# On Identifying Anomalies in Tor Usage with Applications in Detecting Internet Censorship

Joss Wright
University of Oxford
Oxford, United Kingdom
joss.wright@oii.ox.ac.uk

Alexander Darer
University of Oxford
Oxford, United Kingdom
alexander.darer@linacre.ox.ac.uk

Oliver Farnan
University of Oxford
Oxford, United Kingdom
oliver.farnan@balliol.ox.ac.uk

## ABSTRACT

We develop a means to detect ongoing per-country anomalies in the daily usage metrics of the Tor anonymous communication network, and demonstrate the applicability of this technique to identifying likely periods of internet censorship and related events. The presented approach identifies contiguous anomalous periods, rather than daily spikes or drops, and allows anomalies to be ranked according to deviation from expected behaviour.

The developed method is implemented as a running tool, with outputs published daily by mailing list. This list highlights per-country anomalous Tor usage, and produces a daily ranking of countries according to the level of detected anomalous behaviour. This list has been active since August 2016, and is in use by a number of individuals, academics, and NGOs as an early warning system for potential censorship events.

We focus on Tor, however the presented approach is more generally applicable to usage data of other services, both individually and in combination. We demonstrate that combining multiple data sources allows more specific identification of likely Tor blocking events. We demonstrate the our approach in comparison to existing anomaly detection tools, and against both known historical internet censorship events and synthetic datasets. Finally, we detail a number of significant recent anomalous events and behaviours identified by our tool.

## CCS CONCEPTS

• **Networks** → **Network measurement**; • **Social and professional topics** → **Technology and censorship**; • **Security and privacy** → *Pseudonymity, anonymity and untraceability*;

## KEYWORDS

information controls, censorship, filtering, anomaly detection

## 1 INTRODUCTION

Nation states, and others, increasingly employ internet filtering as a means of controlling access to information, and as a tool to limit social and political organisation. Given the central role that the internet plays in communications for a large proportion of the global population, understanding the application and development of filtering technologies, and the effects of these methods on individuals and society, is of great importance. Whilst analyses of known filtering regimes allow us to identify tools, techniques, and limitations of filtering approaches, we consider that discovering internet filtering behaviour in less-studied regions is of great importance.

Much existing research into internet filtering has focused either on observing practices of states already known engage in filtering, or in the development of censorship circumvention tools. Whilst multilateral studies of censorship have been conducted, most notably the seminal work of Deibert et al. [11], these approaches have typically amalgamated manual country-specific investigations. In the case of Deibert et al., countries were hand-ranked according to a number of broad criteria for internet freedom, based on network measurements as well as media reporting and expert interviews.

The work presented here provides a means to alert researchers and activists to developing events that may otherwise have been missed by focusing on patterns of circumvention tool usage around the world. As an initial step our tool currently reports new anomalies and a current ranking of most anomalous countries to a mailing list on a daily basis. The `<infolabe-anomalies>` mailing list has been running publicly since August 2016, has subscribers from academia and civil society organisations, and has provided the first known detection of a number of significant ongoing Tor-related blocking events that we detail in §7.

### 1.1 Contributions

This work presents a theoretical contribution to network anomaly detection, a practical contribution in the form of an implemented tool for detecting anomalous events in Tor usage data, a resource in the form of a public dataset of detected anomalies in historical Tor traffic, and a practical analysis demonstrating the detection of real-world events: we identify known, previously unreported, and newly-detected filtering-related events.

We make the following practical contributions:

- An open tool to detect and highlight anomalies in per-country usage of the Tor network;

- a continually-updated daily ranking of the most anomalous countries in terms of their usage of Tor.

These are built on our key methodological contribution:

- An approach for detecting and quantifying anomalous periods of per-country Tor usage incorporating multiple usage measurements.

We validate the effectiveness of our approach in detecting both a range of artificial anomalies, and known reported filtering events against the Tor network. We also demonstrate our approach's improved detection accuracy compared to the existing Tor metrics anomaly detector, as well showing its additional capabilities in terms of identifying anomalous periods and ranking anomalies by strength.

## 1.2 Problem and Approach

When an entity, such as a state or ISP, chooses to filter or block certain types of information, the resulting patterns of traffic reflect the intervention in the form of statistical anomalies. In a global system, in which many entities may be interfering with traffic or publicising their attempts to do so, it is desirable to identify *localised* anomalies and to gain an understanding of their nature.

To detect anomlies, we model each country's Tor usage *relative* to the behaviour of other countries, not as an individual time series. A given country's usage pattern is judged as anomalous if it deviates from its previous behaviour *relative to other countries*.

The usage patterns of a tool such as Tor, explicitly developed and publicised as a means for bypassing network censorship, are affected by a range of factors such as filtering, social and political unrest, unrelated network outages, and media reporting [4]. The work presented here therefore identifies *statistical anomalies* in Tor usage metrics, but we highlight that such anomalies serve as an indicator, not a proof, of censorship or interference.

In later sections we make use of both standard Tor traffic and blocking-resistant *bridge node* traffic to identify direct blocking of Tor. Combining anomalies across metrics allows identification of declines in normal usage combined with rises in blocking-resistant bridge usage. This corresponds to users being unable to access Tor normally, and so switching to blocking resistant approaches. As we demonstrate in §7, this provides a targeted identification of filtering-related anomalies.

We extend a line of research initially proposed by Jackson and Mudholkar [18] for application in industrial process control, and later employed by Lakhina et al. [22] to detect network-wide traffic anomalies from per-link data in high-performance networks. Our approach differs from that of [22] in a number of ways. Firstly, we do not assume that the underlying set of time series are stationary, but instead allow for series to evolve over time. Secondly, we account for *seasonality* in time series. Most importantly, however, we identify *per-country* anomalies rather than global. Finally, we dynamically adapt our anomaly thresholds for each series to account for long-term evolution of the data.

We directly apply our tool to analysis of Tor usage anomalies, and report on its demonstrated utility for detecting anomalies of practical concern to activists and NGOs working to support censorship circumvention and freedom of expression. A number of such actors subscribe to our public mailing list, and have used our detection results to identify newly-emerging filtering behaviours.

## 2 EXISTING WORK

Internet filtering has received attention from various fields. Technical research has focused on mechanisms of censorship and the development of circumvention approaches. The social sciences have investigated motivations of censors, and their legal, economic, and societal effects.

## 2.1 Technical Analysis

Arguably the most well-known national-level filtering system is that of China, commonly known as the Great Firewall. One of the earliest significant studies of this system was presented by Clayton et al.[5], who isolated one mechanism by which connections were interrupted if particular keywords were identified in traffic. The mechanism discovered by Clayton et al. resulted in TCP RST packets being sent from an intermediary router to both source and destination of a connection if a filtering criterion was met. The authors further demonstrated that if the two endpoints of the connection ignored the TCP RST, the connection could successfully continue.

In more recent work, it has become apparent that the Chinese approach to filtering is both complex and evolving. In two recent papers, a group of anonymous researchers have explored manipulation, or poisoning, of DNS records that pass through China [2, 3]. This work has identified DNS manipulation as one of the most prevalent forms of filtering in China. Similarly, Wright [37] demonstrated that DNS censorship had different effects between different regions within China, with significant variation in the nature of the DNS poisoning seen across the country. Similarly, Farnan et al. [14] showed that the approach taken to DNS poisoning in China resulted in pollution of both network requests and DNS servers themselves.

Crandall et al.[7] make use of *latent semantic analysis* to derive, from known terms blocked in HTTP traffic going into China, semantically related keywords that might also be blocked. These derived keywords can then be verified by the simple process of attempting to make HTTP connections into China containing the suspect words. This approach aims to produce a continually-updated list of blocked terms that could be used to maintain an understanding of those terms most offensive to the filtering authorities. Similarly, Darer et al. [9, 10] have used keyword- and crawling-based approaches to discover previously unidentified blocked domains.

## 2.2 Global Studies

Perhaps the most comprehensive study to date of global filtering practices is given by Deibert et al. [11]. In this work the authors carried out a range of remote and in-country analyses over a number of years, incorporating both technical measurements and interviews with local experts. The resulting research presented a series of snapshots of individual countries, with both an overview of the social, political, and technical landscape, and censorship practices rated on a simple scale in various categories of content: political, social, conflict and security, and internet tools.

Some forms of filtering act not at the network layer, but on application level or social filtering. King et al. [21] studied manual censorship practices in Chinese long-form blogging, and demonstrated that the Chinese censorship authorities were chiefly concerned with preventing calls to *collective action* whilst allowing significant levels of government criticism.

## 2.3 Anomaly Detection

The Tor project maintain a censorship flagging tool, as described by Danezis[8]. This tool uses a particle-filtering approach to model the ratio of daily connections for each country in a seven-day time period. If a country's ratio of current to past users increases or decreases significantly more than the average of the fifty largest Tor-using countries, then an anomaly is flagged. These reported anomalies are available at the Tor Project's metrics portal [29]. We evaluate our approach's accuracy against that of Danezis in §6.

A related approach was used by Lakhina et al. [22] to identify *network-wide* anomalies in high-speed networks. This work assumed that long-term network usage was stable, and made use of data gathered from a restricted set of link-level observation points to detect network-wide anomalies. Our approach relaxes both of these assumptions, neither of which hold for the Tor metrics data. These extensions are discussed in greater detail in §4.1.2.

Several other works have extended or expanded aspects of [22], notably [34], [39], and [16]. These largely focus, however, on using a small number of network observation points to infer network-wide anomalies, and as such typically begin from relatively low-dimensional data. Our approach inverts this concept by detecting per-observation anomalies across a dataset with several hundred dimensions, representing individual countries' usage, in order to highlight states displaying anomalous behaviour.

## 3 CONCEPTS

In this section we discuss the fundamental techniques underlying our approach, and discuss their application to the dataset we use in the rest of this work.

## 3.1 Tor

Tor [12] is an approach to anonymous web-browsing that offers realistic compromises between latency, usability, and the strength of the anonymity properties that it provides. The most visible end-user aspect of Tor is the Tor Browser Bundle, which provides a web-browser that both uses the Tor network for transport, and is tailored to reduce identifiability of end users.

Managed by the Tor Project, Tor has developed into a global network of volunteer-run relays that forward traffic on behalf of other users. The network makes use of an *onion routing* approach that build encrypted circuits between relays, preventing most realistic adversaries from linking Tor users to particular streams of traffic exiting the network.

The most sigificant aspect of the Tor network for the present work is that, by its nature, users' traffic is relayed via third parties. As such, and in addition to its anonymity properties, Tor provides a means to bypass many forms of internet filtering. Censorship circumvention is a core aspect of the Tor Project's goals, and sigificant ongoing research work[26, 33, 36] is aimed at ensuring

that Tor is resilient against attacks and continues to offer means to evade national-level filters.

While the extent and popularity of Tor's use in regions that experience significant levels of filtering, such as China, is open to debate [32], Tor is known to have been blocked actively by a number of states, including China and Iran, that object to its use to bypass local internet restrictions and to act anonymously. Significantly, Tor is also arguably the highest-profile censorship circumvention tool at the international level and has received significant media coverage, making it one of the tools of choice for internet activists.

*3.1.1 Tor Metrics Data.* Tor's role as a high-profile censorship circumvention network make it a useful indicator of global filtering practices. To support analysis of the tool, the Tor project provide estimated daily per-country usage statistics, gathered by counting the number of client requests to central *directory authorities* on a daily basis.

It is assumed that each client, on average, will make ten requests per day, and as such the aggregate user statistics are divided by ten to provide a final estimate of usage. This data is averaged across each 24-hour period to provide the average number of concurrently connected Tor clients for that day[30]. Whilst the number of distinct clients per day cannot be estimated with any accuracy, the methodology of the Tor metrics portal provides a sufficiently stable estimate.

From these estimates we obtain a set of 251 time series representing individual countries according to the GeoIP database used by Tor. These time series comprise daily observations ranging from the beginning of September 2011 to the time of writing. From these, we remove those countries whose Tor usage never rises above 100 users to discount countries whose variance is too high to allow meaningful anomaly detection.

In later sections, we combine normal usage trends in Tor with censorship-resistant *bridge node* usage to identify correlated anomalies. This is discussed in further detail in §4.4.

## 3.2 Principal Component Analysis

Principal component analysis was developed by Pearson[27] as a means to produce tractable low-dimensional approximations of high-dimensional datasets. The original set of variables, which may display correlations, are transformed to a set of linearly uncorrelated variables know as *principal components*.

When data displays a high degree of correlation between variables then a small number of the most significant principal components may be sufficient to describe the original data to a high degree of accuracy. In many practical scenarios, high dimensional data can be described using only two or three of the most significant principal components. See [19] for a detailed treatment of principal component analysis and the various choices and compromises to be made when applying the technique.

The practical result of this is that our results are not influenced by countries with large usage numbers; the principal component analysis considers variance, not magnitude, in calculating the contribution of each country to the model.

## 4 APPROACH

The basic operation of our approach are described here, and are given as pseudocode in Algorithm 1.

---

**1 PCATagAnomaly**

> **input** : *usage* ← Set of per-country time series
> **output**: *anomalies* ← Set of per-country anomaly time series

**2**   (Clean data; remove seasonality)

**3**   *medians* ← {median residual errors for each country}

**4**   *mads* ← {median absolute deviations (MADs) of residual errors for each country}

**5**   **foreach** *day in usage* **do**

**6**   | *pc* ← calculate principal components over all countries' usage[(day-179):day]

**7**   | **foreach** *country* **do**

**8**   | | *recons* ← reconstruct *day* value for *country* using $pc[1:12]$

**9**   | | *obsv* ← observed value for final day for *country*

**10**  | | *err* ← abs( *obsv* - *recons* ).

**11**  | | $medians_{country}$ ← update median using *err*

**12**  | | $mads_{country}$ ← update MADs using $medians_{country}$ and *err*

**13**  | | **if** $abs(err) > abs(mads_{country} \times 2.5)$ **then**

**14**  | | | $anomalies_{country \times day}$ ← 1

**15**  | | **end**

**16**  | | **else**

**17**  | | | $anomalies_{country \times day}$ ← 0

**18**  | | **end**

**19**  | **end**

**20**  **end**

**Algorithm 1:** Basic anomaly tagging algorithm. (Anomaly magnitudes omitted for brevity.)

---

### 4.1 Overview

Starting from Tor's per-country usage data, we initially remove all countries whose usage never rises above 100 users, to avoid the unacceptably high variance in such data. We then apply the STL algorithm to identify and remove any seasonality – in our case weekly trends – in individual countries.

For each 180-day period in the dataset we apply a principal component analysis over the usage time series for all countries, resulting in a set of components for that time window. Taking the true observed usage for each country for the final day of each window, we calculate the *approximated* value from the first 12 principal components. This provides the expected value for each country based on previous behaviour[1].

For each country we calculate the difference between the true value and the reconstructed value, providing a *residual error* that was not captured by the restricted set of principal components.

---

[1]Using the full set of principal components at this stage would result in a perfect reconstruction of the original observed values.

---

We maintain a rolling calculation of both the median observed residual error and the median absolute deviation of the errors for each country. We mark a day as anomalous if the observed residual error falls outside of 2.5 median absolute deviations from the median.

We now detail the individual steps listed above, and justify our choices of parameters.

*4.1.1 Removal of Seasonality.* Per-country Tor usage data, as with much network usage data, exhibits significant *seasonality*, typically on a weekly basis, reflecting changes between usage on weekdays and at weekends. This continual cyclical change in usage can reduce the accuracy of principal component analysis due to varying levels of seasonality exhibited by different countries.

We employ the *Seasonal and Trend Decomposition using Loess* (STL) method of Cleveland et al. [6] to remove the seasonal component of each series, leaving the trend component and the residual noise as inputs to our anomaly detector. In later sections, however, we show the original data with seasonality restored.

*4.1.2 Rolling Analysis.* Principal component analysis does not account for ordering in observations, and as such cannot account for evolution of a dataset according to trends or seasonablity. To account for developing patterns, therefore, we perform a rolling principal component analysis over smaller time windows within the series. For the purposes of our experiments, we make use of a 180-day window as a balance between sufficient data for useful principal component analysis, given the number of dimension in the data, against the evolution of the daily Tor metrics. See Ringberg et al. [31] for a discussion of the sensitivity of PCA to such factors.

*4.1.3 Selection of Components.* For PCA, the full set of principal components allows reconstruction of the full data set. As fewer components are selected, less variance in the original dataset is captured. A common approach to selecting an appropriate number of components for modelling is to make use of *Kaiser's criterion* [20] to select only those principal components with eigenvalue greater than 1, representing those components that provide more information than a single average component. Based on this heuristic, our experimental results suggest twelve principal components as broadly optimal across the dataset.

With appropriately calculated principal components, we can reconstruct an approximate value for each day's Tor usage based on previous behaviour. We highlight that at no point do we *predict* forecasted values for usage in future days. In each case, we reconstruct a day's usage based on principal components in order to compare against the true observed value, and thus to calculate deviance from prior behaviour relative to other countries.

### 4.2 Calculation of Residuals

After reconstructing data from principal components, the result is a set of *residuals* that express variances in the observed data not captured by the current principal component model. A sufficiently large-scale residual represents behaviour that deviates significantly from previous patterns, and is thus of interest.

### 4.3 Identifying Anomalies through Residuals

The residual errors calculated during the reconstruction accounts for variance in the dataset that is not expressed by the chosen principle components in the approximate model.

- Positive residuals represent drops in expected Tor usage for a country.
- Negative residuals represent increases in expected Tor usage for a country.
- Magnitude of residuals expresses how much a country varies from its previous behaviour relative to other countries.

A key advantage of identifying anomalies from residual errors rather than raw usage numbers is that it incorporates the expected *trend* of the data. This identifies anomalous periods even when no visible shift in usage is seen: a flat usage trend where the expectation is a rise or fall is correctly identified as anomalous by our approach. This capacity to identify anomalies in apparently typical usage is an important and unusual aspect of our technique, taking advantage of the relative patterns of usage between countries.

A second advantage of this approach is that each day can be judged as anomalous or not based on a model of behaviour relative to other countries. As such, in contrast to many other anomaly detection approaches, we identify *periods* of anomalous behaviour in which a country may be experiencing ongoing elevated or reduced usage. Other approaches typically flag an individual day as a significant spike or drop, but cannot identify ongoing periods as anomalous. This capability greatly aids our ability to study time-bounded changes in Tor usage.

### 4.4 Combining Features to Identify Targeted Filtering

It is fundamental to the broader goals of this work that usage anomalies in appropriately selected traffic, and in particular from circumvention tools, can be indicative of the imposition or relaxation of filtering. At the same time, it is clear that other types of event, both technical and sociopolitical, can lead to shifting patterns of usage in these tools.

We aim to identify two forms of event: firstly, direct blocking of the Tor network; secondly, changing characteristics of Tor usage in response to exogenous factors. The censorship of a major international website, such as YouTube, has the potential to drive a noticeable number of users to Tor, and as such Tor becomes a useful *proxy variable* [35] for a broader class of filtering behaviour. We discuss this in relation to specific events in §7.

For the first of these classes of event, we detect likely candidates by carrying out anomaly detection on multiple metrics and combining outputs to highlight periods in which anomalies were detected in more than one series. The most useful of these for our purposes is to combine negative trends in standard Tor usage with positive trends in blocking-resistant bridge node usage, reflecting users unable to access Tor normally switching to the tool's blocking-resistant mode.

As such we can identify days in which both standard and blocking-resistant time series were anomalous. Even without refinements, such as allowing time lags between anomalies in each series, this approach already highlight a number of significant cases, which are illustrated in §7.

### 4.5 Expected Error and Anomalous Threshold

A key element in the approach presented in this work is to determine an appropriate threshold for events to be considered anomalous. The size of this threshold value is inherently linked to the expected error in the technique. We here discuss and justify our approach to calculating this threshold, making use of *robust statistics*[17] to minimise false detection rates.

A naïve anomalous threshold can be defined as a proportion of the usage for that day. If the reconstructed value deviates by more than some percentage of the observed value, an anomaly is detected.

This approach is problematic. Critically, different countries may be modelled more or less accurately than others. As such, countries that are typically modelled poorly would produce a high proportion of anomalous periods.

As such, we calculate an ongoing threshold based on the characteristics of each country. By tracking the expected residual value for each country an expected anomalous threshold can be determined based on typical observed errors.

The standard approach of basing this threshold on the mean and standard deviations are, however, not *robust* against outliers in the dataset due to their assumption that errors are Gaussian. We therefore calculate thresholds based on the *median absolute deviation about the median* (MAD) to define the expected error in normal usage [24].

The median is robust against outliers in the dataset; a small number of extreme events do not significantly alter its value. Similarly, by taking the median of the absolute deviations about the median as a measure of the statistical dispersion in the dataset, we avoid anomalies from overly affecting the remaining data points.

As a default, we consider events as anomalous if they fall outside of 2.5 median absolute deviations[2] from the rolling median value. See [24] for a discussion of the robustness of the median and MAD against outliers, and a justification of a 2.5 median absolute deviation threshold.

### 4.6 Ranking of Countries

The size of the residual error from the principal componenet analysis provides a convenient metric by which to rank countries according to the level of anomalous behaviour that they exhibit in a given time period. We make use of the size of the median absolute deviation about the median to rank countries, as shown in Figure 1.

We now proceed to discuss the application of our technique, and the validation of the approach.

In §6 we evaluate our approach against synthetically injected anomalies in the data to analyse the effectiveness of our detection methods as the magnitude and severity of the anomalies vary. We also compare our detection mechanism against the small number of verified reported blocking events against the Tor network.

Finally, in §7 we conduct a series of analyses of the Tor metrics data to identify anomalous countries and specific periods of anomalous behaviour.

---

[2]Corresponding to roughly one expected false positive every 80 days. See §6 for an experimental analysis of false positives in our approach.
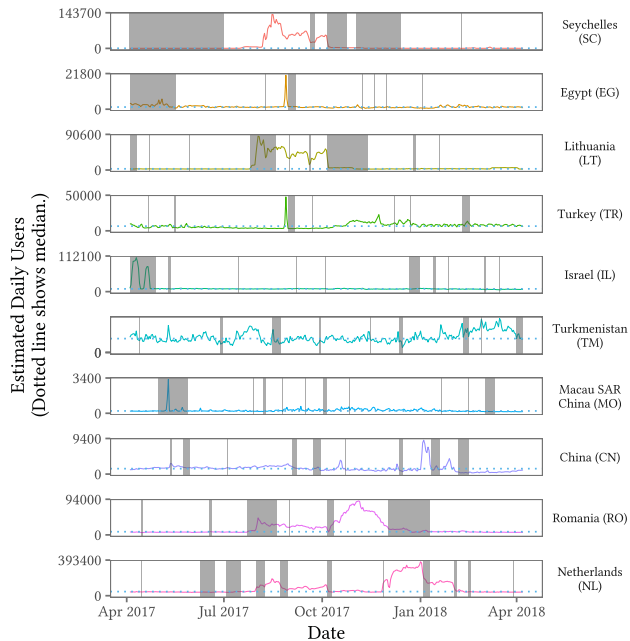
**Figure 1: Ten most anomalous countries according to median absolute deviation of residuals over the previous year. Grey areas highlight detected anomalous periods.**

## 5 ETHICS

Conducting research into network filtering presents a number of ethical issues [38]. The most significant of these is that approaches to investigating network filtering may require direct access to filtered networks. In practice this often involves the participation of in-country experts to conduct local network tests.

Due to the uncertain legal, or quasi-legal, status of violating or investigating state-level network filters, it is generally impossible to quantify the risks to research participants in carrying out network tests. The classic models of informed consent used in many other fields of research can be difficult to apply for a number of reasons, the most important of which is the lack of meaningful informed consent afforded by automated testing on behalf of users, and the legal uncertainty surrounding attempted access to filtered resources on a test subject's network connection.

We therefore assert that, where possible, research into network filtering should make use of passive measurements and existing available data sources. The work in this paper is a deliberate attempt to maximise the effectiveness of such a passive approach.

## 6 VALIDATION

In this section, we judge the efficacy of our method in terms of its ability to detect anomalies, in a variety of circumstances, as well as its false classification rate.

A significant dificulty in validating unsupervised machine learning systems is that it is largely impossible to obtain comprehensive ground truth for internet filtering events, nor are there publicly-available exhaustive lists of filtering events. Indeed, the work here was motivated partially in an attempt to allow a more exhaustive

tracking of such events. Filtering is, by and large, an opaque process that is rarely announced. Even when states do choose to filter connections openly, the details of that filtering are not typically made public.

As observed in [15], this is an inherent problem in *unsupervised* anomaly detection algorithms. In the the following sections we address this lack by injecting artificial anomalies into a synthetic dataset and comparing this to the Tor Project's existing anomaly detection approach, as well as evaluating our method against an existing list of known filtering events.

In the following, we examine both false positive and false negative rates in evaluating detection rates of anomalous behaviour. A false positive in this context is a period in which there is no genuinely anomalous activity, but anomalous activity is reported. A false negative is a period in which there is anomalous activity but is is not detected.

### 6.1 Evaluation in Synthetic Data

To test our approach, and to create a fair comparison against the existing deployed tool from the Tor Project, we inject artificial anomalies into synthetic data generated according to underlying features of real-world Tor usage.

An alternative test for false negatives is to compare the results from our approach with an external list of known censorship events. This allows us to test whether periods exist in which we did not detect anomalous behaviour during a period where external sources believe an event occurred. We take this approach in §6.5.

### 6.2 Generating Synthetic Data

To evaluate our approach against an approximation of real-world data, we use the underlying features of genuine observed Tor data to generate a synthetic set of time series.

To do so, we select a year-long period of Tor data in which no major global events can be observed. This was to avoid an unfair basis for comparison between our approach and that of the Danezis. As such, we selected the year running from the 1st January 2014 to the 31st December 2014.

To remove, as far as possible, genuine anomalies from this dataset we first decompose the series into trend, seasonal, and residual components through use of the STL algorithm [6]. This allowed us to preserve seasonal properties of the data separately from the underlying trend. We emphasise that, whilst STL is also used in our anomaly detection approach, the application of it here preserves, rather than removes, the underlying features of the data and thus is not unfairly biasing the synthetic dataset towards our approach.

The underlying trend data is then smoothed using a 28-day rolling median average. Due to the robust nature of the median against small outliers, this approach preserves broad-scale trends in the data whilst removing, as far as possible, small-scale deviations. Without an objective labelled set of anomalies we cannot guarantee that no anomalies were preserved in the final dataset, but a visual inspection did not reveal any significant causes for concern.

We then calculate, for each country, the mean and the standard deviation of the residual errors after the trend and seasonal components have been removed. This gives a base set of parameters from which to generate random noise to be added to each series.

To create the final synthetic dataset, we recombine the smoothed underlying trend data with the seasonal component and add randomised noise. As it is impossible to characterise the "true" noise process without having labelled anomalies we conservatively add Gaussian noise drawn according to the observed mean and standard deviation. This provides a "clean" dataset without anomalies, based on real-world patterns of behaviour.

## 6.3  Injecting Anomalies

As with the underlying data, we generate anomalies based on properties observed in real-world data. The strength of the injected anomalies is based on the average daily users for each country, and magnified upwards or downwards gradually to create the anomaly.

To create an anomaly, the number of users in each set was increased or decreased by 0–100%. Anomalies are added to the data gradually, ranging over periods from one to four weeks. These parameters were selected based on observation of known anomalies and visual inspection of the original dataset.

In total, for the year of synthetic data, we injected a total of 250 anomalies across all countries, randomly drawn from the space of possible parameters.

This synthetic, labelled dataset provides the basis both for objective evaluation of the effectiveness of our technique, and as an unbiased means of comparison between our approach and that of [8]. We now evaluate the effectiveness of these two approaches.

## 6.4  Comparison of Tools

An evaluation of false positive and false negative rates in detecting anomalous periods allows both an objective judgement on the effectiveness of our approach, and a comparison against the existing tool used by the Tor Project [8]. To carry out this comparison, we formatted the clean synthetic dataset appropriately for each tool and compared the detected anomaly series from each to the injected set of anomalies.

One problematic element of such a comparison is in the nature of event reporting from each tool. As mentioned, our approach reports day-by-day anomalies based on principal component modelling. By comparison, [8] bases its detection on significant spikes and dips on a day-by-day basis. As such, it is far less likely that Tor Project's existing tool will report anomalous periods, but will instead detect only the points at at which an anomaly starts and ends. This should hypothetically result in a much higher detection accuracy rate for our tool on a day-by-day comparison: an anomaly that lasts for ten days will typically only produce two anomalously flagged days in the Tor Project's detection scheme, whereas it may result in ten days for our tool as each day in the anomalous period may be identified. By contrast, however, our tool's approach leaves us open to a potentially higher false negative rate when a period is falsely judged to be anomalous.

We highlight again, however, that this period-based rather than event-based approach is one of the key strengths of our improved approach – we report entire periods as anomalous rather than simply identifying point anomalies.

As such, to compare, we perform a simple analysis: the output of each tool is evaluated according to the ground truth in the labelled synthetic dataset. Days correctly identified as anomalous contribute

|  | Tor Metrics | Principal Component |
|---|---|---|
| True Positives | 8.57% | 20.08% |
| True Negatives | 92.75% | 94.25% |
| False Positives | 7.25% | 5.75% |
| False Negatives | 91.43% | 79.92% |
| Total Days Flagged | 2962 | 2820 |
| Minimal Detection Total[1] | 88 | 139 |

Total anomalous days across entire set was 4214.
[1] Anomalies during which at least one day was identified.

**Table 1: Comparison between Tor Metrics and Principal Component approach on synthetic data.**

to the *true positive* rate, whilst days marked as anomalous that are not in the synthetic data contribute to the *false positive* rate. Similarly, if a day is anomalous in the synthetic data and missed by our tool, it contributes to the *false negative* rate, whilst days correctly identified as not anomalous contribute to the *true negative* rate. These values are reported in Table 1.

Our principal component-based approach significantly outperforms the currently deployed Tor Metrics detector both in marking genuine anomalies and in avoiding marking non-anomalous days incorrectly.

The overall detection rate of our approach is over twice that of the alternative, at 20% of all genuinely anomalous days being identified. This figure is somewhat misleadingly low, however, as this includes many correctly-identified anomalous *periods* for which, however, some individual days were not themselves considered anomalous.

These results suggest that in realistic data generated from observed real-world trends, the proposed principal component analysis-based approach significantly outperforms the existing deployed tool.

*6.4.1  Ranking.* We have attempted, as far as possible, to undertake a fair comparison of the quantitatively comparable elements of these two approaches, despite significant differences in their output. In addition, however, our approach offers a number of advantages for analysis. The most significant of these is the ability to rank countries according to the strength of the anomalies they have demonstrated over time in terms of deviation from expected behaviour. The `infolabe-anomalies` mailing list reports daily the top-ranking anomalous countries for the previous day, week, and month in addition to a list of all countries anomalous for that day.

It is worth highlighting that whilst realtime detection is of great interest to the commmunity, the ability to study historical anomalies in the Tor metrics dataset is also of significant value.

## 6.5  Detection of Known Events

Having calculated anomalous statistics over a synthetic data set, we now aim to validate our approach by comparing anomalies detected in real data against countries and periods in which internet restrictions are known to have been applied, or in which significant events were occurring that may have influenced usage of circumvention tools.

| Date | Country | Description of Event |
|------------|---------|------------------------------|
| 2012-10-18 | Iran | TLS key exchange DPI.[1] |
| 2012-12-16 | Syria | DPI on TLS renegotiation. |
| 2013-01-30 | Japan | Bridge blocked. |
| 2013-03-09 | Iran | SSL handshake filtered. |
| 2013-03-26 | China | Probing obfs2 bridges. |
| 2014-03-28 | Turkey | Tor website blocked. |
| 2014-07-29 | Iran | Block directory authorities. |
| 2015-02-01 | China | Obfs4 bridges blocked. |

[1] See §6.5 for a discussion of this particular anomaly.

**Table 2: Complete list of reported, and detected, Tor blocking events.**

For this purpose we use [1], a list of reported and verified filtering events against the Tor network dating from 2008 to 2015. This list includes a brief description of each reported event, the dates when the event was first reported, and how the blocking was resolved.

The list of events used in this evaluation[1] was compiled through bug reports, talks, examination of blog postings, and the use of machine learning on blog postings to identify reports of censorship automatically. As such, the exact timing of the events is somewhat fuzzy; a blocking event against Tor could have occurred some time before bug reports and blog postings were filed.

In addition, [1] is unfortunately brief, reflecting a significant lack of data available concerning this topic. As discussed, a motivation for this work is to provide a baseline of reliable indicators to allow for potentially censorship-related anomalies to be identified and investigated more thoroughly.

The Tor Project's metrics data does not cover the full time range of the events listed in [1]. For those events that do fall within the available data, we analyse here whether these would be detected by our approach.

As shown in Table 2, only eight reported events coincide with the available published metrics data. Of these, our approach successfully classifies all events as anomalous[3]. In all cases except the Iranian DPI filtering on TLS that occurred in 2012, our anomalies coincide with the reported event from [1]. In the case of Iran in 2012, we detect an anomalous period beginning two weeks *before* the reported event, corresponding to an immediate sharp fall in Tor usage, followed by a longer period of slow decline over the following month.

### 6.6 Recent Events

We have, in the course of investigating Tor metrics data with the tool detailed in this work, discovered and reported a number of significant Tor usage anomalies in countries including Ukraine, Israel, Bangladesh, UAE, and Turkmenistan. In some of these cases anomalies are due to filtering behaviour, such as Bangladesh's blocking of Facebook and chat applications in November 2015. In other cases the anomalies are due to external factors such as Ukraine's blocking of the popular Russian social networking site VKontakte in

[3]Two events corresponded to direct blocking of Tor bridge nodes, and these were identified as anomalous in the bridge usage statistics. All other anomalies were detected in normal Tor usage.



(a) Iran

(b) Turkmenistan

Relay Type — Relay — Bridge
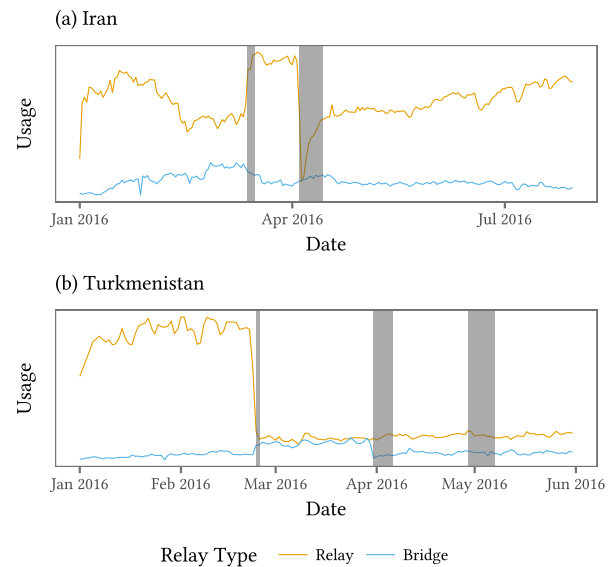
**Figure 2: Combined relay and bridge Tor usage anomalies.**

May 2017 [25] that led to a large spike in circumvention tool usage. Numerous other events have been detected, but space limitations prevent significant discussion of individual cases.

## 7 EXAMPLE RESULTS

Due to space constraints, we will not discuss specific cases in detail. This section shows a number of example outputs that highlight detected anomalies. As far as possible, we have extended the range of time shown in each plot to highlight that detected anomalies are not a frequent occurrence.

### 7.1 Most Anomalous Countries

Figure 1 illustrates the ten most anomalous countries according to their median absolute deviation from the median in the past year. Shaded regions denote periods of anomalous usage, according to our tool.

### 7.2 Combined Tor Metric Anomalies

Figure 2 highlights example combined anomalies that demonstrate periods in which Tor usage via normal relays and access via bridge nodes experienced simultaneous but opposing anomalies.

Over the period included in the available Tor metrics data, which covers late 2011 until the time of writing, our technique identified 485 anomalous periods in which both Tor usage and bridge usage were jointly anomalous, across 102 countries out of the total 251 for which Tor assigns usage statistics. This number is somewhat inflated due to the fact that a number of these anomalous periods are separated only by a small number of days and are likely the result of the same event.

Of these countries, Georgia had the highest number of combined detected anomalies, with 16 anomalous periods identified since 2011. The median number of anomalous periods over the set of all 102 countries that showed any anomalous behaviour was four.
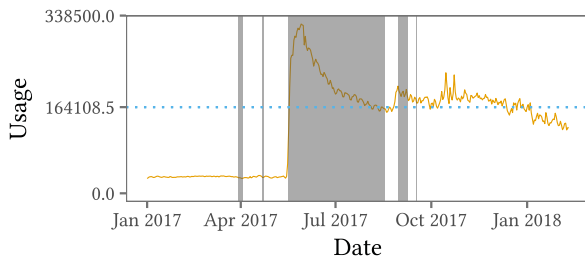
**Figure 3: Anomalous usage following Ukraine's ban on major Russian network services.**

It is possible that this number may increase if the combination of anomalous periods is made more flexible, as discussed in §4.4, however this demonstrates that events that exceed the threshold for combined anomalies are relatively rare.

### 7.3 Ukraine Russian Service Ban

In early May 2017 the Ukrainian government blocked a number of major Russian online services, used by a significant number of Ukrainian citizens, including social network sites VKontakte and Odnoklassniki, mail provider mail.ru, and Yandex, a major search engine[25]. Figure 3 shows a strong surge in Tor usage in the immediate aftermath of this, causing Ukraine to rise to the top of the daily anomaly rankings on the <name-redacted> mailing list. This example represents a significant anomaly in Tor usage related to blocking of standard internet services beyond Tor, and is in direct comparison to the Turkmenistan example of Figure 2 that highlights blocking of the Tor network itself.

### 8 DISCUSSION

The validation and results of §6 and §7 demonstrate that our approach is practically useful for identifying both Tor blocking and, more generally, for identifying periods of anomalous Tor usage. The highlighted anomalies detected by our approach are strong indicators of regions of likely interest to the internet filtering research and activist communities, and in particular in the combination of normal Tor and bridge node usage.

More directly, the experimental validation in the previous section demonstrates that our approach does detect a significant number of anomalies with varying magnitudes and durations.

### 9 FUTURE WORK

A main aspect of future work, for which these techniques were developed, will be to perform analysis on historical filtering behaviour and to maintain an ongoing watch for new potential filtering events. By combination with datasets such as Google's Global Database of Events, Language, and Tone (GDELT) [23], and through collaboration with researchers and activists, the authors hope to develop and maintain a contextualised time series of per-country filtering events for the benefit of future researchers.

Whilst the work presented here has focused on the application of our technique to Tor metrics data, the method is more generally applicable. Applying the techniques presented here to other data

sources is the most obvious direct extension to this work. We have made preliminary analyses based on data from Psiphon, CAIDA, Measurement Lab [13], and the Wikimedia Foundation, as well as evaluating data from the OONI Project [28] for its applicability in detecting filtering. Other data sources, such as social media, are also likely candidates for analysis.

Given the results of combining multiple Tor metrics, an interesting line of enquiry would be to investigate the speed with which users respond to filtering of Tor by adopting bridge nodes, and to understand the proportion of users that make this change. As more data sources are combined, further analysis of filtering's effects in different countries and under different conditions becomes possible.

### 10 CONCLUSIONS

We have developed a principal component analysis-based multivariate anomaly detection system to detect anomalous periods in per-country usage statistics of Tor metrics data. Our approach allows detection of per-country anomalies in time series that are non-stationary and that demonstrate significant seasonality. Our approach discounts global trends and even large-scale global events by considering individual countries' usage patterns as relative to that of others.

We have demonstrated the application of this tool to data from the Tor Project's metrics portal, showing that it provides a means to indicate potential censorship-related events, and others, at the global level. We have further shown that combining multiple metrics to identify jointly-anomalous periods can greatly improve the usefulness of the detected anomalies for identifying periods of direct blocking of Tor.

This work presents a generally applicable tool for detecting a broad class of internet filtering events on a global scale, without the need to focus on individual countries, and that dynamically adapts to changing patterns of usage. Countries exhibiting anomalous behaviour are automatically identified, and can be subjected to further, more targeted, investigation.

We have validated our approach both by evaluating detection rates of injected anomalies in a synthetically-generated time series, and demonstrated that our detection rates are significantly higher than those used in the existing anomaly detector used by the Tor project. Additionally, our tool provides useful ranking of anomalies according to strength, as well as highlighting anomalous periods rather than single-day events.

We have further evaluated our tool by successfully comparing detected anomalous periods with an external list of known Tor blocking events. This evaluation successfully identified each reported blocking event, supporting the tool's practical effectiveness in detecting real-world anomalies.

Using our approach, we have demonstrated that combining anomalies detected in multiple metrics can be an effective means to identify more targeted forms of anomaly that indicate filtering behaviour. Our initial combination of opposite-signed normal Tor usage and bridge node usage anomalies is a key step, but there are other behaviours that could be of specific interest; there is also significant potential for further combination with metrics from other tools and data sources.

Beyond the technique itself, the analyses presented in this work have identified several states that are known to engage in active filtering, but have also highlighted patterns of anomalous behaviour in several states that have not received significant attention from the internet censorship research community. Conducting more detailed investigations of these countries is a promising focus for future research.

Our anomaly detection tool is running actively on a nightly basis, with results output to a dedicated anomaly mailing list. This list has an audience amongst NGOs and research projects working in the field of investigating filtering and circumventing censorship, and has seen active use in detecting emerging real-world filtering events.

In addition to the underlying technique and tool developed to detect anomalous periods of behaviour, we have suggested, and provided initial evidence, that the use of the Tor metrics data, amongst other sources, is of use not only as an indicator of its own usage patterns, but as a practical proxy variable for a much wider class of political and social events. This presents significant potential for researchers, policy makers, and activists investigating global freedom of expression.

## 11 ACKNOWLEDGEMENTS

## REFERENCES

[1] Sadia Afroz and David Fifield. [n. d.]. Timeline of Tor Censorship. www1.icsi.berkeley.edu/~sadia/tor_timeline.pdf. ([n. d.]). Accessed 25th February, 2018.
[2] Anonymous. 2012. The Collateral Damage of Internet Censorship by DNS Injection. *SIGCOMM Comput. Commun. Rev.* 42, 3 (June 2012), 21–27. https://doi.org/10.1145/2317307.2317311
[3] Anonymous. 2014. Towards a Comprehensive Picture of the Great Firewall's DNS Censorship. In *4th USENIX Workshop on Free and Open Communications on the Internet (FOCI 14)*. USENIX Association, San Diego, CA. https://www.usenix.org/conference/foci14/workshop-program/presentation/anonymous
[4] Yana Breindl and Joss Wright. 2012. Internet Filtering in Liberal Democracies. In *Presented as part of the 2nd USENIX Workshop on Free and Open Communications on the Internet*. USENIX, Bellevue, WA. https://www.usenix.org/system/files/conference/foci12/breindl2012foci.pdf
[5] Richard Clayton, Steven J. Murdoch, and Robert N. M. Watson. 2006. Ignoring the Great Firewall of China. In *Proceedings of the 6th International Conference on Privacy Enhancing Technologies (PET'06)*. Springer-Verlag, Berlin, Heidelberg, 20–35. https://doi.org/10.1007/11957454_2
[6] R. B. Cleveland, W. S. Cleveland, J. E. McRae, and I. Terpenning. 1990. STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics* 6 (1990), 3–73.
[7] Jedidiah R. Crandall, Daniel Zinn, Michael Byrd, Earl Barr, and Rich East. 2007. ConceptDoppler: A Weather Tracker for Internet Censorship. Computer and Communications Security. http://www.cs.unm.edu/~
[8] George Danezis. 2011. *An anomaly-based censorship-detection system for Tor*. Technical Report. The Tor Project. https://research.torproject.org/techreports/detector-2011-09-09.pdf
[9] Alexander Darer, Oliver Farnan, and Joss Wright. 2017. FilteredWeb: A Framework for the Automated Search-Based Discovery of Blocked URLs. In *Network Traffic Measurement and Analysis*. IFIP. http://tma.ifip.org/wordpress/wp-content/uploads/2017/06/tma2017_paper32.pdf
[10] A. Darer, O. Farnan, and J. Wright. 2018. Automated Discovery of Internet Censorship by Web Crawling. *ArXiv e-prints* (April 2018). arXiv:cs.CY/1804.03056
[11] Ronald Deibert. 2007. *Access Denied: The Practice and Policy of Global Internet Filtering (Information Revolution and Global Politics Series)* (1 ed.). MIT Press. http://www.worldcat.org/isbn/0262541963
[12] Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. Tor: The Second-Generation Onion Router. In *IN PROCEEDINGS OF THE 13 TH USENIX SECURITY SYMPOSIUM*.
[13] Constantine Dovrolis, P. Krishna Gummadi, Aleksandar Kuzmanovic, and Sascha D. Meinrath. 2010. Measurement lab: overview and an invitation to the research community. *Computer Communication Review* 40, 3 (2010), 53–56. https://doi.org/10.1145/1823844.1823853
[14] Oliver Farnan, Alexander Darer, and Joss Wright. 2016. Poisoning the Well: Exploring the Great Firewall's Poisoned DNS Responses. In *Proceedings of the 2016 ACM on Workshop on Privacy in the Electronic Society*. ACM, 95–98.
[15] Nicolas Goix. 2016. How to Evaluate the Quality of Unsupervised Anomaly Detection Algorithms? (2016). arXiv:arXiv:1607.01152
[16] Ling Huang, Xuanlong Nguyen, Minos Garofalakis, and Joseph M. Hellerstein. 2007. Communication-efficient online detection of network-wide anomalies. In *In IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 134–142.
[17] P.J. Huber. 2004. *Robust Statistics*. Wiley. https://books.google.co.uk/books?id=e62RhdqIdMkC
[18] J. E. Jackson and G. S. Mudholkar. 1979. Control Procedures for Residuals Associated with Principal Component Analysis. *Technometrics* 21, 3 (1979), 341–349.
[19] I. T. Jolliffe. 2002. *Principal component analysis*. Springer, New York. http://link.springer.com/book/10.1007%2Fb98835
[20] Henry F. Kaiser. 1960. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* 20 (1960), 141–151. Issue 1. https://doi.org/10.1177/001316446002000116
[21] Gary King, Jennifer Pan, and Margaret E. Roberts. 2013. How Censorship in China Allows Government Criticism but Silences Collective Expression. *American Political Science Review* 107 (2013), 1–18.
[22] Anukool Lakhina, Mark Crovella, and Christophe Diot. 2004. Diagnosing Network-Wide Traffic Anomalies. In *Proceedings of ACM SIG-COMM 2004*. 219–230. http://www.cs.bu.edu/faculty/crovella/paper-archive/sigc04-network-wide-anomalies.pdf
[23] Kalev Leetaru and Philip A. Schrodt. 2013. GDELT: Global data on events, location, and tone. *ISA Annual Convention* (2013).
[24] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49, 4 (2013), 764 – 766. https://doi.org/10.1016/j.jesp.2013.03.013
[25] Alec Luhn. 2017. Ukraine blocks popular social networks as part of sanctions on Russia. (May 2017). https://www.theguardian.com/world/2017/may/16/ukraine-blocks-popular-russian-websites-kremlin-role-war
[26] Hooman Mohajeri Moghaddam, Baiyu Li, Mohammad Derakhshani, and Ian Goldberg. 2012. SkypeMorph: Protocol Obfuscation for Tor Bridges. In *Proceedings of the 19th ACM conference on Computer and Communications Security (CCS 2012)*.
[27] Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philos. Mag.* 2, 6 (1901), 559–572.
[28] The OONI Project. [n. d.]. The Open Observatory of Network Interference. https://ooni.torproject.org/. ([n. d.]). Accessed 25th February, 2018.
[29] The Tor Project. [n. d.]. Tor Metrics Portal. https://metrics.torproject.org/. ([n. d.]). Accessed 25th February, 2018.
[30] The Tor Project. [n. d.]. Tor Metrics: Questions and answers about user statistics. https://gitweb.torproject.org/metrics-web.git/tree/doc/users-q-and-a.txt. ([n. d.]). Accessed 25th February, 2018.
[31] Haakon Ringberg, Augustin Soule, Jennifer Rexford, and Christophe Diot. 2007. Sensitivity of PCA for traffic anomaly detection. In *SIGMETRICS '07: Proceedings of the 2007 ACM SIGMETRICS international conference on Measurement and modeling of computer systems*. ACM Press, New York, NY, USA, 109–120. https://doi.org/10.1145/1254882.1254895
[32] David Robinson, Harlan Yu, and Anne An. 2013. *Collateral Freedom: A Snapshot of Chinese Users Circumventing Censorship*. Technical Report.
[33] Fatemeh Shirazi, Claudia Diaz, and Joss Wright. 2015. Towards Measuring Resilience in Anonymous Communication Networks. In *Proceedings of the 14th ACM Workshop on Privacy in the Electronic Society (WPES '15)*. ACM, New York, NY, USA, 95–99. https://doi.org/10.1145/2808138.2808152
[34] Augustin Soule, Kavé Salamatian, and Nina Taft. 2005. Combining filtering and statistical methods for anomaly detection. In *In Proceedings of IMC*.
[35] Graham Upton and Ian Cook. 2002. *Oxford dictionary of statistics*. Oxford university press Oxford, UK.
[36] Zachary Weinberg, Jeffrey Wang, Vinod Yegneswaran, Linda Briesemeister, Steven Cheung, Frank Wang, and Dan Boneh. 2012. StegoTorus: A Camouflage Proxy for the Tor Anonymity System. In *Proceedings of the 19th ACM conference on Computer and Communications Security (CCS 2012)*.
[37] Joss Wright. 2014. Regional variation in Chinese internet filtering. *Information, Communication & Society* 17, 1 (2014), 121–141. https://doi.org/10.1080/1369118X.2013.853818 arXiv:http://dx.doi.org/10.1080/1369118X.2013.853818
[38] Joss Wright, Tulio de Souza, and Ian Brown. 2011. Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics. In *Free and Open Communications on the Internet*. USENIX, San Francisco, CA, USA. http://static.usenix.org/event/foci11/tech/final_files/Wright.pdf
[39] Yin Zhang, Zihui Ge, Albert Greenberg, and Matthew Roughan. 2005. Network anomography. In *In IMC*.