Security Now! #988 - 08-20-24

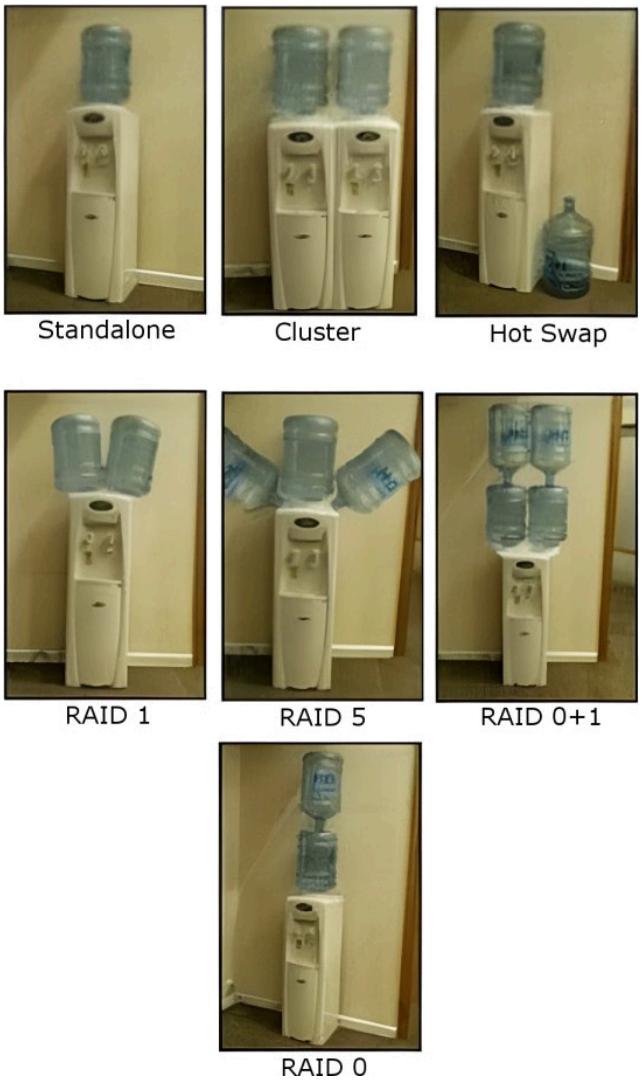
National Public Data

Happy 19th Birthday, Security Now!

This week on Security Now!

As we embark on our 20th year of this weekly Internet security and privacy oriented technical news podcast, we're going to look at some more interesting certificate revocation news and we have an experiment for our listeners. What six 0-days were patched during Microsoft's Patch Tuesday last week? 53 episodes of the 1980's "Famous Computer Café" radio show were recently discovered and are now online – hear Bill Gates before his voice changed. We have release #3 of IsBootSecure and a GRC email update and some interesting listener feedback. Then, to no one's surprise, we're going to take a deep dive into the background, meaning and impact of the largest personal data breach in history; how to look up your own breached records online, what to do and what this means for the future.

If water coolers were RAID arrays...



Security News

Revocation Update

A contributor over in GRC's newsgroups posted some terrific observations following from last week's discussion of revocation. Andrew wrote:

> *I have a sneaky suspicion that Steve's long-term plan with revoked.grc.com is gonna ferret out what software is swimming the waters in this space...naked, shall we say.*

Andrew was addressing my plan, which I had shared there, which was to back GRC's servers away from the use of OCSP to see what happens – and at 2am this morning the last-received and stapled OCSP status expired for the revoked.grc.com site. I'll share what happened after that in a minute. But first let's further examine one last feature of OCSP that we haven't talked about yet and work out OCSP's Achilles heel which made it impractical to require at scale:

As we all just witnessed last week, when GRC's revoked.grc.com site began stapling an OCSP status that loudly said its certificate had been revoked and no one's web browser would show the page, OCSP appears to be working better than anything else ever has so far. But as we also learned, the CA/Browser forum now plans to make OCSP support optional and to switch back to the use of Certificate Revocation Lists, making them mandatory. I replied to Andrew:

If the industry has inexplicably chosen to abandon the system that ALL BROWSERS are currently using with 100% success – as was just demonstrated with GRC's OCSP stapling working PERFECTLY everywhere – then, fine... we'll start testing the replacement system to see how well ***it*** *does.*

Andrew added:

> *With browser providers like Google, Mozilla, and Apple providing the CRL important bits as a centralized service with rapid-update to their browsers in the field...? It sounds like total insanity to me.*

To which I replied:

I 100% agree. If Let's Encrypt and the rest of the CA/Browser forum are suddenly so worried about web privacy due to individual browsers reaching out to query Certificate Authority OCSP services, then start a countdown on the mandatory support for "OCSP Stapling."

The final piece that I have not discussed this time around on the podcast, though we covered it in depth at the time, is one additional bit of perfection for the OCSP system. It's known as "OCSP Must Staple." Here's the problem **that** solves:

A web server obtains a web server TLS certificate signed by its certificate authority. And the whole point of signing is that not a single byte of data can be changed in that certificate or its signature will become invalid. This means that when that web server wishes to include that certificate authority's recent OCSP assurance about the certificate's validity (or lack thereof), the Certificate Authority's signed OCSP status can only be appended to the certificate. It cannot in

any way modify or be incorporated into the certificate. That's why the term "stapling" has been adopted, since stapling is such a good analogy.

The problem is, stapling an up-to-date OCSP status to a web server's certificate is optional. So if a bad guy gets hold of a valid web server certificate they'll gladly staple any OCSP good news to that certificate everytime they send it out to a web browser. But once the certificate has been revoked and its certificate authority's OCSP now contains bad news (as GRC's revoked site's certificate did last week), no bad guy would continue stapling that certificate revocation news to their fraudulently obtained certificate. Instead, they'll remove any stapling and hope that no one's browser checks the current validity of the certificate by querying the certificate authority's OCSP service or CRL directly, themselves. (And, so far as we know, browsers no longer make their own queries to OCSP on their own.)

To prevent that from happening, it's possible for a website that wants the best security for itself, and which always intends to staple – and can commit to that – to ask its certificate authority to include a flag in its certificate which is known as "OCSP Must Staple." Since that "Must Staple" flag is an integral part of the signed certificate itself, it is immutable and once issued can never be changed. And the presence of that flag in any certificate that's received by a browser is an assertion that this certificate must only be honored and trusted if it's accompanied by a current and still valid OCSP assertion to that effect. An OCSP statement **must** be stapled to the certificate or it must not be trusted. That solves the problem of the bag guys choosing not to staple any bad news to their stolen certificate. With OCSP Must Staple, they have no choice.

In my reply in GRC's newsgroup I wrote:

Let's Encrypt (and all other CA's following the CA/Browser forum) can mandate t*hat they will be setting the "OCSP Must Stable" requirement in every certificate they issue after some date certain. And that will force ALL web servers to support stapling and to staple. What's the problem with that?*

With enforceable stapling, after the stapling mandate takes effect, every certificate will have OCSP stapled to it within 397 days once all pre-mandate certificates will have expired.

And the big win for privacy is that with a fresh OCSP response stapled to every certificate, browsers can and will be inhibited from querying for OCSP status directly. They'll have no reason to ask further.

Thus we have fast revocation notification with ZERO privacy risk, since browsers will get their updates from the server's cert, zero performance overhead for the same reason – no need to ask anyone else or look any further – and reduced load on CA's OCSP services since only the servers they have issued certificates to will be querying them... and that querying interval can also be readily changed at the CA end simply by changing the OCSP response's lifetime.

So why isn't this what's being done? After not coming up with any answer that I liked I did some digging in the CA/Browser forum's documents. In the discussion surrounding that CA/Browser ballot measure SC-63 that we discussed last week, the adoption of which was nearly unanimous, I found this text:

"Independent of usage statistics, relying parties can't consistently depend on OCSP stapling for security unless responses are stapled on all connections. Further, even if the web server ecosystem had improved support for OCSP-stapling and we could require the use of the 'must-staple' extension, we'd remain dependent upon robust and highly-reliable OCSP services, which have been an ongoing ecosystem challenge."

And that gave me a clue. On the one hand, what they appear to be saying, all experience to the contrary, is that not that much has changed for OCSP during the past ten years since we first looked at it on this podcast. They're suggesting that OCSP Must Staple cannot be used because the robustness of OCSP services has never been sufficient.

Thinking about this, one thing the plain vanilla certificate system offers is no requirement for real- time communication with anyone other than the client and the server. The server has a signed certificate. The client locates the server by its domain name using hopefully secure DNS. And during the client's connection to the server, the server provides its certificate to prove its identity at that domain name. It really is an elegant system. And it is minimal.

Of course, the one place this beautifully minimal system falls down completely is when the certificate it's sending can no longer be trusted. There's no way for the browser to know that. In the absence of any other facility, that certificate will be trusted until it expires, which will be a maximum of 90 days for Let's Encrypt certs or a maximum of 397 days for traditional annual certificates. If our goal is for this otherwise simple and elegantly minimal system to deal with the need to revoke certificate trust before the certificate's natural end of life, we're going to need to add something else. And what we see is that the industry has been struggling to come up with a solution that works for everyone.

None of GRC's servers have had any problem with OCSP stapling. And I doubt that anyone's would. But it's true that the adoption of Must Staple would place the onus of having that system robustly online squarely on the shoulders of each respective certificate authority. Since web servers generally begin attempting to refresh their soon-to-expire OCSP status the day before it's due to expire, a protracted DDoS which forced an OCSP service offline would cause an OCSP outage with widespread devastating consequences. All of those sites that had "Must Staple" in their certs would be unable to obtain an update from their CA and would effectively go offline because no web browser would trust their expired stapling. Every affected website would look like GRC's deliberately revoked site looked like after our podcast two weeks ago.

Browsers have given up their own checking of OCSP for performance reasons and they've been absolved of any guilt by stating their concern over the privacy of their users. So they rely upon stapling to do the work for them when stapling is present and they do nothing when it's not present. But that means that any certificate that does not use "Must Staple" will always be vulnerable to long term abuse if it's stolen, since no illegitimate server would include a negative OCSP response – this causing all current web browsers to default to trusting the malicious web site. And the industry cannot improve certificate revocation and user privacy by moving to enforce "Must Staple" because a DDoS of the apparently still not very robust OCSP service, which would have then become a requirement for **all** of a certificate authority's customers, would result in widespread web server mistrust and refusal to load pages.

So I believe we've worked out why OCSP is a problem: Unless its stapling is mandatory it can be bypassed simply by not stapling. And the danger of making stapling mandatory is that an OCSP outage, for any reason, malicious or accidental, lasting more than a day, would have devastating consequences as all of that certificate authority's certificates would become untrusted.

Certificate Revocation Lists, imperfect as they may be, are less "online" than OCSP which, after all, stands for **Online** Certificate Status Protocol." The bottom line is that in the reality of today's Internet, "online" is not something that can be made to work. ***Its strength is also its failure.***

GRC's next experiment:

As I mentioned above, last week I followed through with something I had mentioned in passing during last week's podcast: I deliberately blocked all GRC access to DigiCert's OCSP service. Before I did that, I double checked that my certs do **not** have "OCSP Must Staple" enabled.

The last-received OCSP status from DigiCert for the revoked.grc.com site was set to expire around 2am **this** morning. And sure enough, when I checked this morning, GRC's revoked.grc.com server was no longer stapling that expired OCSP status to its thoroughly revoked TLS certificate. And what do you think happened? Yep... every web browser other than Firefox has resumed showing the revoked.grc.com site. Chrome loves it, Safari loves it, Edge proudly shows its page. Everyone's happy with the site despite the fact that its certificate was revoked 21 days ago. Every web browser except Firefox is once again completely happy with the site.

The reason, we have now proven, is that the revoked.grc.com server is no longer stapling a **negative** OCSP status that says "oh by the way, this site's certificate, the one I've been stapled to, has been revoked by its issuer." And in the absence of either a positive or a negative OCSP status, all browsers other than Firefox trust the revoked but otherwise valid certificate. In other words, 21 days after that certificate's revocation, no web browsers other than Firefox demonstrate that they actually have functioning Certificate Revocation List technology today. Firefox stands alone in correctly refusing to show the pages being served under the guise of this well-revoked TLS certificate.

I found some interesting information about what Firefox is doing that I'll share next week. In the meantime, given the amount of feedback I've received on this topic, I imagine that many of our listeners will again be going over to "revoked.grc.com" to see what their browser thinks.

Patch Tuesday

Last Tuesday was August's Patch Tuesday for Microsoft and many for many other publishers who also appear unable to ever get the important bugs out of their code. In this month's installment of trying some more, Microsoft released updates to fix at least 90 security vulnerabilities in Windows and their other software, which included a startling six 0-day flaws that were (or are) already being actively exploited by attackers.

Security flaws were found and fixed in Office, .NET, Visual Studio, Azure, Co-Pilot, Microsoft Dynamics, Teams, Secure Boot, and of course Windows.

Among the six 0-days fixed this month, half are local privilege escalation vulnerabilities which, while severe because they are useful only once an attacker is inside a machine – meaning that at least they were not remote code execution vulnerabilities. Those three allow an attacker to obtain SYSTEM level privileges on a vulnerable machine. We know little more about these, though you can bet that would-be attackers are hard at work reverse engineering these for inclusion in their ever-growing collections of ways to subvert Microsoft's attempts at security.

There is, however, a remote code execution vulnerability when Microsoft's Edge browser is forced to operate in Internet Explorer Mode. Although IE mode is not enabled by default in Edge, the fact that this is or was being actively exploited suggests that there are occasions where an attacker can either arrange to get this enabled or has identified an organization or user who has this enabled for some backward (very backward) compatibility need.

Another 0-day is a bypass of the "Mark of the Web," security feature which causes Windows to be far more mistrustful of files obtained from the Internet. As we've seen in the past, a MOTW bypass is always used a part of a larger exploit chain, but its bypass enables something to be done that was supposed to be impossible.

This month's third and final 0-day is a remote code execution flaw in Microsoft Project. However, Microsoft and multiple security firms have pointed out that this vulnerability is only useful against users who had previously disabled notifications about the security risks of running VBA Macros in Microsoft Project.

So those were the six 0-day flaws out of the kettle of 90 total, that were under active use attacking the users of Microsoft's software products.

"The Famous Computer Café"

Thanks to our listener, Larry Deniston who brought this to my attention, I have some news that might be of interest to both our old timer listeners and and our younger audience who may have heard the names of these people who have grown somewhat legendary within this PC industry that we all love. Audio tapes of a mid-80's technology radio show known as "The Famous Computer Café" were found earlier this year, restored and digitized. Yesterday, the Internet Archive posted:

> *A previously lost cache of celebrity and historical interviews from a long-dormant radio show have been discovered, digitized, and made available for all.*
>
> *The Internet Archive is now home to 53 episodes of **The Famous Computer Cafe**, a 1980s radio show about the new world of home computers. The program included computer industry news, product reviews, and interviews, and aired from 1983 through 1986 on radio stations in southern and central California.*
>
> *The creators of The Famous Computer Cafe saved every episode on reel-to-reel tapes, but over the years the tapes were forgotten, and, ultimately, lost. Earlier this year archivist Kay Savetz recovered several of the tapes in a property sale, and recognizing their value and worthiness of professional transfer, launched a GoFundMe to have them digitized, and made them available at Internet Archive with the permission of the show's creators.*

While full of time-capsule descriptions of 1980s technology news, the most exciting aspect of the show has been the variety and uniqueness of the interviews. The list of people that the show interviewed is a who's-who of tech luminaries of the 1980s: computer people, musicians, publishers, philosophers, journalists. Interviews in the recovered recordings include Timothy Leary, Douglas Adams, Bill Gates, Atari's Jack Tramiel, Apple's Bill Atkinson, and dozens of others. The recovered shows span November 17 1984 through July 12, 1985.

For ease of access, I've made this GRC's shortcut of the week for episode #988. So grc.sc/988 will jump your browser to the Internet Archive's blog posting which includes a link to a Google Docs spreadsheet listing all 53 recordings, who's on them and links to their Archive page. https://blog.archive.org/2024/08/19/archiving-the-famous-computer-cafe/

IsBootSecure

A quick follow-up on GRC's newest "IsBootSecure" freeware: Last Wednesday, Release #2 added keyboard accelerators and tooltips to the buttons and fixed a noncritical misreporting edge case. Release #3 later Wednesday added a "silent" option to cause IsBootSecure to fully suppress its user-interface. So it runs on any Windows machine silently. It examines the machine and immediately exits with an exitcode that batch or powershell scripting can capture and check. This allows the app to be deployed by scripts organization wide to check and inventory an entire organization's inventory of PCs for their boot-time status.

With that finished I got back to work on SpinRite v6.1's documentation. I'm now down to the video walk-throughs that I'm looking forward to creating so that non-owners can see what SpinRite looks like to help them decide whether it might be useful to them.

GRC Email

And I'm so very pleased to report that GRC's email system is working very well. I'm completely happy with this facility and I know – because everyone says so – that many Security Now! listeners appreciate not only being able to directly send things to me that they think will be of interest, but they also very much appreciate receiving a weekly summary of each week's podcast a few hours before we record it.

I also wanted to mention that I finally started bouncing any unregistered email that's addressed to the securitynow@grc.com account. Until now, rather than bouncing, I had been redirecting and monitoring that email while this system was new, and sending notes back to the few people here and there who were, for example, sending email from a different account than the one they had registered. And I've received a few complaints from people who look around for the email address to send to and cannot find it anywhere. I understand the annoyance, but since I want to keep this more or less between us, I'm not advertising the securitynow@grc.com account anywhere on GRC's site. It's just for insider's.

Closing the Loop

Michael French

Hi Steve, I'm stumped on how to get information through wifi firewalls that block everything except HTTP (TCP port 80) and HTTPS (TCP port 443). I run an OpenVPN server at home in Alabama on TCP port 443 but some wifi firewalls outside of my home network still block my communicating with it after only a few seconds of operation -- presumably from their implementing deep packet inspection. I just returned from Europe and found that all the countries that I visited (UK, Netherlands, Belgium, and France) have rigged their wifi firewalls to block everything except HTTP and HTTPS. I would sure appreciate your suggestions on how to get communications through these over restrictive wifi firewalls. Thanks, Mike

This was an interesting puzzle. It cannot be deep packet inspection unless Michael had been installing and trusting a certificate from an access point, which I'm sure cannot be the case. Any TLS connection being made to port 443 will be 100% opaque to anyone monitoring packets. They'll see what looks like any normal connection to a remote web server but only from the outside. They will have no way of knowing what's going on inside.

My best guess about what might be going on comes from Michael's comment that he's getting disconnected after only a few seconds of operation. Although persistent HTTP connections between a web browser and web server can be made and sustained, even then, they are usually dropped after all pending queries and replies have been exchanged. It's unusual for a passive and unused connection to be maintained. So it might be that these WiFi access points are watching the flow of packet traffic and are deliberately dropping any connections that go idle. Doing that would not interfere with normal web browser use while it would be an effective way of blocking other "web-like" traffic such as an HTTPS VPN.

Doug White

Aloha Steve! Listened to the Tuesday podcast and thought I'd mention something when it came to transferring DNS. I switched over to Hover from GoDaddy (something I've put off for years) and wanted to mention that the DNS server entries after the transfer to Hover still pointed at the GoDaddy DNS servers. I had to look find the Hover DNS server names and replace the GoDaddy entries in the Hover settings. I'm guessing it's so that everything still works after the switchover but I wasn't alerted to the fact (that I'm aware of) that I needed to make that change. Cheers!

I'm sure that Doug's correct that in switching his domain registrar from GoDaddy to Hover, Hover would have examined his previous registrar's domain nameserver entries and would have deliberately left them unchanged under his relocated registration. Eventually, presumably, GoDaddy could be expected to suspend their support for his DNS once his registration had been relocated, so this would have probably eventually come to light. So after moving to a new domain registrar you'll want to be sure that it's registered nameserver records are pointing where you intend.

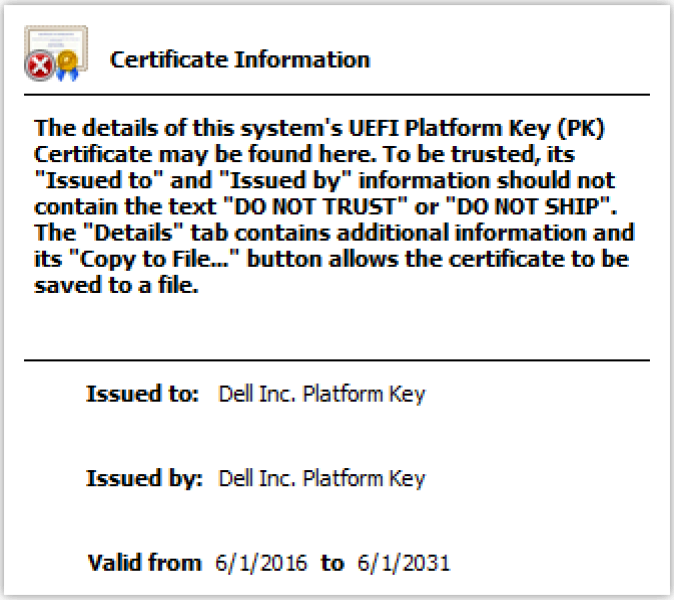
Scott

Is there an advantage of a 7-day stapled OCSP attestation over a TLS certificate with a 7-day expiration? With certificate automation there's no reason an expiry needs to be 30 days or a year or a week. If 7 days is enough time to catch a revocation, just expire the cert that quickly. Revocation only really seems like it makes sense if it's instant.

What I believe we've clearly seen now is just how much all of this revocation and certificate life business involves a trade off. Let's Encrypt has been automated from day one, but they chose 90 days for their certificates when they could have chosen 7 days. Why? One advantage to 90 days, especially when they were starting out, was that it would significantly reduce the load on their certificate issuance and delivery infrastructure by a factor of nearly 13. But it's true that so long as a network outage or attack would **not** hold Let's Encrypt off the air while their certificates were expiring and were unable to renew, then their certificate recycle time could be as short as they like. But I know that if my servers were using Let's Encrypt certs, instead of DigiCert's 397-day certificates, I would want as much life per certificate as I could get just for safety margin. Because one thing that HAS happened to our industry is that you're no longer very effective on the web if you don't have a valid TLS certificate.

Brandon Foust

Will it boot on 6/2/2031?



Brandon sent me a picture of his Dell machine's UEFI platform key which shows that it's valid from 6/1/2016 through 6/1/2031. He wonders what happens on June 1st seven years from now. And the answer is that in this instance the date doesn't matter. It would be up to someone deciding whether or not to trust the root certificate based upon its date, and it's the UEFI firmware that's using this platform key certificate to check the signatures of the other signatures. So distrusting the platform key root certificate would require a deliberate act within the UEFI firmware, which is not something that it cares about or would do.

National Public Data

Before we wrap-up today's podcast, I'm going to provide everyone listening with a URL for a searchable online database containing the records from this breach. You enter your first and last full legal name, your state of residence and the year of your birth... and you'll be immediately presented with a listing of all of the people who share your name, state and date of birth. You'll find that the list is sorted by middle name or initial, and you will almost certainly find yourself listed there, often redundantly many times at different physical addresses all which will be familiar to you, because they will be correct, often with your telephone number and always with the correct last two digits of your otherwise redacted full social security number.

This is as real as it gets. And for that reason I'm going to start out this week with the main takeaway from what is being described, probably accurately by those who would know, as the largest data breach in history. That takeaway is: We have spoken previously, a number of times, about the need to freeze credit reporting at the three primary credit reporting agencies. If that previous coverage was still insufficient to motivate you, your friends or your family to take the steps to do so, then... if this one doesn't do the trick I'm pretty certain that nothing ever will.

Naturally this has been a headline grabber. The Verge's coverage was headlined: "The weirdest '3 billion people' data breach ever." BleepingComputer covered this under their headline: "Hackers leak 2.7 billion data records with Social Security numbers" and Brian Krebs piece was titled: "NationalPublicData.com Hack Exposes a Nation's Data." And if it wasn't too long to be the title of today's podcast, I so much would have loved to use the start of National Public Data's own admission which read: "There appears to have been a security incident." Ya think???

From across all of the coverage, the person who is probably better suited than anyone else to pull all the pieces together, while providing some perspective from his years of involvement with exactly these sorts of incidents is "Have I been Pwned?"s Troy Hunt.

And, indeed, Troy's coverage is the best I've seen anywhere. It was titled: "Inside the 3 Billion People National Public Data Breach. Troy wrote:

> *I decided to write this post because there's no concise way to explain the nuances of what's being described as one of the largest data breaches ever. Usually, it's easy to articulate a data breach: A service people provide their information to have someone snag it through an act of unauthorized access and publish a discrete corpus of information that can be attributed back to that source.*
>
> *But in the case of National Public Data, we're talking about a data aggregator most people had never heard of, where a "threat actor" has published various partial sets of data with no clear way to attribute it back to the source. And national Public Data is already the subject of a class action lawsuit, to add yet another variable into the mix.*

I'll interrupt to note that Bloomberg Law reported that the case is *"Hofmann v. Jerico Pictures, Inc."* filed in the Southern District of Florida. In part of their reporting, Bloomberg wrote:

Jerico Pictures Inc., a background-check company doing business as National Public Data, exposed the personal information of nearly 3 billion individuals in an April data breach, a proposed class action says.

On April 8, a cybercriminal group by the name of USDoD posted a database entitled "National Public Data" on a dark web forum, claiming to have the personal data of 2.9 billion people, according to the complaint filed Thursday in the US District Court for the Southern District of Florida, which said the group put the database up for sale for $3.5 million. If confirmed, the breach could be among the biggest ever, in terms of the number of individuals affected. It's unclear exactly when or how the breach occurred, according to the complaint, and the provider still hasn't provided notice or warning to affected individuals as of the filing.

The complaint said, to conduct its business, National Public Data scrapes the personally identifying information of billions of individuals from non-public sources—meaning plaintiffs didn't knowingly provide their data to the company. Some of the information exposed includes Social Security numbers, current and past addresses spanning decades, full names, information about relatives—including some deceased for nearly two decades—and more, according to the complaint. National Public Data didn't immediately respond to a request for comment.

That is, other than to post that "There appears to have been a security incident." Bloomberg continues:

Named plaintiff Christopher Hofmann, a California resident, said he received a notification from his identity-theft protection service provider on July 24, notifying him that his data was exposed in a breach and leaked on the dark web. He accused National Public Data of negligence, unjust enrichment, breaches of fiduciary duty and third-party beneficiary contract.

Hofmann asked the court to require National Public Data to purge the personal information of all the individuals affected and to encrypt all data collected going forward. In addition to monetary relief, he also asked for a series of requirements, including that National Public Data segment data, conduct database scanning, implement a threat-management program, and appoint a third-party assessor to conduct an evaluation of its cybersecurity frameworks annually for 10 years.

So that's the first of likely many similar expressions of grievance aimed at National Public Data. Let's see what more Troy Hunt has for us. Troy continues:

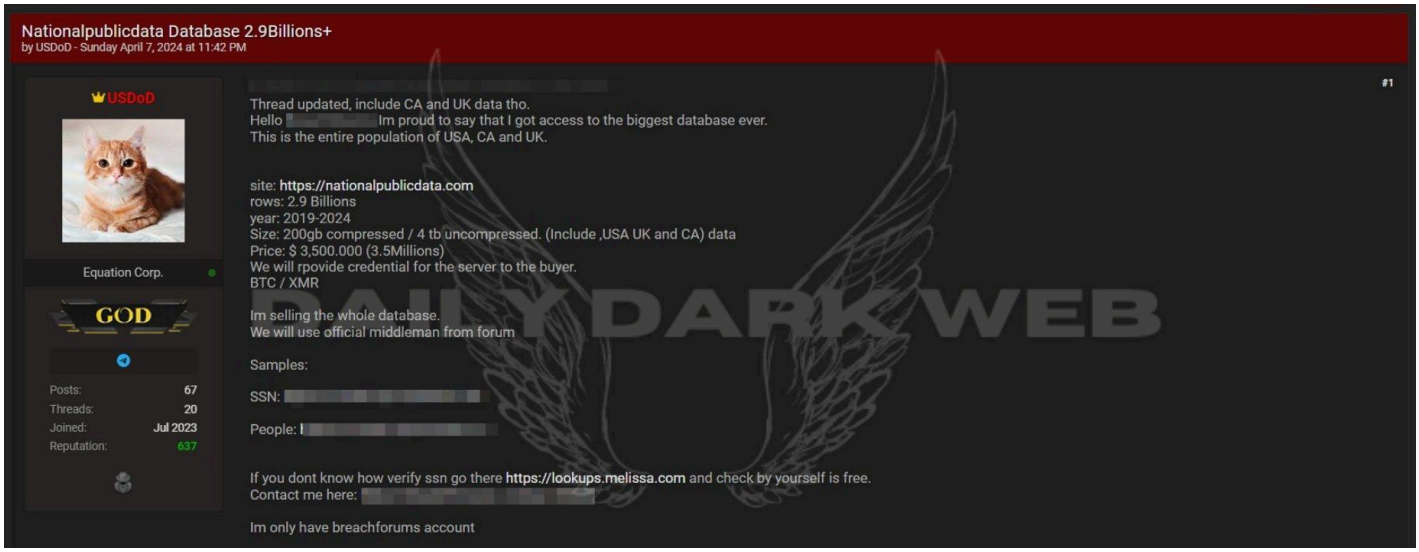
I've been collating information related to this incident over the last couple of months, so let me talk about what's known about the incident, what data is circulating and what remains a bit of a mystery.

Let's start with the easy bit - who is National Public Data (NPD)? They're what we refer to as a "data aggregator", that is they provide services based on the large volumes of personal information they hold. The front page of their website says: Criminal Records, Background Checks and more. Our services are currently used by investigators, background check

websites, data resellers, mobile apps, applications and more.

There are many legally operating data aggregators out there... and there are many that end up with their data in Have I Been Pwned (HIBP). For example, Master Deeds, Exactis and Adapt, to name but a few. In April, we started seeing news of National Public Data and billions of breached records, with one of the first references coming from the Dark Web Intelligence account @DailyDarkWeb:

Troy quotes a tweet from April 8th of this year: "USDoD Allegedly Breached National Public Data Database, Selling 2.9 Billion Records."



Back then, the breach was attributed to "USDoD", a name to remember as you'll see that throughout this post. And this the first reference to the 2.9 billion number we've subsequently seen flashed all over the press, and it's right there alongside the request of $3.5M for the data. Clearly, there is a financial motive involved here, so keep that in mind as we dig further into the story. That image also refers to 200GB of compressed data that expands out to 4TB when uncompressed, but that's not what initially caught my eye. Instead, something quite obvious in the embedded image doesn't add up: if this data is "the entire population of USA, CA and UK" (which is ~450M people in total), what's the 2.9 billion number we keep seeing? Because that doesn't reconcile with reports about "nearly 3 billion people" with social security numbers exposed. Further, SSNs are a rather American construct with Canada having SINs (Social Insurance Number) and the UK having, well, NI (National Insurance) numbers are probably the closest equivalent. This is the constant theme you'll read about in this post, stuff just being a bit... off. But hyperbole is often a theme with incidents like this, so let's take the headlines with a grain of salt and see what the data tells us.

I was first sent data allegedly sourced from NPD in early June. The corpus I received reconciled with what vx-underground reported on around the same time (note their reference to the 8th of April, which also lines up with the previous tweet):

On June 1st, vx-underground (@vxunderground) tweeted: "April 8th, 2024, a Threat Actor operating under the moniker "USDoD" placed a large database up for sale on Breached titled: "National Public Data". They claimed it contained 2,900,000,000 records on United States citizens. They put the data up for sale for $3,500,000."

In their message, they refer to having received data totalling 277.1GB uncompressed, which aligns with the sum total of the 2 files I received. They also mentioned the data contains first and last names, addresses and SSNs, all of which appear in the first file among other fields.

These first rows also line up precisely with the post Dark Web Intelligence included in the earlier tweet. And in case you looking at the data and think "that's the same SSN repeated across multiple rows with different names", those records are all the same people, just with the names represented in different orders and with different addresses (all in the same city).

In other words, multiple rows, in case 6 rows all represent one person, which got me thinking about the ratio of rows to distinct numbers. Being curious, I took 100 million samples and found that only 31% of the rows had unique SSNs, so extrapolating that out, 2.9 billion would be more like 899 million. This is something to always be conscious of when you read headline numbers: "2.9 billion" doesn't necessarily mean 2.9 billion people, it often means rows of data. Speaking of which, those 2 files contain 1,698,302,004 and 997,379,506 rows respectively for a combined total of 2.696B. Is this where the headline number comes from? Perhaps, it's close, and it's also precisely the same as Bleeping Computer reported a few days ago.

At this point in the story, there's no question that there is legitimate data in there. From the aforementioned Bleeping Computer story: "Numerous people have confirmed to us that it included them and their family members' legitimate information, including those who are deceased. And in vx-underground's tweet, they mention that: "It also allowed us to find their parents, and nearest siblings. We were able to identify someone's parents, deceased relatives, Uncles, Aunts, and Cousins. Additionally, we can confirm this database also contains information on individuals who are deceased. Some individuals had been deceased for nearly 2 decades.

And then in his posting, Troy Hunt wrote:

"A quick tangential observation in the same tweet: "The database DOES NOT [Troy put in all caps] contain information from individuals who use data opt-out services. Every person who used some sort of data opt-out service was not present."

Now this is such an obvious lay up for one of this network's sponsors that now would be the time to pause for that... **(DeleteMe)**

The point Troy was making about all data being absent from those using data opt-out services was interesting, he followed-up by writing:

This is what you'd expect from a legally operating data aggregator service. It's a minor point, but it does support the claim that the data came from NPD. None of the data discussed so far contains email addresses. That doesn't necessarily make it any less impactful for those involved, but it's an important point I'll come back to later as it relates to Have I Been Pwned.

So, this data appeared in limited circulation as early as 3 months ago. It contains a huge amount of personal information (even if it isn't actually 2.9 billion different people), and then to make matters worse... it was posted publicly last week:

On August 6th, Wolf Technology Group (@WolfTech) tweeted:

"National Public Data, a service by Jerico Pictures Inc., suffered a massive breach. Hacker "Fenice" leaked 2.9 billion records with personal details, including full names, addresses, and SSNs in plain text. The breach poses significant risks for identity theft and financial fraud. Jerico Pictures Inc. faces potential lawsuits and legal challenges due to the incident."

Who knows who "Fenice" is and what role they play, but clearly multiple parties had access to this data well in advance of last week. I've reviewed what they posted, and it aligns with what I was sent 2 months ago, which is bad. But on the flip side, at least it has allowed services designed to protect data breach victims to get notices out to them.

Inevitably, breaches of this nature result in legal action, which, as I mentioned in the opening, began immediately a couple of weeks ago. It looks like a tip-off from a data protection service was enough for someone to bring a case against NPD.

Up to this point, pretty much everything lines up, but for one thing: Where is the 4TB of data? And this is where it gets messy as we're now into the territory of "partial" data. For example, an 80 gigabyte corpus was recently posted to a popular hacking forum. While it's not clear whether that's the size of the compressed or extracted archive, either way, it's still a long way short of the full alleged 4TB.

Earlier this month, a 27-part corpus of data alleged to have come from NPD was posted to Telegram. The compressed archive files totalled 104GB and contained what feels like a fairly random collection of data. Many of these files are archives themselves, with many of those containing yet more archives. I went through and recursively extracted everything which resulted in a total corpus of 642GB of uncompressed data across more than 1k files. If this is "partial", what was the story with the 80GB "partial" from last month? Who knows, but in those files were 134M unique email addresses.

Just to take stock of where we are, we've got the first set of SSN data which is legitimate and contains no email addresses yet is allegedly only a small part of the total NPD corpus. Then we've got this second set of data which is larger and has tens of millions of email addresses yet is pretty random in appearance. The burning question I was trying to answer is "is it legit?"

The problem with verifying breaches sourced from data aggregators is that nobody willingly - knowingly - provides their data to them, so I can't do my usual trick of just asking impacted Have-I-Been-Pwned subscribers if they'd used NPD before. Usually, I also can't just look at a data aggregator breach and find pointers that tie it back to the company in question due to references in the data mentioning their service. In part, that's because this data is just so damn generic. We have first and last name, address, SSN. Attributing a source when there's only generic data to go by is extremely difficult.

The kludge of different file types and naming conventions worried me. Is this actually all from NPD? Usually, you'd see some sort of continuity, for example, a heap of .json files with similar names or a swathe of .sql files with each one representing a dumped table. The presence of a uniquely named CSV file ties this corpus together with the one from the earlier tweet, but then there's stuff like "Accuitty_10_1_2022.zip"; could that refer to Acuity (single "c", single "t") which I wrote about in November? HIBP isn't returning hits for email addresses in that folder against the Acuity I loaded last year, so no, it's a different corpus. But that archive alone ended up having over 250GB of data with almost 100M unique email addresses, so it forms a substantial part of the overall corpus of data.

The 3,608,086 KB "criminal_export.csv.zip" file caught my eye, in part because criminal record checks are a key component NPD's services, but also because it was only a few months ago we saw another breach containing 70M rows from a US criminal database. And see who that breach was attributed to? USDoD, the same party whose name is all over the NPD breach. I did actually receive that data but filed it away and didn't load it into HIBP as there were no email addresses in it. I wonder if the data from that story lines up with this file? Let's check the archives:

Different file name, but hey, it's a 3,608,086KB file so exactly the same size! Given the NPD breach initially occurred in April and the criminal data hit the news in May, it's entirely possible the latter was obtained from the former, but I couldn't find any mention of this correlation anywhere. (Side note: this is a perfect example of why I retain breaches in offline storage after processing because they're so often helpful when assessing the origin and legitimacy of new breaches).

Continuing the search for oddities, I decided to see if I myself was in there. On many occasions now, I've loaded a breach, started the notification process running, walked away from the PC then received an email from myself about being in the breach I'm continually surprised by the places I find myself, including this one!

Yep, it's an email address of mine, yet oddly, none of the other data is mine. Not my name, not my address, and the numbers shown definitely aren't familiar to me. I suspect one of those numbers is a serialized date of birth, but of the total 28 rows with my email address on them, the two unique DoBs put "me" as being born in either 1936 or 1967. Both are a long way from the truth.

A cursory review of the other data in this corpus revealed a wide array of different personal attributes. One file contained information such as height, weight, eye color, and ethnicity. The "uk.txt" file merely contained a business directory with public information. I could have dug deeper, but by now, there was no point. There's clearly some degree of invalid data in here, there's definitely data we've seen appear separately as a discrete breach, and there are several different versions of "partial" NPD data (although the 27-part archive discussed here is the largest I saw and the one I was most consistently directed to by other people). The more I searched, the more bits and pieces attributed back to NPD I found. If I were to take a guess, there are two likely explanations for what we're seeing: This incident got a lot of press due to the legitimacy of the initial dump of SSNs, and the subsequent partial dumps are riding on the coattails of breach hysteria. It appears that NPD may have siphoned up a heap of publicly circulating data to enrich their offering, and it got snagged along with the initially released SSN data.

These conclusions are purely speculative, though, and the only parties that know the truth are the anonymous threat actors passing the data around and the data aggregator that's now being sued in a class action, so yeah, we're not going to see any reliable clarification any time soon.

Okay. So Troy's focus is understandably on his Have I Been Pwned web service and the email addresses his site uses to allow users to lookup their own records. And I thought that Troy's obsessive pursuit of, and verification of, the exact source of this breached data was interesting.

As we launched into this topic I said:

"Before we wrap-up today's podcast, I'm going to provide everyone listening with a URL for a searchable online database containing the records from this breach."

To aid everyone's memory, I made a shortcut for it, which begins with our standard GRC.SC for GRC's shortcut redirector, then slash and "NPD" – for National Public Data: https://grc.sc/npd Using that shortcut will bounce your web browser over to: https://npd.pentester.com

Pentester.com is a 100% legitimate penetration testing subscription service that can be used by websites to checked their site's security. Pentester describes itself as: *"We are a cybersecurity technology platform that has sourced the tools, methods, and techniques attackers use. Our system allows owners and operators to find potential risks and exposures before attackers do."*

So in the wake of this massive National Public Data breach, these guys jumped on the opportunity to drive a great deal of traffic to their site and to make people aware of their services, simply by loading these 3 billion records into the cloud – which are after all, all now in the public domain – and allowing anyone to easily search these 3 billion records by entering their full legal first and last name, the state of their residence and their year of birth. I say "full legal first and last name" because these records are from various legal documents of the sort that will include one's social security number. I did that and the results took my breath away. The search for me revealed multiple records for every address that had been associated with me through the years, in every case with my correctly associated social security number.

It's important for everyone to understand that what's shown here is just meant to be a proof of presence of everyone's highly personal data within this massive database. The fact that this search only displays the last two digits of everyone's social security number does not mean that that's all there is. No. Nor that these search fields are all there is. Again, No. Everything about us, our full physical addresses, our various past and present telephone numbers, our complete social security number, and a massive amount of other highly personal data is all there for the taking.

It's unclear what the exact geographic spread of the population of this database is. From the reporting it appears to at least be the US, Canada and perhaps the UK. And since this podcast has a global listening audience we may have many listeners who will not find themselves listed here. But given what I've seen from poking around at people whose year of birth I know, I would be surprised if all of our domestic US and Canadian listeners were not quite chagrined by their lookup of their own data and that of others whom they know and care about.

It is my sincere hope that the ridiculously massive scope and scale of this breach, and of the press attention it has deservedly received, will shine a very bright light upon this otherwise dark personal data aggregating corner of the Internet and that we might see some changes made in the laws that have allowed this practice to evolve and thrive.

The data that has escaped already, will now be forever public. There's nothing that can be done about that now. And it's clearly wrong that it's up to individuals to pay to out-out of this personal data collection process. But Troy's finding that the data of those who had previously done so was conspicuously absent from this breach, serves as a true testament to the effectiveness of today's opt-out services.

It actually works. So if I were a young person today, for whom it is not too late to prevent one's entire personal life and history from being aggregated, sold, and eventually leaked out onto the Internet, I would seriously consider adopting paid-for measures – today – to require all data aggregators to remove any records they already have, and then bide my time until the slow-moving legal and governmental system catches up with what's going on. This event clearly demonstrates that the way things are now needs to be changed.

As for the rest of us whose last several decades of personal data is now so obviously out on the Internet flapping in the breeze, all we can really do is tightly lock down all access to our credit histories so that no one can apply for credit in our names.

We last talked about this need on April 15th and at that time I created the GRC shortcut of "credit" – so https://grc.sc/credit – which will bounce your web browser over to the investopedia page which talks about freezing credit and provides updated links to each of the three main credit reporting bureaus.

It may also be that the policies of the credit bureaus will need to change so that rather than, by default, passively allowing anyone to access our credit histories it will be incumbent upon them to first obtain our clear real-time permission to allow anyone to have any access. But regardless, we know that any change will be slow and incremental and will take time to even get started.

So as this podcast enters its 20th year it appears that one of our favorite phrases "what could possibly go wrong" will continue to keep us on our toes for quite some time to come.