VITAL STATISTICS IN CYBER PUBLIC HEALTH

Technical Report 22-02

The role of vital statistics in public health and the challenges of gathering and generating these statistics in cybersecurity

Prepared by the CyberGreen Institute
March 2022

TABLE OF CONTENTS

1. INTRODUCTION

This paper is part of a continuing effort to improve the rigor and grounding of a Cyber Public Health project, and does so by introducing the concept of vital statistics, their role in public health, and the challenge of gathering and generating this data in cyber public health.

2. VITAL STATISTICS IN PUBLIC HEALTH

Vital statistics are systematically collected data on births, deaths, marriages, divorces and other life events. More broadly, they include data on health and disease. The then-new collection and publication of vital statistics enabled some of the early successes of public health, by allowing comparisons of prevalence of disease and comparison to normal rates.

THE ROLE OF VITAL STATISTICS

There are a great many ways in which data about diseases and health problems are measured, and many of them rely on data about populations to assess measurements from incidence to prevalence to lifetime likelihood to mortality rate.

As an example of how prevalent the use of these statistics is, one article asserts that "Congenital heart disease (CHD), the most common congenital birth defect, has long been known as one of the main causes of infant death during the first year of life. More than one million of the world's approximately 135 million newborns are born each year with CHD" (Xu et al. 2021). Note the use of statistics to contextualize the work: how many children are born, how many of those are born with CHD, and that it is the most common congenital birth defect. We can see why the authors care about this problem.

3. CYBERSECURITY'S CHALLENGE

We have no vital statistics in cybersecurity. There are some obvious analogs. For example, we might define births as new computers sold. If we use the number of computers sold in a year, for which numbers are available and compiled by industry analysts (e.g. Doshi 2021), it may be easy. However, the numbers are there for desktops, notebooks and workstations. They exclude phones, and tablets, servers, and connected devices in the Internet of Things (IoT). Each of those is a device, and so may be counted, perhaps at point of sale. There are very rough statistics, with resolution on the scale of a hundred thousand hosts, or networks. Examples can be found in Geer and Vixie (2018).

WHAT "BIRTHS" TO COUNT?

Servers are more interesting in two ways. First, many of them are now built by the large tech companies like Amazon or Google for use in their data centers, and second, the software they run is different.

Amazon's data centers run an incredible variety of virtual machines. A mere twenty years ago, the software on those would have been tied to a server, and rarely changed. Now, it is commonplace to

create, run, and destroy the software stack that runs on those virtual machines perhaps even daily. So, do we count each creation of an Amazon Machine Image (AMI) as a birth? Is each new instance of an AMI like a birth?

Some of the virtual machines which Amazon runs host Lambdas: short lived functions where Amazon provides many of the services, and simply calls a routine in a Python or Java program. The rest of the server is managed by Amazon on the customer's behalf.

Google, meanwhile, is reputed to consider their computers to all be part of a giant computer which spans the planet, and has internal software that partitions resources in that massive system. For example, "Google's Borg system is a cluster manager that runs hundreds of thousands of jobs, from many thousands of different applications, across a number of clusters each with up to tens of thousands of machines" (Verma et al. 2015). Those tens of thousands of machines are being managed as if they're parts of a single, larger supercomputer system. So, is adding a computer to a Borg cluster like adding a laptop, or more like adding an extra hard drive to a laptop?

It may also be useful to have counts of accounts, because as passwords or "records" are leaked, such data may contextualize that.

Counting IP addresses does not directly help since IP addresses are often behind Network Address Translation (NAT) technology.

New domains created may be typosquatting or impersonating (cybergreem.net, cybergreen.com). The proliferation of such domains is a problem we have not yet studied within the Cyber Health metaphors.

WHAT "DEATHS" TO COUNT?

We make statements like "my computer died." However, we also resurrect them, either by repairing that physical computer, or by getting a new one and installing from a backup. If the reasons to count deaths include understanding the rate of deaths, and thus their causes, then the hardware no longer working certainly counts. What of the hardware being sold onward, and the hardware that's literally sitting unplugged on a shelf?

Software that is at a formal "end of life" or "end of support" can still be used, and there may be an analogy to the increasing medical cost of last year of life: software that is unsupported has higher costs and more emergencies. It is often kept around not out of love or respect, but the high costs of replacement. Related to this is the software or deployed system that can no longer be replicated, because the build system or deployment tools are gone, or because of untracked changes to the deployed system.

Virtual machines present an interesting challenge to the metaphor of "death." An approach of "treat systems like crops, not pets" is gaining traction for many reasons, including that it forces us to track system configuration in a way that will promulgate to the next crop. But what of a planned system retirement? Much like counting hardware failures, the operational reality that systems exist or not may be part of vital statistics and help us understand other things.

There are infrastructures, like botnets, which we choose to "takedown." This seems different from vital statistics, and closer to the idea of cleaning out standing water. However, infection rates and other details may contribute to vital statistics.

ALL MODELS ARE WRONG, SOME ARE USEFUL

The model of births and deaths clearly has limits. That doesn't mean that knowing how many computers are out there in more reliable ways would not be helpful. Being consistent in discussion has value: 1 in 135 has this problem, and we don't have to worry about the quality of that data.

Not everyone is born or dies in the same place. Many people migrate during their lifetimes, and as we think about public health, there is an interesting analogy. When a chip is made in China or Mexico, and then "migrates" to another country, there are import and export duties to be paid, and those may offer an interesting opportunity to use data already being collected to kickstart the creation of cybersecurity vital statistics.

Marriages and divorces are counted because they relate to both the likelihood of children and are predictive of people's health. There are not obvious equivalents for cyber public health. At the enterprise level, perhaps mergers, acquisitions or spin-offs are equivalent, and also may serve a purpose analogous to twin studies: what happens after the spin-out is that the choices made by the leadership of each are applied to the same baseline. Similarly, private equity imposing requirements may be similar to a family having similar health behaviors or beliefs. Software being "forked" may also be an opportunity to study divergence.[1]

4. A MODEST PROPOSAL FOR CYBERSECURITY VITAL STATISTICS

All of the measurements of vital statistics serve to inform us about the population. That is, the public in public health. The complexity that technology brings means we may need a few more numbers. Many of these are available, but not collected. Searchers may discover different sets. The sets may rely on different definitions of terms. For example, we might get statistics on hardware via PC shipments, or via sales from Intel and AMD, or do we include ARM or TSMC?

We propose the following categories as useful for measuring cybersecurity vital statistics. The main categories are listed as bullets, while possible sub-categories are listed in parentheses.

Births:

- Hardware creation (PC, phone, IoT, chips, other)
- Hardware sale (retail and bulk sales, new or used)
- Operating system installs
- Re-install of operating systems
- Accounts
- New domain registrations

[1] Such forks are sometimes the result of acrimonious discussion or "irreconcilable differences." Beyond the divorce metaphor, such acrimony may result in competitive analysis.

Deaths:

- Hardware fail (recovered, trash)
- Virtual machines reap – planned versus not planned

CHALLENGES

- Devices that cannot be repaired, or the software that cannot be reliably re-installed.
- Modern "server" deployment patterns including:
    - Virtual machine instances at big service providers are cached and cloned.
    - Lifecycle analysis for systems. What about a large system like Google Borg, or a sharded system which slowly adjusts the mix of operating systems in use in a gated or progressive rollout deployment model.
    - Virtual machines get spun up differently than desktop computers.

OPPORTUNITIES

Technological systems may have other "vital statistics" which could be gathered. For example, data from crashes is now commonly collected either by operating system creators like Microsoft and Apple, by cloud operators, and by system designers who focus on "observability". We should be open to opportunities created by the difference between technical and biological systems, and aware of the risk of focusing our search for our lost keys only where a lamp happens to be casting light.

Crashes may be an equivalent of emergency room visits. Those are important statistics, gathered broadly, and also there are various statistical programs which gather additional data at a small subset of emergency rooms, selected to be representative. More broadly in cybersecurity, there is a discipline of disaster recovery engineering which may have analogies, surveillance or tracking, or other possible valuable contributions or partnerships.

CONCLUSION

The absence of vital statistics in cybersecurity limits the questions that can be asked about populations. Such population information is needed to consider disease prevalence or incidence, to judge long term impacts of programs, or even to "check our own work," such as excess mortality assessments (Wallace-Wells 2022).

There are certainly complexities and limits to the metaphor, but much of the data may be gathered already. That data could be assembled and published with relatively low cost, and allow us to demonstrate its value and inform the question of how much more could be done.

REFERENCES

Doshi, Rushabh. "Global PC Shipments Pass 340 Million in 2021 and 2022 Is Set to Be Even Stronger." Canalys, January 12, 2022. https://www.canalys.com/newsroom/global-pc-market-Q4-2021.

Geer, Dan and Paul Vixie. Nameless Dread. Usenix ;login:, Vol. 43 No 4., Winter 2018. http://geer.tinho.net/fgm/fgm.geer.1812.pdf.

Verma, Abhishek, Luis Pedrosa, Madhukar R. Korupolu, David Oppenheimer, Eric Tune, and John Wilkes. Tech. Large-Scale Cluster Management at Google with Borg. Google Inc., 2015. https://static.googleusercontent.com/media/research.google.com/en//pubs/archive/43438.pdf.

Wallace-Wells, David. "What a Single Metric Tells Us about the Pandemic." Intelligencer. New York Magazine, March 26, 2022. https://nymag.com/intelligencer/2022/03/covid-excess-mortality.html.

Xu, Xiaowei, Hailong Qiu, Qianjun Jia, Yuhao Dong, Zeyang Yao, Wen Xie, Huiming Guo, et al. "AI-CHD: An AI-Based Framework for Cost-Effective Surgical Telementoring of Congenital Heart Disease." Communications of the ACM 64, no. 12, December 2021. https://cacm.acm.org/magazines/2021/12/256931-ai-chd/fulltext.

ACKNOWLEDGEMENTS

This report was written by Adam Shostack for the CyberGreen Institute.

We would like to acknowledge Dan Geer who provided several points including mergers and acquisitions, software build systems as a metaphor for birth and "end of life," and questions about domains. Shawn Hernan pointed out the existence of crash data and unlocked the opportunity to think further about what we could do uniquely in technical systems. Arastoo Taslim prompted more thinking about why divorces are included as vital statistics, and if there are analogs.