# Chocolatine: Outage Detection for Internet Background Radiation

Andreas Guillot*, Romain Fontugne†, Philipp Winter‡, Pascal Merindol*
Alistair King‡, Alberto Dainotti‡, Cristel Pelsser*
∗ICube, University of Strasbourg – France    †IIJ Research Lab – Japan    ‡CAIDA, UC San Diego – USA

*Abstract*—The Internet is a complex ecosystem composed of thousands of Autonomous Systems (ASs) operated by independent organizations; each AS having a very limited view outside its own network. These complexities and limitations impede network operators to finely pinpoint the causes of service degradation or disruption when the problem lies outside of their network. In this paper, we present `Chocolatine`, a solution to detect remote connectivity loss using Internet Background Radiation (IBR) through a simple and efficient method. IBR is unidirectional unsolicited Internet traffic, which is easily observed by monitoring unused address space. IBR features two remarkable properties: it is originated worldwide, across diverse ASs, and it is incessant. We show that the number of IP addresses observed from an AS or a geographical area follows a periodic pattern. Then, using Seasonal ARIMA to statistically model IBR data, we predict the number of IPs for the next time window. Significant deviations from these predictions indicate an outage. We evaluated `Chocolatine` using data from the UCSD Network Telescope, operated by CAIDA, with a set of documented outages. Our experiments show that the proposed methodology achieves a good trade-off between true-positive rate ($90\%$) and false-positive rate ($2\%$) and largely outperforms CAIDA's own IBR-based detection method. Furthermore, performing a comparison against other methods, i.e., with BGP monitoring and active probing, we observe that `Chocolatine` shares a large common set of outages with them in addition to many specific outages that would otherwise go undetected.

*Index Terms*—Outage detection, Internet Background Radiation, ARIMA

## I. Introduction

Connectivity disruptions caused by physical outages, software bugs, misconfiguration, censorship, or malicious activity, occur repeatedly on the Internet [1]. Monitoring the state of Internet connectivity is useful to raise public awareness on events of intentional disconnection due to censorship [14]. It further helps operators pinpoint the location of an outage, *i.e.,* the place where there is a loss of connectivity, when it happens outside their reach. This enables to speed up recovery as the correct network operator team can be contacted directly instead of reaching out to the global network operator's community via mailing lists or personal contacts. Fast outage detection is also useful to locally switch to backup routes, when available [16].

A few methods exist to detect connectivity outages. Monitoring for withdrawals of BGP prefixes is a commonly used approach, but it can only observe outages that affect the control plane [10], [2]. Data-plane approaches solve this problem, and can be either based on active measurements – *e.g.,*

Trinocular [21] sends pings to 4 M remote /24 address blocks to measure their liveness – or on passive traffic analysis – Disco [27] relies on the long-running TCP connections between RIPE Atlas probes and their controlling infrastructure to identify bursts of disconnections.

Another data-plane approach for the detection of connectivity outages, is based on the analysis of Internet Background Radiation (IBR) [5]. IBR is unsolicited traffic captured by darknets (also known as network telescopes), which announce unused IP prefixes on BGP, *i.e.,* there are no actual services running in the prefix nor eyeballs. IBR is composed of a constantly evolving mix of various phenomena: network scans, the results of malware infections, DoS attacks using spoofed IPs from the range announced by the telescope [4], packets from misconfigured (or with a polluted DHT) BitTorrent clients, etc. [31]. By leveraging the pervasiveness of IBR sources, and the consistent presence of traffic, we can infer a connectivity outage for a given geographic area or Autonomous System (AS) based on a significant reduction of IBR traffic that originates from them. In addition, Dainotti et al. [8], [5] demonstrated that IBR can effectively complement both control-plane and active probing data-plane approaches: both in terms of coverage (not all networks respond to pings) and in terms of information that it provides (*e.g.,* confirming outbound connectivity for a remote network even when inbound connectivity is disrupted).

The IODA system from CAIDA [17] has recently operationalized this method for extracting time series, i.e. "signals", at different spatial grain (e.g. countries or ASs). However, IODA's current automated detection algorithm is simplistic (a threshold based on the last 7 days moving median) and unable to take into account the IBR's noise and the intensity variability of the signal. Indeed, in order to avoid an overwhelming amount of false positives, the threshold is currently set to raise an outage alert when the signal intensity drops under 25% of the intensity of the median value observed in the last 7 days. That is, an outage is detected only when there is a severe connectivity loss, leaving many connectivity losses undetected [18]. In particular, the test remains the same whatever the period of the day and the week, such that a drop occurring in an usually busy period is treated the same as if it was occurring during an inactive one. In one word, this naive model is static, and as such challenging to calibrate, as it does not take into account any trends in the traffic.

In this work, we take these trends into account by applying

Seasonal ARIMA (SARIMA) [9], a popular technique that forecasts the behavior of the time series extracted at the UCSD Network Telescope [29]. More specifically, we analyze the number of unique source IP addresses that try to reach the darknet of different countries/ASs. `Chocolatine` is sensitive and robust, respectively to the seasonality and noise observed in the data. We show that it is able to detect outages with a true positive rate of $90\%$ and a false positive rate of $2\%$ with a detection delay of only 5 minutes. Additionally, the comparison with CAIDA's method showed that `Chocolatine` can detect a large share of outages seen by other data sources, as well as additional specific outages. Another benefit of `Chocolatine` is that its algorithm automatically self-tunes on time series exhibiting very different magnitudes and levels of noise (*e.g.,* time series of IBR extracted for ASs and countries of different sizes and with different compositions of IBR-generating sources). As a result, `Chocolatine` can be applicable to other seasonal and noisy data sources related to Internet traffic activity.

The remainder of the paper is structured as follows: some background on main outage detection methods is first provided in Section II. We then introduce the dataset we use in Section III, and explain why it is suited for outage detection. We describe how `Chocolatine` is designed at a high level in Section IV. We also illustrate our outage detection process with a case study of the censorship that occurred during the Egyptian revolution in Section V. We evaluate `Chocolatine` in Section VI, validating it with ground truth data and also comparing its performances against several current outage detection algorithms. Lastly, we address the reproducibility of our experiments in Section VII.

## II. BACKGROUND

Outage detection can be achieved with different measurement techniques analyze different performance indicators. A recent survey [1] provides a taxonomy of most existing techniques, including three main categories: *active* monitoring, *passive* monitoring, and *hybrid* monitoring, which is a combination of both active and passive. We reuse this terminology here.

Active monitoring techniques generate traffic in order to collect information and examine the state of networks. Most active monitoring approaches are based on variants of *ping* and *traceroute*, and rely on a set of *vantage points* (*i.e.,* the devices that perform the measurements) that are usually distributed across different networks. For example, RIPE Atlas [24] is a popular platform for network measurement that is composed of over 10 000 probes. In [12], Fontugne *et al.* detect significant link delay changes and rerouting from the RIPE Atlas built-in measurements. Dasu [28], on the other hand, is more versatile than RIPE Atlas. It has been used for diverse measurements, such as broadband performance measurements, as well as the mapping of the Google CDN infrastructure. Thunderping [26] measures the connectivity of residential Internet hosts before, during, and after forecast periods of severe weather.

Passive monitoring techniques collect existing traffic and infer the state of networks from it. Generally speaking, they analyze real-user traffic to be close to the user experience. It ensures that the inferred statistics correspond to real traffic, thus granting a view of a network's current state. Different datasets have been leveraged for passive analysis, such as CDN traces [23], or darknets [3].

Outage detection methods also rely on different theoretical modeling techniques to discriminate outages from normal network conditions. Trinocular [21] leverages Bayesian inference to estimate the reachability of /24 subnetworks. Disco [27] detects surge of Atlas probe disconnections using a burst modeling algorithm. Using also Atlas data, authors of [12] rely on the central limit theorem to model usual Internet delays and identify network disruptions.

In the present work, we rely on passive measurements collected from CAIDA's network telescope [29] and employ SARIMA models to forecast IBR time series and detect outages.

## III. DATASET

The data used for this study is obtained from the UCSD network telescope [29]. The goal of this section is to provide an overview of the characteristics of this dataset, and to motivate why it is suitable for outage detection.

The collected data consists exclusively of unsolicited traffic caused by both benign and malicious activities. For instance, software and hardware errors, such as bit-flipping or hardcoded IP addresses, result in IBR traffic. Network scans and backscatter traffic are another common source of IBR traffic. Backscatter traffic is usually the consequence of malicious spoofed traffic sent to a victim and whose replies are returned to unused addresses monitored by the network telescope. Consequently, IBR data has been extensively used to study worms [30], virus propagation [15], and Distributed Denial of Service (DDoS) attacks [11].

CAIDA's IODA [17] aggregates UCSD network telescope data geographically and topologically, respectively using NetAcuity [19] IP geolocation datasets and longest prefix matching against BGP announcements from public BGP data [20]. Consequently, we obtain IBR streams per country, regional area (*e.g.,* states in the US, provinces in France, etc.), and AS. IODA also pre-filters the traffic that reaches the telescope, removing large components of potentially spoofed-source-IP traffic (since their presence would significantly alter inference about originating ASs and geographical areas) using a set of heuristics derived semi-manually [7].

Traffic from these streams can be summarized in different ways, the most common being the *number of bytes*, the *number of packets*, and the *number of unique source IP addresses*. The number of unique source IP addresses [8] is defined as the number of IP addresses originating from the same location that contact the network telescope during a given time interval. It is an adequate metric to study Internet outages because it counts the number of devices that send traffic at a geographical or topological location, while abstracting the need to analyze
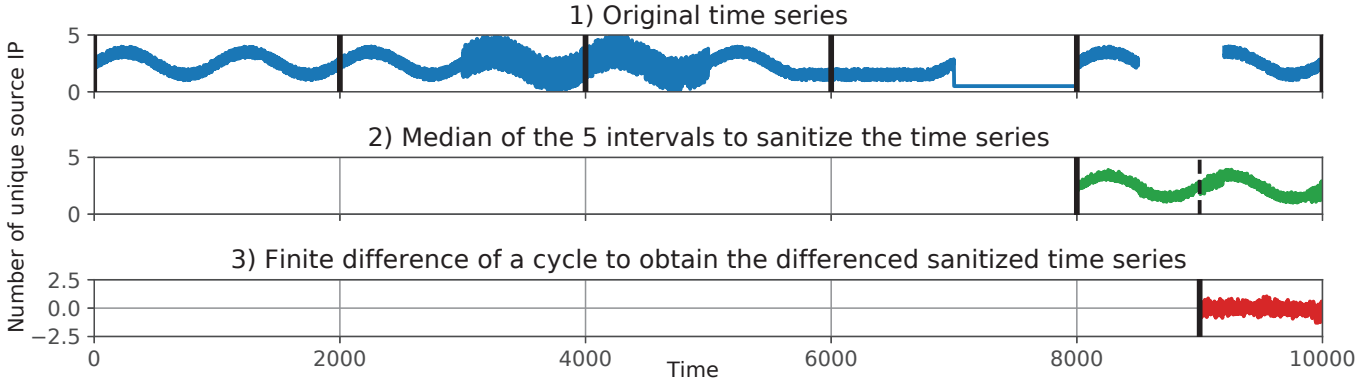
Fig. 1: Illustration of preprocessing and seasonal integration of the training data

traffic. In the event of an outage, some of these devices get disconnected from the Internet, so we expect to observe drops in the number of unique source IP addresses monitored with the network telescope.

The usage of IBR to detect outages is particularly pertinent since it is pervasive. Indeed, the amount of IBR packets that reaches network telescopes is considerable, incessant, and originates from a variety of applications [31]. In [4], Benson et al. performed a spatial analysis and determined that IBR provided an Internet-wide view. All countries, except for 3 with a population of less than 4000 inhabitants, and more than half of all ASs are observed in their dataset. Note that half of the ASs that do not show up in the dataset are small, as they only advertise a /24 prefix, while 86% of ASs that advertise a /16 or more are visible. A fifth of the remaining 14% that do not generate IBR traffic are unused blocks that belong to the US government.

The temporal analysis done in [4] shows that most networks frequently generate IBR traffic, in particular when considering coarse grain aggregations. Indeed, the median time between observations is shorter than 1 minute for over 90% of countries, and is shorter than 10 minutes for about 75% of the ASs.

To summarize, IBR traffic is ubiquitous, and thus can be used to detect and analyze large-scale network events. It is continually sent by a variety of sources all around the world, which makes it a suitable source to make opportunistic world-wide Internet measurements and specifically for efficiently detecting outages.

## IV. METHODOLOGY

In this section, we describe how Chocolatine forecasts the number of unique IP addresses in IBR traffic and detects outages. Among the numerous approaches available to forecast time series, Autoregressive Integrated Moving Average (ARIMA) models are a popular choice thanks to their simplicity and efficiency [32]. For this study we select Seasonal-ARIMA (SARIMA) [9] models in order to deal with weekly patterns observed in IBR time series. We propose an outage detection method composed of four main steps. First,

we sanitize the training part of the dataset (Section IV-A) and we eliminate non-stationarity in the data by differencing the data with a lag of one week (Section IV-B). Second, we compare results with multiple sets of parameters to find the best parameters for modeling each time series, and compute the corresponding prediction intervals (Section IV-C). Finally, we detect and report outages based on the differences between the computed predictions and the actual data (Section IV-D).

### A. Data preparation

In the following, the IBR time series are split into three sets: *training*, *calibration*, and *test*. These are used differently for the modeling (Section IV-C) and detection phases (Section IV-D). The training and calibration sets are used for the modeling, *i.e.,* to learn the best set of parameters for the ARMA model. These parameters are then used on the test set to detect potential outages.

The training data is used as the basis of the predictive model, and we need to sanitize it. There are three problems that need to be addressed:

- Missing values that we need to fill to have a working model,
- Extreme values, which will bias the model by greatly influencing the statistical properties of the time series,
- The presence of an outage inside the training data, which will lead to a model considering outages as the norm.

To overcome these problems we assume that the occurrence of missing and extreme values are uncommon so we can synthesize ten weeks of data into two weeks of sanitized data. Our data preparation process is illustrated in Figure 1 with a time series having the three problems mentioned above. We consider five intervals of two weeks (top plot in Figure 1), and compute the median values across all five intervals to obtain two weeks of clean synthesized data (middle plot in Figure 1). This sanitized time series is then used as the training set for our SARIMA model.

### B. SI: Seasonal Integration

ARMA models assume that the data is *stationary*, that is, the statistical properties of the data (*e.g.,* mean and variance)
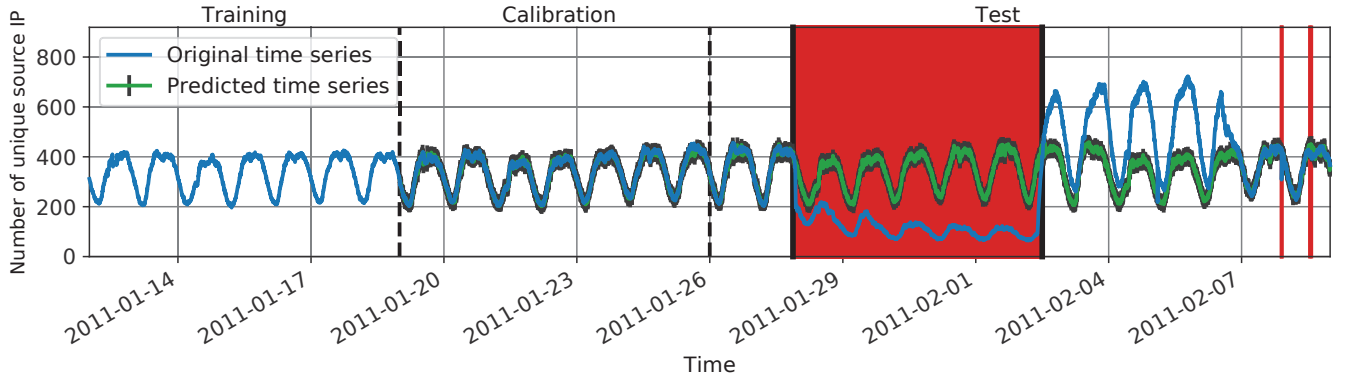
Fig. 2: Analyzing the test set with the best model ($AR = 4, MA = 1$)

are constant over time. Because of the strong daily and weekly patterns present in IBR data, our time series are non-stationary (*e.g.,* there is less traffic at night and during weekends, because more devices are turned off or disconnected during these periods of time [22]). This is the reason why a simple predictive model would not work with IBR time series. As a result, we make our time series stationary by filtering these trends with seasonal differencing. In practice, our time series contain a weekly and daily trend which we both remove by applying a seasonal differencing (SI part of SARIMA) of a week (*e.g.,* bottom plot in Figure 1).

The computed training data, which is now sanitized and stationary, can then be used in the following step to create a predictive model and to make predictions on the calibration data.

### C. ARMA: Autoregressive Moving Average

In this step, we estimate the best parameters for any given time series. In pratice, `Chocolatine` will compute a different set of parameters for each analyzed time series, which will increase the adaptability of the solution and the quality of the predictions. To achieve this goal, we have to precisely estimate the values for the two key parameters of ARMA, that is the order of the autoregressive model (named $p$), and the order of the moving-average model (named $q$). We use the sanitized training data for that purpose, as ARMA models only work on the condition that the training data is anomaly free and stationary. In order to find the best combination of parameters for any given time series, we make predictions on a second set of data that we refer to as the calibration data. In practice, we use the period following the training data for defining such a calibration. We consider several predictive models, each with their own set of $(p, q)$ parameters, to evaluate the performances of various distinct predictions. We finally compare the accuracy of these predictive models on the data used for this calibration. The *Root Mean Square Error* (RMSE) is used to compute the error between the real time series and the one obtained from the predictive model. We chose the RMSE to penalize predictive models that made predictions that are significantly far from the actual data. The

predictive model (*i.e.,* the set of $(p, q)$ parameters) with the lowest error will thus be used for future predictions.

Now that we have the best parameters to use within the ARMA model, we also compute a prediction interval. It defines the boundaries of our predictions and it is used for the outage detection process. We compute $99.5\%$ prediction intervals using the residual variance. The residual variance is computed using the *Median Absolute Deviation* (MAD), a robust measure of data variability used for anomaly detection [12] (the RMSE being not suitable enough in this case). This is essential, as the prediction intervals should be both robust to false positives but still able to capture extreme values introduced by measurement errors and outages.

The model, and its associated prediction interval, are used to detect outages, as described in the next section.

### D. Detection

The steps described above provide us with stationary data and an optimized predictive model for each time series.

The next step is to detect outages with the predictive models. We define an outage as a point in time where a value of a time series is smaller than the lower bound of the prediction interval. The severity of this alarm will be determined by computing the following distance:

$$d = (\hat{X} - X)/(\hat{X} - L),$$

where $\hat{X}$ is the predicted value, $X$ is the actual value from the time series, and $L$ is the lower bound of the prediction interval. Distances $d > 1$ and $d < -1$ mean that the time series is outside of the prediction interval, whereas the time series is within the prediction interval when $-1 \leq d \geq 1$. The only cases that are reported as outages are cases where $d > 1$, that is, when the actual values are outside of the prediction interval and are smaller than the lower bound of the prediction interval, which translates in a significant drop in the number of IPs observed in the time series. Cases where $d < -1$ (*i.e.,* points that are greater than the upper bound of the prediction interval) are considered as extreme values, but they do not fall into our definition of an outage, and are thus not reported.

4

Every hour (*i.e.,* 12 data points) we make predictions for the next hour and compare the actual data to these predictions as explained above. Each time we move forward in the data, ARMA takes into account the new data points for the future predictions. However, we take particular precautions to maintain the quality of the predictive model. Data identified as part of an outage should not be used for future predictions, which brings us back to the problems discussed in Section IV-A, where missing values, extreme values, and outages would diminish the quality of the predictive model. In this phase, we solve these problems differently, by doing what we refer to as *inpainting*: if a new sample of data is considered to be part an extreme value (*i.e.,* $d < -1$ or $d > 1$), we feed the predictive model with the predicted value instead of the real value.

## V. Case study

To illustrate the functioning of the proposed method and some of its benefits, this section provides thorough results for a case study.

On January 25th 2011, the Mubarak regime ordered network operators to shut down Internet connectivity during the Egyptian revolution in an attempt to silence the opposition. The chronology of this event has been described in [8]. The authors used BGP routing data, ping, traceroute, and IBR data. The IBR data was manually analyzed to shed light on the massive packet-filtering mechanisms that were put in place, and to identify denial-of-service attacks related to the political events happening in Egypt during the same period. In this section, we present how our solution analyzes the same IBR data but allows us to systematically detect the beginning and the end of the connectivity loss, and to estimate the severity of the outage.

Figure 2 shows the time series of unique source IP addresses from Egypt reaching the UCSD Network Telescope (plotted in blue). The disconnections occurred between the 28th of January and the 3rd of February, 2011, as it can be seen by the loss of intensity of the signal depicted in the figure. Here, we chose to include in our analysis also the values of the time series after the outages, because of an interesting phenomenon that was occurring: the values of the time series are higher than usual during the days that follow the Egyptian revolution and go back to normal around the 7th of February. In [6], the authors revealed that a botnet covertly (and massively) scanned the Internet during those days.

This time series is analyzed as follows. The training set, to the left, is sanitized following the methods discussed in IV-A. Multiple sets of ARMA parameters are then going to be used to predict the calibration set. The predictions are plotted with a green line. The set of parameters that resulted in the lowest error ($p = 4, q = 1$ in this case) will be used for the rest of the analysis. The difference between the predicted time series and the original time series allowed us to compute prediction intervals using the MAD. These intervals are plotted with gray bars that surround the predictions.

Then the test set is compared to the ARMA model and the prediction intervals computed in the previous step. The sudden drop that occurs when the outage starts puts the time series below the prediction intervals, which means that an outage is reported. Visually, this is shown with a red vertical line. Additionally, it also means that the *inpainting* process described in Section IV-D will take place, which is clear here, since the trend of the predicted time series stays similar to that of the original time series, even if an outage is occurring at the same time. No alarm is reported during the botnet attack that follows the outage because the original time series is higher than our prediction intervals, which means that the data is again inpainted and it will not count as an anomaly.

## VI. Validation, Calibration and Comparison

We evaluate the limits, and performance of `Chocolatine` through a validation and a comparison. We start by considering a set of verified outages from our ground-truth dataset, which we use to assess the accuracy of our outage detector, and look for the best threshold, e.g., the one determining the minimal number of IPs required to make accurate predictions. We then use a different set of outages in order to compare `Chocolatine` against CAIDA's outage detection techniques (using BGP dumps, active probing and the network telescope data).

### A. Validation

In this section, we evaluate the reliability of our technique using a reference dataset and gathering 130 time series containing outages. These time series contain three different types of spatial aggregates — ASs, countries, and regions within countries — from various years (2009 to 2018). The duration of these outages spans from an hour to a week. The comprehensive list of time series that compose this dataset is given in Table II. As an example, the RIPE NCC and Duke University BGP experiment [25] caused several outages in different ASs worldwide by triggering a bug in some Cisco routers.

We evaluate `Chocolatine` by computing the True Positive Rate (TPR) and the False Positive Rate (FPR), and show our calibration results with a ROC curve. Our purpose is twofold: we look into the accuracy of our approach, and we search for its best parameters by exploring its calibration spectrum. In particular, we determine which confidence level should be used to assess whether an outage is occurring or

TABLE I: Number of time series per IP threshold and per spatial scale

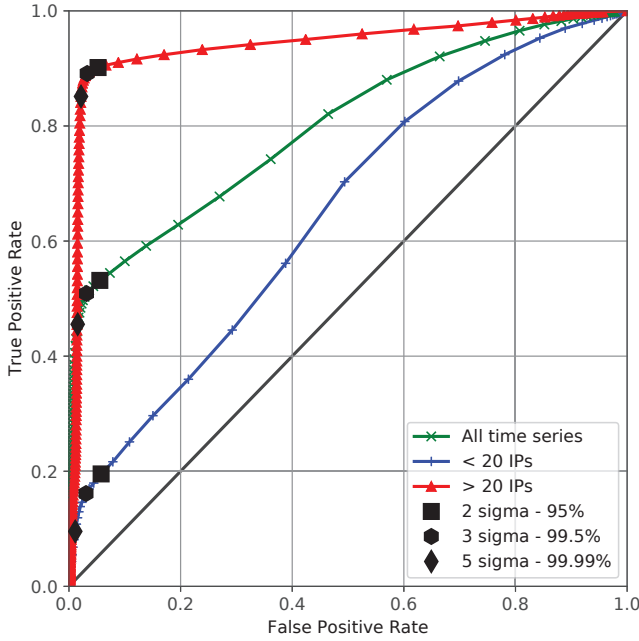|  | > 10 | > 15 | > 20 | > 25 | Total |
|---|---|---|---|---|---|
| Countries | 144 (56.9%) | 135 (53.3%) | 128 (50.5%) | 120 (47.3%) | 253 |
| Regions | 1,038 (21.4%) | 879 (18.1%) | 778 (16.0%) | 704 (14.5%) | 4,846 |
| ASs | 1,157 (1.8%) | 867 (1.4%) | 719 (1.1%) | 621 (1.0%) | 61,639 |

Fig. 3: ROC curve with a 5 minutes time grain and a threshold of 20

not. Our aim is to find the best trade-off between the TPR and the FPR by considering our collection of documented outages as the ground truth.

Moreover, to quantify the ability of our method to maximize the TPR while keeping the FPR low, we need to set two evaluation parameters used in our ROC analysis. On the one hand, we need to find out the minimal intensity required in the time series for our method to finely operate, and on the other hand, the smallest time granularity at which we can accurately detect outages. The intensity of time series is measured as the median number of observed IPs in a week. Trying multiple thresholds showed us that Chocolatine yielded better results with a threshold of 20 IPs, and that increasing this number had little effect on the accuracy. The results are presented in Figure 3, where three different ROC curves are plotted:

- The green curve plots the accuracy for all time series;
- The red curve plots the accuracy for time series with a median of IP addresses in a week that is greater than 20;
- The blue curve plots the accuracy for time series with a median of IP addresses in a week that is smaller than 20.

On the one hand, using a low number of IP addresses provides performance only slightly better than using a random model, which is expected, as the central limit theorem does not hold for samples that are too small. On the other hand, the higher the number of IPs is, the better the performance. (the red curve yields much better results than the blue one). The accuracy of our method for all time series (the green curve), is not satisfactory because of the influence of the time series contained in the blue curve. As a result, we have chosen to limit our analysis to the time series that had a median of more

than 20 IPs per week.

Table I summarizes the impact of this threshold on the number of remaining time series. Setting this threshold to 20 limits the number of time series that we can analyze to 1625, but it significantly increases the accuracy of our detector. Here, we make the assumption that network operators will want to have a low FPR, even if it means missing smaller outages. We also found that the size of the time bins we use can be relatively small (around 5 minutes) without impacting the performance much. This analysis is not included due to space constraints.

To conclude this section, we recommend to use a threshold of 20 IPs for the time series and 5 minutes long time-windows as in Figure 3. These two parameters can of course be tuned according to the data collection's specificity. Using such a threshold and time granularity (we can estimate outage durations at a 5 min granularity), the best confidence level for the prediction intervals is $99,5\%$ $(3\sigma)$. With these settings we obtain an acceptable true positive rate of $90\%$ while keeping the false positive rate under $2\%$.
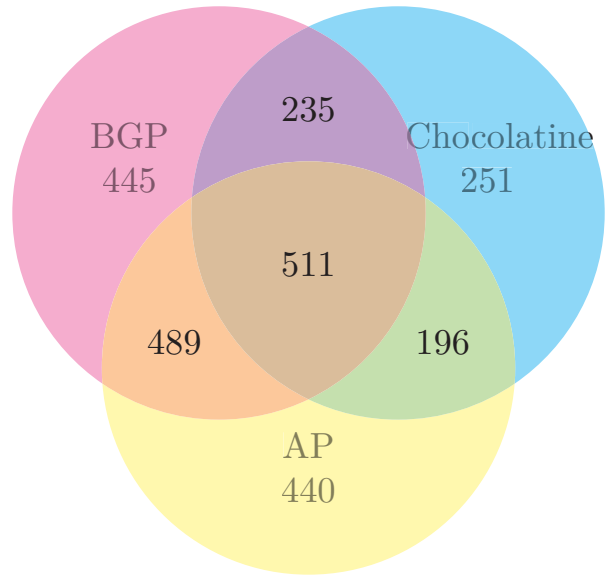
### B. Comparison

In this section, we compare the performance of our detector to three other techniques hosted in IODA: CAIDA's darknet detector (DN), CAIDA's BGP detector (BGP), and a technique based on active probing (AP), Trinocular [21]. A description of the integration of these 3 detectors in IODA can be found in [18]. In order to compare the detectors, we use a second ground-truth sample to emphasize the versatility of Chocolatine on different time series. Its set of outages is distinct from the previous one, but still decomposed in 5 minutes bins (see Table III). We ran the 4 detectors and enumerated the number of 5 minutes time bins where an outage is detected, for each detector. Fig. 4 a plots the number of outages detected by IODA's components, and Fig. 4 b plots how Chocolatine compares against BGP and Active Probing (AP) detectors. Note that the number of events given below the name of each detector are events detected only with that technique. The intersections depict the number of events detected by multiple detectors. For example, there are 1680 BGP events (the sum of each intersection combination in the magenta based set), 985 of which are also detected by the active probing technique.

Comparing Chocolatine with the IODA's darknet detector, one can observe that Chocolatine detects two order of magnitude more outages, 1193 compared to 71. This result highlights the much higher sensitivity of our approach, while CAIDA's darknet detector is extremely conservative by nature. By modeling weekly, and *a fortiori* daily, patterns our predictions are adaptively following the time series oscillations, while this is not the case in CAIDA's detector, which uses a global threshold approach.

Another way to evaluate Chocolatine is to cross-reference the set of alarms it is able to detect compared to the other detectors (and look at all the intersections). When there are intersections, the corresponding events are very likely
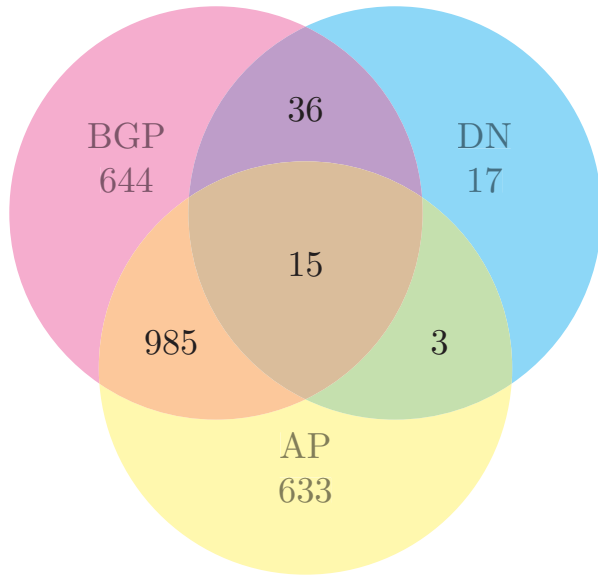
(a) BGP IODA versus AP IODA versus DN IODA      (b) `Chocolatine` versus BGP IODA versus AP IODA

Fig. 4: Comparison of the number of 5 minutes time bins identified as outage per detector

to be actual outages, *i.e.,* they are true positives. Fig. 4 b shows that the outages detected by `Chocolatine` are likely to intersect the outages of the other sources. Indeed, there are only 251 alarms that are specific to `Chocolatine`. The analysis of these alarms shows us that 59% of them occur in a range of 1 hour around alarms detected by other data sources. Generally speaking, these results suggest that our tool is complementary to the two others (BGP and AP) and clearly outperforms IODA's current darknet detector.

## VII. REPRODUCIBILITY

All results presented in this paper are easily reproducible by the research community. Our source code is made publicly available [13] and it automatically fetches and processes IBR data, which means that the dataset is also available. The code is structured in such a way that one simply needs to format its data according to our data format to be able to launch `Chocolatine` on different data sources.

## VIII. CONCLUSION

In this paper we proposed `Chocolatine`, which detects remote outages using Internet Background Radiation traffic. The underlying predictive methodology is based on SARIMA models. Both the method and the data are easy to respectively deploy and collect in most ISP. We show that our method detects outages as quickly as 5 minutes after their occurrence with a 90% true positive rate and a small percentage of false alarms (2%). `Chocolatine` is able to detect outages in time series with as little as 20 IP addresses. Moreover, we compare its performance against other passive and active detectors. Its

share of common events with other sources is high, while each technique seems able to reveal specific events. We observe that the shares of common events, the overall and two-by-two intersections, are the most significant, while each technique seems able to reveal specific events too.

Our method is tailored to seasonal data and is robust to noise. It is therefore applicable to many other data sources reflecting Internet activity. For example, we plan to experiment its deployment on access logs of widely popular content. Its operational integration into the CAIDA's IODA outage detection system [17] is already in progress.

## REFERENCES

[1] Aceto, G., Botta, A., Marchetta, P., Persico, V., Pescapé, A.: A comprehensive survey on internet outages. Journal of Network and Computer Applications (2018)
[2] akamai. https://www.akamai.com/uk/en/about/our-thinking/state-of-the-internet-report/
[3] Benson, K., Dainotti, A., Claffy, K.C., Aben, E.: Gaining insight into as-level outages through analysis of internet background radiation. In: Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on. pp. 447–452. IEEE (2013)
[4] Benson, K., Dainotti, A., Snoeren, A.C., Kallitsis, M., et al.: Leveraging internet background radiation for opportunistic network analysis. In: Proceedings of the 2015 Internet Measurement Conference. pp. 423–436. ACM (2015)

## TABLE II: Ground truth — validation (Section VI-A)

| Event | Detection Time Frame | Time series |
|---|---|---|
| Czech ISP | 16–02–2009 16:20–17:20 | AS={62, 135, 158, 166, 223, 291, 348} |
| AfNOG | 03–05–2009 12:00–13:00 | AS={3, 242, 467} |
| CNCI | 17–08–2009 18:00–18:40 | AS={80, 149, 333, 360, 524, 580, 585} |
| RIPE-Duke | 27–08–2010 08:30–09:30 | AS={48, 54, 56, 63, 95, 143, 153, 204, 209, 210, 283, 310, 374, 377, 384, 385, 397, 398, 443, 474, 475, 483, 488, 497, 509, 564, 575, 595, 674, 676, 694, 714, 788, 791} |
| JunOS bug | 07–11–2011 14:00–15:00 | AS={7, 45, 68, 71, 73, 93, 119, 160, 177, 181, 187, 209, 215, 229, 257, 260, 273, 278, 297, 314, 316, 317, 320, 322, 324, 325, 328, 332, 335, 336, 337, 347, 392, 414, 415, 425, 429, 431, 479, 485, 490, 493, 501, 504, 529, 535, 569, 597, 624, 628, 636, 647, 650, 654, 655, 697} |
| Egypt | 01–27–2011 21:00 – 02–02–2011 12:00 | Countries={Egypt}, Regions={ 978, 971, 984, 985, 974, 979, 980, 993}, AS={8452, 36992, 24863, 24835} |
| Brazil | 21–03–2018 18:45 – 22–03–2018 10:00 | Countries={Brazil}, Regions={Amazonas, Bahia, Caera, Distrito Federal} |
| Syria | 27–05–2018 22:00 – 28–05–2018 06:00 | Countries={Syria} |
| Syria | 30–05–2017 00:00–06:00 | Countries={Syria} |
| Azerbaijan | 02–07–2018 12:00–03–07–2018 18:00 | Countries={Azerbaijan} |
| DRC | 23–12–2017 15:00–26–12–2017 09:00 | Countries={Democratic Republic of the Congo} |
| Gambia | 30–11–2016 17:00–04–12–2016 22:00 | Countries={Gambia} |

## TABLE III: Ground truth — comparison (Section VI-B)

| Event | Detection Time Frame | Time series |
|---|---|---|
| Angola | 07–09–2018 16:57 – 08–09–2018 06:20 | Countries={Angola} |
| Iraq | 13–10–2018 15:10 – 18–10–2018 19:32 | Countries={Iraq} |
| Venezuela | 15–10–2018 18:00 – 19–10–2018 05:00 | Countries={Venezuela} |
| Tajikistan | 26–10–2018 10:00–14:20 | Countries={Tajikistan} |
| Ivory Coast | 28–10–2018 23:00 – 29–10–2018 08:00 | Countries={CI} |
| Argentina | 17–11–2018 11:00 – 18-11-2018 01:00 | Countries={Argentina} |
| Syria | 18–11–2018 22:00 – 19–11–2018 03:00 | Countries={Syria} |
| Taiwan | 19–11–2018 00:00–06:00 | Countries={Taiwan} |
| Armenia | 20–11–2018 11:00–15:00 | Countries={Armenia} |
| Algeria | 30–11–2018 03:00–17:00 | Countries={Algeria} |
| Gabon | 11–12–2018 17:00–23:55 | Countries={Gabon} |
| Kyrgyzstan | 11–12–2018 22:00 – 12–12–2018 02:00 | Countries={Kyrgyzstan} |
| AS 209 | 27–12–2019 15:00 – 28–12–2018 01:00 | AS={209} |
| Ethiopia | 03–01–2019 11:00–15:00 | Countries={Ethiopia} |
| Cameroon | 14–01–2019 11:00 – 15–01–2019 10:00 | Countries={Cameroon} |
| Indonesia | 14–01–2019 05:00 – 15–01–2019 08:00 | Countries={Indonesia} |
| Zimbabwe | 15–01–2019 04:00 – 17–01–2019 12:00 | Countries={Zimbabwe} |
| Zimbabwe | 17–01–2019 20:00 – 18–01–2019 16:00 | Countries={Zimbabwe} |
| Panama | 20–01–2019 15:00 – 21–01–2019 01:00 | Countries={Panama} |
| Laos | 24–01–2019 17:00–21:00 | Countries={Laos} |
| Panama | 29–01–2019 13:00–23:55 | Countries={Panama} |
| Morocco | 11–02–2019 06:00–16:00 | Countries={Morocco} |

[5] Dainotti, A., Amman, R., Aben, E., Claffy, K.: Extracting benefit from harm: using malware pollution to analyze the impact of political and geophysical events on the Internet. ACM SIGCOMM Computer Communication Review (CCR) **42**(1), 31–39 (Jan 2012)

[6] Dainotti, A., King, A., Claffy, K., Papale, F., Pescapè, A.: Analysis of a "/0" Stealth Scan from a Botnet. In: Internet Measurement Conference (IMC). pp. 1–14 (Nov 2012)

[7] Dainotti, A., Benson, K., King, A., Kallitsis, M., Glatz, E., Dimitropoulos, X., et al.: Estimating internet address space usage through passive measurements. ACM SIGCOMM Computer Communication Review **44**(1), 42–49 (2013)

[8] Dainotti, A., Squarcella, C., Aben, E., Claffy, K.C., Chiesa, M., Russo, M., Pescapé, A.: Analysis of country-wide internet outages caused by censorship. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. pp. 1–18. ACM (2011)

[9] De Gooijer, J.G., Hyndman, R.J.: 25 years of time series forecasting. International journal of forecasting **22**(3), 443–473 (2006)

[10] Dyn. https://dyn.com/

[11] Fachkha, C., Bou-Harb, E., Debbabi, M.: Fingerprinting internet dns amplification ddos activities. In: New technologies, mobility and security (NTMS), 2014 6th international conference on. pp. 1–5. IEEE (2014)

[12] Fontugne, R., Aben, E., Pelsser, C., Bush, R.: Pinpointing anomalies in large scale traceroute measurements. In: Internet Measurement Conference (IMC 2017) (2017)

[13] Guillot, A., Fontugne, R.: Chocolatine: darknet outage detector (May 2019). https://doi.org/10.5281/zenodo.2784847

[14] Hall, J., Aaron, M., Jones, B., Feamster, N.: A Survey of Worldwide Censorship Techniques (May 2018), internet Draft, draft-hall-censorship-tech-05 (section 4.6)

[15] Harder, U., Johnson, M.W., Bradley, J.T., Knottenbelt, W.J.: Observing internet worm and virus attacks with a small network telescope. Electronic Notes in Theoretical Computer Science **151**(3), 47–59 (2006)

[16] Holterbach, T., Molero, E.C., Apostolaki, M., Dainotti, A., Vissicchio, S., Vanbever, L.: Blink: Fast Connectivity Recovery Entirely in the Data Plane. In: USENIX NSDI 2019 (February 2019)

[17] IODA. https://ioda.caida.org/

[18] IODA Help. https://ioda.caida.org/ioda/help

[19] Net acuity. https://www.digitalelement.com/geolocation/, accessed: 02/07/2018

[20] CAIDA Prefix2AS. https://www.caida.org/data/routing/routeviews-prefix2as.xml

[21] Quan, L., Heidemann, J., Pradkin, Y.: Trinocular: Understanding internet reliability through adaptive probing. In: ACM SIGCOMM Computer Communication Review. vol. 43, pp. 255–266. ACM (2013)

[22] Quan, L., Heidemann, J., Pradkin, Y.: When the internet sleeps: Correlating diurnal networks with external factors. In: Proceedings of the 2014 Conference on Internet Measurement Conference. pp. 87–100. IMC '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2663716.2663721, http://doi.acm.org/10.1145/2663716.2663721

[23] Richter, P., Padmanabhan, R., Spring, N., Berger, A., Clark, D.: Advancing the art of internet edge outage detection. In: Internet Measurement Conference (IMC 2018) (2018)

[24] Ripe atlas. https://atlas.ripe.net/, accessed: 27/04/2018

[25] RIPE duke experiment. https://labs.ripe.net/Members/erik/ripe-ncc-and-duke-university-bgp-experiment

[26] Schulman, A., Spring, N.: Pingin'in the rain. In: Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference. pp. 19–28. ACM (2011)

[27] Shah, A., Fontugne, R., Aben, E., Pelsser, C., Bush, R.: Disco: Fast, good, and cheap outage detection. In: Network Traffic Measurement and Analysis Conference (TMA), 2017. pp. 1–9. IEEE (2017)

[28] Sánchez, M.A., Otto, J.S., Bischof, Z.S., Choffnes, D.R., Bustamante, F.E., Krishnamurthy, B., Willinger, W.: A measurement experimentation platform at the internet's edge. IEEE/ACM Transactions on Networking (TON) (2014)

[29] The UCSD network telescope. https://www.caida.org/projects/network_telescope/

[30] Wang, Q., Chen, Z., Chen, C.: Darknet-based inference of internet worm temporal characteristics. IEEE Transactions on Information Forensics and Security **6**(4), 1382–1393 (2011)

[31] Wustrow, E., Karir, M., Bailey, M., Jahanian, F., Huston, G.: Internet background radiation revisited. In: Proceedings of the 10th ACM SIGCOMM conference on Internet measurement. pp. 62–74. ACM (2010)

[32] Yu, Q., Jibin, L., Jiang, L.: An improved arima-based traffic anomaly detection algorithm for wireless sensor networks. International Journal of Distributed Sensor Networks **12**(1), 9653230 (2016)