7 Taking the Growth of the Internet Seriously When Measuring Cybersecurity

Eric Jardine

No one seriously thinks of the Internet as a static system. By both design and effect, the Internet ecosystem exists in an inherently plastic and fluid state. In two very particular ways—in terms of both its scale and composition—the Internet has morphed from its early beginnings as a small set of networked research computers into something much bigger and far more diverse.

Each day, for example, thousands of new entrants connect to the Internet for the very first time. While year-over-year growth of the Internet-using population is slowing, it remains above 10 percent per annum (Meeker 2017, 6). With 3.7 billion users in 2016 (Internet World Stats 2017), this rapid pace of change resulted in potentially over 4 billion users in 2018. By dint of global demographics, many of these new users hail from developing-world countries and more autocratic regimes, potentially leading to growing levels of contention in the global Internet governance regime (Bradshaw et al. 2015).

Increasingly, the Internet is also losing its human face as it rapidly becomes predominately an interconnection of nonhuman devices. By dint of sheer numbers, the Internet of Things (IoT), as it is commonly known, is fundamentally changing the nature of the online environment. In 2014, for example, there were some 3.8 billion IoT devices (Gartner Research 2015). By 2016, that number was predicted to have increased to 6.39 billion. In 2015, Gartner Research (2015) forecasted that almost 21 billion IoT devices could be connected to the Internet by 2020. In 2020, Statista Research puts the number of IoT devices at 30.73 billion and offers a prediction of 74.44 billion devices online by 2025 (Statista 2020). Moreover, while the pace of new-user growth is slowing, the rate of new IoT interconnections is still in an acceleratory phase, exhibiting a classic S-shaped pattern of innovation diffusion that was earlier followed by Internet connectivity as a whole (Hampson and Jardine 2016; Rogers 2010).

These changes to the underlying user base and composition of the Internet are fundamental. The wider Internet governance literature has accommodated the changing Internet landscape well. For example, as detailed in chapter 3 by Mueller and Badiei, Internet governance—and literature on the governance of the Internet more generally—has undergone several distinct phases of work and focus since the early 1990s, with a big move toward securitization and concerns over privacy since 2010. Likewise, Cogburn in chapter 9 shows with a quantitative textual analysis of Internet Governance Forum (IGF) transcripts that the dominant themes at the IGF have changed over time. In 2006, the three dominant themes were "mobile," "youth," and "ICANN" (the Internet Corporation for Assigned Names and Numbers). In 2011, the focus had shifted toward more rights-based issues, with "human rights," "young people," and "freedom of expression" being the most expressed phrases. Lastly, in 2017, changes again occurred with "women," "cybersecurity," and "news" being the top phrases in the corpus of IGF text. Other research has also started to contemplate what will happen to Internet governance when all devices are interconnected (Van Eeten 2017) or the implications of the IoT for law and policy within the wider Internet governance regime (DeNardis and Raymond 2017). In short, coping with change is a big part of Internet governance practice, and it is well captured in many strands of the literature.

And yet, as I argue here, coping with ecosystem change at the level of cybersecurity measurement is less often done. When developing metrics to measure and assess the state of cybersecurity, the changing scale and composition of the Internet is often ignored. This error is implicit in the measures themselves. Two errors, in particular, are potentially distorting our sense of the state of cybersecurity. First, trends in security incidents as reported by information technology (IT) security vendors and governments tend to be depicted as absolute count (e.g., 1 million web-based attacks in 2016). This number fails to account for the significant increase in new users of the Internet (Jardine 2015, 2018). As the ecosystem grows, attacks, hacks, or vulnerabilities will also increase, everything else being equal. For measures of cybersecurity incidents to be meaningful, they need to be expressed as a rate, which requires that the absolute count of security incidents, say, be normalized around the size of the ecosystem, be it users, websites, data flows, or emails.

The second way in which current measures of cybersecurity are potentially misleading is a result of their often extreme level of aggregation (Jardine

2017). Devices and users of every kind are often grouped together as if they are effectively the same thing. The trouble here is that different parts of the ecosystem are growing faster than others, and moreover, the fastest-growing parts, such as the IoT, tend to be the most susceptible to being hacked (HP 2017). This pairing creates a very real danger of what is known as Simpson's Paradox, in which aggregate trends in security incidents show things getting worse while the disaggregated trends show the state of online security to be improving over time (Jardine 2017). Data from US data breaches can showcase the problem, but simulation results suggest that the phenomenon occurs in any situation in which groups are differentially vulnerable, and the fastest-growing group is the most vulnerable of the lot.

In the chapter's first section, I discuss trends in current cybersecurity metrics. I illustrate several potential perils and pitfalls in this section, too, and propose some cautionary steps that can help researchers produce the best possible metrics of cybersecurity. In the second section, I show that even the best cybersecurity measures will produce biased over-time trends if they do not take into account the ongoing growth Internet denominator, to use the mathematical term. The third section points to existing over-time trends in cybersecurity measures. It shows further how Simpson's Paradox (or aggregation bias, as it is also known) exists in data breach data and plausibly biases many aggregate indicators of insecurity. I conclude with thoughts on why accurate measurement of trends in cybersecurity matters and how cybersecurity researchers can approach their topic to form the best possible descriptive inferences.

What We Know about Measuring Cybersecurity

Measures of insecurity are never in short supply within the Internet ecosystem. Yet many such measures are hardly worth the bytes on which they are stored. Even the simple estimate of the total cost of cybercrime is hugely variable. One study by the Center for Strategic and International Studies and the antivirus vendor McAfee (2014), for example, estimated that the likely global cost of cybercrime was on the order of $400 billion. More recent estimates have ballooned these already significant numbers further still. A year after the center and McAfee produced their figure, Juniper Research (2015) estimated that the global costs of cybercrime could reach as high as $2 trillion per annum as early as 2019. Another report, by the firm Cybersecurity

Ventures, predicted that global cybercrime costs could actually have been as high as $3 trillion in 2015 and may well potentially increase by stunning proportions to as much as $6 trillion in 2021 (Morgan 2016).

Much of these costs are the result of an accumulation of a thousand small cuts. Each year, firms of all stripes and sizes are breached and their records are stolen or otherwise compromised. The Ponemon Institute conducts an annual survey of the cost of data breaches among hundreds of firms globally. As detailed in figures 7.1 and 7.2, both the average cost per breached record and the average total cost that a firm faces in the aftermath of a breach are undergoing a general rise.[1] In 2011, for example, the cost per breached record was $130. By 2017, the average breach cost had increased to $148 per record, amounting to a not inconsiderable increase of 13.9 percent. The total average cost that firms need to absorb in the fallout of a data breach is also increasing, rising from $3.13 million in 2011 to roughly $3.86 million in 2017 (Ponemon Institute 2013–2018).

These trends in the cost of data breaches are roughly mirrored by the movement in their frequency. Within the United States, the Privacy Rights Clearinghouse (2020) has collated an ongoing record of US data breaches from 2005 onward. Both the number of breaches and the summed total number of compromised records is increasing over time. The total number of US data breaches (from all threat vectors), for example, increased from

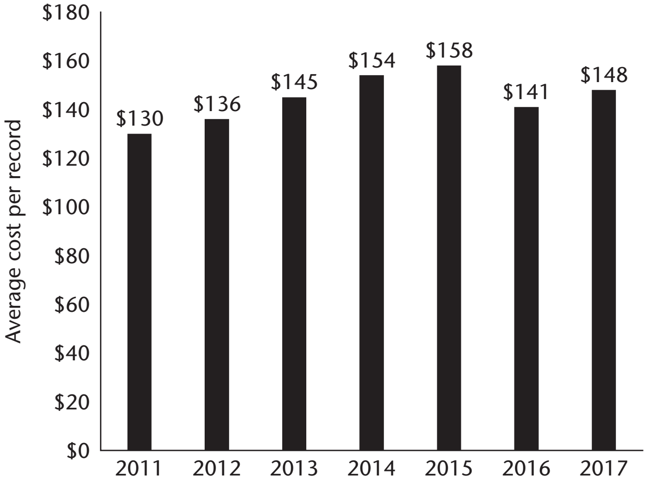

Figure 7.1
Ponemon Institute average cost per record breached.

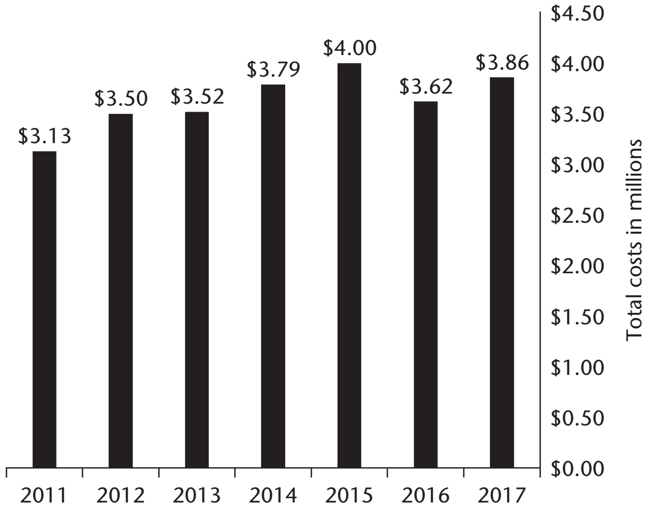

Figure 7.2
Ponemon Institute global data breach costs.

136 in 2005 to 832 in 2016, amounting to a total increase of 512 percent. The total number of breached records, on the other hand, has increased at a far more prodigious rate. Here, the total number of compromised records[2] has increased from 55,101,241 in 2005 to 4,626,238,665 in 2016, which includes the 2013 megabreach of 1 billion Yahoo! users. This change amounts to an increase of 8,296 percent. (We return to these numbers in the third section, as the story they tell can be deceptive.)

These numbers suggest that one massive ongoing shift in the cybersecurity environment is toward more pronounced outcomes, what Nassim Taleb (2010, 33–34) would call the world of "extremistan." This notion is broadly in line with much of the growth in data breaches being accounted for by the existence of so-called heavy tails, or extreme outliers that drive a lot of the variation (Edwards, Hofmeyr, and Forrest 2016). Growing digitization, interconnectivity, and market concentration all affect cyber risk and help make the occurrence and fallout of a data breach far more pronounced (Geer, Jardine, and Leverett 2020).

These measures all point in the same approximate direction, indicating roughly that the state of cybersecurity is worsening. IT security practitioners have certainly internalized this message. The online Index of Cybersecurity (2020), for example, conducts a monthly survey of IT security professionals to gauge their concern over the state of the online environment. From a base value of 1,000 in March 2011, the index score has exhibited a decidedly worsening trajectory, moving to a score of 4,762 on July 4, 2019. These collective sentiments are broadly in line with the idea of former high-level policy makers who openly fret about the potentially severe consequences of a cyberattack. Former Homeland Security secretary Tom Ridge, for example, argued in late 2016, "Notwithstanding the pain and horror associated with a physical attack, … the potential for physical, human, and psychic impact with a cyberattack, I think, is far more serious" (Roberts 2016). Clearly, the general sentiment is that efforts to improve cybersecurity outcomes are, so far at least, a losing battle.

At the same time, however, there are a number of telling reasons to suggest that we might know less than we think about the state of online security. The broad point of the literature is that most measures are, as is often the case, biased—and only a very select few useful. What this means in practice is that our view that cybersecurity is a mug's game might not be as well grounded in empirical reality as we might like.

Anderson and his colleagues, for example, produced a rigorous estimate of the cost of cybercrime, which is far less than the trillions upon trillions of dollars cited earlier. This research team started by disaggregating costs into defense costs, indirect losses, and the direct losses from cybercrime (2012, 5). The disaggregation of costs is an important step, as direct costs to victims or the benefits to attackers are often only a small portion of the overall social costs due to cybercrime. The difficulty is the working of network externalities, only in reverse (Anderson and Moore 2006). Rather than the benefits of network interconnection increasing at the pace of Metcalfe's Law (i.e., number of connected devices, n^2), the tendency is for the costs of cybercrime to increase disproportionately as network interconnection goes up. But a majority of these social costs do not translate into dollars in the bank accounts of malicious actors.

The point is well shown by the huge gap that can exist between what cybercriminals make and the destruction that they can bring. In one study of the individual economics of phishing campaigns (Herley and Florêncio 2009), the authors persuasively argue that many estimates of the high value of phishing campaigns for individual attackers (sometimes on the order of thousands of dollars per day) is likely woefully inaccurate. Since phishing attacks target a common pool resource (users' wallets) that cannot regenerate at an infinite pace, the phishing activity of multiple attackers is likely to drive down the benefits until each attacker is basically earning what their skills would return in the noncriminal world. In more concrete dollar terms, Herley and Florêncio (2009, 67) suggest that rather than earning thousands a day, as some accounts suggest, phishers in 2008 likely shared a $61 million pot some 113,000 ways, or roughly $540 per attacker. At the same time, the social cost of phishing campaigns is enormous, with billions being poured into defense and remediation.

Overall, these numbers suggest that estimates of the high-flying benefits of cybercrime might be greatly exaggerated and that there is likely an asymmetry, due to network externalities, between malicious actor benefits and the damage done. But troubling estimations are not confined simply to metrics for the monetary gain to malicious actors. The problems with correctly measuring cybersecurity indicators of various stripes are manifold.

One such example is the lack of a common measurement foundation from which conclusions can be drawn (Brecht and Nowey 2013). A good concrete instance of this problem of sample composition and metric design is to look again at the cost of data breach studies produced by the Ponemon Institute. The object of Ponemon's efforts are twofold: provide a static snapshot of what is going on in the ecosystem and provide an over-time measure to assess whether data breaches are getting better or worse. The Ponemon data breach studies are helpful on the first count but suffer from a shifting-sample problem on the second.

As detailed in table 7.1, the Ponemon results are based on a growing population of both countries (increasing from 8 in 2011 to 19 in 2017) and firms (from 199 in 2011 to 477 in 2017). The growing sample increases

Table 7.1
Ponemon Institute's shifting sample.

	Countries	Firms
2011	8	199
2012	9	277
2013	10	314
2014	11	350
2015	12	383
2016	13	419
2017	19	477

coverage. But this shifting-sample population can also introduce biases when looking at the trends in the data over time. If a newly added country is from a highly targeted part of the world, then the average will be pulled up, especially since data breach outcomes are highly power-law distributed and characterized by extreme variance (Edwards, Hofmeyr, and Forrest 2016). All the other countries could be performing the same or even better than before, but the new entrant to the sample makes the over-time trend appear as though things are worsening. Likewise, if the new firms that join the sample are highly targeted (firms from the finance or health care sectors, say), then the averages could again be biased upward. The point, really, is that with a shifting-sample population it is hard to say if an increase in the average cost per breached record is really due to an increased cost per firm or due to the inclusion of new actors whose average cost per record is higher than the rest of the population under question (more on this problem of aggregation in the third section).

In some ways, the law of large numbers—in which averages become stable over a large enough population size—suggests that the shifting-sample base might not have much effect. Randomly sampling (which is not really what is being done) from the population of firms should eventually lead to a stable sample mean and normal distribution. However, stable means are common only in a Gaussian, or bell-shaped, distribution. When the phenomenon that is being surveyed follows a more extreme distribution, such as an exponential or power-law distribution, then measures of average costs can become terribly misleading (Florêncio and Herley 2013). Indeed, when power laws are at work, extreme values of outliers in the sample can cause shifts in the mean so radical that the average becomes nearly meaningless as a guide for policy.[3] Median cost—which Ponemon does not, but probably should, use—is then a simple and far better estimation of a middling value.

A host of problems are added on to issues of incorrectly estimated costs, shifting samples, and extreme distributions. Many areas of the social world, for example, use voluntary reporting mechanisms to tally up who has been subjected to what and why. Surveys are a classic example. The problem here is that cybercrime surveys can be marred by a curious admixture of both under- and overreporting (Florêncio and Herley 2013; Kshetri 2006). For some, being the victim of a cybercrime is a traumatic event and one that people want to express out loud, in print or online, to whomever will listen. For others, incentives might align to discourage disclosure of a hack, which probably happens in countries lacking mandatory data breach disclosure regulations for firms (Laube and Böhme 2016). Additionally, a lack of cybersecurity loss data stemming from incidents complicates risk pricing, necessitating cleverly innovative (yet somewhat error-prone owing to extrinsic factors) actuarial models (Woods, Moore, and Simpson 2019).

Finally, it is more than possible that the production of cybersecurity metrics can be captured by vested interests. The most obvious potential set of culprits, but by no means the only ones, are IT security vendors. These companies are in the uncomfortable position of researching a topic and producing estimates about the state of cybersecurity while also selling a product the demand for which is based, at least in part, on perceptions of online insecurity. This tension does not always play out in favor of bias, but it often can. As Anderson and colleagues point out,

> One problem with Symantec is a conflict of interest: their reports are published principally for marketing reasons. It is not surprising, then, that sometimes their data are consistently over-reported. For example, we studied more closely one statistic measured in several reports, which tracks the proportion of malicious code that exploits confidential information. In volume 12 of the report, covering January to June 2007, 65% of malicious code exploits confidential information, compared to just 53% in the previous six months. However, the earlier report claimed 66% for this period of July to December 2006. (2008)

There are, clearly, a host of different methodological and pragmatic biases that can cloud our view of cybersecurity. However, while improved measurement techniques that account for missing evidence, shifting samples, extreme distribution types, under- and overreporting of statistics, and vested interests would all be a step in the right direction, even the best metrics as they currently stand will remain biased. The trouble here is that some fundamental changes to the Internet ecosystem—most notably its growing scale and shifting demographic constitution—have not been accounted for well enough in the formulation of cybersecurity metrics. In the next two sections, I detail the distorting effects that these oversights can have on our view of cybersecurity.

Expressing Cybersecurity Incidents as a Rate

In the analogue world, indicators of insecurity or crime are always expressed as a rate (for example, 1,000 armed robberies, murders, or burglaries per

100,000 people). Expressions of this sort are useful because they correct for the simple fact that a larger population size should see more activity, both good and bad (see Jardine 2015, 2018). In an ecosystem with a denominator that is growing as fast as the Internet in terms of users, interconnections, traffic volume, and more, failing to correct for the changing size of the system will bias even the best-devised metrics of cybersecurity. The publication of statistics like the count of breached records is, in other words, inherently distortive and misleading.

Indeed, if the denominator (i.e., a measure of the size of the ecosystem) is growing fast enough, then the divergence between counts (number of records breached) and rates (number of records breached per 100,000 records or Internet users) can be significant, often so much so that one set of numbers could actually indicate that the state of cybersecurity is getting worse while the other indicates that things are getting better.

Consider, as a useful case in point, a comparison of the count of phishing websites taken from reports by the Anti-Phishing Working Group (2008–2016) with that same count normalized around the number of websites (see Jardine 2018 for more details on this analysis). As detailed in figure 7.3, phishing websites per quarter are steadily increasing in number. Certainly, more attack vectors do not always mean more insecurity necessarily. And phishing websites are a good example of that principle. With low latency blacklisting and other technical protections built into most modern browsers, malicious websites are not likely to remain effective for very long and are far less likely now than in the past to be able to infect a passerby in a drive-by attack. That being said, for a constant level of technological protections, more threats should roughly translate into more insecurity, so leaving

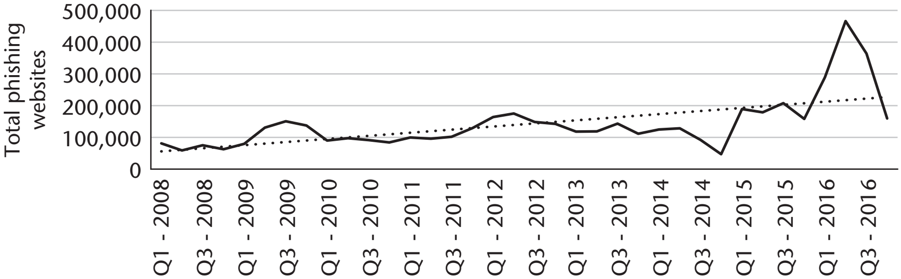

Figure 7.3
Observed phishing websites.

aside potential remedial countermeasures for a moment, let's just assume that more phishing websites could mean more insecurity online. Under that assumption, figure 7.3 clearly shows a worsening state of online security.

A starkly different picture emerges if you look, not at the count of the total number of observed phishing websites, but instead at per year rate of phishing websites per 100 websites or per 1 million Google searches.[4] These two denominators measure two separate aspects that help contextualize trends in the number of phishing websites. Obviously, phishing websites exist within a pool of all websites, so they are some proportion of the whole, just as the victims of crimes are some proportion of the whole population in a given area. Google searches are a measure of online activity. While Google's PageRank algorithm ensures that people are not simply wandering around randomly online, Google search activity could capture the amount of time people are spending online and so the amount of time within which they might stumble across a malicious website.

Counts and rates, in this case, tell very different stories. As shown in figure 7.4, rather than seeing a worsening situation as was the case with more phishing websites, both normalized measures of the rate of phishing websites are in decline. This declining rate suggests a couple of things. First, it means that the number of websites or Google searches is growing

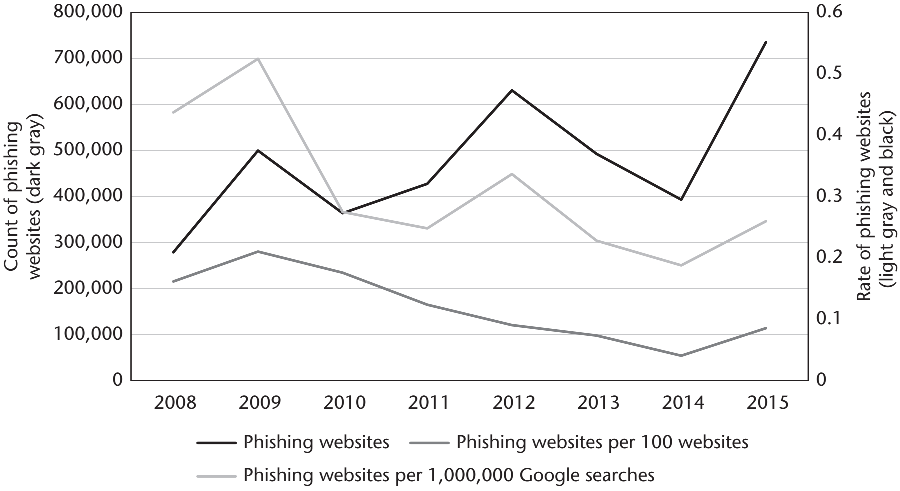

Figure 7.4
Counts versus rates of phishing websites.

faster than the number of malicious websites. This differential growth rate effectively entails that the population within which phishing websites live (all websites) or people's time online (Google searches) are outpacing the growth of phishing websites. Second, framed in propensity-based terms, these differential growth rates mean that it is increasingly *much less likely* that an average person doing average things online is going to encounter a malicious website. In this very important sense, using rates that capture the growing size of the Internet ecosystem when measuring cybersecurity tells a story of improved security.

Rates may or may not say that online security is getting better—if they deterministically did, then something is wrong with the measure. But, by simple dint of the mathematics involved, so long as the Internet ecosystem continues to grow, the rate will always paint a better picture than the absolute count of incidents, vulnerabilities, or costs. Indeed, as I have shown before (Jardine 2015), when the Internet is getting bigger, rates will favorably contrast with counts for year-over-year comparisons in one of the following three ways:

- The count says things are getting worse, but the rate says cybersecurity is getting better.
- The count says things are getting better, but the rate says cybersecurity is improving even faster.
- The count says things are getting worse, but the rate says that cybersecurity is getting worse at a slower pace.

These differences are crucial, as they imply that a year-over-year count of zero-day exploits, security incidents, phishing attacks, and so on will be biased. Using the best-designed numerator measure in the world—a design that is challenging in its own right—helps avoid some of the problems sketched out in the first section, but a failure to normalize these very same metrics around the growing size of the Internet ecosystem will result in a flawed measure that is not only potentially exaggerating the poor state of cybersecurity but that may well say everything is getting worse when it is actually getting better.

Aggregation and Bias in Cybersecurity Measures

There is, too, another way in which cybersecurity researchers may fail to adequately grapple with the idea that the Internet is a dynamic system that is growing and changing right under our feet. In this case, the pressing problem is the level of aggregation of most publicly available cybersecurity data. When data from various groups are aggregated, the results can fall prey to what is known as Simpson's Paradox (Simpson 1951). When Simpson's Paradox is at work, aggregate trends that show a worsening state of cybersecurity can actually be based on underlying data that show the state of cybersecurity to be improving year-over-year (Jardine 2017). Overaggregation, in other words, can produce bias.

Unfortunately, a lot of currently available trend data in cybersecurity exist at just such an extreme level of aggregation. IT security software vendors, such as Kaspersky Labs or Symantec, produce annual reports detailing the volume of threats that they contend with each year. While the data are a bit noisy, the general trend line observed by both companies is decidedly positive, tending toward more and more attacks over time.

Figure 7.5, for example, plots two forms of online attacks as observed by Kaspersky and Symantec. In Kaspersky's case, its annual Security Bulletin (2008–2016) contains a count of the total number of attacks launched from web-based sources, which would exclude some forms of malicious activity such as insider threats or USB-based malware. Symantec, for its part, includes a count for the number of blocked attacks in its annual Internet Security Threat Report (2013–2016).

In both cases, observed attacks are generally going up over time. Kaspersky's measure, for instance, has increased from 23,680,646 attacks in 2008

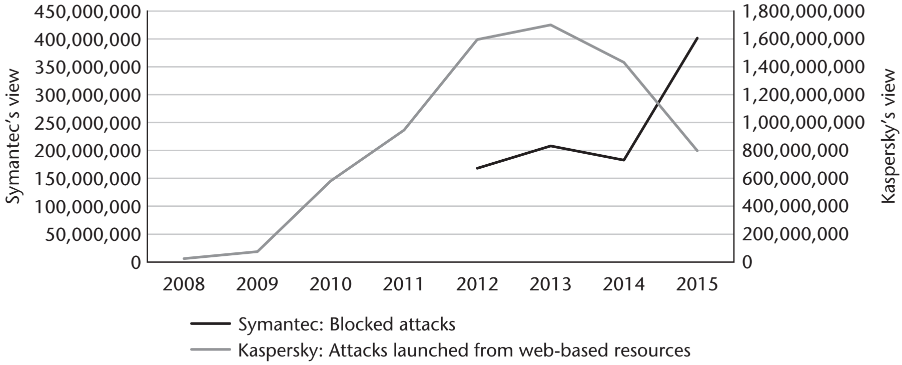

Figure 7.5
Aggregate attacks over time.

to 798,113,087 attacks in 2015, amounting to a growth in attacks of some 3,270 percent. Symantec's blocked attacks have likewise increased, by some 139 percent, growing from 167,900,000 blocked attacks in 2012 to 401,500,000 attacks in 2015. Obviously, the former point about the perils of count measures (as opposed to rates) suggest that these worsening trends might not be as bad as they seem. But these trends might be misleading for an additional reason: they fail to factor in the changing face of the Internet ecosystem. These statistics pay no attention to the type of devices being targeted, the baseline vulnerability of different user groups, the sectors within which these devices are housed, or even the region of the world where the devices reside. Aggregating these differences is potentially a major source of bias.

Take another example of seemingly bad overall trends, where sufficiently disaggregated data allow for a more fine-grained analysis of the problem of aggregation. The Privacy Rights Clearinghouse provides a record of publicly known data breaches in the United States. Asking what the trend is in data breaches and in the number of breached records shows the problem of overaggregation. Figure 7.6 shows the trend in the number of disclosed data breaches that resulted from all recorded attack vectors. There are generally more data breaches now than there were in the middle of the first decade of the 2000s. Figure 7.7 does the same for both the total count of breached

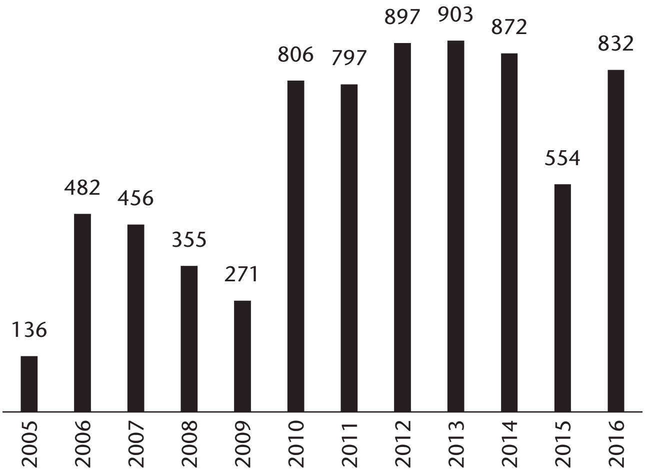

Figure 7.6
Total number of breaches.

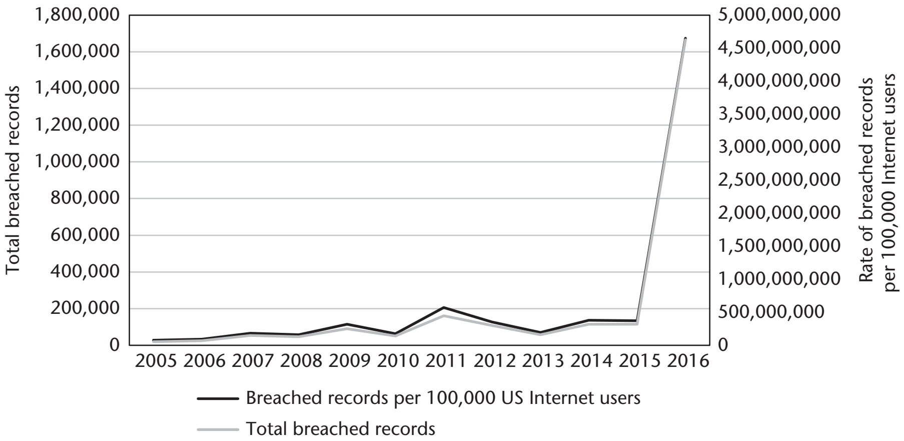

Figure 7.7
Total counts and rates of number of breached records.

records and the rate of breached records per one hundred thousand US-based Internet users. The trends track one another and are decidedly positive. From these three aggregate figures, whether the count of breaches or counts or rates of the number of compromised records, the primary conclusion to be drawn seems to be that the trend in cybersecurity is heading in a worsening direction.

But to its credit, the Privacy Rights Clearinghouse also provides a record of attacks disaggregated by sector. This disaggregation allows an analysis of over-time trends at a finer-level detail. In other words, it allows a sense of what is happening in specific parts of the economy, not just overall. Using data from the Privacy Rights Clearinghouse (2020), Table 7.2 highlights what is happening in the aggregated trend in the number of data breaches, the count of breached records, and the rate of breached records per one hundred thousand US Internet users. It also shows what is happening in each disaggregated sector.

Each aggregate measure tells the same story: cybersecurity is getting worse. The disaggregated story is far more nuanced and often the opposite. When counting the total number of breaches, the aggregate trend is positive, but only three (other business, retail, and health care) experienced more breaches in 2005–2016. In short, the total count of data breaches per year is declining in the majority of sectors. Likewise, the overall trend in the number of breached records is positive, and is similarly positive for all but

Table 7.2

Aggregate versus disaggregate breach data, 2005–2016.

Trend direction	Aggregate	Financial services	Other business	Retail	Education	Gov/mil	Health care	Nongovernmental
Total breaches	Positive	Negative	Positive	Positive	Negative	Negative	Positive	Negative
Counts of records	Positive	Positive	Positive	Positive	Negative	Positive	Positive	Negative
Breached records per 100,000 US Internet users	Positive	Negative	Positive	Positive	Negative	Negative	Positive	Negative

the educational and nonprofit sectors. But as discussed in the previous section, counts of compromised records are deceptive and are best normalized around the changing size of the Internet ecosystem. Doing so with a crude measure of the number of US-based Internet users taken from World Bank Indicators (a proxy for the extent of digitization) still produces a positive overall trend in the rate of breached records. The disaggregate story is radically different, however. Now, four of the seven categories exhibit a negative trend in the breached records rate per one hundred thousand Internet users from 2005 to 2016.

The point is that the level of aggregation in the data can decidedly affect the view that you get. Further disaggregation within each sector would likely review additional nuances. Grouping firms on the basis of their level of information security sophistication would be even more revealing. These results are also not confined to data breaches in which some semblance of sufficiently disaggregated data allow an illustration of the problem. In fact, this problem of Simpson's Paradox, or aggregation bias, emerges whenever three simple conditions hold (Jardine 2017):

- The online population can be divided into groups.
- Some of these groups are more vulnerable than others.
- Each group's rate of growth mirrors its vulnerability, with the most susceptible group growing the fastest.

To phrase these conditions a bit differently, if groups of online users or devices can be rank ordered in terms of vulnerability and the most vulnerable groups grow faster than the less vulnerable groups, then extensive simulation results in both normal and power-law-distributed settings indicate that aggregate trends will show a worsening situation, even as the disaggregate results all show security improving (Jardine 2017). The massive growth rate of highly vulnerable IoT devices almost ensures that overaggregation is affecting cybersecurity statistics today, making things seem worse than they really are. More generally, if a get-to-market-first-and-do-security-later attitude characterizes the future ethos of technology expansion (as it has in the past), then Simpson's Paradox is likely to be a persistent issue for cybersecurity metrics. Aggregate trend data, in other words, that fail to take into account the changing composition of the Internet are likely to be biased, because the focus on the forest misses all the trees.

Conclusions for Policy and Research

This chapter advances a simple idea. While cybersecurity researchers are highly attuned to some discrete changes in the Internet ecosystem (shifting malware variants, for example), many of our current metrics tend to miss the forest for the trees by ignoring the very real measurement implications of an ever-changing online world. In this chapter, I show that a failure to recognize these fundamental shifts in the metrics we use can produce measurement bias in two ways: via a lack of normalization and through overaggregation.

Drawing on some of my past research (Jardine 2015, 2018), I have shown here how using a count measure to express insecurity introduces bias in year-over-year trends by failing to grapple with the simple fact that a growing Internet should have more malicious activity. Instead, I argue that cybersecurity researchers need to adopt an age-old method from the analogue world and start normalizing cybercrime statistics around the growing Internet. The Anti-Phishing Working Group, whose phishing website data I present here, could usefully begin normalizing its counts to great effect—assuming a good measure of an agreed-on denominator could be found.

Of course, the most effective denominator for each measure of insecurity is still an open question. The number of Internet users offers a simple generic measure that will suffice in many cases. Yet in some discrete instances, such as zero-day exploits, a better denominator might be either lines of code (if we knew that number) or computer programs. Zero-day exploits in the code of mobile applications, for example, could be normalized around the total number of mobile applications to produce a better measure than just a count of vulnerabilities. For phishing websites, the population of domains, search traffic, or websites might work best in other cases, as discussed earlier.

I also show that more attention needs to be paid to the changing face of the Internet ecosystem in the process of devising cybersecurity metrics and collecting and analyzing data. Overly aggregate measures can point in one direction, while disaggregated numbers can show another story altogether. This phenomenon of Simpson's Paradox is quite general and likely holds whenever groups have different levels of vulnerability and growth rates (Jardine 2017). Data breach data from the United States exemplifies the potential problem. On the basis of total counts of breached records, for example, a policy maker, regulator, or corporate C-suite executive might decide that more needs to be done to improve cybersecurity. But the disaggregate trend in that firm's sector might show an improving scenario, suggesting that current measures and levels of investment are sufficient. All policies and investments have opportunity costs. Missing out by spending too much on security can reduce the usability of services and forgo other, potentially more beneficial activities.

The results of this analysis are important for both policy makers (broadly defined) and Internet governance researchers.

For policy makers, the results of the analysis suggest that current cybersecurity trends might be misleading. While more and better data would be needed to say for sure, the results presented earlier often indicate an improvement, rather than worsening, of the state of cybersecurity in both phishing websites and US-based data breaches across many discrete sectors of the economy. Data ideally inform public policy development, as bad policy can cause a host of unintended consequences and waste resources countering problems that are only partially real. Good policy follows only from good metrics, however, and this chapter proposes two discrete ways in which stakeholders, such as governments, computer security incident response teams, and security vendors, can collect more meaningful data.

For researchers, the analysis speaks to a few practices that ought to be adopted during data collection and analysis: normalize counts and, as much as possible, disaggregate data to control for potential so-called lurking confounders that give rise to Simpson's Paradox. Beyond that, the research prescriptions also speak to many of the core themes of this research compendium volume. Proprietary or closed systems, for example, diminish the transparency of collected data. Such a lack of transparency makes assessing data quality more difficult, but it also makes research-initiated corrections difficult to do after the fact. Large (big) data—an increasingly common feature of today's online environment—are potentially a boon, since they can include sufficient information to properly normalize incident counts and disaggregate trends to sufficiently fine-grained levels. Last, growing market concentration and increasing technical complexity are persistent trends online that directly affect cybersecurity risk (Geer, Jardine, and Leverett 2020) and promise to continue the constant ebb and flow of the scale and composition of the Internet ecosystem in a way that will, undoubtedly, affect cybersecurity measures. Researchers need to be ever vigilant to ongoing changes online, as good policy and research requires good measures and data.

Notes

1. The Ponemon Institute studies are published the year after the year in which the data are collected. For example, the data in the 2016 report is typically for 2015.

2. The Privacy Rights Clearinghouse draws a distinction between a breached record involving financial data and records involving nonsensitive material such as emails. I use the total number of records here. These breaches are also only in deliberately malicious attacks (not from accidental disclosures or in unknown attacks) and with a known target.

3. In a power-law probability distribution, the smaller the alpha, the more pronounced the instability in the mean (see Neumann 2015).

4. The per 100 and per 1,000,000 designators are used to make the figure intelligible. They have no material effect on the analysis.

References

Anderson, R., Barton, C., Bohme, R., Clayton, R., van Eeten, M. J. G., Levi, M., Moore, T., & Savage, S. (2012). *Measuring the cost of cybercrime*. WEIS 2012. Retrieved from http://www.econinfosec.org/archive/weis2012/papers/Anderson_WEIS2012.pdf

Anderson, R., Bohme, R., Clayton, R., & Moore, T. (2008). *Security economics and European policy*. The Workshop on the Economics of Information Security, WEIS 2008. Retrieved from http://www.econinfosec.org/archive/weis2008/papers/MooreSecurity.pdf

Anderson, R., & Moore, T. (2006). The economics of information security. *Science, 314*, 610–613. Retrieved from http://tylermoore.ens.utulsa.edu/science-econ.pdf

Anti-Phishing Working Group. (2008–2016). *Phishing attack trends reports*. Retrieved from http://www.antiphishing.org/resources/apwg-reports/

Bradshaw, S., DeNardis, L., Hampson, F. O., Jardine, E., & Raymond, M. (2015). *The emergence of contention in global Internet governance*. Global Commission on Internet Governance Paper Series (Paper no. 17). Centre for International Governance Innovation/Chatham House. Retrieved from https://www.cigionline.org/sites/default/files/no17.pdf

Brecht, M., & Nowey, T. (2013). A closer look at information security costs. In R. Böhme (Ed.), *The economics of information security and privacy*. Berlin, Germany: Spring Science & Business Media.

Center for Strategic and International Studies (CSIS) & McAfee. 2014. *Net losses: Estimating the global cost of cybercrime*. Centre for Strategic and International Studies. Retrieved from https://csis-prod.s3.amazonaws.com/s3fs-public/legacy_files/files/attachments/140609_rp_economic_impact_cybercrime_report.pdf

DeNardis, L., & Raymond, M. (2017). The Internet of Things as a global policy frontier. *UC Davis Law Review, 51*(2), 475–497.

Edwards, B., Hofmeyr, S., & Forrest, S. (2016). Hype and heavy tails: A closer look at data breaches. *Journal of Cybersecurity, 2*(1), 3–14. Retrieved from https://doi.org/10.1093/cybsec/tyw003

Florêncio, D., & Herley, C. (2013). Sex, lies and cyber-crime surveys. In B. Schneier (Ed.), *Economics of information security and privacy III* (pp. 35–53). New York, NY: Springer. Retrieved from https://link.springer.com/chapter/10.1007/978-1-4614-1981-5_3

Gartner Research. (2015). *Gartner says 6.4 billion connected "things" will be in use in 2016, up 30 percent from 2015*. Retrieved from http://www.gartner.com/newsroom/id/3165317

Geer, D., Jardine, E., & Leverett, E. (2020). On Market Concentration and Cybersecurity Risk. *Journal of Cyber Policy, 5*(1).

Hampson, F., & Jardine, E. (2016). *Look who's watching: Surveillance, treachery, and trust online*. Waterloo, Canada: Centre for International Governance Innovation Press.

Herley, C., & Florêncio, D. (2009). A profitless endeavor: Phishing as tragedy of the commons. *NSPW '08 proceedings of the 2008 New Security Paradigms Workshop* (pp. 59–70). Retrieved from http://dl.acm.org/citation.cfm?id=1595686

HP. (2017). *HP study reveals 70 percent of Internet of Things devices vulnerable to attack*. Retrieved from http://www8.hp.com/us/en/hp-news/press-release.html?id=1744676#.WSMGKmjysuU

Index of Cybersecurity. (2020). Retrieved from https://wp.nyu.edu/awm1/

Internet World Stats. (2017). *Internet usage statistics: The Internet big picture*. Retrieved from https://www.internetworldstats.com/stats.htm

Jardine, E. (2015). *Global cyberspace is safer than you think: Real trends in cybercrime*. Global Commission on Internet Governance Paper Series (Paper no. 16). Centre for International Governance Innovation/Chatham House. Retrieved from https://www.cigionline.org/sites/default/files/no16_web_0.pdf

Jardine, E. (2017). *Sometimes three rights really do make a wrong: Measuring cybersecurity and Simpson's paradox*. Paper presented at the 16th Workshop on the Economics of Information Security, WEIS 2017. Retrieved from http://weis2017.econinfosec.org/wp-content/uploads/sites/3/2017/05/WEIS_2017_paper_18.pdf

Jardine, E. (2018). Mind the denominator: Towards a better measurement system for cybersecurity. *Journal of Cyber Policy, 3*(1), 116–139.

Juniper Research. (2015). *Cybercrime will cost businesses over $2 trillion by 2019*. Retrieved from https://www.juniperresearch.com/press/press-releases/cybercrime-cost-businesses-over-2trillion

Kaspersky Labs. (2008). *Kaspersky security bulletin 2008*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/36241/kaspersky-security-bulletin-statistics-2008/

Kaspersky Labs. (2009). *Kaspersky security bulletin 2009*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/36284/kaspersky-security-bulletin-2009-statistics-2009/

Kaspersky Labs. (2010). *Kaspersky security bulletin 2010*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/36345/kaspersky-security-bulletin-2010-statistics-2010/

Kaspersky Labs. (2011). *Kaspersky security bulletin 2011*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/36344/kaspersky-security-bulletin-statistics-2011/

Kaspersky Labs. (2012). *Kaspersky security bulletin 2012*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/36703/kaspersky-security-bulletin-2012-the-overall-statistics-for-2012/

Kaspersky Labs. (2013). *Kaspersky security bulletin 2013*. Retrieved from http://media.kaspersky.com/pdf/ksb_2013_en.pdf

Kaspersky Labs. (2014). *Kaspersky security bulletin 2014*. Retrieved from https://securelist.com/files/2014/12/Kaspersky-Security-Bulletin-2014-EN.pdf

Kaspersky Labs. (2015). *Kaspersky security bulletin 2015*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/73038/kaspersky-security-bulletin-2015-overall-statistics-for-2015/

Kaspersky Labs. (2016). *Kaspersky security bulletin 2016*. Retrieved from https://securelist.com/analysis/kaspersky-security-bulletin/72771/kaspersky-security-bulletin-2016-predictions/

Kshetri, N. (2006). The simple economics of cybercrimes. *IEEE: Security & Privacy, 4*(1), 33–39.

Laube, S., & Böhme, R. (2016). The economics of mandatory security breach reporting to authorities. *Journal of Cybersecurity, 2*(1), 29–41.

Meeker, M. (2017). Internet trends, 2017—code conference. *Kleiner Perkins*. Retrieved from http://dq756f9pzlyr3.cloudfront.net/file/Internet+Trends+2017+Report.pdf

Morgan, S. (2016). Hackerpocalypse: A cybercrime revelation. *Cybersecurity Ventures*. Retrieved from https://www.herjavecgroup.com/hackerpocalypse-cybercrime-report/

Neumann, J. (2015). Power laws in venture. *Reaction Wheel*. Retrieved from http://reactionwheel.net/2015/06/power-laws-in-venture.html

Ponemon Institute. (2013). *2013 cost of data breach study: Global analysis*. Retrieved from https://www.ponemon.org/local/upload/file/2013%20Report%20GLOBAL%20CODB%20FINAL%205-2.pdf

Ponemon Institute. (2014). *2014 cost of data breach study: Global analysis*. Retrieved from https://www-935.ibm.com/services/multimedia/SEL03027USEN_Poneman_2014_Cost_of_Data_Breach_Study.pdf

Ponemon Institute. (2015). *2015 cost of data breach study: Global analysis*. Retrieved from https://nhlearningsolutions.com/Portals/0/Documents/2015-Cost-of-Data-Breach-Study.PDF

Ponemon Institute. (2016). *2016 cost of data breach study: Global analysis*. Retrieved from https://www.ibm.com/downloads/cas/7VMK5DV6

Ponemon Institute. (2017). *2017 cost of data breach study: Global overview*. Retrieved from https://www.ibm.com/downloads/cas/ZYKLN2E3

Ponemon Institute. (2018). *2018 cost of data breach study: A global overview*. Retrieved from https://www.ibm.com/downloads/cas/AEJYBPWA

Privacy Rights Clearinghouse (2020). Data breaches. Retrieved from https://privacyrights.org/data-breaches

Roberts, D. (2016). Tom Ridge: Cyber attacks are now worse than physical attacks. *Yahoo! Finance*. Retrieved from http://finance.yahoo.com/news/tom-ridge-cybersecurity-attacks-are-now-worse-than-physical-attacks-170426390.html?soc_src=social-sh&soc_trk=tw

Rogers, E. M. (2010), *Diffusion of innovations*. New York, NY: Simon & Schuster.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13*(2). 238–241.

Statista. (2020). Internet of Things (IoT) connected devices installed base worldwide from 2015 to 2025. Retrieved from https://www.statista.com/statistics/471264/iot-number-of-connected-devices-worldwide/

Symantec. (2013). *2013 Internet security threat report*. Retrieved from http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v18_2012_21291018.en-us.pdf

Symantec. (2014). *2014 Internet security threat report*. Retrieved from http://www.symantec.com/content/en/us/enterprise/other_resources/b-istr_main_report_v19_21291018.en-us.pdf

Symantec. (2015). *2015 Internet security threat report*. Retrieved from https://www.symantec.com/content/en/us/enterprise/other_resources/21347933_GA_RPT-internet-security-threat-report-volume-20-2015.pdf

Symantec. (2016). *2016 Internet security threat report*. Retrieved from https://www.symantec.com/security-center/threat-report

Taleb, N. (2010). *The black swam: The impact of the highly improbable*. New York, NY: Random House Trade Paperbacks.

van Eeten, M. J. (2017). Patching security governance: An empirical view of emergent governance mechanisms for cybersecurity. *Digital Policy, Regulation and Governance, 19*(6), 429–448.

Woods, D., Moore, T., & Simpson, A. (2019, June). *The county fair cyber loss distribution: Drawing inferences from insurance prices*. Paper presented at the 18th Workshop on the Economics of Information Security, WEIS 2019, Boston, MA. Retrieved from https://www.researchgate.net/publication/332861796_The_County_Fair_Cyber_Loss_Distribution_Drawing_Inferences_from_Insurance_Prices