Sex, Lies and Cyber-crime Surveys

Dinei Florêncio and Cormac Herley
Microsoft Research
One Microsoft Way
Redmond, WA, USA
{dinei,cormac}@microsoft.com

ABSTRACT

Much of the information we have on cyber-crime losses is derived from surveys. We examine some of the difficulties of forming an accurate estimate by survey. First, losses are extremely concentrated, so that representative sampling of the population does not give representative sampling of the losses. Second, losses are based on unverified self-reported numbers. Not only is it possible for a single outlier to distort the result, we find evidence that most surveys are dominated by a minority of responses in the upper tail (*i.e.*, a majority of the estimate is coming from as few as one or two responses). Finally, the fact that losses are confined to a small segment of the population magnifies the difficulties of refusal rate and small sample sizes. Far from being broadly-based estimates of losses across the population, the cyber-crime estimates that we have appear to be largely the answers of a handful of people extrapolated to the whole population. A single individual who claims \$50,000 losses, in an $N = 1000$ person survey, is all it takes to generate a \$10 billion loss over the population. One unverified claim of \$7,500 in phishing losses translates into \$1.5 billion.

1. INTRODUCTION

In the 1983 Federal Reserve Survey of Consumer Finances an incorrectly recorded answer from a single individual erroneously inflated the estimate of US household wealth by \$1 trillion [10]. This single error added 10% to the total estimate of US household wealth. In the 2006 FTC survey of Identity Theft the answers of two respondents were discarded as being “not identity theft” and “inconsistent with the record.” Inclusion of both answers would have increased the estimate by \$37.3 billion [14]; *i.e.*, made a 3× difference in the total estimate. In surveys of sexual behavior men consistently report having had more female sex partners than women report having had male sex partners (which is impossible). The difference ranges from a factor of 3 to 9. Morris [27] points out that a tiny portion of men who claim, *e.g.*, 100 or 200 lifetime partners account for most of the difference. Removing the outliers all but eliminates the discrepancy.

How can this be? How can an estimate be so brittle that a single transcription error causes a \$1 trillion difference? How can two answers (in a survey of 5000) make a 3× difference in the final result? These cases have in common that the estimates are derived from surveys, that the underlying quantity (*i.e.*, wealth, ID theft losses, or number of sexual partners) is very unevenly distributed across the population, and that a small number of outliers enormously influenced the overall estimate. They also have in common that in each case, inclusion of the outliers, caused an enormous error to the upside, not the downside. It does not appear generally understood that the estimates we have of cyber-crime losses also have these ingredients of catastrophic error, and the measures to safeguard against such bias have been universally ignored.

The common way to estimate unknown quantities in a large population is by survey. For qualities which are evenly distributed throughout the population (such as voting rights) the main task is to achieve a representative sample. For example, if the achieved sample over- or under-represents any age, ethnic or other demographic group the result may not be representative of the population as whole. Political pollsters go to great lengths to achieve a representative sample of likely voters.

With surveys of numeric quantities things are very different. First, some quantities, such as wealth, income, *etc*, are very unevenly distributed across the population. A representative sample of the population (*i.e.*, all people have equal likelihood of being chosen) will give an unrepresentative picture of the wealth. For example, in the US, the top 1% and the bottom 90% of the population each controls about one third of the wealth [25]. A *representative* sample of 1000 people would end up estimating the top third of the wealth from the answers of about ten people, and the bottom third from the answers of about 900 people. Thus, there are two orders of magnitude difference in the sample size for equivalent fractions of the wealth. We have far greater accuracy at the bottom than at the top. Second, for numeric quantities even a single outlier can greatly effect the survey estimate. The survey mean can be affected to an arbitrary extent by a single lie, transcription error or exaggeration. Self-reported numbers are known to have large sources of bias [26] and there is no guarantee that any survey respondent accurately reports the truth. If errors cancel then this error is unbiased (*i.e.*, in expectation neither pulls the estimate up nor down). However, for non-negative quantities (*e.g.*, prices, wealth, cyber-crime losses, number of sex partners *etc*) errors have a lower bound, but no upper bound, so errors do not cancel and the bias is always upward. Finally, there are unique difficulties when surveying rare phenomena. Non-response error can be large, there is significant reduction in effective sample-size and it is difficult to overcome the fact that some fraction of the population routinely lies, exaggerates and misreports. If the phenomenon we wish to survey is rarer than the frequency of liars, our signal is effectively overwhelmed with noise, and no accurate estimate can be formed, at any survey size.

These three sources of error, that a representative sample of the population doesn't give a representative picture of the surveyed quality, that outliers can cause catastrophic errors, and for rare phenomenon we are measuring a signal weaker than the noise in which it is embedded pose a serious threat. In this paper we show that the estimates we have of cyber-crime come from surveys that suffer from all three of these sources of error. Cyber-crime losses follow very concentrated distributions where a representative sample of the population does not necessarily give an accurate estimate of the mean. They are self-reported numbers which have no robustness to any embellishment or exaggeration. They are surveys of rare phenomena where the signal is overwhelmed by the noise of misinformation. In short they produce estimates that cannot be relied upon. The difficulties presented have long been recognized in the areas of Robust Statistics [33] and Survey Science [25]. However safeguards against producing erroneous results seem largely ignored in cyber-crime surveys.

2. SEX AND LIES

We begin with an example which illustrates one of the major sources of error. Surveys of sexual behavior consistently show a large gender discrepancy. Men report having had more female sex partners than women report having had male sex partners. The difference ranges from a factor of 3 to 9 (see Wiederman [34] and references therein). This discrepancy is repeated across many different surveys and countries (*e.g.*, US, Britain, France, New Zealand and Norway). In a closed population with equal numbers of men and women, of course, this is impossible. The average lifetime number of heterosexual partners for men and women is the same.

Thus, the surveys of men and women give independent estimates of the same quantity, yet those estimates are mutually inconsistent. Clearly, there are sources of significant error in one, other or both of the estimates. Further, since men reliably report more partners than women, in surveys performed in different countries at different times and using different methodologies, those errors appear to pull consistently in one direction. This strongly suggests that each of the surveys has the same source of error. There are various possibilities. Selection bias which excludes women who have had many male partners might occur for this difference. Response bias, where women under- and men over-report their number of sexual partners, might also account for this error.

Morris [27] points out that the data has a heavytail distribution and most of the discrepancy is generated by a very small fraction of respondents who report large numbers of partners. Among the 90% of respondents who report having fewer than 20 partners the discrepancy between the reports of men and women all but disappears. This suggests a very simple explanation which accounts for most of the bias. The majority of women tell the truth, but perhaps under-report by a little. The majority of men also tell the truth, but perhaps over-report by a little. However, a small fraction of men tell whoppers: they exaggerate the number of partners they have had, not by a little, but by a lot. A man who claims to have had 100 lifetime sex partners (as about 1% in the dataset that Morris examines do) when the actual number is 50, adds enormous response error. It would take 16 men with the median number of partners understating by 2× to cancel this single 2× overstatement. Thus there is great asymmetry in the response error.

What has this to do with cyber-crime? Cyber-crime, like sexual behavior, defies large-scale direct observation and the estimates we have of it are derived almost exclusively from surveys. The sexual partner surveys are unique in that, while we don't know the correct answer, we have a cross-check (*i.e.*, the result from the women) that shows that the estimate procedure is producing inaccurate answers. These surveys serve to illustrate two of the problems that are present also in cyber-crime surveys: the difficulty of achieving a representative sample of heavytail distributions, and the difficulty of telling representative outliers, which should be included, from unrepresentative ones (*e.g.*, lies and exaggerations) which should not. A third difficulty, that of surveying very rare phenomenon amplifies both of these difficulties.

2.1 Sources of Error in Survey Research

When we wish to estimate any numerical quantity, x, over a large population we select some portion of

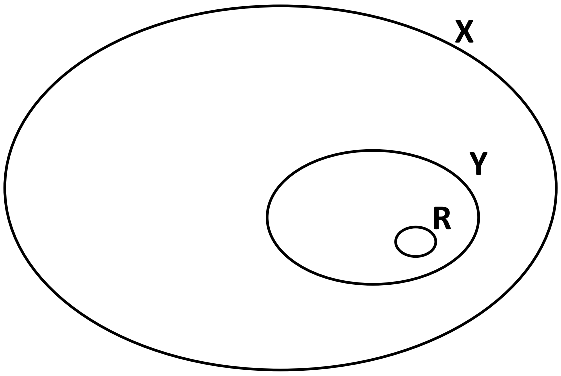

Figure 1: Venn Diagram. X is the whole population, Y is the contacted population, and R is the achieved sample.

the population. Call X the whole population and Y the contacted population (*i.e.*, the set of people who are asked to respond). However, some of those who are contacted refuse, so we end up with a smaller responding population R. Clearly, $R \subset Y \subset X$. We call the averages over these populations $\overline{x}, \overline{y}$, and $\overline{r}$. When surveying a self-reported numerical quantity (such as salary, or hours of exercise per week) the observed answer is not necessarily the the true answer. Thus, the true mean of those who respond is $\overline{r}$ but we observe $\overline{f[r]}$.

If the goal of the survey is to estimate $\overline{x}$, the mean of X, the survey error is $\overline{x} - \overline{f[r]}$. This error can be broken down into sampling error and non-sampling error [9]. The sampling error is $\overline{x} - \overline{y}$, or the difference between the mean of the whole population and that of the contacted population. The non-sampling error, $\overline{y} - \overline{f[r]}$, reflects the difference between the mean of the contacted population and that observed of the responding population. This in turn is generally split into non-response error and response error. Non-response error, $\overline{y} - \overline{r}$, is the difference between the mean of the contacted population and that of the responding population. Finally, response error, $\overline{r} - \overline{f[r]}$, is the difference between the true mean of the responding population and the observed mean. The total survey error is then [9]:

$$\overline{x} - \overline{f[r]} = (\overline{x} - \overline{y}) + (\overline{y} - \overline{r}) + (\overline{r} - \overline{f[r]}).$$

Non-response error, $(\overline{y} - \overline{r})$, is known to be particularly important where the refusal rate is high (*i.e.*, the number of people in R is small relative to the number in Y). This has long been known in the crime survey literature. If the refusal rate is high there is a possibility that victims respond at a much higher or lower rate than the rest of the population which causes over- or under-estimation. For example if 10% of non-victims, and 50% of victims respond then R contains 5× as many victims as Y. We examine this in Section 3.3.1.

Response error, $(\overline{r} - \overline{f[r]})$, is especially problematic when dealing with self-reported numbers. When there is no ability to verify the reported answers then there is no protection against lying or mis-statement, and the potential error can dwarf sampling error. We examine the role that this plays in Section 3.2. Sampling error is examined next.

3. LIES AND CYBER-CRIME

3.1 The survey mean need not approximate the true mean, even when the survey is representative

3.1.1 Heavytail distributions

Many qualities are very unevenly distributed across the population. Some of them, such as height, weight, *etc*, are well-approximated by the familiar bell-curve, or normal, distribution. Of course, non-negative quantities such as height cannot precisely follow a normal distribution as the distribution has tails that extend infinitely in both directions: neither negative nor infinite heights are admissible. Heights nonetheless follow a normal pattern fairly closely. In particular, heights are more or less symmetrically distributed about the mean.

For some qualities the distribution is much more uneven. For height, even a factor of two difference is extreme. Wealth, income, fame *etc*, by contrast, are obvious examples where the quality is heavily concentrated among a small fraction of the population. A small number of people have a great deal (*e.g.*, wealth or fame) and most have very little or none. These qualities are much better captured by heavytail distributions such as Pareto or Log-normal. Heavytail distributions have infinite tails that contain a large fraction of the probability mass. Because of the large mass in the tail the mean is generally much higher than the median. These are also know as distributions with positive skew.

The Pareto is a family of concentrated distributions, containing for example the well-known 80/20 distribution, which indicates that 80% of the phenomenon is concentrated among 20% of the samples. It is used, for example, to model the wealth distribution of households in the US [25, 12]. In the Pareto distribution the probability of a randomly chosen individual having amount x is:

$$p(x) = Cx^{-\alpha}, \text{ for } \alpha > 2.$$

The fraction of the phenomenon accounted for by the top fraction P of the population is

$$W = P^{(\alpha-2)/(\alpha-1)}. \quad (1)$$

Observe that as $\alpha \rightarrow 2$, an arbitrarily small fraction P will control and arbitrarily large fraction W of the wealth. That is, $W \rightarrow 1$: more and more of the phenomenon will be held by a small fraction P of the population. For US wealth $\alpha \approx 2.32$. We now show that as the concentration increases even representative samples of the population will fail to give a representative picture of its statistics.

3.1.2 Representative sampling gives an unrepresentative estimate

Quantities that are unevenly distributed across the population are harder to survey than those that are evenly distributed. For a uniform distribution, every individual has an equally important contribution to make to the survey. Concentrated distributions are at the other extreme: a representative sample of the population gives a very unrepresentative picture of the quantity of interest. If we uniformly sample the population we end up with many samples from the part of the population that has little or no wealth, and very few samples from the part that has most of the wealth. Figure 2 shows the distribution of wealth among households in the US. The top 1% control approximately 33% of the wealth. In a sample of 1000 where all households respond with equal likelihood we'll end up estimating one third of the wealth from the answers of ten households. If the average of those ten is not the true average of the upper 1% we end up with a misleading estimate.

The problem does not end there. The third that is held by the top 1% is just as unevenly distributed as the overall wealth [29]. Approximately a third of one third is held by the top 1% of 1%. That is 0.01% of the population holds 11% of the wealth. Table 1 summarizes the wealth concentration in the upper tail for the Pareto that closely models US wealth [25]. As can be seen, the concentration continues at different scales.

In fact, a representative sample of the population does not guarantee that the sample mean approximates well the true mean. That is, when things are very skewed we have $\overline{r} \not\approx \overline{x}$. This is so, since it is hard to achieve a representative sample with very few samples. And when a large portion of the wealth is concentrated among few hands the sample-size in that fraction of the wealth is tiny. Table 1 shows that for US wealth an $N = 1000$ survey should expect ten and one respondents respectively for the top 33% and 19% of the wealth. Further, there is only a one in ten, and one in a hundred chance respectively of having a respondent from the top 11% and 6% of the wealth.

It is not possible to get a representative picture of that portion of the wealth with minuscule sample-sizes.

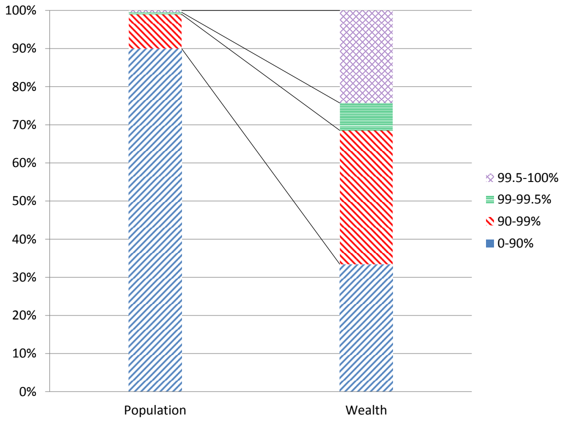

Figure 2: Fraction of the wealth controlled by segments of the population [25]. The top 1% and bottom 90% each control about one third of the wealth. A survey that is representative of the population will be very unrepresentative of the wealth (having 90× fewer samples for the top third than the bottom).

While we can gather survey responses and average them, this can fail to give a representative picture of the wealth. If we repeat the trial we can get a very different answer. Figure 3 shows 100 trials of the sample mean of 1000 representative samples of a Pareto ($\alpha = 2.32$, *i.e.*, US wealth) distribution. As can be seen the sample mean varies considerably from the true mean (which is 4.125). This picture is simply for the Pareto that approximates US wealth distribution. If the concentration increases (*i.e.*, $\alpha \rightarrow 2$) or the sample-size decreases the variations become more extreme. We will see that both of these conditions apply in the case of cyber-crime surveys (Sections 3.1.3 and 3.3.1). The great variability is simply an artifact of the unreliability of the sample mean. As Newman writes [29]: "while we can quote a figure for the average of the samples we measure, that figure is not a reliable guide to the typical size of the samples in another instance of the same experiment."

The concentration is a problem for two main reasons. First, since so much of the phenomenon is in the tail it is difficult to adequately sample it unless a truly enormous survey is conducted. Second, the estimate is extremely brittle. An inordinate fraction of the estimate is coming from the answers of a handful of respondents. If those respondents are not representative, mis-remember, exaggerate, or entirely invent their answers the effect on the overall estimate is catastrophic. As, the 1983 Consumer Finances [10] and 2006 ID Theft [14] surveys show, an error or two can cause enormous increase. Expressed differently, since the majority of the estimate comes from a handful of people, great faith is being placed in their answers. The estimate is reliable to the degree that their answers are both representative and reliable.

The extreme difficulty of surveying heavytail phenomena has long been recognized. In the US the Survey of Consumer Finances a multi-layer sampling approach is used [25]. A first sample of 3824 households were selected with equal probability, which gave a broad overview of wealth and finances in the overall population. A second sample of 438 households from two higher strata was conducted (the median net worth of households in these two strata were $50 million and $300 million). This allows formation of a far more accurate picture of the upper tail of the wealth distribution than is possible from a uniform sample. Considerable effort was taken to keep the refusal rate among those in the upper strata low (not surprisingly wealthy individuals have a far higher refusal rate than the population average).

3.1.3 *Concentration in cyber-crime surveys*

Concentration in cyber-crime surveys is not merely a possibility. In fact those surveys that give enough information make clear that the distribution of losses is enormously concentrated, with a small fraction of respondents accounting for the bulk of the losses. For example, the Gartner 2007 phishing survey finds a median loss of $200, but a mean of $857. A factor 4.5× difference between mean and median is indicative of greater concentration than even the US wealth distribution. A Pareto distribution with this skew concentrates 59% of the wealth in the hands of the top 1%.

The FTC in 2006 report [14] great differences between mean and median, both of money and time lost, and the value the thief obtained. Even with the exclusion of the two outliers mentioned in the introduction the survey found a mean loss of $1876 and median of $500, which is roughly comparable to the degree of concentration of US wealth. “The median value for the number of hours spent resolving problems by all victims was 4. However, 10 percent of all victims spent at least 55 hours resolving their problems. The top 5 percent of victims spent at least 130 hours.”

The IC3 survey [5] finds a 9.7× ratio of mean/median: “Of those complaints reporting monetary loss that were referred to law enforcement, the mean dollar loss was $5,580 and the median was $575. The significant difference between the mean and median losses is reflected by a small number of cases in which hundreds of thousands of dollars were reported to have been lost by the complainant.” This is simply an eye-popping level of concentration, indicating that almost all the losses were endured by a tiny number of complainants. In a Pareto distribution with this level of skew the top 1% controls 78% of the wealth.

Top Fraction of population	Percent of wealth ($\alpha = 2.32$)
1%	32.7%
0.1%	18.7%
0.01%	10.7%
0.001%	6.1%

Table 1: Concentration of Pareto distribution that approximates US wealth.

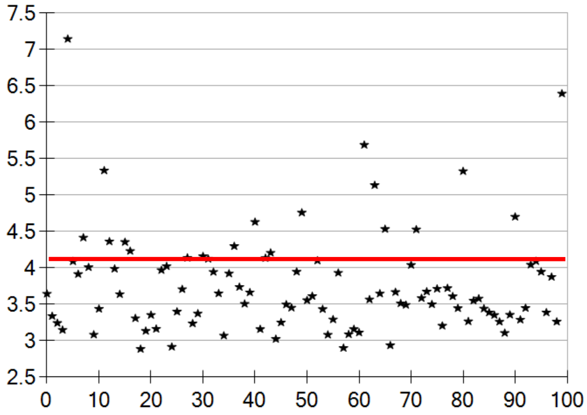

Figure 3: Instances of sample mean for a Pareto ($\alpha = 2.32$, *i.e.*, US wealth) distribution. There are 100 trials each with $N = 1000$ samples. Even though there is no measurement error and the sampling is representative the sample mean shows considerable variance. This problem gets worse as concentration increases or sample-size decreases.

The US Bureau of Justice Statistics produce bi-annual reports of Identity Theft [11]. The mean/median ratio varies across different theft categories, with 10× being typical. In some categories the ratio of the mean out-of-pocket loss to the median is as high as 14, indicating that almost the entire reported loss for the survey is derived from the answer of a single respondent.

Unfortunately, the majority of cyber-crime surveys give only the mean, $\overline{x}$, or total estimate $|X| \cdot \overline{x}$. While they refer to the concentration of the losses, failure to provide the median makes it impossible to do further analysis.

3.2 The survey mean need not approximate the true mean when there is measurement error

The average response of the responding population **R** is:

$$\overline{f[r]} = \frac{1}{|R|}\sum_{i \in R} f[r_i]. \qquad (2)$$

If $\overline{f[r]} \approx \overline{r}$ and $\overline{r} \approx \overline{x}$ then we can approximate the mean response of the responding population for that of x over the overall population.

As we saw above, for heavytail distributions, we can't assume that $\overline{r} \approx \overline{x}$. This, it turns out is only the beginning of our difficulties. Neither can we assume that $\overline{f[r]} \approx \overline{r}$. Unfortunately the sample mean is an extremely non-robust quantity: a single outlier is sufficient to distort the value.

There are various reasons that can produce measurement error (*i.e.*, $f[r_i] \neq r_i$). Transcription error was the cause of a massive error in the 1983 consumer finance survey. It's possible that respondents mis-remember, or misunderstand the survey question. And, of course, not everyone tells the truth. The discrepancy in sexual partner reports emphasizes the lack of robustness when forming estimates of self-reported numbers. Morris' work shows that even when the majority of self-reports are accurate, the sample mean can be wildly inaccurate. A respondent who lies (*i.e.*, $f[r_i] \neq r_i$) affects the average by $(f[r_i] - r_i)/|R|$. Since answers must be positive, the erroneous contribution to the mean is bounded below by $-r_i/|R|$, but is unbounded above. The larger $(f[r_i] - r_i)/|R|$ the larger the response error introduced. For example, if an individual has wealth r_i but it is incorrectly recorded as $f[r_i] = 10r_i$ no other individual understatement cancels this error. We needn't stop there, with self-reported numbers exaggerations by $100\times, 1000\times$ or more are not merely feasible, but have been observed. Recall that the FTC 2006 survey excludes answers from two respondents who appear to be relating fictitious losses which (if included) would have added \$37.3 billion to the estimate. Since \$10k in a survey of $N = 1000$ people translate into \$2 billion when applied to a population of 200 million (see Section 4.1) the estimates are extremely fragile.

The extreme concentration of distributions of wealth (and cyber-crime losses) raises the stakes considerably. Since so much of the phenomenon is concentrated among a small fraction of respondents the accuracy of the estimate depends on the accuracy of their answers. Indeed, when concentration is high enough, most of the estimate is coming from a tiny fraction of the responding population. Just how much is coming from the highest reporting respondents is tabulated in Table 2. This shows the factor difference made to the entire estimate by the fraction P with the highest reports. That is, for example, how much higher the estimate is for inclusion of the top 1% as opposed to an estimate based solely on the other 99%. When $\alpha = 2.05$ for example (the concentration found in the IC3 survey [5]) the top 1% increased the estimate by $5.1\times$. Here we tabulate $1/(1-W)$, where W is as defined in (1). For increasing concentration a very small fraction of the population has an outsized influence. For example, when the survey size is small, and the phenomenon is rare a single respondent can be 5% of the response pool (and thus account for a $1.9, 3.4$ or $7.5\times$ increase).

Percent	$\alpha = 2.32$	$\alpha = 2.13$	$\alpha = 2.05$
1%	1.5×	2.4×	5.1×
5%	1.9×	3.4×	7.5×
10%	2.4×	4.3×	9.6×

Table 2: Factor difference that misrepresentation by a small fraction of respondents can make. For $\alpha = 2.32$, approximating the concentration of US wealth, invented numbers from the top 5% result in a $1.9\times$ increase in the overall estimate.

Of course whether 1% of the survey is 100 people, ten, one or (in expectation) less than one depends on the sample-size. We'll see in Section 3.3.2 how 1% of the sample-size, on which 80% or so of the estimate is depending can be as little as one person.

The essential problem we face, that, for non-negative quantities, the sample mean can be increased (but not decreased) by an arbitrary amount by the answer of a single respondent has long been known in Robust Statistics. In the best of circumstances (measurement error is rare, the phenomenon is evenly distributed and errors cancel) Tukey writes [33]: "If contamination is a real possibility (and when is it not?), neither the mean nor variance is likely to be a wisely chosen basis for making estimates from a large sample." However cyber-crime surveys are far from the best of circumstances. Sampling of heavytail distributions is far less robust than the normal distributions of which Tukey was writing.

Further evidence of the upward rather than downward bias of sample mean is found in a recent examination of the wisdom of the crowd effect by Lorenz *et al.* [20]. They find that the median gave a more accurate measure than the arithmetic mean of answers from a crowd. Of the six phenomena surveyed, the mean of the crowd answers always over-estimated, by an amount ranging from 59% to 1365%.

3.2.1 Self-reported numbers

If we had no measurement or reporting errors (*i.e.*, we always have $f[r_i] = r_i$) things would be relatively simple. We would then merely have sampling error, $(\overline{x} - \overline{y})$, and non-response error, $(\overline{y} - \overline{r})$, to contend with. However, self-reported numbers are known to be generally inaccurate.

Self-reported numbers on calorie consumption and exercise are known to generally err on the optimistic side. In a weight loss survey [26]: "subjects under-reported their actual food intake by an average ($\pm$ SD) of 47 ± 16 percent and over-reported their physical activity by an average of 51 ± 75 percent."

The problem is twofold. First, that we have no ability to check the accuracy of any of the responses offered. Second, in concentrated phenomena most of the effect is reported by a handful of people. If the answers of those at the top are exaggerated or inaccurate we produce wildly inaccurate answers. There are numerous reasons why people may report inaccurately. In several cyber-crime surveys [14, 16] it appears the total estimate was based on how much respondents believe the thief obtained (rather than how much the victim lost). For example the median answer for the former was \$500 but the latter was \$0 in the FTC 2006 survey. Since respondents are being asked something of which they have no direct knowledge, over-estimation is highly likely. Vague and unclear categories may encourage respondents to "throw in" experiences that were not part of the survey intent. For example, an unsatisfactory online auction experience or dispute with a merchant might easily be conflated with "online fraud." The FTC survey which finds an individual respondent trying to report a claimed loss of \$999999 "theft of intellectual property" as ID theft is just such an example. Victims may be angry, and when offered an opportunity to complain be tempted to over-state rather than under-state their true losses. Finally, some percent of the population just lies and make things up.

3.3 Surveying Rare Phenomena

We've seen some of the difficulties of surveying unevenly distributed phenomena such as wealth. There is one further complication that makes accurate estimation of cyber-crime losses even harder: surveying rare phenomena is hard. Wealth and income may be unevenly distributed, but most of the population is involved and most responses can be used (although some answers are many times more useful than others in forming the overall estimate). If 1000 people respond to a survey on wealth the answers of all of them will be useful in forming an estimate. For rare phenomena this isn't the case. For a phenomenon that affects 5% of people, 95% of the population will have nothing useful to say: their answers contribute nothing to the estimate. This complicates things in three respects. First, non-response bias can be high. When the phenomenon is rare there is a real risk that those who are affected respond at a much higher or lower rate than the overall population. Second, there is a raw reduction of sample-size. Third, some fraction of the population routinely lies and fabricates answers. This can cause our signal to be lost in the noise.

3.3.1 Achieving a representative sample

Suppose a small fraction of the population, X, are affected by phenomenon V. That is $|V|/|X|$ is small. Let's call the members of V victims, and all others non-victims. In doing a survey it is of paramount importance that the percent of victims in the responding population, R, be similar to that in X. It is not hard to imagine that people affected by phenomenon V may respond at a higher or lower rate than the rest of the population. Gamblers may be more likely than non-gamblers to respond to a survey on gambling, for example. People who have been victimized by a certain type of crime may be significantly more likely (or less) to respond to a survey on that crime.

The victimization rate is $V/(V+N)$. But if only a fraction V_r and N_r of victims and non-victims respectively respond we estimate the rate as

$$\frac{V \cdot V_r}{V \cdot V_r + N \cdot N_r}.$$

When the overall victimization rate is low (*i.e.* $V \ll N$ so that $V \cdot (V_r/N_r) + N \approx V + N \approx N$) we get [17]:

$$\frac{V \cdot V_r}{V \cdot V_r + N \cdot N_r} \approx \frac{V}{V+N} \cdot \frac{V_r}{N_r}.$$

Thus, our estimate of the victimization rate is the true rate, multiplied by V_r/N_r. Any difference in the victim and non-victim response rates enormously influences the estimate. So, if $V_r = 5N_r$ (victims are 5× more likely to respond) then the estimated victimization rate is about 5× the true rate. Exactly such a bias appears to occur in the Gartner 2007 phishing survey which estimates the victimization rate a full factor of ten higher than the non-survey estimates of Florêncio and Herley [17], Clayton and Moore [32] and Trustseer [6].

3.3.2 Sample-size reduction

A further difficulty comes from the sheer reduction in effective sample size that surveying a rare phenomenon brings. If a phenomenon affects 5% of the population then in a representative sample of 1000 people we expect only 50 answers that are of interest.

In Sections 3.1 we saw the difficulty of surveying quantities that are unevenly distributed. It is almost impossible to avoid under-sampling the tail in a concentrated distribution. In addition we now find that rare phenomena are hard to survey, as most of the responses are wasted and cannot contribute to the estimate. However, cyber-crime losses suffer from both these problems: they are rare phenomena that are also extremely concentrated. That is, only a few percent of people suffer from ID theft. Even among those that do suffer from it the losses are extremely concentrated as we saw in Section 3.1.3. Thus cyber-crime losses are both confined to a small segment of the population, but also, have very uneven distribution within that segment. The rareness gives a reduction in the sample size. The concentration adds to the fragility of the sample.

To be concrete, consider a $N = 1000$ survey of a phenomenon that affects 2% of the population. Our effective sample-size is now 20, not 1000. A single individual counts for 5% of the response pool. Further suppose that the phenomenon is concentrated to the same degree as US wealth (*i.e.*, Pareto with $\alpha = 2.32$). In this case 48% of the phenomenon is concentrated in the top 5%. Thus, we expect that fully one half of our estimate will be coming from a single individual.

Let's examine examples from actual surveys. The FTC 2006 survey [14] reached 4917 respondents and found 3.7%, 1.4% and 0.8% rates of all ID theft, misuse of existing accounts, and misuse of new accounts respectively. However, these appear to correspond to sample sizes of 181, 68 and 39 respectively. Thus, for new account fraud the top 1% of respondents is less than one person. From Table 2, if these losses are as concentrated as US wealth, the top 5% (*i.e.*, approximately 2 people) double the entire estimate.

As we move on from the FTC survey things only get worse. Gartner's 2006 survey [16] found a 3.2% phishing victimization rate. In a survey of 4000 people this means approximately 128 claimed to be victims (recall we argue in Section 3.3.1 above that they over-estimate the true victimization rate by 10×). Thus the top 1% (which at the concentration level that Gartner finds accounts for 59% of losses) is about one person. Javelin in a survey of 4000 [22] finds 4.25% have been ID theft victims and 1.7% of those have been phishing victims. This gives an effective sample size of three individuals!

3.3.3 Liars

Finally, in surveying rare phenomena it is hard to avoid the subject of liars [18]. There can be little doubt that some fraction of the population embellish, exaggerate and tell whoppers, even when there is no clear motive for doing so. We examined the difficulty that outliers present in Section 3.2. There, however, we tackled the general problem, where people report $f[r_i] = 10r_i$ or so (*i.e.*, multiply their real wealth or number of sexual partners by 10). If there are a percent or two of liars in the population, they affect the estimate modestly unless any of them are outliers in the tail.

However, when surveying rare phenomena most of the population are unaffected, that is they have nothing to report. If the phenomenon affects 1% of the population and 1% of people are habitual liars then our survey can have up to 50% contributions from people who are offering pure invention by way of answers.

4. DISCUSSION

4.1 Total estimate

We've seen that, when estimating $\overline{x}$, the survey error, $\overline{x} - \overline{f[r]}$, can be enormous. Often, however, it is the total, rather than the mean of X that we wish to estimate. That is we want $|X| \cdot \overline{x}$ rather than $\overline{x}$. This is the case, for example, in estimating total US household wealth [12], and losses in all cyber-crime surveys. Now, the response errors are amplified by backing into the overall population. The estimate becomes $|X| \cdot \overline{f[r]}$. Thus, from (2), each respondent adds $|X|/|R| \cdot f[r_i]$ to the estimate. For example, if the population size is $|X| = 200$ million and the survey size is $|R| = 1000$ then each dollar of losses claimed is multiplied by $|X|/|R| = 200,000$. In other words every dollar of claimed losses translates into \$200,000 in the estimate. A respondent who claims \$50,000 in ID theft losses adds \$10 billion to the overall loss estimate. Indeed five individuals, each of whom claim \$50,000 is all that is required to generate a \$50 billion loss estimate. Similarly, a single respondent who claims to have lost \$7,500 to phishing is all it takes to generate \$1.5 billion in estimated population-wide losses. Two such individuals is all it takes to give a loss estimate in the \$3 billion range.

4.2 Lack of Consistency

The variability of cyber-crime surveys is not merely theoretical. The FTC estimated Identity theft at \$47 billion in 2004 [13], \$15.6 billion in 2006 [14] and \$54 billion in 2008 [23]. Either there was a precipitous drop in 2006, or all of the estimates are extremely noisy.

The vagueness and lack of clarity about what has been measured allows for a large range of interpretation. In the last two years alone we find the following claims, which value cyber-crime at anywhere from \$560 million to \$1 trillion. "The spoils of cyber crime almost doubled in 2009. As a whole, losses totaled \$560m," Patrick Peterson, Cisco Fellow [1]. "Cyber crime costs corporate America \$10 billion every year!" [2]. "Damage caused by cyber-crime is estimated at \$100 billion annually," said Kilian Strauss, of the Organization for Security and Cooperation in Europe (OSCE) [3]. "Cyber-crime revenues are worth approximately \$1 trillion," Edward Amoroso, CSO, AT&T (written testimony to the US Senate Commerce, Science, and Transportation Committee, March 17, 2009).

4.3 Other Analyses of Cyber-crime Surveys

Our assessment of the quality of cyber-crime surveys is harsh: they are so compromised and biased that no faith whatever can be placed in their findings. We are not alone in this judgement. Most research teams who have looked at the survey data on cyber-crime have reached similarly negative conclusions. Ryan and Jefferson [21], who perform a meta-study of fourteen cyber-crime surveys, write "In the information security arena, there is no reliable data upon which to base decisions. Unfortunately, there is unreliable data that is masquerading as reliable data." Anderson *et al.*[30] find "there has long been a shortage of hard data about information security failures, as many of the available statistics are not only poor but are collected by parties such as security vendors or law enforcement agencies that have a vested interest in under- or over-reporting." Moitra produces a survey of various cyber-crime surveys [31]. He observes that "a lack of reliability and validity checks on the data that have been collected" and singles out exaggeration of losses, and self-selection bias as major sources of error not accounted for in the methodology. Brenner, in arguing that accurate measures and estimates for the incidence of computer-related crime are necessary writes: "We have never done this, even though the term 'cybercrime' and its various correlates [...] have been in use for decades." Herley and Florêncio [17] say that the cyber-crime survey estimates they examine "crumble upon inspection." Shostack and Stewart [7] write "today's security surveys have too many flaws to be useful as sources of evidence." The lack of faith in existing surveys is not limited to research teams. At the keynote at Workshop on Economics of Information Security (WEIS) 2010 Tracey Vispoli, VP and head of CyberSecurity Infrastructure at Chubb Insurance stated that [4] the insurance industry has "no expected loss data and no financial impact data."

4.4 Recommendations

What general conclusions can we draw from this? Survey science is hard. Mistakes can be made even when every care is taken (as the $1 trillion mistake in the Consumer Finance survey shows). The very term "survey" creates the impression of a broadly-based study which gives a representative snapshot of what is going on. When we deal with simple evenly distributed quantities, such voting intentions, this is the case. When we deal with concentrated phenomena, such as wealth, it is very far from the case. Extreme care (such as multi-layer sampling [25]) is required for concentrated phenomena. When we deal with phenomena that are both confined to a small segment, and concentrated within that segment all of the difficulties are amplified.

How may we recognize the danger signs in a survey? First, no weight can be given to surveys that fail to disclose methodology. The risks of catastrophic error are great even when things are done with care. Ensuring that the sample is representative, that concentration is not too great, that the upper tail has been adequately sampled and that outliers have been checked for gross error or fabrication: these are not matters on which benefit of the doubt can be extended. Second, evidence of the degree of concentration is important. The ratio of the mean to the median is a simple figure of merit for the concentration. For US wealth this number is about 4.12. At this level of concentration multi-layer sampling is essential. Ratios higher than this imply the need for infeasibly large sample-sizes. For example, the 2008 US Department of Justice ID theft survey [11] had a sample size of 56,480. ID theft is largely dominated by low-tech means (e.g. a credit card run twice, stolen wallet, etc.), and affects a rather large fraction of the population (*i.e.*, up to 5%). The survey also indicates approximately 0.2% (i.e., 4% of the 5% ID theft victims) responded to a phishing e-mail or phone call. Thus, to achieve an estimate of phishing comparable in accuracy to the estimate of credit-card fraud would require a 25× larger sample size (*i.e.*, over 1 million people). If losses from cyber-crime are more concentrated than those from credit-card fraud then surveys of several million people would be required.

Estimates which fail to disclose the median as well as the mean, or which fail to give some measure of concentration, can be discarded. The reliability of the survey is inversely related to the concentration. Failure to declare concentration is as serious a failing as failure to state the sample size. In fact, as the concentration (*i.e.*, the ratio of mean to median) increases the sample mean is not stable [29]: "while we can quote a figure for the average of the samples we measure, the figure is not a reliable guide to the typical size of the samples from another instance of the same experiment."

5. RELATED WORK

Despite their ubiquity analyses of cyber-crime surveys have been relatively few. Andreas and Greenhill [8] examine the effect that bad estimates can have on policy and resource allocation. Ryan and Jefferson [21], perform a meta-study of fourteen cyber-crime surveys and are largely unimpressed with the methodologies. Moitra produces a survey of various cyber-crime surveys [31]. He observes that "a lack of reliability and validity checks on the data that have been collected" and singles out exaggeration of losses, and self-selection bias as major sources of error not accounted for in the methodology. Herley and Florêncio [17] provide an extensive study of various phishing and ID theft surveys and conclude that all are considerable over-estimates.

The field of Robust Statistics has long studied the problem of estimating distributions from samples. Tukey was among the first to examine the difficulties of measurement (or response) error [33]. Morris [27] appears to have been the first to draw attention to the potential for extreme error when dealing with heavytail distributions and self-reported numbers. A series of papers by Kennilick and co-workers [24, 25] address the difficulties of estimating concentrated distributions from samples.

6. CONCLUSION

The importance of input validation has long been recognized in security. Code injection and buffer overflow attacks account for an enormous range of vulnerabilities. "You should never trust user input" says one standard text on writing secure code [19]. It is ironic then that our cyber-crime survey estimates rely almost exclusively on unverified user input. A practice that is regarded as unacceptable in writing code is ubiquitous in forming the estimates that drive policy (see, *e.g.*, [28]). A single exaggerated answer adds spurious billions to an estimate, just as a buffer overflow can allow arbitrary code to execute. This isn't merely a possibility. The surveys that we have exhibit exactly this pattern of enormous, unverified outliers dominating the rest of the data. While we can sum user responses, and divide to get an average, the resulting calculation is not worthy of the term "estimate" unless we can have confidence that it reflects the underlying phenomenon. For the cyber-crime surveys that we have, statistical difficulties are routinely ignored and we can have no such confidence. Are we really producing cyber-crime estimates where 75% of the estimate comes from the unverified self-reported answers of one or two people? Unfortunately, it appears so. Can any faith whatever be placed in the surveys we have? No, it appears not.

7. REFERENCES

[1] \url{https://vishnu.fhcrc.org/security-seminar/IT-Security-Landscape-Morphs.pdf}.

[2] \url{http://www.ssg-inc.net/cyber_crime/cyber_crime.html}.

[3] http://www.newscientist.com/article/dn16092-cybercrime-toll-threatens-new-financial-crisis.html.

[4] http://taosecurity.blogspot.com/2010/07/brief-thoughts-on-weis-2010.html.

[5] Internet crime complaint center.

[6] Measuring the Effectiveness of In-the-Wild Phishing Attacks. 2009. http://www.trusteer.com/sites/default/files/Phishing-Statistics-Dec-2009-FIN.pdf.

[7] A. Shostack and A. Stewart. The New School of Information Security Research. 2008.

[8] P. Andreas and K. Greenhill. *Sex, Drugs, and Body Counts: The Politics of Numbers in Global Crime and Conflict*. Cornell Univ Pr, 2010.

[9] H. Assael and J. Keon. Nonsampling vs. sampling errors in survey research. 1982.

[10] R. Avery, G. Elliehausen, and A. Kennickell. Measuring wealth with survey data: An evaluation of the 1983 survey of consumer finances. *Review of Income and Wealth*, 34(4):339–369, 1988.

[11] Bureau of Justice Statistics. Victims of Identity Theft. http://bjs.ojp.usdoj.gov/content/pub/pdf/vit08.pdf.

[12] Federal Reserve Board. Survey of Consumer Finances. http://www.federalreserve.gov/pubs/oss/oss2/scfindex.html.

[13] Federal Trade Commission. Identity Theft Survey Report. 2003. http://www.ftc.gov/os/2003/09/synovatereport.pdf.

[14] Federal Trade Commission. Identity Theft Survey Report. 2007. www.ftc.gov/os/2007/11/SynovateFinalReportIDTheft2006.pdf.

[15] D. Florêncio and C. Herley. Where Do Security Policies Come From? *SOUPS 2010, Redmond.*

[16] Gartner. Phishing Survey. 2007. http://www.gartner.com/it/page.jsp?id=565125.

[17] C. Herley and D. Florêncio. A Profitless Endeavor: Phishing as Tragedy of the Commons. *NSPW 2008, Lake Tahoe, CA.*

[18] C. Herley and D. Florêncio. Nobody Sells Gold for the Price of Silver: Dishonesty, Uncertainty and the Underground Economy. *WEIS 2009, London.*

[19] M. Howard, D. LeBlanc, and I. Books24x7. *Writing secure code*, volume 2. Microsoft press, 2003.

[20] J. Lorenz and H. Rauhut and F. Schweitzer and D. Helbing. How social influence can undermine the wisdom of crowd effect. *Proceedings of the National Academy of Sciences*, 108(22):9020, 2011.

[21] J. Ryan and T. I. Jefferson. The Use, Misuse, and Abuse of Statistics in Information Security Research. *Proc. 23rd ASEM National Conference*, 2003.

[22] Javelin. Identity Theft Survey Report. 2003. http://www.javelinstrategy.com/uploads/505.RF_Phishing.pdf.

[23] Javelin. Identity Theft Survey Report. 2009. http://www.javelinstrategy.com/uploads/505.RF_Phishing.pdf.

[24] A. Kennickell. Multiple imputation in the Survey of Consumer Finances. In *Proceedings of the Section on Business and Economic Statistics, 1998 Annual Meetings of the American Statistical Association, Dallas, Texas.* Citeseer, 1998.

[25] A. Kennickell. Getting to the Top: Reaching Wealthy Respondents in the SCF. *Washington, DC: Federal Reserve Board of Governors*, 2009.

[26] S. Lichtman, K. Pisarska, E. Berman, M. Pestone, H. Dowling, E. Offenbacher, H. Weisel, S. Heshka, D. Matthews, and S. Heymsfield. Discrepancy between self-reported and actual caloric intake and exercise in obese subjects. *New England*

Journal of Medicine, 327(27):1893–1898, 1992.

[27] M. Morris. Telling tails explain the discrepancy in sexual partner reports. *Nature*, 1993.

[28] National Strategy for Trusted Identities in Cyberspace. Why We Need It. `http://www.nist.gov/nstic/NSTIC-Why-We-Need-It.pdf`.

[29] M. Newman. Power laws, Pareto distributions and Zipf's law. *Contemporary physics*, 46(5):323–351, 2005.

[30] R. Anderson and R. Boehme and R. Clayton and T. Moore. Security Economics and the Internal Market. *Report for European Network and Information Security Agency*, 2007.

[31] S.D. Moitra. Cyber Security Violations against Businesses: A Re-assessment of Survey Data. `http://www.iimcal.ac.in/res/upd%5CWPS%20571.pdf`.

[32] T. Moore and R. Clayton. Examining the Impact of Website Take-down on Phishing. *Proc. APWG eCrime Summit*, 2007.

[33] J. Tukey. A survey of sampling from contaminated distributions. *I. Olkin*, 1960.

[34] M. Wiederman. The truth must be in here somewhere: Examining the gender discrepancy in self-reported lifetime number of sex partners. *Journal of Sex Research*, 34(4):375–386, 1997.