Origin Disturbances in BGP

Dan Pei, William Aiello, Anna Gilbert, Patrick McDaniel

Paper ID: E00-1465139405
Number of Pages: 14

Abstract—**This paper develops an empirical profile of BGP prefix announcements that originate from multiple ASes, so-called MOAS announcements. Analysis of Oregon RouteViews data over one year shows that a small fraction of prefixes are responsible for a very large fraction of all origin AS transitions observed at RouteViews. Moreover, these heavy-hitter prefixes oscillated between two origin ASes. The prevalence of this behavior indicates that a clear profile of its characteristics will inform a larger understanding of MOASes and ultimately BGP.**

The central contribution of this paper is a detailed analysis of these MOAS multihoming oscillations at different time scales. We empirically derive a model of AS disturbance periods during which the origin AS observed oscillates with heavy tailed holding times. We demonstrate that these disturbances arrive according to a Poisson process. We also show that the update stream within these disturbances exhibits long range dependence. Using simulations, and physical-based modeling of events at origin to drive these simulations, we demonstrate that heavy-tailed oscillation at the origin is a possible explanation for our observations (while the complex interplay of the BGP protocol and network topology is not such an explanation). Comparison with BGP beacon data verifies our simulations that discrete and singular events at the origin do not generate heavy-tailed oscillations at the viewpoint. In sum, we find that AS oscillations driven by heavy-tailed oscillations between different multihomed providers are a widespread and important BGP phenomenon with complex but recognizable signatures such as heavy-tailed holding times and long-range dependence.

I. Introduction

This paper develops an empirical profile of BGP prefix announcements that originate from multiple ASes, so-called MOAS(Multiple Origin AS) announcements [1]. Our model is driven by solid data analysis and physical-based simulations. We analyze BGP updates observed at Oregon RouteViews [2] from August, 2002 to July, 2003 for origin AS transitions. Surprisingly, a small fraction of prefixes are responsible for a very large fraction of all transitions. Moreover, these heavy-hitter prefixes oscillated between two origin ASes. In fact, in our dataset 23% of origin transitions were associated with the 473 prefixes that oscillated at least 50 times. The prevalence of this behavior indicates that a clear profile of its characteristics will inform a larger understanding of MOASes and ultimately BGP.

Several varieties of multihoming can lead a prefix to be legitimately announced by two (or more) ASes. We hypothesize that the origin AS oscillations in our dataset are the result of these types of multihoming which we denote MOAS multihoming. The central contribution of this paper is a detailed analysis of these MOAS multihoming oscillations at different time scales. We hypothesize the following model. There are disturbance periods during which the origin AS (observed at distant viewpoints) oscillates with heavy tailed holding times. The "arrival" of these disturbances is given by a Poisson process. That is, the distribution of the times between the beginning of consecutive disturbance events is exponential.

To test this hypothesis we break up the oscillations into disturbances periods that are separated by quiescent periods; *i.e.*, periods without an origin transition. We provide empirical evidence for the model, demonstrate that the heavy-tailed AS holding times are stable over time, and postulate that the tail parameter of the AS holding times may be a useful signature of legitimate MOAS for such prefixes. We also discover that within each disturbance the number of updates as a function of time exhibits long range dependence over seven or eight time scales (roughly one hour).

We postulate two potential causes of the heavy-tailed oscillations within disturbances at the RouteViews viewpoint. In the first model, there is a single discrete cutover from the AS of one MOAS multihoming provider to the other. Then interactions between BGP and complex AS topology result in the heavy-tailed AS oscillations at the viewpoint. The second model postulates heavy-tailed oscillations between the two MOAS multihoming ASes which drive the oscillations at RouteViews.

To investigate the role of topology in the structure of the AS oscillations and to test the first hypothesis, we develop a simplified discrete-time simulation of BGP. We show that simulations of discrete singular switches between AS origins even on complex network topologies do not demonstrate any heavy-tailed behavior at distant viewpoints for AS holding times. In this same simulator, we oscillate the originating AS of a prefix between two ASes with heavy-tailed holding times, the parameters of which are derived from the heavy-tailed BGP session down times measured at the origin. Indeed, distant viewpoints do observe heavy-tailed origin oscillations. These simulations suggest the latter as a potential explanation for the oscillations (and eliminate the former).

We compare and contrast our analysis with update streams as seen from RouteViews of the BGP Beacon project [3]. The beacon data verify our simulations that

Dan Pei is with UCLA, E-mail: peidan@cs.ucla.edu. William Aiello, Anna Gilbert and Patrick McDaniel are with AT&T Labs. E-mail: {aiello,agilbert,pdmcdan}@research.att.com

discrete and singular events at the origin do not generate heavy-tailed oscillations at the viewpoint. On the other hand, these discrete events are on such small time scales that the data are, in some sense, incomparable with MOAS AS oscillation observations. If we do, however, adjust our time scales and define a much smaller time scale at which update disturbances come to quiescence, the behavior we observe in the MOAS update stream is consistent with the beacon data. It is clear that using this smaller time scale misses the larger scale of the AS origin oscillations and their associated long range dependence.

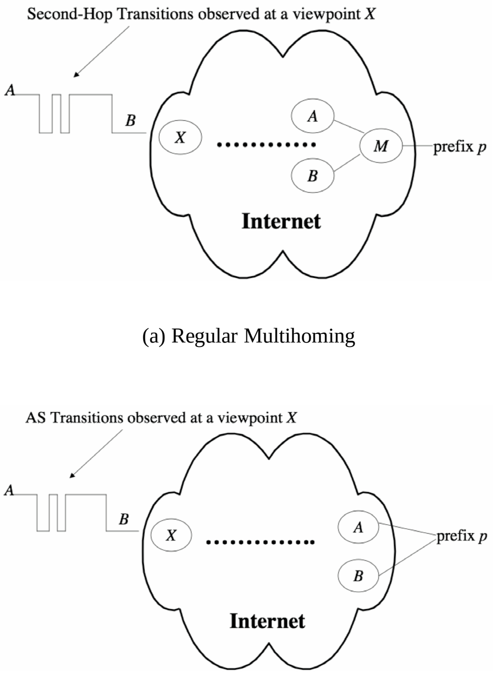
We also briefly study analogous behavior of the oscillations of the second hop AS for regular multihoming to ascertain whether our observations about MOAS multihoming apply to the much more prevalent regular multihoming. In sum, we find that AS oscillations driven by heavy-tailed oscillations between different multihomed providers are a widespread and important BGP phenomenon with complex but recognizable signatures such as heavy-tailed holding times and long-range dependence.

The rest of the paper is organized as follows. Section II defines *origin transition*, explains the potential causes of origin transitions, and describes our dataset. In Section III, we present our model of AS oscillations. To identify physical explanations of our model, we use simulations and empirical observations to drive input models for our simulations in Section IV. In Section V, we compare and contrast our results with those from the BGP Beacon project. We briefly discuss the applicability of our observations to the more widespread regular multihoming in Section VI. In Section VII, we summarize related work, and we conclude with Section VIII.

II. Origin Transitions and MOAS Multihoming

BGP [4] is the global routing protocol running between Autonomous Systems(ASes) in the Internet. A destination network in BGP is called a prefix p, and its *origin AS* is the AS who originates the announcement of a prefix p. Fig. 1(a) is an example of regular multihoming in which AS M is the origin for prefix p. In Fig. 1(b), both AS A and B are the origins for prefix p and we say that prefix p has multiple origin AS multihoming(or *MOAS multihoming*).

To understand more thoroughly MOAS behavior, we study how the origin AS changes at a "viewpoint." For a public BGP monitor (such as Oregon RouteViews[2] and RIPE [5]), a viewpoint is a router in a participating AS that has exterior BGP sessions with the monitor. We say that we have an *origin transition* for a particular prefix p if two consecutive updates for p from the same viewpoint have two different origins. Fig. 1(b) provides one example time series of the origin of prefix p at viewpoint X(The second-hop transitions in regular multihoming in Fig. 1(a) is analogous to the origin transitions in MOAS multihoming).


Second-Hop Transitions observed at a viewpoint X
A
B
X
A
M
B
prefix p
Internet

(a) Regular Multihoming

AS Transitions observed at a viewpoint X
A
B
X
A
B
prefix p
Internet

(b) MOAS Multihoming

Fig. 1. Regular Multihoming and MOAS Multihoming.

A. Causes of MOAS and Origin Transitions

Several recent works [1][6] identify the causes of MOAS behavior. These causes include illegitimate reasons such as misconfigurations and malicious attacks. According to [6] there are roughly 200 MOAS conflicts caused by misconfigurations daily. Legitimate reasons for MOAS behavior include aggregation (rare in practice according to [1]) and several varieties of MOAS multihoming.

MOAS multihoming covers the following cases: (i) a prefix can be multihomed to two or more ASes through some backdoor connections (e.g., prefix p in Fig. 1(b)); (ii) a prefix is multihomed to multiple ASes using a private AS number (which is stripped off by its provider ASes before announced to the global Internet); (iii) an exchange point prefix is shared by multiple ASes; or (iv) a prefix is temporarily multi-homed to both the old provider and a new provider when it switches providers.

Similarly, the origin transitions can happen for both legitimate and illegitimate reasons. Misconfigurations and attacks will have some origin transitions but the final origin AS will be the correct one; switching providers is an infrequent event for a fixed prefix and the origin AS of the prefix will stabilize with a new one. For MOAS multihoming prefixes, the origin transitions can be more frequent since they reflect BGP's automatic reaction to (more frequent) network topology changes.

B. Collection of Origin Transitions

We collected the AS transitions from all the viewpoints[1] of RouteViews Server 2 from August 2002 to July 2003(a one year study period). Note that we must clean some of the data. First, there are some reserved address blocks where are incorrectly announced by two or more origin ASes. Such prefixes are called bogon prefixes [7]. Second, some origin ASes in the BGP update message appears as AS SET, thus we are not able to clearly define an origin transition. These two AS inconsistencies contribute to about 5% of the total AS transitions and we remove them from consideration. The dataset composition is shown in Fig. 2.

We observe more than 6 million origin transitions in total. We find 16,160 prefixes (more than 10% of about 140,000 prefixes) and 5,144 ASes (more than one third of about 14,000 ASes) are involved in at least one origin transition. These results from the continuous BGP update stream are consistent with other origin stability studies using BGP table snapshots such that in as [8].

dataset	#transitions(percent)	# prefix(percent)
Complete	6,898,383 (100%)	16,474 (100%)
bogon	351(0.005%)	12 (0.07%)
AS SET	351,791 (5.1%)	302 (1.8%)
Cleaned	6,546,055 (94.9%)	16,160 (98.1%)
2-transition	3,514,696(50.9%)	10,211 (62.0%)
MOAS	1,630,064 (23.6%)	473 (2.87%)

Fig. 2. Data Set

C. MOAS Multihoming

Surprisingly, a small fraction of prefixes are responsible for a very large fraction of all transitions. We first rank prefixes according to the number of transitions they experience. Fig. 3 shows that the top 0.1% of prefixes are responsible for 18% of the total transitions, the top 1% of prefixes generate 45% of the total transitions, and the top 10% of prefixes have 81% of the total transitions.

Because a small number of prefixes generate a large number of the total AS transitions, it's worthwhile to ask how many unique transitions there are and how many each prefix has. To this end, we examine the unique transitions from *previous origin* to *current origin* for each prefix, independent of the view points. A prefix with two unique transitions means this prefix only has transitions from A to B and from B to A during the study period.

The solid curve in Fig. 4 shows the CCDF of the number of unique transitions per prefix. The dashed curve shows the CCDF of number of total transitions contributed by prefixes with a particular number of unique transitions. We see that about 88% of the prefixes have only one or two unique transitions, but they contribute more than 60% of the total transitions.

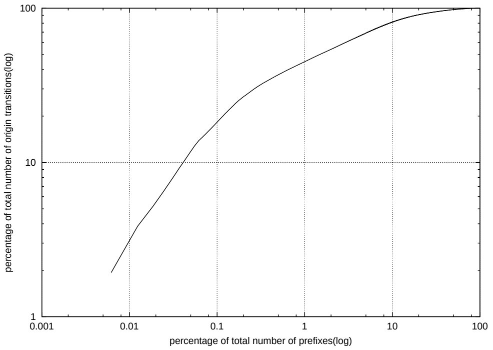


Fig. 3. Top x percent prefixes' contribution to the total origin transitions.

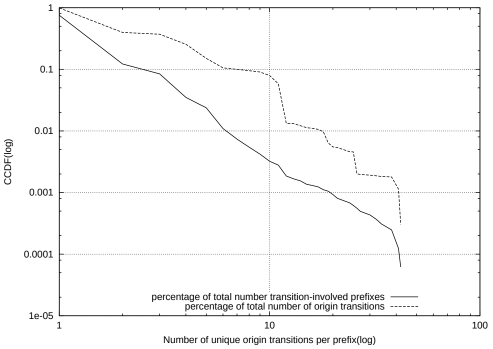


Fig. 4. Distribution according to the number of unique transitions per prefix.

Prefixes with two unique transitions give us a good starting point for studying origin transitions and MOAS behavior. The dataset is large–60.2% of prefixes that have two unique transitions contribute about 50.9% of all the AS transitions. Unfortunately, a prefix can experience two unique transitions for a variety of reasons, not necessarily because that prefix is a MOAS prefix. While we cannot verify that a prefix is MOAS multihomed directly, we can construct a simple heuristic to narrow the prefixes which we will study in the rest of the paper. Among those prefixes with two unique transitions, we consider those have more than 50 transitions observed at AT&T viewpoint as MOAS prefixes.

With this definition, we obtain a set of 473 prefixes and these prefixes contribute about 23% of total origin transitions. Furthermore, these 473 prefixes contributes a reasonably large fraction of BGP updates. We used the same dataset of AT&T internal BGP updates as [9] and found that these 473 prefixes (about 0.35% of total prefixes) contribute to about 0.8% percent of the total updates. Among the top 100 prefixes with the most updates, four are from these 473 prefixes. Therefore, even from the traditional BGP stability point of view, understanding these MOAS

[1] The number of viewpoints hovered around 30 during our period of study.

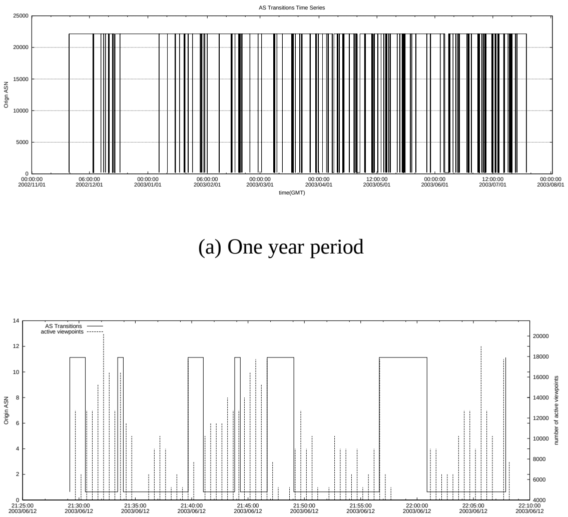
multihoming prefixes are important to understand BGP in general. In the rest of the paper, unless specified otherwise, we focus on the AT&T viewpoint,and the results we present are representative to this set of 473 prefixes, which we denote *MOAS multihoming dataset*.

III. Statistical Structure of AS Transitions

In this section, we present the statistical structure of the AS transitions and the analysis of the BGP updates amongst these AS transitions.

A. AS holding times

We observe that the AS transitions can be numerous and they can come in short bursts. In addition, there are long stretches of time with no AS transitions. Fig. 5 shows the origin time series of a representative prefix 198.69.224.0/22 (whose two origin ASes are AS 22143 and AS 209) for a one year period and for a two hour period. Observe that at both large and small time scales, we observe AS transitions.



(a) One year period

(b) Two hour period

Fig. 5. AS transition time series for prefix 198.69.224.0/22

To understand the different time scales of the origin transitions, we define the *AS holding time* of one origin A of prefix p at one viewpoint X as the time that p's origin stays with A at view point X. We say that a random variable Y is *heavy-tailed* or follows a heavy-tailed distribution if

$$Pr(Y > x) \sim x^{-\alpha}, \quad \text{as } x \to \infty, 0 < \alpha < 2.$$

Note that for α values less than 2, the random variable X has infinite variance and for $\alpha < 1$, infinite mean. We define two random variables; each one is the holding time over the entire dataset in one of the two origin ASes of a MOAS multihoming prefix. These two random variables' empirical complementary cumulative distribution functions (CCDFs) ($\bar{F}(x) = 1 - F(x) = Pr(Y > x)$ for random variable Y) on log-log axes are shown in Fig. 6. The particular MOAS prefix is 198.69.224.0/22. On the one hand, this CCDF appears to be linear on a log-log scale for several orders of magnitude; on the other hand, there is typically an abrupt transition where the probability density function appears to tail off much more quickly. That is, the CCDF appears to be an admixture of a heavy tailed distribution and an exponential distribution. We found such a CCDF is common is our MOAS multihoming dataset, and we use prefix $198.68.224.0/22$ in all of our analyses in the rest of this section.

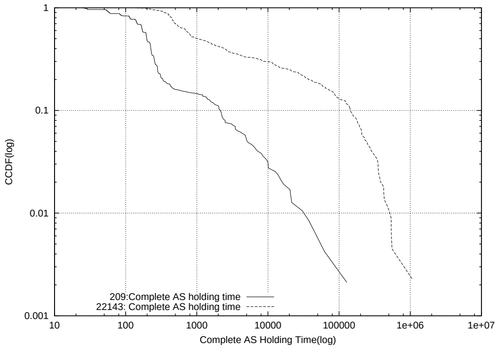


Fig. 6. AS holding time CCDF over the entire study period

B. AS disturbances

We posit the following model of AS disturbances. There are disturbance periods during which the origin AS oscillates with heavy tailed holding times. The "arrival" of these disturbances is given by a Poisson process. That is, the distribution of the times between the beginning of consecutive events is exponential. To verify our claim, we break up our AS oscillations into disturbances that are separated by periods of quiescence.

For a fixed viewpoint X and a fixed prefix p, we define *AS disturbance* as follows. The definitions are illustrated in Fig. 7.

- *AS disturbance* is a period of time in which consecutive origin transitions of p observed at X are separated by no more than Δ seconds.
- *AS quiescence* is a period of time of at least Δ seconds during which viewpoint X observes no AS transitions of prefix p.
- *Interdisturbance time* is the time between the beginnings of two consecutive AS disturbances of prefix p at viewpoint X.

We choose Δ according to the "knee" in the overall AS holding time CCDF in log-log between the power law and the exponential law. For example, we chose $\Delta = 25,000$ seconds for prefix 198.69.224.0/24 for AS 22143. We can

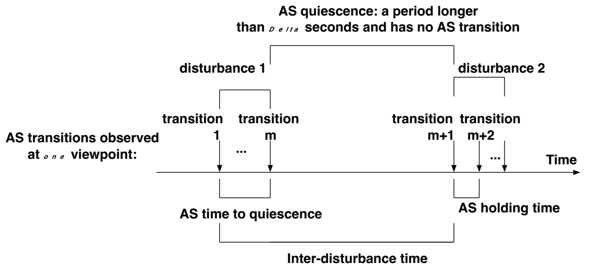


Fig. 7. An illustration of the definitions for AS disturbance.

see in Fig. 8 which is the CCDF of the interdisturbance times (plotted on semi-log axes), that these times are consistent with an exponential distribution. In addition, Fig. 9 is a plot of the CCDF of the AS holding times within the disturbances. This plot is consistent with a powerlaw CCDF (with a cutoff at $\tau = \Delta$). Thus, our data and our model are consistent.

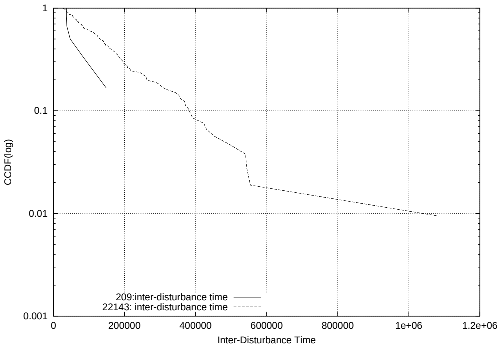


Fig. 8. Exponential interdisturbance time for prefix 198.69.224.0/22.

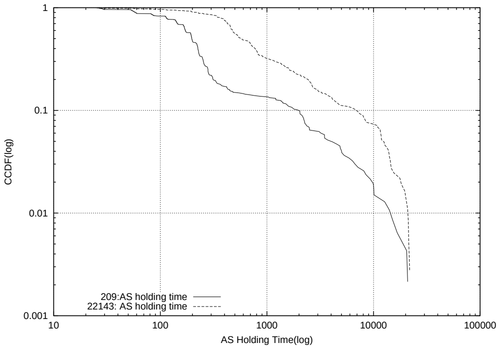


Fig. 9. Heavy-tailed AS holding time within the disturbances for prefix 198.69.224.0/22.

In addition, we find that the CCDF of the AS holding times is stable over long periods of time. We compared the CCDF of the holding times in first third transitions of any disturbance and that in the last third transitions of any disturbance, and found that they are very similar. Fig. 10 shows the CCDF of AS holding times in all transitions, first third transitions and last transitions within any disturbances of AS22143 for prefix 198.168.224.0/22. The stability of the holding times CCDF suggests that the tail parameters may be a useful signature of legitimate MOAS for such prefixes.

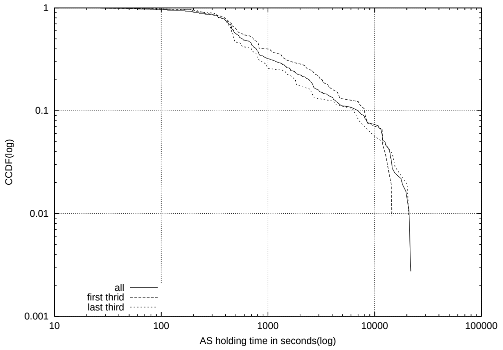


Fig. 10. CCDF of AS holding times of any transitions, first third transitions, and last third transitions of any disturbances.

C. Refined analysis of AS holding time

Fig. 9 are log-log complementary cumulative distribution function plots. They show the function $\bar{F}(x) = 1 - F(x) = Pr(Y > x)$ on log-log axes. In these plots, heavy-tailed distributions display linear behavior with slope equal to $-\alpha$. Unfortunately, the eye is easily deceived by CCDF plots; distributions which do not have truly heavy-tailed behavior "appear" to have linear log-log CCDFs. To bolster our intuition that the AS holding time within disturbances is heavy-tailed, we perform a second, more reliable test of heavy-tailed behavior. Fig. 11 shows the cumulative variance plots of the AS holding times for AS 209. We choose the holding times for this AS because they seem to be the most consistent with a heavy-tailed distribution. The cumulative or sample variance is one of the oldest tests for determining whether data has infinite variance or not. In this plot, we plot the sample variance S_n^2 from the first n observations as a function of n. If the data are drawn from a distribution with finite variance, the plots will converge to a finite value. If the data are from an infinite variance distribution, then the plots will diverge and show large jumps. This is exactly the behavior we see in the figure.

While the log-log CCDF plot of the AS holding time does show a linear relationship (and, hence, a heavy-tailed distribution), it is difficult to calculate the exponent α or the tail weight exactly from this plot. To do so, we use the standard Hill estimator. Let $Y_1, Y_2, \ldots, Y_n$ be a sequence of n observations drawn from a stationary *i.i.d.* process with probability distribution F, unknown. We assume that F is heavy-tailed with tail weight α. Let $Y_{(1)} \geq Y_{(2)} \geq \ldots \geq Y_{(n)}$ be the descending order statistics from our se-

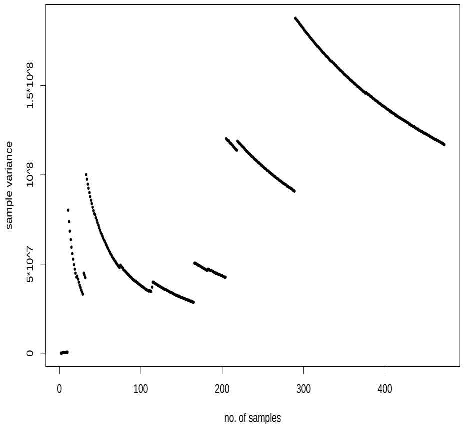


Fig. 11. Sample variance of AS holding time as a function of sample size for the prefix 198.69.224.0/22 and the AS 209.

quence of observations. The Hill estimator [10] is

$$\alpha(k) = \frac{1}{k-1}\sum_{i=1}^{k-1} \log Y_{(i)} - \log Y_{(k)} \quad \text{for } k > 1$$

and gives an estimate of α as a function of k. We plot the estimate as a function of k and if it stabilizes to a consistent value of α, then this value provides an estimate of the tail weight. Fig. 12 shows Hill estimator results for the holding times corresponding to AS 209. The estimator seems to give a consistent estimate of α about 0.8. Note that for α in this range, our probability distribution F has neither a finite variance nor a finite mean.

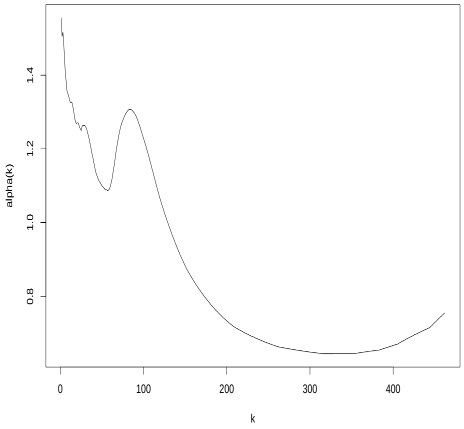


Fig. 12. Hill estimator of AS holding time as a function of sample size for the prefix 198.69.224.0/22 and the AS 209.

D. *Time to quiescence of AS disturbances*

The previous subsections describe the spacing between AS disturbances and the behavior of the AS holding times within disturbances but they do not address how long disturbances last. For that analysis, we turn to the *AS time to quiescence*. For a viewpoint X and prefix p, the *AS time to quiescence* of a disturbance is the time between the first and the last AS transitions of this disturbance observed at viewpoint X for prefix p. In other words, AS time to quiescence for a disturbance is simply the sum of the all the AS holding times within the disturbance. Fig. 13 shows that the CCDF of the AS time to quiescence on semi-log axes. Surprisingly, it appears to have an exponential distribution. In other words, the time to quiescence is the sum of K random variables H_i, where H_i's distribution is heavy-tailed and K's is exponential.

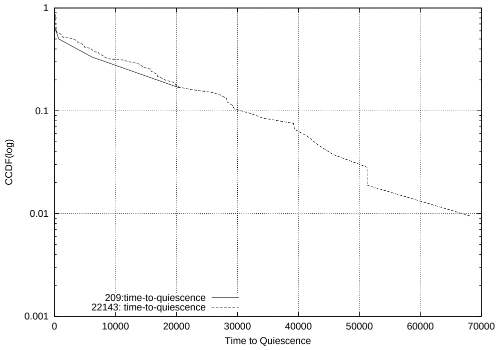


Fig. 13. AS time to quiescence for prefix 198.69.224.0/22.

We then plot the distribution of the number of transitions within a disturbance in Fig. 14 to see whether it can give us some hints. On semi-log axes, this distribution appears to be exponential and the average disturbance has only a small number of transitions and most have fewer than 20 (*i.e.*, K is small). Our conjecture is that when K is large enough, we have a reasonable chance of drawing a large H_i and that large value swamps the sum. If K is small, then with reasonable probability most of the H_i's are small. Thus the distribution of the sum of H_i can look like exponential. This conjecture seems to be consistent with what we see in Fig. 13 and Fig. 14, but a more careful verification of this conjecture is our future work.

E. *Long range dependence within disturbances*

All of the analysis in the previous subsections concentrated on the AS transitions themselves rather than on the stream of updates that imply the AS transitions. In this subsection, we focus on the statistical behavior of these updates. We create a time series of the number of updates for the given prefix received from all viewpoints of RouteViews in 30 second bins. This time series is divided up into disturbance periods as implied by our definition of AS disturbance. That is, the time boundaries are the same as the

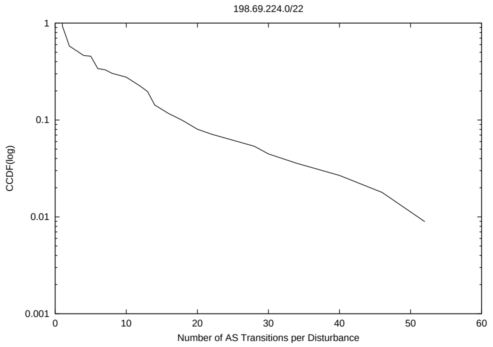


Fig. 14. Number of transitions per AS disturbance for prefix 198.69.224.0/22.

AS disturbance periods for that prefix. We show that the update counts within disturbances exhibit long range dependence (LRD) over seven or eight time scales (roughly, an hour).

Let $Y = \{Y(i), i \geq 1\}$ be a stationary sequence. Let

$$Y^{(m)}(k) = \frac{1}{m} \sum_{i=(k-1)m+1}^{km} Y(i), \quad k = 1, 2, \ldots,$$

be the sequence corresponding to averages of Y over blocks of size m. If

$$Y \overset{d}{=} m^{1-H} Y^{(m)}$$

where the equality is in the sense of finite dimensional distributions, then we say that Y is self-similar with Hurst parameter H. If equality holds over large values of m (over large scales), we say that Y is asymptotically self-similar or long range dependent.

Wavelets with their built-in scale-localization ability provide an ideal mathematical tool for investigating the scaling behavior of self-similar processes across all (or a wide range of) time scales. (See [11] for more information about wavelets.) Abry and Veitch [12] show that if Y is a self-similar process with Hurst parameter $H \in (0.5, 1)$, then the expectation of the average energy E_j that lies within a given bandwidth 2^{-j} around frequency $2^{-j}\lambda_0$ is given by

$$\begin{aligned} \mathbf{E}[E_j] &= \mathbf{E}\left[\frac{1}{N_j} \sum_k |d_{j,k}|^2\right] \qquad (1) \\ &= c|2^{-j}\lambda_0|^{1-2H} \int |\lambda|^{1-2H} |\hat{\psi}(\lambda)|^2 \, d\lambda, \end{aligned}$$

where $\hat{\psi}(\lambda)$ is the Fourier transform of the wavelet $\psi(t)$, λ_0 is a baseline frequency parameter that depends on the wavelet ψ, and N_j is the number of wavelet coefficients at each scale j. By plotting $\log_2 E_j$ against scale j (where $j = 1$ is the finest scale and $j = N > 1$ is the coarsest) and identifying scaling regions, breakpoints and non-scaling behavior, we have an unbiased scaling analysis of a given signal Y that is simple, computationally efficient and informative. For example, the scaling analysis of a signal which is exactly self-similar will yield a linear plot of $\log_2 E_j$ vs. j for all scales. On the other hand, for an asymptotically self-similar signal a linear relationship between $\log_2 E_j$ and scale j will be apparent only for large times or scales. We can also estimate the Hurst parameter from the scaling analysis. Experience with (asymptotically) self-similar time series and those arising in data networks [13] suggests that "dips" or bends in the the scaling analysis plots indicate periodic behavior or components in the time series.

In Fig. 15, we perform a scaling analysis on one long disturbance from the prefix 198.69.224.0/22. We use the Haar wavelet for this analysis. We can see from the figure that over a large number of time scales, the average energy is a linear function of scale so this disturbance is consistent with long range dependence. We note that the Hurst parameter estimate for this disturbance is 0.8, which is consistent with long range dependence. A complete qualitative analysis of the scaling analysis is beyond the scope of this paper and is a direction for future research.

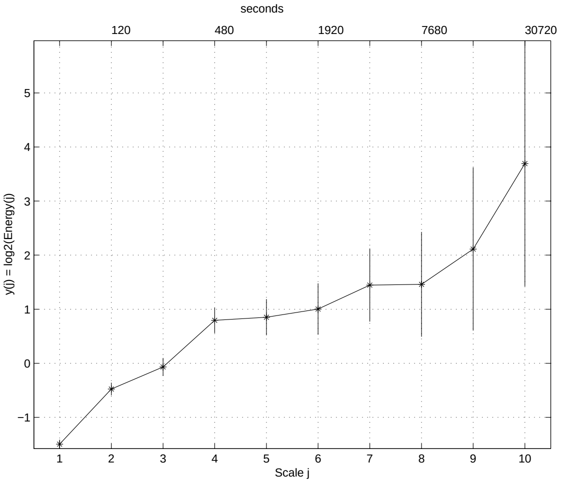


Fig. 15. Energy plot analysis for an individual disturbance for prefix 198.69.224.0/22.

For most of the disturbances from this particular prefix, our Hurst parameter estimator generated values which are consistent with long range dependence ($0.5 < H < 1$); however, the qualitative analysis of their energy plots was less convincing. Many of these energy plot analyses did not show clear linear behavior over many time scales. A closer look at the time series shows many very long runs of no updates. That is, there are small bursts of updates interspersed amongst long runs of zero updates. These disturbances show a delineation of time scales: bursts of updates on small time scales and correlated updates over long time scales.

IV. Causes of Heavy-tailed Oscillations

In the previous section we demonstrated that AS disturbances are an exponential arrival process and AS holding times within disturbances are heavy-tailed. We postulate two potential causes of the heavy-tailed AS holding times within disturbances at the RouteViews viewpoint. In the first setting, there is a single discrete cut-over from the AS of one MOAS multihoming provider to the other that is amplified to produce large time scale oscillations. These transient events might interact with the powerlaw AS graph to introduce additional propagation delay, and ultimately the observed heavy tailed AS disturbances. A second setting postulates that heavy-tailed oscillations between the two MOAS multihoming ASes drive the oscillations at RouteViews; e.g., flapping between origins. That is, the distribution at the viewpoint is the result of some recurrent condition(s) at the origin. This section uses both simulation and investigation of real origin events to rule out the first hypothesis and to verify the second.

A. Simulating Disturbances

Our first hypothesis is that heavy-tailed AS disturbances are the result of UPDATE propagation oddities occurring within the powerlaw topology [14] of the Internet AS graph. As shown in Appendix A, a single origin transition can be amplified by network conditions and BGP policy to result in multiple transitions at remote viewpoints. We posit that these events could resonate to create long disturbances when occurring in bursts and propagating across structured AS graphs.

We attempt to replicate resonance events by simulation. The *agpmsim* discrete-time-step simulator built for this study implements various models of BGP. The most complex model, the *the Mini-path-vector model*, integrates elements of asynchronous update propagation and path selection policy over arbitrary AS topologies. All experiments in this section use the mini-path-vector model as detailed in Fig. 16.

There are two central parameters of the simulation that affect the replication of origin disturbances on long time scales. The propagation rate R is the probabilistic rate at which update information is forwarded to its neighbors. Propagation delays frequently occur in practice due to BGP timer asynchrony, network congestion, router load, or other factors. Any simulation lacking delay would fix quiescence as distance from an origin, and hence its inclusion is essential to a complete model. The second parameter S is the probabilistic stability of the origin. That is, S defines the AS transition distribution at the origin. While R affects oscillation and the length of fine-grained (update) disturbances, S will dictate the length and inter-arrival times of origin disturbances.

The Mini-path-vector Model

A network is an undirected graph G with vertex set V and edge set E. The neighborhood $N(v)$ of a node $v \in V$ is the set of nodes that share an edge with v. We define time is in discrete steps $\{t_0, t_1, \ldots\}$.

A path P is defined as a totally ordered subset of V. Given a node v and a path P, we assign a policy value $T(P, v)$ to the pair. $T(\emptyset, \cdot)$ is undefined. We fix a hash function H and assign $T(P, v)$ the value of H applied to P and v. This ensures that the policy (i.e., the hash function) remains fixed over time, but its value is different for each vertex and, possibly, for each time step. At time t_i, the state of a node consists of a local best path $P(v, t_i)$ and a path state for each neighbor $u \in N(v)$, denoted $Q(u, v, t_i)$.

An experiment is defined as follows. Initially, all state variables are set to null (e.g., $P(v, t_0), Q(u, v, t_0) = \emptyset$ for all $v \in V$ and for all $u \in N(v)$). A primary node r_1 and a backup node r_2 are selected at random. At time t_0, r_1's state is set to contain a single node, itself (e.g., $P(r_1, t_0) = r_1$). At time $t_i, i > 0$ each node v copies its neighbors' local best path into the neighbor state with constant probability R. That is,

$$Q(u, v, t_i) = \begin{cases} Q(u, v, t_{i-1}) & \text{with probability } R, \\ P(v, t_{i-1}) & \text{with probability } 1 - R. \end{cases}$$

At time t_i, after updating all its neighbor state, each node computes $P(v, t_i)$ as follows:

1. if $\forall u \in N(v), Q(u, v, t_i) = \emptyset$, then $P(v, t_i) = \emptyset$, (if all neighbor state empty, no shortest path)

2. else,

(a) compute set M of shortest non-empty paths from neighbor state, $M = \mathrm{shortest}\{Q(u, v, t_i) | u \in N(v)\}$, (use the shortest path, if unique)

(b) select $P(v, t_i) = Q(u^*, v, t_i)$ where u^* is the neighbor with the largest policy value $T(Q(u^*, v, t_i)$ over all paths in M, (break ties with policy).

The simulator models oscillation by withdrawing r_1 (e.g., $P(r_1, t_{i+1}) = \emptyset$), and announcing r_2 as the origin (e.g., $P(r_2, t_{i+1}) = r_2$), or vice-versa, according to a random variable S. A node is quiescent when it converges on a shortest path to the originating node (and it performs no subsequent transitions).

Fig. 16. The mini-path-vector model as implemented in the *agpmsim* simulator. This models policy and asynchronous update propagation.

An initial suite of tests establishes that origin oscillation (as described in Appendix A) at viewpoints can be replicated in simulation on very simple graphs. For example, Fig. 17 shows the AS transitions seen by a single viewpoint in a ten node network in response to a single AS transition.

A second set of experiments seeks to replicate heavy-tailed AS disturbances at the viewpoints. We perform

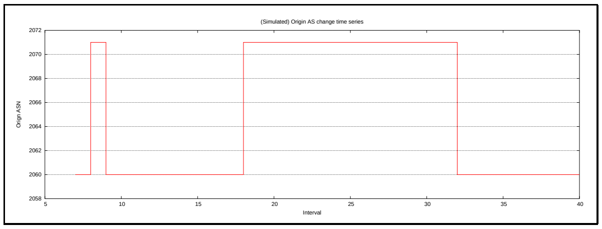


Fig. 17. Origin time series (simulated): origin oscillation seen by a viewpoint on ten node network in response to a single AS transition.

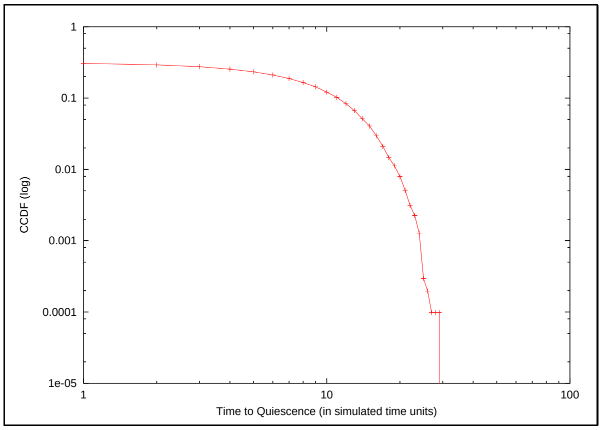


Fig. 18. CCDF of simulated AS time to quiescence.

these tests on a 3,038 node (powerlaw) graph generated by INET [15]. We chose an $R = .5$ and modeled S as a binary random variable, where the probability of an AS transition between the two origins (per time step) was fixed at 10% (and a 90% chance of no AS transition). The test oscillates between ASes according to S for 100 time-steps and measures the time to quiescence for each of 10 viewpoints following the oscillation period.

We measure the AS time to quiescence by recording the last transition seen by a viewpoint following the 100 interval oscillation (a transition was always performed at time step 100). Fig. 18 shows a CCDF of the AS time to quiescence for one viewpoint over 10,000 repetitions of the test. Note that there is no significant behavior at large time scales; the range of time to quiescence can readily be explained by the probabilistic delay of update propagation. The distribution is exponential and with high probability the system achieves quiescence in a short number of steps. We performed a raft of experiments on different network graphs (Erdös-Renyi random graphs, powerlaw random graphs, and hypercubes) with a variety of R and S parameters. Thee results are similar and no experiment yielded heavy-tailed oscillations at a viewpoint.

The absence of large time scales indicates that while propagation may increase AS time to quiescence at small times scales, it does not directly result in long disturbances or oscillation(such as those shown in Fig. 9). Hence, we conclude that the powerlaw topology, the propagation delay (and its potential interaction with topology), and the use of policy are not a key factors in generating heavy-tailed behavior. This led us to investigate another hypothesis: something at the origin could be driving the observed behavior.

B. *Physical modeling of events at Origin*

Our second hypothesis for the heavy-tailed AS holding times is that there are recurrent events or conditions at the origin that drive the heavy-tailed distribution we see at the viewpoint. Note that in theory the origin oscillations may be the result of BGP session bounces at the origin ASes or elsewhere in along the AS paths. But in practice the origin is most vulnerable to a single router's BGP session bounces; other ASes along AS paths are much resilient to single router's session bounces since they usually have multiple peering sessions between them. In addition we were able to confirm the former for some of our prefixes for which we had sufficient visibility into logs for BGP sessions at one or more of the origin ASes. We leave it as our future work to look into more details the impact of failures at intermediate ASes.

Rather than postulating origin conditions with no supporting empirical evidence, we turn to measurements to drive our modeling. Thus, we practice physical-based modeling. We begin by identifying a set of AS environments we can reasonably profile. We identify 11 of the 437 MOAS prefixes which are originated by AT&T(AS 7018) and other ASes. Among these 11 prefixes, 203.76.160.0/20 was originated by both AS 7018 and AS 4858. AS 4858 is adjacent to AS 7018 and shares a single peering link with AT&T. We extracted 79 days (May 16, 2003 to July 31, 2003) of timestamped system logs from the routers supporting peering sessions associated with the prefixes.[2] We extract BGP specific events from the system logs and reconstruct the session state over the 79 day period.

A CCDF of the BGP session down times associated with the router over which 203.76.160.0/20 was indirectly originated is shown in Fig. 19. Even with only 162 observations, the down times appear to be heavy-tailed. To verify this claim, we show in Fig. 20 the sample variance and Hill estimator plots for this time series. Because the extended length of some of the session down times is surprising, we spot-check several of the session outages and correlate them with network events. Almost all of the dozen or so studied outages could be directly correlated with layer 2 outages; e.g., ATM failures. Interestingly, the session uptimes for this router were not similarly heavy tailed.

Note that down times of several days or weeks may be quite natural, either as the result of some catastrophic event (e.g., backhoe through fiber) or as part of a planned outage

[2] 79 days was the longest uninterrupted feed of system logs during the year long experiment, and hence was used to reconstruct session information.

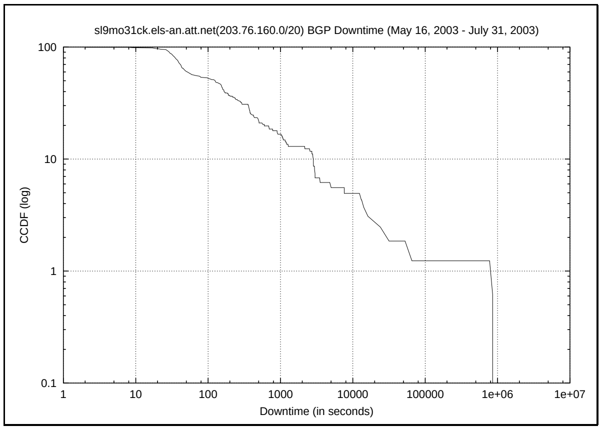


Fig. 19. CCDF of single router down times as measured over 79 day period in 2003.

(e.g., network upgrade). A session outage does not indicate that the network is necessarily partitioned from the larger Internet. The link will fail-over to an backup link if one is available (as is the case in most multi-homing situations).

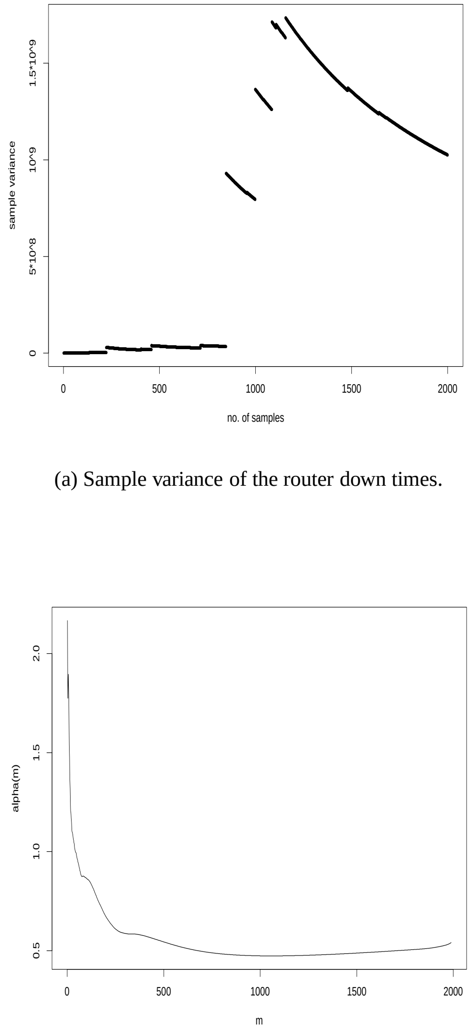
We conclude that the empirical evidence is consistent with heavy-tailed BGP session times (especially down times) at the origin and that we are justified in using a heavy-tailed distribution in our simulation to generate oscillations at the origin. Our goal is to determine whether these heavy-tailed oscillations give rise to heavy-tailed holding times at the viewpoint.

A new suite of tests models S as a Pareto distribution[3] with α and β parameters similar to those exhibited by the session down-times. The test repeated 162 oscillations whose inter-transition time was randomly selected from a Pareto distribution S with $\alpha = 0.85$ and $\beta = 60$.[4]

A CCDF of the AS holding times in response as seen by a single viewpoint in the simulated environment is shown in Fig. 21. Clearly, the Pareto generation of oscillation leads to heavy tailed holding times. Hence, we conclude that it is highly likely that the heavy tails in the AS holding times are resulting from highly variable session downtimes. In addition, we can conclude from our first suite of tests that events at the origin which are short-lived (such as single discrete cut-overs from one AS to another) are not sufficient, not even in conjunction with propagation delay on a powerlaw graph, to yield heavy-tailed AS holding times.

V. Comparison with BGP Beacon

Section IV shows that origin oscillations drive the viewpoint oscillations and that heavy-tailed oscillations at the origin give rise to heavy-tailed oscillations at the viewpoint (at least in simulation). In this section, we compare and contrast our analysis with the update streams of prefix 192.83.230.0/24 from the BGP Beacon project [3] as seen from RouteViews.



(a) Sample variance of the router down times.

(b) Hill estimator of the router down times.

Fig. 20. Sample variance and Hill estimator for BGP session down times.

BGP Beacon prefixes [3] are prefixes established to be announced or withdrawn according to publicly known schedules and their updates are readily available at monitoring point such as Oregon RouteViews or RIPE. Among all the beacon prefixes, 192.83.230.0/24 is a *regular* multihoming prefix (Fig. 1(b) has illustrated the regular multihoming) and its second-hop AS transitions (analogous to origin transitions in MOAS multihoming) is precisely driven by a *discrete* second-hop failure/recovery every two

[3] The distribution function F of a general Pareto distribution is given by $F(x) = 1 - (\beta/x)^{\alpha}$ for $x \geq \beta$.

[4] The α and β values were selected from estimates of similar prefixes within our dataset. Note that the precise values of the distribution are not important. The regeneration of a similar distribution for times to quiescence is sufficient to support our thesis.

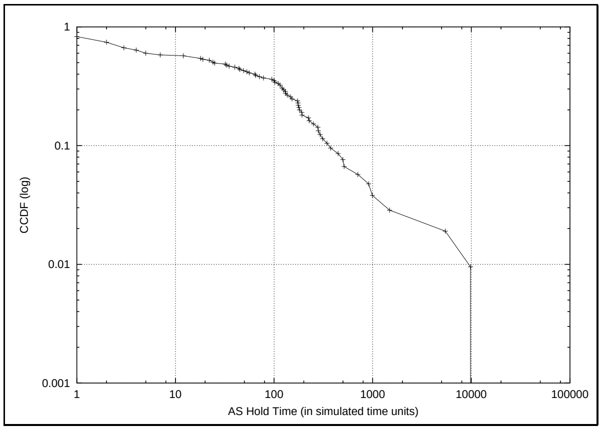


Fig. 21. CCDF of AS hold times resulting from simulated Pareto-distributed AS transitions.

hours *scheduled* according to the setup shown in [16]. Thus this prefix's second-hop transition is a natural comparison point with the origin transition in MOAS multihoming. The dataset is from June 2003 (when it first became available) to April 2004.

A. Basic Observation

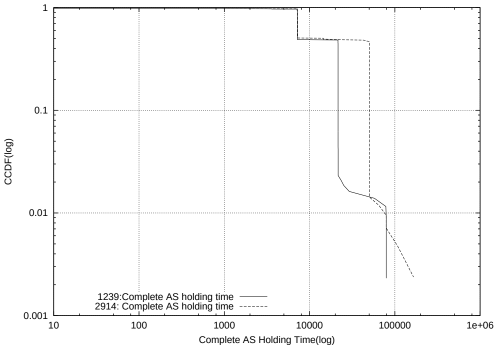


Fig. 22. Second-hop AS Holding time for beacon 192.83.230.0/24

For the beacon data, Fig. 22 shows that the second-hop AS holding times are indeed quite discrete at the viewpoints. Not only do they show distinct jumps in time but they also are not heavy-tailed. We also found that in almost all the cases a viewpoint has only one second-hop transition around the scheduled time, and in only very few cases, a viewpoint can see two or more second-hop transitions around the scheduled time. In other words, we see little amplification of individual failure/recovery. Therefore, it seems clear that the discrete disturbance at the second-hop that beacon creates on a regular basis cannot cause heavy-tailed AS transitions/second hop transitions at the viewpoint. We thus conclude that the BGP beacon data verify our simulations that discrete and singular origin events do not give rise to heavy-tailed oscillations at the viewpoint.

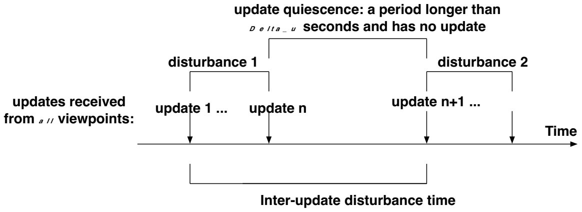

update quiescence: a period longer than
D e l t a _ u seconds and has no update
disturbance 1
disturbance 2
updates received
from a l l viewpoints:
update 1 ...
update n
update n+1 ...
Time
Inter-update disturbance time

Fig. 23. An illustration of the definitions for update disturbance.

B. Comparison using Update disturbances

For a direct comparison of the beacon data and the MOAS dataset, we should break up the beacon data into AS disturbances (as defined in Section III); however, the beacon data at any large scale (two hours or more) are discrete and the range of time scales in the beacon data is simply not large enough to afford such a disturbance definition(e.g., with $\Delta = 25,000$). That is, the behavior in the beacon data and that in the MOAS data are incomparable in the AS disturbance level.

There is a more complex analysis we can perform, however, if we adjust our definition of disturbance accordingly. For this analysis we break origin (or second-hop) oscillations up according to an *update disturbance*, a different definition of disturbance than the AS disturbance we use earlier. The updates received by multiple viewpoints are triggered by the same failure/recovery at the second-hop in beacon. Previous measurements in [17] [3] show that the majority of Internet route convergence takes less than 180 seconds. Thus it should be a clear indication that the network has converged to the new sets of the stable path after the triggering failure/recovery if there are no updates from any viewpoints for a threshold of time (e.g. 120 seconds). We thus define the *update disturbance* for a fixed set of viewpoints Z at RouteViews and a fixed prefix p, as follows. The definitions are illustrated in Fig. 23.

- *Update disturbance* is a period of time in which viewpoint set Z receives consecutive updates for prefix p that are separated by no more than Δ_u seconds.
- *Update quiescence* is a period of time of at least Δ_u seconds during which viewpoint set Z receives no updates from any viewpoint for prefix p.
- For a fixed viewpoint set Z, the *time to quiescence* is the time between the first and the last updates this disturbance observed at the viewpoint Z for prefix p.

For the results in the rest of the paper, we choose a threshold of $\Delta_u = 120$ seconds. We observe that our results are robust to changes of Δ_u as they did not change significantly with different values of $\Delta_u = 45, 60, 90, 120, 150, 180$.

Measuring beacon's second-hop AS holding time within each update disturbance generates no data points (thus not shown), confirming our early finding that there is at most one second-hop transition at one viewpoint per failure/recovery. Similarly, for the MOAS prefix there are too few data points for the AS holding time within the update disturbances to show a clear trend in Fig. 24. This is because most of the AS holding times in our data are significantly longer than the time required for updates to come to quiescence after a single AS transition at the origin. It is clear from this that using the time scale at which updates come to quiescence misses the larger scale of the AS origin oscillations and their associated long range dependence. The above observation, plus the difference between Fig. 22 and Fig. 6, indicate that the update disturbances in the beacon data are distinctly different from the AS disturbances in the MOAS data with respect to holding times.

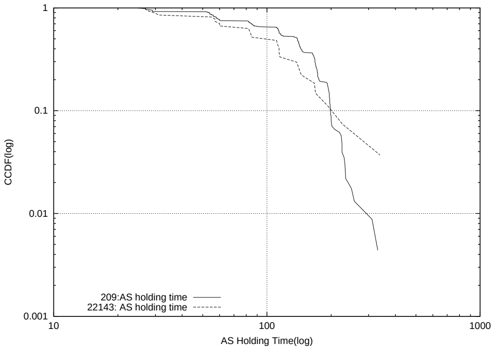


Fig. 24. Not enough data points for AS holding time within update disturbances for MOAS prefix.

On the other hand, we observe that the time to quiescence of update disturbances for both the beacon data and the MOAS data are similar. Both CCDFs show an exponential tail with most disturbances dying out within two to three hundred seconds.[5] MOAS prefix's time to update quiescence is slightly longer in the tail, and can be caused by two or more consecutive transitions at the origin, which is consistent with Fig. 24 which has some (although very few) data points for holding time within update disturbances.

We conclude that, from one perspective, the BGP beacon data verify our simulations that discrete and singular origin events do not give rise to heavy-tailed oscillations at the viewpoint. From another perspective, these discrete events are on such small time scales that they are incomparable with the long time scales we observe in the MOAS oscillations. If we do, however, adjust our time scales and define an update disturbance (that is naturally, a small time event), the behavior we observe in the update stream of the MOAS data is consistent with that we see in the beacon data. We point out that this smaller time scale misses the larger scale of the AS origin oscillations and their associated long range dependence.

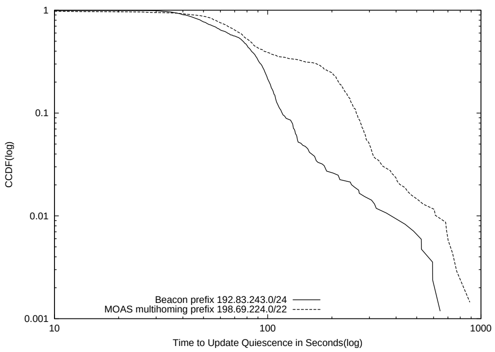


Fig. 25. A comparison of the time to quiescence for update disturbances in the beacon and MOAS datasets.

VI. Analogous behavior in Regular Multihoming

In the previous sections, we describe several important phenomena of MOAS multihoming and provide solid experimental and empirical evidence of models for this behavior. MOAS multihoming is, however, a special type of multihoming while regular multihoming is much more prevalent in the Internet. As a first step to determine how universal our observations on MOAS multihoming are and how applicable they are to the much wider practice of regular multihoming, we take a small (random) set of 30 prefixes whose origin ASes are multihomed to both AS 701 and AS 1239, two tier-1 ASes in the Internet. In the figures below, we give results only for the prefix 129.121.0.0/16; however, the results are similar for the other 29 prefixes. We use the definitions of AS disturbance from Section III and use the notion of second-hop AS transition as in the previous section, and $\Delta = 10,000$ seconds.

We see in Fig. 26 that the second-hop AS holding times have a similar distribution as that in Fig. 6; the CCDF appears to be the admixture of a heavy-tailed distribution and and an exponential distribution. Fig. 27 shows that, similar to Fig. 8, the interdisturbance time has approximately an exponential distribution. The second-hop AS holding times within disturbances shown in Fig. 28 has approximately a heavy-tailed distribution(similar to Fig. 9), and the time to second-hop AS quiescence in Fig. 29 has approximately an exponential distribution(similar to Fig. 13).

Given above similarities between regular multihoming and MOAS multihoming, and that the behavior of MOAS multihoming at the viewpoints is driven by the heavy-tailed disturbances at the origin, it's likely that for regular multihoming the structure is also driven by the link failures/recoveries between the origin AS and its providers. We leave a detailed study of second-hop transitions in regular multihoming our future work. In sum, our findings suggest that AS(or second-hop AS) oscillations driven by heavy-tailed oscillations between different multihomed providers are a widespread and important BGP phe-

[5] Our definition of time to update quiescence does not reflect the false damping effect [18] as our Δ_u is much shorter than the minimal damping time(20 minutes).

nomenon with complex but recognizable signatures such as heavy-tailed holding times and long-range dependence.

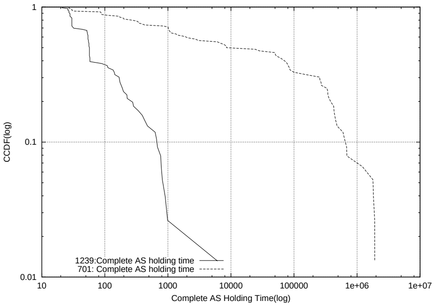


Fig. 26. Second-hop AS holding time of regular multihoming prefix 129.121.0.0/16 in the entire study period.

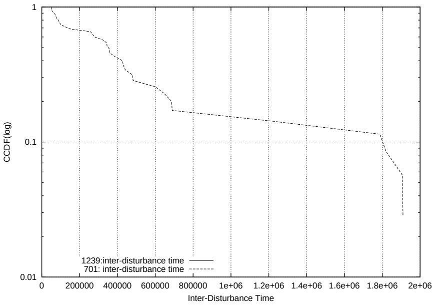


Fig. 27. Interdisturbance time of regular multihoming of 129.121.0.0/16.

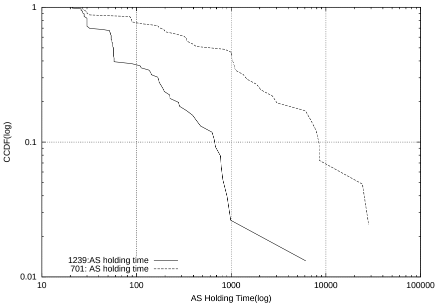


Fig. 28. Second-hop AS holding time within disturbances of regular multihoming prefix 129.121.0.0/16.

VII. Related Work

In this section, we briefly review related work. Zhao, *et al.* measure the MOAS conflicts using BGP table snapshots and provide possible causes of the MOAS conflicts in [1]. Previous work on the stability of origins of prefixes includes [19] and [8]. Kent, *et al.* [19] claim origins are very stable as measured by the number of new announcements they encountered. Aiello, *et al.* [8] also study the stability of origins using BGP table snapshots, and found 70-90% stability over a five month period and monthly churn of around 5% (of 143,215 prefixes).

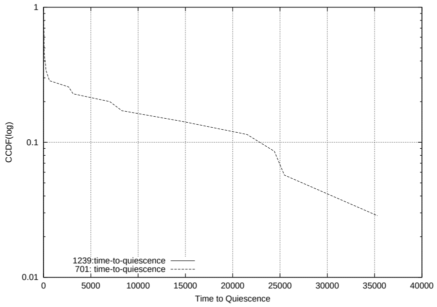


Fig. 29. Second-hop AS time to quiescence of regular multihoming prefix 129.121.0.0/16.

The AS transition dynamics are closely related to the BGP update dynamics as they are both triggered by underlying topology changes (or policy changes). Thus, we briefly review recent efforts on BGP update dynamics measurement and modeling. Labovitz, et al. [20] measure the update interarrival time of the tuple (viewpoint, prefix). Labovitz, *et al.* [21] also show that Internet backbone paths exhibit a mean time to fail-over (due to either physical failure or policy changes) of roughly two days and only roughly 20% of paths unchanged in five days. Rexford, *et al.* [9] show that a large portion of the interarrival time is around 30 seconds (the default BGP update rate limiting timer value). This observation is confirmed by the measurement of the controlled failures of beacon data by Mao, *et al.* [3]. Based on the interarrival time distribution, Rexford, *et al.* [9] use a time window to divide sequence of updates into sequence of “events”. The authors measure interarrival time of events and observe that most events are short-lived. They also show that a small percent of the prefixes contribute to majority of the events and conjecture that one possible explanation is flapping devices.

Mao, *et al.* [3] provides interarrival modeling of the beacon prefixes (prefixes with controlled failure/recovery). They model the observed interarrival CCDFs using a combination of mass distribution, geometric distribution, and exponential distribution. They conjecture that the exponential tail is caused by BGP false damping (as demonstrated by [18]); a legitimate sequence of updates can trigger the BGP damping mechanism to stop announcing a prefix and re-announce this prefix up to one hour later.

Labovitz, *et al.* [17] discover in BGP it can take as long as 15 minutes for the network to converge to the new set of paths after a topology change but majority of time-to-quiescence is below 180 seconds.

VIII. Conclusion

The central contribution of this paper is a detailed analysis of the MOAS multihoming oscillations at different time scales. We empirically derive a model of AS disturbance periods during which the AS oscillates with heavy tailed holding times. We demonstrate that these disturbances arrive according to a Poisson process. We also show that the update stream within these disturbances exhibits long range dependence. Using simulations and physical-based modeling to drive these simulations, we demonstrate that heavy-tailed oscillation at the origin is a possible explanation for our observations (while the complex interplay of the BGP protocol and network topology is not such an explanation). That is, “unusual” or heavy-tailed “operational events” at the origin might be the root cause of our observations rather than the intricacies of BGP. Research to “fix” this behavior of BGP might be misguided.

Our work suggests that the tail parameters of AS holding times may be a useful signature of legitimate MOAS for prefixes. We also identify interesting scaling behavior in BGP update streams–scaling behavior that might otherwise be missed with datasets such as the Beacon data. A thorough investigation of this scaling behavior is just one possible direction for future work.

References

[1] X. Zhao, D. Pei, L. Wang, D. Massey, A. Mankin, S. Wu, and L. Zhang, “An Analysis BGP Multiple Origin AS(MOAS) Conflicts,” in *Proceedings of the ACM IMW2001*, Oct 2001.

[2] “The Route Views Project,” http://www.antc.uoregon.edu/route-views/.

[3] Z. Morley Mao, Randy Bush, Tim G. Griffin, and Matthew Roughan, “BGP Beacon,” in *Proceedings of ACM IMC 2003*, October 2003.

[4] Y. Rekhter and T. Li, “Border Gateway Protocol 4,” RFC 1771, SRI Network Information Center, July 1995.

[5] “The RIPE Routing Iformation Services,” http://www.ris.ripe.net.

[6] R. Mahajan, D. Wetherall, and T. Anderson, “Understanding bgp misconfiguration,” in *Proceedings of ACM Sigcomm*, August 2002.

[7] “The Team Cymru Bogon Reference Page,” http://www.cymru.com/Bogons/index.html.

[8] W. Aiello, J. Ioannidis, and P. McDaniel, “Origin Authentication in Interdomain Routing,” in *Proceedings of 10th ACM Conference on Computer and Communications Security*. ACM, October 2003, pp. 165–178, Washington, DC.

[9] Jennifer Rexford, Jia Wang, Zhen Xiao, and Yin Zhang, “BGPRouting Stability of Popular Destinations,” in *Proceedings of ACM IMW 2002*, October 2002.

[10] B. M. Hill, “A simple general approach to inference about the tail of a distribution,” *Annals of Statistics*, vol. 3, no. 5, pp. 1163–1173, 1975.

[11] S. Mallat, *A Wavelet tour of signal processing*, Academic Press, 1998.

[12] P. Abry and D. Veitch, “Wavelet analysis of long-range dependent traffic,” *IEEE Transactions on Information Theory*, vol. 44, pp. 2–15, 1998.

[13] A. Feldmann, A. C. Gilbert, P. Huang, and W. Willinger, “Dynamics of IP traffic: A study of the role of variability and the impact of control,” 1999, ACM SIGCOMM Conf. Proc.

[14] Michalis Faloutsos, Petros Faloutsos, and Christos Faloutsos, “On power-law relationships of the internet topology,” in *Proc. of ACM SIGCOMM ‘99*, 1999, pp. 251–262.

[15] Jared Winick and Sugih Jamin, “Inet-3.0: Internet topology generator,” Tech. Rep. CSE-TR-456-02, University of Michigan, 2002.

[16] “Schedule of BGP Beacon 5 at PSG,” http://www.psg.com/ zmao/beacon5.jpg.

[17] C. Labovitz, A. Ahuja, A. Bose, and F. Jahanian, “Delayed Internet Routing Convergence,” in *Proceedings of ACM Sigcomm*, August 2000.

[18] Z. Mao, R. Govindan, G. Varghese, and R. Katz, “Route Flap Damping Exacerbates Internet Routing Convergence,” in *Proceedings of ACM Sigcomm*, August 2002.

[19] S. Kent, C. Lynn, J. Mikkelson, and K. Seo, “Secure Border Gateway Protocol (S-BGP) — Real World Performance and Deployment Issues,” in *Proceedings of Network and Distributed Systems Security 2000*. Internet Society, February 2000.

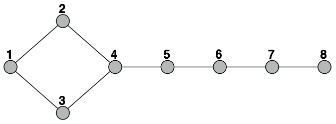
[20] C. Labovitz, G. Malan, and F. Jahanian, “Internet Routing Instability,” in *Proceedings of ACM Sigcomm*, September 1997.

[21] C. Labovitz, A. Ahuja, and F. Jahanian, “Experimental Study of Internet Stability and Wide-Area Network Failures,” in *Proceedings of FTCS99*, June 1999.

Appendix A - Single Event Origin Oscillation

Two factors can collude to amplify a single origin transition into origin oscillation at an observer: delay and BGP policy. Delays, whether caused by slow links, overloaded routers, or topology, slow the propagation of prefix announcements. Because propagation time of an announcement or withdrawal can vary greatly, the same information about origin can arrive at an viewpoint through different paths over time. Policy contributes directly to oscillation by adding local autonomy to the decision process. For example, a LOCAL PREF attribute can be used to prefer one path over another. Where such preferences are used, The path selection algorithm can fluctuate between paths quickly where many new announcements with preferences arrive.

To illustrate, consider the following network of ASes:



Initially, assume that a multihomed prefix p is originated by AS 1, and that the network has reached quiescence for p. All ASes have a path to p. Arbitrarily, AS 4 chooses the path $[2 - 1]$ for the path to p, which is then propagated to ASes 5, 6, and 7. We denote AS 5 as our observer. AS 5’s path to p is $[4 - 2 - 1]$.

1. The origin of prefix p switches from AS 1 to AS 8. In response, AS 1 sends a withdrawal to AS 2 and 3.

2. AS 8 propagates its path to AS 7, and then AS 5 receives the path $[6 - 7 - 8]$ from 6 and selects it over its previous route (($[4 - 2 - 1]$) because of local policy.

3. AS 4 receives a withdrawal from AS 2 for $[2 - 1]$, and selects $[2 - 1]$. Note that the withdrawal of AS 1 as origin may be delayed between AS 1 and AS 3 or AS 3 and AS 4. However, AS 4 does not know yet that $[3 - 1]$ route is not valid and hence selects it.

4. AS 5 receives $[4 - 3 - 1]$ from AS 4 as a backup route and selects it over $[6 - 7 - 8]$ due to policy.

5. The withdrawal of $[3 - 1]$ finally arrives at AS 4, which in turn sends a withdraw of $[4 - 3 - 1]$ to AS 5.

6. AS 5 selects $[6 - 7 - 8]$ because it is the only route available.

Even in the absence of origin flapping or multiple origin prefixes, this simple network saw oscillations for four distinct origin changes. More complex topologies can amplify the effects of origin oscillation.