Statement of Troy Hunt

For the House Committee on Energy and Commerce

"Identity Verification in a Post-Breach World"

30 November 2017

Summary

1. Data breaches occur via a variety of different "vectors" including malicious activity by attackers exploiting vulnerabilities, misconfiguration and behalf of system owners and software products intentionally exposing data by design.
2. There is frequently a long lead-time (sometimes many years) between a data breach and the service owner (and those in the breach) learning of the incident. We have no idea of how many incidents have already occurred but are yet to come to light.
3. The industry has created a "perfect storm" for data exposure. The rapid emergence of cheap, easily accessible cloud services has accelerated the growth of other online services collecting data. Further to that, the rapidly emerging "Internet of Things" is enabling us to digitise all new classes of information thus exposing them to the risk of a data breach.
4. An attitude of "data maximisation" is causing services to request extensive personal information well beyond the scope of what is needed to provide that service. That data is usually then retained for perpetuity thus adding to an individual's overall risk.
5. Lack of accountability means that even in the wake of serious breaches, very little changes in the industry and we continually see other organisations repeat the same mistakes as their peers.
6. Data breaches are redistributed *extensively*. There's an active trading scene exchanging data both for monetary gain and simply as a hobby; people collect (and thus replicate) breaches.
7. Many of the personal data attributes exposed in breaches cannot be changed once in the public domain, nor can these breaches be "scrubbed" from the internet once circulating.
8. Even without data breaches, we're willingly exposing a huge amount of personal information publicly via platforms such as social media.
9. The prevalence with which our personal data is exposed has a fundamental impact on the viability of knowledge based authentication. Knowledge which was once personal and could be relied upon to verify an individual's identity, is now frequently public knowledge.

Opening

Vice Chairman Griffith, Ranking Member DeGette, and distinguished Members of the House Energy and Commerce Committee, thank you for the opportunity to testify.

My name is Troy Hunt. I'm an independent Australian Information Security Author and Instructor for Pluralsight, an online learning platform for technology and cybersecurity professionals. I'm commissioned on a course-by-course basis to create training material that has been viewed by hundreds of thousands of students over the last 5 years. I'm also a Microsoft Regional Director (RD) and Most Valuable Professional (MVP), both titles of recognition rather than permanent roles. I've been building software for the web since 1995 and specialising in online security since 2010.

Of particular relevance to this testimony is my experience running the data breach notification service known as Have I Been Pwned (HIBP). As a security researcher, in my analysis of data breaches I found that few people were aware of their total exposure via these incidents. More specifically, I found that many people were unaware of their exposure across *multiple* incidents (one person appearing in more than 1 data breach) and indeed many people were unaware of *any* exposure whatsoever. In December of 2013, I launched HIBP as a freely accessible service to help people understand their exposure. Over the last 4 years, the volume of data in the service has grown to cover more than 250 separate incidents and over 4.8 billion records. What follows are insights drawn largely from running this service including the interactions I've had with companies that have been breached, those who have had their personal data exposed (myself included) and law enforcement in various jurisdictions around the world.

Data Breach Vectors

Data breaches have become a fact of modern digital life. Our desire to convert every aspect of our beings into electronic records has delivered both wonderful societal advances and unprecedented privacy risks. It's an unfortunate yet unavoidable reality that the two are inextricably linked and what follows describes the risks we are now facing as a result.

The term “data breach” is used broadly to refer to many different discrete vectors by which data is exposed to unauthorised parties. Some are as a result of malicious intent, some occur due to unintentional errors and yet others are inadvertent by-products of software design; they’re “features”, if you will.

Malicious incidents are the events we immediately associate with the term “data breach”. In this case, a “threat actor” has deliberately set out to gain unauthorised access to a protected system, often with the intention of causing harm to the organisation and their subscribers. We frequently see successful attacks mounted through exploitation of very well-known vulnerabilities with equally well-known defences. They exploit flaws in our software design, our security measures and indeed our human processes. They may be as sophisticated as leveraging previously unknown flaws or “zero days”, yet they’re frequently as simple as exploiting basic human shortcomings such as our propensity to choose poor passwords (and then to regularly reuse them across multiple services).

Especially in recent years with the growing ubiquity of easily accessible cloud services, data breaches often take the form of unintentionally exposed data. The ease today with which a publicly facing service can be provisioned and large volumes of data published to it is unprecedented – it can take mere minutes. Equally unprecedented is the simplicity with which an otherwise secure environment can be exposed to the masses; a single firewall setting or a simple access control change performed in mere seconds is all it takes.

The very design of some online services predisposes them to revealing large volumes of data about their subscriber base. Particularly in systems intended to make people discoverable such as social media or dating sites, we’ve seen many precedents of large volumes of publicly accessible information collated in an automated fashion in order to build a rich dataset. Some may be reluctant to even call this a “data breach”, yet the end result is largely consistent with the previous two examples of malicious intent and unintentionally disclosed data.

We Often Don't Know Until Years Later

We simply have no idea of the scale of data that has been breached. We can measure what we know and conclude that there's an alarmingly large amount of personal information having been exposed, but it's the extent of the "unknown unknowns" that is particularly worrying.

Increasingly, we're realising the significance of the problem. During 2016 and 2017 in particular, we saw many incidents where large data sets belonging to well-known brands appeared after having been originally obtained years earlier. These incidents were frequently of a scale numbering in the millions, tens of millions or even hundreds of millions of customers. In some cases, the organisations involved were aware of a successful attack yet consciously elected not to disclose the incident. Many of the recent large breaches involved companies that *were* aware of unauthorised access to their systems, yet the scope of the intrusion was not known until years later when large volumes of data appeared in the public domain. In other cases, intrusions were entirely unknown until the organisation's data appeared publicly.

I've been personally involved in the disclosure of multiple incidents of this nature directly to the organisations involved. They're consistently shocked – *shocked* – that a breach had taken place and had not seen prior indicators that their data may have fallen into unauthorised hands. The passage of time frequently means that root cause analysis isn't feasible and indeed many of these systems have been fundamentally rearchitected since the original event.

It begs the questions – how much more data is out there? And what are we yet to see from events that have already occurred? We simply don't know nor is there any feasible way of measuring it. The only thing I can say with any certainty is that there is still a significant amount of data out there that we're yet to learn of.

A Perfect Storm of Data Exposure

Data breaches have been increasing in regularity and the incidents themselves have been increasing in terms of the volume of records impacted. There are a variety of factors contributing to what can only be described as a "perfect storm" of data exposure:

Firstly, as mentioned above, the rapid emergence of cloud services has enabled organisations and individuals alike to publish data publicly with unprecedented ease, speed and cost efficiency. The low barrier to entry has meant that it's never been easier to collect and store huge volumes of information and very little technical expertise is required to do so.

Then we have the ever-increasing array of online services collecting data; social media sites, e-commerce, education, even cooking – every conceivable area of human interest has an expanding array of online services. In turn, these services request personal information in order to subscribe or comment or interact with others. As a result, the number of pools of user data on the internet grows dramatically and so too does the total attack surface of information.

The more recent emergence of the class of device we refer to as the "Internet of Things" or IoT is another factor. We're now seeing data breaches that expose information we simply never had in digital format until recently. In recent times, we've seen security vulnerabilities that have exposed data in cars, household appliances and even toys (both those targeted at children and those designed for consenting adults to use in the bedroom). All internet connected and all leaking data that didn't even exist in digital form a few years ago.

Data Maximisation as a Feature

Exacerbating both the prevalence and impact of data breaches is a prevailing attitude of "data maximisation", that is the practice of collecting and retaining as much data as possible. We constantly see this when signing up for services with requests for information that is entirely unnecessary for the function of the service itself. For example, requests for personal attributes such

as date of birth and physical address, both data points that frequently provide no functional benefit to the service.

Further compounding the data maximisation problem is the fact that the retention period of the data usually extends well beyond the period in which the service is used by the owners of the data. (Indeed, even that term – "data ownership" – can be interpreted to mean either the service retaining it or the individuals to whom the data relates.) For example, signing up to an online forum merely to comment on a post means the subscriber's personal data will usually prevail for the life of the service. There are many precedents of data breaches occurring on sites where those who've had their personal data exposed haven't used the service for many years.

Individuals' personal data is also frequently collected without their informed consent, that is it's obtained without them consciously opting in to the service and the purpose for which it's being used. Our data is aggregated, "enriched" and sold (often entirely legally) as a commodity; the people themselves have become the product and alarmingly, we're seeing the aggregation services themselves suffering data breaches both in the US and abroad. In this environment, it's the organisations holding personal data that control it, not the people to whom that data rightfully belongs.

I frequently hear from subscribers of HIBP that they have no recollection of using a service that's suffered a data breach. The alert they receive after the data is exposed is often the first they've heard of the service in many years. In fact, so much time has often passed that they frequently reject the notion that they were members of the site until they discover the welcome email in their archives or perform a password reset and logon to the service. The site was providing zero ongoing value to them yet it still retained their data and subsequently exposed it in a breach.

Data maximisation prevails as a practice for a variety of reasons. One is that it's increasingly cost effective to simply retain everything possible, once again due to the emergence of cloud services as well as rapidly declining storage costs. Another is that purging old data comes at a cost; this is a

feature that has to be coded and supported. It also creates other challenges around technical constraints such as referential integrity; what happens to records such as comments on a forum when the creator of that comment has their record purged? Organisations view data on their customers as an asset, yet fail to recognise that it may also become a liability.

Attempts by individuals to *reduce* their data footprint often lead to frustration. There's frequently no automated way of purging their own personal information and in some cases, organisations have even imposed a financial barrier in a "user-pays to delete" model. Even then, the purging of data from a live system is unlikely to purge that same data from backups that may stretch back years and we've seen many cases of the backups themselves being exposed in breaches.

We need to move beyond an attitude of data maximisation and instead embrace the mantra of "you cannot lose what you do not have".

There's a Lack of Accountability and a Propensity to Repeat Mistakes

Time and time again, we see serious data breaches that impact people's lives around the world and we ask "Is this the watershed moment?" "Is this the one where we start taking things more seriously?" Yet clearly, nothing fundamental has changed and we merely repeat the same discussion after the next major incident.

There's a lack of accountability across many of the organisations that suffer breaches as they're not held strictly liable for the consequences. Despite the near-daily headline news about major security incidents, there remain fundamental shortcomings in the security posture of most organisations. They trade off the cost of implementing security controls against the likelihood of a data breach occurring and inevitably, often decide that there's not a sufficient return on investment in further infosec investments. This attitude contributes to both the frequency and severity of serious security incidents and without greater accountability on behalf of the organisations involved, it's hard to see the status quo changing. There's not enough incentive to do things *right* and not enough disincentive to do them *wrong* therefore the pattern repeats.

Data Breach Redistribution is Rampant

An important factor exacerbating the impact of data breaches is the prevalence with which the data is redistributed once exposed. Data breaches often spread well beyond the party that originally obtained it and the ease with which huge volumes of digital information can be replicated across the globe means that once it's exposed, it spreads rapidly.

There are multiple factors driving the spread of data that has been breached from a system. One is commercial incentives; data breaches are often placed for sale in marketplaces and forums where they may be sold many times over. The personal information contained within these breaches poses value to purchasers ranging from the ability to compromise other accounts of the victims' (frequently due to the prevalence of password reuse unlocking other unrelated services) to value contained within the accounts themselves (such as the ability to acquire goods at the victims' expense) through to outright identity theft (the accounts contain data attributes that help attackers impersonate the victim). In short, there is a return on investment for those who pay for data breaches therefore it has created a thriving marketplace.

More worrying though in terms of the spread of data breaches is the prevalence with which they're redistributed amongst individuals. Data breach trading is rampant and I often liken it to the sharing of baseball cards; two people have assets they'd like to exchange so they make a swap. However, unlike a physical commodity, the trading of data breaches replicates the asset as each party retains their original version, just like making a perfectly reproduced photocopy. Most of those involved in the redistribution of this data are either children or young adults, doing so as a hobby. Often, they'll explain it away as a curiosity; they wanted to see if any of their friends (or sometimes, enemies) were involved. Other times they're experimenting with "hash cracking", the exercise of determining the original passwords when a system stores them as cryptographic hashes. They rarely believe there are any adverse consequences as a result of redistributing the data.

The exchange of data breaches is enormously prevalent. Sites hosting hundreds or even thousands of separate incidents are easily discoverable on the internet; there's often terabytes of data simply sitting there available for anyone to download. Forums dedicated to the discussion of data breaches frequently post links to new breaches or old data which may have finally surfaced. These are not hidden, dark web sites, these are easily discoverable mainstream websites.

Exposed Data is (Often) Immutable and (Usually) Irrevocable

Many of the data classes exposed in breaches are immutable, that is they cannot be changed. For example, people's names, their birth dates, security questions such as their mother's maiden name or even the IP address they were using at the time (which can be used to geographically locate them and potentially tie them to other exposed accounts). Other data attributes may be mutable albeit with a high degree of friction; an email address or a physical address, for example. They may both change over time but the effort of doing so is high and it's unlikely to happen merely because that data has been exposed in a breach.

Paradoxically, the data that is most easily changed is frequently the data people are most concerned about. Credit cards, for example, are often referenced in disclosure statements as not having been impacted by a breach yet a combination of fraud protection by banks and the ability to cancel and refund fraudulent transactions whilst issuing a new card means the real-world impact on card holders is frequently limited and short lived.

Exposed passwords are also easily changed and the impact of them falling into unauthorised hands can be minimal, albeit with one major caveat: The prevalence of password reuse means that the exposure of one system can result in the compromise of accounts on totally unrelated systems. But the password itself is readily changed and unlike immutable personal attributes, doing so immediately invalidates its usefulness.

Frequently, I'm asked how someone's data can be removed from the web; they're a victim of a data breach, now how do they retrieve that data and ensure it's no longer in unauthorised hands? In

reality, that's a near impossible objective, exacerbated by the aforementioned redistribution of data breaches. Digital information replicates so quickly and is so difficult to trace once exposed, there's no putting the data breach genie back in the bottle.

The Emerging Prevalence of OSINT Data and the Power of Aggregation

Data available within the public domain is often referred to as "Open Source Intelligence" or OSINT data. OSINT data can be collated from a range of sources including social media, public forums, education facilities and even public government records to name but a few. It's data we either willingly expose ourselves or is made publicly available by design. Often, the owner of the data is not aware of its publicly available presence; they inadvertently published it publicly on a social media platform or had it put on public display without their knowledge by a workplace or school. In isolation, these data points may appear benign yet once aggregated from multiple sources they can expose a huge amount of valuable information about individuals.

Data aggregation – whether it be from OSINT sources alone or combined with data breaches – is enormously powerful as it can result in a very comprehensive personal profile being built. One system may leak an email address and a name in the user interface, another has a data breach and exposes their home address then that's combined with an OSINT source that lists their profile photo and date of birth. Suddenly, many of the ingredients required to identify and indeed impersonate the individual are now readily available.

The Impact on Knowledge-Based Authentication

Knowledge-based authentication (KBA) is predicated on the assumption that an individual holds certain knowledge that can be used to prove their identity. It's assumed that this knowledge is either private or not broadly known thus if the individual can correctly relay it then, with a high degree of confidence, they can prove their identity. KBA is typically dependent on either static or dynamic "secrets" with the former being the immutable data attributes mentioned earlier (date of birth, mother's maiden name, etc.) and the latter being mutable such as a password.

The risks associated with static KBA have changed dramatically in an era of data breaches and an extensive array of OSINT sources. Further to that is the frequency and effectiveness of phishing attacks which provide nefarious parties with yet another avenue of obtaining personal data from unsuspecting victims. In years gone by, personal data attributes used for verification processes had very limited exposure. For example, one's date of birth or mother's maiden name would normally only be known within social circles which in the past, meant people you physically interacted with. A government issued ID was typically only provided to professional services that had limited exposure.

Now, however, the availability of static KBA data has fundamentally changed yet its use for identity verification prevails. The threat landscape has progressed much more rapidly than the authentication controls yet we're still regularly using the same static KBA approaches we did before the extensive array of OSINT sources we have available today and before the age of the data breach.

Closing

Data breaches will continue to grow in both prevalence and size for the foreseeable future. The rate at which we willingly share personal data will also continue to grow, particularly with an increasing proportion of the population being "internet natives" who've not known a time where we *didn't* willingly share information online. Increasingly, the assumption has to be that everything we digitise may one day end up in unauthorised hands and the way we authenticate ourselves must adapt to be resilient to this.