Ethics, Products, Top Lists - and their Use at Internet Measurement Conferences

Luca Ciprian
Advisor: Quirin Scheitle
Seminar Innovative Internet Technologies and Mobile Communications SS2017
Chair of Network Architectures and Services
Departments of Informatics, Technical University of Munich
Email: luca.ciprian@tum.de

ABSTRACT

In this paper, we evaluate ethical considerations, reproduction considerations, the use of geolocation products, and the use of top lists in the three major Internet Measurement Conferences. We take all the publication of the past three years into account, making a total of 260 evaluated papers. We further dive in deeper into subtopics for each of the four categories mentioned above, comparing differences and commonalities of papers and outline special characteristics of certain papers. In the end we can see that top lists are more frequently used than the other inspected subtopics. Additionally it stands out that the consideration of ethics and reproduction increases every year.

Keywords

Ethics, Reproduction, Geolocation, Top Lists

1. INTRODUCTION

Internet Measurement is conducted in order to understand and assess many different sectors of the web. Whether it is the security and privacy of users, the behavior of applications or simply analyzing the traffic, there are many fields that need thorough exploration to keep the information up to date. This paper examines the considerations of certain subtopics in Internet Measurement. Beginning with looking at Ethics, the focus is on the role ethical considerations play during Internet Measurement. In addition the opportunities to reproduce researchers' results are analyzed. Lastly the use of geolocation products and top lists during Internet Measurement research is evaluated. The base for this paper are the publications of annual Internet Measurement Conferences.

1.1 Conferences

Internet Measurement Conferences are held yearly and give researchers from all over the world the possibility to submit papers documenting their work. The best papers, selected by the conference's jury, are then presented by their authors at the conference. This paper focuses on three Internet Measurement Conferences: The "Internet Measurement Conference" (IMC), the "Passive and Active Measurement Conference" (PAM) and the "Network Traffic Measurement and Analysis Conference" (TMA). All the papers published at these three conferences in the course of the past three years are covered, totaling 260 papers.

2. ETHICS

Ethical considerations play an important role in internet measurement. Many topics call for an assessment previous to getting started. If you want to obtain certain data and measure things important to your research, you often affect servers and resources that belong to parties not being part of the research team and not knowing about the measurement.

It is common to ask for internal approval by the university's Institutional Review Board or Ethics Committee in order to conduct such a research. In that process you explain what you are intending to do, what the effects of your actions could be and what you will do to cover these effects.

This section will evaluate how many papers at the three Internet Measurement Conferences we focused on, consider the ethical points of their research.

The selected keywords for evaluating how many papers considered the ethics of their work in the past three years were: "ethic", "ethics", "ethical", "moral", "morally" and "righteous". If they appeared in a context that was not relevant for this paper, those sentences were ignored.

2.1 Ethics in scientific papers

Out of the 260 conference papers, 44 take into account the ethical consideration of their work. They clarify why their research was ethically acceptable and explain the measures taken to assure that. Mostly the measures included securing the privacy of people while collecting data. Additionally, some papers emphasize that it is important to give thought to ethics, especially being aware that potential harm could be caused to non-participants through research.

While reading the papers, it occurs that there are two ways the authors express their consideration of ethics. Nine of them briefly talk about how they try to sustain ethical correctness. For example: "Given the ethical considerations involved, we recruited only volunteers and took an informed consent from them all after explaining and demonstrating the entire process." [14] On the other hand 35 papers enlarge into ethics in an own chapter. They talk about the importance of ethical consideration in research in a more extensive manner and project it onto their work by explaining all the steps taken they felt were necessary. Independent of how comprehensive ethics were discussed, 17 papers additionally mention going through the process of obtaining their university's approval through an Institutional Review Board.

doi: 10.2313/NET-2018-03-1_04

Year	Total papers at all conferences	Papers	Ethical consideration	Chapter on Ethics	IRB approval
2015	87	4 (4.6%)	1	3	1
2016	92	16 (17.4%)	3	13	6
2017	81	24 (30.0%)	5	19	10

Table 1: Ethical Consideration in papers

2.1.1 References to publications on ethics

Some papers refer to other researcher's work regarding ethical measures in research: "As with any active scanning research, there are many ethical considerations at play. We followed the best practices defined by Durumeric et al. [...] and refer to their work for more detailed discussion of the ethics of active scanning." [32] Another example: "Partridge and Allman [16] propose to evaluate whether the active measurements themselves or the release of the resulting data can harm an individual." [11] Partridge and Allman [25] and Durumeric et al. [10] were named by most papers referring to other researcher's work.

2.1.2 E-Mails, websites and blacklists

In order to increase the transparency of their measurement, some research teams took additional measures. This paper explains how transparency was facilitated: "To minimize the intrusiveness of our active network measurements we implemented several procedures: [...] Second, we set up a website on the scanning machines which explains our measurement activity in detail. Third, we maintain a blacklist of hosts and networks which will not be scanned in any of our measurements. Throughout the experiment, we received one e-mail out of curiosity and another one asking to be blacklisted. We complied with the blacklisting request and did not probe this network anymore." [12] Other research teams took similar measures.[35, 2, 28, 4, 29]

2.1.3 Disclosure of security weaknesses

An other ethical issue, especially in security research, is the disclosure after finding a vulnerability. On the one hand it is important to inform the community about discovered security flaws, on the other hand it could cause damage to a company's reputation. This research group informed the developers about a bug they found in their product and advised a way to correct it. They then gave them time to fix the bug before disclosing: "Finally, once we confirmed the vulnerability, we notified both Periscope and Meerkat about this vulnerability and our proposed countermeasure (directly via phone to their respective CEOs). We also promised any disclosures of the attack would be delayed for months to ensure they had sufficient time to implement and deploy a fix." [36]

2.2 Conclusion

Table 1 shows the development of ethical discussions in papers for the past three years. It should be remarked that not all paper's topics are of content that needs an ethical consideration, wherefore the percentage values can't be compared to a target percentage of 100%.

While the total number of yearly papers at the conferences barely changed, the amount of papers mentioning ethical aspects, strongly increased. The percentage of papers discussing ethics per conference underlines this improvement. Equally to the raising percentage, there is an increase in the Institutional Review Board's engagement into the ethics of a research. This overall improvement makes it clear that the Internet Measurement community considers ethical valuation in research evermore important.

3. REPRODUCTION

Reproducibility is very important in scientific research. It allows fellow researchers to acquire a more extensive understanding of a topic and gives them the opportunity to build their academic project on other researcher's work. The provided data also allows a thorough review and help's verifying the integrity of the published results.

This section focuses on papers that are careful of making their work easily reproducible. It checks the validity of internet links and the availability of data and code.

The selected keywords to find information on the reproducibility of a paper's work in the past three years were: "reproducible", "reproducibility", "reproduction", "reproduce", "reproduced", "reproducing".

3.1 Reproducibility of results

Searching the 260 papers for reproducibility, brings to light that 19 papers mention reproducible results. 15 of these papers have their focus on topics allowing direct reproduction through code and data. Four papers don't offer data, but address reproducibility or offer theoretical reproduction. Looking at the papers with topics allowing active reproduction, 14 of the 15 papers specify a direct Internet link to required code and data for reproducing their results. The paper not disclosing a link for reproductional purposes is consciously not sharing the used code. The research topic focuses on network vulnerabilities, making it irresponsible and dangerous to reveal the used implementation to the public without limitations: "For ethical reasons, we do not publish our code: it will be shared with fellow researchers and interested anti-abuse projects on a request basis." [16]

From the 14 papers providing their code and data, each paper offers a valid link. All links directly lead to a website without redirections, with 13 links leading to an online hosting servive, whereas one of the links leads to a personal university website.[7] Table 2 illustrates these results.

From the four papers not offering reproducible data, one describes the reproduction of previous experiments.[26] Furthermore two papers present methods based on publicly available data making their presented methods easily reproducible.[21, 3] The last of these four papers addresses, that there could be legal consequences making certain data available to the public, especially malware. But on the contrary also saying that "[...] the sharing and reuse of ex-

links for reproduction	
working	14
dead	0
redirecting	0
personal site	1

Table 2: Papers offering Reproduction

Year	Total papers at all conferences	Papers with instructions on reproduction	%
2015	87	4	4.6%
2016	92	5	5.4%
2017	81	10	12.3%

Table 3: Reproduction in papers

isting datasets aids reproducibility, an important scientific goal."[33]

3.2 Conclusion

The importance of reproducibility in scientific research is indisputable, therefore scientific results should be repeatable, replicable and reproducible. Nevertheless many publications in computer science still lack reproducible information, making it more desirable in the academic community for the future.
Favorably all the links supporting reproduction are valid. However, the utilization of a personal website for reproduction purposes is not useful, since there is a high probability that the link won't be working in a couple of years.
Table 3 shows the amount of papers facilitating reproduction in the past years. The amount increased every year, especially in 2017. Even though the possibility of reproducing a result is strongly dependent on the paper's topics, it seems that the awareness to support reproducibility in research raises.

4. GEOLOCATION PRODUCTS

The use of IP-based geolocation databases to determine the physical location of a server or router is a well known practice in computer science research. Those databases are very convenient to use, making it possible to refer measured data to specific cities, countries or areas.
The market offers many different products, some are fee-based, others are free. Most of them provide particular levels of location accuracy, generally differentiating between city-level and country-level accuracy.
This chapter looks at various public and commercial IP Intelligence products that offer location based information and their use in research. It puts in contrast the use of paid and free products and evaluates which geolocation accuracy levels are most commonly prefered.
The selected keywords to obtain information on the utilization of geolocation products in research in the past three years were: "geolocation", "geo-location", "geo location", "IP2Location", "MaxMind", "GeoIP", "GeoLite", "DRoP", "NetAcuity" and "Neustar".

4.1 Geolocation product use

Of the 260 examined papers, 33 mention the use of geolocation databases, which are offered by four different providers:
-IP2Location [17]: IP2Location offers a comprehensive variety of geolocation databases. The price of a database depends on the features and informations provided. The free version, IP2Location LITE [18], is limited in accuracy and number of records compared to the commercial version.
-MaxMind: MaxMind also offers commercial and free options. GeoIP [22] is the commercial and GeoLite [23] is the free database. They are both available with city-level and country-level accuracy.
-Digital Element - NetAcuity Edge [9]: The NetAcuity databases are all fee-based databases. They are available in versions with different extent. All of the researchers that were found working with NetAcuity were using NetAcuity Edge
-Neustar IP GeoPoint [24]: Neustar's geolocation database is called Neustar IP GeoPoint and is fee-based.
Reading through the papers, it stands out quite fast that MaxMind is the predominant provider. In 26 of 36 cases the researchers selected a MaxMind geolocation database.

4.1.1 Free vs Paid and City vs Country

Table 4 demonstrates a detailed overview about the accuracy levels used with each database. Comparing the use of MaxMind's free database against it's paid database, the free one is slightly favored, with GeoLite being used ten times and GeoIP nine times. Seven papers mention the utilization of MaxMind for location purposes of their measures, but do not specify the used database. Investigating the accuracy, 13 of the 24 MaxMind users utilize city-level accuracy, whereas the country-level accuracy satisfies five research groups. The other papers do not provide information on the accuracy level used and it is not possible to deduce it from the context. Regarding the four papers using IP2Location, three choose the paid database and one uses the IP2Location LITE database. Likewise do three use city-accuracy and one uses country-accuracy.
Table 5 depicts, that most researchers favor using a paid geolocation database instead of a free one and city-level accuracy rather than country-level accuracy.

4.1.2 Reflected use

Some of the papers warn of the sole use of geolocation databases and address their inaccuracies: "Unfortunately, prior work has established that IP geolocation databases are often rather inaccurate [...]. To fix inaccuracies we use two other sources to manually estimate location for 5,172 unique, routable IP addresses observed over the course of the experiment." [30]
Another example, an entire paper on geolocation databases: "Our main contributions in this paper are: (1) we show that the studied databases have many inconsistencies, especially at city-level." [13]

4.2 Conclusion

Even though there are plenty of options, a lot of research groups utilize MaxMind's services for IP-based localization. Their database is by far more frequently used than the other databases. No paper mentions a reason for their decision. One cause could be that many researchers believe it to be the most accurate IP-based localization service. An other

		MaxMind									
IP2Location		**GeoIP**		**GeoLite**		**No Info**		**NetAcuity**		**Neustar**	
4		**9**		**10**		**7**		**5**		**1**	
City	Country	City	Country	City	Country	City	Country	City	Country	City	Country
3	1	6	1	6	3	1	1	2	2	1	0

Table 4: Accuracy level used with every geolocation database

city vs. country accuracy		paid vs. free services	
city	19	paid	17
country	8	free	12
no information	9	no information	7

Table 5: geolocation accuracy and payment

reason for MaxMind's predominance could be, that because of the frequent use, it is the best-known and benefits from that scenario. Since there are many papers using geolocation services but not informing about the service they select, it is harder to assess the numbers. Focusing on the papers, that say which database is utilized, the paid version and city accuracy is preferred. This decision probably results from the researchers' need of city-accuracy and the paid databases being more precise. Although even the paid ones might have high inaccuracy on city level. Table 6 shows that the use of geolocation products decreased. Since it depends strongly on the research topics presented at the conferences, the decrease has no significance of geolocation databases being used less in internet measurement research.

Two papers did not disclose information about the uses geolocation databases [19, 5] One paper [29] used both, MaxMind GeoLite City and IP2Location. Another paper uses 4 databases (all 5 from Table 4 except Neustar). For this reason the numbers in Table 4 total 36.

5. TOP LISTS

To conduct website measurements it is favorable to know which the most popular websites on the internet are. That way the measurement is done on relevant pages and results in a meaningful outcome. Different companies focus on ranking websites, creating a list of the most popular websites, called top lists. An arbitrary number of the top websites on these lists is often selected by researchers to be included into their studies. Most top lists are created by ranking websites by page views and people using the website in a certain time period. This section evaluates the most commonly used lists in internet measurement research. Furthermore the number of selected top sites is examined and compared against each other. Lastly the researchers' motivation to chose a particular list is looked at.

Year	Total papers at all conferences	Papers using geolocation products	%
2015	87	15	17.2%
2016	92	9	9.8%
2017	81	9	11.1%

Table 6: Use of geolocation products

The selected keywords to find the top lists used by researchers in the past three years were: ”Alexa”, ”Umbrella”, ”Quantcast”, ”SimilarWeb”, ”toplist”, ”top list” and ”top domain”.

5.1 Top list use

From the 260 examined papers, 56 use a top list as base for their Internet measurement. Table 10 shows the distribution over the past three years. These 56 papers utilize the top lists of only three different companies. Most common are the Alexa to plists [1]. Alexa offers a free of charge Top 1 Million global list. Furthermore Alexa provides a top list for every country it has enough data for, as well as a top list for various categories. The other two lists used in the papers are the Umbrella 1 Million list [6] and the Quantcast Top Websites list [27]. The Umbrella 1 Million list is a free list provided by Cisco with a global ranking of websites. The Quantcast Top Websites list offers a top list for every country and is also free of charge.

Surprisingly, every single research group used a top list ranked by Alexa for their analysis. The Umbrella list and the Quantcast list are only combined with an Alexa list in a single case respectively [2, 8], meaning that 54 of 56 research groups chose an Alexa list and two groups decided to amplify an Alexa list. One group used the Alexa Top 1000 domains from the traditional top lists and extended it with the websites of the 500 world's biggest companies according to Forbes Magazine.[37]

5.1.1 *Comparison of individual top lists*

Among the Alexa lists, the global lists were used more frequently than the country lists and the category lists. For the global lists the domains of the Top 1 Million list were applied the most, with 27 papers using it. Second are the domains from the Top 500 list, being used by seven papers. Table 7 shows which top domains were used by how many research groups.

The top list of a country was used in eight papers and the top list of a category was utilized by six research groups. Table 8 shows the different category lists used, while the News & Media category was applied twice . One paper uses the Top 500 list of each category [20], it is excluded of Table 8.

5.1.2 *Motivation and reflected use*

From 56 papers using Alexa's top domains, two papers describe what the top lists are and briefly elucidate the functionality of the website's ranking. [31, 15] None of the 56 papers explain the reasons for choosing an Alexa list over other top lists. Reflecting about the use of an Alexa top list, there are two papers mentioning a negative perception: ”While Alexa provides high-quality rankings for the most popular sites, our experience has shown it to be less reliable

Global top lists	
Top 500	7
Top 1K	5
Top 10K	3
Top 100K	4
Top 1M	27
No information	2

Table 7: Alexa's TopX domain usage

Category lists	
Adult	Top500
E-Mail	Top10K
News & Media	Top500
Region	Top500

Table 8: Category lists used with the amount of top domains

for the long tail of the distribution.” [8]Another example: ”We explicitly avoid using the Alexa ranking since it includes services which are questionable for some categories.” [34]

5.2 Conclusion

It becomes clear that Alexa's top lists are prevalent in the Internet Measurement community. Almost all the papers mention the utilization of Alexa's ranking without explaining what it is and does, like it was natural to use that particular ranking. This implicitness is probably also the point why barely any paper describes the reason for selecting Alexa's top lists over another one. The Alexa Top 1 Million list is by far the most frequently used. Since all the papers use Alexa, the Top 1 Million list is also the most frequently used in research. Using many domains enables researchers to measure more data that still is significant, since it is ranked high. Looking at Table 10 there is a frequent use of top lists at Internet Measurement Conferences. Although they are not necessary for every research topic, their use has increased. Nevertheless, to state that more researchers use top lists for their internet scans seems difficult, since the use depends a lot on the research topics. Therefore the increment could just be arbitrary and not meaningful.

Year	Total papers at all conferences	Papers using top list	%
2015	87	16	18.4%
2016	92	21	22.8%
2017	81	19	23.5%

Table 10: Use of top lists

6. CONCLUSION

On the whole every category is represented with a good amount of papers regarding the distribution of research subjects and their topics. Table 9 depicts the complete distribution of papers for all the categories and all the conferences in the past three years. The percentage values refer to the total of 260 papers. It should be noted again, that none of the four categories could possibly be discussed in every paper. The percentage values are just an information, they should not be compared to a target percentage of 100%. Top lists are most commonly used in Internet Measurement, while the opportunity for reproduction is mentioned the least. As said in the previous chapters, this is a matter of topics and can not be compared. Ethics and reproduction opportunities strongly increased at the conferences in the past three years. These two categories are important to mention if they fit the research topic. More and more research teams acknowledge that fact. On the other hand the use of geolocation products and and top lists had a more random variation of mentions in the three years, most likely due to arbitrary topic decisions.

Conference	Papers total	Ethics	Reproduction	Geolocation	Top lists
TMA 2015	16	1	1	2	2
TMA 2016	16	4	1	1	3
TMA 2017	19	8	4	2	3
PMA 2015	27	0	1	4	4
PMA 2016	30	1	1	4	4
PMA 2017	20	3	2	2	5
IMC 2015	44	3	2	9	10
IMC 2016	46	11	3	4	14
IMC 2017	42	13	4	5	11
Total	**260**	**44**	**19**	**33**	**56**
Rate	-	**16.9%**	**7.3%**	**12.7%**	**21.5%**

Table 9: Complete statistics on all papers

7. REFERENCES

[1] Alexa Top 500 sites on the web. https://www.alexa.com/topsites.

[2] J. Amann, O. Gasser, Q. Scheitle, L. Brent, G. Carle, and R. Holz. Mission Accomplished? HTTPS Security after DigiNotar. IMC, nov 2017.

[3] S. Benitez-Baleato, N. B. Weidmann, P. Gigis, X. Dimitropoulos, E. Glatz, and B. Trammell. Transparent estimation of internet penetration from network observations. PAM, 2015.

[4] R. Beverly. Yarrp'Ing the Internet: Randomized High-Speed Active Topology Discovery. IMC, 2016.

[5] I. N. Bozkurt, A. Aguirre, B. Chandrasekaran, P. B. Godfrey, G. Laughlin, B. Maggs, and A. Singla. Why Is the Internet so Slow?! PAM, 2017.

[6] Cisco Umbrella. https://umbrella.cisco.com/blog/2016/12/14/cisco-umbrella-1-million/.

[7] G. Comarela, E. Terzi, and M. Crovella. Detecting Unusually-Routed ASes: Methods and Applications. IMC, 2016.

[8] J. DeBlasio, S. Savage, G. M. Voelker, and A. C. Snoeren. Tripwire: Inferring Internet Site Compromise. IMC, 2017.

[9] Digital Element - NetAcuity Edge Premium. https://www.digitalelement.com/solutions/netacuity-edge-premium/.

[10] Z. Durumeric, E. Wustrow, and J. A. Halderman. ZMap: Fast Internet-wide Scanning and Its Security Applications. USENIX Security 13.

[11] O. Gasser, F. Emmert, and G. Carle. Digging for Dark IPMI Devices: Advancing BMC Detection and Evaluating Operational Security. TMA 2017.

[12] O. Gasser, Q. Scheitle, S. Gebhard, and G. Carle. Scanning the IPv6 Internet: Towards a Comprehensive Hitlist. TMA 2016.

[13] M. Gharaibeh, A. Shah, B. Huffaker, H. Zhang, R. Ensafi, and C. Papadopoulos. A Look at Router Geolocation in Public and Commercial Databases. IMC, Nov 2017.

[14] P. Gupta, M. Patel, and K. Chebrolu. *Cutting Internet Access Costs Through HTTPS Caching: A Measurement Study*. PAM 2017.

[15] T. Halvorson, M. F. Der, I. Foster, S. Savage, L. K. Saul, and G. M. Voelker. From .academy to .zone: An Analysis of the New TLD Land Rush. IMC, 2015.

[16] L. Hendriks, R. de Oliveira Schmidt, R. van Rijswijk-Deij, and A. Pras. On the Potential of IPv6 Open Resolvers for DDoS Attacks. PAM, 2017.

[17] IP2Location. https://www.ip2location.com/.

[18] IP2Location LITE - Free Databases for Download. https://lite.ip2location.com/.

[19] M. Javed, C. Herley, M. Peinado, and V. Paxson. Measurement and Analysis of Traffic Exchange Services. IMC, 2015.

[20] A. J. Kaizer and M. Gupta. Characterizing Website Behaviors Across Logged-in and Not-logged-in Users. IMC, 2016.

[21] A. Marder and J. M. Smith. MAP-IT: Multipass Accurate Passive Inferences from Traceroute. IMC, 2016.

[22] MaxMind GeoIP2 Databases. https://www.maxmind.com/en/geoip2-databases.

[23] MaxMind GeoLite2 Free Downloadable Databases. https://dev.maxmind.com/geoip/geoip2/geolite2/.

[24] Neustar IP GeoPoint. https://www.risk.neustar/ip-intelligence/ip-address-data.

[25] C. Partridge and M. Allman. Ethical Considerations in Network Measurement Papers. *Commun. ACM*, Sept. 2016.

[26] D. A. Popescu and A. W. Moore. Reproducing Network Experiments in a Time-controlled Emulation Environment. TMA, 2016.

[27] Quantcast Top Websites. https://www.quantcast.com/top-sites/.

[28] Q. Scheitle, O. Gasser, M. Rouhi, and G. Carle. Large-Scale Classification of IPv6-IPv4 Siblings with Variable Clock Skew. TMA, 2017.

[29] Q. Scheitle, O. Gasser, P. Sattler, and G. Carle. HLOC: Hints-Based Geolocation Leveraging Multiple Measurement Frameworks. TMA, 2017.

[30] P. Schmitt, M. Vigil, and E. Belding. *A Study of MVNO Data Paths and Performance*. PAM. 2016.

[31] P. Snyder, L. Ansari, C. Taylor, and C. Kanich. Browser Feature Usage on the Modern Web. IMC, 2016.

[32] D. Springall, Z. Durumeric, and J. A. Halderman. Measuring the Security Harm of TLS Crypto Shortcuts. IMC, 2016.

[33] D. R. Thomas, S. Pastrana Portillo, A. Hutchings, R. N. Clayton, and A. R. Beresford. Ethical issues in research using datasets of illicit origin. 2017.

[34] S. Traverso, M. Trevisan, L. Giannantoni, M. Mellia, and H. Metwalley. Benchmark and Comparison of Tracker-blockers: Should You Trust Them? TMA, 2017.

[35] B. VanderSloot, J. Amann, M. Bernhard, Z. Durumeric, M. Bailey, and J. A. Halderman. Towards a Complete View of the Certificate Ecosystem. IMC, 2016.

[36] B. Wang, X. Zhang, G. Wang, H. Zheng, and B. Y. Zhao. Anatomy of a Personalized Livestreaming System. IMC, 2016.

[37] H. Zhang, M. Gharaibeh, S. Thanasoulas, and C. Papadopoulos. BotDigger: Detecting DGA Bots in a Single Network. TMA, 2016.