Durbin: Internet Outage Detection with Adaptive Passive Analysis

Asma Enayet
USC/ISI and
the Thomas Lord CS Dept.
Los Angeles, USA

John Heidemann
USC/ISI and
the Thomas Lord CS Dept.
Los Angeles, USA

ABSTRACT

Measuring Internet outages is important to allow ISPs to improve their services, users to choose providers by reliability, and governments to understand the reliability of their infrastructure. Today's active outage detection provides good accuracy with tight temporal and spatial precision (around 10 minutes and IPv4 /24 blocks), but cannot see behind firewalls or into IPv6. Systems using passive methods can see behind firewalls, but usually, relax spatial or temporal precision, reporting on whole countries or ASes at 5 minute precision, or /24 IPv4 blocks with 25 minute precision. We propose Durbin, a new approach to passive outage detection that *adapts spatial and temporal precision* to each network they study, thus providing good accuracy and wide coverage with the best possible spatial and temporal precision. Durbin observes data from Internet services or network telescopes. Durbin studies /24 blocks to provide fine spatial precision, and we show it provides good accuracy even for short outages (5 minutes) in 600k blocks with frequent data sources. To retain accuracy for the 400k blocks with less activity, Durbin uses a coarser temporal precision of 25 minutes. Including short outages is important: omitting short outages underestimates overall outage duration by 15%, because 5% of all blocks have at least one short outage. Finally, passive data allows Durbin to report this results for outage detection in IPv6 for 15k /48 blocks. Durbin's use of per-block adaptivity is the key to providing good accuracy and broad coverage across a diverse Internet.

1 INTRODUCTION

Internet outages are an economic and societal challenge. An outage costs Amazon $66k on 2013-08-19 [2]. Data-centers lose $5k per minute when users can not reach them [1]. Natural disasters, political events, software and hardware failure, human error, and malicious activity can cause Internet outages [3, 7, 9, 28]. Prior outage detection has shown outages are rare but ubiquitous [9, 10, 24, 26, 37].

Monitoring approaches use active measurement or passive traffic analysis. Active monitoring has vantage points (VPs) query destinations, responders prove reachability [17, 18, 23, 27, 33]. Passive methods instead infer outages by the absence of prior network traffic [14, 19, 29, 34].

Today's active outage detection provides good accuracy with tight temporal and spatial precision (around 10 minutes and IPv4 /24 network blocks), but they face two limitations. First, traffic from active observations draws abuse complaints and blocking from those who consider it intrusive. Second, active methods cannot see behind firewalls.

Passive outage detection today usually relaxes spatial or temporal precision, reporting on whole countries or ASes at 5 minute precision [14], or /24 IPv4 blocks with 30 minute precision [29], or require physical devices and so have limited coverage [34]. However, passive systems have some advantages: they pose no additional traffic on targets and so do not draw abuse complaints or blocking. In addition, they can see networks behind firewalls when those networks send traffic.

This paper proposes Durbin, a system that detects Internet outages based on passive analysis of data sources. Unlike prior passive systems, Durbin's detection algorithm is parameterized based on the historical data of each block, allowing it to cover both /24 IPv4 blocks and extend coverage to /48 IPv6 blocks. Durbin can vary the spatial and temporal precision of detection which provides high accuracy, even in networks with weaker signals, enhancing Durbin's effectiveness across a wider range of network conditions. Our approach provides a new, systematic approach to passive analysis to address these problems, making the following three contributions:

Detecting short outages: We know brief outages occur (short-burst DDoS attacks or pulse attacks, but prior systems do not detect outages shorter than 10 minutes for individual /24 blocks. Prior active detection systems use active probing and probes every 11 minutes, and cannot increase temporal precision without becoming excessively intrusive (which could result in abuse complaints or silent discard of

draft (under submission), June 2020, Marina del Rey, CA, USA

measurements). Prior passive detection systems detect short-duration outages, but at the cost of providing only much coarser, AS-level spatial precision.

Our new approach interprets passive data and can employ exact timestamps of observed data, allowing both fine spatial and temporal precision when possible. Our measurements show in §5.6 that around 5% of total blocks have 5 minute outages that were not seen in prior work. These short outages add up—when we add the outages from 5 to 10 minutes that were previously omitted to observations, we see that total outage duration increases by 20%.

Optimizing across a diverse Internet: Variation in Internet use means that outage detection systems should be tuned to operate well in each region. Our second contribution is to describe the first passive system that optimizes parameters for each block to provide fine spatial and temporal precision when possible, but can fall back to coarser temporal precision when necessary. By contrast, although prior passive systems optimize some parameters, they operate with a homogeneous global sensitivity, and therefore provide only coarse spatial coverage (at the country or AS level [14], or decreasing coverage). We instead exploit the ability to trade-off between spatial and temporal precision (§5.5), allowing some blocks to have less temporal precision. This flexibility means we can retain accuracy and increase coverage by reporting coarser results for blocks that otherwise would be ignored as unmeasurable. Our hybrid approach detects around 20% more outage duration than fixed parameters (§5.5).

Extending to IPv6: Our third contribution is to show that our new approach applies to IPv6 (§6). Since passive data comes from live networks, we avoid the active methods requirement for an accurate IPv6 hitlist. Although there has been considerable work on IPv6 hitlists [5, 6, 11, 12, 22, 36], their IPv6 coverage remains incomplete, particularly in the face of client preference for privacy-preserving addresses that are intentionally hard to discover. We show that coverage using passive data from `B-root` and an example of network services like DNS, is about 17% of the Gasser hitlist [12], and our approach could cover 5× more than Gasser if it used data from Wikipedia or $10^5\times$ more if using NTP [31].

Data availability and ethics: Our work uses publicly available datasets. Datasets for the input and results from our experiments are available at no charge [41]. Our analysis uses data from services, not individuals, so it poses no privacy concerns. We discuss research ethics in detail in Appendix A.

2 RELATED WORK

Internet outage detection systems can use either active data monitoring or passive data monitoring. We highlight how the outage detection system we propose compares to existing active and passive monitoring techniques for both IPv4 and IPv6 address blocks.

Outage Detection Systems using Active Monitoring probe the Internet from a set of vantage points, typically distributed across different networks, sending pings or traceroutes to most or all of the Internet.

ThunderPing first used active measurement to track weather-related outages [33]. They probe many individual addresses in areas with severe weather from around ten vantage points and report outages for individual addresses. Padmanabhan et al. later showed that outages sometimes occur at spatial scales smaller than a /24 block [24]. Like this prior work, we are interested in edge networks, but our goal is wide coverage, not just areas under severe weather.

Hubble finds potential Internet outages by probing all .1 addresses,triggering traceroutes to localize a potential outage [17]. LIFEGUARD extends Hubble to work around local outages caused by routing [18], detecting outages per routable prefix. We instead do passive traffic observations from network-wide services, and target outages in edge networks, using finer IPv4 /24s and also adding IPv6.

Trinocular provides precise measurements of Internet reliability to all "measurable" edge networks, 5.1 million /24 blocks in current datasets [27]. It uses adaptive ICMP probing every 11 minutes dynamically adapting how many probes are sent to balance traffic and accuracy. Our passive work adds coverage for firewalled regions and extends to IPv6.

Our passive monitoring technique does not increase traffic on target networks, unlike active systems. It also supports outage detection behind firewalls and we are the first to report results for IPv6. Finally, we support finer temporal resolution (up to 5 minutes) when possible, while retaining fine spatial resolution (IPv4 /24s).

Outage Detection Systems using Passive Monitoring infer outages by the disappearance of previously observed traffic. Passive systems watch traffic from some global network service such as a CDN, a DNS service, or Internet background radiation seen in network telescopes.

Dainotti et al. examined outages from censorship [9] as detected in observations from both BGP and Internet background radiation (IBR) as seen in network telescope traffic [21]. Chocolatine [14] formalizes this approach, uses SARIMA models to detect outages in IBR [25]. Each of these passive systems improves temporal resolution, either by using more input traffic or by growing spatial precision. For example, Chocolatine has 5 minute temporal precision, but only at the scale of entire ASes. We instead infer outages using data from passive sources which provide important insight into networks that block active probes.

CDN-based analysis provides /24 spatial precision at 1 hour temporal precision [29]. As a passive approach, it also provides global results for firewalled blocks, with broad coverage (2M /24 blocks). It optimizes expected response from the history of each block, but sets overall detection parameters globally. We provide good spatial *and* temporal precision, in part with additional per-block customization, and our approach could apply to CDN data to similar coverage.

Fontugne et al. use RIPE Atlas data to identify network outages [38]. Blink detects failures without controller interaction by analyzing TCP retransmissions [16]. Blink creates a characteristic failure signal when multiple flows aggregate retransmission information.

Our proposal differs from prior passive systems for several reasons. First, protocols such as DNS or NTP can provide much larger coverage than RIPE: we infer outages from existing networks instead of explicit status provided by the network service providers. Second, we can ensure finer temporal precision (5 minutes or less) by interpreting passive data and using exact timestamps of observed data. We optimize parameters for each block to provide precise results but may use coarser temporal precision when required. We discuss precision in detail in §5.4 and §5.5.

Hybrid Active and Passive Detection Systems: Disco detects outages by passively detecting correlated disconnection events across actively maintained connections from about 10k sites [34]. We instead detect outages by analyzing passive traffic with IP and timestamps and search for a gap in that traffic without injecting excess traffic. Our system customizes parameters for each block to optimize the performance of our model.

IPv6 Coverage: Prior work on outage detection only considers IPv4. IPv6 hitlists [5, 11–13, 22, 36] are a potential step towards IPv6 outage detection. Gasser et al. reported the first large IPv6 hitlist [13] using prior data and traceroutes to find 25.5k prefixes, 21% of what was announced at the time.

Building on this work, Beverly et al. use random probing to discover 1.3M IPv6 routers [5]. Entropy/IP models IPv6 use with information-theoretic and machine-learning techniques [11]. After training on 1k addresses, 40% of their 1M candidate addresses are active. Murdock et al. search regions near known addresses, discovering 55 M new active addresses [22]. AddrMiner also grows the hitlist from a seed set [35], finding 1.7 B addresses after dealiasing. Rye and Levin instead turn to passive addresses in NTP traffic, finding an impressive 7.9 B addresses [31], showing that prior approaches missed many client addresses.

We directly use passive data, like Rye and Levin, avoiding the need to build a hitlist for active probing. We provide the first reports of outage detection in IPv6.

3 METHODOLOGY

Durbin methodology is to observe passive data from some service (§3.1). From history, it models the probability traffic arrives from any address in some time period (§3.2). It then detect deviations from this model with Bayesian inference, reporting traffic reduction as an outage (§3.3). Finally, when possible, it combines observations for multiple observers in a block (§3.4). We optimize parameters to trade off spatial and temporal precision for each block (§3.5).

3.1 Data Requirements

Durbin uses passive traffic observations from network services like DNS or darknets. Many network services could serve as input to Durbin: large websites like Google, Amazon, or Wikipedia; web infrastructure like CDNs; infrastructure services like DNS or NTP. darknets are an alternative, with darknets operated by CAIDA and Merit and smaller telescopes operated by many parties. Durbin's requirement is that the data source accurately reports communication from a client IP address at some specific time. Ideally, it provides such data real-time or near-real-time. Our approach works equally well for IPv4 and IPv6, as we show in §6.

Although we can use many possible data sources, we evaluate our system using two specific systems: First, we use traffic arriving at `B-root` [40], one of the 13 authoritative DNS Root services [30]. Second, we evaluate it using passive traffic arriving at the Merit darknet [20]. These two very different data sources show that Durbin generalizes and can apply to many potential passive data sources.

Each data source collects traffic and shares the time and partially-anonymized source IP address of each flow. We take several steps to minimize any privacy risks to users generating the data. For `B-root` we omit other fields (such as query name) that are not required for detection. For the Merit darknet we retain most fields only until we filter for spoofing. Durbin models traffic at the block level, so we preserve the network portion of the IP address (IPv4: /24, IPv6: /48) and anonymize the remaining bits to shuffle individual users.

We see `B-root` as representative of a large Internet service that receives global traffic. For `B-root`, we see about 700M queries from about 7M unique locations per day. While large, this coverage is much smaller then large websites like Google, Amazon, or Wikipedia. In §6.2 we quantify `B-root` coverage, show that Wikipedia would provide better coverage than today's public IPv6 hitlists. Thus the smaller coverage of `B-root` represents a limitation in the data that we have access to, but *not* a fundamental limitation in our approach. Although `B-root` does receive spoofed traffic when it is attacked, when not under attack all queries arriving at `B-root` are from legitimate recursive resolvers, and we assume the source address indicates a valid, active IP address. We believe the Durbin algorithm can apply to other data sources (such as NTP [31]), although any new source will require tuning Durbin parameters, as we do for our current sources (§3.5).

We also evaluate over the Merit darknet, an example darknet. A darknet does not run real services and so should receive no legitimate traffic. Incoming traffic is often network scanners, malware attempting to propagate itself, or

Observation	Prior	(Observation | Prior)
Negative	Down	1
Negative	Up	$1-\pi(a)$
Positive	Down	$\pi(a)$
Positive	Up	$1-P(\text{neg}\|\text{down})$

Table 1: Responses when a timebin has traffic or not

backscatter, where someone spoofed the darknet as the source IP for traffic sent to another party. A darknet's source addresses suggest a live network, but could be spoofed. We filter darknet traffic to discard traffic where we do not believe the source address is legitimate. We follow CAIDA's filtering rules [8], discarding packets with TTL exceeding 200 which are not ICMP; those with IPv4 sources ending in .0 or .255 or identical source and destinations; and protocols 0 and 150.

3.2 Learning From History

Durbin models expected traffic from address a from long-term observations. Each address has three parameters: the timebin duration for detection, $T(a)$; the historical probability we see traffic in that timebin, $\pi(a)$, and model has enough data to be consistent or *measurable*, $M(a)$. We discuss how we generalize from addresses to blocks in §3.4, and how set these parameters in §3.5.

We divide the timeline into specific timebins for each address, each lasting $T(a)$ seconds. In principle, $T(a)$ may range from 1 to 60 minutes, but currently we select from short and long options (typically 5 or 25 minutes, §3.5).

The *active probability* $\pi(a)$, is the probability that traffic arrives in timebin $T(a)$. We compute $\pi(a)$ for each address from from long-term observation of address a, based on the last d days. We currently use d of two days.

Finally, we define measurability $M(a)$, when the address has enough data to provide signal, when $\pi(a) > \theta_{measurable}$.

Durbin works well for short-term outages, but outages lasting longer than the training period will disappear when training considers only the outage. Long-term outages are difficult to handle in most outage detection systems. In general, external information is required to distinguish long outages from changes in usage.

3.3 Address-Level Outage Detection

To detect outages we estimate the belief $B(a)$ as probability that address a is reachable, from 0.0 to 1.0, certainly down to certainly up. We classify an address as unreachable (down) when $B(a) < \theta_a$ and reachable (up) when $B(a) > 0.95$. We consider middle values ($\theta_a < B(a) < 0.95$) to indicate uncertainty, with the hope that information in the next timebin will resolve its state. We set the threshold θ_a to 0.6, and validate each of these parameters in §5.3.

Table 1 shows how belief changes according to conditional probabilities. We compute belief $B(a)$ by applying Bayesian inference on a stream of observations in each timebin $T(a)$ if the address has traffic (positive) or not (negative). For each timebin with a positive and negative observations compute new belief $B'(a)$ from prior belief $B(a)$ as:

$$B'(a) = \frac{\pi(a)B(a)}{\pi(a)B(a) + (1 - P(no|down))(1 - B(a))} \quad (1)$$

$$B'(a) = \frac{(1 - \pi(a))B(a)}{(1 - \pi(a))B(a) + (P(no|down))(1 - B(a))} \quad (2)$$

We illustrate shifts in belief with case studies in §4. These equations get stuck when $B(a)$ reaches 0 or 1, so we limiting $B(a)$ to the range B_{min} to B_{max}, currently set to 0.1 and 0.95.

3.4 Block-Level Outage Detection

We next merge address-level belief ($B(a)$) of all addresses in a block to determine block-level belief $B(b)$.

We study /24 address blocks as the smallest unit of spatial coverage, following prior work [27, 29]. Combining results from multiple addresses in one block can improve accuracy since we can get more information about the block.

Following our definition of addresses, we consider block status in timebins of duration $T(b)$, and that value can vary by block. This approach follows address analysis with $T(a)$ in §3.2. We define $B(b)$ as the belief in the status of each /24 block b. Unlike addresses $B(b)$ is not inferred from data, but instead, it combines all address beliefs in that box, defining:

$$B(b) = \max(B(a_i)) \forall a_i \in b$$

We merge address-level detection results for each $T(b)$ to get block-level results. Block detection uses a potentially different threshold (θ_b), identifying an outage when $B(b) < \theta_b$, a reachable block when $B(b) > 0.95$, and otherwise identifying the block as uncertain. We set $\theta_b = \max_{a \in b} \theta_a$.

When considering block-level outage detection we select other block-level parameters from the address-specific parameters, taking the minimum timebin and considering a block measurable $M(b)$ if any address is measurable:

$$T(b) = \min_{a \in b} T(a) \quad \text{and} \quad M(b) = \bigvee_{a \in b} M(a)$$

3.5 Optimizing Parameters for Each Block

We optimize Durbin parameters for each block to trade-off accuracy and coverage. We adjust timebin duration ($T(b)$) and thresholds for measurability (θ_b, θ_{sparse} and $\theta_{measurable}$) based on amount of traffic to each block.

Durbin first selects timebin duration ($T(a)$) to provide rapid detection for addresses with frequent traffic and reliable detection for addresses with sparse traffic. Currently we select between two timebin durations, based on the threshold θ_{sparse}, labeling addresses where $\theta_{measurable} \leq \pi(a) <$

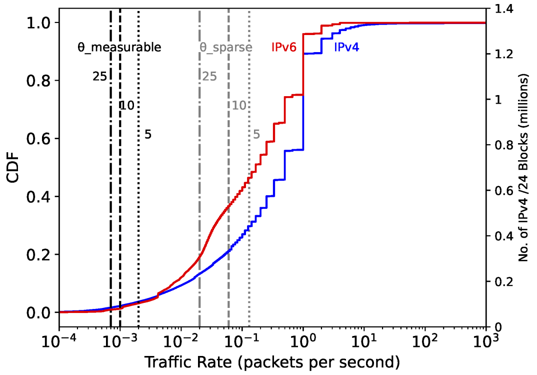


Figure 1: Traffic per address for B-root on 2019-01-10

θ_{sparse} as sparse and those with $pi(a) > \theta_{sparse}$ as frequent. With B-root as our data source we set θ_{sparse} as 0.6, and use 5 and 25 minutes as short and long timebins. With the Merit darknet as the source we set $\theta_{measurable}$ and θ_{sparse} as 0.6 and $T(b)$ is 20 minutes. We use the same Durbin algorithm for each source, but choose parameters (θ_{sparse} and $T(\cdot)$) based on analysis of coverage and accuracy in §5.4 and §5.5 as we vary $T(a)$. As future work, we plan to vary $T(b)$ for merit and to explore allowing $T(a)$ to vary continuously.

Finally, we define $T(b) = \min(T(a_i)), \forall a_i \in b$: the block can be as sensitive as the best address, analogous to belief following the most reliable address (§3.4).

Currently thresholds (θ_a and θ_b), are fixed (both at 0.6). Because belief ranges from 0 to 1, fixed values here seem appropriate. However, belief adapts based on block history (§3.3) and so it is customized for each block.

4 DEMONSTRATING VIABILITY

We next show that Durbin feasible, first by confirming that a network-wide service has enough data to provide coverage of many blocks, then by showing how Durbin handles sources with different amounts of traffic.

4.1 Traffic Rates per Address

We first characterize traffic in our data sources.

4.1.1 Traffic Distribution in B-root. Figure 1 shows a cumulative distribution of traffic arrival rates for each address in B-root by IPv4 (the blue line) and IPv6 (red) over one day.

Both IPv4 and IPv6 show similar distributions, although IPv4 reports on 1.2M blocks while IPv6 is only 13k. The similarity of these distributions suggests the same algorithm will apply to both IPv4 and IPv6. This similarity is important because we show good accuracy for IPv4 (§5.7) where we can compare to alternative methods. As the first public IPv6 approach, we cannot compare to alternatives there, but since it is the same algorithm, we expect IPv4 accuracy to apply to IPv6 as well.

This data allows us to estimate coverage of Durbin-with-B-root. Coverage depends on our parameters for timebin ($T(b)$, §3.5) and measurability ($\theta_{measurable}$, §3.2), as we evaluate in §5.4. Here we see measurability for three different timebin sizes (25, 10, and 5 minutes, vertical black lines from left to right). Almost all B-root sources are measurable (90% for 5 minutes, and 95% for 25 minutes).

The gray vertical lines show the division between sparse and frequent sources, determined by θ_{sparse}. Most sources (80% of IPv4 and 50% of IPv6) and have frequent data (to the right of the gray vertical line). The ability to pick up sparse sources (blocks between the black and gray lines) adds coverage for another 20% or 50% of all blocks (100k to 400k for IPv4, and 1000 to 6500 for IPv6).

4.1.2 Traffic Distribution in the Merit darknet. Traffic in the Merit darknet differs because it is not an active service, but receives only unsolicited traffic. Similar analyses reveal a wider variability in traffic arrival probabilities for IPv4.

The distribution of traffic in the Merit darknet is similar to that of B-root, suggesting that the same algorithms apply, although perhaps with different parameters. However, the Merit darknet has many more sparse blocks (80%, compared to 30% for B-root). The smaller amount of traffic in a darknet shows the importance of observing network services, but scanning detected in darknets can can provide information about the status of otherwise silent and firewalled blocks.

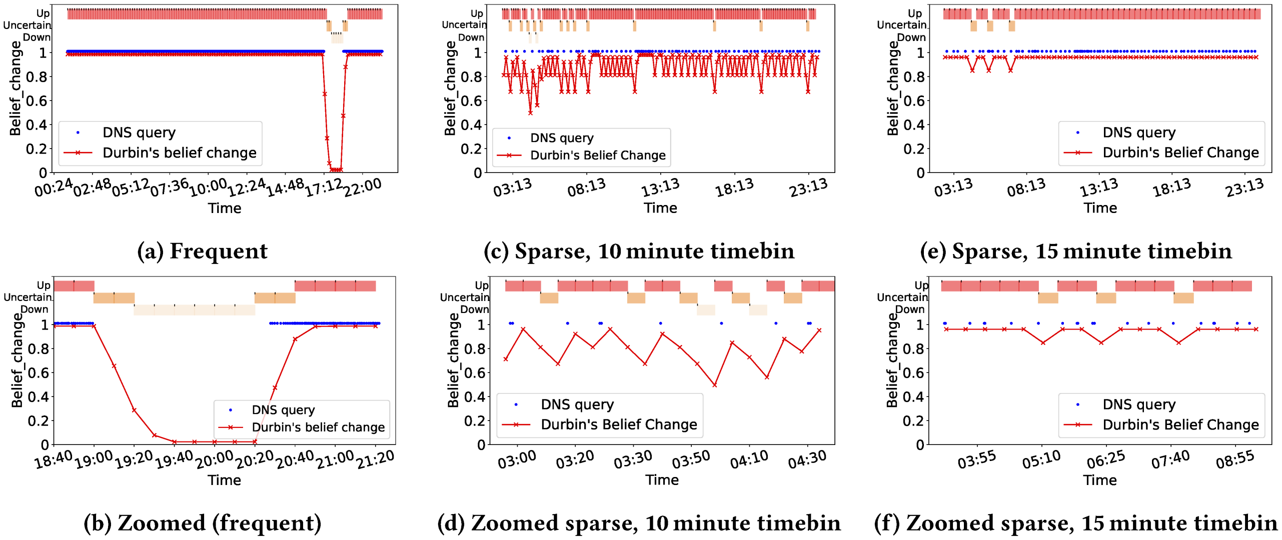
4.2 Belief for Frequent Traffic

We next show how Durbin can react very quickly when an address has frequent traffic. From §3.3, an address a has frequent traffic if $\pi(a) > \theta_{sparse}$ Here we pick one example address where $\pi(a)$ is 0.9; we see similar results in the hundreds of other addresses that have similar traffic levels.

For addresses with frequent traffic, gaps are very unusual, so belief changes quickly from certainly up to down. Figure 2a shows belief (the red line) for our example addresses with frequent traffic (blue dots). In Figure 2a, blue dots indicate traffic to an address at B-root on 2019-01-12 with a gap from 19:20 to 20:20.

We show timebins as boxes at the top of Figure 2a, showing reachable with red boxes (top row, with traffic), uncertain as dark orange in the middle, and unreachable (no traffic) with light orange, on the third row. We then zoom on the outage period in Figure 2b.

Figure 2a shows that frequent traffic (the solid blue dots before 19:00) yield confident of reachability (belief at the 0.90 maximum). However, the traffic gap from 18:56 to 20:26 (see Figure 2b) causes belief to drop through uncertainty (two timebins at 19:00) to unreachable (at 19:20). This example shows how Durbin reacts quickly given frequent traffic.



(a) Frequent

(c) Sparse, 10 minute timebin

(e) Sparse, 15 minute timebin

(b) Zoomed (frequent)

(d) Zoomed sparse, 10 minute timebin

(f) Zoomed sparse, 15 minute timebin

Figure 2: Belief change in Durbin-with-B-root, varying traffic frequency and timebin duration. Data: 2019-01-10.

4.3 Belief for Addresses with Sparse Traffic

For addresses with sparse traffic (when $\pi(a) < \theta_{sparse}$), Durbin must be more cautious in determining when an outage occurs. Here we pick one example with $\pi(a) = 0.6$. Again we see similar results in the hundreds of other, sparse addresses where $\theta_{measurable} < \pi(a) < \theta_{sparse}$.

Figure 2c shows one day (2019-01-13) for this representative address, with sparser traffic (blue dots) causing belief (the red line) to vary. To see Durbin coping with gaps in data, Figure 2d zooms in a 90 minute period starting at 2:50. Here belief often drops as multiple timebins pass without traffic, all the way to the outage at 3:40 and 4:10. While relatively rare, these false outages are a problem that comes from attempting to track blocks with sparse traffic—we simply do not have enough information in a short timebin to make consistently good decisions. We can discard such blocks, but that reduces coverage. In the next section, we show how to gather more information to correct this situation.

4.4 Sparse Traffic with a Longer Timebin

The false outages in Figure 2d result from short timebins not observing enough information to make good decisions. We next show that a longer timebin addresses this problem.

Figure 2e shows one day for the same address as Figure 2c), now with a 25 minute timebin.

With longer timebins, $\pi(a)$ rises to 0.9, enough information that we never detect false outages, and we see only three unknown periods for the same day results in several false outages and many unknown periods with a shorter timebin. Figure 2f zooms 5 hours starting at 3:00 to show how a longer timebin bridges gaps.

We use these examples to motivate our overall design choices: using Bayesian Inference to adapt belief based on multiple rounds of observation, and adapting a belief threshold and tuning parameters based on address history (§5.3).

5 VALIDATION OF OUR APPROACH

We next validate our results, examining accuracy with positive predictive value (PPV), recall, and true negative rate, sensitivity to the parameters of belief threshold and timebin duration. We use *PPV* instead of information retrieval's *precision* to avoid confusion with terms temporal and spatial precision, both key "knobs" in our design. We then demonstrate that our approach can detect short outages and trade-off spatial and temporal precision.

We validate our results by comparing them to prior data sources. We run Durbin with both both 7 days of B-root and 7 days of darknet as data sources. We compare to Trinocular [27], a system using active analysis with very broad coverage (about 5M IPv4 /24s). We validate our short-duration outage results by comparing them to Disco [34]. It uses RIPE Atlas data to detect correlated disconnections, making it sensitive to short-duration outages. We would like to compare to other passive systems, but Durbin's much greater spatial sensitivity makes direct comparison to Chocolatine impossible [14], and CDN data is not publicly available [29].

We show Durbin system has good accuracy (§5.1), how belief threshold (§5.3) and time-bin (§5.4) affects accuracy and coverage, we can detect short outages (§5.6), Durbin can trade-off between temporal and spatial precision (§5.5) and our system is consistent over a long period of time (§5.7).

5.1 Accuracy of Durbin-with-network services

5.1.1 Direct comparison of Durbin-with-`B-root`. To compare Durbin-with-`B-root` to Trinocular we first find all /24 blocks observed both, yielding about 880k blocks in both datasets for the 7 days starting on 2019-01-09. We then compare the block-durations (in seconds) each system identifies as reachable or not.

Table 2a shows this confusion matrix, defining Trinocular outages as ground truth. We define a false outage (*fo*) for a block when Durbin predicts the block is unreachable, but Trinocular can reach it, and have similar definitions for false availability (*fa*), true availability (*ta*), and true outages (*to*).

PPV is uniformly good ($(ta/(ta + fa)) = 0.9999$): Durbin's reported availability is almost always correct.

We next consider true negative rate (TNR) to quantify what duration of outages we report are true ($(to/(to + fa))$. Our TNR is good, at 0.8417, but lower than the PPV. Strong TNR means we correctly estimate outage duration, but TNR is lower than PPV because outages are rare, making small differences between Durbin and Trinocular more noticeable.

5.1.2 How does imprecise comparison affect accurate results? Both Trinocular and Durbin measure with fixed timebins, and misalignment between the two systems inevitably results in small differences (just like comparisons measured in whole numbers of meters and feet will never be identical). Measurement precision results in lower-than-expected recall ($(ta/(ta + fo)) = 0.6282$).

Recall suggests that we often find shorter outages than Trinocular. Trinocular's temporal precision is coarser than Durbin (±330 s vs. ±150 s), so Durbin can detect shorter outages. Enhancing Durbin's precision using exact timestamps is a potential future research direction.

Temporal precision is affected by several factors. Most important is choice of timebin duration, $T(b)$. If we select precise timing with a small timebin, then we will either lose accuracy (because many timebins have no responses and so we need to use a lower belief threshold), or we must increase our spatial scale to provide reliable decisions. Also, actual outages do not always line up with timebins, so any timebin-based system may report outages up to one timebin late. Finally, depending on block history, it make take multiple timebins to shift belief.

5.1.3 Re-evaluating with Precision-Aware Comparison: Our measurement accuracy is determined by temporal precision, but smaller differences provide metrics that can be misleading, exaggerating differences that reflect random phase of measurements rather than differences in the underlying conclusions. To factor out the measurement system and get at the underlying phenomena we next consider precision-aware comparison.

We define *precision-aware comparison* as ignore differences that are shorter than the measurement timebin for a given block. Ignoring these short differences is justified timebin phase is arbitrary, and it reflects more on quantization of outage detection into timebins than on the actual correctness of the underlying method. We keep any differences lasting longer then the block's timebin, since those represent real disagreement in results. We then computing PPV, recall and TNR on these "precision-aware" observations. (Our precision-aware comparison is analogous to how CDN-based outages were comparing to only Trinocular outages lasting longer one hour, the CDN quantum [29], however we present both the full data §5.1.1 and precision-aware comparisons here.)

Table 2b shows comparisons with observations with precision-aware time bins. Now, PPV is uniformly good (inference of blocks being reachable is nearly always correct) as before which is 0.9999. Also, Recall rises to near-perfect 0.9985 (from 0.6282), because alignment eliminates what would otherwise be many short, false outages.

The number and duration of false outage events drops to one-quarter of before, from 31.09 Gs to 78.16 Ms. We believe these improved results better reflect the true ability of passive observation to detect events, once with rounded time bins.

5.1.4 Comparison with event counts . Our prior comparisons consider block-seconds, giving longer differences heavier weight. In Table 2c we instead count events (state changes).

When considering events instead of time, recall is dominated by many blocks that never change state, and recall and TNR are disproportionately by a few blocks that frequently change state. Since a stable block has one correct event, but a block that frequently changes state may have hundreds of changes, events magnify the effects of frequently changing blocks. Thus events show lower recall and TNR. As future work, we plan to look for these frequently changing blocks to account for them with more conservative parameters.

5.1.5 IPv6 Correctness. We validate the Durbin algorithms above for IPv4, finding excellent PPV and good TNR. For IPv4, we can validate against other systems, but there are no prior results for IPv6 against which to compare.

Fortunately, our *IPv4 and IPv6 algorithms are identical*, and we showed in Figure 1 that many addresses have similar traffic rates. Since correctness depends on traffic rate and regularity and these are similar in IPv4 and IPv6, we expect our accuracy for IPv4 outage detection to apply to IPv6.

5.2 Accuracy of Durbin-with-darknet

To confirm Durbin's accuracy the Merit darknet, a different data source, we repeat these comparison.

5.2.1 Directly comparing Durbin-with-darknet. To compare Durbin using the Merit darknet data with Trinocular, we find all /24 blocks in both and compare the duration for each system labels as reachable or not.

	seconds
True availability = TP	52,525,765,695
True outage = TN	13,147,965
False availability = FP	2,471,178
False outage = FN	31,087,360,212
PPV	0.9999
Recall	0.6282
TNR	0.84178

(a) Direct comparison of duration

	seconds
True availability = TP	52,525,765,695
True outage = TN	13,147,965
False availability = FP	2,471,178
False outage = FN	78,163,261
PPV	0.9999
Recall	0.9985
TNR	0.8417

(b) Precision-aware comparison of duration

	events
True availability = TP	359,415
True outage = TN	508
False availability = FP	241
False outage = FN	24,218
PPV	0.9976
Recall	0.9368
TNR	0.6782

(c) Precision-aware comparison of events

Table 2: Confusion matrix for long-duration outages for Durbin-with-`B-root` (Dataset: 2019q1)

We see 66,776 blocks in Durbin-with-the Merit darknet, compared to 5,210,923 blocks in Trinocular, with 59,979 blocks in the intersection. the Merit darknet has 66,776 blocks, but 3,198 are only active because of spoofing. We, therefore, compare the 63,578 remaining to the 5,210,923 blocks in Trinocular, finding 59,979 in the intersection.

We compared seven days of data starting from 2021-01-10. Table 3a shows the confusion matrix from this analysis when we define Trinocular outages as ground truth.

PPV is very good (0.9810), showing that Durbin-with-the Merit darknet's reported availability is almost always correct. True negative rate (TNR) estimates how many outages are correct. TNR (0.7334) is lower than PPV because outages are rare, so small differences between Durbin-with-darknet and Trinocular are noticeable. Finally, recall (0.8488) is also quite good, indicating that Durbin successfully detects a high proportion of the outages identified by Trinocular.

5.2.2 Durbin-with-the Merit darknet, by events. We next compare the number of outage events, comparing Durbin-with-the Merit darknet against Trinocular.

Similar to Durbin-with-`B-root`, many blocks are always up, giving us good recall (0.8137). A small fraction of blocks (about 10%) are detected poorly in Durbin and produce many false events. These blocks have sparse traffic, and when multiple consecutive timebins have no traffic, false outages result. These sparse-traffic blocks reduce TNR. As future work we plan to examine making timebin duration more adaptive to provide more reliable results for these blocks. This problem has observed [29] before in active detection, where it was resolved by observing more data to confirm or reject the outage, a rough equivalent to increasing the timebin duration [4].

5.3 Sensitivity of Belief Threshold

We next examine the sensitivity of our results to the belief threshold (θ_b) an important parameter discussed in §3.3. In our model, true outage detection varies with the change in belief threshold. A higher belief threshold can guarantee not to get false reports on short gaps. It also guarantees the detection of true outages. But too high a threshold will miss very brief outages. We customize parameters to find a middle ground to balance these two competing requirements.

To study the belief threshold we vary belief threshold θ_b and hold the time bin at $T(b) = 10$ minutes. We compare the impact of belief threshold θ_b of Durbin-with-`B-root` and Durbin-with-the Merit darknet against Trinocular, quantitatively (Figure 3a and Figure 3b) and graphically (Figure 4).

Figure 3a and Figure 3b show the trend change by looking at parameters PPV, recall, and TNR with varying belief thresholds for 0.3, 0.6, and 0.8, and Figure 4 compares TNR as threshold varies for more values between 0.2 and 0.9.

TNR is similar for threshold 0.6 or more. We chose $\theta_b = 0.6$ to maximize sensitivity to short-duration outages. We see the benefits of more sensitive detection in Figure 3a where $\theta_b = 0.6$ reports shorter outage duration (78.16 Ms less of the 87.85 Ms before) and larger true availability (52.52 Gs duration than 50.83 Gs before.) compared to $\theta_b = 0.8$.

This evaluation also shows the cost of a low threshold. False outages and false availability are higher with $\theta_b = 0.3$ compared to thresholds of 0.6 or 0.8. (For example, in Figure 3a, we see 2× more false availability and 26× more false outages.) Both recall and TNR are lower for $\theta_b = 0.3$ because Durbin is detecting short inactive periods as outages.

5.4 Sensitivity to Timebin Duration

Durbin optimizes parameters to provide temporal precision when possible, but falls back on coarser temporal precision when necessary to improve coverage and accuracy for both sparse and dense blocks. §5.3 compares PPV, recall and TNR as belief thresholds vary while setting $T(b)$ as constant. In this section, we will vary $T(b)$ and set the belief threshold as constant (0.6) to see the trend change in PPV, recall and TNR.

Table 4a shows Durbin's recall and TNR are very sensitive to timebin duration ($T(b)$). Longer durations of $T(b)$ are more likely to miss short outages but improve coverage by including addresses with sparse data.

	seconds
True availability = TP	3,064,574,810
True outage = TN	9,833,309
False availability = FP	59,215,390
False outage = FN	546,190,291
PPV	0.9810
Recall	0.8488
TNR	0.7334

(a) Comparison of durations

	events
True availability = TP	72,787
True outage = TN	524
False availability = FP	210
False outage = FN	16,656
PPV	0.9971
Recall	0.8137
TNR	0.7138

(b) Comparison of events

Table 3: Long-duration outages after precision-aware comparison for Durbin-withthe Merit darknet (Dataset: 2021q1)

	Belief Threshold (θ_b)		
	0.3	0.6	0.8
TA	50,834,003,114	52,525,765,695	55,374,729,345
TO	16,964,568	13,147,965	12,934,568
FA	5,452,416	2,471,178	2,923,456
FO	2,347,645,465	78,163,261	87,853,262
PPV	0.9999	0.9999	0.9999
Recall	0.9558	0.9985	0.9984
TNR	0.7567	0.8417	0.8191

(a) Durbin-with-B-root

	Belief Threshold (θ_b)		
	0.3	0.6	0.8
TA	3,264,854,345	4,908,712,735	4,345,576,243
TO	7,824,539	4,711,215	3,528,342
FA	5,026,362	1,826,015	1,124,462
FO	475,252	361,615	198,573
PPV	0.9910	0.9914	0.9997
Recall	0.9988	0.9981	0.9996
TNR	0.6088	0.7206	0.7586

(b) Durbin-with-the Merit darknet.

Figure 3: Confusion matrix as belief threshold varies for Durbin-with-B-root and Durbin-with-the Merit darknet. Times in seconds.

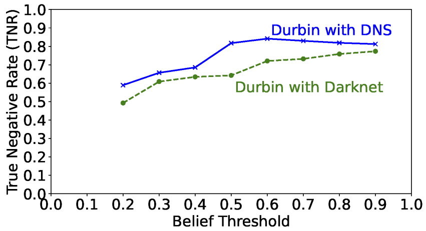


Figure 4: True outage detection rate for Durbin-with-B-root and Durbin-with-the Merit darknet.

	Time bin duration ($T(b)$)		
	15 minute	10 minute	5 minute
TA	56,492,461,162	52,525,765,695	47,456,373,912
TO	11,234,345	13,147,965	14,092,345
FA	3,043,362	2,471,178	2,001,769
FO	72,036,450	78,163,261	128,124,934
PPV	0.9999	0.9999	0.9999
Recall	0.9999	0.9985	0.9973
TNR	0.7821	0.8417	0.8756

(a) Durbin-with-B-root

	Time bin duration ($T(b)$)		
	30 minute	20 minute	10 minute
TA	4,974,983,931	4,908,712,735	3,064,574,810
TO	4,025,307	4,711,215	9,833,309
FA	2,021,539	1,826,015	1,615,440
FO	267,706	361,615	546,190
PPV	0.9997	0.9914	0.9910
Recall	0.9996	0.9981	0.9988
TNR	0.6656	0.7206	0.8588

(b) Durbin-with-the Merit darknet

Table 4: Confusion matrix as $T(b)$ varies for Durbin-with-B-root and Durbin-with-the Merit darknet. Times in seconds.

	events
True availability = TP	31,115
True outage = TN	2,030
False availability = FP	735
False outage = FN	1,799
PPV	0.9769
Recall	0.9453
TNR	0.7341

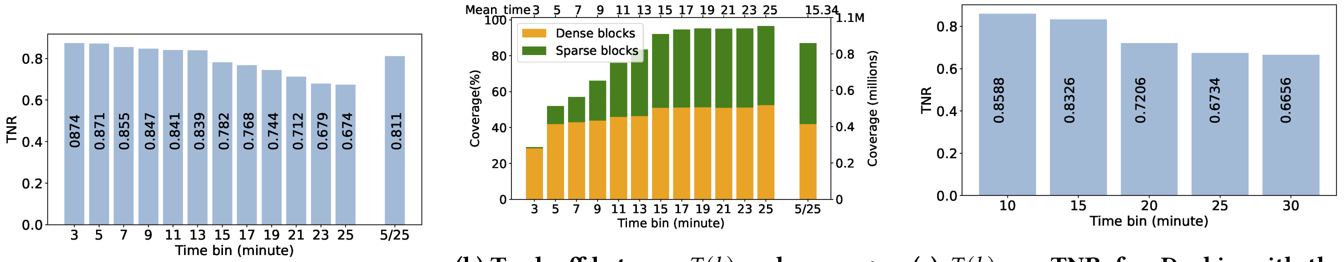
Table 5: Short-duration outages for Durbin-with-B-root (events)

On the contrary, short $T(b)$ can reduce coverage because it means that we are analyzing a smaller amount of traffic data at once making it more difficult to identify patterns or anomalies in the data, especially for sparse blocks or regions with low traffic volume. In Table 4a we show accuracy for three timebin durations in detail and in Figure 5 we add more reference points and see the trend change of TNR and coverage.

5.4.1 How does accuracy vary with timebin? Timebin duration is an important parameter to Durbin. We next vary timebin duration's influence on accuracy to justify Durbin's choice of a 10 minute timebin. We hold the belief threshold constant at $\theta_b = 0.6$.

In Table 4a, with a 10 minute timebin for both sparse and dense blocks, the performance of PPV, Recall, and TNR is outstanding: 0.9999, 0.9985, and 0.8417, respectively. In comparison, using a 5 minute timebin reduces Recall slightly (0.9973).

Recall is lower for 5 minute timebins because it increases the number of false outages. We see this change as the false outage duration increases to 128.12 Ms from 78.16 Ms as we go to 5 minutes from 10 minute timebins. These false outages occur in blocks with infrequent traffic (§5.1). Although a shorter timebin results in some false outages, it also allows Durbin to detect previously missed short outage. Durbin identifies one empty timebin without traffic as an outage



(a) $T(b)$ **vs. TNR for Durbin-with-B-root**

(b) Tradeoff between $T(b)$ **and coverage for Durbin-with-B-root**

(c) $T(b)$ **vs. TNR for Durbin-with-the Merit darknet**

Figure 5: Sensitivity of the timebin for Durbin-with-B-root and Durbin-with-the Merit darknet.

when the timebin duration is short, even if there is no actual outage shown in §4.

When timebin duration is longer, the true outage duration (in seconds) falls somewhat. Again, this reduction is because some short-duration outages are missed. Table 4a shows short timebins have more true outages (compare $T(b)$ of 15 minutes vs 10 or 5). Outage duration is reduced from 14.09 Ms (when $T(b)$ is 5 minute) to 11.23 Ms (when $T(b)$ is 15 minute).

Here we examined only outage duration. In future work, we plan to evaluate counts of outage events in addition to durations.

We conclude that $T(b) = 10$ minutes provides the best choice when possible. However, we select a longer duration from block with sparse traffic, as described in §5.5.

5.4.2 How coverage varies with timebin? The prior section examined three different timebins. We next consider a wider range of options and study how TNR and coverage vary with timebin.

We can get both good coverage and TNR by setting short-duration timebin for dense blocks and long-duration timebin for sparse blocks. We next experiment with different timebin durations to see how both sparse and dense blocks can achieve good coverage and TNR.

In Figure 5b we vary $T(b)$ and study coverage. (We hold the belief threshold constant at $\theta_b = 0.6$). We vary $T(b)$ from 3 minutes to 30 minutes and observe the coverage. Here we define coverage as the percentage of observed B-root blocks in a time bin with respect to the total number of existing B-root blocks.

Our coverage varies as a function of timebin because we pick coverage based on history discussed in §3.2. In short-duration $T(b)$, we only observe dense blocks, which results in a high true outage detection rate (0.874) but low coverage (around 30% dense blocks and 2% sparse blocks of all measurable blocks). The short duration captures only a small amount of traffic in one $T(b)$, which excludes sparse blocks and leads to low coverage.

With long-duration $T(b)$ the coverage improves by including more sparse blocks, but the TNR suffers.

Figure 5a shows TNR for timebin sizes from 3 to 25 minutes (left bars). As timebin increases, more blocks become measurable and we increase coverage as we describe next. When $T(b)$ is 25 minute the coverage is around 95% but the TNR is 0.674 because of the inclusion of sparse blocks. By this analysis, we can say that Durbin can trade-off between spatial and temporal precision which is described in the next section §5.5

In future work, we will perform two separate analyses: holding the coverage constant and observing changes in TNR as we varied $T(b)$ and allowing the coverage to vary while we monitored changes in TNR as we varied $T(b)$. By comparing the effects of $T(b)$ under these two conditions, we can determine how much each factor contributes to changes in TNR and understand how timebin and coverage interact with each other.

5.4.3 How accuracy varies with timebin, in Durbin-with-darknet? In this section, we show how varying the timebin duration influences outage detection performance, specifically in terms of metrics such as PPV, Recall, and TNR. Shorter timebins improve the detection of shorter outages, especially for dense blocks, while longer timebins enhance reliability for sparse blocks by reducing false positives. In Table 4b, with a 10 minute timebin for both sparse and dense blocks, the performance of PPV, Recall, and TNR is outstanding, achieving values of 0.9910, 0.9988, and 0.8588, respectively.

When timebin duration is longer, we observe a reduction in the true outage duration (in seconds), indicating that some outages are missed. Similar to Table 4a, Table 4b shows short timebins have more true outages (compare $T(b)$ of 30 minutes vs 20 or 1).

Therefore, if we have a fixed timebin then 20 minutes time bin seems good for Darknet. We can see around 5% more true outages when the time bin is 20 minutes. But if we have a variable timebin then different timebins for dense and sparse blocks are good §5.5.

5.5 Trading Between Spatial and Temporal Precision

We exploit the ability to trade-off between spatial and temporal precision while preserving accuracy. We customize parameters to treat each block differently, allowing different regions to have different temporal and spatial precision. As a result, we can get coverage in sparse blocks, although to get good accuracy we must use coarser temporal precision.

In Figure 5b we evaluate this trade-off, showing that we have fine precision for the dense blocks, but require coarser precision to cover blocks with sparse traffic across the left bars. The rightmost bar, labeled 5/25, shows a hybrid system where blocks with dense traffic use $T(b)$ of 5 minutes, while those with sparse traffic have $T(b)$ of 25 minutes.

With varying $T(b)$ values (customized to block traffic), we obtain broad coverage: 85% of all blocks with B-root traffic. Varying $T(b)$ by block also provides a good TNR (0.811). By contrast, a strict 25 minute $T(b)$ has TNR 0.647, because coarser precision can miss short outages. Comparing the rightmost two bars in Figure 5a, this is about a 20% improvement in TNR.

This comparison shows the advantage of tuning parameters to each block to maximize coverage *and* accuracy.

5.6 Can We Detect Short-Duration Outages?

We next demonstrate that Durbin can detect shorter outages than prior systems, in part because we can trade-off spatial and temporal precision with coverage and accuracy. Here, we examine 5 minute outages, with a belief threshold of 0.6 and $T(b)$ to 5 minute.

To validate our short-duration outage results, we compare them to Disco [34] using RIPE Atlas data [38] as ground truth. Using RIPE data, Disco infers that multiple concurrent disconnections of long-running TCP connections in the same AS indicate a network outage.

Although its coverage is only about 10k /24s, we use Disco for ground truth to compare short outages because it reports 5 minute outages (unlike Trinocular's 11 minutes).

We study all 10.5k /24 blocks observed from both Durbin using B-root data that also have data from RIPE Atlas over 7 days of data starting on 2019-01-09. Table 5 shows the confusion matrix, testing Durbin against Atlas disconnections as ground truth. Our model can correctly detect outages with short lengths which can be as little as 5 minutes or less than 5 minutes. When routing changes there can be transients on the internet which can cause brief outages. We observe that we have great PPV (0.9769), recall (0.9453) and TNR (0.7341) for short-duration outages (5 minutes or more). Our measurements show that on that week, around 5% of total blocks that have 5 minute outages that were not seen in prior work. The duration of outages from 5 to 11 minutes, omitted from previous observations, increases total outage duration by 20%.

5.7 Are These Results Stable?

We next validate the consistency of our results over a week, examining accuracy with PPV, recall, and TNR for long-term observation to show the results are stable. We run Durbin on B-root data everyday for continuous monitoring.

5.7.1 Observation for a week. In Figure 9 we observe seven days of Durbin's accuracy parameters (PPV, recall and TNR) (We put the figure in appendix Appendix B). PPV is the same for all seven days (0.9999) because of the high true availability on each day. Recall and TNR are also generally stable, with recall ranging from 0.96 to 0.99 and TNR from 0.8 to 0.9. Both are lowest on 2019-01-11, because on that day we see many sparse blocks which gives more false outages.

This data suggests that our results are consistent over multiple days.

5.7.2 Continuous Monitoring. As Durbin matures, we are shifting to running it continuously. Thus far we have revised our implementation to run against B-root continuously. We currently run it once at the end of the day. It generates and caches training data for the current day, and then runs detections on the current day using cached training data from the prior two days. Thus far we have run Durbin-over-B-root against a full month of data, and we are currently bringing up 24x7 processing.

Since our IPv6 B-root data mirrors the structure of IPv4 network services, the validation of Durbin performance for IPv4 gives confidence that IPv6 accuracy will be similar.

6 RESULTS

Having established that Durbin works in §5, we next explore what it says about the Internet. Our results show the outage rate on IPv4 and IPv6 in §6.1 and our IPv6 coverage in §6.2.

6.1 How Many IPv6 Outages?

We evaluate IPv6 outage rate based on seven days of passive data from B-root. We established Durbin's accuracy in IPv4 (§5.1), and showed that IPv6 sources have a similar traffic rate as IPv6 (§4.1). Using this result, we next provide the first results for IPv6 outages.

6.1.1 Outage Duration. First, we show internet outage duration (seconds of outages for fraction of time for outages) of IPv4 and IPv6 address blocks and Trinocular's outage Figure 6 for seven days. The fraction of outage duration follows prior work and allows us to compare IPv4 outages against IPv6.

In §5.6 see /48 IPv6 blocks are out about 9% of the time.

To show that both Durbin and Trinocular have similar outages in IPv4, we compare Durbin's outage duration with Trinocular's outage duration for IPv4 blocks. By comparison, Durbin sees around 1% outage duration in /24 IPv4 blocks.

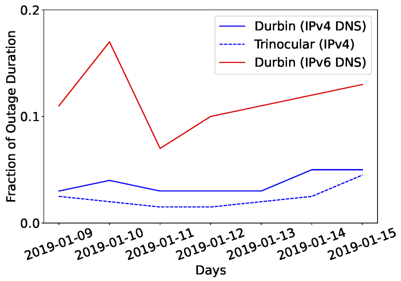


Figure 6: IPv4 and IPv6: outage fraction

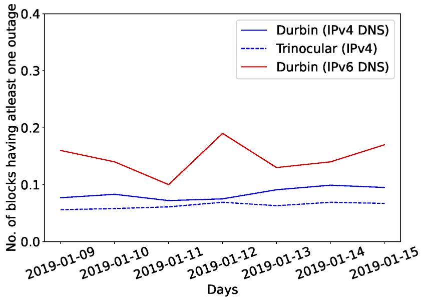


Figure 7: IPv4 and IPv6: blocks with at least one outage

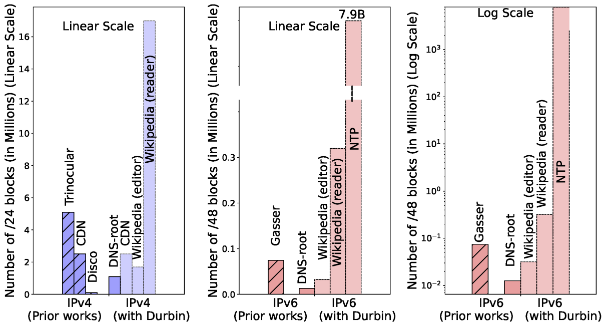


Figure 8: Current (dark shades) and potential (light) coverage in IPv4 and IPv6 for prior work (hashed) and Durbin (solid).

Durbin seems to have a bit higher outage duration than Trinocular, due to its detection of short outages (see §5.6).

We first examine how often Durbin infers networks are down.

We evaluate a week of data using Durbin-with-`B-root` in Figure 6. Here the Durbin-inferred fraction of outages are shown for IPv4 (blue, solid line) and for IPv6 (the red line). We compare to the publicly shared outage rate from Trinocular for IPv4 (the blue dashed line).

First, we see that IPv4 outage fractions are similar for both Durbin and Trinocular, both around 2% to 4%. Durbin's results are from 1M /24 IPv4 blocks, less coverage than Trinocular's 5M, but a similar fraction.

Durbin also provides *the first data for IPv6 outages* (the red line). We see IPv6 outages are much higher: from 0.6 to 1.5, rates 10× what we see in IPv4. This data is a report of 1338 outages in /48 IPv6 blocks. The absolute number of outages is significantly higher for IPv4 (1M) compared to IPv6 (30k) due to the larger number of measurable IPv4 blocks.

We expect that IPv6 outages will be more than IPv4, but evaluating why the difference is this large is ongoing work. Recent work that studied outage rates in RIPE Atlas [32], finding that IPv6 DNS query rates fail at a rate 3× what IPv4 sees (about 9% failures in v6 vs. 3% in v4). They point to a combination of end-system misconfiguration and long-term peering disputes as reasons for the difference. Our results show a larger difference, something we are looking into.

6.1.2 *Blocks With At Least One Outage.*

We next consider how many blocks see at least one outage each day.

We consider this metric because outage durations are heavily influenced by a few blocks that are unreachable for a long time. Blocks with long-term outages are unlikely to have active users, but users remember times their work was stopped by an Internet outage—something captured by this metric.

In Figure 7 shows the fraction of blocks with at least one 10-minute outage on each day. For IPv4, Durbin sees 3 to 4% of /24 IPv4 blocks have at least one outage on any given day (the solid blue line). Trinocular shows a similar fraction, with about 2% of blocks out at least once (the dashed blue line), a similar fraction as what Durbin sees. For IPv6, that fraction rises to 12% to 13% of all measurable /48 IPv6 blocks (the solid red line).

We can use this metric to compare IPv4 and IPv6: we see that the outages for IPv6 seems somewhat greater than for IPv4, suggesting IPv6 reliability can improve. The absolute number of outages is much larger for IPv4 (1M) than for IPv6 (30k) because there are many more measurable IPv4 blocks.

6.2 How Broad Is Our IPv6 Coverage?

We next evaluate Durbin's current coverage along with Durbin's *potential* coverage given access to other data sources, comparing both to the prior work.

Durbin's coverage depends on its input data—any data source that supports the passive observation of global sources is suitable input (§3.1). Potential coverage is maximized using data sources that see many source addresses. Most top-10 websites (Google, Facebook, Wikipeida, etc.) more than meet this requirement, as would many global CDNs (Akamai, Amazon Cloudfront, Cloudflare, etc.), and some global services (public DNS resolvers, NTP services, etc.). While we do not currently have access to this data, and it is unlikely a commercial service would share their data with researchers, we consider potential Durbin coverage to show how well the method *could* work, given proper input. (It is always possible that a top website or CDN would choose to implement Durbin for their own purposes.)

Published work shows Wikipedia sees 25M unique IP addresses [39], which suggests they likely see millions of /24 IPv4 prefixes. Analysis of "a major CDN" states they have 2.3M "trackable" /24s address blocks (where trackable means blocks with at least 40 active IP addresses, allowing outage detection by their method). NTP sees *billions* of IPv6 addresses [31]. Below we evaluate B-root, a CDN, Wikipedia, and NTP as potential data sources for Durbin.

6.2.1 Comparing Durbin Coverage to Prior Work: Figure 8 compares prior systems (left, hashed bars) against Durbin with B-root (the first darker blue bar in the right cluster). We show data for both IPv4 (the left graph) and IPv6 (the right graph), normalizing both to 100% as the best possible current result.

For IPv4 (the left graph with blue bars), we see that Durbin with B-root provides good coverage: about 1M /24s blocks. We find this coverage surprisingly good, given B-root only sees traffic from DNS servers, not end-users. Durbin's IPv4 coverage is about one-fifth of the 5.1M in Trinocular (the largest current outage detection system), half of CDN-based detection, and 10× more than Disco.

Durbin's *current* coverage is determined by B-root, and which /24 blocks report traffic that is frequent enough to evaluate (§3.2). We evaluate Durbin's IPv6 coverage based on one representative day (2019-01-10) of passive data from B-root, comparing results in IPv4 and IPv6.

We evaluate Durbin's IPv6 coverage based on one representative day of passive data from B-root, comparing results in IPv4 and IPv6. We show Durbin's coverage of both IPv4 and IPv6 address blocks and compare the coverage with prior works Trinocular, Akamai and Disco. We use Trinocular, Akamai, and Disco coverage as the prior work for IPv4 and Gasser hitlist for IPv6 coverage in Figure 8.

6.2.2 Durbin IPv4 Coverage with Potential Alternative Sources: Durbin's current coverage is limited by not seeing clients, but if Durbin were run with a major website's logs as input, its coverage can equal or exceed current systems. We next consider what Durbin coverage *would* be if it was applied to CDN, Wikipedia, or NTP traffic.

CDN: We estimate potential coverage with CDN data from published work [29]. However, their paper provides only IPv4 coverage.

Wikipedia: We use Wikipedia as an example top-10 website. Wikipedia does not provide public access to browsing traffic, but all Wikipedia edits are public, and about half are logged with IP addresses (as disclosed to the editor, so with their consent).

We downloaded the entire Wikipedia edit history and the logging history and extracted all "IP users". We count 9,264,603 unique IPv4 addresses and 94,042 IPv6 addresses, for 1,694,599 IPv4 /24 blocks and 30,257 IPv6 /48 blocks.

Of course, Wikipedia has *far* more readers than editors. One analysis observes that although the read: edit ratio is not known, the number of page views can provide an upper bound on the number of readers. All Wikipedia sees about 85 billion page views per month, and Hill suggested 35 page-views per reader [15], implying about 686M readers per month, or 75× more than editors. We suggest this offers a loose *upper bound* on the number of unique IPs that Wikipedia sees per month. We assume a more conservative 10× multiplier from editors, implying readers will show 1.7M IPv4 /24 blocks and 30.2k IPv6 /48 blocks.

Implications for Durbin: IPv4 coverage with the CDN will roughly match coverage with prior work [29], but Durbin will be able to report 5-minute temporal precision for frequent-traffic blocks. This analysis uses only blocks reported as measurable by their outage detection system. It is possible that Durbin could provide coarse-time results for blocks that are unmeasurable by their method, thereby increasing coverage.

These results suggest that the Durbin algorithm could provide at least as good coverage as CDN-based outage detection, when applied to a data source like a major website.

6.2.3 Actual and Potential IPv6 Coverage: Durbin coverage is even more promising when one considers IPv6. Here we compare the IPv6 hitlist as the best possible option (although we have not seen published work using IPv6 hitlists for outage detection). We cover slightly less than one-fifth of the Gasser hitlist, but use only passive data from B-root.

Our analysis of Wikipedia IPv6 edits suggests Durbin would see 30.2k /48 blocks, roughly double B-root, and half of the Gasser IPv6 hitlist. Projecting edits to readers with the same ratio as in IPv4, we expect around 300k /48 blocks.

Finally, while our analysis of Wikipedia is conservative, Rye and Leven took data from NTP, a global service touched by billions of IPv6 addresses. The right-most set of red bars in Figure 8 add NTP, but with a *log-scale y*-axis: with 7.9 billion IPv6 addresses, NTP exceeds all other sources.

7 CONCLUSION

We have describe Durbin, a system to detect Internet outages with a new adaptive algorithm using passive sources. The challenge to outage detection from passive data is balancing accuracy with spatial and temporal precision and coverage; Durbin provides good accuracy (0.811 TNR) at constant spatial precision (/24 IPv4 and /48 IPv6 blocks) by adapting temporal precision for each block (5 or 25 mintues). We evaluated Durbin with two different data sources: B-root and the Merit darknet, examples representing network services like DNS and darknets. Coverage of IPv4 with our data sources is good (about 1M /24 blocks with `B-root`). IPv4 coverage with `B-root` is large as current active methods, but a top website could use Durbin to see IPv4 coverage equal to or exceeding active methods. Finally, Durbin provides the first published data reporting IPv6 outages (30k /48 IPv6 blocks with `B-root`), showing the promise passive methods to track outages in IPv6.

REFERENCES

[1] Data-centers lose on about $5,000 per minute when users can not connect due to outages. In *https://www.businesswire.com/news/home/20110510006495/en/At-a-Cost-of-More-Than-5000-Per-Minute-Data-Center-Outages-Can-Be-Painful-Emerson-Study-Shows.*

[2] Amazon down 15 minutes, loses over $66,000 per minute. In *https://smallbiztrends.com/2013/08/amazon-down-custom-error-page.html*, 2013.

[3] Giuseppe Aceto, Alessio Botta, Pietro Marchetta, Valerio Persico, and Antonio Pescapé. A comprehensive survey on internet outages. *Journal of Network and Computer Applications*, 113:36–63, 2018.

[4] Guillermo Baltra and John Heidemann. Improving coverage of Internet outage detection in sparse blocks. In *Proceedings of the Passive and Active Measurement Conference*, Eugene, Oregon, USA, March 2020. Springer.

[5] Robert Beverly, Ramakrishnan Durairajan, David Plonka, and Justin P. Rohrer. In the IP of the beholder: Strategies for active IPv6 topology discovery. In *Proceedings of the ACM Internet Measurement Conference*, pages 308–321, Boston, Massachusetts, USA, October 2018. ACM.

[6] Robert Beverly, Matthew Luckie, Lorenza Mosley, and Kc Claffy. Measuring and characterizing ipv6 router availability. In *Passive and Active Measurement: 16th International Conference, PAM 2015, New York, NY, USA, March 19-20, 2015, Proceedings 16*, pages 123–135. Springer, 2015.

[7] Ryan Bogutz, Yuri Pradkin, and John Heidemann. Identifying important internet outages. In *2019 IEEE International Conference on Big Data (Big Data)*, pages 3002–3007. IEEE, 2019.

[8] Alberto Dainotti, Karyn Benson, Alistair King, KC Claffy, Michael Kallitsis, Eduard Glatz, and Xenofontas Dimitropoulos. Estimating internet address space usage through passive measurements. *ACM SIGCOMM Computer Communication Review*, 44(1):42–49, 2013.

[9] Alberto Dainotti, Claudio Squarcella, Emile Aben, Kimberly C Claffy, Marco Chiesa, Michele Russo, and Antonio Pescapé. Analysis of country-wide internet outages caused by censorship. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 1–18, 2011.

[10] Asma Enayet and John Heidemann. Internet outage detection using passive analysis (poster abstract). In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 772–773, 2022.

[11] Pawel Foremski, David Plonka, and Arthur Berger. Entropy/IP: Uncovering structure in IPv6 addresses. In *Proceedings of the ACM Internet Measurement Conference*, pages 167–181, Santa Monica, CA, USA, November 2016. ACM.

[12] Oliver Gasser, Quirin Scheitle, Pawel Foremski, Qasim Lone, Maciej Korczyński, Stephen D. Strowes, Luuk Hendriks, and Georg Carle. Clusters in the expanse: understanding and unbiasing ipv6 hitlists. In *Proceedings of the Internet Measurement Conference 2018*, pages 364–378, 2018.

[13] Oliver Gasser, Quirin Scheitle, Sebastian Gebhard, and Georg Carle. Scanning the IPv6 internet: Towards a comprehensive hitlist. In *Proceedings of the*, Louvain La Neuve, Belgium, April 2016.

[14] Andreas Guillot, Romain Fontugne, Philipp Winter, Pascal Merindol, Alistair King, Alberto Dainotti, and Cristel Pelsser. Chocolatine: Outage detection for internet background radiation. In *2019 Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–8. IEEE, 2019.

[15] Benjamin Mako Hill. Editor-to-reader ratios on wikipedia. blog post https://mako.cc/copyrighteous/editor-to-reader-ratios-on-wikipedia, February 2011.

[16] Thomas Holterbach, Edgar Costa Molero, Maria Apostolaki, Alberto Dainotti, Stefano Vissicchio, and Laurent Vanbever. Blink: Fast connectivity recovery entirely in the data plane. In *16th USENIX Symposium on NSDI 19*, pages 161–176, 2019.

[17] Ethan Katz-Bassett, Harsha V Madhyastha, John P John, Arvind Krishnamurthy, David Wetherall, and Thomas E Anderson. Studying black holes in the internet with hubble. In *NSDI*, volume 8, pages 247–262, 2008.

[18] Ethan Katz-Bassett, Colin Scott, David R Choffnes, Ítalo Cunha, Vytautas Valancius, Nick Feamster, Harsha V Madhyastha, Thomas Anderson, and Arvind Krishnamurthy. Lifeguard: Practical repair of persistent route failures. *ACM SIGCOMM Computer Communication Review*, 42(4):395–406, 2012.

[19] Ege Cem Kirci, Martin Vahlensieck, and Laurent Vanbever. "is my internet down?" sifting through user-affecting outages with google trends. In *Proceedings of the 22nd ACM Internet Measurement Conference*, pages 290–297, 2022.

[20] MERIT. About the Orion network telescope. web page https://www.merit.edu/initiatives/orion-network-telescope/, 2021.

[21] David Moore, Colleen Shannon, Geoffrey M. Voelker, and Stefan Savage. Network telescopes: Technical report. Technical Report TR-2004-04, UCSD CAIDA, July 2004.

[22] Austin Murdock, Frank Li, Paul Bramsen, Zakir Durumeric, and Vern Paxson. Target generation for internet-wide IPv6 scanning. In *Proceedings of the ACM Internet Measurement Conference*, pages 242–253, San Diego, CA, USA, October 2017. ACM.

[23] Ramakrishna Padmanabhan. *Analyzing internet reliability remotely with probing-based techniques.* PhD thesis, University of Maryland, College Park, 2018.

[24] Ramakrishna Padmanabhan, Aaron Schulman, Alberto Dainotti, Dave Levin, and Neil Spring. How to find correlated internet failures. In *Passive and Active Measurement: 20th International Conference, PAM 2019, Puerto Varas, Chile, March 27–29, 2019, Proceedings 20*, pages 210–227. Springer, 2019.

[25] Ruoming Pang, Vinod Yegneswaran, Paul Barford, Vern Paxson, and Larry Peterson. Characteristics of internet background radiation. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement*, pages 27–40, 2004.

[26] Lin Quan, John Heidemann, and Yuri Pradkin. Detecting internet outages with precise active probing (extended). *USC/Information Sciences Institute, Tech. Rep*, 2012.

[27] Lin Quan, John Heidemann, and Yuri Pradkin. Trinocular: Understanding internet reliability through adaptive probing. *ACM SIGCOMM Computer Communication Review*, 43(4):255–266, 2013.

[28] Lin Quan, John Heidemann, and Yuri Pradkin. When the internet sleeps: Correlating diurnal networks with external factors. In *Proceedings of the 2014 Conference on Internet Measurement Conference*, pages 87–100, 2014.

[29] Philipp Richter, Ramakrishna Padmanabhan, Neil Spring, Arthur Berger, and David Clark. Advancing the art of internet edge outage detection. In *Proceedings of the Internet Measurement Conference 2018*, pages 350–363, 2018.

[30] Root Operators. http://www.root-servers.org, April 2016.

[31] Erik Rye and Dave Levin. IPv6 hitlists at scale: Be careful what you wish for. In *Proceedings of the ACM SIGCOMM Conference*, pages 904–916, New York, NY, USA, September 2023. ACM.

[32] Tarang Saluja, John Heidemann, and Yuri Pradkin. Differences in monitoring the DNS root over IPv4 and IPv6. pages 194–203, December 2022.

[33] Aaron Schulman and Neil Spring. Pingin'in the rain. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, pages 19–28, 2011.

[34] Anant Shah, Romain Fontugne, Emile Aben, Cristel Pelsser, and Randy Bush. Disco: Fast, good, and cheap outage detection. In *2017 Network Traffic Measurement and Analysis Conference (TMA)*, pages 1–9. IEEE, 2017.

[35] Guanglei Song, Jiahai Yang, Lin He, Zhiliang Wang, Guo Li, Chenxin Duan, Yaozhong Liu, and Zhongxiang Sun. AddrMiner: A comprehensive global active IPv6 address discovery system. In *2022 USENIX Annual Technical Conference (USENIX ATC 22)*, pages 309–326, 2022.

[36] Guanglei Song, Jiahai Yang, Zhiliang Wang, Lin He, Jinlei Lin, Long Pan, Chenxin Duan, and Xiaowen Quan. Det: Enabling efficient probing of ipv6 active addresses. *IEEE/ACM Transactions on Networking*, 30(4):1629–1643, 2022.

[37] Sejun Song and Jim Huang. Internet router outage measurement: An embedded approach. In *2004 IEEE/IFIP Network Operations and Management Symposium (IEEE Cat. No. 04CH37507)*, volume 1, pages 161–174. IEEE, 2004.

[38] RIPE NCC Staff. Ripe atlas: A global internet measurement network. *Internet Protocol Journal*, 18(3), 2015.

[39] Khoi-Nguyen Tran and Peter Christen. Cross language prediction of vandalism on wikipedia using article views and revisions. In *Advances in Knowledge Discovery and Data Mining: 17th Pacific-Asia Conference, PAKDD 2013, Gold Coast, Australia, April 14-17, 2013, Proceedings, Part II 17*, pages 268–279. Springer, 2013.

[40] USC. B-Root DNS service. https://b.root-servers.org/.

[41] USC. Passive outage data. https://www.ant.edu/datasets/outage.

APPENDIX A RESEARCH ETHICS

Our work poses no ethical concerns. In evaluating the risks of or work relative to its benefits, it poses minimal risks, while there are significant benefits to a new method to detect Internet outages and thereby improve Internet service.

We believe our work poses minimal risk because our approach analyzes passive traffic to look for activity on networks. The primary risk is that such traffic analysis may reveal personal information about individuals. To avoid revealing such information, our data provider provides only an IP address and timestamp of activity, not the actual user activity. (For example, in DNS data, the query and reply are removed.) We discuss data handling in §3.1: while we require tracking specific sources, we do not need to know actual IP addresses, just correct address blocks. Our data provider therefore anonymizes the least-significant bits of IP addresses.

Finally, our use of `B-root` poses minimal privacy risk because nearly all DNS queries are from infrastructure (recursive resolvers), not directly from individuals, any requests from individuals are mixed in with queries from infrastructure.

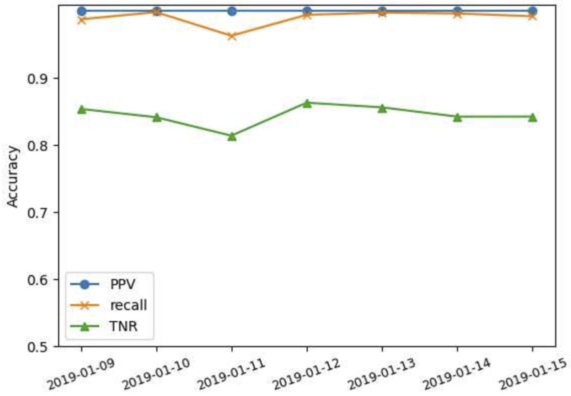


Figure 9: PPV, recall and TNR for seven days for Durbin-with-`B-root`

We have submitted an IRB review request proposing an analysis of new data sources (beyond `B-root`) with the above anonymization as non-human-subjects research. This IRB is currently under review.

APPENDIX B ARE THESE RESULTS STABLE?

In Figure 9, we observe seven days of Durbin's accuracy parameters (PPV, recall, and TNR). PPV is consistent for all seven days (0.9999) because of the high true availability on each day. Recall and TNR are generally stable, with recall ranging from 0.96 to 0.99 and TNR from 0.8 to 0.9. Both are lowest on 2019-01-11 due to many sparse blocks causing more false outages.