Crowd-Sourced Probability Estimates: A Field Guide Article

Recently I had the pleasure of presenting at the Society of Information Risk Analysts Conference at the Starbucks Corporation Headquarters in Seattle, Washington. It was quite an exciting conference, and I would like to thank all of you who were able to attend. It was certainly a delight to interact with all of you. For those of you unable to attend, I gave a presentation on a field guide for Probability Estimates.

I started off with giving a hypothetical example of a man named Bob, who works for a mid-sized San Francisco-based FinTech company. In this example, Bob works as a risk analyst for his firms information security department, and is trained in quantitative risk analysis. Bob's Board of Directors are very concerned with the risk of Ransomware and the negative effects an infection would have on the company. Several members of the Board recently attended a cybersecurity conference and learned about the "Ransomware Epidemic", and as such would like to know how this ranks with other risks in the enterprise. Bob is asked to perform a risk analysis and with the enlisted help of Natalie, an incident analyst that works in the company's Security Operations Center, he figures out the probability portion. Together they gather media reports on the ransomware epidemic, find research reports that states 91% of clients had been victimized by ransomware, and discover their own company had one infection the previous year. Bob then asks Natalie: "The research shows that the ransomware epidemic affects over 91% of companies. What is the probability of this happening here?". Natalie reviews the research reports, reads some news articles, and then gives her assessment. She states that there is a 100% probability of this occurring at their company.

With this information, Bob packages up his research, reports the probability as 100%, finishes the risk assessment, and sends it up the Board. The Board focuses in on the probability portion of the assessment and are pretty incredulous that a ransomware infection probability could be as high as 100%. Essentially stating that it would be a sure thing. Several of these Board members are experienced in risk analysis, others are familiar with probability theory, so they ask Bob for his work papers. They want to see the research, assumptions, and detailed analysis. Bob shows them the research report that cites the 91% figure, and recounts his workshop with Natalie. The Board reviews this and agrees that this is not sufficient – the risk analysis does not have the rigor they expect and does not have enough data to back up the probability claim. They ask him to go back, examine the fundamental problems in his analysis, and perform it again.

What were the problems of Bob's risk assessment? What all did Bob do wrong? The first thing we noted was the he used only one source for research, the 91% figure, which is not easy to defend. There was no other research Bob could've used to support his claim. He also did not recognize, identify, or control for cognitive bias. Who had contributed to this figure? Were they calibrated sources, or uncalibrated? In addition, he also only relied on one expert. He had no other expert probability estimates who could challenge both his and Natalie's assumptions. I pointed out that these are common mistakes made, however, even just one of them can really skew a risk analysis and misrepresent the entire range of risk.

After stating this example, I then began to discuss how to fix this. You have to gather and vet research; taking a very critical look at the sources we use for research. It needs to be vetted for quality while weeding out any junk, and identifying any biases. The most effective way is to use crowd-sourced probability estimates. What I mean by this is to ask many people for their opinion instead of just the risk analyst doing it their self or just one person. Over all of risk assessments I've done over the years, I've really come to appreciate diversity of opinions that comes from asking many people for probability

estimates. The more people you talk to, the better your underlying assumptions are, and therefore the better your risk assessment is. The hallmark of any good risk assessment is if you hand the same data, same research, and similar experts from the same field to a totally different risk analyst and they come up with the same or very similar results. That makes for a good risk assessment that you can defend.

Following this, I then gave a demonstration. I sent out a call for help to SIRA and FI and asked for a few things, with 15 experts answering the call. I first gauged their level of calibration and asked them for a probability assessment on the ransomware question. The entire process I talked about – gathering research, gathering experts, calibrating them, and combining results – is called eliciting expert judgment. It is an interdisciplinary practice that can be done in other practices such as medicine, biology, climate science, and engineering as well as risk management. Using expert judgment makes sense when you are trying to forecast something but there is any degree of uncertainty, such as a high level of conjecture or missing data.

We then returned to the hypothetical situation of Bob's data sources to see how we could improve. We looked into Bob's primary source, an interesting statistic from a firm called Datto in a 2016 global ransomware report. A scary statistic, but we found that there was more than met the eye. When reading the fine print, it was found that 91% figure came from not regular firms such as the one Bob worked at, but were M.S.P.'s, with a huge intake of incidents. We also discovered that it was not stat significant. We know that in survey science, stat significant means that the people who constructed the survey did their best to randomize the respondents to minimize and control for response bias. This enables both the survey firm and the reader to take the results and apply to a larger and more general population. Without such a distinction, the survey results only apply to the respondents themselves.

Another problem we found in Bob's risk assessment was that it only focused on what happened the previous year. While it is true that we can learn much about observing one thing, as described in Doug Hubbard's book "How to Measure Anything", in this situation this wouldn't apply. As a risk analyst, the fact that Bob's company, or any company for that matter, had one ransomware infection the previous year is interesting. What would be more interesting, however, is what happened in the previous five years?

In addition, using the word infection without context creates too much room for interpretation for the expert. Was it an old cryptolocker strain that was blocked by an antivirus or was it a multi-day, multi-server infection and we had to pay 500 bitcoin to a criminal gang in Belarus? This leads to the issue of using the term "ransomware epidemic" that Bob asked the expert. Now I may not be a lawyer, but I've learned enough from Jack McCoy from Law and Order to know that this sounds like leading the witness! If you want the expert to think there's a ransomware epidemic, you ask them about the epidemic. If you are asking someone to forecast the probability of a future ransomware incident, we create bias by using terms such as this.

The solution lies in a method utilized in Social Sciences known as "Literature Review", which is the cornerstone and key element of any research project. In a "literature review", you read all existing or subset of research and articles on a particular topic, and write this up and provide the direction for beginning your research. A similar framework works for risk management. You look into existing research on a subject, evaluate it, research methodology, and how it applies to the question. It's a valuable skill for risk analysts, and much better than letting an expert fend for themselves.

I then referred to Jay, who had covered this topic prior to myself, and used his own literature review. We turned to some existing research from the Cyentia Institute, a meta analysis of twelve different

studies on ransomware prevalence rates. We discussed whether the research was based on data or a survey. As a risk analyst, I obviously will favor data based. After, we aggregated the most important data points out of each research source, an improvement over Bob's singular research item. Taking all of this information and research, I stated that it had been sent over to our 15 experts to read and start thinking about how they would use this data to perform a probability assessment. While this was taking place, I moved onto discussing three forms of cognitive bias; availability bias, the overconfidence effect, and a form of anchoring/group think and confirmation bias that I dubbed the "infosec folklore effect".

Availability bias, also known as availability heuristic, is the mental shortcut when making a decision where we tend to favor very recent information. To illustrate this, I gave an example of a workshop that I had conducted a few weeks prior. In this workshop, we were gathering expert judgment for a variety of incident remediation efforts, but focusing on incidents that caused system outages. During this, people became laser focused on DDoS, cyber criminals, nation states, and North Korea. Internal and external data shows that the most dangerous threat actor this area are system admins that don't follow change control. This is something that people don't talk about or see in the news. How do we work with an example such as this? We encourage people to think in the long-term trends, tell them about the bias, and bring in data and research to present instead of letting people rely on memory alone.

The next bias discussed was the cognitive bias, also called the overconfidence effect. This tends to happen when you are eliciting expert judgment. I gave the example of asking an expert for the annualized probability of a data breach. You want the best case, worst case, and most likely case for this situation. The catch would be that you also told the expert that you wanted them to be at least 90% confident that the "correct" answer fell into that range. The overconfidence effect comes into play for this example because most people will overestimate their ability to give a correct estimate. The good news for this particular bias is that it is possible to correct with a technique called calibration, a real life example from the SIRA and FI volunteers later in the discussion.

The last bias discussed was something that I have been studying up on for quite some time that I have called the infosec folklore effect. It's in part availability, group-think, and confirmation bias. It's present when those of us in InfoSec believe a statistic to be true, even when there is contradictory evidence or research that tells us otherwise. For this, I listed several examples that would apply to this particular effect:

- "60% of small companies that suffer a cyber attack are out of business within six months."
- "80% of all cyber attacks originate from the inside."
- "75% of companies have experienced a data breach in the past 12 months."

For all three of these examples, infosec folklore can be perfectly seen. All three of these examples all contain some form of misquotation, evidence to prove to the contrary, or are the product of several opinion surveys. Yet all three examples are not true. Though they may seem accurate or plausible, they are not accurate and yet still continuously persist. I stated that avoiding the infosec folklore effect is difficult, but can be minimized with good research and a good vetting process.

We then directed our attention back to the experts from SIRA and FI, to which I had posted asking for their assistance. I requested that they were to read the Cyentia Institute blog post, participate in an exercise so that I could see how calibrated they were (control for over/under confidence), and answer one question on the prevalence of ransomware.

One topic I wanted to dig into was the ten question trivia exercise, containing seed questions, an interesting technique used to try and control for the overconfidence effect. A vast majority of people have this bias and have to work hard to overcome it, however there are three groups of people that show very little of this effect. Those groups are: bookies/bookmakers, meteorologists, and professional bridge players. The reason why these individuals are calibrated is due to the fact that they are constantly receiving feedback on the quality of their prior estimate. A bookmaker will know within hours or days if the odds they gave on a particular race horse were good. A weather person will know within days or even hours if their weather forecast was accurate. Those of us in cyber risk, however, have to wait anywhere from five to twenty years to see if any of our probability estimates come to fruition, and it's unlikely we'll receive any feedback from this. Self-calibration or calibration in a business setting is achievable. The idea with seed questions is to ask the expert general trivia questions to determine someone's level of calibration. They don't have to be InfoSec related at all to be effective. Doug Hubbard and Roger Cooke, both individuals who have both done influential work on this topic, advocate for these types of questions.

The first two questions that I sent to the SIRA and FI members were then shown and discussed. The first portion were true/false questions while the second portion was a self-assessment about how confident they were about getting the questions correct. The focus isn't about the trivia questions, or if the individuals completely bomb out on the answers. The focus is if the individual knows that they bombed out, and people who are calibrated will reflect this in the test. There was a total of ten questions asked, using Hubbard's method in "How to Measure Anything". Cooke generally recommends around 50 questions, but according to Hubbard, ten questions is enough to get an idea about over/under confidence. The results of this found that of the 15 people who responded to the questionnaire: 9 were perfectly calibrated and 6 were not, with varying degrees. Of those not calibrated, 5 were overconfident while 1 was under confident. Next, I asked the participants to answer a question on ransomware. I asked for them what the probability of a significant, company-impacting ransomware event would be for a typical company in the next year. They were to give the minimum (best case), maximum (worst case), and the mode (most likely case) of their probability. This allowed the experts the opportunity to express their uncertainty about their estimate.

Before moving on to combining the estimates, I briefly pivoted the topic to what to do with vastly differing opinions. I gave a checklist for vastly differing opinions and how to deal with them. The first on the checklist is are the individuals calibrated? For this situation, you can a few different things. You can discard their probability estimates, coach them on the ranges and on calibration, or you can integrate them into your final assessment but giving their estimates a lower weight. The next on the checklist is whether the individual misunderstood the question, research, or assumptions. To handle this particular situation, you will need to follow-up with the expert and review their understanding of the request. If there's a misunderstanding, ask for a reassessment. The last on the checklist is if the individual has a different world-view than others. Here, you can let the expert challenge your own assumptions and consider multiple risk assessment. A real world example of this is the field climate science on the topic of global warming. The vast majority, 97%, of climate scientists agree that humans are causing global warming, while 3% do not. The 97% of scientists have the same or similar worldviews, so it's reasonable to combine and aggregate their estimates. If the other 3% are included, they would skew the results, so they are put into their own assessment.

After this discussion, we returned back to the probability estimates and how we would combine them. There are two methods as to how to do this, behavioral and mathematical. The behavioral method usually entails the facilitator working through the problem with people until they reach a consensus.

The benefit of such a method is that it is fast and can quickly get people on the same page. The disadvantage of this method is the obvious issue of group-think. A lot of people will subconsciously, or consciously, adopt the same opinions are their leader/manager. The mathematical method most commonly averages out distributions. I cautioned not to use averaging for this method, as you can't average out distributions. I focused on the Linear Opinion Pool for the mathematical method, utilizing the Vose Software's Model risk for this. We set a baseline for all of the respondents with equal weight and went through several graphs based off of this, including some software that I recommended for combining all of the data. We then were able to finally wrap up Bob's story, as we were able to finish the risk assessment. Bob was able to present his new findings to the board, thanks to us, making everyone extremely happy.

To conclude my presentation, I thanked everyone who was able to participate and stated that there were several things we could do to promote these techniques. Many organizations utilize these techniques now, certainly outside of cybersecurity. Hosting local riskathons could be beneficial; getting a group of analysts together and working on a problem until we have a good probability estimate for a list of incidents. Seeing quant risk shops as well as FAIR shops could also help bring these techniques into their programs. It is up to us to help promote these techniques to others.