BGP in 2009

Geoff Huston
APNIC R&D
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:

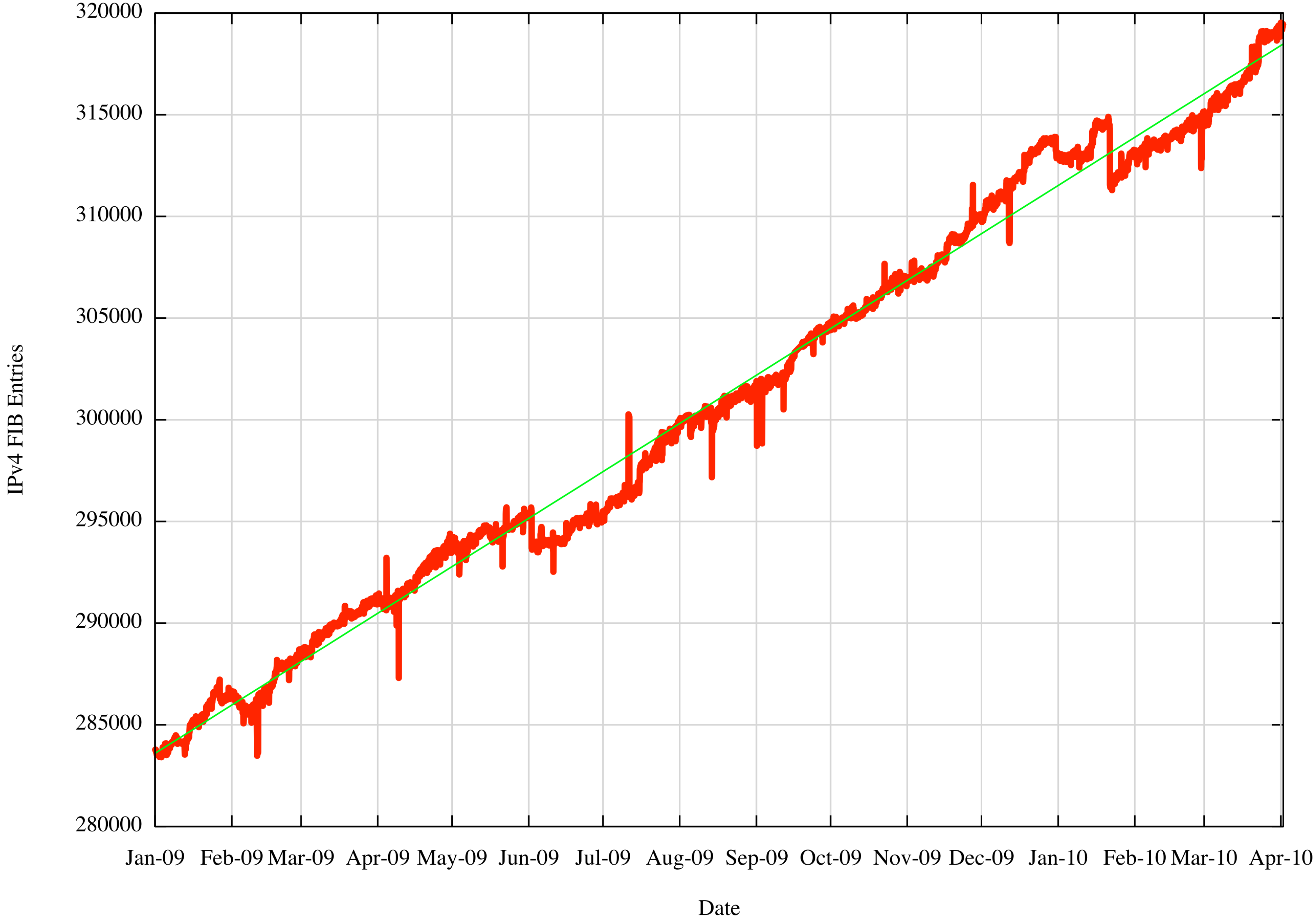
1. Assemble a large set of BGP peering sessions and record everything
   - RIPE NCC's RIS service
   - Route Views
2. Perform carefully controlled injections of route information and observe the propagation of information
   - Beacons
   - AS Set manipulation
   - Bogon Detection and Triangulation
3.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

AS2.0 BGP measurement

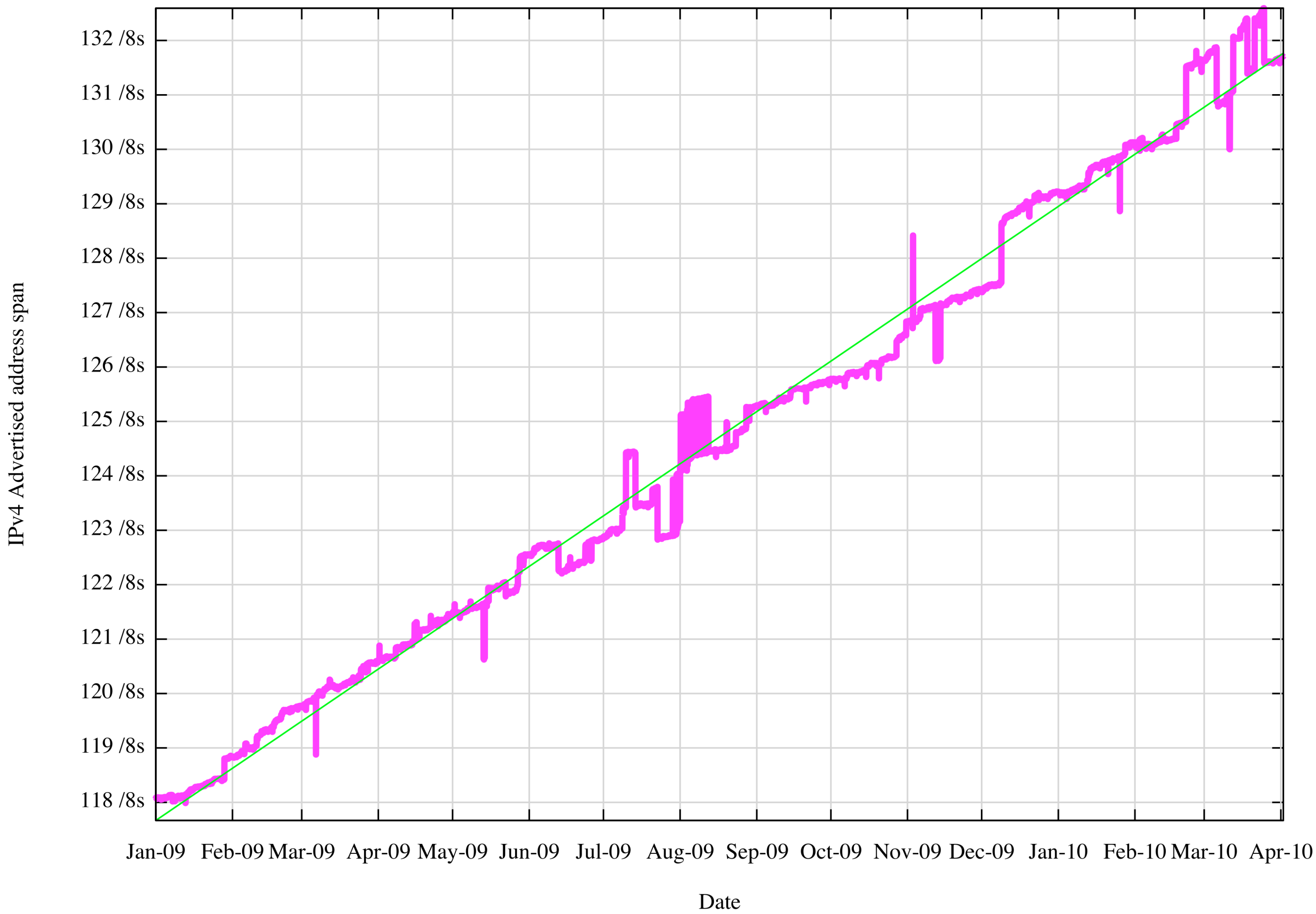
- Data collection since 1 July 2007
- Passive data measurement technique (no advertisements or probes)
- Quagga platform, connected to AS4608
- Dual Stack operation
- Archive of all BGP updates and daily RIB dumps
- Data and reports are continuously updated and published: http://bgp.potaroo.net

BGP in 2009

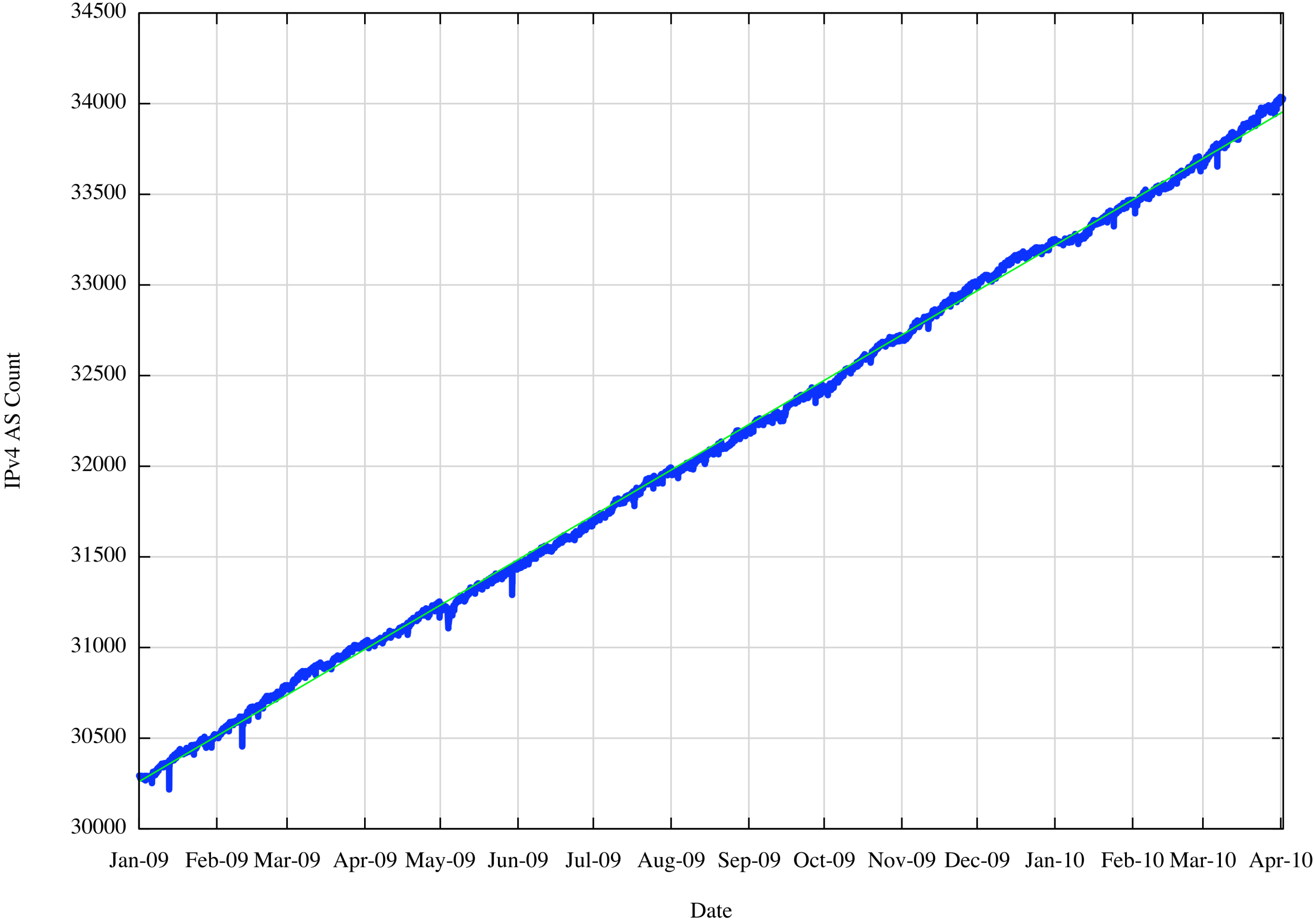
IPv4 BGP Prefix Count



IPv4 Routed Address Span



IPv4 Routed AS Count



IPv4 Vital Statistics for 2009

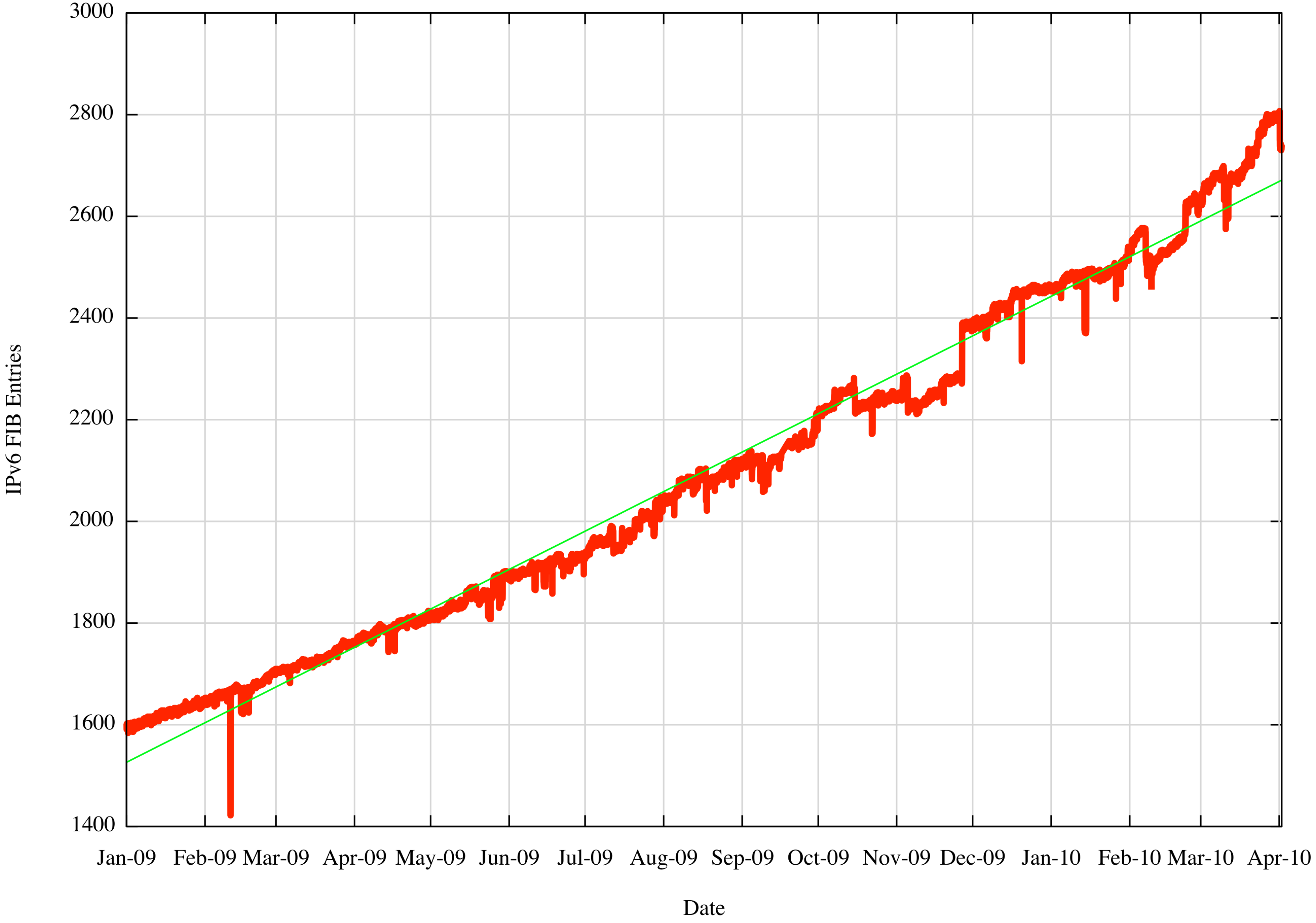
	Jan-09	Dec-09	
Prefix Count	283,000	312,000	**+10%**
Roots	135,000	151,000	+12%
More Specifics	148,000	161,000	+ 9%
Address Span	118/8s	129/8s	**+ 9%**
AS Count	30,200	33,200	**+10%**
Transit	4,000	4,400	+10%
Stub	26,200	28,800	+10%

The Internet in 2009

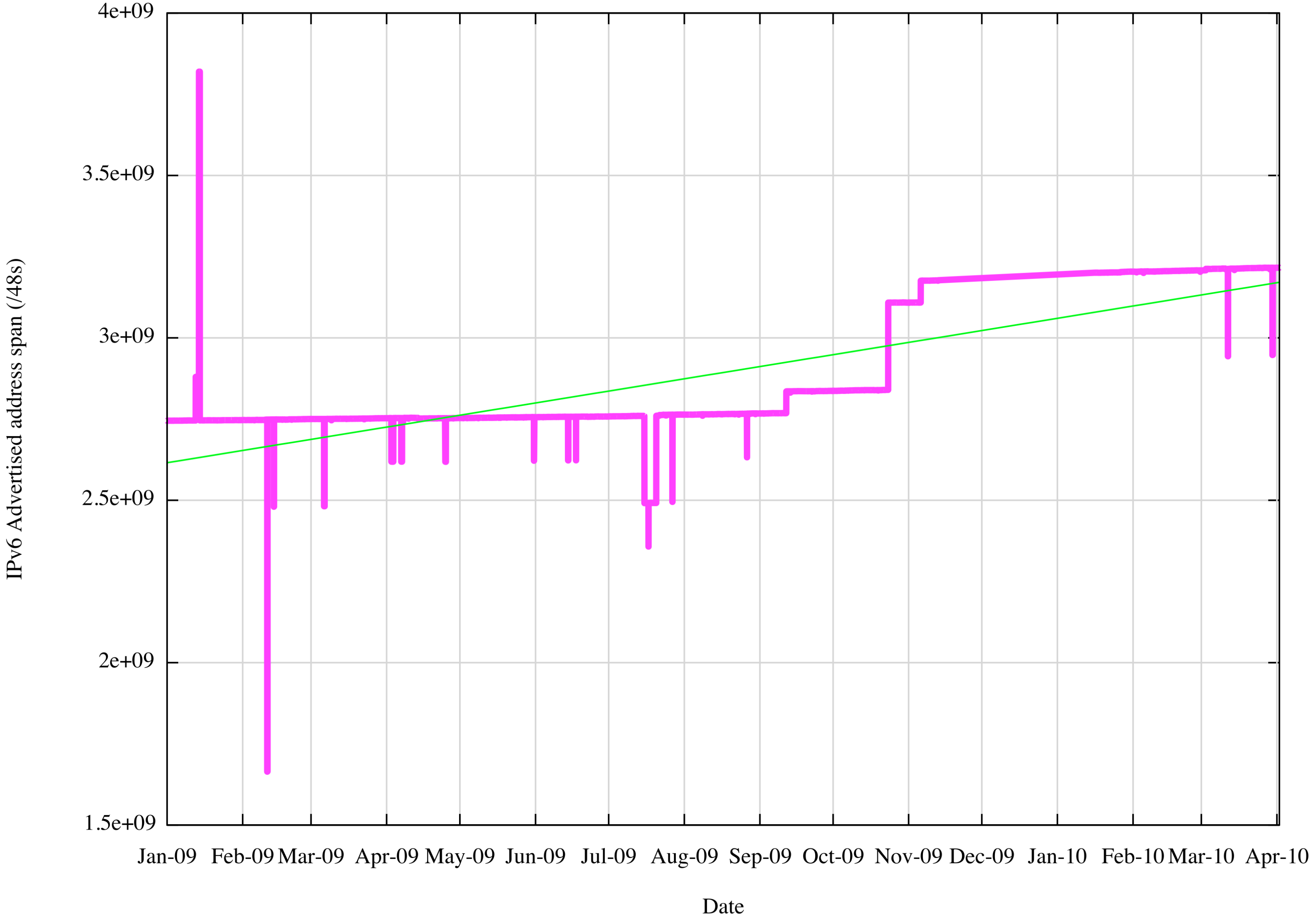
The IPv4 Routing table grew by 10% over 2009

- compared with 12% - 15% growth in 2008
- Is this an indicator of reduced growth overall in the Internet?
- Or an indicator of reducing diversity in the supply side, and increasing market dominance by the larger providers?

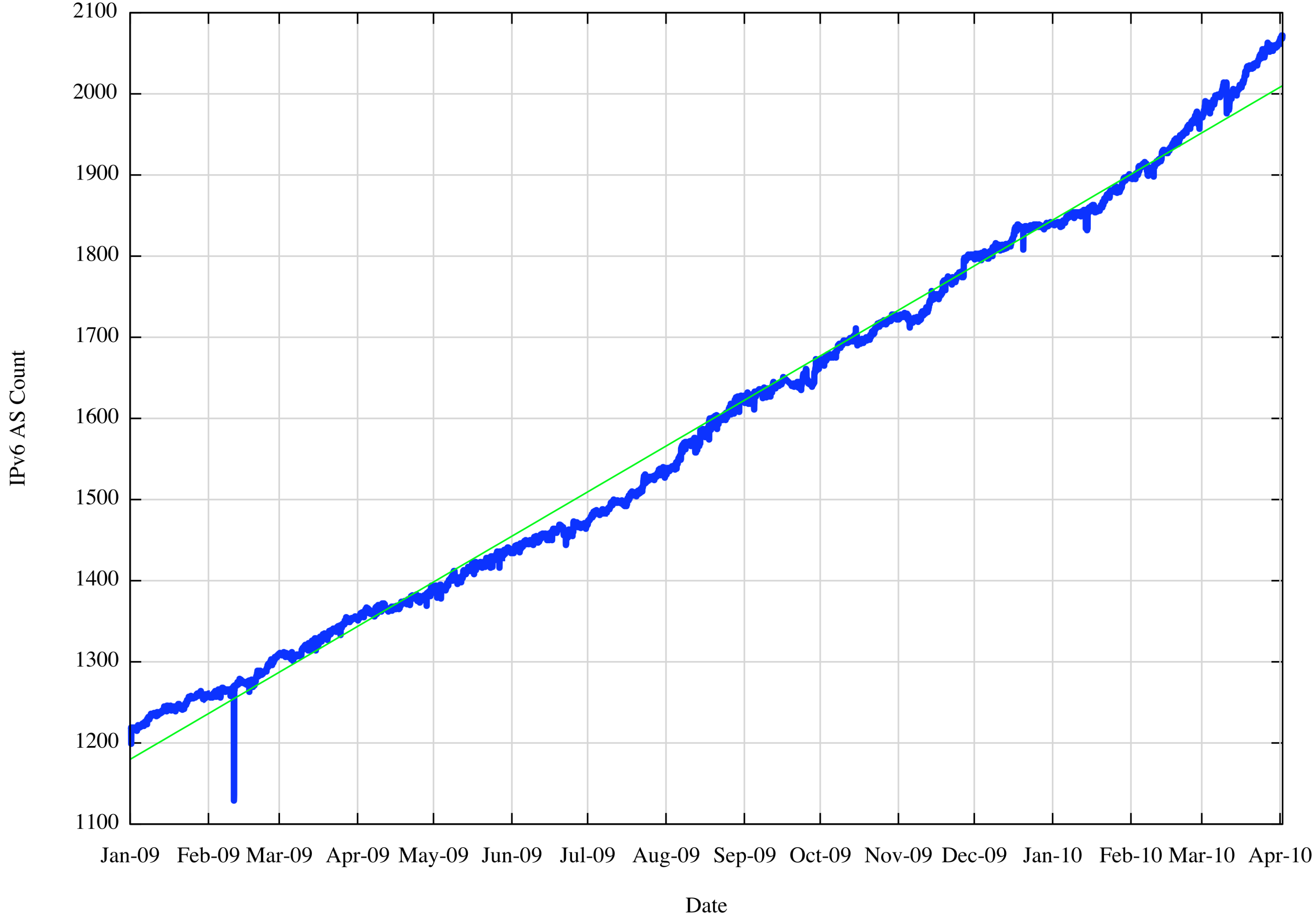
IPv6 BGP Prefix Count



IPv6 Routed Address Span



IPv6 Routed AS Count



IPv6 Vital Statistics for 2009

	Jan-09	Dec-09	
Prefix Count	1,600	2,460	**54%**
Roots	1,310	1,970	50%
More Specifics	290	490	69%
Address Span	/16.64	/16.25	**31%**
AS Count	1,220	1,830	**50%**
Transit	300	390	30%
Stub	920	1,440	56%

The Internet in 2009

The IPv6 Routing table grew by 50% over 2009

- compared with 50% growth in 2008
- The momentum of growth of IPv6 is:
  - higher than IPv4 – which is good
  - not increasing – which is perhaps not so good

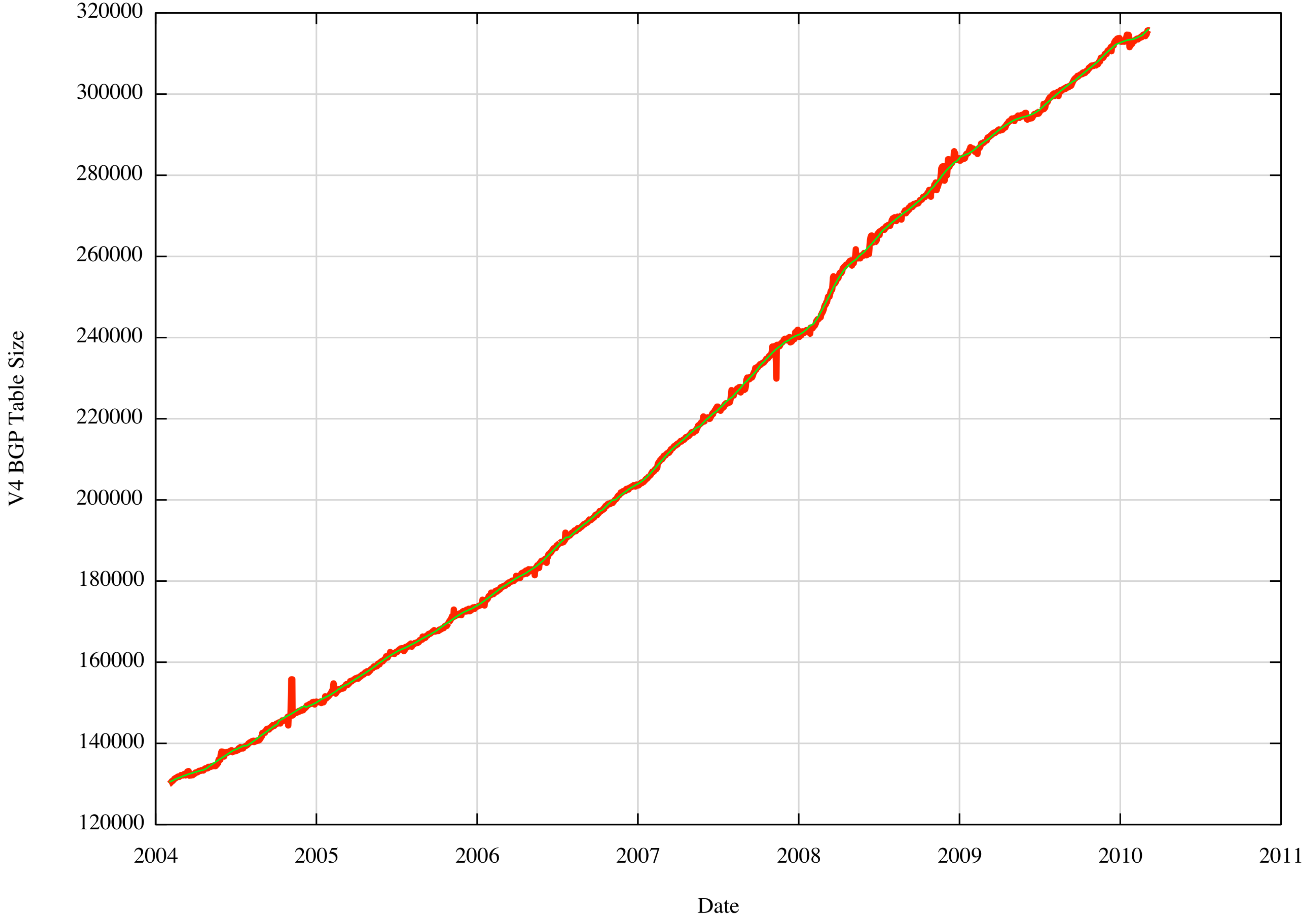
Where is this heading?

BGP Size Projections

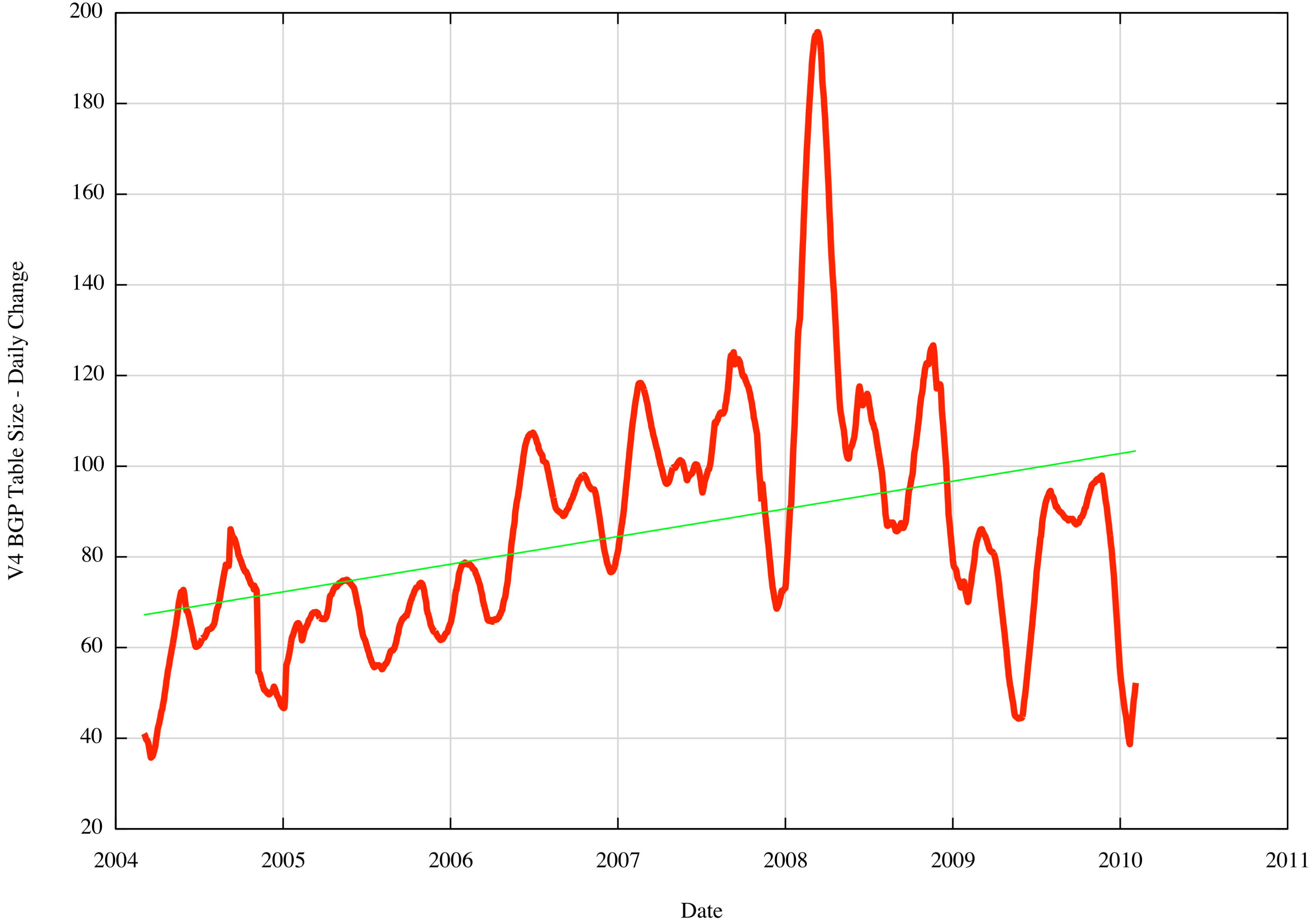
Use IP BGP table size data to generate a 4 year projection of the IPv4 routing table size

- 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 - 75 months data window

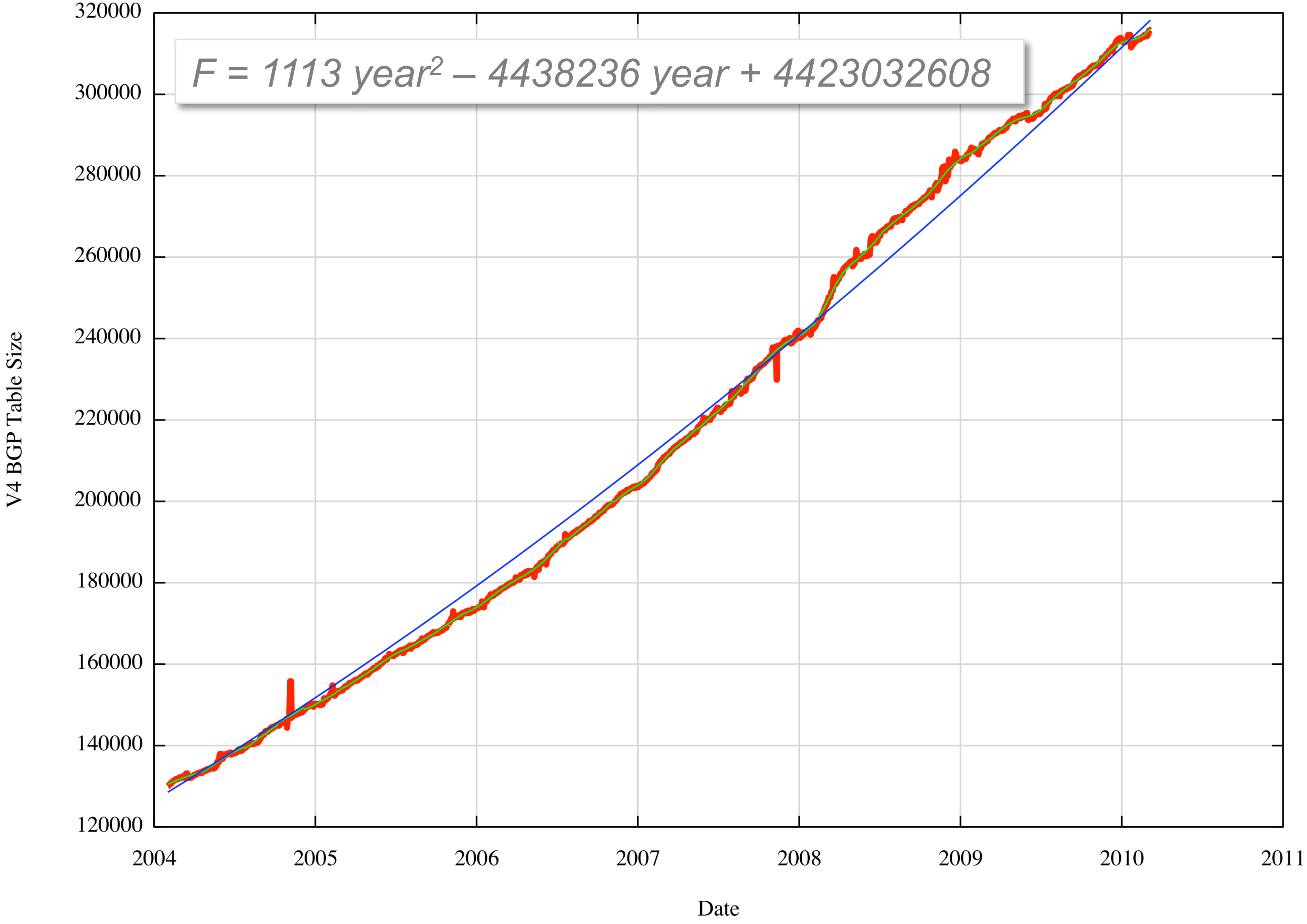


Daily Growth Rates

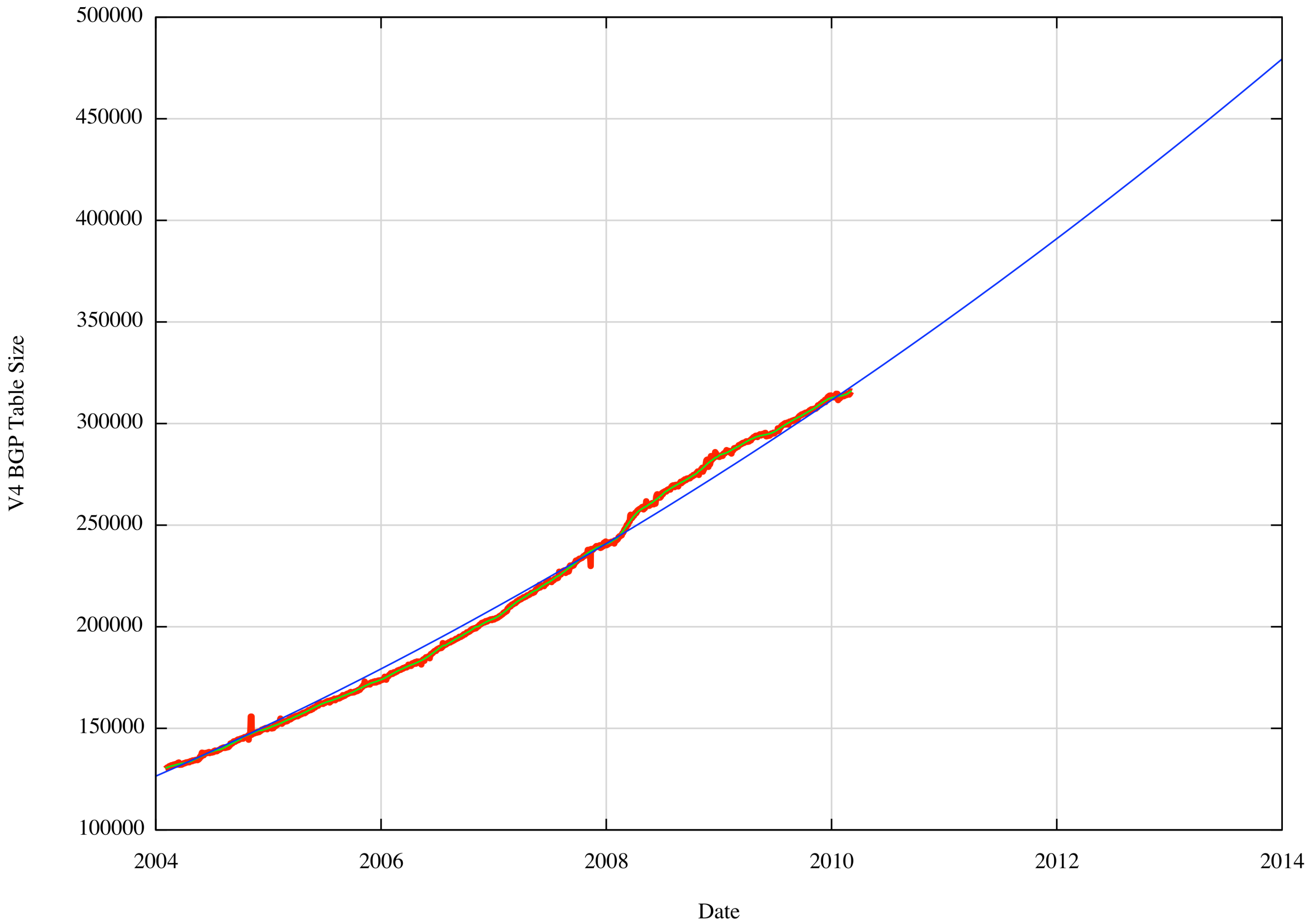


First order differential of the smoothed data

IPv4 Table Size Quadratic Growth Model



IPv4 Table Size Quadratic Growth Model - Projection

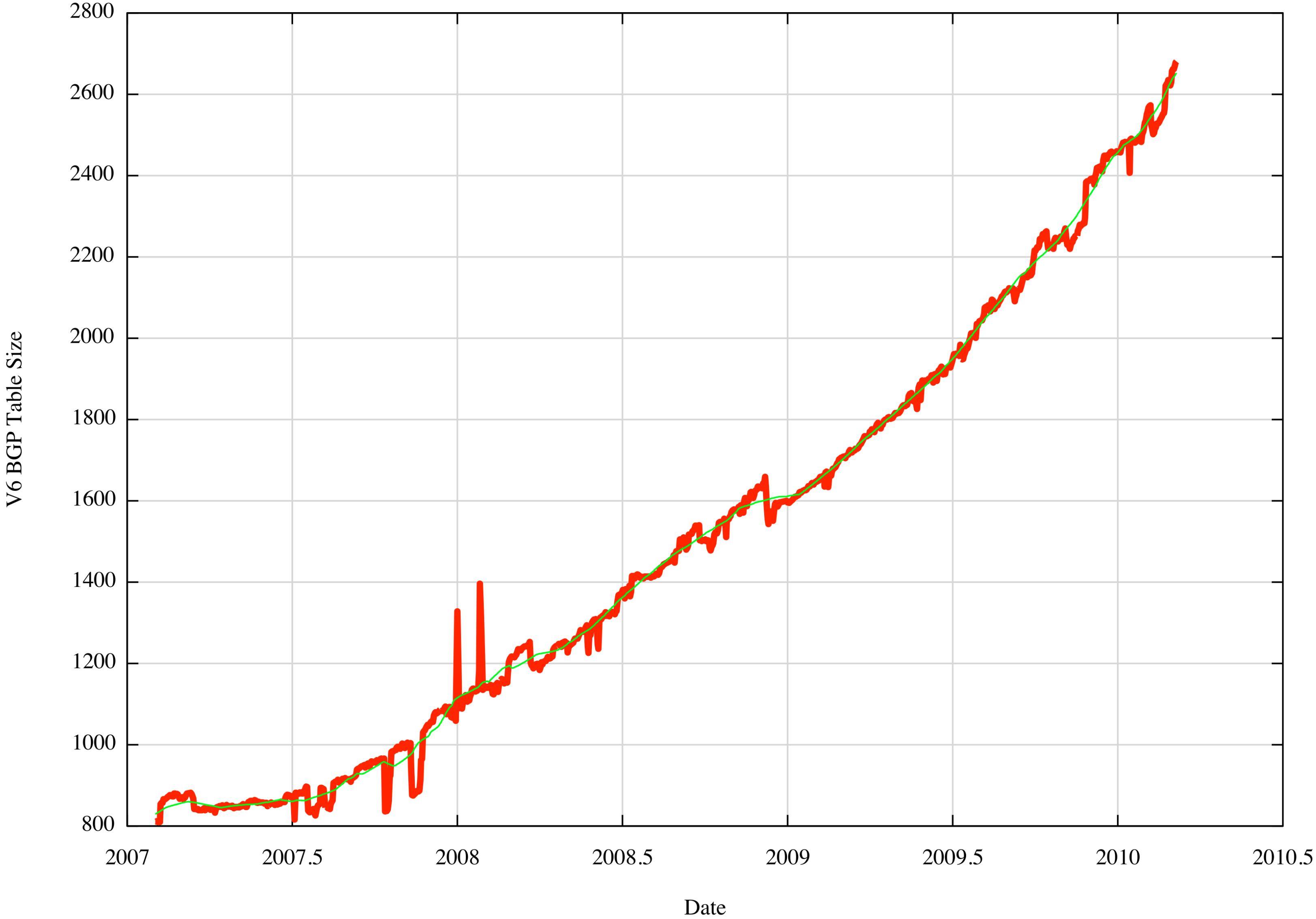


IPv4 BGP Table Size Predictions

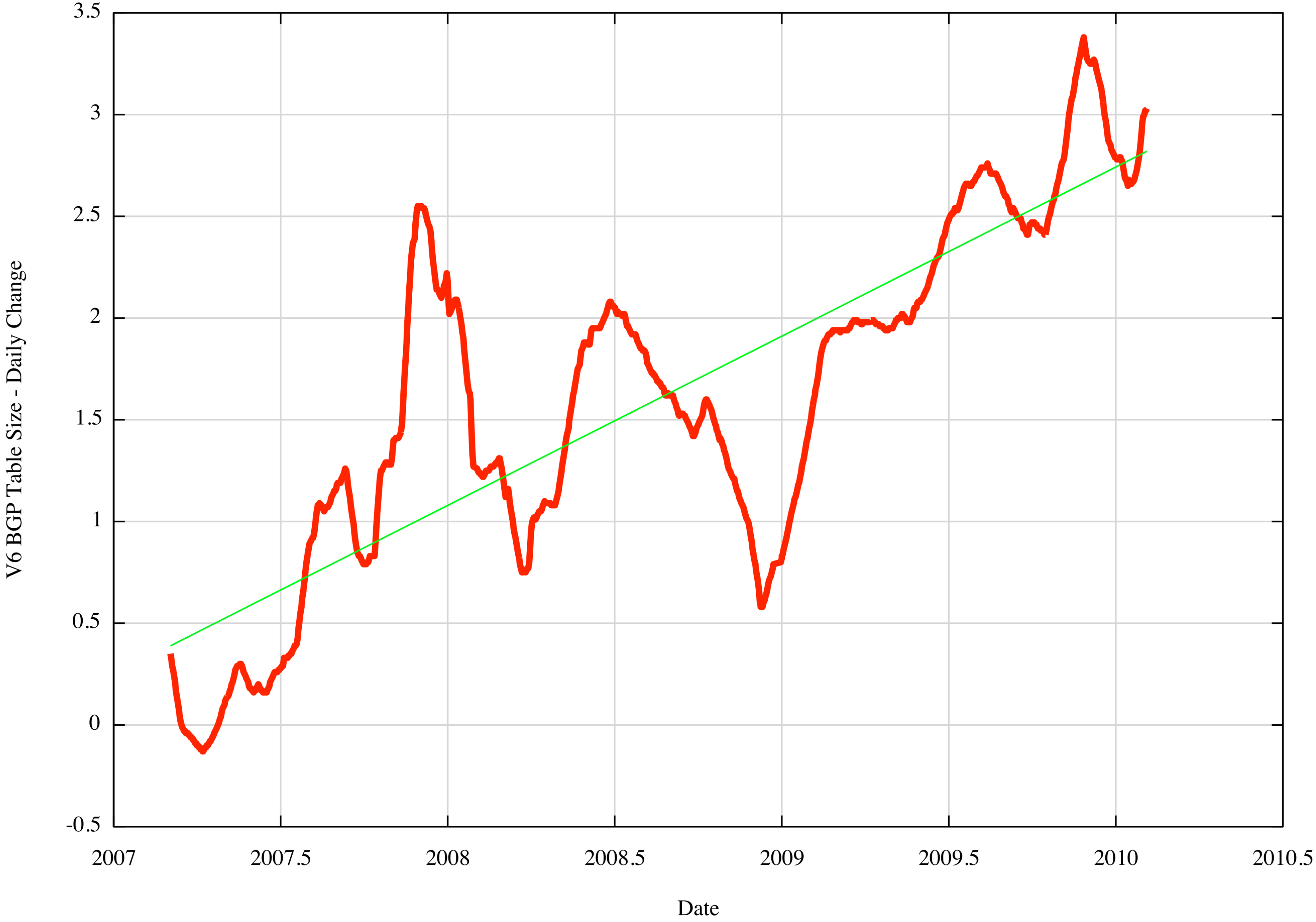
Jan 2010	313,000 entries
2011	350,000 entries
2012	391,000 entries
2013*	434,000 entries
2014*	479,000 entries

** 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.*

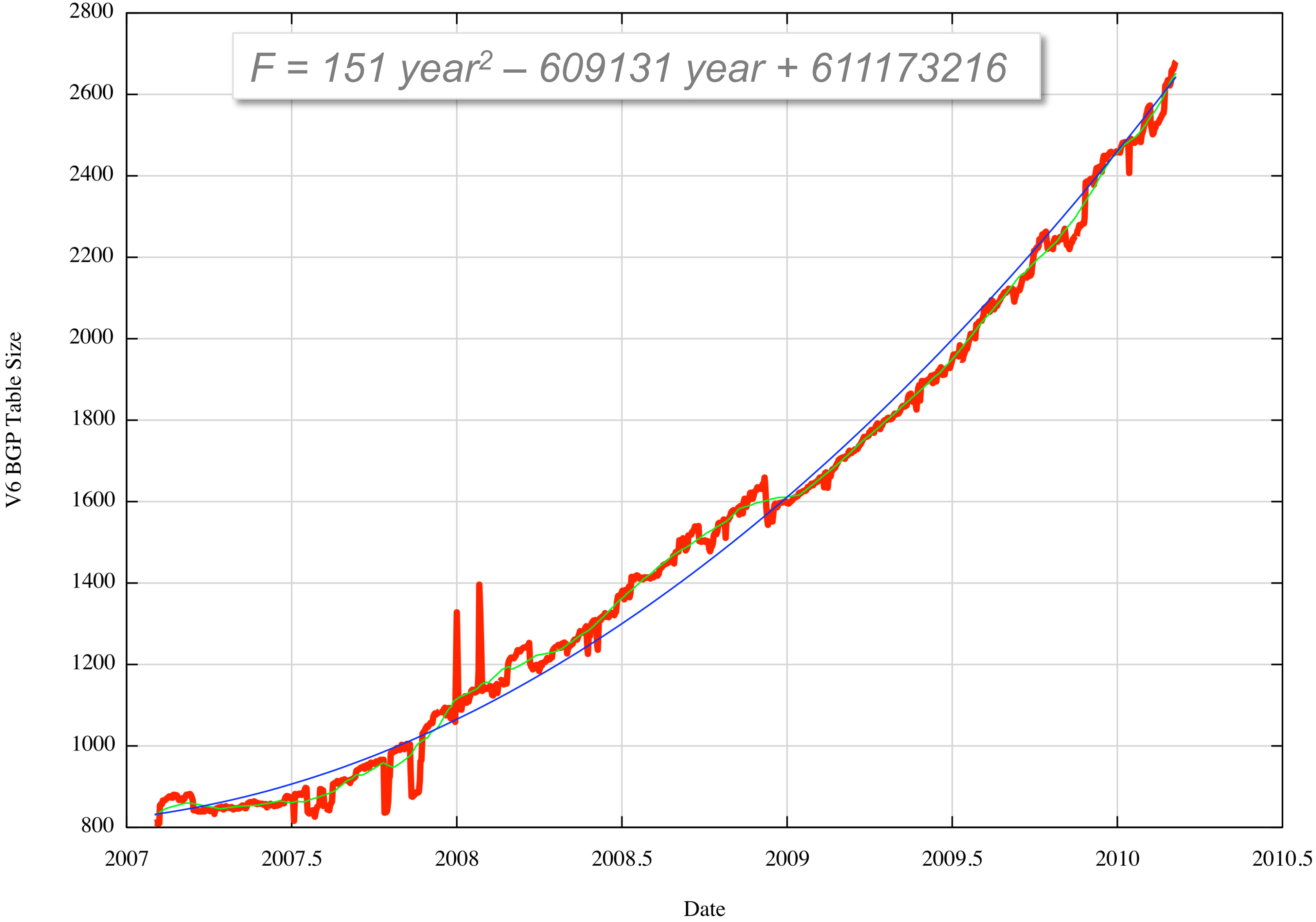
IPv6 Table Size - 39 months data window



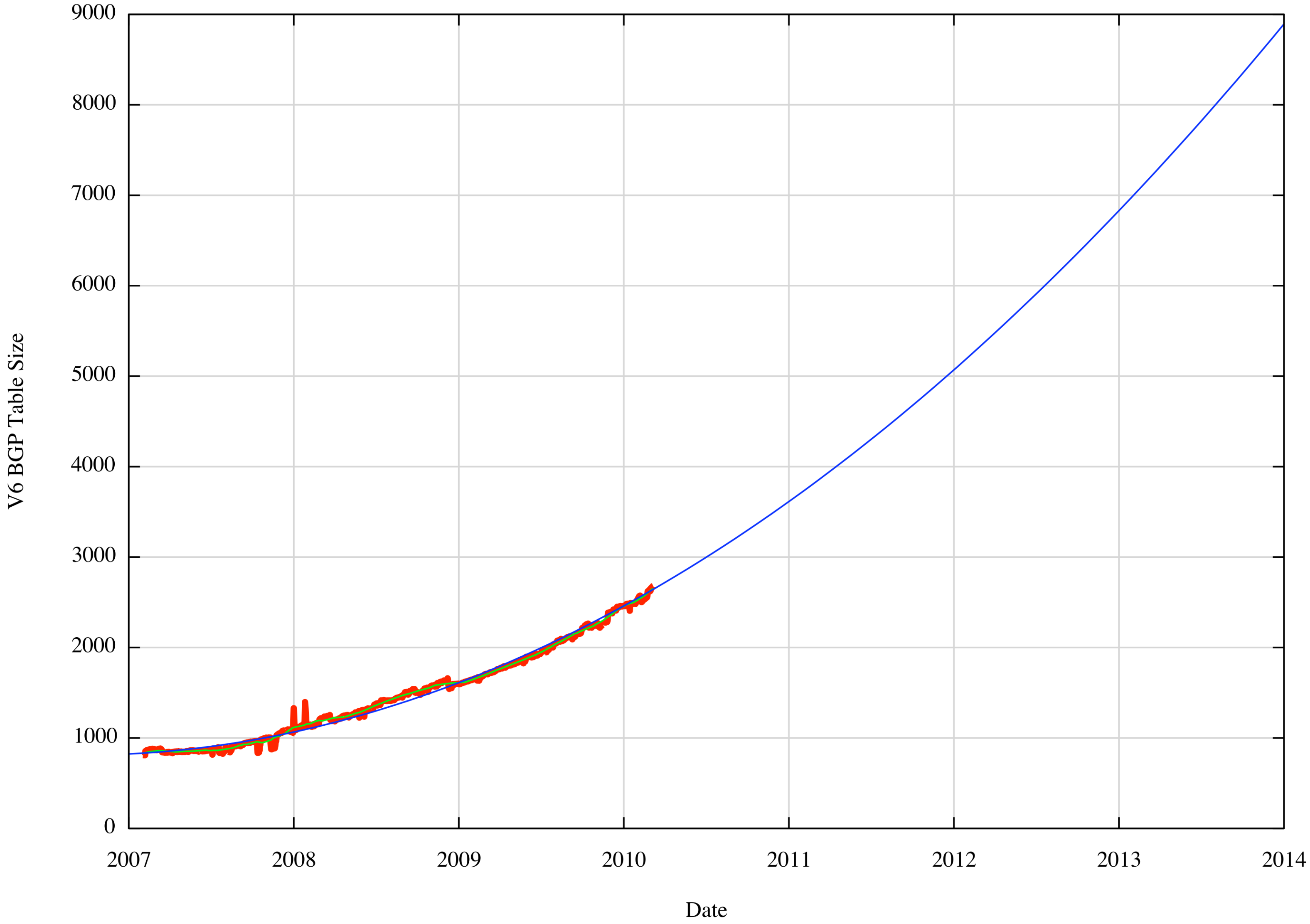
IPv6 Daily Growth Rates



IPv6 Table Size
Quadratic Growth Model



IPv6 Table Size Quadratic Growth Model - Projection



IPv6 BGP Table Size Predictions

Jan 2010	2,400 entries
2011	3,600 entries
2012*	5,000 entries
2013*	6,800 entries
2014*	8,800 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.*

BGP Table Size Predictions

Jan 2010	$313{,}000_4 + 2{,}400_6$ entries	
2011	$350{,}000_4 + 3{,}600_6$ entries	+ 12%
2012*	$391{,}000_4 + 5{,}000_6$ entries	+ 12%
2013*	$434{,}000_4 + 6{,}800_6$ entries	+ 11%
2014*	$479{,}000_4 + 8{,}800_6$ entries	+ 11%

** 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.*

BGP Scaling and Table Size

- As we get further into the IPv6 transition we may see:
  - accelerated IPv4 routing fragmentation as an outcome from the operation of a V4 address trading market that starts to slice up the V4 space into smaller routed units
  - parallel V6 deployment that picks up pace
- These projections of FIB size are going to be *low*.
- Just how low it will be is far harder to estimate.

Is this a Problem?

Is this a Problem?

- What is the anticipated end of service life of your core routers?
- What's the price/performance curve for forwarding engine ASICS?
- What's a sustainable growth factor in FIB size that will allow for continued improvement in unit costs of routing?
- A growth factor of 20% p.a. is the upper bound of anticipated trend unit cost improvements of routing hardware

Is this a Problem?

- What is the anticipated end of service life of your core routers?
- What's the price/performance curve for forwarding engine ASICS?
- What's a sustainable growth factor in FIB size that will allow for continued improvement in unit costs of routing?
- A growth factor of 20% p.a. is the upper bound of anticipated trend unit cost improvements of routing hardware

BUT:

- *What is a reasonable margin of uncertainty in these projections?*

BGP Scaling and Stability

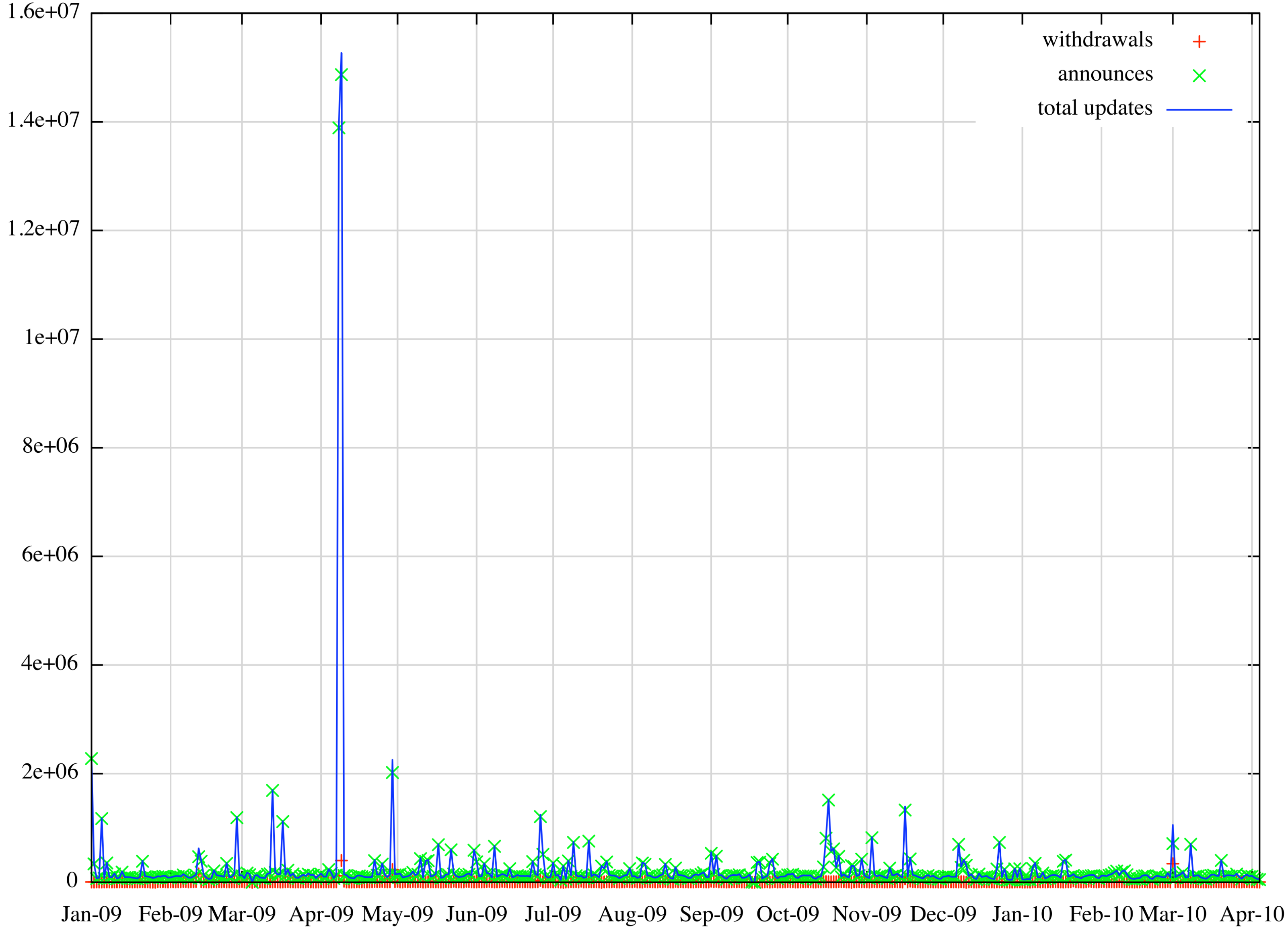
Is it the **size of the RIB** or the **level of dynamic update and routing stability** that is the concern here?

BGP Scaling and Stability

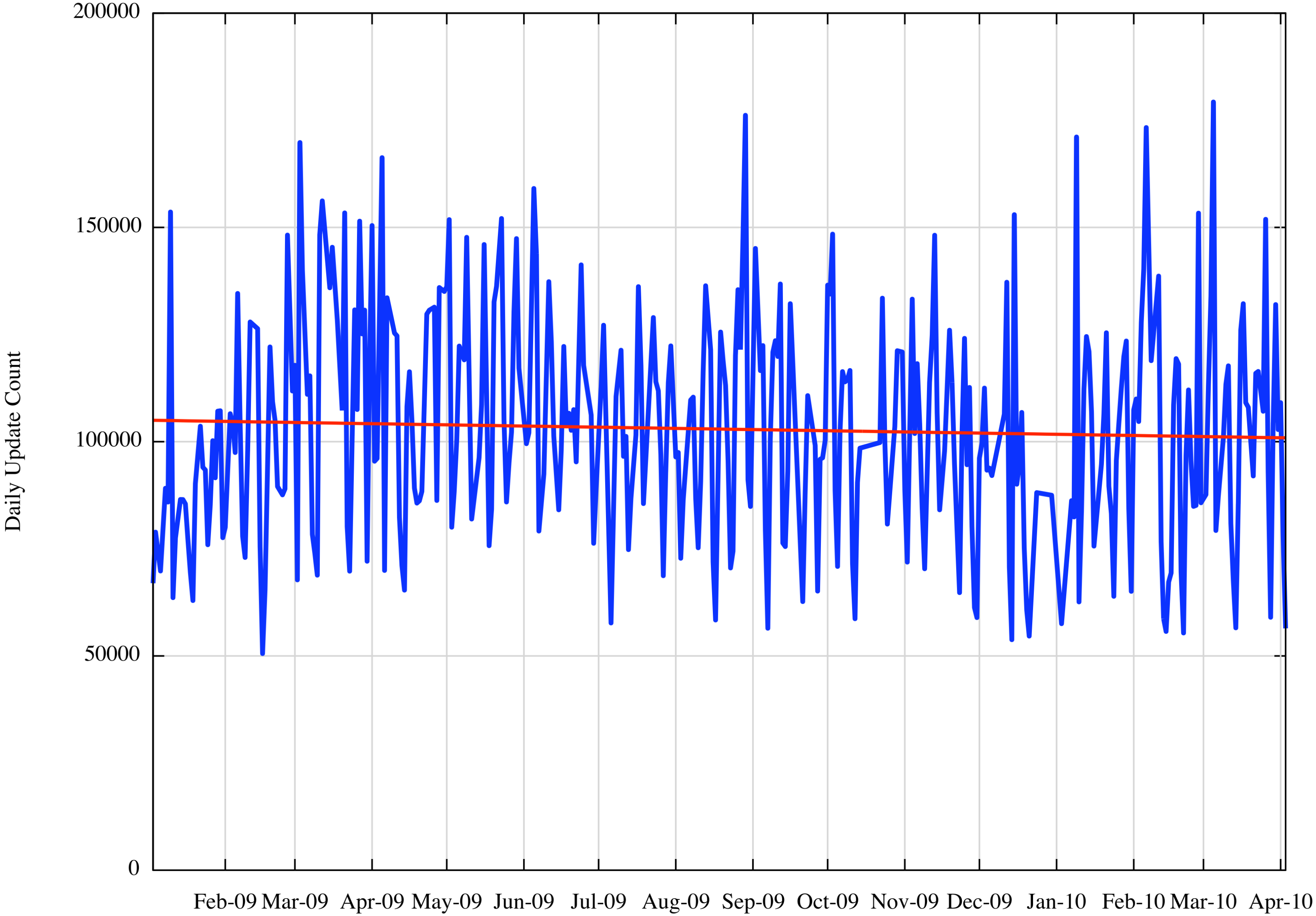
Is it the **size of the RIB** or the **level of dynamic update and routing stability** that is the concern here?

So lets look at update trends in BGP…

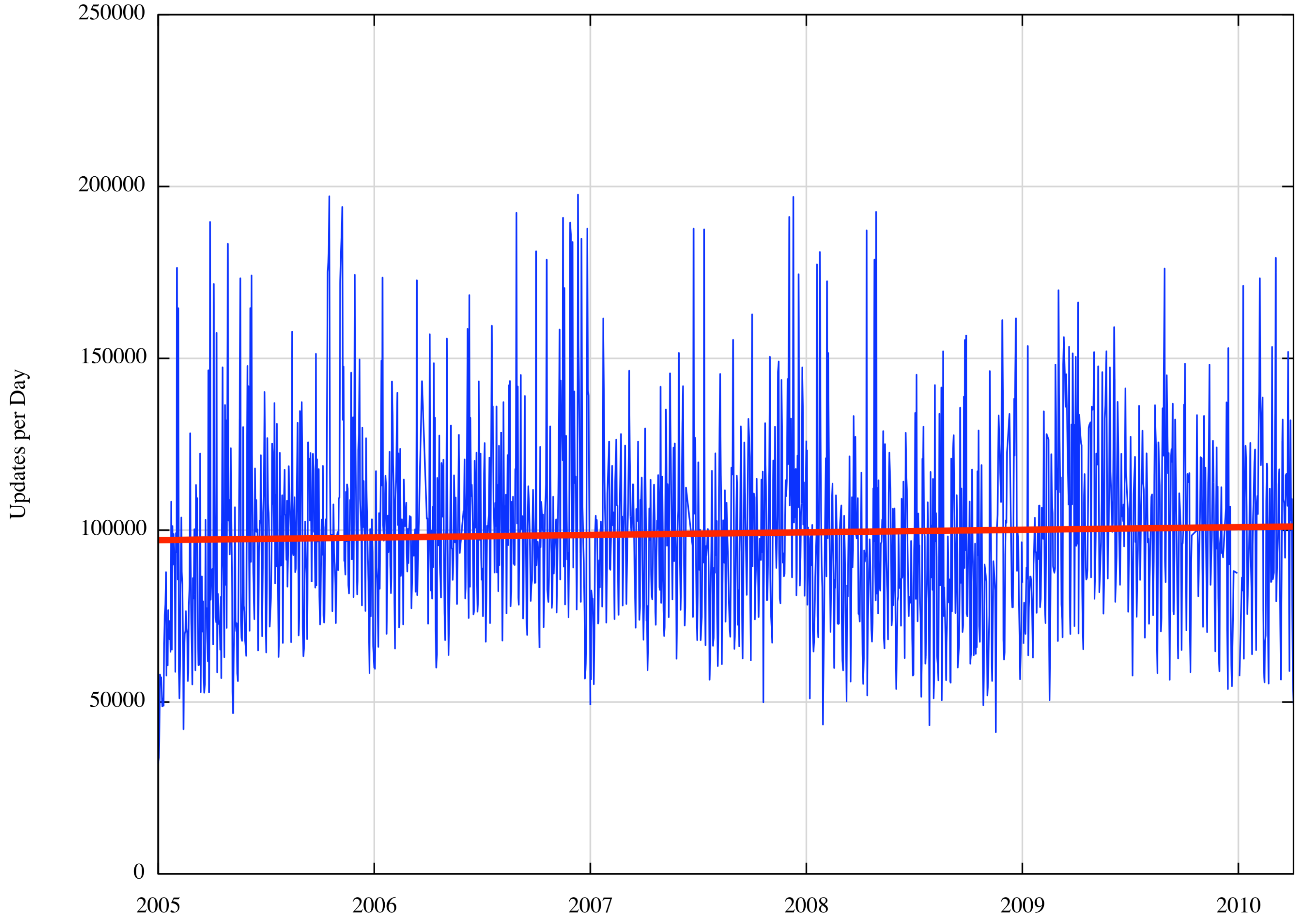
Daily Announce and Withdrawal Rates



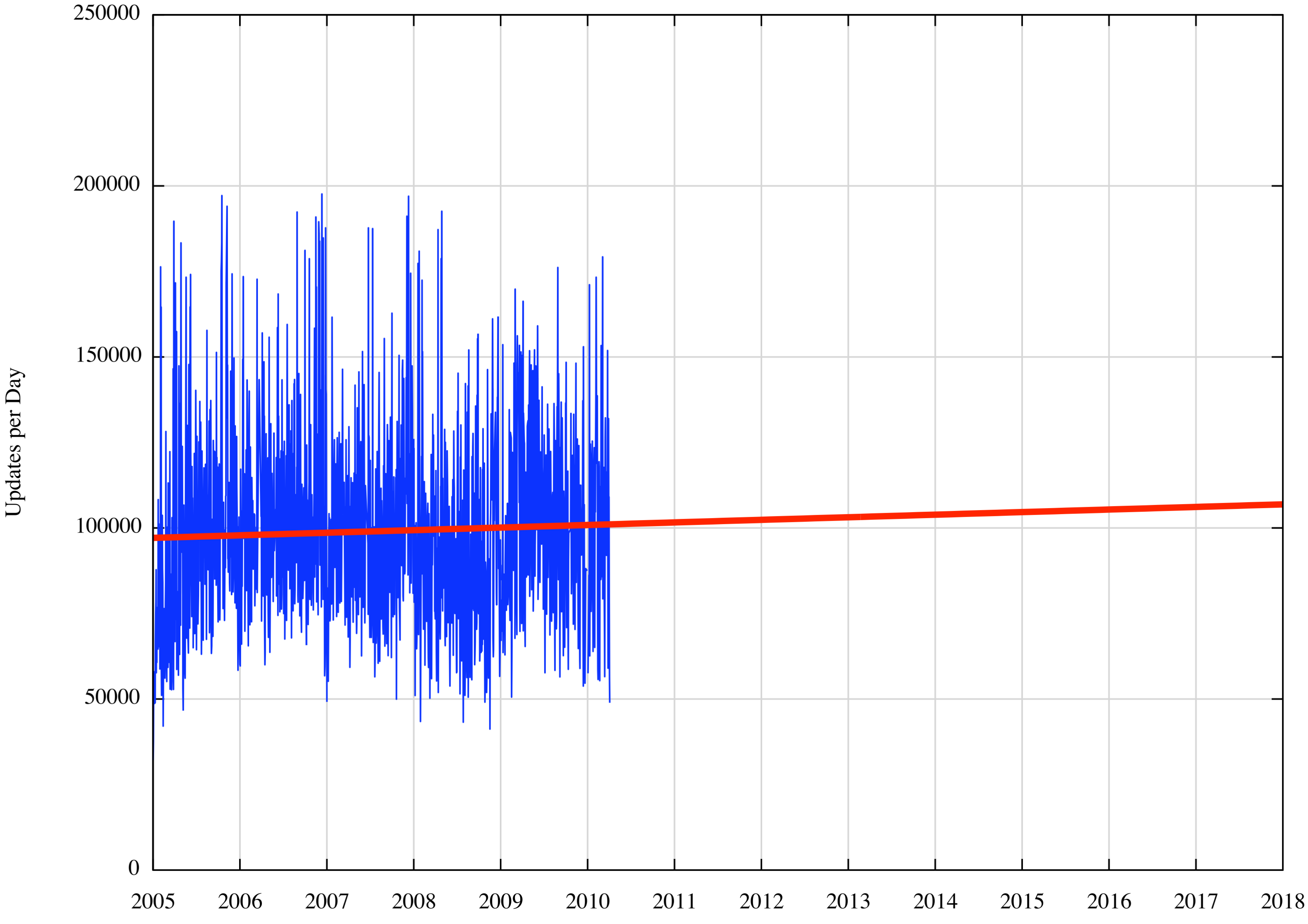
Daily Updates - 2009



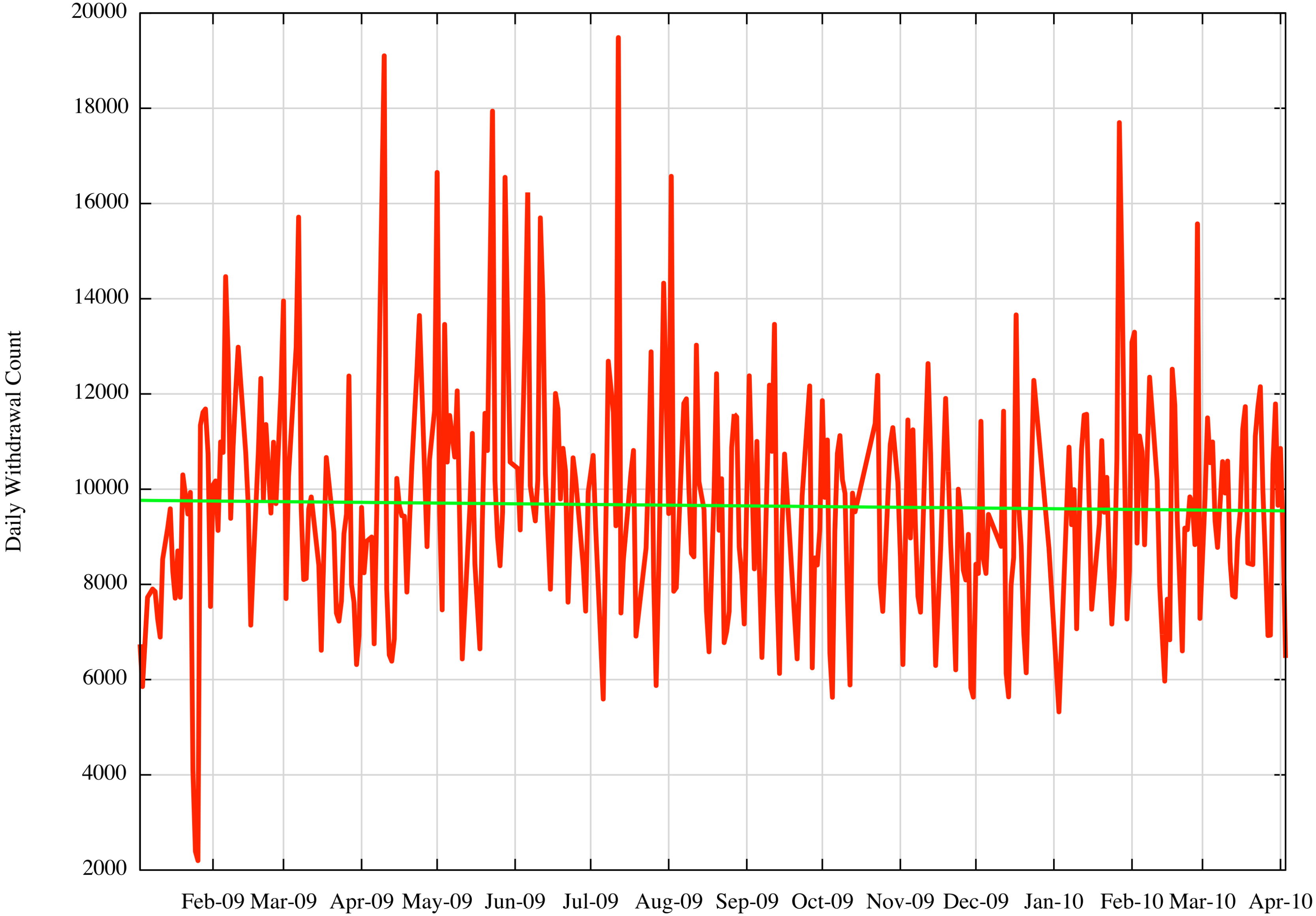
BGP Updates - 2005 - 2010
Extended Data Set



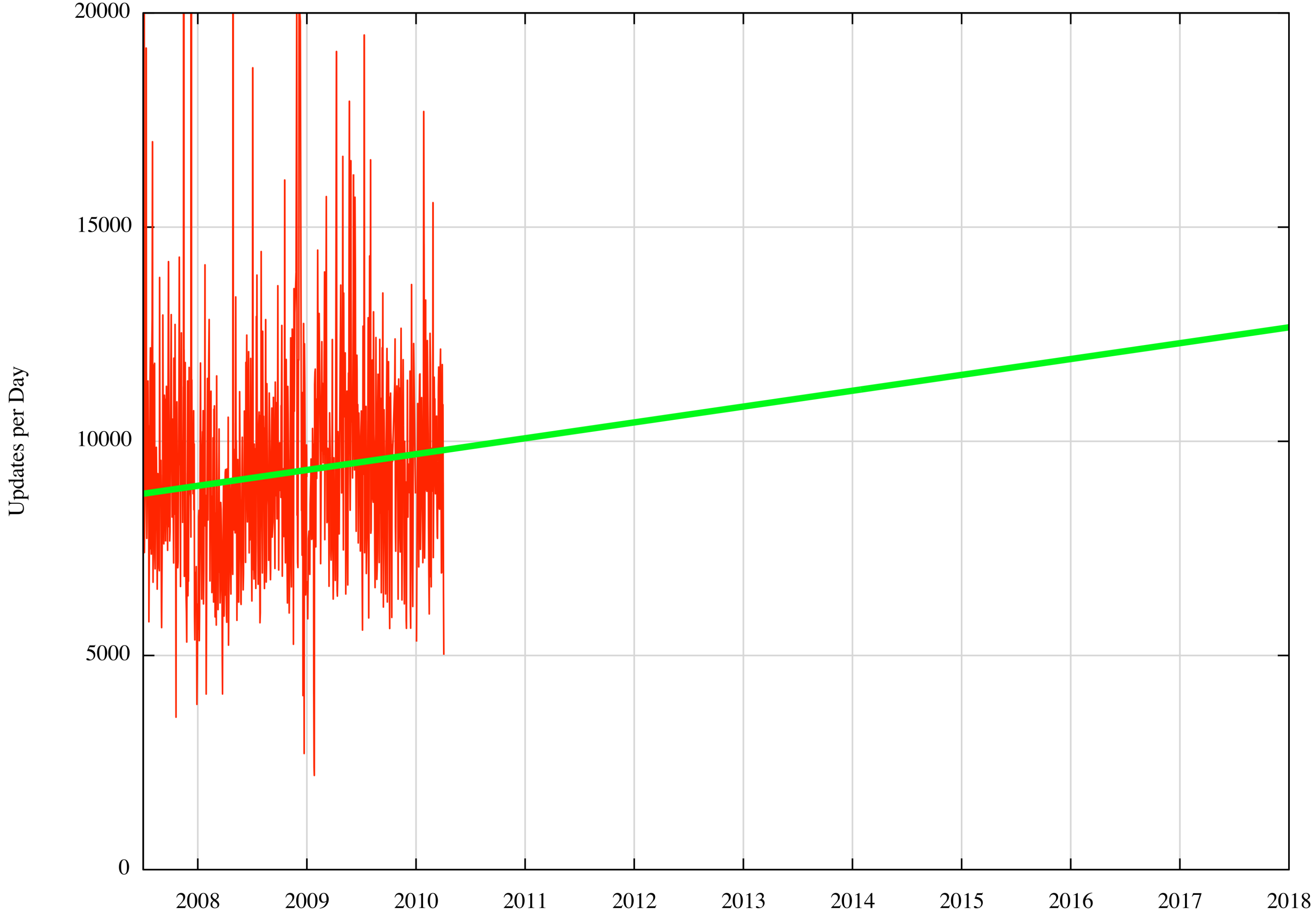
BGP Update Projection



Daily Withdrawals - 2009



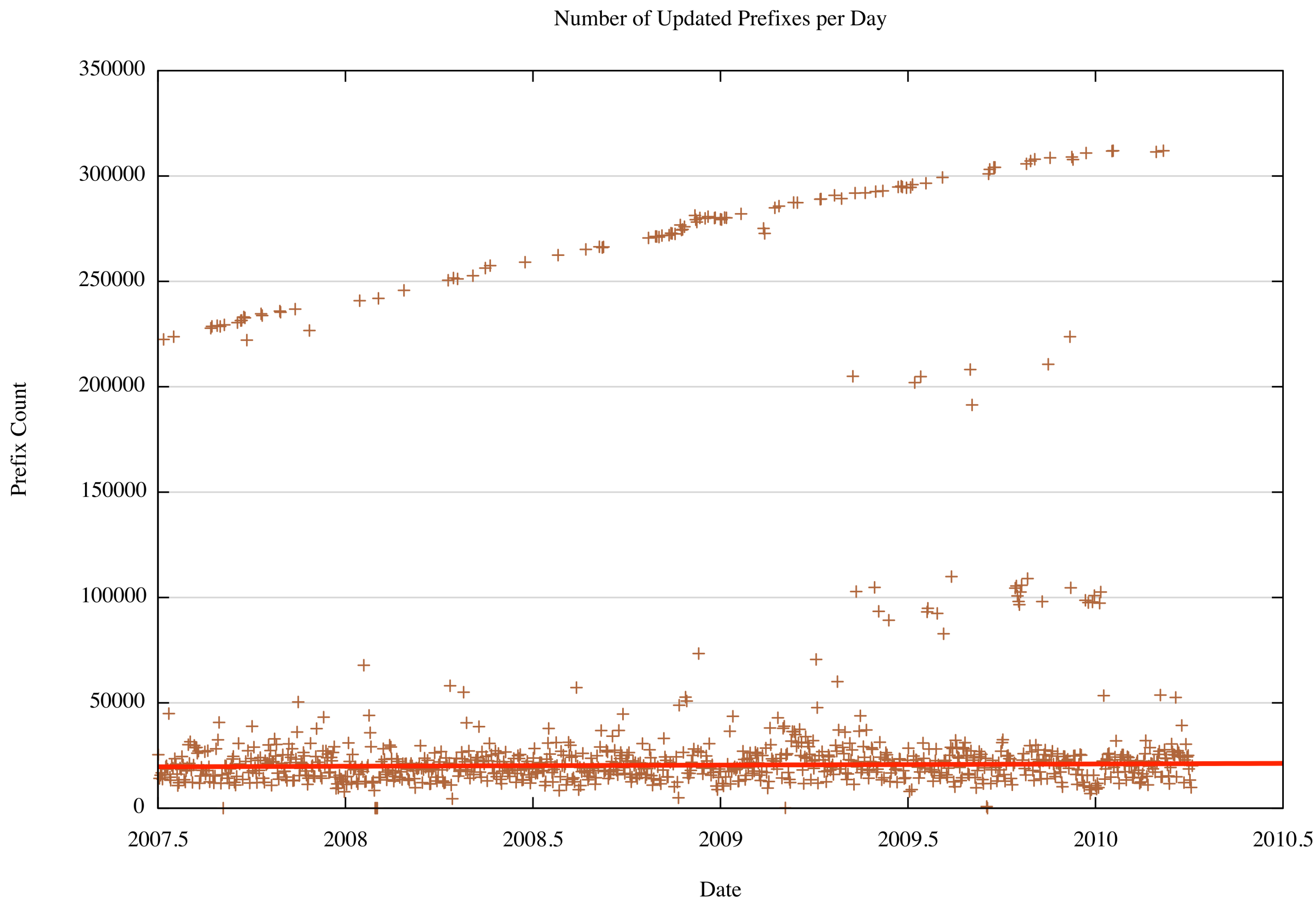
BGP Withdrawal Projection



Why is this so flat?

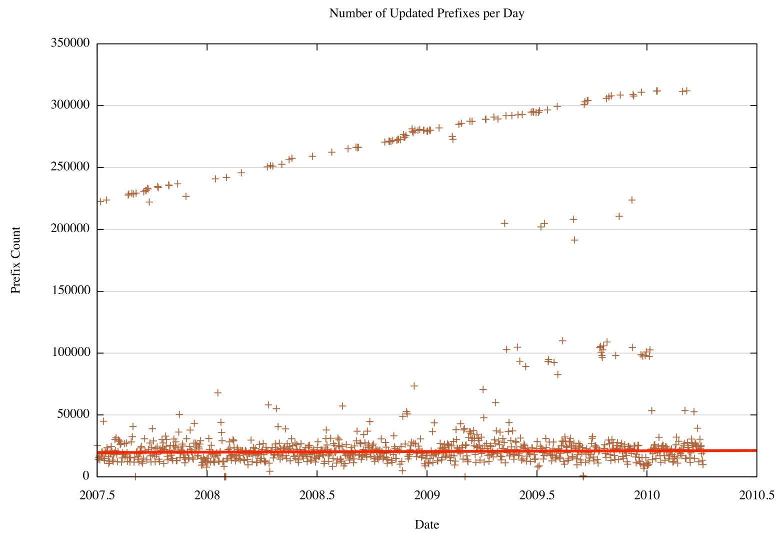
- Growth rates of BGP update activity appear to be far smaller than the growth rate of the routing space itself
- Why are the levels of growth in BGP updates ***not*** proportional to the size of the routing table?

(In)Stability



(In)Stability

- Over the past 1,000 days the number of announced prefixes increased by 40% (225,000 to 320,000)
- But the average number of unstable prefixes on any day increased by only 7% in 1,000 days(19,600 to 21,000)
- Routing instability is **not** directly related to the number of advertised objects
- What is routing instability related to?



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

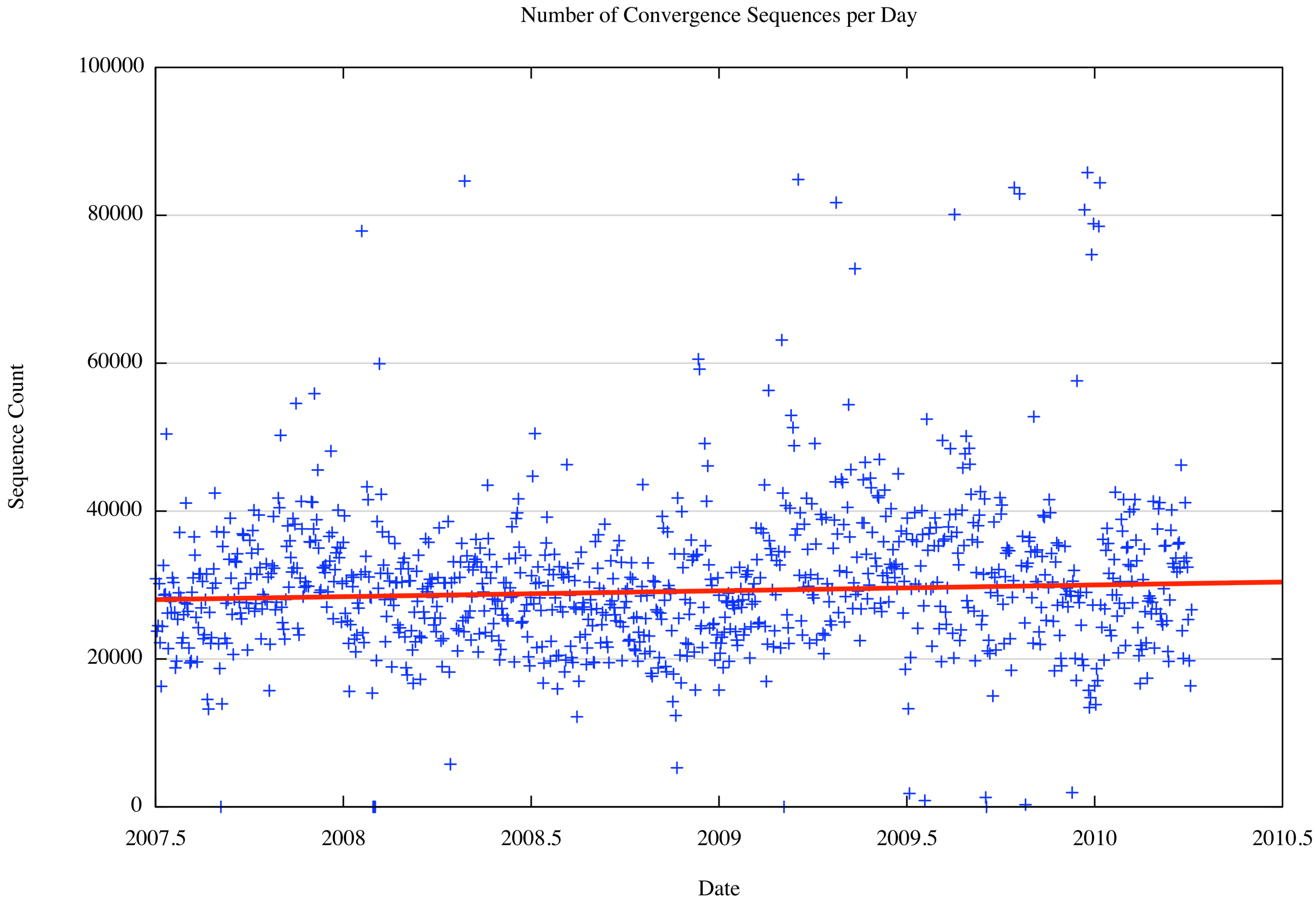
(In)Stability

- There are two types of updates:
  - updates that are part of a convergence sequence
  - updates that are single isolated events that are not part of a convergence sequence - solitons

(In)Stability

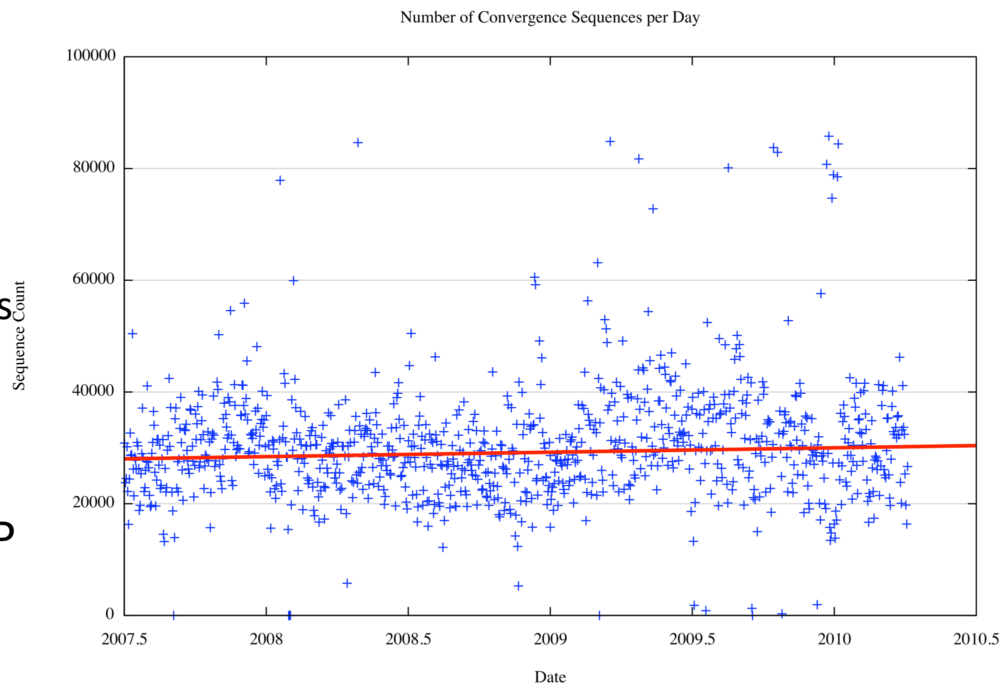
- There are two types of updates:
  - updates that are part of a convergence sequence
  - updates that are single isolated events that are not part of a convergence sequence - solitons

Measurement Approach for stability behaviour

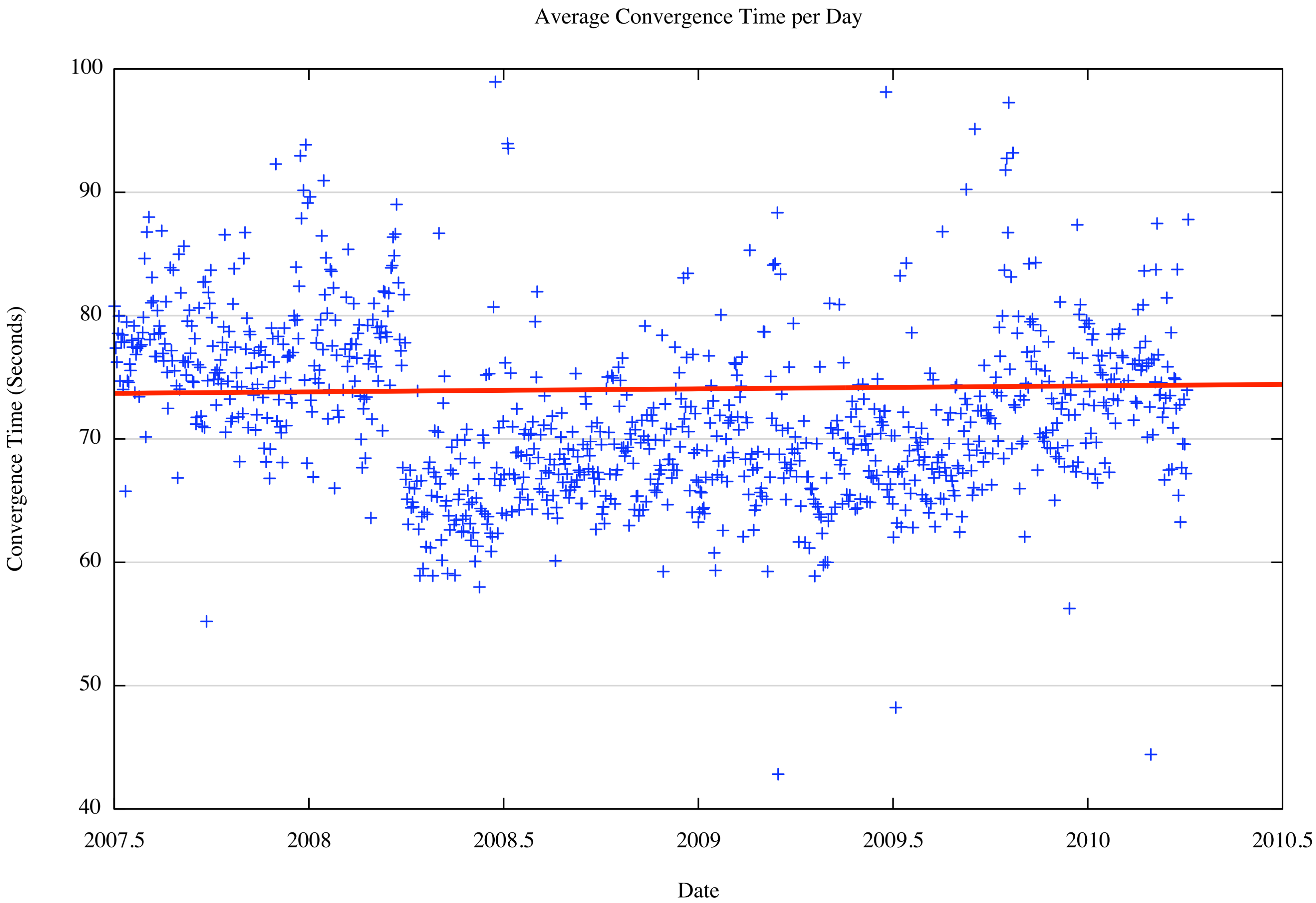


Measurement Approach for stability behaviour

- 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)
- The number of "convergence sequences" per day has been steady between 20,000 to 40,000 over the past ~3 years

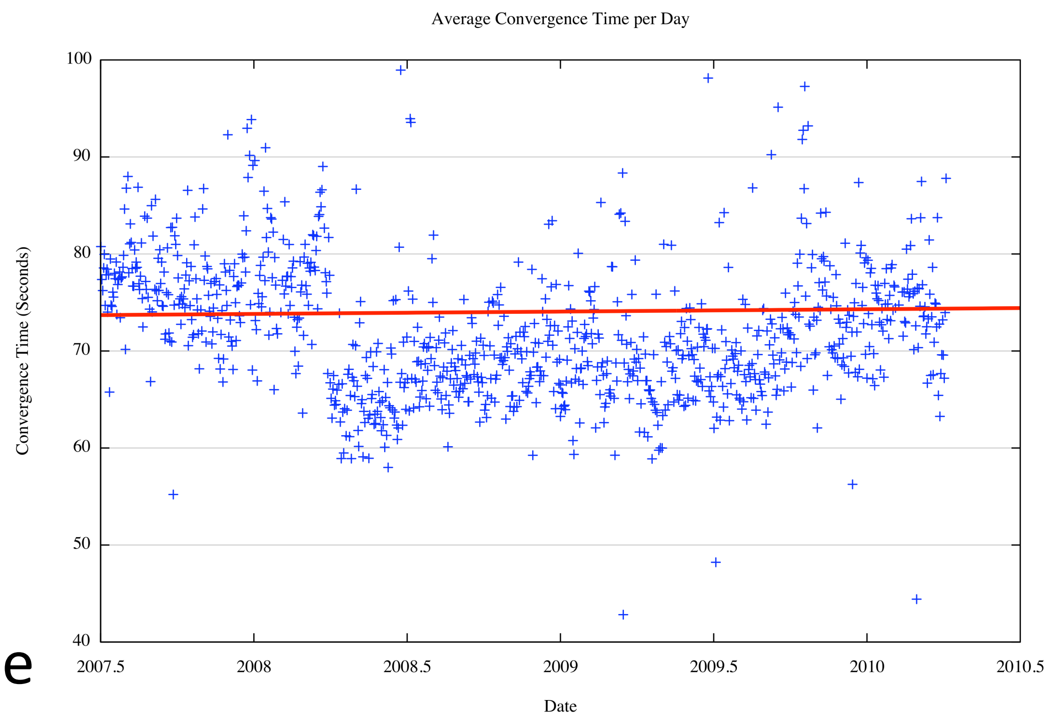


Average Convergence Time

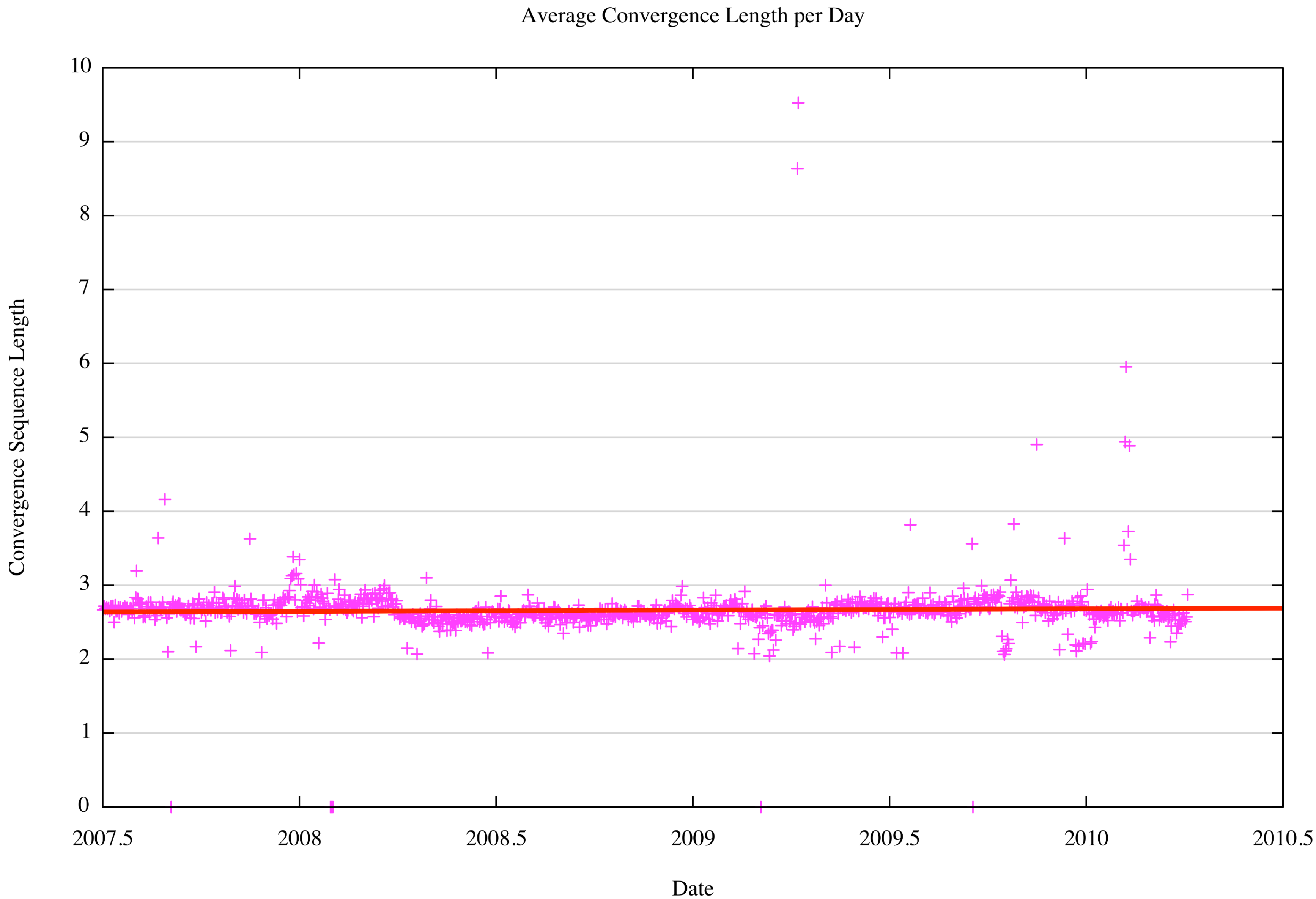


Average Convergence Time

- An unstable prefix takes, on average around 70 seconds to reach a stable state
  - given the 30 second MRAI timer constraints this approximates to between 2 and 3 MRAI intervals.
- This has remained constant for almost two years
- As the network expands, the distance vector operation to achieve convergence is taking the same elapsed time

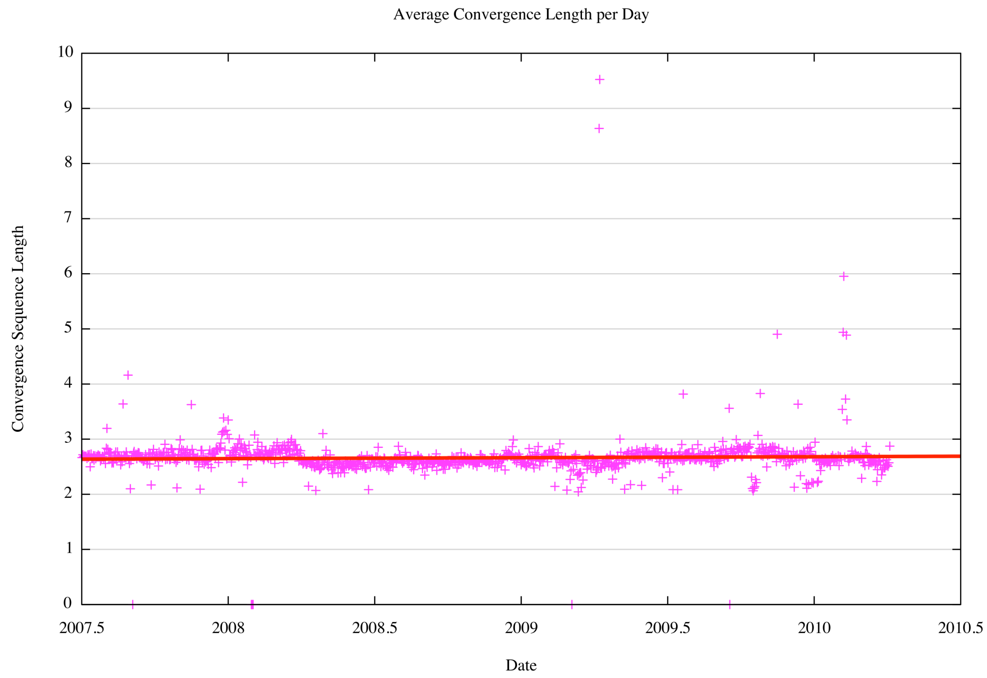


Average Convergence Updates



Average Convergence Updates

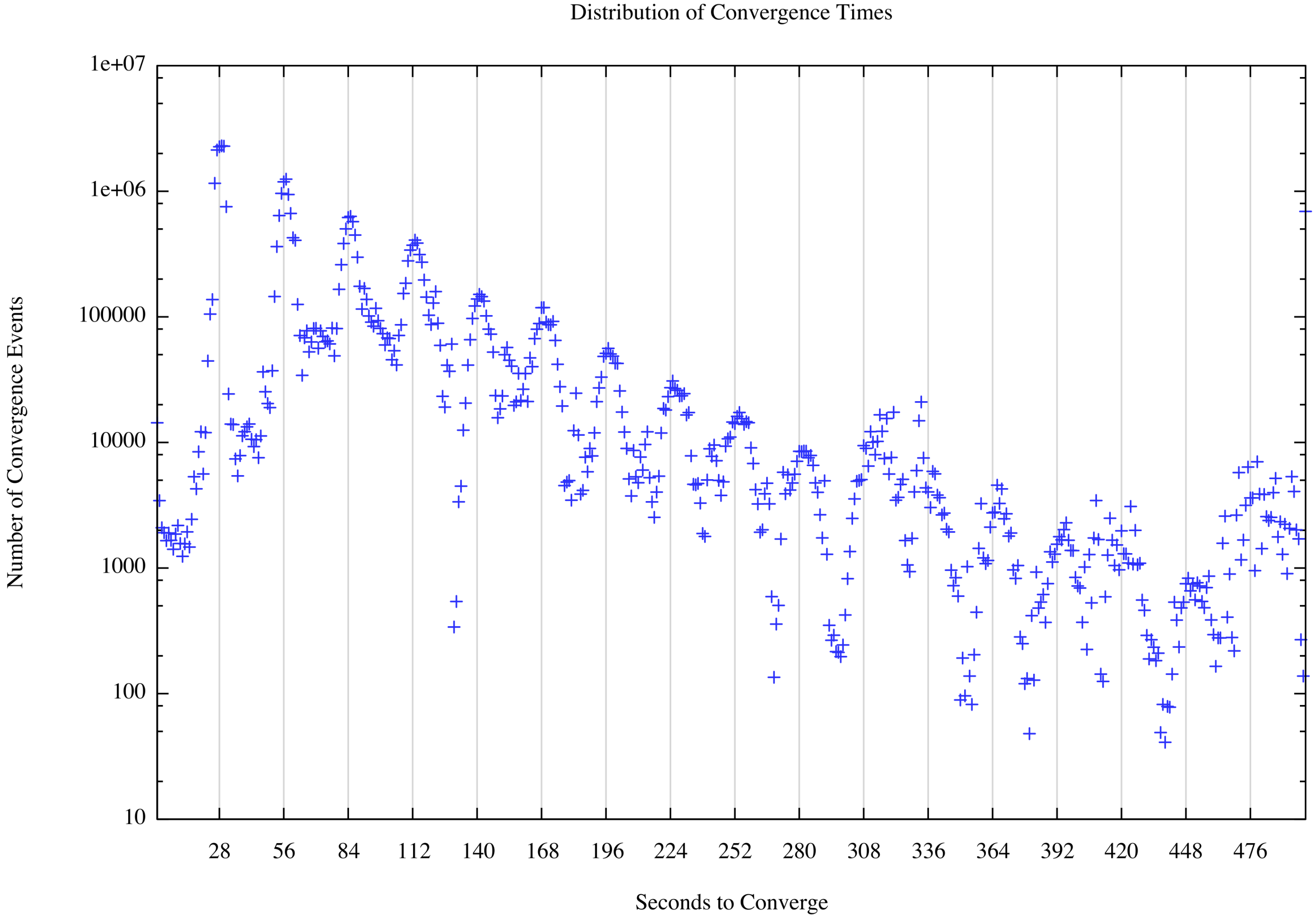
- The average number of updates to reach a converged state has remained constant for the past 2 ½ years at 2.7 updates
- The growth of the network appears to have been achieved by increasing the density of connectivity, rather than increasing the network's diameter



Convergence Trends

- Why is BGP so stable in terms of convergence behaviour?
- Why is convergence behaviour not directly related to the size of the network?
- Is the a general trend, or a case of a skewed distribution driving the average values?

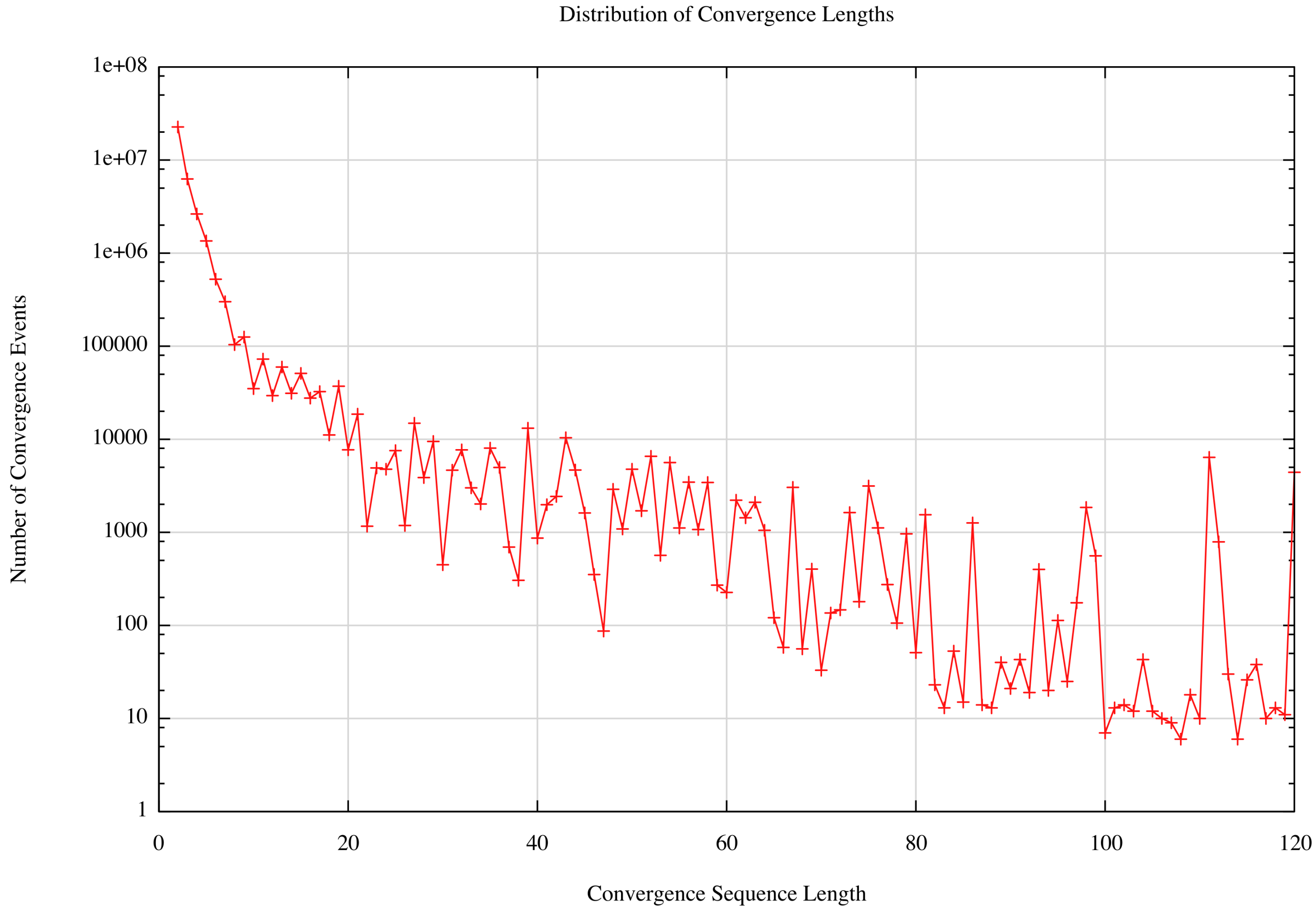
Convergence Distribution



Time to reach converged state has strong 28 second peaks

Default 27 -30 second MRAI timer is the major factor here

Convergence Distribution



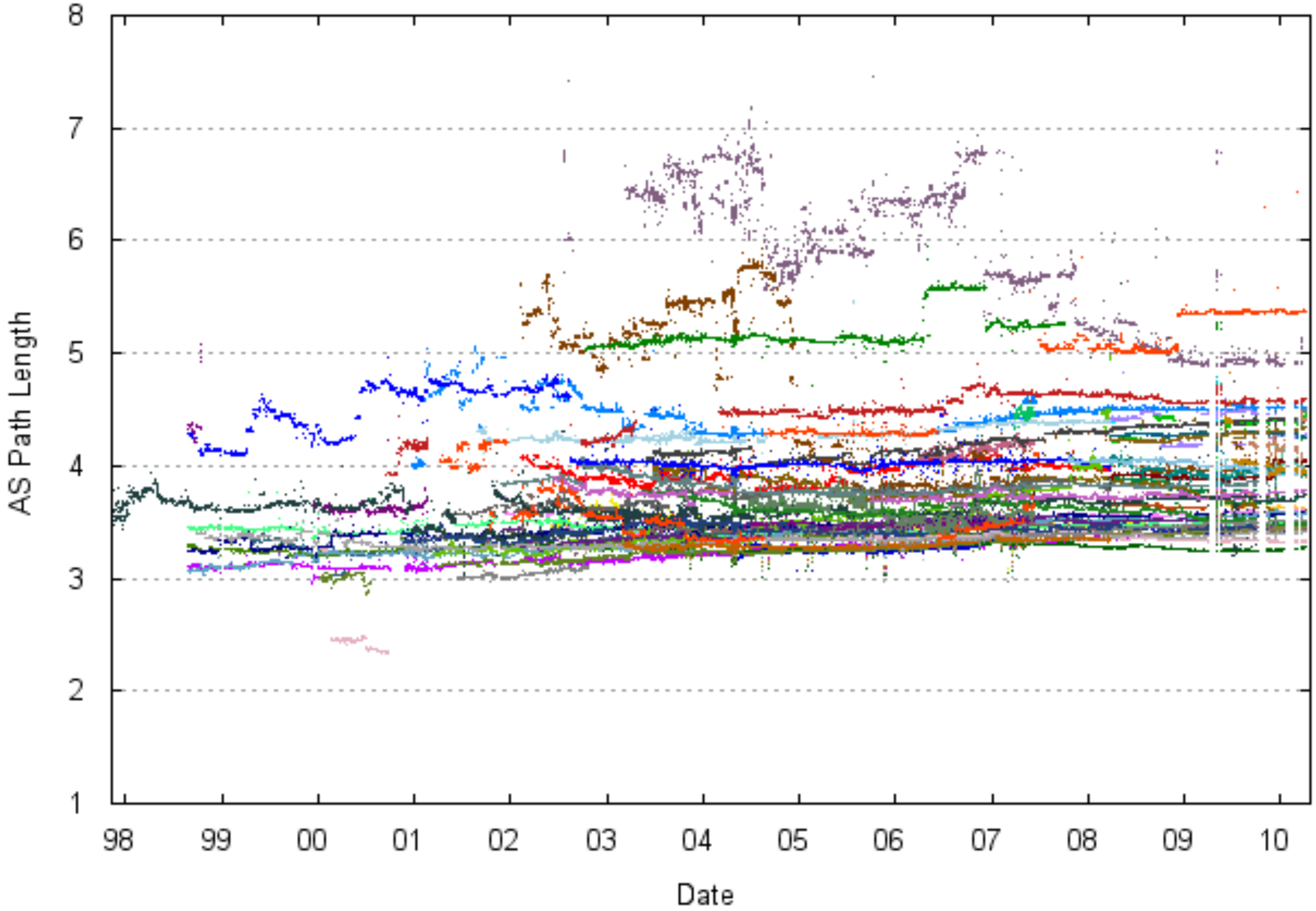
Number of updates to reach convergence has exponential decay in the distribution.

Convergence Distribution

Observations

- There is a reasonable correlation between AS Path Length Distribution and Convergence Update Distribution
- The number of updates to reach convergence and the time to reach convergence is related to AS Path Length for most (98.66%) of all instability events
- Persistent instability events (1.3% of all such events) are probably related to longer term instability that may have causes beyond conventional protocol convergence 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 of 30,000 ASes is more densely interconnected, but not more "stringier" than the internet of 5,000 ASes of 2,000
- This is consistent with the observation that the number of protocol path exploration transitions leading to convergence to a new stable state is relatively stable over time

Is BGP Scaling?

Is BGP Scaling?

So Far, So Good!

Will BGP Continue to Scale?

Will BGP Continue to Scale?

Only if:

- the address system continues to maintain strong alignment with network topology
  - provider-based addressing policies appear to assist in maintaining a viable global routing infrastructure
  - continued awareness of address aggregation in the operational community
- further growth of the network is matched with increased inter-connectivity
  - Local Exchanges and Regional / Global Transit Providers both play beneficial roles in limiting the diameter of a constantly growing network

Thank You