HAL
open science

Weaknesses and Challenges of Network Neutrality Measurement Tools

Ximun Castoreo, Patrick Maillé, Bruno Tuffin

To cite this version:

Ximun Castoreo, Patrick Maillé, Bruno Tuffin. Weaknesses and Challenges of Network Neutrality Measurement Tools. 16th International Conference on Network and Service Management (CNSM), Nov 2020, Bordeaux, France. 10.23919/CNSM50824.2020.9269077 . hal-02542689

HAL Id: hal-02542689

https://inria.hal.science/hal-02542689

Submitted on 14 Apr 2020

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Weaknesses and Challenges of Network Neutrality Measurement Tools

Ximun Castoreo
Inria, Univ Rennes, CNRS, IRISA
Rennes, France
ximun.castoreo@inria.fr

Patrick Maillé
IMT Atlantique, IRISA, UMR CNRS 6074
F-35700 Rennes, France
patrick.maille@imt.fr

Bruno Tuffin
Inria, Univ Rennes, CNRS, IRISA
Rennes, France
bruno.tuffin@inria.fr

***Abstract*—Network neutrality has been the subject of a hot debate worldwide. Regulation has been implemented in many countries to enforce the principle of a "neutral" network. But compliance to the rules has to be checked thanks to specific measurement tools. This paper aims at highlighting the weaknesses of current network neutrality measurement tools and at providing hints on challenges to be addressed on the topic.**

***Index Terms*—Net neutrality, monitoring, measurements.**

I. Introduction

Network neutrality is basically "the principle that traffic should be treated equally, without discrimination, restriction or interference, independent of the sender, receiver, type, content, device, service or application" (European parliament positioning on April 3rd 2014[1], other definitions –but related ones– exist). The issue has been first raised in 2005 when the CEO of AT&T complained that distant content providers (CPs) were using AT&T's network without financially participating to its maintenance and investment, and threatened to block or slow down traffic [23]. It has since been the topic of a hot debate worldwide with user associations and CPs asking for an "open Internet" and claiming among other things that a non-neutral network would prevent innovation. For more on the history about neutrality, its stakes, and how it is handled worldwide, the readers are advised to look at [14], [16]–[18], [23], [24], [26].

Regulation has been passed in most countries all over the world (see the survey [9]), but the debate has recently been exacerbated with the US repealing neutrality in 2018 [7] (creating new issues about the interest to maintain it elsewhere as a consequence [1], [19]). Imposing neutrality to Internet Service Providers (ISPs) means that tools need to be put in place to monitor their behavior and verify that they stick to the rules and do not try to bypass them.

Our goal in this paper is to point out the weaknesses of existing network neutrality measurement tools and what are the challenges to be addressed by the research in the area. We do not aim to detail and survey the state of the art; for that, we refer to the recent nice and complete survey [9].

The remaining of the paper is organized as follows. Section II describes the categories and requirements of network neutrality measurement tools, and briefly lists the characteristics of the available softwares. Section III then summarizes what is lacking with the existing tools to highlight what should be developed in the future.

II. Network neutrality measurement tools

This section classifies and reviews the main characteristics of tools developed during the last couple of decades and very briefly explains what each tool does.

A. Blocking or degradation

A non-neutral behavior can be of two kinds: traffic blocking, i.e., censorship, or quality of service (QoS) degradation, also known as Traffic Differentiation. Censorship can be achieved by various ways: simply dropping any packet of the targeted traffic, manipulating DNS to make a domain name inacessible, terminating TCP connections with RST packets, filtering search engine queries... As traffic blocking is fairly easy to detect, most tools focus on Traffic Differentiation.

B. Focusing on service/type of traffic/destination

Many monitoring tools focus on specific service-based Traffic Differentiation. The service can be characterized by the protocol (BitTorrent, HTTP, FTP...), the service type (video streaming, video gaming, file transfer...) or the specific application (YouTube, DOOM, Facebook...). To identify which traffic has to be targeted among others (known as Traffic Classification), ISPs can rely on several factors: the destination or source IP address and port numbers, the data transmitted (signature in the packet payload, as a keyword in a search engine query), or the flow shape [3], [12] are among the commonly used information.

The tools differ in the range of traffic they monitor. They often focus on one type among HTTP, BitTorrent, SMTP, PPLive and VoIP (Skype, Vonage), very few on a wide predefined range of traffic types (POPI below being an exception), while some others accept any type of traffic to be tested/measured.

[1]European Parliament (2014), European Parliament legislative resolution of 3 April 2014 on the proposal for a regulation of the European Parliament and of the Council laying down measures concerning the European single market for electronic communications and to achieve a Connected Continent, and amending Directives 2002/20/EC, 2002/21/EC, 2002/22/EC, and Regulations (EC) No 1211/2009 and (EU) No 531/2012 (COM(2013)0627 - C7 -0267/2013 - 2013/0309(COD)).

C. Considered metrics

Tools also differ by the metrics they measure from the targeted traffic. The commonly used ones are:

- Packet integrity/modification (a QoS downgrading can be made for instance by modifying a packet header);
- Packet loss rate. It is then compared with the loss rate of "mainstream" traffic;
- Transmission rate. It is compared again with a reference traffic. The rate can be calculated in various layers, resulting on network throughput or application throughput (goodput). For throughput and packet loss rate, a statistical test has to be put in place and a limited differentiation might not be easy to detect;
- DNS consistency for traffic blocking. One can check the answer origin server or also compare the answer with a "reference DNS" query.

D. Measurement type

Network measurements are divided into two major categories: active measurements and passive measurements.

1) Active measures: Active measurements methods inject traffic to investigate the performance. Thanks to active measurements, one can measure exactly what is needed. But it may be challenging in various aspects: the generated traffic has to correctly reflect the legit traffic, it adds network load, a program generating traffic sometimes needs special permissions (specially on mobile devices)... Measurements are usually made by replaying traffic under "normal" and differentiated conditions.

2) Passive measures: Passive measurements monitor existing user traffic. They do not add artificial traffic to the network. This method has the advantage of being silent and measures a real traffic when active measures may not accurately represent "real" traffic. However, passive measures are inefficient if the targeted trafic is not being played.

E. List of tools

We do not provide details but just the general principles; the reader is advised to go to the references if requiring more information.

- *Switzerland* [6] was the first widely deployed automated tool, in 2008. By hashing each packet in a communication between two nodes, it detects if packets are injected, dropped or modified in a communication between two nodes. The first n bits of these hashes (from the packet at origin and destination) are sent with a timestamp to a server, that performs a comparison. Modified traffic is detected from different hashes, dropped packet from received hash from the sender and not the receiver, and injected traffic if a hash is received from the receiver only. This method is semi-passive in the sense that no extra traffic is generated between two clients, although "control" traffic is generated between the clients and the server. Also, hashing can be expensive, and the process suffers from a lack of anonymity.
- *NetPolice* [27] investigates potential differentiation in the backbone of the network through measurements at the edge. It seeks whether packets are routed through different paths according to their type or origin, a potential discrimination in terms of performance. Paths–and losses through them–are determined by sending packets with an adjusted Time To Live (TTL) field: packet drops due to TTL reaching 0 are notified to the sender while those due to differentiation are "silent". Comparison for various sources or destinations allows to detect if a differentiation is made at a node of the backbone. The comparison is made thanks to a Kolmogorov-Smirnov statistical test; Jackkniffe resampling is also used and claimed to "reduce" noise.
- *NANO* [4] is a passive measurement tool, trying to isolate service differentiation from performance differences due to "natural" factors, called *confounding factors*, such as operating system, geographical location, source and destination addresses, time of the day, type of physical link, etc., grouped into two categories: client-based features and network-based features. Application throughput is measured for several clients and the results sent to a server with user confounding factors. Note that this client-server traffic is generated by NANO, making the tool a semi-passive measurement tool, like Switzerland. Measures with similar confounding factors are grouped into strata, and statistical comparisons are made within each stratum to detect a potential different behavior. The tool is quite involved, with the necessity to identify confounding factors, which can be dynamic; also, the required sample size to get accurate results is large.
- *Glasnost* [5] is an active measurement tool that sends specific traffic (peer-to-peer, web, mail and SSH) first without modification and then replays it by modifying elements that may cause differentiation (port number for example). It compares the results to detect a potential hindrance to Net Neutrality. Analysis is performed through a web interface and no program is required at the user place. The tool arbitrarily evaluate the external noise as $(\mathrm{bitrate}_{\mathrm{max}} - \mathrm{bitrate}_{\mathrm{mean}})/\mathrm{bitrate}_{\mathrm{max}}$ and does not proceed to analysis if it exceeds 20%; otherwise it decides that differentiation takes place if the difference between flow rates is larger than 50% of the maximum flow rate. BonaFide [2] is a mobile terminal variant of Glasnost, operating the same way but adapted to the mobile environment.
- *DiffProbe* [11] assumes that the ISP separates the traffic in two classes: high (H) and low (L) priority. Differentiation is detected by sending the service flow and a probing flow simultaneously. The packets of the probing flow should have characteristics similar to the tested flow, which may be difficult to realize. The rate of the probing flow is increased to saturate the path and observe results with congestion. Instead of a Kolmogorov-Smirnov test, Kullback-Leibler divergence is considered to measure the gap between distributions of measures.

Tool	Differention	Traffic	Metrics	Measure	Test
Switzerland	packet integrity	any	packet hash	passive	comparison
NetPolice	type/routing-based	HTTP, BitTorrent, SMTP, PPLive, VoIP	packet loss rate	active	Kolmogorov-Smirnov with Jackknife
NANO	type-based	any	throughput	passive	causal inference
Glasnost	type-based	BitTorrent (can add others)	throughput	active	maximal throughput comparison
DiffProbe	type-based	Skype, Vonage	packet loss rate, delay	active	Kullback-Leiber
POPI	type-based	ICMP, FTP, Telnet, POP3, BGP, HTTPS, Fasttrack, Donkey, Gnutella, BitTorrent	packet loss rate	active	ranking, averaging and clustering
Packsen	type-based	BitTorrent	inter-arrival packet time	active	Mann-Whitney U
OONIProbe	type-based	web, DNS, Tor, messaging applications	DNS resolution, connection success	active	comparison
ChkDiff	type-based	any	packet loss rate, delay	active	Kolmogorov-Smirnov
Wehe	type-based	any	packet loss rate, throughput, delay	active	custom Kolmogorov-Smirnov-inspired
CONNEcT	type-based	any	packet loss rate	passive	no analysis

TABLE I
COMPARISON OF TOOLS

- *POPI* [15] is an active measurement tool detecting forwarding prioritization between packets. Bursts of packets are sent for different types of traffic, and packet losses are computed for each type. A statistical test is then performed to detect if differences in losses are significant.
- *Packsen* [25] also tries to detect prioritization between types of traffic, comparing traffic features at origin and destination. A reference flow and another one using the suspected differentiation characteristic (port, time of day, etc.) are sent, and both compared with the traffic at the origin. Statistical tests are performed: Mann-Whitney U test for the inter-arrival of the flows (test of the median of distributions), and another one to try to infer the used differentiation scheduler in case differentiation is detected.
- *OONIProbe* [22] is an active measurement tool aiming at detecting the blocking of web, instant messaging application, or overlay network traffics, through DNS resolution. The first test is a connectivity test, trying to connect to the given website and comparing the DNS resolution with a neutral-assumed DNS resolution. The authors use as neutral reference the Google DNS servers and assume there is no DNS manipulation during the process. If the connection fails, the tool concludes TCP-level blocking occurred. If the connection works, the tool requests a resource to the server; if the request is aborted, it concludes on HTTP-level blocking. Those results can help the client understand at which level blocking occurs and find how to circumvent it.
- *Wehe* [10] is an active measurement tool primarily aimed at mobile terminals, as they operate in a resource-constrained network with constantly increasing demands, and in which network operators have historically acted against net neutrality. The tool connects to servers where traffic mimicking that of specific applications has been uploaded using a VPN. Traffic is sent "in clear" and replayed encrypted. The goal is to detect potential differences in terms of throughput, the encrypted traffic being assumed non-differentiated, using a custom variant of a Kolmogorov-Smirnov test. The current implementation of Wehe is limited because it only allows users to test a limited number of services (Hangouts, Netflix, Skype, Spotify, Viber and Youtube).
- *CONNEcT* [20] is a measurement model designed to escape ISP oversight and the potential tentatives to bypass measurements from users (the above existing tools do not study the case of an active ISP, that could modify its behavior when it detects it is being tested). For example, the Time-to-Live value of a packet is commonly 64 or 128 (depending on the operating system) [20]. Therefore, an uncommon TTL value (in a NetPolice [27] test packet for instance) is easily identifiable by the ISP, and traffic differentiation during the measurement campaign can then be postponed. To counter ISP traffic monitoring, CONNEcT makes use of a meta-communication called covered channel; a hidden synchronization allows two machines to insert messages into the legit traffic data with an offset only known from both of them. This hidden data contains measures concerning the previous packet such as a timestamp, a packet pseudo-hash...

The characteristics of the above tools are summarized in Table I.

III. WEAKNESSES AND CHALLENGES TO BE ADDRESSED

The above tools all have their own characteristics. Our goal here is to emphasize what is lacking and what should be addressed by the community to ensure that neutrality is properly monitored.

A. Sustainability

We have listed several tools developed in the literature. Many were implemented by research projects and are not maintained anymore, or worse not available anymore. This is for instance the case of Switzerland, Glasnost, DiffProbe and BonaFide. Similarly, NetPolice and NANO were tested on PlanetLab, but have not been publicly released, similarly to POPI, Packsen, ChkDiff. Moreover CONNEcT is for now just a proof of concept without implementation. It is therefore clear that even if many have been implemented, very few can be used. Available tools are limited to OONIProbe and Wehe in our list. A support by a regulator (ARCEP, the french telecomunication regulator, is associated to the development of Wehe) or maybe by user associations (for OONIProbe) is probably a requirement for a perennial implementation of a tool.

An initiative from New America's Open technology Institute, PlanetLab Consortium, Google and a group of academic researchers, named Measurement-Lab (M-Lab) was created in 2009 to collect and save Internet measurements from various tools. Most of the presented tools are part of the M-Lab initiative and measures can be accessed via the database. The gathered data is not analyzed, the purpose of M-Lab being to make measures available.

B. Usage

In the design of a tool, a question to be addressed is the target in terms of usage. If neutrality is to be tested by institutions or associations, the tool/program can be involved. But if it is to used by end users, which may lead to a broader impact, it has to be easy to install, easy to manipulate, and results have to be obtained quickly. For example, Glasnost authors had to cut the measurements by half as the users where abandoning because of the test duration [5]. Testing by mobile users (for which a non-neutral behavior is more likely) is also not that obvious for active measurements since measuring traffic will be counted in the user's data plan, which might deter from implementation.

Some authors advocate for crowd-sensing tools [21] to gather a huge amount of measures, following the Open Data paradigm. It would allow to precisely construct a global overview of the network. But related issues have to be addressed: data anonymity, big data processing, citizen participation and incentives, etc.

Dissemination of a tool might also be an issue: in the case of a mobile terminal, an application needs to be downloaded through the Google Play Store or the Apple Store. The Wehe application was, at first, refused in the Apple Store [13].

C. Specific tools

All tools address specific types of neutrality hindrance (traffic blocking, or throughput reduction for example). The tools consider particular metrics, but also particular and limited types of traffic (P2P, HTML, etc.). Therefore a single tool may not be sufficient to certify a neutral network since many kinds of hindrance are possible. A big challenge would be to propose a tool looking at many (if not all) potential infringements. Such a general tool would probably be provided first to regulatory bodies. It also has to adapt, if possible ex ante, to network evolutions (the SDN paradigm being an example).

Another major issue is the ability for the ISPs to detect measurement tools. Most of the tools generate specific traffic (TTL-adjusted packets, client to server measure reports, control traffic with known values and keywords...). Therefore, the ISP has the possibility to: i) not differentiate service when those tools are used or ii) bypass the tested differentiation by using other ways to discriminate. Behavior ii) is probably easy to implement due to the very specific tests of most tools. Issue i) is related to the development of CONNEcT to avoid ISP oversight.

D. Limited statistical ground

A major issue not highlighted in the literature is the lack of theoretical grounds for the (mostly statistical) tests to detect service differentiation. This is notably the case for Glasnost with 20% as the maximum percentage of noise to allow a test, and 50% as the threshold for declaring service discrimination. There are tools without any decision process, left to the user, such as CONNEcT. A tool like Switzerland based on passive measures detects differences between packet content, but what if those differences are for a small proportion of packets only? Can it just be due to network processing errors and can we define thresholds, justified theoretically based on observed errors on a "regular" traffic? NANO requires complicated identification of and is very sensitive to confounding factors; making its use limited to experts.

Most of the above active measurement tools base their decision process on the Kolmogorov-Smirnov (KS) test. This test computes the maximum distance between two empirical cumulative distribution functions (cdf) and says that the two cdf are statistically different if the max distance is larger than a threshold (which depends on the sample size to get the two empirical cdf). In several cases the authors claim to suffer from "noise" and modify the test, but then in general lose its theoretical guarantees. Wehe used KS, but seemed to encounter issues with the result of the test; it then creates a new test looking at the surface, instead of the max distance, separating the two cdf, with an arbitrary threshold for decision without any theoretical validation but probably with a conservative bias to avoid false positives. Probably more could be done with a simple KS test to justify its valid use: sample points are throughput estimations over short and successive periods of time during a replay, but are the provided values independent? This independence could or should be tested and the intervals adjusted. This is similar to what is done for variance estimation using the batch means technique [8]. NetPolice use Jackniffe resampling to "reduce noise"; Jackniffe is theoretically used to reduce the bias of an estimator; more information would be interesting to see the link with noise.

DiffProbe uses the Kullback-Leibler divergence to compute the "distance" between distributions, and rejects the matching distribution if above a threshold. While a nice other idea, no theoretical test exists, making the decision arbitrary, and the Kullback-Leibler divergence is in the literature used to simplify derivations with respect to the distance between cdf, something not clear in the current context. Packsen is based on the Mann-Whitney U test, which only looks at medians, much more limited than the full distributions.

This illustrates that much more can be done to provide a valid statistical test for network neutrality measurement tools.

References

[1] K. Agarwal, P. Maillé, and B. Tuffin. Impact of Heterogeneous Neutrality Rules with Competitive Content Providers. working paper or preprint, available at https://hal.inria.fr/hal-02463485, February 2020.

[2] V. Bashko, N. Melnikov, A. Sehgal, and J. Schöwälder. Bonafide: A Traffic Shaping Detection Tool for Mobile Networks. *2013 IFIP/IEEE International Symposium on Integrated Network Management*, May 2013.

[3] L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian. Traffic Classification on the Fly. *SIGCOMM Comput. Commun. Rev.*, (4):23–26, April 2006.

[4] M. Bin Tariq, M. Motiwala, N. Feamster, and M. Ammar. Detecting Network Neutrality Violations with Causal Inference. *CoNEXT '09: Proceedings of the 5th International Conference on Emerging Networking Experiments and Technologies*, pages 289–300, 2009.

[5] M. Dischinger, M. Marcon, S. Guha, K.P. Gummadi, R. Mahajan, and S. Saroiu. Glasnost: Enabling End Users to Detect Traffic Differentiation. *Proc. of USENIX Symposium on Networked Systems Design and Implementation*, 2010.

[6] P. Eckersley. Switzerland Design. May 2008. https://www.eff.org/files/2018/06/21/design.pdf.

[7] Federal Communications Commission. Restoring internet freedom, Jan 2018. available at https://docs.fcc.gov/public/attachments/FCC-17-166A1.pdf.

[8] G.S. Fishman. *Monte Carlo: Concepts, Algorithms and Applications*. Springer-Verlag, 1996.

[9] T. Garrett, L.E. Setenareski, L.M. Peres, L.C. ErpenBona, and E. Procópio Duarte Jr. Monitoring Network Neutrality: A Survey on Traffic Differentiation Detection. *IEEE Communications Surveys & Tutorials*, March 2018.

[10] A.M. Kakhki, A. Razaghpanah, A. Li, H. Koo, . Golani, D. Choffnes, P. Gill, and A. Mislove. Identifying Traffic Differentiation in Mobile Networks. *Proceedings of the 2015 Internet Measurement Conference*, pages 239–251, October 2015.

[11] P. Kanuparthy and C. Dovrolis. Diffprobe: Detecting ISP Service Discrimination. *Proceedings - IEEE INFOCOM*, 2010.

[12] T. Karagiannis, K. Papagiannaki, and M. Faloutsos. BLINC: Multilevel Traffic Classification in the Dark. *Proceedings of the 2005 Conference on Applications, Technologies, Architectures, and 'rotocols for Computer Communications*, (12):229–240, 2005.

[13] J. Koebler. Apple Is Blocking an App That Detects Net Neutrality Violations From the App Store. https://www.vice.com/en_us/article/j5vn9k/apple-blocking-net-neutrality-app-wehe. Accessed March 5, 2020.

[14] T.M. Lenard and R.J. (Eds.) May. *Net Neutrality or Net Neutering: Should Broadband Internet Services be Regulated*. Springer, 2006.

[15] G. Lu, Y. Chen, S. Birrer, and F.E. Bustamante. POPI: A User-Level Tool for Inferring Router Packet Forwarding Priority. *IEEE/ACM Transactions on Networking*, 18(1):1–14, February 2010.

[16] P. Maillé, P. Reichl, and B. Tuffin. Internet governance and economics of network neutrality. In A. Hadjiantonis and B. Stiller, editors, *Telecommunications Economics - Selected Results of the COST Action IS0605 EconTel*, pages 108–116. Lecture Notes in Computer Science 7216, Springer Verlag, 2012.

[17] P. Maillé, G. Simon, and B. Tuffin. Toward a net neutrality debate that conforms to the 2010s. *IEEE Communications Magazine*, 54(3):94–99, 2016.

[18] P. Maillé and B. Tuffin. *Telecommunication Network Economics: From Theory to Applications*. Cambridge University Press, 2014.

[19] P. Maillé and B. Tuffin. Neutral and non-neutral countries in a global internet: What does it imply? In Karim Djemame, Jörn Altmann, José Ángel Bañares, Orna Agmon Ben-Yehuda, and Maurizio Naldi, editors, *16th International Conference on Economics of Grids, Clouds, Systems, and Services (GECON)*, pages 111–123, Leeds, UK, 2019.

[20] A. Maltinsky, R. Giladi, and Y. Shavitt. On Network Neutrality Measurements. *ACM Transactions on Intelligent Systems and Technology*, 8(4), May 2017.

[21] D. Miorandi, I. Carreras, E. Gregori, I. Graham, and J. Stewart. Measuring Net Neutrality in Mobile Internet: Towards a Crowdsensing-based Citizen Observatory. *IEEE International Conference on Communications*, June 2013.

[22] Network Neutrality Squad. Open Observatory of Network Interference. https://ooni.org. Accessed March 5, 2020.

[23] A. Odlyzko. Network neutrality, search neutrality, and the never-ending conflict between efficiency and fairness in markets. *Review of Network Economics*, 8(1):40–60, 2009.

[24] H. Schulzrinne. Network neutrality is about money, not packets. *IEEE Internet Computing*, 22(6):8–17, 2018.

[25] U. Weinsberg, A. Soule, and L. Massoulie. Inferring Traffic Shaping and Policy Parameters using End Host Measurements. *Proceedings - IEEE INFOCOM*, May 2011.

[26] T. Wu. Network neutrality, broadband discrimination. *Journal of Telecommunications and High Technology*, 2003.

[27] Y. Zhang, Z. Morley Mao, and M. Zhang. Detecting Traffic Differentiation in Backbone ISPs with NetPolice. *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 103–115, November 2009.