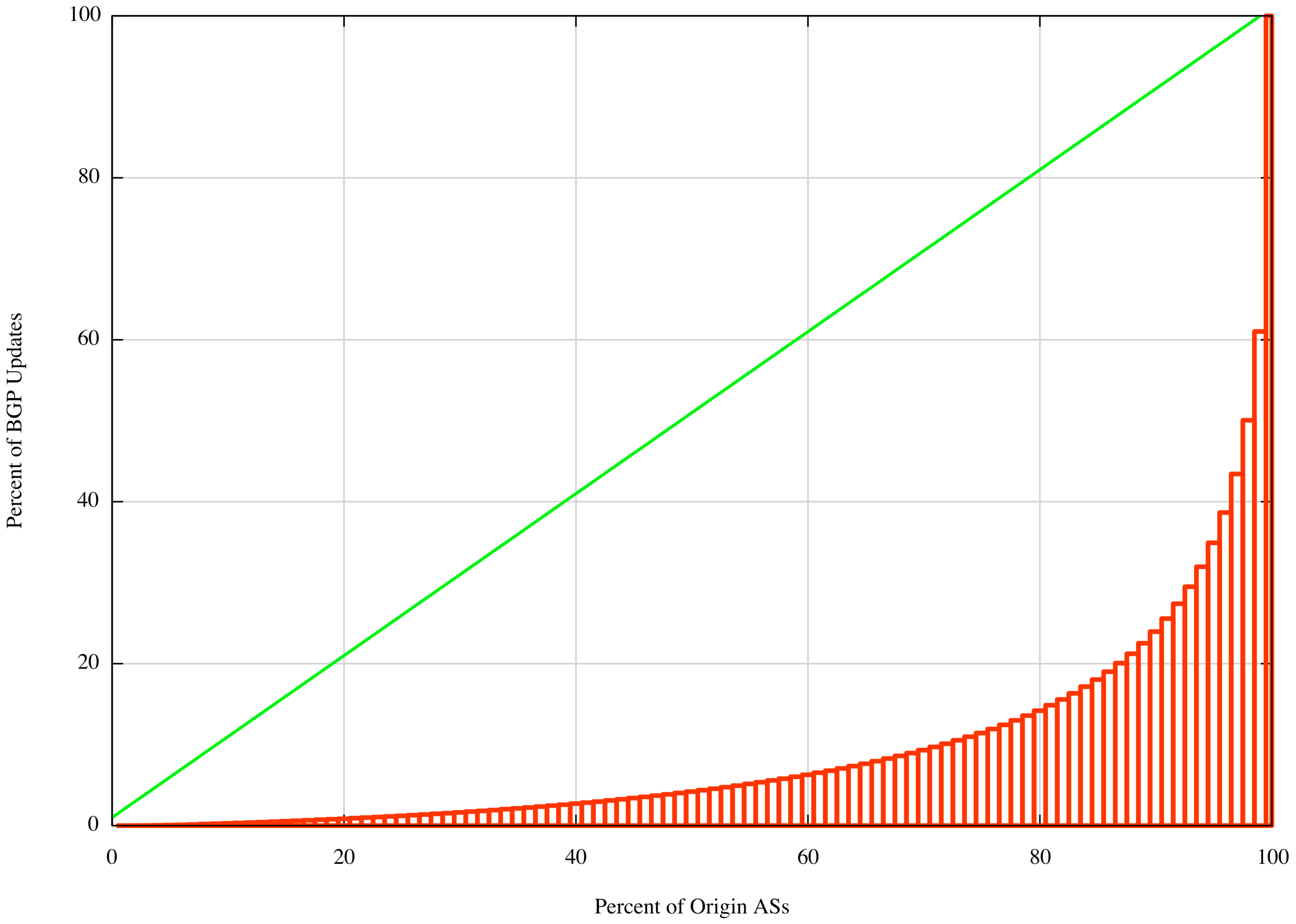
What is going on?

- The convergence instability factor for a distance vector protocol like BGP is related to the AS path length, and average AS Path length has remained steady in the Internet for some years
- Taking MRAI factors into account, the number of received Path Exploration Updates in advance of a withdrawal is related to the propagation time of the withdrawal message. This is approximately related to the average AS path length
- Today's Internet is more densely interconnected, but is not any more "stringier"
- This implies that the number of protocol path exploration transitions leading to a prefix withdrawal should be relatively stable over time

What is going on?

- But that’s not exactly what we see in the data
- The average duration and number of updates per instability “event” appears to be slowly increasing over time
- Why?

The update distribution of BGP is heavily skewed



What is going on?

- A significant component of dynamic BGP load is not an artifact of the larger routing space, but a case of relatively intense levels of BGP path manipulation at or close to origin for TE purposes from a very small subset of origin AS's at the "edge" of the network
  - the dominant factor behind what is being measured in updates is not implicitly related to network component stability, but more likely to be related to path manipulation associated with TE

Some Closing Opinions

- The BGP sky is **not** falling

  The 2008 BGP data appears to indicate that the prospects of the imminent death of BGP through routing table inflation appear to be vastly exaggerated

  - The inflation rate of the routing table remains well under Moore's law
  - The rate of increase of processed updates is minimal
  - The stability of the network is improving over time

  – The network is, on the whole, very stable and BGP is not under immediate stress in terms of scaling pressures

A Word of Caution

- This is a simple exercise in statistical curve fits, not a demand level simulation of the players routing environment.
- This exercise does not factor in any IPv4 address exhaustion considerations and scenarios around address movement that may alter the picture of fragmentation of the routing space.
- However the AS growth projections are a strong indicator of underlying industry dynamics in terms of discrete routing entities, and these projections show a modest growth component

  This means that while the projections are very weak in the period of 2011 and beyond, there are reasonable grounds to take a conservative view of BGP growth in this phase of the Internet's evolution

Thank You

Questions?