Fabio Massacci
Associate Editor in Chief

The Holy Grail of Vulnerability Predictions

Scoring vulnerabilities is a hard task, and we build standards for this purpose: the Common Vulnerability Scoring System (CVSS) used by NIST is the oldest of them. It was invented by Peter Mell and Karen Scarfone, among others, to assess *severity*.[1]

After almost 10 years at the CVSS Special Interest Group of FIRST, I can say that each time a vulnerability was brought to a meeting to be discussed and dissected, eventually, everybody agreed on whether the impact was high or whether this vulnerability required an active interaction by the user and so on. This anecdotal evidence is also confirmed with experiments: experts and students quickly converge.[2] Assessing the facets of severity is doable.

The point where stronger disagreement starts is when one needs to order them along the real line: Is this worse than that? Most importantly, should I worry? Scoring metrics have always been hotly debated.[3] In the desperate quest for popularity, a security researcher has even started to use ChatGPT to triage scoring metrics and reported "anomalies" in the scoring (https://www.linkedin.com/posts/parisel_cvss-activity-7112793826427031553-VpGS/). These anomalies are of course simple hallucinations[4] and misinterpretations of Pareto frontiers.

Yet severity is the only start of the journey. It does not tell what everybody really wants to know: Will this vulnerability be exploited any time soon? This information might have practical significant implications for a company, for example, on adopting just patching (if they are too many) or rather updates and bug hunting (if they are few).[5] In a retrospective study with my former student Giorgio Di Tizio, we found that updates do not always work. You can update all the time instantaneously, or you can be just a normal company following the industry standard lag of 30 days. However, in the latter case, your risk profile, i.e., the odds of succumbing, will not be better off than the company who just patches high-risk vulnerabilities, even against advanced persistent threats.[6] So why bother toiling as the hare if you fall prey to the same wolves as the tortoise? You can watch the video (https://vimeo.com/853062910) accompanying the *Communications of the ACM* shorter piece[7] or read the rebuttal by Steve Lipner and John Pescatore warning against such dangerously heretical ideas.[8]

The point where stronger disagreement starts is when one needs to order them along the real line: Is this worse than that? Most importantly, should I worry?

Mehran Bozorgi and his colleagues have shown long ago that CVSS scores alone are not effective for predictions,[9] albeit they can be a reasonable proxy (if the hole is big enough and comfortable enough, somebody might be tempted to lodge there). Similarly, in a paper with Luca Allodi, we showed that combining the CVSS with information on whether the affected vulnerabilities being sold in the black markets was a better proxy.[10] Ours was a retrospective study, but what about prospective studies?

Finding prospective metrics or more alluring "*predictive metrics*" is the new quest for the

Digital Object Identifier 10.1109/MSEC.2023.3333936
Date of current version: 19 January 2024

Holy Grail where new metrics and standards proliferate.[11] The most popular one is the Exploit Prediction Scoring Systems (EPSS), started by Jay Jacobs and Sasha Romanosky, which aggregated several community metrics.[12] Yet there are many more metrics. People debate on LinkedIn whether the pet metric of their company is better than the other one. All these studies are based on some form of threat intelligence whose value is often even more debatable than scoring mechanisms.[13]

What I found surprising is that even reputed data scientists, when they presented their dataset-based scoring mechanisms (or other even more patchy mechanism), failed to see that they were not predicting but just retrospectively measuring notices of exploits, a typical foundational mistake in security measures.[14] As a data scientist you can use tweets mentioning a CVE and argue that they "predict" exploits. If you reflect on it, why should people tweet about a vulnerability among the hundreds discovered daily if not to report that this has been exploited in the wild? This is not a prediction. It is just an after the fact highlight.

The illusion of being able to predict the future is too hard to let go of, and reviewers even at top conferences are equally prone to the temptation of seeing predictions where there is only the ability to summarize the news. The most egregious recent example is the metrics of expected exploitability,[15] which "predicts" the ability to have a functional exploit by using as input the presence of functional instructions on the web… How come we should not be surprised that they achieved >80% accuracy?

Governments are not helping either: CISA has come out with Known Exploited Vulnerabilities (KEVs), and the Dutch government proposes likelihood (or chance) metrics where various indexes are added up but have no mathematical background. In this Mare Magnum of metrics, companies have to disentangle themselves, and "best practices" increasingly look like Macbeth's witches' recipes: one plate of CVSS, two cups of EPSS, a sprinkle of KEVs, add your pet vulnerability scoring system, shake well, and drink in one sip.

The illusion of being able to predict the future is too hard to let go of, and reviewers even at top conferences are equally prone to the temptation of seeing predictions where there is only the ability to summarize the news.

The question is whether we should really continue to look for *the* metrics to predict future exploits. I am afraid these noble attempts are better described by Mark Twain in *A Connecticut Yankee in King Arthur's Court*:

> "The boys all took a flier at the Holy Grail now and then. It was a several years' cruise. They always put in the long absence snooping around, in the most conscientious way, though none of them had any idea where the Holy Grail really was, and I don't think any of them actually expected to

Digital Object Identifier 10.1109/MSEC.2023.3337935

> find it, or would have known what to do with it if he had run across it. [...]. Every year expeditions went out holy grailing, and next year relief expeditions went out to hunt for them."

Are you planning an expedition any time soon?

Rather than doing it, you might reconsider the question I asked at the very beginning: "Will this be exploited soon?" You may want to conclude that the question is unanswerable. And if it is unanswerable, what are the consequences? Setting prospective studies aside, is there even value, then, in continuing work on retrospective studies?

I am a bit skeptical of prospective studies, but I see value in retrospective studies as they provide information on what happened, and we could use this information to build systems that address a different question: do not ask "What I do to avoid being exploited?" but "What will I do when I am exploited?" Defense in depth is the long-term answer, and in this respect, you may want to read the piece by Eric Bodden and his colleagues in the "Building Security In" department.[A1] ■

Appendix: Related Article

A1. E. Bodden, J. Pottebaum, M. Fockel, and I. Gräßler, "Evaluating security through isolation and defense in depth [Building Security In]," *IEEE Security Privacy*, vol. 22, no. 1, pp. 69–72, Jan./Feb. 2024, doi: 10.1109/MSEC.2023.3336028.

References

1. K. Scarfone and P. Mell, "An analysis of CVSS version 2 vulnerability scoring," in *Proc. 3rd Int. Symp. Empirical Softw. Eng. Meas.*, Oct. 2009, pp. 516–525, doi: 10.1109/ESEM.2009.5314220.
2. L. Allodi, M. Cremonini, F. Massacci, and W. Shim, "Measuring the accuracy of software vulnerability assessments: Experiments with students and professionals," *Empirical Softw. Eng.*, vol. 25, pp. 1063–1094, Mar. 2020, doi: 10.1007/s10664-019-09797-4.
3. J. Spring, E. Hatleback, A. Householder, A. Manion, and D. Shick, "Time to change the CVSS?" *IEEE Security Privacy*, vol. 19, no. 2, pp. 74–78, Mar./Apr. 2021, doi: 10.1109/MSEC.2020.3044475.
4. M. Zhang, O. Press, W. Merrill, A. Liu, and N. A. Smith, "How language model hallucinations can snowball," 2023, *arXiv:2305.13534*.
5. J. M. Spring, "An analysis of how many undiscovered vulnerabilities remain in information systems," *Comput. Secur.*, vol. 131, Aug. 2023, Art. no. 103191, doi: 10.1016/j.cose.2023.103191.
6. G. Di Tizio, M. Armellini, and F. Massacci, "Software updates strategies: A quantitative evaluation against advanced persistent threats," *IEEE Trans. Softw. Eng.*, vol. 49, no. 3, pp. 1359–1373, Mar. 2023, doi: 10.1109/TSE.2022.3176674.
7. F. Massacci and G. Di Tizio, "Are software updates useless against advanced persistent threats?" *Commun. ACM*, vol. 66, no. 1, pp. 31–33, 2022, doi: 10.1145/3571452.
8. S. Lipner and J. Pescatore, "Updates, threats, and risk management," *Commun. ACM*, vol. 66, no. 5, pp. 21–23, 2023, doi: 10.1145/3587826.
9. M. Bozorgi, L. K. Saul, S. Savage, and G. M. Voelker, "Beyond heuristics: Learning to classify vulnerabilities and predict exploits," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 105–114, doi: 10.1109/TSE.2022.3176674.
10. L. Allodi and F. Massacci, "Comparing vulnerability severity and exploits using case-control studies," *ACM Trans. Inf. Syst. Secur. (TISSEC)*, vol. 17, no. 1, pp. 1–20, 2014, doi: 10.1145/2630069.
11. T. H. Le, H. Chen, and M. A. Babar, "A survey on data-driven software vulnerability assessment and prioritization," *ACM Comput. Surv.*, vol. 55, no. 5, pp. 1–39, 2022, doi: 10.1145/3529757.
12. J. Jacobs, S. Romanosky, O. Suciu, B. Edwards, and A. Sarabi, "Enhancing Vulnerability prioritization: Data-driven exploit predictions with community-driven insights," in *Proc. IEEE Eur. Symp. Secur. Privacy Workshops (EuroS&PW)*, Jul. 2023, pp. 194–206, doi: 10.1109/EuroSPW59978.2023.00027.
13. X. Bouwman, H. Griffioen, J. Egbers, C. Doerr, B. Klievink, and M. Van Eeten, "A different cup of {TI}? The added value of commercial threat intelligence," in *Proc. 29th USENIX Secur. Symp. (USENIX Security)*, 2020, pp. 433–450, doi: 10.5555/3489212.3489237.
14. H. Chen, R. Liu, N. Park, and V. S. Subrahmanian, "Using Twitter to predict when vulnerabilities will be exploited," in *Proc. 25th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2019, pp. 3143–3152, doi: 10.1145/3292500.3330742.
15. O. Suciu, C. Nelson, Z. Lyu, T. Bao, and T. Dumitraș, "Expected exploitability: Predicting the development of functional vulnerability exploits," in *Proc. 31st USENIX Secur. Symp. (USENIX Security)*, 2022, pp. 377–394.