Automatic Control of Configuration of Web Anonymization

Tomas Sochor University of Ostrava
Department of Computer Science
Ostrava, 30. dubna 22, 701 03 Czech Republic
e-mail: tomas.sochor@osu.cz

ABSTRACT

Anonymization of the Internet traffic usually hides details about the request originator from the target server. Such a disguise might be required in some situations, especially in the case of web browsing. Although the web traffic anonymization is not a part of the http specification, it could be achieved using a certain extra tool. Significant deceleration of anonymized traffic compared to normal traffic is inevitable but it can be controlled in some cases as this article suggests. The results presented here focus on measuring the parameters of such deceleration in terms of response time, transmission speed and latency and proposing the way how to control it. This study focuses on TOR primarily because recent studies have concluded that other tools (like I2P and JAP) provide worse service. Sets of 14 file locations and 30 web pages have been formed and the latency, response time and transmission speed during the page or file download were measured repeatedly both with TOR active in various configurations and without TOR. The main result presented here comprises several ways how to improve the TOR anonymization efficiency and the proposal for its automatic control. In spite of the fact that efficiency still remains too low compared to normal web traffic for ordinary use, its automatic control could make TOR a useful tool in special cases.

KEYWORDS

Web anonymization, I2P, JAP, TOR, WWW.

1 INTRODUCTION IN WEB ANONYMIZATION

There are cases when Internet users would want to hide details about their activity, especially on specific web sites. A part of the task could be done by encryption (e.g. VPN) but the traces of the activity still remain in the user's computer as well as in the web server. In case that users desire to eliminate them too they need a tool performing so-called ″anonymization″. The web anonymization consists in both hiding the client IP address from the server and the traces of the network communication with the server from the client computer.

There are various anonymization tools. Our study focuses to free tools among which TOR (The Onion Routing) belongs to the best known tools. Recently other free tools (JAP, I2P) have also emerged. Recent studies [1] and [2] were focused on a comparison of all the three tools mentioned. Their results showed that TOR represents the best option for the most cases (see the diagram in Figure 1). It can be seen from the diagram that I2P load times are the longest (worts) in most cases. The most important problem with I2P is not the worse efficiency but the poor stability of the client program disregard its security that is in question due to the fact that only a single tunnel is avalable for

I2P operation, see [5]). This was the reason why only 3 sets of measurements were completed for I2P while 15 sets were measured for TOR and JAP.

This study focuses to deeper analysis of TOR behaviour and especially to potential capability to improve the TOR behavior and efficiency. The primary focus of the study was to perform a set of measurements of the latency and response time inside the web browser using TOR-anonymized web traffic having various configuration, to compare results and propose solutions for TOR application.

1.1 Onion Routing Anonymization

Anonymization of Internet traffic can be done by various tools. Our main focus is given to free tools. Nevertheless the recent work [2] has shown that TOR (The Onion Routing) seems to be the best option among public domain tools. Therefore only TOR is studied here in details. The majority of anonymization tools including TOR rely on concealing the IP address of the originator. Operation of public anonymization tools is possible thanks to the fact that many people across the world allow to use their computers as anonymization nodes (in the case of TOR called onion routers). The basic principle of TOR is that the TOR client installed on the user′s computer initiates forming the networks (called ″TOR circuit″) composing three onion routers. The communication between these TOR routers is encrypted so that every onion router can directly communicate only with their two neighbors in the TOR circuit. The result is that the other (non-neighboring) nodes do not know anything about the node (including IP address). More detailed information about TOR operation can be found in [3]. There are other anonymization tools like I2P (for details see [5]) and JAP (see [4]) but for reasons summarized in [2] the main focus of this paper is the use of TOR.

An obvious drawback associated to the use of anonymization tools is the increase in latency as well as the total response time (or in other words the decrease of transmission speed). Significant worsening of transmission quality expressed in latency and total response time is obviously not only due to adding more nodes into the path between the client and the target but also due to encryption of data traffic between intermediate nodes. The main aim of the study described here was to quantify such decrease for various TOR configurations so as to allow the assessment of efficiency of using TOR.

2 Measurement METHODS

2.1 Model traffic selection

Similarly to the case of the previous works [1] and [2], it was decided to focus the study on web traffic only. The web traffic was chosen not only because of its significant and stable share in the total Internet traffic (e.g. [6]). There were also other reasons, primarily the fact that the http communication is also used for other purposes than simple web browsing (e.g. web services in information systems, see e.g. [7]). The nature of web traffic causes that it is much more likely to require anonymization as well. The most important reason was limitation of most of anonymization tools consisting in the fact that only web traffic is supported (in some cases even only to selected web browsers). This was the reason why our previous studies were focused on web traffic so we decided to keep this model to achieve comparability with older results.

Our measurements were intended to measure latency and total response (or download) time in various TOR configurations, thus the significant part of the study was proper selection of the test sets. Two separate test sets were formed, the first test set was a set of web pages and the other was a set of locations of a single file available via http. The set composition was based on the sets used for previous measurements but it demonstrated to be difficult to keep the substantial part of the previous sets because of the fact that a significant part of the sets has changed (mostly disappeared) since the last measurements have been performed.

2.1.1 Web pages test set

Web pages to test were chosen so that they are stable in time both from the point of view of

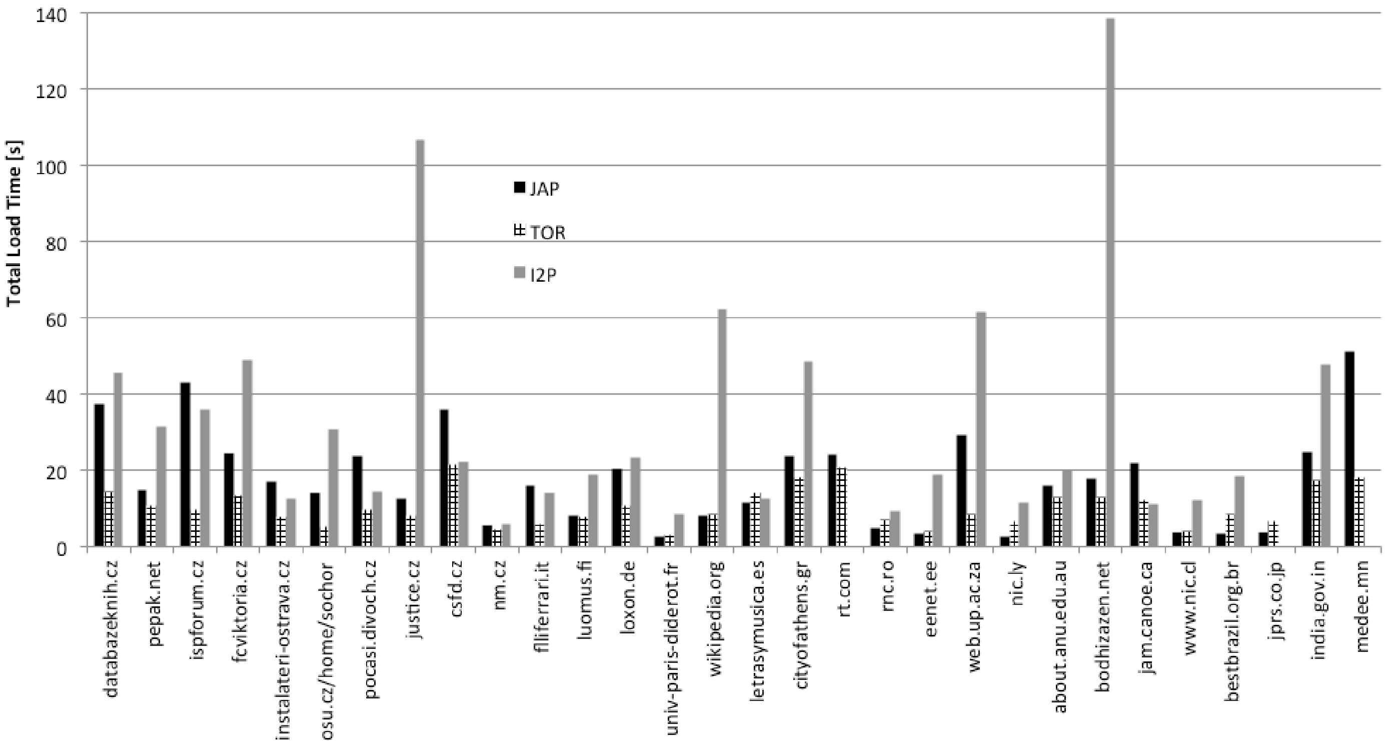


Fig. 1. Load time of web pages for among JAP, TOR and I2P anonymization tools. The data in the diagram are averages of 15 measurements (JAP and TOR) and averages of 3 measurements for I2P.

availability and size. Multimedia-rich or active-contents web pages were avoided as well as pages requiring active plugins (e.g. Silverlight, Flash, Java etc.) to display properly. It should be noted that many plug-ins could pose a problem in respect to anonymity because web browser plugins could operate independently and potentially perform actions resulting in the loss of anonymity, e.g. by ignoring the redirection to anonymization proxy or separate cookies management. The overall selection process of web pages was rather subjective-casual one with trial&error refinement however.

The resulting list of 30 web pages where all measurements have been done is described in brief in Table 1. The URL did not fit but they can be found in the legend of Figures 3 and 4. It should be noted that all web pages were composed of many (typically 10 or more) elements (images, scripts and style files). This fact influences the total load time of the web page because in most web browsers the total number of simultaneous TCP connections downloading the individual files from the server. This influence is discussed in the Results section in more details.

The test set is composed in a way that the first 10 web pages are in the Czech Republic, the next third of the webpages are located in Europe (at least according to the domain names and verified according to IP addresses) and the final third is located in the rest of the world (i.e. in non-European countries). This way of web page test set composition is intended to allow taking geographical influences to results into account. The selection process was rather subjective with trial&error refinement, however.

2.1.2 Testing file download

Testing files were chosen in a different way. Just one single file was chosen for testing (Firefox installation package sized 19,329 kB). The file was downloaded from 14 different locations selected from approx. 20 mirrors offered by sourceforge.net for the specific file. The subset of 14 location was formed in order to get the most stable files. Geographic distribution of the mirrors was also taken into account when

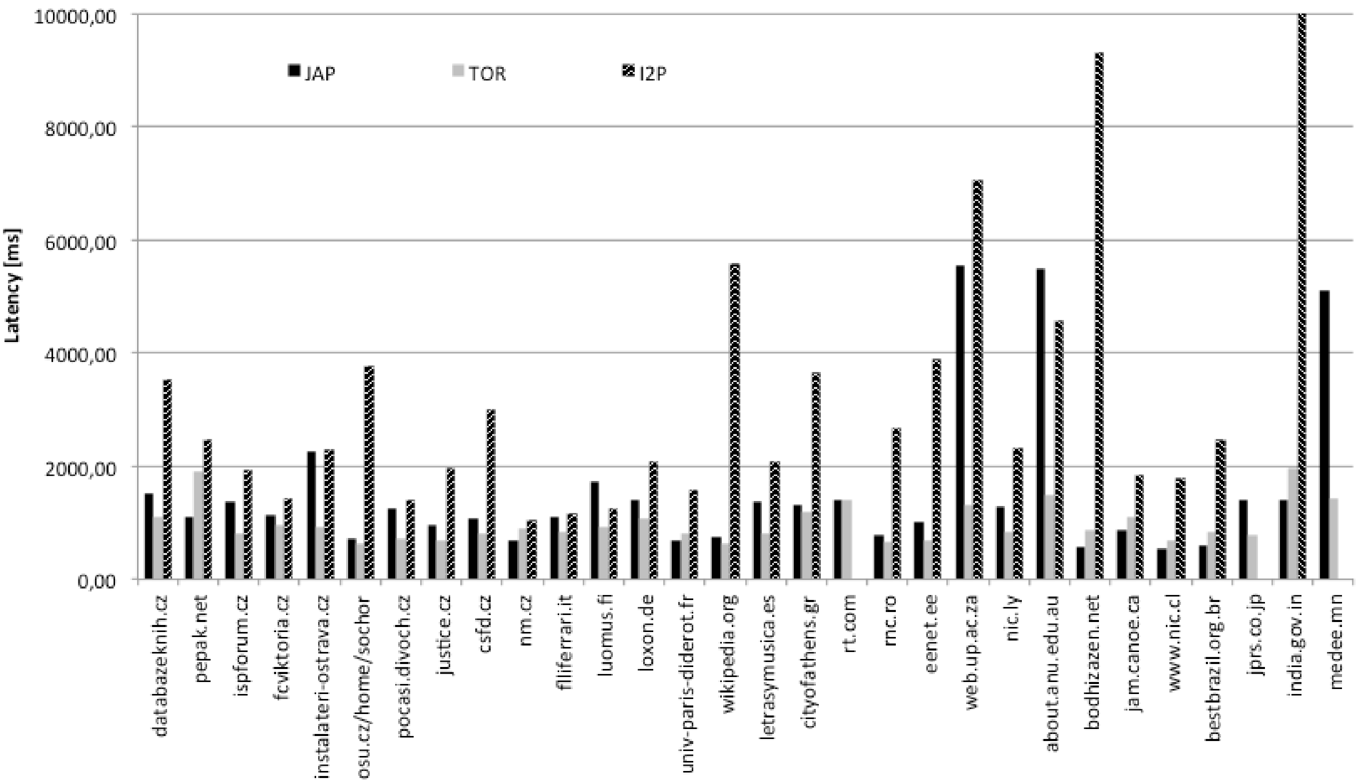


Fig. 2. Latency when loading web pages: comparison among JAP, TOR and I2P anonymization tools. The data in the diagram are averages of 15 measurements (JAP and TOR) and averages of 3 measurements for I2P.

selecting to keep as uniform worldwide distribution as possible.

2.2 Measurement of Latency and Transmission Speed

Latency (for the purpose of this study defined as “round-trip latency” because all measurements were made locally) is a parameter that is quite difficult to measure exactly because of its ’soft’ (i.e. insufficiently exact) nature. The definition itself (the time difference between sending the request and the first byte of the response) looks exact enough. The inexactness lies in the problem what time to accept as the end of the measured period. Almost the same is true for measurement of response time. Moreover, the effect of DNS resolution also had to be taken into account. If DNS requests are not eliminated, the latency measurement would be affected. The exact determination of the end of the load period is particularly complicated due to the nature of the world wide web (http) communication where the traffic consists of downloading multiple files consecutively initiated by the web client.

Therefore this study assumed certain additional limitations. One of the main limitations was that only a single web client software was used (namely Mozilla Firefox, see details below) so as to eliminate extrinsic fluctuations in measured data due to different implementation of client http communication functions. The fact that the behavior of the web browser could affect the results led to the decision to incorporate adjusting certain parameters of the web browser into the fine-tuning of the anonymization process as mentioned in Chapter 4.

It is necessary to remark that all the measurements were performed solely on the application layer despite the fact that in some cases the L3 measurement tools could be more precise. The limitation of the measurement to L7 is due to the nature of the desired comparison. The main obstacle making the L3 measurement useless was the fact that all anonymization tools studied here perform anonymization on

TABLE 1
Basic information on the test set of web pages (omitted URL can be found in the legend of Figures 3 and 4.)

Item	Size (kB)	Country	HTML	CSS	JS	Images
1	406.58	Czech	1	5	12	28
2	197.83	Czech	1	8	6	31
3	289.24	Czech	1	5	3	21
4	494.65	Czech	1	6	16	59
5	331.22	Czech	1	3	4	18
6	230.17	Czech	1	1	0	3
7	148.32	Czech	4	6	4	21
8	306.14	Czech	1	1	2	30
9	549.58	Czech	6	2	3	19
10	125.59	Czech	1	0	0	4
11	233.28	Italy	1	2	1	8
12	390.89	Finland	2	5	0	23
13	275.02	Germany	1	8	7	9
14	43.37	France	1	1	0	2
15	80.42	Netherlands	1	1	1	12
16	81.14	Spain	3	1	2	18
17	332.2	Greece	1	29	8	47
18	508.19	Russia	1	4	10	39
19	46.73	Romania	1	2	0	15
20	65.77	Estonia	1	1	1	9
21	299.85	South Africa	2	1	4	14
22	35.5	Libya	1	1	0	16
23	298.8	Australia	1	3	2	13
24	505.35	USA	1	4	2	16
25	107.82	Canada	1	2	16	38
26	120.85	Chile	1	1	0	5
27	281.97	Brazil	1	2	3	16
28	49.9	Japan	1	6	1	16
29	376.88	India	1	4	5	22
30	517.31	Mongolia	3	3	2	16

the application layer so the measurement of L3 parameters would not produce any results applicable to this study.

2.2.1 Software used for measurement

As mentioned the only web client was be used for measurement. Mozilla Firefox 16 was chosen primarily due to the most extensive support of this web browser for TOR (e.g. Firefox configurability, available plugins). All measurements were done using the following software:

- TOR 0.2.2.39
- Vidalia 0.2.20 Control panel
- Mozilla Firefox 16.0.2 with the following Mozilla Firefox plugins
  - AdBlock Plus 2.1.2
  - Firebug 1.10.5
  - iMacros for Firefox 7.6.0.2
  - NoScript 2.6
  - Torbutton 1.4.6.3

In addition to the software listed above the cURL 7.28.0 was used for file download measurement as a replacement of Firefox. Also Wireshark (version 1.8.3) was also used but not for routine measurements, only for verification of results where necessary.

Both latency and response time were measured. To eliminate undesired fluctuations in results due to caching of previously downloaded data, the use of the web client's cache was completely deactivated (using Firebug plugin, see [9]).

The latency (RTT) was defined here as time from the end of establishing the TCP connection (i.e. sending the first HTTP request) to the target till the beginning of the data reception from the target. These times were measured by Firebug and the RTT is then defined as follows:

$$Latency = Sending\_req + Waiting\_for\_response \quad (1)$$

It means that the time required for DNS lookup and establishing the TCP connection are not considered to be a part of the *Latency* (RTT).

The total response time was defined as the time from sending the first request till the reception of the complete web page (all its files). In order to obtain comparable results, the time reported by Firebug was compared to the time measured by Fasterfox and Lori used in previous studies. It was observed (and verified using multilayer measurement using Wireshark too) that data measured by Fasterfox and Lori are systematically incorrect (longer) than results from Firebug verified by Wireshark. Therefore the previous results presented in [2] in terms of transmission speed contain a systematic error. Due to the systematic nature of the error, it is obvious that all measured times were slightly longer and it means that comparison of transmission speed between normal and anonymized traffic gave slightly lower speed ratio between normal and anonymized traffic. The results presented here do not contain any such error. A certain measurement error in respect to the DNS resolution has been observed in Firebug but it can be neglected due to the fact that DNS resolution time is not a part of the measured times.

The web page download has been measured in 15 sets. Each set of measurements comprising gradual displaying of each of the webpage from the test set described in Table 1 was performed in different weekday and time of the day to avoid systematic circadian and hebdomadal errors.

The file download was measured using cURL instead of Firefox because of easier automation of measurements. The latency (RTT) measured in the case of file download was defined in the same way as in the case of web pages. The cURL values time_starttransfer and time_connect were used for RTT calculation. The file response time was calculated using time_tolal cURL value. The response time was calculated as follows:

$$response\_time = time\_total - time\_starttransfer \tag{2}$$

Thanks to the subtraction of the time_starttransfer value the resulting time could be directly used for transmission speed estimation as follows.

$$transmission\_speed = file\_size/response\_time \tag{3}$$

Similarly to the case of webpages, each set consisted of subsequent downloading of the test file files from 14 various locations.

The precision of the cURL program measurement was tested, too, and the time difference between data from cURL and Wireshark were approx. 0.01% so it was considered as neglectable.

3 RESULTS

The measured data was preprocessed by computing average and standard deviation values. The extreme values at the 5% significance level were eliminated. In this case the both-sided interval of 95% reliability is the interval $(\bar{x} - 1.96 \cdot std(x), \bar{x} + 1.96 \cdot std(x))$. All values beyond this interval were excluded from further processing. This elimination was applied for 140 measurements of RTT and 112 values of response time from total 4500 measured values for web pages (i.e. approx. 5.6% measurement eliminated) while 40 measurement of RTT and 34 ones of transmission speed from total 1260 measured values for file download measurement (i.e. approx. 5.9% measurement eliminated).

3.1 Results for web page download

Web pages were downloaded in Firefox 16 with all plugins inactive (except for Firebug and iMacros for Firefox). The latter plugin was used for loading all DNS records required for all 30 web pages from the test set 1 to be loaded. This was done in order to avoid DNS request to be sent during measurement. This also caused establishing the TOR circuit causing that it is not necessary to initiate the establishing the TOR circuit after first measurement is started. In the case of multiple sets of subsequent measurements, a new Vidalia[1] identity was used (causing to form a new TOR circuit).

It should be noted that TOR efficiency depends not only on TOR configuration settings but on the web client (Mozilla Firefox) detailed configuration as well. In means that the aim of measurement was not only to find the ways how to configure TOR but the optional configuration of Firefox, too. Fifteen sets of measurements were performed, each of them with all 30 web pages and with 5 different configurations. i.e. web page download without anonymization, web page download with TOR in default setting, web page download with TOR in configured setting (Firefox in default), web page download with TOR in default setting and Firefox configured, and the last scenario: web page download with TOR in configured setting and Firefox configured.

The comparison of latency in the cases with TOR in default as shown in Figuer 3 demonstrates that the default TOR anonymization is worse (slower) compared to the TOR-anonymized web traffic with TOR configured and TOR with Firefox configured. The comparison of the same cases in terms of the total response time is shown in Figure 4.

[1] Vidalia is a TOR control panel, see Section II.

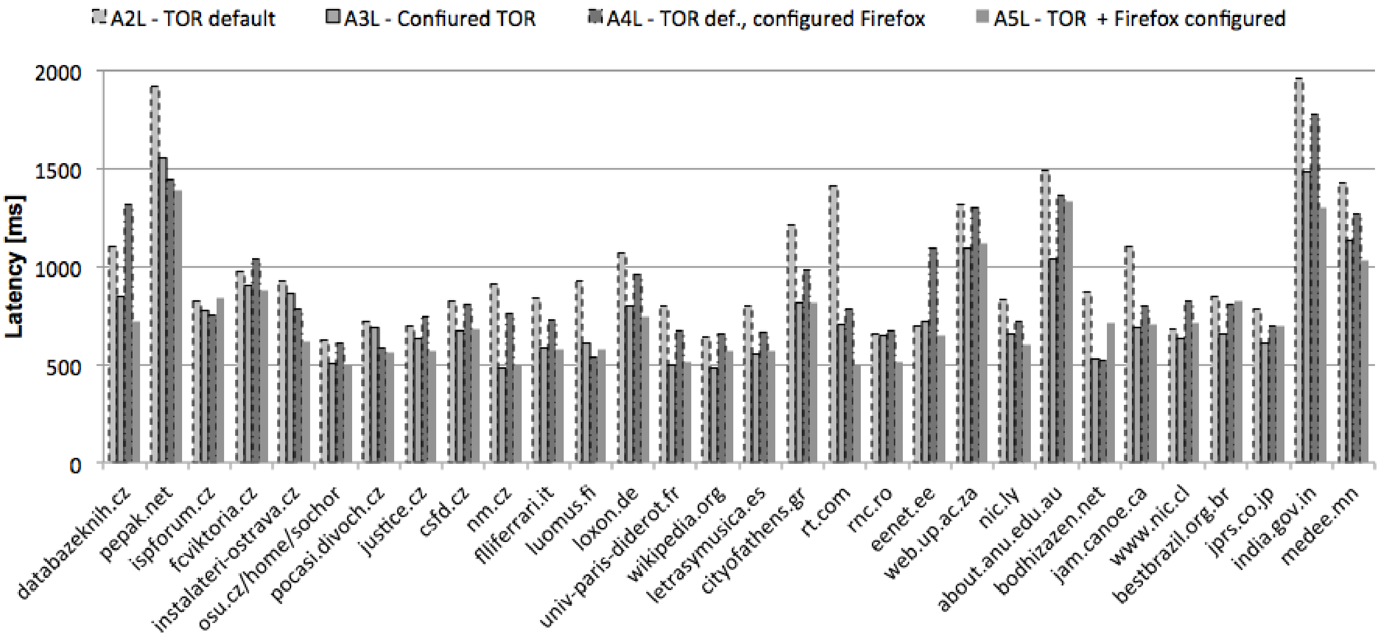


Fig. 3. Latency (RTT) of web pages comparison: TOR-anonymized in default (A2L), TOR-anonymized in configured setting (A3L), TOR.anonymized with Firefox configured, and TOR.anonymized in configured setting with Firefox configured (A5L)

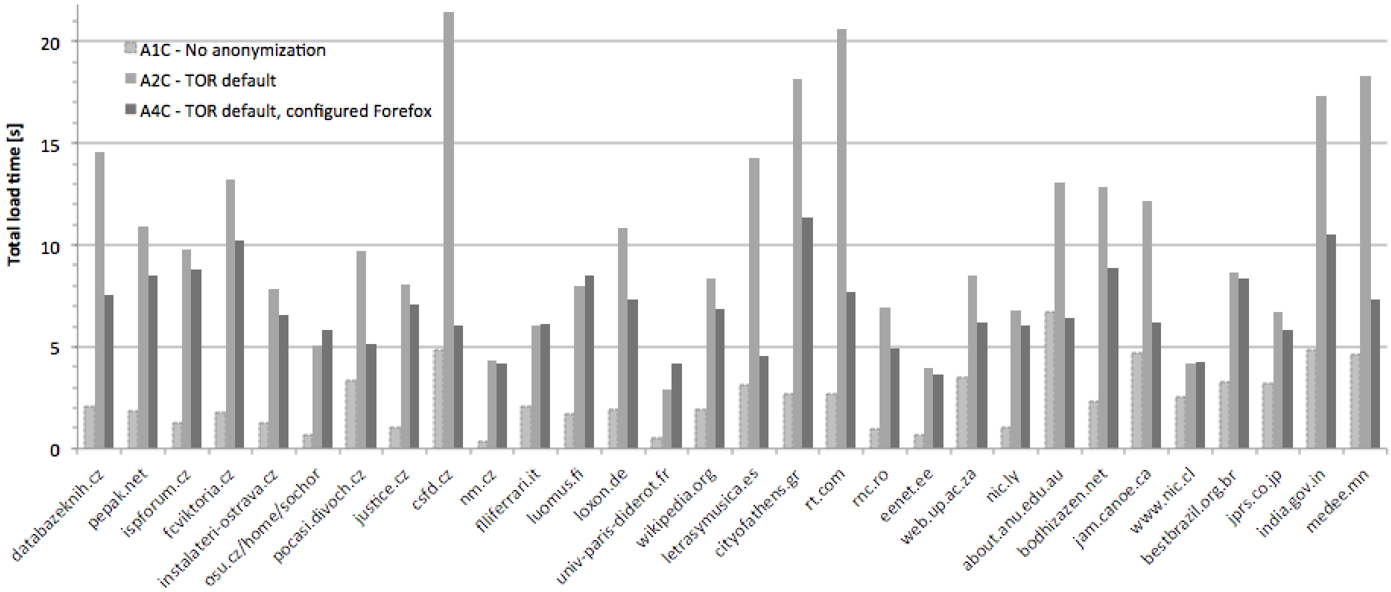


Fig. 4. Response time of web pages compared among TOR-anonymized in default with Firefox in default setting (A2C - middle column) and Firefox configured (A4C - right column). Non-anonymized traffic is also shown (A1C-left column).

3.2 Configuration of TOR and Firefox

There are several TOR configuration parameters for fine tuning. These parameters are accessible by edition of `.torrc` file. This is a common text file where each parameter can be set individually. The most influential parameter is CircuitBuildTimeout with a default value of 60 but the optimum value found during the experiments was 3. This parameter controls the timeout used in building TOR circutits (the smaller value the faster new TOR circuits are built).

Firefox has a lot of configuration parameters, too. These parameters are accessible via Firefox GUI after entering *about:config* into the address line. There is a variety of configuration parameters affecting various aspects of Firefox behavior and some of them can have an effect on the total TOR operation efficiency (i.e. on the efficiency of web page loading when using TOR). The parameter that plays the most significant role in cooperation with TOR anonymization is, as our experimental results showed, the `network.http.proxy.pipelining`. This is a Boolean parameter with FALSE default value. When set to TRUE it efficiently shortens the load time because it allows simultaneous loading of more web page elements. This parameter effect is limited to web pages with multiple (or better numerous) elements (e.g. images). As it is shown in Table 1 this is true for the majority of web pages in the test set.

An important role of Firefox configuration was played by elimination of advertisements using the AdBlock Plus plugin and deactivation of scripts using NoScript. The size of web pages after ad and script deactivation was reduced up to almost 90% in some cases (while it had no or neglectable effect in the others).

3.3 Results of File Download

The measurements of file download were made in the following three modes of operation: file download without TOR, file download with TOR in default setting,and the last case file download with TOR in the configured setting that is expected to produce the best results. The results of file download in terms of latency (RTT) is shown in Figure 5 and the transmission speed in Figure 6.

It can be easily seen that the latency increase is significant for all anonymization tool studied so far here and in our recent work [2]. The most favorable results were obtained in the case of TOR configured (increase factor 1.92) while TOR default (increase factor 3.72) was even worse than JAP. The behavior of I2P was completely unsatisfactory (increase factor almost 28) disregarding its poor stability and reliability (see [2]).

4 ANONYMIZATION PARAMETER CONTROL

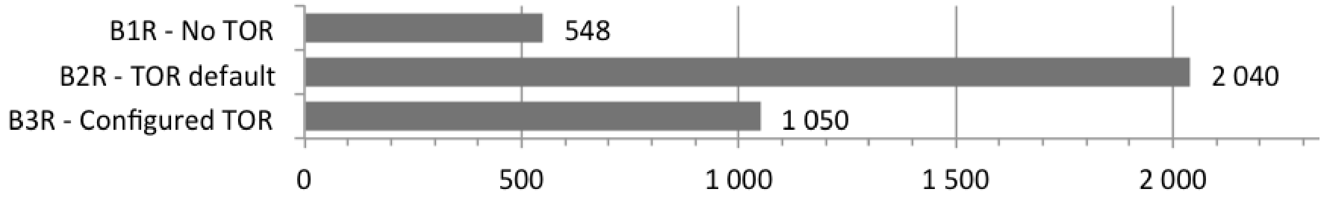
The results presented in previous sections confirmed that the application of TOR causes significant deceleration of communication comparing to the normal web browsing or file downloading. This is in accordance with the results presented in previous studies ([1], [2]). The new experimental data shows that the efficiency of TOR could be improved much by the fine-tuning of both TOR and the web client configuration.

The results in this article showed that the latency increase due to TOR was approx. 3.7 times or more compared to normal web traffic. The increase ratio was reduced to less than 2 using proper fine-tuning of both TOR and Firefox parameters. A similar conclusion holds for the response time, too. The average increase of the file download time using TOR was more than 80%. The improvement obtained due to the fine-tuning of parameters is not so significant here (approx. 78%). Setting proper configuration parameters seems to have a potential to be an important factor in the latency of web browsing anyway. Bearing a permanently changing nature of http communication in mind, it seems to be difficult to conclude with stable configuration recommendations.

From the results above, two parameters with the most significant effect to the efficiency of anonymization were selected, namely the TOR's Circuit establishing time (default value is 60 seconds) and Firefox's maximum number of pipelined requests (default value 4 simultaneous requests but this value is not applied in Firefox default setting because the pipelining is switched off by default). The best results were produced by setting shown in Table 2. together with the default values. Other configuration parameters were neglected because their influence is weak.

4.1 Automatic Control Proposal

It is necessary to note that the results in the sense of establishing the optimum values of the control parameters are not permanent and stable enough. This could be due to the changing nature of data in www pages of other



(a) Case I

Fig. 5. Latency (RTT) of file download compared among non-anonymized (B1R), TOR-anonymized in default (B2R) and TOR-anonymized in configured setting (B3R)

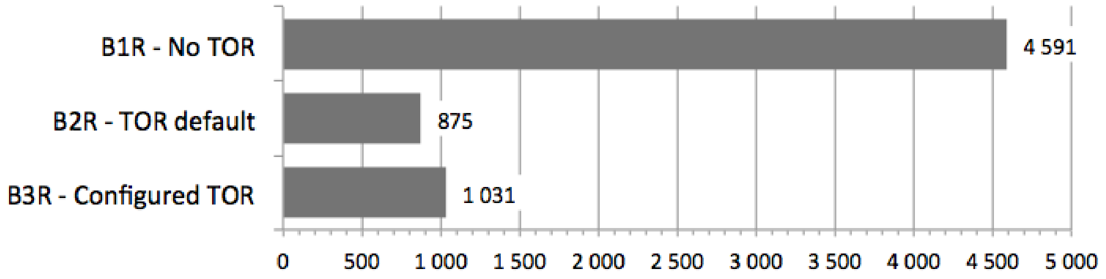


Fig. 6. Transmission speed of file download

TABLE 2
The most influential parameters of anonymization configuration in TOR and Firefox

Variable Name	Default	Optim.	Origin
Circuit establishing time [s]	60	3	TOR
Max. No. of pipelined req.s	4 (Off)	6 (On)	FF

factors beyond the scope of this study. Because of the fact the automatic control of the most influential parameters is proposed. The proposal expects the application of Linguistic Fuzzy Logic Controller (see [8]). There are also other options how to control this process, e.g. using a self-adapting neural network (e.g. [10]). The detailed proposal of the application of the control mechanism is beyond the scope of this study, however.

The automatic control of the most influential Firefox and TOR parameters described above

5 CONCLUSIONS

The results presented in Section III confirmed that the TOR anonymization affects the www client behavior in the sense of both latency increase and transmission speed decrease. At least the undesired latency increase could be controlled through fine-tuning of the most significant parameter of TOR (namely TOR circuit establishment timeout) and the most influential parameter of Firefox that controls pipelining of http requests. It was demonstrated that the latency results with controlled parameters are much better than in default TOR setting. In order to achieve and keep the best results for TOR ordinary use in the environment of permanently changing www communication the automatic control using a fuzzy controller oran artificial neural network has been proposed with expectation to allow automatic tuning of anonymization parameters.

REFERENCES

[1] Liska T., Sochor T., Sochorova H. Comparison between normal and TOR-anonymized web client traffic. Procedia - Social and Behavioral Sciences. 2010, vol. 9, pp 542-546

[2] Sochor T. Anonymization of web client traffic efficiency study. In: 19th International Conference on Computer Networks. Berlin Heidelberg: Springer Verlag, 2012. pp. 237-246.

[3] Dingledine R., Mathewson N., Syverson P., Tor: The Second-Generation Onion Router. [online]. Available at www: https://svn.torproject.org/svn/projects/design-paper/tor-design.pdf

[4] JAP - Anonymity & privacy [online] Project AN.ON Anonymity Online. Available at www: http://anon.inf.tu-dresden.de/index_en.html

[5] I2P Anonymous Network - I2P [online]. Available at www: http://www.i2p2.de/index.html

[6] Ipoque. Internet Study 2008/2009 [online]. (quot. 2013-04-08). Available at www: http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf

[7] Zacek J., Melis Z., Hunka F. Extendable Domain Specific Generator Based on Web Services. ECON 2011, Vol. 19, pp 98-104.

[8] Dvorak, A., Habiballa, H., Novak, V. , Pavliska, V. The concept of LFLC 2000 - Its specificity, realization and power of applications. Computers in Industry. Vol. 51, Issue 3, August 2003, pp. 269-280

[9] Odvarko, J. Firebug Net Panel Timings. Software is hard. [online]. [quot. 2012-08-01]. Available at: http://www.softwareishard.com/blog/firebug/firebug-net-panel-timings/

[10] Volna, E. Using Neural network in cryptography. In P. Sincak, J. Vascak, V. Kvasnicka, R. Mesiar (eds.): The State of the Art in Computational Intelligence. Physica-Verlag Heidelberg. 2000 pp.262-267. ISBN 3-7908-1322-2, ISSN 1615-3871.