Detecting Internet Filtering from Geographic Time Series

Joss Wright
Oxford Internet Institute
University of Oxford
1 St. Giles
Oxford. OX1 3JS. UK
joss.wright@oii.ox.ac.uk

Alexander Darer Oliver Farnan
Cyber Security Centre
Department of Computer Science
University of Oxford
Wolfson Building, Parks Road
Oxford OX1 3QD. UK
alexander.darer@linacre.ox.ac.uk,
oliver.farnan@balliol.ox.ac.uk

ABSTRACT

We propose an approach based on principle component analysis to identify per-country anomalous periods in traffic usage as a means to detect internet filtering, and demonstrate the applicability of this approach with global usage statistics from the Tor Project. In contrast to previous country-specific investigations, our techniques use deviation from global patterns of usage to identify countries straying from predicted behaviour, allowing the identification of periods of filtering and related events in any country for which usage statistics exist. To our knowledge the work presented here is the first automated approach to detecting internet filtering at a global scale.

We demonstrate the applicability of our approach by identifying known historical filtering events as well as events injected synthetically into a dataset, and evaluate the sensitivity of this technique against different classes of censorship events. Importantly, our results show that usage of circumvention tools, such as those provided by the Tor Project, act not only as direct indicators of network censorship but also as a meaningful proxy variable for related events such as protests in which internet use is restricted.

Categories and Subject Descriptors

D.4.6 [**Security and Protection**]: Information flow controls; K.4.1 [**Computers and Society**]: Public Policy Issues—*Use/abuse of power*

General Terms

Security, Measurement

Keywords

Filtering, censorship, principal component analysis

1. INTRODUCTION

Nation states, and others, have increasingly begun to employ internet filtering as a means of controlling access to information, and as a tool to limit certain forms of social and political organisation. Given the central role that the internet plays in communications for a large and increasing 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 infrastructures provide useful data for identifying tools, techniques, and limitations of these entities, it is also necessary to incorporate social and political dimensions toward understanding the motivations of censors and the reaction of populations to filtering.

Much existing research into internet filtering has focused either on observing the specific practices of states already known engage in filtering, or in the development of circumvention tools. Whilst multilateral studies of censorship have been conducted, most notably the seminal work of Deibert et al. [7], these approaches have typically amalgamated country-specific investigations. In the case of [7], 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.

A core problem in research into network filtering is the discovery of filtering events, and their targets. To date, most technical analyses of filtering have focused on known filtering countries, or on word of mouth or media reports regarding new events. The work presented here provides a means to alert researchers and activists to potential developing events that may otherwise have been missed.

The work presented here is motivated by a desire to identify global patterns of internet filtering[1] through technical network measurements, and to link these events to their broader social and political context. In the current work we focus on the former whilst considering the means to achieve the latter.

The main contributions of this work are:

- an automated approach to detecting worldwide filtering-related events from public data sources;
- the application of this approach to a real-world dataset, in this case the Tor project's metrics data;
- a validation of the approach through detection of known real-world censorship events, as well as test events artificially injected into the data;

[1]The term *censorship* has become commonly used in the field to refer to manipulation of network traffic for social or political purposes. To avoid making moral judgements on the nature of particular events, we prefer the more neutral terms *filtering*, *blocking*, or *manipulation*.

- a tool for producing an ongoing global overview of anomalous activity showing both "known" filtering nations, as well as less well-known examples.

As an addition to the above, we demonstrate that usage data for tools such as the Tor project, along with others, can act as practical and effective proxy variables for detecting social and political events around the world.

1.1 Problem and Approach

When an entity chooses to filter or block certain types of content, the resulting traffic exhibits detectable anomalous patterns of traffic. In a complex 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.

We seek to detect anomalous behavior resulting, directly or indirectly, from filtering practices. It is substantially harder to assert any form of causal relationship between an observed anomaly and a technical intervention, however we do seek to link real-world events to anomalies that appear simultaneously in our observations.

We consider the problem of detecting internet filtering from the perspective of anomalous behaviour in traffic flows. Key to our approach is the modelling of each country's behaviour in terms of its relative behaviour to other countries, with time periods judged as anomalous if they deviate from these patterns.

The intuition behind this work is that, in absence of interference, certain network statistics for countries are likely to fall into one of a small number of classes, and that this classification will remain relatively stable over time. If an individual country begins to deviate from its prior classification, it suggests external interference in the flow of traffic.

Through principal component analysis we construct an approximation of observed traffic that accounts for the majority of the variance in the dataset. Differencing the observation against the approximate model produces residuals that represent localised anomalies outside of them modelled behaviour.

This approach, which splits the set of principal components into *normal* and *anomalous* subspaces, was initially proposed by Jackson and Mudholkar [13] for application in industrial process control. It was later employed by Lakhina et al. [18] to detect network-wide traffic anomalies from per-link data in high-performance networks.

We discuss principal component analysis in §4, and the details of the PCA-subspace methodology in §4.2.

2. EXISTING WORK

Internet filtering, and more broadly censorship, has received attention from a number of fields of study. Technical research has focused on analysing mechanisms of censorship, and the development of censorship circumvention approaches. At the same time, researchers from the social sciences have investigated the motivations of censors, and the legal, economic, and societal effects of such systems. We argue that a holistic understanding of internet filtering, and the interaction between technical capabilities and human factors, is necessary to influence its future development.

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.[4], who isoolated 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 [35] demonstrated that DNS censorship was experienced differently in different regions within China, with significant variation in the nature of the DNS poisoning seen across the country.

Crandall et al.[5] 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 "Concept Doppler" approach aimed 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.

Perhaps the most comprehensive study to date of global filtering practices is given by Deibert et al. [7]. 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.

Whilst the approach of [7] is far more comprehensive in its scope than other studies, it relies on a largely manual process that would require significant ongoing resources to maintain as a continuous overview of the state of internet filtering.

More recently, the OONI project[22] has developed a platform to conduct a variety of tests for network filtering, with the intention of building a global network of volunteer operators. Whilst the project has produced a number of useful analyses of filtering events, the project is still in its infancy and is subject to significant ethical concerns with respect to the risks to participants in the network. Ethical issues in the research of network filtering have been discussed by Wright [36], and we discuss them in greater detail in §5.2.

Certain types of filtering act less on network traffic in transit, and more on application level or social filtering. King et al. [17] 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.

From the perspective of detecting anomalies in traffic data, the most similar approach to our own is that taken by Lakhina

et al. [18], who make use of principal component analysis to detect network-wide anomalies in high-speed networks via data gathered from a restricted set of link-level observation points. This is in contrast to our approach, which uses global patterns to identify per-country anomalies.

Several other works have extended or expanded aspects of this approach, notably [31], [38], and [12]. 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 specifically focuses on per-observation anomalies across a dataset with several hundred dimensions in order to highlight individual states displaying anomalous behaviour.

In addition, to counter the effects of significant long-term shifts in the underlying data, we also perform analyses over a rolling time window to cancel out large-scale developing patterns. This is discussed in greater detail in §4.3.

We will now examine various aspects of our approach.

3. TOR

Tor [9] 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 significant ongoing research work[20, 34] is aimed at ensuring that Tor 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 [29], 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 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.

Gathering statistics in the Tor network is inherently a difficult task, as the anonymity properties of the network preclude the identification of individual users via their connections. Instead, the Tor Project make use of client requests to central *directory authorities* to estimate overall user numbers.

When a Tor client connects to the network, or desires to refresh its view of the network, it connects to one of a small number of directory authorities that store a list of all active Tor relays. These directory servers count the number of requests received each day, and geolocate the requesting IP addresses[19]. The resulting aggregate request statistics are passed to a centralised Tor *metrics portal*, from which data is freely available[26].

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[27]. 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[2]. From this, we remove those countries whose Tor usage is sufficiently low as to make statistical analysis unreliable. In practice, we remove all those countries whose usage never rises above 1000 daily users.

In the next section, we describe principal component analysis, and show how this is applied to the Tor metrics data.

4. PRINCIPAL COMPONENT ANALYSIS

Principal component analysis was developed by Pearson[21] 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*.

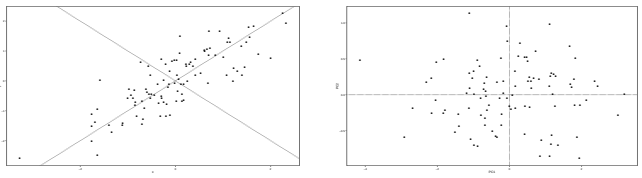
Principal component analysis transforms a dataset to a new coordinate system in which orthogonal axes, the *principal components*, successively represent the direction of greatest variance in the data. The majority of the variance lies along the first coordinate axis, with the orthogonal coordinate axis that describes the next highest degree of variance being the second principal component. This process continues until the data is fully described by a set of principal components of equal cardinality to the original set of dimensions. Figure 1 demonstrates this transformation on a simple two-dimensional dataset.

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.

Before application of principal component analysis, data is typically transformed on a per-dimension basis, to be zero-mean; and scaled to unit variance to ensure that each dimension contributes equally to the result. See [14] for a detailed treatment of principal component analysis and the various choices and compromises to be made when applying the technique.

4.1 Application to Tor Metrics Data

[2]Earlier data is available, but was gathered using a different methodology and has not yet been analysed with the techniques presented in this paper.



(a) Data with first two principal components indicated.

(b) Data rotated onto principal component axes.

Figure 1: Example principal component transform.

Our set of observations can be considered as an $m \times n$ matrix X, in which the n columns correspond to individual countries, and the m rows correspond to date-indexed observations. Each row x_i of X represents a point in n-dimensional space. As noted above, each column is transformed to have a zero mean.

The purpose of a principal components analysis is to form a projection of X into a p-dimensional subspace, where $p \leq n$, such that the sum of the squares of the distances between points and their projection in the subspace are minimized. There are various ways to achieve this, but the most well-known is by maximising the covariance matrix of the projected data.

For our matrix of observations, X, we wish to calculate a set of principal components W, where $|W| = n$. This is achieved by an iterative procedure. The first principal component is defined as the vector that maximises the variance in X:

$$\mathbf{w}_1 = \arg\max_{\|\mathbf{w}\|=1} \{\|\mathbf{X}\mathbf{w}\|^2\}$$

Subsequent components are those vectors that account for the maximum variance in the residuals between X and the projection of X onto the current set of components. The residuals from the first $k-1$ principal components can be expressed as:

$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{X}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}$$

Thus, the k^{th} principal component is given by:

$$\mathbf{w}_k = \arg\max_{\|\mathbf{w}\|=1} \left\{ \|\hat{\mathbf{X}}_k\mathbf{w}\|^2 \right\}$$

We now discuss the application of this approach to detecting anomalies in time series data.

4.2 Subspace Analysis

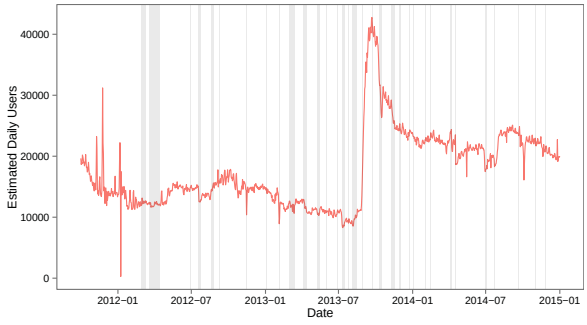
As noted above, principal component analysis is a means to project n-dimensional data onto a p-dimensional subspace, which accounts for the majority of the variance in the original dataset. As we are interested in anomalies in data rather than underlying trends, it is possible to consider principal component anaysis as dividing the data into two subspaces: a modelled *normal* subspace that accounts for overall trends, and an *anomalous* subspace that is not accounted for by the selected components.

Subspace analysis focuses on the anomalous subspace by inverting the transform using a restricted subset of the principal components. This results in an approximation of the data in which the residual errors are the key item of interest, similarly to the original iterative step of the principal component calculation.

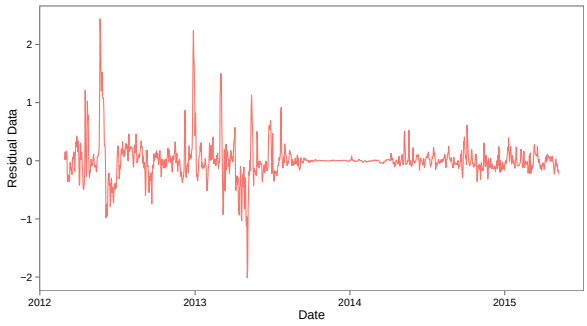
$$\hat{\mathbf{X}}_k = \mathbf{X} - \sum_{i=1}^{k-1} \mathbf{X}\mathbf{w}_i\mathbf{w}_i^{\mathrm{T}}$$

In contrast to a reduced-dimensional approximation of the data, we are left with a set of *residuals* that express variances in the data not captured by the chosen set of principal components. We scale the residual errors in proportion to the original data to produce a set of proportional residuals for each of the n countries.

To illustrate, Figure 2 demonstrates overall usage patterns for China, against calculated residuals.



(a) Overall Tor usage for China, with anomalous periods highlighted.



(b) Proportional residuals from Chinese usage projected from 12 principal components.

Figure 2: Chinese usage statistics and calculated residuals over entire dataset.

4.3 Rolling Analysis

As discussed in §1.1, we are interested in evaluating the deviation of individual countries from their predicted patterns over time. In order to highlight developing patterns we therefore do not calculate principal components over the entire time series, which would tend to obscure developing anomalies, but instead perform a rolling principal component analysis over smaller time windows within the series.

A further justification for this approach is that principal component analysis treats each observation within a time window of X as a point in $\mathbb{R}^n$, but does not capture the sequential relationship between successive observations. By reducing the time window of observations we focus the analysis, on each successive window, on observations occurring temporally close to each other. This allows the analysis to ignore past and future relative shifts in trends.

We conduct the calculation of residuals only on the final row of each window, corresponding to the most recent observation. Residual anomalies for each day are threfore expressed according to the principal components for a prior fixed-length time window. As a reasonable compromise between short-term sensitivity of results and the dimensionality of the data, all experiments in §6 make use of a rolling 180-day time window.

This approach makes no guarantee that the specific form of the principal components calculated over each time window are consistent. Whilst the most significant principal components are likely to be broadly similar across time windows, less significant principal components are likely to be differing linear combinations of observations. As we are interested in the aggregated principal components in the anomalous subspace, however, this is unlikey to have significant effects on our results unless a major network event pollutes the normal subspace, as discussed in Ringberg et al.[28].

4.4 Selection of Components

For many applications, a relatively small number of principal components will be sufficient to describe a large proportion of the variance in a dataset. Lakhina et al. [18] note that multivariate network traffic time series often exhibit low overall dimensionality, and are the well described by small numbers of principal components.

One accepted technique when selecting an appropriate number of components is to calculate the proportion of variance explained by each component, and identify the "elbow" point at which successive principal components add comparitively little extra information about of the dataset. This is typically achieved through visual inspection of a scree plot of principal component variance scores.

A second common approach is to make use of Kaiser's criterion[15] to select only those principal components with eigenvalue greater than 1, representing those components that provide more information than a single average component.

It is worth noting that most approaches to selection of principal components are focused on an appropriate tradeoff between minimising the number of dimensions to be used in further analysis and maximising the explained variance in the data. As the approach taken here is explicitly focused on the residual subspace, however, there is an argument that erring on the side of less principal components, and thus preserving a greater number of residuals, may be of use.

Figure 3 shows the variance explained by differing selections of principle components over the Tor metrics data. In line with Kaiser's criterion, and keeping a minimum variance explanation of 0.8 over the dataset, the experiments conducted in the remainder of this work make use of twelve principal components. It is worth noting, however, that this is a key tuning parameter of the technique, and that a different set of principal components may be much more appropriate for other data sources.

Having discussed the various aspects of the approach, we now detail the specific methodology employed in the experiments.

5. METHODOLOGY

The previous section detailed the underlying approach proposed for detecting filtering as per-country anomalies.

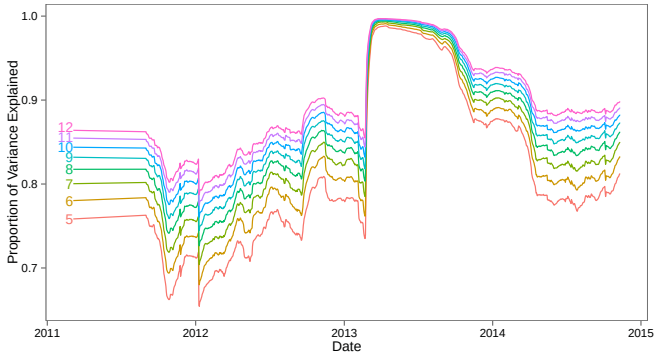


Figure 3: Variance explained by different numbers of principal components over series.

We will now discuss how this approach can be applied in practice to the Tor metrics data as the basis for the experimental results shown in §6.

As noted above, our approach differs from most other subspace analyses in that we focus specifically on anomalies appearing in individual dimensions of the time series, rather than reconstituting system-wide anomalies from a small number of observation points.

Crucially, our approach eliminates global trends in the set of observations by considering only anomalies that occur within individual time series outside of global trends. This, as we show in §7.2 allows us to discount significant large-scale events in global Tor usage.

5.1 Censorship Events as Usage Anomalies

It is fundamental to the approach described here that usage anomalies in appropriately selected traffic can be indicative of filtering events. In our experiments we make use of the Tor metrics data, from which we aim to identify two forms of event: firstly, direct blocking of the Tor network resulting in anomalies in Tor usage; secondly, changing characteristics of Tor usage in response to exogeneous factors. The censorship of a major international website, such as YouTube in Turkey, has the potential to drive a noticeable minority of users to Tor, and as such Tor becomes a useful *proxy variable*[33] for a broader class of filtering behaviour.

The Tor project currently maintain a censorship flagging tool, as described by Danezis[6], that assesses the ratio of daily connections on a per-country basis over a seven-day time period. If a country's ratio of users falls outside of the globally-observed usage trends for Tor, based on the fifty largest Tor-using countries, which are largely presumed not to filter connections. This tool is focused on detecting Tor anomalies, and there is no explicit consideration of external events that may also affect Tor usage. The work presented here is intended, in part, as a replacement censorship detector for use by the Tor project. Our experimental code is already usable for this purpose, and will shortly be made available for ongoing flagging of anomalous countries and analysis of historical trends.

Direct effects on usage related to censorship events may therefore be expected to fall into three major classes:

- Sharp drops in Tor usage, indicative of direct blocking of the network.

- Gradual drops in usage, which maybe indicative of more subtle filtering but may also represent the relaxation of other practices, resulting in a decrease in users that considering Tor a necessity.
- Sharp spikes in usage, indicative of the blocking of key resources such as major websites that drive a large number of users to the Tor network.

All of these classes may display greater or lesser effects on the size of the residuals from the subspace analysis. As such, when judging a particular effect as anomalous, the nature of the specific class of anomaly should be considered; when attempting to isolate direct blocking of Tor large residuals should be expected, whilst attempts to explore the effects of political debates surrounding censorship are likely to result in much more subtle anomalies.

An assumption of our approach is that relative patterns of Tor usage are largely consistent worldwide, and this is supported by our experimental results. We reject manual selection of "stable" countries as a baseline for Tor usage, and instead employ principal components analysis to identify trends directly.

In the remainder of this work, we conduct a series of analyses of the Tor metrics data to identify anomalous countries, and specific periods of anomalous behaviour. We evaluate our approach against injected anomalies in a single country in the data, and demonstrate the limitations of the anomaly detection when anomalies are small-scale and gradual.

5.2 Ethics

Conducting research into network filtering presents a number of ethical issues, as discussed by Wright [36]. 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. Firstly, approaches to broad-scale network testing preclude intensive training of research participants due to the time and resource constraints, and providing a disclaimer warning of possible risks is not regarded by most ethics bodies as an appropriate level of informed consent.

Secondly, the possible implications of a user's device being identified as conducting filtering tests are potentially severe, and hard to quantify. Whilst an attempt to perform a DNS request for a suspected blocked social networking site may be considered relatively innocuous by the remote researcher, it may be far more significant to the censor. More seriously, testing for blocking of socially unacceptable or illegal content, such as hate speech or images of sexual abuse, carry more obvious risks to the participant.

As such, 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.

5.3 Tunable Parameters

Our approach to analysis and detection of anomalies relies on a number of parameters that affect the nature of detected events. As suggested above, different classes of anomaly may result in more or less subtle effects in the data. At the same time, behaviours that develop gradually over a long time period are likely to be harder to detect than immediate shifts in traffic.

The main parameters are:

- **Size of the normal subspace**

  The number of principal components selected can have significant effects on which effects are accounted for in the model. As discussed in [28], an overly generous selection of principal components may draw large-scale anomalous events into the normal subspace, and thus produce a false negative. Figure 3 shows that, during a large-scale attack on the Tor network in August 2013, the proportion of variance explained by even small numbers of principal components rose sharply towards 1. This will be discussed in more detail in §7.2.

- **Anomalous Threshold**

  The size of the residual between the original data and the reconstructed data using the normal subspace is a direct measure of the scale of the anomalous behaviour, and anomalous periods are classified according to this residual rising above a given threshold. For direct blocking of Tor, or prominent events such as blocking of major websites, a relatively large threshold may be appropriate. For more subtle effects, the threshold for considering behaviour anomalous may need to be much lower.

- **Width of time window**

  To prevent long-term per-country trends from being incorporated into the normal subspace, we calcuate principal components over a restricted time window of the original data. Shorter time periods will be more sensitive to changing trends in the dataset, as new data will contribute comparitively more to the overall trend. Longer time windows may, however, produce more robust results and reduce potential false positives.

- **Selection of countries**

  Our approach is intended, and functions effectively, on a global scale. By reducing the set of countries of interest, however, it may be possible to focus on specific regions of the world. This would remove a level of automation from the approach, as countries would necessarily have to be carefully chosen to ensure that their correlations are not overly affected by other countries outside the dataset, but could provide advantages for more detailed regional studies.

With these parameters in mind, we now describe the results of analysing the Tor metrics dataset.

6. EXPERIMENTAL RESULTS

We analyse data from the Tor metrics portal to identify anomalous periods, and to isolate those countries that consistently exhibit unusual behaviour. Our results show that a small number of countries present consistently anomalous

Country	Median Residual
People's Republic of China	0.07779491
South Africa	0.06903181
Bangladesh	0.05099613
Islamic Republic of Iran	0.04977244
Syrian Arab Republic	0.04800895
Mongolia	0.04688284
India	0.04589252
Nepal	0.04389294
Senegal	0.04223453
Republic of Moldova	0.04125283

Figure 5: Ten most anomalous countries by mean residual score.

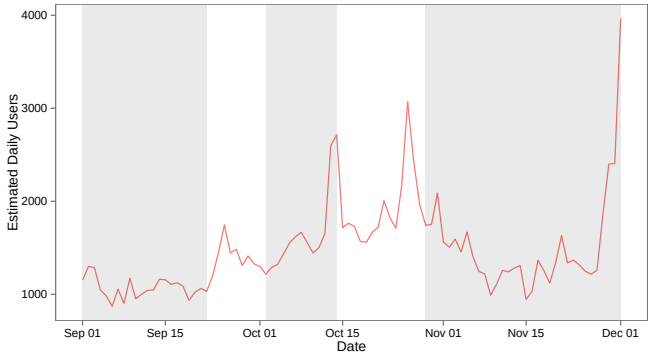


Figure 6: Highlighted anomalies before and during the Chinese Communist Party Congress in November 2012.

behaviour over the entire period, whilst others demonstrate more punctuated anomalies.

Following our overall analysis, we isolate particular countries that are known to have experienced unusual periods of internet usage, and examine the effects on Tor usage in these countries during suspect periods.

To demonstrate the applicability of the approach against a known ground truth, we introduced two differing periods of anomalous behaviour into Belgium's usage statistics.

Unless otherwise stated, all experiments were performed with 12 principal components in the normal subspace, a 180-day model time window, and a threshold of 0.1 as the proportional residual score to indicate an anomaly.

Note also that those countries whose usage remained below 1000 daily users over the entire period of observations are excluded from the analysis. This removes 124 countries with current data, the majority of which are located in Africa.

6.1 Most Anomalous Countries

Figure 4 illustrates the ten most anomalous countries according to their median residual threshold[3]. In all diagrams, shaded regions denote detected anomalous periods.

In order, the ten highest-scoring countries over the period from September 2011 until March 2015, and their residuals, are shown in Figure 5. For this set of ten countries, the median residual score was ≈0.047, compared with a median score of ≈0.025 over all countries. Scores within individual time windows vary considerably.

While this list verifies some prior assumptions regarding expected states, it is also surprising to note that South Africa scores highly by this metric. According to Deibert et al.[7], South Africa shows no evidence of filtering; nor has it had significant political or social events that immediately suggest such anomalous behaviour. While the possibility exists that this is a significant false positive in our results, it is perhaps the most interesting immediate avenue of future exploration.

6.2 Detection of Known Events

Having calculated anomalous statistics over the historical Tor data set, we now aim to validate our approach by comparing detected anomalies 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.

We are wary of taking an overly confirmatory approach to the results of these experiments. With sufficiently low thresholds for consideration of residuals, it would be possible to identify seeming anomalies at almost any point in any country. Positive correlation between our results and known events can only suggest a potential effect for further investigation. Equally, an anomalous period should not be interpreted as anything more than evidence that a particular state at a particular period may warrant further investigation.

With that caveat in mind, we now examine particular cases of interest.

6.2.1 *China*

China is known to be arguably the most extensive and significant filtering infrastructure on the modern internet. Filtering of political and social content is widespread, and is carried out through both automated network tools and an extensive manual human effort. In particular, content related to Tibetan independence is widely blocked.

In-country reports[24, 30] have claimed a particular intensification of filtering efforts surrounding the 2012 18th National Congress of the Communist Party of China, a five-yearly event in which leadership changes occur in the ruling Communist Party of China.

Figure 6 shows Chinese Tor usage data leading up to, and during, the National Congress, which began on the 8th November 2012. Anomalies are clearly shown during this period.

This result is perhaps less striking than other experiments due to the high degree of anomalies experienced in Chinese usage, as shown for the overall series in Figure 2a.

6.2.2 *Iran*

Similarly to China, Iran is know to engage in active filtering of a broad category of internet content, including direct blocking of the Tor network. The most well-known period of this blocking occurred in 2011, which falls before the earliest available observations[4].

Despite this, Iran has one of the highest internet penetration rates in the region: 42 million internet users, 53.3% of

[3]We make use of the median in order to provide a level of robustness against outliers in the calculated residuals.

[4]Earlier data exists but was sadly unavailable at the time of writing.

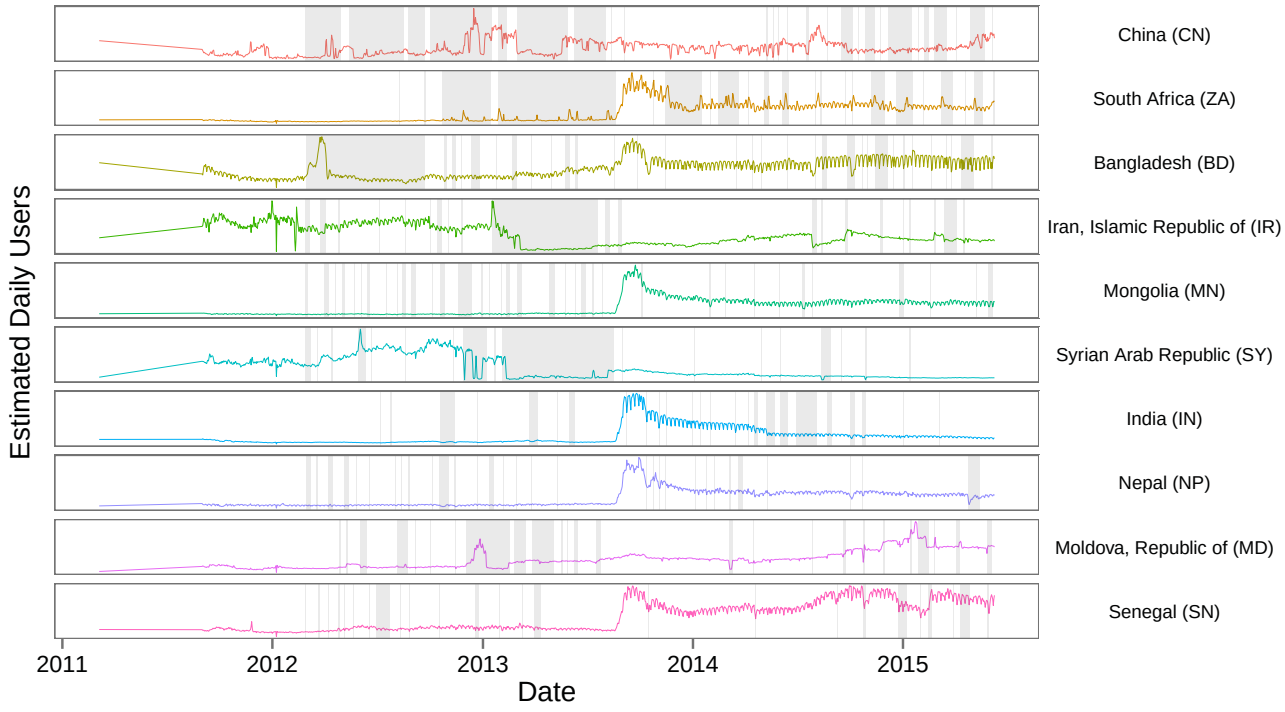


Figure 4: Ten most anomalous countries according to median residual score over observed time period.

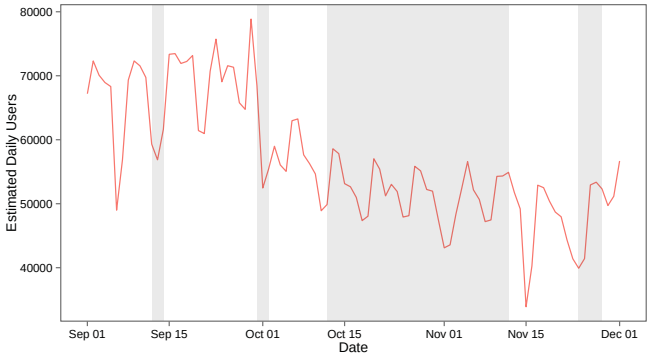


Figure 7: Iranian Tor usage and highlighted anomalies during the currency protests of 2012.

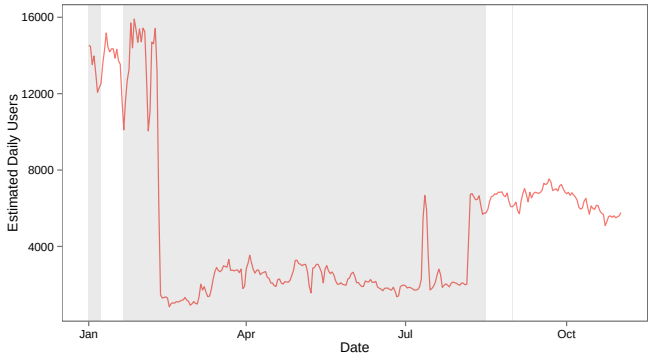


Figure 8: Syrian Tor usage during 2013.

the population, have access to the internet as of the 30th June 2012 [32].

In relation to ongoing economic sanctions, Iran experienced currency protests during October 2012 [25, 16], during which time observers noted significant bandwidth throttling, particularly of connections to international news and media outlets. This was in line with earlier research by Anderson [1] that had identified this mechanism in use during the contested Iranian presidential election in 2009.

Figure 7 shows Iranian Tor usage during this time period.

6.2.3 *Syria*

With an ongoing civil war in the Syrian Arab Republic, internet disruptions are relatively common, as is political protest. Before the political upheaval the government was already known to engage in active internet surveillance and filtering, which has reportedly intensified in recent years [23].

A country-wide disabaling of the internet was reported in May 2013. Such a dramatic event is clearly identified in Figure 8, and can easily be verified by inspection of the usage statistics.

6.2.4 *Turkey*

In late May 2013, Turkish activists conducting a sit-in to protest against urban development of the Taksim Gezi Park in Istanbul were forcefully evicted from the park. This event led to a series of demonstrations and protests concerning ongoing controversial topics in Turkish public life, of which regular blocking of websites such as Twitter and

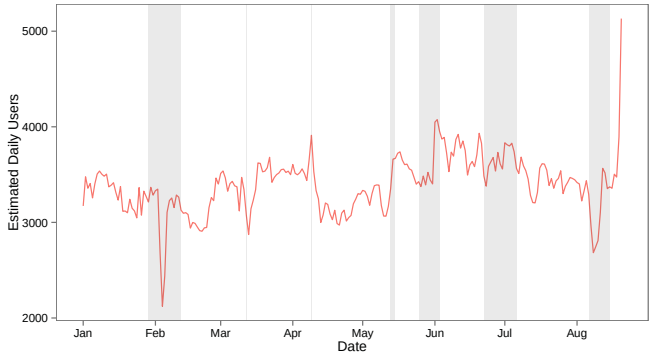


Figure 9: Turkish Tor usage before and during the Gezi Park protests.

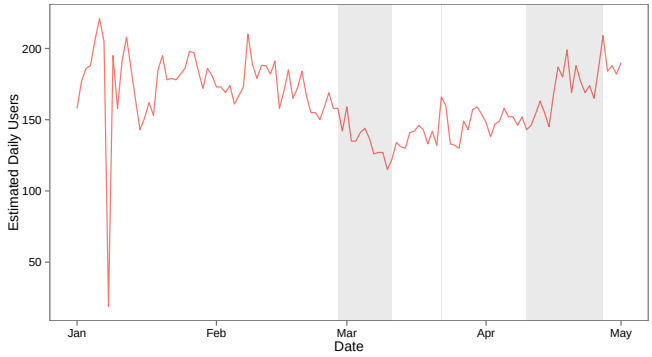


Figure 10: Increases in Palestinian Tor usage after government block political websites.

YouTube featured. The protests continued during the summer of 2013, before eventually dying down during August of that year.

While there were accusations of general media censorship of the protests, there was no overt blocking of internet traffic during the period.

Figure 9 shows identified anomalies in Turkish Tor usage during the period of the protests. As might be expected, this is initially represented by a spike in Tor usage that slowly declines over the summer, before finally dropping significantly in August. Coincidentally, this occurred directly prior to a large global spike in Tor usage, reflected at the extreme right of Figure9, which is consequentlly not identified as anomalous. This is discussed further in §7.2.

6.2.5 *Palestine*

In April 2012, the Ma'an News Agency, through an investigation conducted in collaboration with the OONI Project [22], reported that the Palestinian Authority had ordered Hadara, the state's main internet service provider, to block eight news outlets that were known to be critical of the government.

The decision was widely reported [37, 11], and received significant attention both locally and internationally. The corresponding period of Tor usage, with anomalous periods highlighted can be seen in Figure 10.

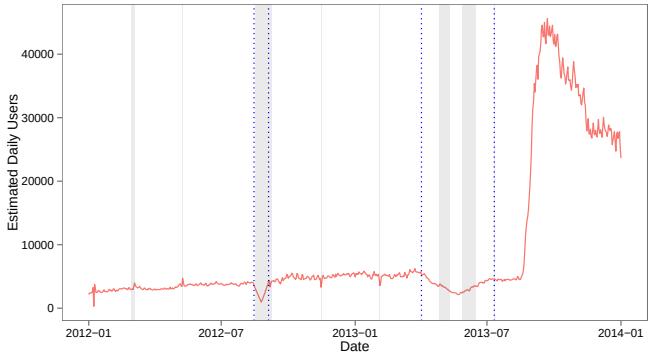


Figure 11: Belgian Tor usage with injected anomalies. Dotted lines show true position of injected anomalies.

7. ANALYSIS

Having demonstrated detection of real world anomalous periods, we examine here the effectiveness of the technique in a more controlled setting, and comment on the likelihood of false detection rates when analysing data.

7.1 Detection of Injected Events

The above events provide evidence for the effectiveness not only of our approach to detecting anomalies from Tor's metrics data, but also to Tor's previously unrecognised potential as an indicator for political and social events occurring around the world.

Due to the nature of these events, however, it is challenging to validate our approach against an objective list of filtering events; no such list exists, nor can events be labelled with certainty in the general case.

We here show results for a country that, on the whole, neither demonstrates nor is expected to demonstrate significant anomalous traffic patterns. By injecting two differing anomalies into the traffic, we demonstrate that the results are clearly distinguished by the techniques described in this work.

Figure 11 demonstrates detected anomalies injected into Belgian data, showing the effects of unusual traffic patterns in an otherwise 'normal' country.

7.1.1 *Sharp Drop*

The first of the two injected anomalies in the Belgian Tor usage statistics aims to model a short-term blockage or interference with the Tor network. The event was injected during August 2012, and comprises a twenty-day anomalous period in which usage steadily declines by 7%, along with a small amount of gaussian white noise, before returning to its previous level.

As can be seen in the first of the two anomalous periods highlighted in Figure 11, this relatively sharp anomaly is detected very quickly at a threshold of 0.1 for the proportional residual threshold.

The anomalous period began on the 15th August 2012; the identified anomalous period began the following day, resulting in a single day's lag. The true anomalous period lasted until the 4th September, and our model again follows this with one day's lag.

7.1.2 *Slow Decline*

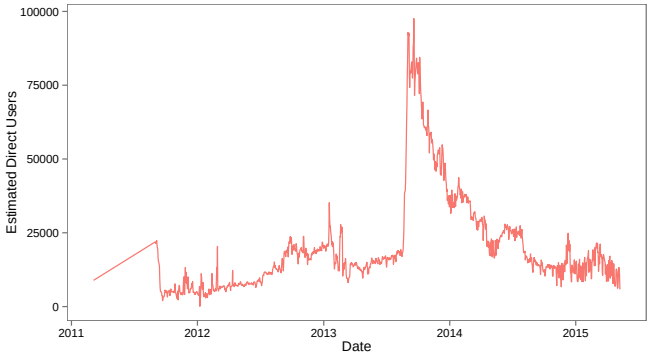


Figure 12: Aggregate global Tor usage showing Sefnit botnet spike in August 2013.

In contrast to the relatively sharp drop in the previous anomaly, we also injected a drop of slightly lower magnitude, with an overall decrease of 6% of the total volume, occurring over a 100 day time window.

While this alteration to the traffic was flagged as anomalous by our technique, it highlighted the limitations of the approach in continuing to regard as anomalous longer trends that deviate from earlier models.

The second injected anomaly began on the 2nd April 2013, and was initially detected at the 0.1 residual threshold level on the 27th April – a lag of 25 days.

More importantly, the model ceased to regard this behaviour as anomalous by the 11th May, leaving a 17-day gap before detecting the returning trend as anomalous on the 28th May. This second detected anomalous period within the overall anomaly lasted until the 15th June, 26 days before the true end of the anomalous period.

This slow decline highlights that more subtle effects on usage patterns cannot be captured in the same way as more significant effects. It is worth noting that the slow decline would have been detected much more effectively with a reduced residual threshold for tagging a period as anomalous, however this would have increased the likelihood of false positives at other points in the series.

Ultimately, these two example demonstrate that, while relatively small anomalies can be effectively detected using this approach, different classes of anomaly require appropriate selection of the tuning parameters discussed in §5.3.

7.2 Discounting of Global Anomalies

A final useful aspect of our approach is its ability to discount global effects in the time series in favour of small local effects. This is most notable in the Tor usage statistics in the period beginning in mid-August 2013, at which point a large-scale botnet began to use the Tor network for its command-and-control infrastructure [8].

As can be clearly seen from Figure 12, and also from Figures 9 and 4, global Tor usage experienced a sharp increase of almost an order of magnitude during this period. Despite this extreme alteration in global usage patterns, however, the botnet and is subsequent decline are not highlighted as anomalous periods in our analyses.

7.3 False Detection Rates

As with any unsupervised machine learning technique it is difficult to present a convincing analysis of falsely-detected positives and negatives. This is compounded by the subjective nature of the phenomena that we investigate in this work; false negatives in periods known to be anomalous can be shown, as demonstrated in §7.1.2, however there is no objective basis against which to judge a detected anomaly as a false positive in the general case.

It is, of course, possible to make judgements with respect to the effectiveness of detection rates, however these must be considered in light of the tuning parameters of §5.3 and the nature of the required analysis.

A simple method to judge the most significant anomalies at a per-country level is to consider those whose residual falls outside of a given number of standard deviations from the mean residual over the series, and we have experimented with this approach. In practice, however, we have found that the 0.1 threshold for residual anomalies produces useful and effective results.

Jackson and Mudholkar [13] suggest a *Q-statistic* for expressing the squared prediction error in terms of the number of selected principal components in the normal subspace. A more formal evaluation of this statistic could be of use in future work, but is ultimately a re-expression of the tuning parameters we have already discussed.

8. FUTURE WORK

We have demonstrated that principal component analysis can detect various forms of censorship-related events. It would be useful to classify these types of event more explicitly, and to determine appropriate parameters for detecting different classes of filtering-related event.

Whilst the work presented here has focused on the application of our technique to Tor metrics data, the method is more generally applicable. Appying the techniques presented here to other data sources is the most obvious direct extension to this work, and we have made preliminary analyses based on data from the Measurement Lab[10], as well as evaluating data from the OONI Project[22] for its applicability in detecting filtering. Other data sources, such as social media, are also likely candidates for analysis.

More interestingly, combining multiple data sources into the model to strengthen the power of the predictions is of great interest, although such an extension would be non-trivial.

Perhaps the most significant extension to this work, however, is in a more detailed analysis of the relative behaviours of countries in the dataset. One limitation of principal component analysis is that components are largely non-explanatory, and do not indicate the nature of the relationships between countries. Analysing such relationships is of great importance to understanding the global spread and development of filtering practices.

The software developed in the course of this research is already usable, and will shortly be made available, as a means for activists and the research commnity to detect suspiciuos events in global internet usage for further research.

9. CONCLUSIONS

In this work we have presented an approach based on principal component analysis to detect anomalous periods in per-country usage statistics from globally-gathered time series. We demonstrate that application of this approach

to statistics gathered by the Tor Project's metrics portal provides a means to detect filtering and censorship-related events at the global level.

To our knowledge, this work provides the first generally applicable tool for detecting a broad class of internet filtering events on a global scale, without the need to focus on individual countries. Countries exhibiting anomalous behaviour are automatically identified, and can therefore be subjected to further, more targeted, investigation.

We have demonstrated that our approach is effective at identifying known internet filtering events, both for direct blocking of the Tor network and also for related social and political events. Whilst our approach cannot prove causation between events and observed effects, it provides an effective means to highlight regions whose behaviour may be cause for concern.

We have validated our approach by correlating detected anomalous periods with known events, and have also demonstrated the effectiveness of the technique by injecting artificial anomalies into an otherwise normal time series and demonstrating the sensitivity and lag experienced in detecting these anomalies.

Our approach avoids many of the difficult ethical issues with censorship research by relying entirely on passive measurements, using statistics gathered from existing data sources. We thus avoid exposing users or research participants to unquantifiable risks.

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.

In addition to the underlying technique and tool developed to detect anomalous periods of behaviour, we have hypothesised and provided 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.

References

[1] C. Anderson. "Dimming the Internet: Detecting Throttling as a Mechanism of Censorship in Iran". In: *CoRR* abs/1306.4361 (2013). URL: http://arxiv.org/abs/1306.4361.

[2] Anonymous. "The Collateral Damage of Internet Censorship by DNS Injection". In: *SIGCOMM Comput. Commun. Rev.* 42.3 (June 2012), pp. 21–27. ISSN: 0146-4833. DOI: 10.1145/2317307.2317311. URL: http://doi.acm.org/10.1145/2317307.2317311.

[3] Anonymous. "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)*. San Diego, CA: USENIX Association, Aug. 2014. URL: https://www.usenix.org/conference/foci14/workshop-program/presentation/anonymous.

[4] R. Clayton, S. J. Murdoch, and R. N. M. Watson. "Ignoring the Great Firewall of China". In: *Proceedings of the 6th International Conference on Privacy Enhancing Technologies*. PET'06. Cambridge, UK: Springer-Verlag, 2006, pp. 20–35. ISBN: 3-540-68790-4, 978-3-540-68790-0. DOI: 10.1007/11957454_2. URL: http://dx.doi.org/10.1007/11957454_2.

[5] J. R. Crandall et al. "ConceptDoppler: A Weather Tracker for Internet Censorship". In: Computer and Communications Security. Oct. 2007. URL: http://www.cs.unm.edu/\~{}crandall/concept_doppler_ccs07.pdf.

[6] G. Danezis. *An anomaly-based censorship-detection system for Tor*. Tech. rep. The Tor Project, 2011. URL: https://research.torproject.org/techreports/detector-2011-09-09.pdf.

[7] R. Deibert. *Access Denied: The Practice and Policy of Global Internet Filtering (Information Revolution and Global Politics Series)*. 1st ed. MIT Press, Dec. 2007. ISBN: 0262541963. URL: http://www.worldcat.org/isbn/0262541963.

[8] R. Dingledine. *How to handle millions of new Tor clients*. https://blog.torproject.org/blog/how-to-handle-millions-new-tor-clients. Accessed 15th May, 2015.

[9] R. Dingledine, N. Mathewson, and P. Syverson. "Tor: The Second-Generation Onion Router". In: *IN PROCEEDINGS OF THE 13 TH USENIX SECURITY SYMPOSIUM*. 2004.

[10] C. Dovrolis et al. "Measurement lab: overview and an invitation to the research community". In: *Computer Communication Review* 40.3 (2010), pp. 53–56. DOI: 10.1145/1823844.1823853. URL: http://doi.acm.org/10.1145/1823844.1823853.

[11] A. FilastÃš. *Hadara Palestine Report*. https://ooni.torproject.org/post/hadara-palestine/. Accessed 11th May, 2015.

[12] L. Huang et al. "Communication-efficient online detection of network-wide anomalies". In: *In IEEE Conference on Computer Communications (INFOCOM)*. IEEE, 2007, pp. 134–142.

[13] J. E. Jackson and G. S. Mudholkar. "Control Procedures for Residuals Associated with Principal Component Analysis". In: *Technometrics* 21.3 (1979), pp. 341–349.

[14] I. T. Jolliffe. *Principal component analysis*. New York: Springer, 2002. ISBN: 0387224408 9780387224404. URL: http://link.springer.com/book/10.1007%2Fb98835.

[15] H. F. Kaiser. "The Application of Electronic Computers to Factor Analysis". In: *Educational and Psychological Measurement* 20 (1 1960), pp. 141–151. DOI: 10.1177/001316446002000116.

[16] M. B. Kelley. *Evidence Of Iran 'Throttling' The Internet Points To An Ingenious Form Of Censorship*. http://www.businessinsider.com/how-iran-slows-down-its-internet-2013-6?IR=T. Accessed 11th May, 2015.

[17] G. King, J. Pan, and M. E. Roberts. "How Censorship in China Allows Government Criticism but Silences Collective Expression". In: *American Political Science Review* 107 (2013), pp. 1–18.

[18] A. Lakhina, M. Crovella, and C. Diot. "Diagnosing Network-Wide Traffic Anomalies". In: *Proceedings of ACM SIGCOMM 2004*. Portland, OR, Aug. 2004, pp. 219–230. URL: http://www.cs.bu.edu/faculty/crovella/paper-archive/sigc04-network-wide-anomalies.pdf.

[19] MaxMind Inc. *MaxMind GeoIP City Database*. http://www.maxmind.com/app/city. Accessed 14^{th} May 2015.

[20] H. M. Moghaddam et al. "SkypeMorph: Protocol Obfuscation for Tor Bridges". In: *Proceedings of the 19th ACM conference on Computer and Communications Security (CCS 2012)*. 2012.

[21] K. Pearson. "On lines and planes of closest fit to systems of points in space". In: *Philosophical Magazine* 2.6 (1901), pp. 559–572.

[22] T. O. Project. *The Open Observatory of Network Interference*. https://ooni.torproject.org/. Accessed 15th May, 2015.

[23] British Broadcasting Corporation. *Syria 'cut off from the internet'*. http://www.bbc.co.uk/news/world-middle-east-22446041. Accessed 11th May, 2015.

[24] Committee to Protect Journalists. *Tibetan voices censored around China's Party Congress*. https://cpj.org/blog/2012/11/tibetan-voices-censored-around-chinas-party-congre.php. Accessed 11th May, 2015.

[25] Los Angeles Times. *Tehran Currency Protests*. http://latimesblogs.latimes.com/world_now/2012/10/protests-erupt-in-iran-over-plunging-value-of-currency.html. Accessed 11th May, 2015.

[26] The Tor Project. *Tor Metrics Portal*. https://metrics.torproject.org/. Accessed 14th May, 2015.

[27] The Tor Project. *Tor Metrics: Questions and answers about user statistics*. https://gitweb.torproject.org/metrics-web.git/tree/doc/users-q-and-a.txt. Accessed 11th May, 2015.

[28] H. Ringberg et al. "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*. New York, NY, USA: ACM Press, 2007, pp. 109–120. DOI: 10.1145/1254882.1254895. URL: http://dx.doi.org/10.1145/1254882.1254895.

[29] D. Robinson, H. Yu, and A. An. *Collateral Freedom: A Snapshot of Chinese Users Circumventing Censorship*. Tech. rep. 2013.

[30] N. Rovnick. *The Great Firewall of China looms higher around the Communist Party congress*. http://qz.com/26360/great-firewall-china-communist-party-congress-censorship-internet/. Accessed 11th May, 2015.

[31] A. Soule, K. Salamatian, and N. Taft. "Combining filtering and statistical methods for anomaly detection". In: *In Proceedings of IMC*. 2005.

[32] I. W. Stats. *Internet World Stats – Iran*. http://www.internetworldstats.com/me/ir.htm. Accessed 15th May, 2015.

[33] G. Upton and I. Cook. *Oxford dictionary of statistics*. Oxford university press Oxford, UK, 2002.

[34] Z. Weinberg et al. "StegoTorus: A Camouflage Proxy for the Tor Anonymity System". In: *Proceedings of the 19th ACM conference on Computer and Communications Security (CCS 2012)*. 2012.

[35] J. Wright. "Regional variation in Chinese internet filtering". In: *Information, Communication & Society* 17.1 (2014), pp. 121–141. DOI: 10.1080/1369118X.2013.853818. eprint: http://dx.doi.org/10.1080/1369118X.2013.853818. URL: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2265775.

[36] J. Wright, T. de Souza, and I. Brown. "Fine-Grained Censorship Mapping: Information Sources, Legality and Ethics". In: *Free and Open Communications on the Internet*. San Francisco, CA, USA: USENIX, 2011. URL: http://static.usenix.org/event/foci11/tech/final_files/Wright.pdf.

[37] J. York. *Palestinian Authority Found to Block Critical News Sites*. https://www.eff.org/deeplinks/2012/04/palestinian-authority-found-block-critical-news-sites. Accessed 11th May, 2015.

[38] Y. Zhang et al. "Network anomography". In: *In IMC*. 2005.