Falsifiability of network security research: the Good, the Bad, and the Ugly

Dennis Gamayunov
Lomonosov Moscow State University
gamajun@cs.msu.su

ABSTRACT

A falsifiability criterion helps us to distinguish between scientific and non-scientific theories. One may try to raise a question whether this criterion is applicable to the information security research, especially to the intrusion detection and malware research fields. In fact, these research fields seems to fail to satisfy the falsifiability criterion, since they lack the practice of publishing raw experimental data which were used to prove the theories. Existing public datasets like the KDD Cup'99 dataset and VX Heavens virus dataset are outdated. Furthermore, most of current scientific research projects tend to keep their datasets private. We suggest that the scientific community should pay more attention to creating and maintaining public open datasets of malware and any kinds of computer attack-related data. But how can we bring this into reality, taking into account legal and privacy concerns?

Categories and Subject Descriptors

Security and privacy [**Intrusion/anomaly detection and malware mitigation**]: Malware and its mitigation; Security and privacy [**Intrusion/anomaly detection and malware mitigation**]: Intrusion detection systems

General Terms

Security, Experimentation

Keywords

network security, intrusion detection, malware analysis, research methodology

1. INTRODUCTION

Back in the 1930-s the philosopher of science Karl Popper suggested a simple criterion for distinguishing scientific theories from everything else. He proposed falsifiability – the ability of theories to come in conflict with observation – as the landmark of empirical theories, and falsification – the search for observations that conflict with the theory – as an empirical method for replacing verifiability and induction by purely deductive notions. He claimed that there may be one universal method for checking any new theory: the negative method of criticism, trial and error [9]. And we should say that one of the keystones of falsifiability is a proper description of an experiment, including its conditions and all natural phenomena involved.

TRUST'14 June 09-11 2014, Edinburgh, United Kingdom

ACM 978-1-4503-2951-4/14/06 ...$15.00.
http://dx.doi.org/10.1145/2618137.2618141

2. NETWORK SECURITY AS A SCIENCE

Speaking of the information security field we should acknowledge that the state-of-the-art here is not quite smooth. Part of information security subfields such as cryptography and cryptanalysis satisfy the falsifiability criterion. But in the field of intrusion detection and malware analysis the situation is different: according to the current research practice it is common for researchers to publish only final results even at the highest rank conferences, while they rarely publish raw experimental data, if ever. As a result, trying to verify and possibly falsify published results becomes a challenging task for the community, infeasible in some cases. Researchers have to trust the results relying on the author's reputation or try to get similar results using their own experimental data. Therefore, we can state that, currently, the intrusion detection field does not satisfy Popper's criterion.

This problem could be addressed by creating some public datasets which most of researchers would treat as reliable enough to use them for comparing each other's results. Recently there were several attempts to create such public datasets. For example, in 1999 the KDD Cup'99 [6] dataset for intrusion detection systems and the VX Heavens virus collection [3] were published. Unfortunately, the KDD'99 dataset is completely outdated, and using it in intrusion detection research is not only useless but strongly not recommended – any paper which relies on this dataset would be rejected by the peer review process. The VX Heavens dataset now in 2014 could possibly be used in student-level research work, but it surely cannot be treated as a strong basis for making conclusions about modern malware.

There are several ongoing research projects which collect malware from different sources, both binary malware and web-specific JavaScript – for example, CWSandbox [12], Anubis [4, 1], Wepawet [5]. For some reason none of these projects provide public access to their malware collection, they only publish the final results of malware analysis. In case of CWSandbox such behavior is quite understandable, because the project became a core of Sunbelt Software's

Table 1: Datasets citation rates according to Google Scholar

Dataset name	Number of citations	Year of initial publication	Average citations per year
KDD Cup 99 dataset	4,380	1999	292
VX Heavens	8,450	1999	563
Metasploit Framework	1,630	2003	148
Anubis	482	2007	68
CWSandbox	641	2006	80
Wepawet	153	2008	25

business, and the malware collection is a significant part of its assets. Maybe authors of Anubis and Wepawet have similar reasons for making their collection closed to the public. But as the result the scientific community receives a number of research papers, which do not seem to fully satisfy the falsifiability criterion, because malware demonstrates high variability rates and direct reproduction of the malware dataset is near to impossible. For example, any observations about computer attacks or malware propagation characteristics revealed from some large months-long traffic dump from East-Asian Tier-2 ISP may turn out to be unobservable at the same Tier-2 level but in Northern Europe – because of regional differences in distribution of popular applications, social networks and so on. On the other hand, in those cases when community does have publicly available benchmark, it requires an order of magnitude less effort for researchers to check, prove or falsify results of one another.

We could also make an important notice – any publicly available dataset is a great stimulus for research activity by itself. Let us compare the citation rates for the datasets mentioned above counted using Google Scholar. The citation rates in the table 1 show that the overall number of research papers which use publicly available datasets by at least one decimal exponent outnumbers the research papers which use private datasets, including citations. Metasploit Framework [7] is a good borderline example of a publicly available benchmark for malware and attack detection tools. It has a very limited set of real-life exploits for popular application vulnerabilities, which is presumably far richer in reality. However it does have a number of very nice features. One of them, for example, is the generation of numerous polymorphic variations of the same exploit which enables researchers to test and compare their algorithms with something close to the ground-truth data. As a result the overall amount of publications using Metasploit as a source of malware samples is quite high in comparison to those who utilize real malware collections keeping them private.

There are also research projects whose primary result is an alternative implementation of some known method (for example, CUDA API port of popular regular expression matching algorithms) or implementation of some author's idea. For such projects, the availability of implementations to the community is essential for the falsifiability of the published results. But it is also quite common for this kind of research not to publish the implementation, even under non-disclosure agreement, and the most popular response for one's request is often: *"Thanks for writing. We won't be releasing the implementation. Sorry that I can't be more helpful"*.

2.1 Impact of ignoring the falsifiability

When it comes to the question of how we learn about computer attacks, network worms propagation, distributed denial of service and any kind of cybercrime activity, the answer is - commercial intrusion detection systems. Nowadays there is quite a number of intrusion detection systems available on the market (or intrusion prevention systems - IPS, or unified threat management - UTM, the distinction is irrelevant in this paper). Major networking equipment vendors like Cisco Systems, Intel, and IBM have intrusion detection solutions in their hardware list. There are also many specialized vendors like Sourcefire, Arbor, Narus and so on. Actually, the intrusion detection field is now over 30 years old, and one could name over 200 research projects in intrusion detection and over 30 currently existing vendors of commercial systems of this kind. A list of 132 research project can be found in [11], and this list is not thorough at all.

Despite such a significant number of research projects and solutions available on the market, there is no benchmark or common methodology which could give us a short answer to the simple question - how could we compare the efficiency of two given intrusion detection systems? Can we actually trust to those characteristics which hardware vendors list at their websites and in whitepapers? What will be the actual false positives rates and percentage of missed attacks (including 0-days) if we put a given IDS into our network? No one knows. As we have pointed out earlier, the network security research lacks falsifiability, and as a result we now neither have any international standards for intrusion detection, nor any kind of common benchmark or methodology for evaluation of IDS efficiency. There is even no standard definition for the "attack", and no clear common view of what we should monitor and analyze to distinguish between "attack" and "normal". But we do use these systems for monitoring security state of the real-world networks and networking channels, and we do use results of such monitoring for real-world decision making.

Actually, there are a number of commercial laboratories which try to fill this gap. Companies like NSS Labs develop their own methodologies and datasets for benchmarking different kinds of network security equipment, including intrusion detection systems [8]. But again, they are doing it to earn money - the reports are only available for a fee, and the datasets used for testing are not publicly available.

There are also documents like NIST recommendation for intrusion detection [10], which try to give answers to such question as: where should IDS be placed? How do we choose appropriate type of IDS according to our needs? How do we

tune it to gain optimal efficiency? But the answers given are too general and abstract, and the result of IDS tuning may alter its behavior and efficiency dramatically, by tens of percents. That is why we cannot use such documents as international standards.

3. CONCLUSIONS

Having studied hundreds of intrusion detection and malware detection projects one may come to a conclusion that the network security field, especially intrusion detection, has stuck at the point where major research works have already been done and no breakthrough is taking place. In fact there are certain developments which drive it beyond what we saw ten years ago and earlier: modern hardware architectures (and even CPU architectures) have started embedding anti-malware techniques and trusted platforms have been facing an obvious rise. But still, many modern commercial IDS/IPS systems repeat each other in many ways, and signature-based detection remains the primary method of detection. Any direct attempt to standardize this field would come to be faced with the extreme complexity of the task of building complete and thorough benchmarks. But it also seems that we could make the situation much better if we stimulate research by providing open and public datasets, and also by stimulating sharing of raw experimental data between researchers from different countries.

It seems reasonable for the information security community and national governments to support creating open and public collections of up-to-date malware along with results of its preliminary analysis. And what seems to be most important - it is necessary to recover the practice of publishing raw experimental data, on which the research results rely. The overall experience of the information security field and other natural sciences demonstrates that publicity of this kind always greatly encourages both quality and quantity of research projects. The open science data movement [2] is actively promoted in many other natural sciences like chemistry or biology. Some of these fields face the same privacy concerns as the information security field. Maybe we should try to bring their more than 50 years of experience into our own?

There are several obvious issues regarding data sharing of security-related data. There are legal issues: it is often illegal to openly share malware. For example, Russian law-enforcing agencies tend to shutdown any malware sharing websites hosting in Russia, even when malware is shared for research only. And it is even more difficult to share meaningful traffic dumps containing real attacks because of privacy concerns and existing corpus of laws which protect privacy – here the situation is similar in EU, USA and Russia. But on the other hand, cyber-criminals are not bound to comply with these laws, and therefore they are always at least one step ahead of researchers. Should not the community develop legal frameworks which would allow us to change the situation? Supposedly, we should.

The research community might take some steps to broaden data sharing. At first it might be helpful to extend network security academic conference acceptance rules to put emphasis on raw data sharing. At the same time reviewers could take into account the availability of the datasets which were used to prove key evaluation results. Moreover academic institutions and companies may provide free storage space for sharing experimental data. These actions require very little effort though may draw researchers' attention to the importance of data sharing and making their theories more falsifiable.

4. REFERENCES

[1] Anubis - Malware Analysis for Unknown Binaries. http://anubis.iseclab.org/, 2014. [Online; accessed 05-May-2014].

[2] Open Science Data Movement. http://en.wikipedia.org/wiki/Open_science_data, 2014. [Online; accessed 05-May-2014].

[3] VX Heavens virus collection. http://vxheaven.org/, 2014. [Online; accessed 05-May-2014].

[4] U. Bayer, I. Habibi, D. Balzarotti, E. Kirda, and C. Kruegel. A view on current malware behaviors. In *USENIX workshop on large-scale exploits and emergent threats (LEET)*, 2009.

[5] S. Ford, M. Cova, C. Kruegel, and G. Vigna. Wepawet. http://wepawet.cs.ucsb.edu/, 2009. [Online; accessed 05-May-2014].

[6] S. Hettich and S. Bay. The UCI KDD archive. http://kdd.ics.uci.edu/databases/kddcup99/, 1999.

[7] H. Moore. Metasploit framework. http://www.metasploit.com/, 2004.

[8] NSS Labs. Network intrusion prevention test reports. https://www.nsslabs.com/reports/categories/test-reports/network-intrusion-prevention, 2014. [Online; accessed 05-May-2014].

[9] K. Popper. *The logic of scientific discovery*. Routledge, 2005.

[10] K. Scarfone and P. Mell. Guide to intrusion detection and prevention systems (IDPS). *NIST special publication*, 800(2007):94, 2007.

[11] S. Schmerl and M. Meier. Intrusion detection systems list and bibliography. https://www-rnks.informatik.tu-cottbus.de/en/node/209, 2010. [Online; accessed 05-May-2014].

[12] C. Willems, T. Holz, and F. Freiling. Toward automated dynamic malware analysis using CWSandbox. *IEEE Security and Privacy*, 5(2):32–39, 2007.