An anomaly-based censorship-detection system for Tor

George Danezis
gdane@microsoft.com

September 9, 2011

1 Introduction

The Tor project is currently the most widely used anonymity and censorship resistance system worldwide. As a result, national governments occasionally or regularly block access to its facilities for relaying traffic. Major blocking might be easy to detect, but blocking from smaller jurisdictions, with fewer users, could take some time to detect. Yet, early detection may be key to deploying countermeasures. We have designed an "early warning" system that looks for anomalies in the volumes of connections from users in different jurisdictions and flags potential censorship events. Special care has been taken to ensure the detector is robust to manipulations and noise that could be used to block without raising an alert.

The detector works on aggregate number of users connecting to a fraction of directory servers per day. That set of statistics are gathered and provided by the Tor project in a sanitised form to minimise the potential for harm to active users. The data collection has been historically patchy, introducing wild variations over time that is not due to censorship. The detector is based on a simple model of the number of users per day per jurisdiction. That model is used to assess whether the number of users we observe is typical, too high, or too low. In a nutshell the prediction on any day is based on activity of previous days locally as well as worldwide.

2 The model intuition

The detector is based on a model of the number of connections from every jurisdiction based on the number of connections in the past as well as a model of "natural" variation or evolution of the number of connections. More concretely, consider that at time t_i we have observed C_{ij} connections from country j. Since we are concerned with abnormal increases or falls in the volume of connections we compare this with the number of connections we observed at a past time t_{i-1} denoted as $C_{(i-1)j}$ from the same country j. The ratio $R_{ij} = C_{ij}/C_{(i-1)j}$

summarises the change in the number of users. Inferring whether the ratio R_{ij} is within an expected or unexpected range allows us to detect potential censorship events.

We consider that a ratio R_{ij} within a jurisdiction j is “normal” if it follows the trends we are observing in other jurisdictions. Therefore for every time t_i we use the ratios R_{ij} of many countries to understand the global trends of usage of Tor, and we compare specific countries’ ratios to this model. If they are broadly within the global trends we assume no censorship is taking place, otherwise we raise an alarm.

3 The model details

We model each data point C_{ij} of the number of users connected at time t_i from country j as a single sample of a Poisson process with a rate λ_{ij} modelling the underlying number of users. The Poisson process allows us to take into account that in jurisdictions with very few users we will naturally have some days of relatively low or high usage—just because a handful of users may or may not use Tor in a day. Even observing zero users from such jurisdictions on some days may not be a significant event.

We are trying to detect normal or abnormal changes in the rate of change of the rate λ_{ij} between time t_i and a previous time t_{i-1} for jurisdiction j compared with other jurisdictions. This is $\lambda_{ij}/\lambda_{(i-1)j}$ which for jurisdictions with a high number of users is very close to $C_{ij}/C_{(i-1)j} = R_{ij}$. We model R_{ij}’s from all jurisdictions as following a Normal distribution $N(m, v)$ with a certain mean (m) and variance (v) to be inferred. This is of course a modelling assumption. We use a normal distribution because given its parameters it represents the distribution with most uncertainty: as a result the model has higher variance than the real world, ensuring that it gives fewer false alarms of censorship.

The parameters of $N(m, v)$ are inferred directly as point estimates from the readings in a set of jurisdictions. Then the probability of a given country ratio R_{ij} is compared with that distribution: an alarm is raised if the probability of the ratio is above or below a certain threshold.

4 The model robustness

At every stage of detections we follow special steps to ensure the detection is robust to manipulation by jurisdictions interested in censoring fast without being detected. First the parameter estimation for $N(m, v)$ is hardened: we only use the largest jurisdictions to model ratios and within those we remove any outliers that fall outside four inter-quartile ranges of the median. This ensures that a jurisdiction with a very high or very low ratio does not influence the model of ratios (and can be subsequently detected as abnormal).

Since we chose jurisdictions with many users to build the model of ratios, we can approximate the rates λ_{ij} by the actual observed number of users C_{ij}.

On the other hand when we try to detect whether a jurisdiction has a typical rate we cannot make this assumption. The rate of a Poisson variable λ_{ij} can be inferred by a single sample C_{ij} using a Gamma prior, in which case it follows a Gamma distribution. In practice (because we are using a single sample) this in turn can be approximated using a Poisson distribution with parameter C_{ij}. Using this observation we extract a range of possible rates for each jurisdiction based on C_{ij}, namely $\lambda_{ij_{min}}$ and $\lambda_{ij_{max}}$. Then we test whether that full range is within the typical range distribution—if not we raise an alarm.

5 The parameters

The deployed model considers a time interval of seven (7) days to model connection rates (i.e. t_i - $t_{i-1} = 7$ days). The key reason for a weekly model is our observation that some jurisdictions exhibit weekly patterns. A 'previous day' model would then raise alarms every time weekly patterns emerge. We use the 50 largest jurisdictions to build our models of typical ratios of traffic over time—as expected most of them are in countries where no mass censorship has been reported. This strengthens the model as describing "normal" Tor connection patterns.

We consider that a ratio of connections is typical if it falls within the 99.99 % percentile of the Normal distribution $N(m, v)$ modelling ratios. This ensures that the expected rate of false alarms is about 1/10000, and therefore only a handful a week (given the large number of jurisdictions). Similarly, we infer the range of the rate of usage from each jurisdiction (given C_{ij}) to be the 99.99 % percentile range of a Poisson distribution with parameter C_{ij}. This full range must be within the typical range of ratios to avoid raising an alarm.

6 Further work

The detector uses time series of user connections to directory servers to detect censorship. Any censorship method that does not influence these numbers would as a result not be detected. This includes active attacks: a censor could substitute genuine requests with requests from adversary-controlled machines to keep numbers within the typical ranges.

A better model, making use of multiple previous readings, may improve the accuracy of detection. In particular, when a censorship event occurs there is a structural change, and a model based on modelling the future of user loads before the event will fail. This is not a critical problem, as these "false positives" are concentrated after real censorship events, but the effect may be confusing to a reader. On the other hand, a jurisdiction can still censor by limiting the rate of censorship to be within the typical range for the time period concerned. Therefore adapting the detector to run on longer periods would be necessary to detect such attacks.