How to Calculate the Probability of a Third-Party Data Breach

Thomas Lee[1] PhD and Timothy Smith[2] CPA, CPA/CITP, CISA

Draft: September 19, 2025

Acknowledgements

The authors would like to acknowledge the contributions of the following individuals:

Spencer Graves[3] PhD, who has guided our <u>understanding and use of probability theory</u> and our <u>empirical regression modeling.</u>

Christine Dewhurst[4], for her many contributions including the co-discovery of the <u>similarities between organizations in their P_{ave} values</u>, <u>the derivation of Equation 4</u>, <u>categorizing third parties by potential breach size</u> and many hours of discussion regarding the current practices for managing third-party data breach risk and how organizations can begin managing the number of third parties.

Patricia Drooff[5], for her assistance in articulating concepts such as the <u>nature of third-party data breach risk revealed by Equation 4</u>, for the <u>Answers to Frequent Questions and Comments</u> section and for the general organization of this white paper.

Peter Davis[6], for advice on the current practice for <u>third-party risk management</u> and how this risk is understood in standards such as COBIT; for shepherding the first talk for measuring third-party risk based upon probability theory, and for helping to clarify concepts within this white paper in terms of the current cybersecurity standards.

Download the latest versions

Table of Contents

Why Calculate Third-Party Risk

There is a common assumption by business leadership that the current approach to **Third-Party Risk Management** (TPRM)—vetting each third party in isolation—is sufficient to mitigate the risk of a third-party data breach. In reality, there is also a large **systemic**[7] **cumulative risk** that stems from the **sheer number of third parties** that have access to large amounts of sensitive data. Failing to measure this **cumulative risk** not only hampers the organization's capacity to harness valuable third-party technologies but also undermines its competitive edge.

Third-party breaches are not hopelessly random. Their **expected frequency** can be well characterized with minimal effort using **Probability Theory**. Regularly calculating the **expected frequency** as a function of breach size can give leadership **confidence in using third parties** even though small third-party breaches do occur regularly. These small breaches that might happen every couple of years can be considered the cost of doing business. Simultaneously, leadership can be **reasonably assured**[8] that a very large breach[9] will never happen by enforcing calculated limits[10] on the number of third parties handling the largest volumes of sensitive data. Business leaders can have confidence in this calculation because it is solidly grounded in **Probability Theory** and because it is tested through tracking the actual occurrence of small third-party breaches.

For a business leader that wants to weigh the value versus the effort, consider that the effort can be focused primarily on identifying the (hopefully) small number of third parties with very large amounts of data. The number of these third parties is typically fewer than one hundred. The value, then, is the confidence to use more third parties without the fear of experiencing an impactful third-party breach.

How to Use This Document

This document is intended for two distinct audiences. The first group includes those responsible for evaluating the foundation and limitations of the calculations (i.e., model risk), such as the CRO, CISO,

---

[7] Systemic risks are **expected** but initially unknown risks that derive from the complexity of a system. The expectation of systemic risks is the reason software engineers perform integration testing, for example. We reveal the nature of this systemic risk using mathematics.

[8] In this context, reasonably assured means a probability that is below 1% (100 years) and is very unlikely to occur. See Judging Acceptability from the Organization Perspective.

[9] What constitutes a small or a large breach is for business leadership to decide and is within their power to control, once they understand the nature of the risk.

[10] Business leaders are able to decide an achievable expected frequency as a function of potential breach size and then calculate the number of allowed third parties (see How Business Leaders can Calculate Third-Party Thresholds).

CIO, Board members, Internal Audit, and model validation teams. If you are in this first group, skip to the section titled [The Mathematical Basis of Third-Party Data Breach Risk](#).

The second group consists of individuals tasked with implementing the approach, including professionals in Data Governance or Third-Party Risk Management (TPRM). If you are in the second group, begin with the next section.

How to Implement the Calculations

In this section, we will step you through how to calculate probability for third-party data breach. Probability will be calculated as a function of data breach size. Implementing the calculations can be divided into the following steps:

1. **Data collection**, discovering third parties and the amount of data each can expose
2. **Data organization,** by the amount of data third parties can expose
3. **Calculating probabilities**, as a function of potential data breach size
4. **Judging acceptability**, by the impact to your customers

We provide three examples:

- **Unknown Breach Size,** typical for a small organization with an immature TPRM program
- **Two data breach sizes**, typical for an organization with a strong TPRM program
- **Three data breach sizes,** typical for a large organization

Data Collection

Create a list of all third parties which 1) have access to sensitive data and 2) could expose this data if they were to have an internal data breach or a data breach by one of their own third parties. Following are considerations:

1. Accuracy of the calculation depends on the accuracy of the relevant third-party list and an accurate assessment of the amount of data that would be exposed for each third party. If the list is not comprehensive, the calculation will yield a probability which is too small and an organization's business leadership might be overconfident in using third parties. Including third parties that cannot expose your data will produce a probability which is too high and might limit your organization's ability to use additional third parties.

2. Some large organizations use a *federated* approach for third-party risk management (TPRM), where TPRM is divided by country. Some organizations in the pharmaceutical industry divide third parties as GMP (required to implement or maintain "Good Manufacturing Practice") or non-GMP, and TPRM is performed by different groups. In some organizations, the CIO might promote or at least not discourage shadow-IT, since this is a way to obtain additional funding for IT projects for which the CIO cannot get funding. This shadow-IT might include SaaS, PaaS, IaaS and even low-code, no-code third-party services that could expose data. All of these separate third-party lists must be combined in order to generate accurate calculations.

3. Some organizations fail to require that third parties purge the organization's data post engagement. If these *historical* third parties could still affect your organization were they to experience a data breach, they should be included in the list.

4. **Fourth parties should not be included** in the calculation, since many third parties use the same fourth parties and this will result in an overestimation of the probability. For example, many third parties use the same cloud service provider, such as Microsoft Azure^{tm}. Counting Azure^{tm} multiple times will overestimate probability. Also, P_{ave} in effect already includes fourth-party risk, since a third-party breach is a very common way for your third parties to experience a data breach.

5. For each third party, a qualified cybersecurity expert should be consulted to determine if sensitive data could actually be exposed in the event of a third-party breach. For example, if the third party is providing virtual infrastructure such as virtual servers, and the virtual servers are encrypted and only your organization has the encryption keys, then this third party should not be included in the calculation. As another example, if a third party has access to your organization's data but the sensitive portions are obfuscated, then this third party should not be included in the calculation.

6. The quantity of data that could be exposed by each third party should also be collected. If the data that could be exposed is nonpublic PII data (see Glossary), then the quantity of data should be determined simply as the number of people that would be affected if there were to be a third-party data breach.

7. In the case of PII, do not subdivide this data into different kinds of PII, since this will result in many smaller **probabilities** rather than one large **probability** which better reflects your organization's likelihood for a third-party breach.

8. Since the accuracy and credibility of the calculation is no better than the data used, documentation should be maintained regarding the accuracy of the list and the amount of data shared with each third party. Table 1 shows example documentation.

Table 1, Example Documentation of a Third-Party List

Third Party	People affected	Access	Affirmation
Acme Virtual Servers	11,567,876 customers	Unencrypted database on an unencrypted server	John Doe, DBA, SQL query, 2/10/2025
Acme Hosted HR Systems	2,501 employees	Unencrypted benefits on all current and past employees	Jane Doe, head of HR, summary page from HR system, 1/12/2025
Acme web hosting	11,567,876 customers	Web portal for our customers to manage their data	John Doe, DBA, SQL query, 2/10/2025
Acme IT services	11,567,876 customers + 2,501 Employees	Remote IT services	Sum of all HR and customer data
Acme PayCheck	2,501 Employees	Unencrypted employee payment system	No longer use this vendor but **data was not purged**. Jane Doe, head of HR, 1/12/2025

Data Organization

Before we calculate probabilities, third parties must be organized into groups based upon the amount of data they can expose. The organization begins with creating ***Breach Size Categories*** that are useful for business leaders to decide how often third-party breaches should be allowed to occur.

Create Breach Size Categories

Create breach size categories that will be used to organize third parties. There should be at least three categories:

1. A breach size that is not impactful and can occur with some frequency, for example **one thousand** people affected every **2 years** on average,
2. A breach size that would be impactful but could be tolerable if it occurred rarely, for example **10 thousand** people affected every **10 years** on average,
3. A breach size that leadership is willing to make significant effort to ensure will not happen, for example **10 million** people affected with a frequency of **200 years**.

The purpose of these categories is to allow leadership to use the maximum number of third parties while remaining within their [quantitatively determined risk tolerances](#). For example, if leadership can tolerate a small third-party breach as often as every two years, they could expose this small amount of data through as many as 620 third parties (see [Precalculated Probabilities](#)). Similarly, if leadership could tolerate a third party breach affecting ten thousand people as often as 10 years, then they could expose this amount of data through as many as 146 third parties. Finally, to ensure a breach affecting ten million people will never occur, they might choose a target frequency of 200 years. Such a risk tolerance would limit the organization to exposing this large amount of data through just 8 third parties.

Group Third Parties

After defining breach size categories, the next step involves organizing third parties according to these categories.

The probability of a third-party data breach is always calculated for a range of breach sizes and we will organize third parties so that breach size ranges follow the natural pattern. The natural pattern is for large breaches to be less common, while smaller breaches are more frequent, making it logical to organize third parties in a way that reflects this pattern—from small, common breaches to large, rare ones.

To accomplish this, we propose visualizing third parties in nested circles (see Figure 1). The outermost circle encompasses all third parties, and therefore has the highest probability of a breach. Ideally, if good security practices for sharing data are followed[11], the majority of third parties in this outer circle will hold minimal amounts of data. Since these smaller data holders are the most numerous, the most probable breach will be small. This outer circle should therefore be labeled to denote the smallest breach (e.g., "1K+"), covering probabilities for breaches from 1,000 people affected and up, with 1,000 people affected being most probable.

The next inner circle represents third parties with the subsequent breach size category and beyond. For example, this circle could be labeled "10K+" and account for probabilities related to breaches starting at 10,000 people affected or more. The nested structure continues, reflecting increasingly larger breach size ranges as the circles move inward.

---

[11] For example, ***Need to Know*** which is about limiting data access based on relevance or role-specific requirements. This is similar to ***Principle of Least Privilege*** (PoLP) which usually refers to system or permission levels.
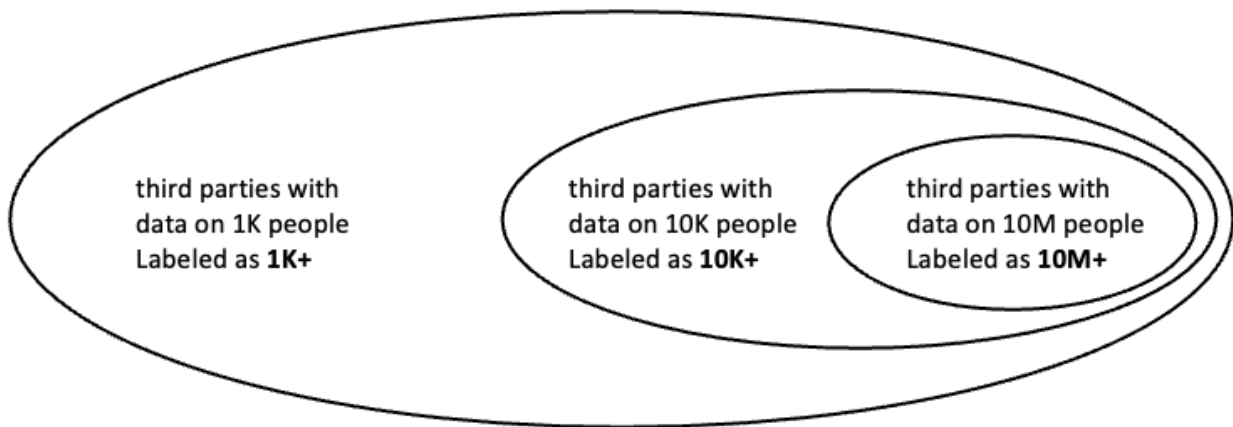
Figure 1, Think of third parties as organized into nested circles[12].
- The largest circle will contain all third parties and will be labeled with the smallest amount of data shared with a third party. For example, if the smallest amount of data could affect one thousand people if it were exposed, we would label this circle 1K+, to indicate it represents a third-party data breach affecting one thousand <u>or more people</u>. Since this circle contains all third parties, it will have the highest probability, and will generally represent the probability for a small third-party data breach.
- The next circle, which will be within the 1K+ circle, might enclose all third parties with enough data to affect 10 thousand or more people, and we would label this circle 10K+.
- Within the 10K+ circle is another circle which might enclose all third parties with enough data to affect 10 million or more people and we would label this circle 10M+.

Calculate Probabilities

Use <u>Equation 4</u>, which is $E_{cum} = N \times P_{ave}$ to calculate a probability (or **expected value)** for each circle, based upon the number of third parties within each circle. E_{cum} is the probability (or <u>expected value when probability is high</u>), N is the number of third parties within each circle and P_{ave} is the average probability for third parties to cause a third party data breach. As we <u>explain below</u>, we estimate P_{ave} to be the annual probability 0.066%. For each circle, set N to the number of third parties within the circle and the nested circles within. If P_{ave} = 0.066% is used, you can save time by using <u>Precalculated Probabilities</u> from the table below.

For example, using Figure 1 above, if there are 200 third parties that are only in the outer circle which is labeled 1K+, and there are 50 third parties only within the first inner circle labeled 10K+, and there are 5 third parties within the innermost circle, labeled 10M+, then we would combine 200 with 50 and 5 to obtain 255 total third parties enclosed by the outer circle and calculate (200 + 50 + 5) x 0.066% = 0.168 or 17% annual probability for the outer circle. If we invert 17% (1/0.168) to get the expected number of years between occurrences, we get 5.9 years. This could also be understood as there being a small

---

[12] Of course these are ovals and not circles, since ovals take up less space in the white paper.

breach every year, on average, among six companies with a 17% probability. So, we should expect a third party breach every six years on average, and this breach is most likely to be a small breach since the number of third parties with small amounts of data is 3.6 times more numerous (3.6=200/(50+5)).

For the circle labeled 10K+, we would combine 50 with 5 to obtain 55 total third parties enclosed by the middle circle and calculate (50 + 5) x 0.066% = 0.036 or 4% annual probability for the middle circle. If we invert 4% we get 27.5 years. This could be understood as a breach every year, on average among 28 organizations with a 4% probability.

Finally, for the circle labeled 10M+ we calculate 5 x 0.066% = 0.0033 or 0.3% annual probability. If we invert 0.3% we get 303 years. This can be understood to mean a breach every year, on average among 303 organizations with a 0.3% probability.

Judging Acceptability

Business leadership should set policies and decide probabilities (or expected values) that are acceptable and achievable. We present two ways to judge the acceptability: 1) from the perspective of the organization and 2) from the perspective of the public whose data might be exposed.

From the Organization Perspective

An acceptable **expected value** or **probability** depends on the data breach size. For small data breaches and large organizations, a data breach every few years might be considered acceptable and simply the consequence of normal business. This would especially be true if business leadership understands that the large **expected value** is the consequence of a large number of third parties with only small amounts of data, and that the very large breach is **reasonably assured** not to happen because there are very few third parties with very large amounts of data.

An acceptable probability for a very large data breach might be below 1% (100-years), and **reasonable assurance** might be a probability below 0.5% (200-years). One way to consider a 1% probability is to understand that this breach would be expected every year on average among 100 organizations with a 1% probability.

From the Public Perspective

One way to consider acceptability to the public is to consider how often a member of the public should expect to have their personal data exposed. A 1% probability might be acceptable to an organization, but consider that a member of the public might do business with, for example, 20 organizations. The **Linearity of Expectations (**section below**)** also applies to members of the public, so this 1% probability would build up across the 20 organizations and a member of the public would see a much higher probability. We can use **Equation 4,** which is a general purpose equation derived from the **Linearity of Expectations**, to calculate what this *cumulative* probability would be for a member of the public.

When we use the equation for third-party data breach, we used 0.066% for P_{ave} because this is the average annual probability we find empirically among third parties. To apply **Equation 4** to the public, N will be the number of organizations with which a member of the public has a relationship and which could expose their personal data and P_{ave} will be the average probability for an organization to expose data, for a member of the public. In other words, when we applied Equation 4 to third parties, we used the P_{ave} that we find empirically for Business to Business (B2B) organizations. When we apply Equation 4 for a member of the public, we will be asking what P_{ave} is acceptable for Business to Consumer (B2C) organizations.

In this case, we want to answer the question: would a 1% annual probability be acceptable, so we will set P_{ave} to 1%. If we assume that a member of the public does business with 20 organizations that could expose their data and P_{ave} is 1% then a member of the public should expect to have their personal data exposed every 5 years on average (E_{cum} = N x P_{ave} = 20 x 1% = 20% or 5 years). Whether this is a concern for the organization's customers depends on how many customers the organization could affect.

How Business Leaders can Calculate Thresholds

Business leaders can decide *achievable* values for expected frequencies for third-party data breaches and then calculate the number of third parties that would be allowed. We use the word *achievable* because leadership would of course like the likelihood to be zero, but this would be too costly. The cost comes in terms of being limited in the use of third parties in order to maintain thresholds, and the efforts to encrypt or obfuscate data that is shared with third parties beyond these thresholds. There is also the cost to ensure that any data shared is purged post engagement with a given third party. Once thresholds are determined by leadership, the organization can focus on achieving these thresholds.

Beginning with Equation 4, we can solve for *N*:

$$N = E_{cum} \div P_{ave} = 1 \div (EF \times P_{ave})$$

Where *EF* is the expected frequency, and the inverse of E_{cum}. Business leaders can perform this calculation for breach sizes that make sense for their business. Table 2 shows example calculations.

Table 2, Example calculated target *N* thresholds as a function of breach size

Breach size (people affected)	Achievable Expected Frequency (*EF*)	*N* threshold 1 ÷ (*EF* × 0.066%)
1K+	5-years	303 third parties
10K+	10-years	152 third parties
100K+	30-years	50 third parties
10M+	500-years	3 third parties

The table shows, for example, that an expected frequency of 500 years would allow only exposing records for 10 million or more people through three third parties, but **reasonably assures** a breach affecting 10M+ people will not occur. Note that more third parties can be used with such large amounts of data, but the data would need to be encrypted or obfuscated. See the table [Precalculated Probabilities](#) as a fast way for business leaders to decide achievable frequencies based upon the $\overline{P}_{ave}$ of 0.066%.

Such goals can be aspirational; achieved and even modified over time. These thresholds are of course only for third parties that have unencrypted or unobfuscated data.

Examples

Unknown Breach Size

We begin with an example that is not ideal. It is not uncommon for smaller organizations to not track the amount of data shared with third parties. The organization can produce a list of third parties that can expose data but the amount and even kind of sensitive data is unknown. This results in only a single *circle* of third parties, but calculating the expected value can be the first step in deciding when further action is worth the effort.

Example

An organization has 13 third parties that could expose sensitive data. The amount of data that each third party has is unknown.

Probability for a third-party data breach

Using Equation 4 and the 0.066% value for P_{ave}, we can calculate 13 x 0.066% = 0.86% or once in 117 years (117 = 1/0.0086). Using the standard deviation of 0.027%, probability can range from 0.51% or once in 197 years to 1.2% or once in 83 years.

Table 3, probability for a third-party data breach affecting 1+ people

People Affected	Third Parties (N)	+1σ N x (0.066% + 0.027%)	Median N x 0.066%	-1σ N x (0.066% - 0.027%)
1+	13	1.2% (83 years)	0.86% (117 years)	0.51% (197 years)

Comment

This example demonstrates how to calculate likelihood for an organization that has not documented the amount of data shared with third parties. The likelihood is small, with an expected frequency in 117-years due to the small number of third parties and is likely acceptable to the organization's business leadership (see below: *Judge Acceptability*). This is a small organization (we know from the small number of third parties) and at this point in the organization's growth, no further action is warranted. The amount of data was not specified so it is hard to judge acceptability for the public.

Two Data Breach Sizes

The following is a more representative example: a larger organization that has tracked the amount of data for each third party.

Example

A company has 400 third parties, each of which could expose data for 1K+ people and a subset of 10 third parties that could expose data for 1M+ people.

In this example, we will divide third parties into two nested circles. The outer circle contains all third parties that have any amount of data but which is largely made up of third parties that can expose small amounts of data. The inner circle contains just ten third parties that can expose large amounts of data. We use the notation 1K+ for the outer circle, with the "+" indicating a breach could affect one-thousand **or more people** because this outer circle includes within it the smaller inner circle. The inner circle we will label 1M+ because the amount of data is likely approximate and will likely grow over time.

Probability for a data breach affecting 1K+ people (outer circle)

Using Equation 4 and the 0.066% value for P_{ave}, we can calculate 400 x 0.066% = 26% or once in 3.7 years (3.7 = 1/0.26). Using the standard deviation of 0.027%, probability can range from 16% or once in 6.4 years to 37% or once in 2.7 years.

Probability for a data breach affecting 1M+ people (inner circle)

Using Equation 4 and the 0.066% value for P_{ave}, we can calculate 10 x 0.066% = 0.66% or once in 152 years. Using the standard deviation of 0.027%, probability can range from 0.29% or once in 256 years to 0.93% or once in 108 years.

Table 4, probability (expected value) for a third-party data breach affecting 1K+ and 1M+ people

People Affected	Third Parties (N)	Expected Value (Probability)		
		+1σ N x (0.066% + 0.027%)	**Median** N x 0.066%	**-1σ** N x (0.066% - 0.027%)
1K+	400	37% (2.7 years)	26% (3.8 years)	16% (6.4 years)
1M+	10	0.93% (108 years)	0.66% (152 years)	0.239% (256 years)

Comment

This example demonstrates how to calculate likelihood for an organization that has tracked the amount of data shared with third parties. This is also an example of an organization that has effectively managed its potential exposure of large amounts of data, even as it benefits from a large number of third parties. Table 2 shows that while a small third party breach is expected (3.8-year frequency), a large third-party breach is **reasonably assured** not to happen since third parties' access to large amounts of data is limited. This is also likely acceptable to the public since a member of the public should only expect to have their personal data exposed every eight years on average if they did business with 1020 similar organizations (see below: *Judging Acceptability*).

Three Data Breach Sizes

In this example, a large organization has third parties with access to both sensitive customer and employee data. We will step through the process of eliminating and consolidating third parties as well as performing the calculations.

Example

A large organization has organized third parties with access to sensitive data according to Table 5.1.

Table 5.1, Initial third-party data.

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged
11,221	Below 500	No	HR	Active	No
1,127	500 to 1000	No	HR	Active	
57	10,000 to 50,000	Yes	HR	Active	
137	10,000 to 50,000	No	HR	Active	

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged
87	80,000,000+	Yes	Customer	Active	
19	80,000,000+	No	Customer	Active	
411	500 to 1000	No	HR	Inactive	No
2	80,000,000	NA	Customer	Inactive	Yes
11	10,000 to 50,000	No	HR	Inactive	No
55	10,000 to 50,000	No	HR	Inactive	Yes

Step 1: Imagine that business leadership has set a threshold of 500 people, below which we will ignore third parties for this calculation. Perhaps 500 people affected was chosen as a threshold because many reporting laws make public only data exposures that affect 500 or more people. So, we will begin by eliminating third parties that will not be part of the calculation. This eliminates more than 11,000 third parties (see Table 5.2)

We will also eliminate all third parties that are inactive and where data has been purged. Finally, we will eliminate all third parties where data has been obfuscated or encrypted since, if these third parties were to experience a data breach, our data will not be exposed. Note that encrypted means data is encrypted in transit and at rest and only our organization has the keys for decrypting the data.

Table 5.3 shows the third parties remaining after removing rows from table 5.2 which met one or more of the conditions referenced above. Following is a summary of eliminated third parties:

Eliminated	Justification
Third parties that can affect fewer than 500 people	Limit set by business leadership. The point where breaches are made public.
Third parties where data is obfuscated or encrypted	Impactful data cannot be exposed if the third party experiences a data breach
Inactive and data has been purged	There is no data that would be exposed if the third party experiences a data breach

Table 5.2, Third parties from Table 5.1, that will be eliminated because the number of people affected is below a threshold of 500 or whose data is encrypted, obfuscated or purged, as indicated in the last column right.

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged	Reason for eliminating
11,221	Below 500	No	HR	Active	No	<500
1,127	500 to 1000	No	HR	Active		

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged	Reason for eliminating
57	10,000 to 50,000	Yes	HR	Active		Encrypted or obfuscated
137	10,000 to 50,000	No	HR	Active		
87	80,000,000+	Yes	Customer	Active		Encrypted or obfuscated
19	80,000,000+	No	Customer	Active		
411	500 to 1000	No	HR	Inactive	No	
2	80,000,000	NA	Customer	Inactive	Yes	Purged
11	10,000 to 50,000	No	HR	Inactive	No	
55	10,000 to 50,000	No	HR	Inactive	Yes	Purged

Table 5.3, Third parties remaining from Table 5.2

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged
1,127	500 to 1000	No	HR	Active	
137	10,000 to 50,000	No	HR	Active	
19	80,000,000+	No	Customer	Active	
411	500 to 1000	No	HR	Inactive	No
11	10,000 to 50,000	No	HR	Inactive	No

Step 2: Reorder rows from Table 5.3 and combine third-party counts by potential breach size (people affected). Results are presented in Table 5.4.

Table 5.4, Order rows with the same amounts of data

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged
1,127	500 to 1000	No	HR	Active	
411	500 to 1000	No	HR	Inactive	No
1,538	Total third parties				

Third Parties	People Affected	Encrypted or obfuscated	Data Type	Status	Purged
137	10,000 to 50,000	No	HR	Active	
11	10,000 to 50,000	No	HR	Inactive	No
148	**Total third parties**				
19	80,000,000+	No	Customer	Active	
19	**Total third parties**				

Step 3: Form nested circles (figuratively of course) of third parties (see Table 5.5) by adding the number of third parties for larger breaches to smaller breach sizes. For example, the table shows that for the number of third parties for the outermost circle 500+, we will add the number of third parties for breaches sizes from 10,000 to 50,000 (148 third parties) and 80,000,000+ (19 third parties).

In labeling each size range, we will use the smaller number. For example, for the range from 500 to 1000 people affected, we will use the label 500+. The reason is because if the organization is following **Principle of Least Privilege** (PoLP) and sharing as little data as possible, most third parties will have the smaller amount of data.

Table 5.5, Calculate total number of third parties for potential breach size ranges

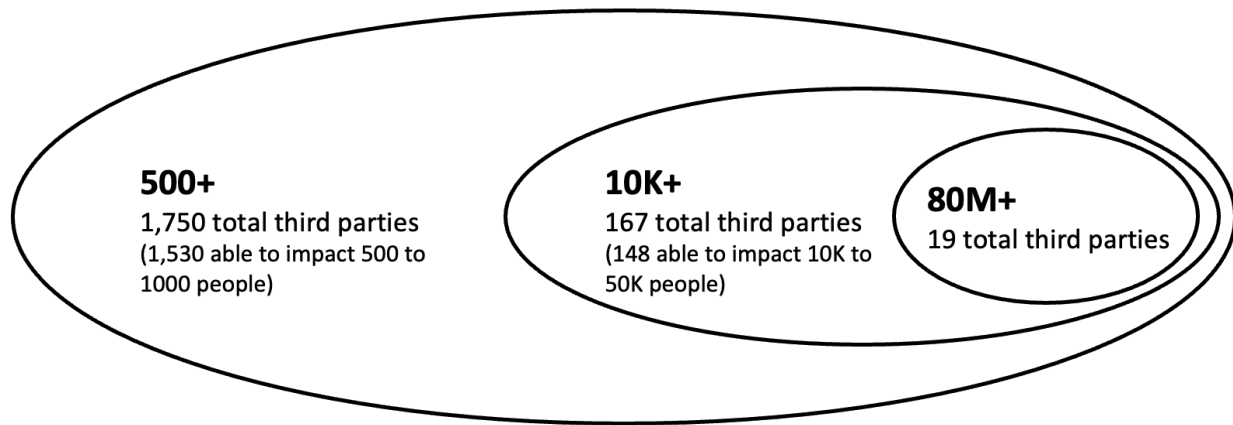| Labels (nested circles)	People affected	Number of Third Parties
500+	500 to 1000	1,538
	10,000 to 50,000	148
	80,000,000+	19
	Total	**1705**
10,000+	10,000 to 50,000	148
	80,000,000+	19
	Total	**167**
80,000,000+	80,000,000+	19
	Total	**19**

Figure 2, Totals from Table 5.5 presented in nested circles, similar to the circles in Figure 1 above.

Step 4: Finally calculate the range of **Expected Values** for each third-party breach size for the third party totals in Table 5.5 using Equation 4 and the 0.066% value for P_{ave}. These values are shown in Table 5.6. Frequency in years was calculated by inverting the **expected values**.

Table 5.6, Calculated Expected Values for potential breach size ranges

People affected	Third parties (N)	Expected Value (Probability)		
		+1σ N x (0.066% + 0.027%)	**Median** N x 0.066%	**-1**σ N x (0.066% - 0.027%)
500+	1,705	159% (0.6-years)	113% (0.9-years)	66% (1.5-years)
10,000+	167	15.5% (6.4-years)	11% (9-years)	6.5% (15-years)
80,000,000+	19	1.8% (57-years)	1.3% (80-years)	0.74% (135-years)

Comment

This example demonstrates the process of culling and combining third parties into figurative nested circles or groups, then calculating expected values for these groups so that business leadership can understand *why* and *how often* they should expect a third-party data breach as a function of data breach size.

Table 5.6 shows that a small third-party breach affecting 500+ people is calculated to be nearly every year on average (median frequency 0.9-years). We know these will be mostly small data breaches because the majority of the 1,705 third-parties that compose this category have records on from 500 and 1000 people (see first three rows of Table 5.5). Business leadership will likely view this as acceptable, since the organization is benefiting from a very large number of third parties. From the number of third parties, we know this is a very large company so a data breach affecting one thousand people is not of great concern. Leadership also knows this high frequency is restricted to small third-party breaches because the number of third parties with much larger amounts of data is greatly restricted.

Table 5.6 also shows that a larger breach affecting 10,000+ people is rarer, with a median frequency of every 9-years and this is because this category of third parties is one tenth the number of third parties in the first category.

Finally, table 5.6 shows us that a massive third-party breach affecting 80 million people is unlikely to happen since there are only 19 third parties that could expose this very large amount of data. However, that's not to say it cannot happen. This probability is not acceptable to the public (see: Judging Acceptability *from the Public Perspective*) and leadership would prefer a lower probability.

Reference Table of Probabilities per Number of Third Parties

Following are calculated values using the $\overline{P}_{ave}$ from Figure 5 for the P_{ave} in Equation 4 from Figure 3. Table values can be used for each *nested circle* after third parties have been organized (see *Calculate Probabilities*). Based upon the table below, the authors regard a third-party breach as **reasonably assured not to happen** with 1 - 7 third parties, and **unlikely to happen** with 8 - 16 third parties. Please read the *Disclaimer* section below.

One way to evaluate the risk is to divide the frequency in half and understand there is a fifty-fifty chance within that period of time. For example, if the frequency is 10 years, then there is a fifty-fifty chance in 5 years.

Table 6, Expected Values as a Function of the Number of Third Parties

Third Parties (N)	+1σ N x (0.066% + 0.027%)	Median N x 0.066%	-1σ N x (0.066% - 0.027%)
1	0.09% (1075-years)	0.07% (1515-years)	0.04% (2564-years)
2	0.19% (538-years)	0.13% (758-years)	0.08% (1282-years)
3	0.28% (358-years)	0.2% (505-years)	0.12% (855-years)
4	0.37% (269-years)	0.26% (379-years)	0.16% (641-years)
5	0.47% (215-years)	0.33% (303-years)	0.2% (513-years)
6	0.56% (179-years)	0.4% (253-years)	0.23% (427-years)
7	0.65% (154-years)	0.46% (216-years)	0.27% (366-years)
8	0.74% (134-years)	0.53% (189-years)	0.31% (321-years)
9	0.84% (119-years)	0.59% (168-years)	0.35% (285-years)
10	0.93% (108-years)	0.66% (152-years)	0.39% (256-years)
11	1.02% (98-years)	0.73% (138-years)	0.43% (233-years)

Third Parties (N)	+1σ N x (0.066% + 0.027%)	Median N x 0.066%	-1σ N x (0.066% - 0.027%)
12	1.12% (90-years)	0.79% (126-years)	0.47% (214-years)
13	1.21% (83-years)	0.86% (117-years)	0.51% (197-years)
14	1.3% (77-years)	0.92% (108-years)	0.55% (183-years)
15	1.4% (72-years)	0.99% (101-years)	0.59% (171-years)
16	1.49% (67-years)	1.06% (95-years)	0.62% (160-years)
17	1.58% (63-years)	1.12% (89-years)	0.66% (151-years)
18	1.67% (60-years)	1.19% (84-years)	0.7% (142-years)
19	1.77% (57-years)	1.25% (80-years)	0.74% (135-years)
20	1.86% (54-years)	1.32% (76-years)	0.78% (128-years)
21	1.95% (51-years)	1.39% (72-years)	0.82% (122-years)
22	2.05% (49-years)	1.45% (69-years)	0.86% (117-years)
23	2.14% (47-years)	1.52% (66-years)	0.9% (111-years)
24	2.23% (45-years)	1.58% (63-years)	0.94% (107-years)
25	2.33% (43-years)	1.65% (61-years)	0.98% (103-years)
26	2.42% (41-years)	1.72% (58-years)	1.01% (99-years)
27	2.51% (40-years)	1.78% (56-years)	1.05% (95-years)
28	2.6% (38-years)	1.85% (54-years)	1.09% (92-years)
29	2.7% (37-years)	1.91% (52-years)	1.13% (88-years)
30	2.79% (36-years)	1.98% (51-years)	1.17% (85-years)
31	2.88% (35-years)	2.05% (49-years)	1.21% (83-years)
32	2.98% (34-years)	2.11% (47-years)	1.25% (80-years)
33	3.07% (33-years)	2.18% (46-years)	1.29% (78-years)
34	3.16% (32-years)	2.24% (45-years)	1.33% (75-years)
35	3.26% (31-years)	2.31% (43-years)	1.37% (73-years)
36	3.35% (30-years)	2.38% (42-years)	1.4% (71-years)
37	3.44% (29-years)	2.44% (41-years)	1.44% (69-years)
38	3.53% (28-years)	2.51% (40-years)	1.48% (67-years)
39	3.63% (28-years)	2.57% (39-years)	1.52% (66-years)
40	3.72% (27-years)	2.64% (38-years)	1.56% (64-years)
41	3.81% (26-years)	2.71% (37-years)	1.6% (63-years)
42	3.91% (26-years)	2.77% (36-years)	1.64% (61-years)
43	4% (25-years)	2.84% (35-years)	1.68% (60-years)

Third Parties (N)	+1σ N x (0.066% + 0.027%)	Median N x 0.066%	-1σ N x (0.066% - 0.027%)
44	4.09% (24-years)	2.9% (34-years)	1.72% (58-years)
45	4.19% (24-years)	2.97% (34-years)	1.76% (57-years)
46	4.28% (23-years)	3.04% (33-years)	1.79% (56-years)
47	4.37% (23-years)	3.1% (32-years)	1.83% (55-years)
48	4.46% (22-years)	3.17% (32-years)	1.87% (53-years)
49	4.56% (22-years)	3.23% (31-years)	1.91% (52-years)
50	4.65% (22-years)	3.3% (30-years)	1.95% (51-years)
52	4.84% (21-years)	3.43% (29-years)	2.03% (49-years)
54	5.02% (20-years)	3.56% (28-years)	2.11% (47-years)
56	5.21% (19-years)	3.7% (27-years)	2.18% (46-years)
58	5.39% (19-years)	3.83% (26-years)	2.26% (44-years)
60	5.58% (18-years)	3.96% (25-years)	2.34% (43-years)
64	5.95% (17-years)	4.22% (24-years)	2.5% (40-years)
68	6.32% (16-years)	4.49% (22-years)	2.65% (38-years)
72	6.7% (15-years)	4.75% (21-years)	2.81% (36-years)
76	7.07% (14-years)	5.02% (20-years)	2.96% (34-years)
80	7.44% (13-years)	5.28% (19-years)	3.12% (32-years)
84	7.81% (13-years)	5.54% (18-years)	3.28% (31-years)
88	8.18% (12-years)	5.81% (17-years)	3.43% (29-years)
92	8.56% (12-years)	6.07% (16-years)	3.59% (28-years)
98	9.11% (11-years)	6.47% (15-years)	3.82% (26-years)
106	9.86% (10-years)	7% (14-years)	4.13% (24-years)
114	10.6% (9-years)	7.52% (13-years)	4.45% (22-years)
122	11.35% (9-years)	8.05% (12-years)	4.76% (21-years)
132	12.28% (8-years)	8.71% (11-years)	5.15% (19-years)
146	13.58% (7-years)	9.64% (10-years)	5.69% (18-years)
160	14.88% (7-years)	10.56% (9-years)	6.24% (16-years)
180	16.74% (6-years)	11.88% (8-years)	7.02% (14-years)
210	19.53% (5-years)	13.86% (7-years)	8.19% (12-years)
240	22.32% (4-years)	15.84% (6-years)	9.36% (11-years)
280	26.04% (4-years)	18.48% (5-years)	10.92% (9-years)
340	31.62% (3-years)	22.44% (4-years)	13.26% (8-years)

Third Parties (N)	+1σ N x (0.066% + 0.027%)	Median N x 0.066%	-1σ N x (0.066% - 0.027%)
460	42.78% (2-years)	30.36% (3-years)	17.94% (6-years)
620	57.66% (2-years)	40.92% (2-years)	24.18% (4-years)
1100	102.3% (1-years)	72.6% (1-years)	42.9% (2-years)

Mathematical Basis of Third-Party Data Breach Risk

只知其然 不知其所以然

Only knowing that it is so, without knowing why it is so.
Zhu Xi (1130–1200) Chinese philosopher and historian

Zhu Xi's thoughts from over eight hundred years ago are relevant to the cyber security industry today with regard to continuing data breaches. Cybersecurity practitioners have long attempted to provide the same level of controls to entities outside the firewall as they had for their own organization, as more and more non-core business functions have been outsourced to third parties. It has been assumed that carefully selecting third parties, sending them internal control questionnaires and requiring them to undergo external security audits, such as SOC 2s and / or sending staff to monitor them would reduce or even eliminate the risk of breaches for such outsourcing. Nonetheless, despite the effort and expense directed to reducing the risk of third parties with corporate data having a breach, they still happen. And, in organizations with a significant number of third parties, these traditional TPRM procedures do not scale well, especially if an organization does not want to spend an outsized portion of the budget on TPRM.

So, coming back to Zhu Xi, we know that third party breaches happen despite our best efforts, but we do not know why. We analyze third parties in what is often great detail, individually. We may even find one or two, which, in our due diligence, appear to be weakest links from a security standpoint, so work with them to remediate. Yet breaches happen, sometimes in those third parties we would least expect them. The truth of the matter is that third-party data breaches are a non-zero probability event. They are going to happen. Why?

In this section, we will examine why, by analyzing third-party risk from the perspective of **Probability Theory**. Third-party data breaches are random events and **Probability Theory** is a branch of mathematics that provides the framework for understanding random events.

Generally speaking there are two kinds of outcomes that we try to predict with **Probability Theory**: *binary outcomes* and *count outcomes*. **Binary outcomes** happen once, and the calculated **probability** will be a number between zero and one, often expressed as a percent chance of the event occurring. An example binary outcome might be rolling a "one" with the roll of a die. If this is a fair die, the probability is one-in-six or 16.7% ($1 \div 6 = 0.167 = 16.7\%$).

Count outcomes are the other kind of outcomes we try to predict. An example **count outcome** might be the total number of "ones" we obtain rolling multiple dice at once over many rolls. Imagine that we roll twelve dice at the same time. For any particular roll, we can obtain from zero to 12 "ones" across the twelve dice, but over many rolls, we will obtain two "ones" on average. In probability theory, this is called the **expected value**. We can calculate the **expected value** by summing the probabilities of each individual die. This property is called the ***linearity of expectation***. In the case of rolling twelve dice at once, over many rolls, the **expected value**, *E*, is:

$$E = \sum_{i=1}^{12} P_i = \sum_{1}^{12} 0.167 = 2$$

Where P_i is the probability of rolling a "one" with die *i*, and the symbol $\sum$ is a mathematician's shorthand for adding all of the P_i values together ($P_1 + P_2 + P_3 \ldots P_{12}$). If all dice are fair, this probability is 16.7% for each die. The calculation finds an expected value of 2, which means that over many rolls, we should expect to roll two "ones" on average.

Third-Party Breaches are Count Outcomes

Third-party data breaches can be treated as **count outcomes**, with each third party having some non-zero probability for causing a third-party data breach. We can calculate an **expected value** in the same way we did with the dice. Equation-1 below calculates the **expected value** E_{cum}[13] for a third-party data breach as the sum of the individual P_i probabilities for each third-party *i* over the number of third parties *N*,.

Equation 1 requires knowing the probability for **each** third party to cause a third-party data breach. But the **expected value** can also be calculated using Equation 4 which only requires knowing the **average** probability for a third party to cause a third-party data breach. Figure 3 shows how Equation 4 can be derived from Equation 1. It is important to remember that Equation 4 is an algebraic simplification and is therefore mathematically equivalent to Equation 1. All of the individual P_i values are still present, and captured in the value for P_{ave}. In the next section we will explain the insights that this simplification reveals about third-party risk.

---

[13] We have added *cum* just to remind you that probabilities are cumulative when calculating **expected value**

$$E_{cum} = \sum_{i=1}^{N} P_i = \frac{N}{N} \times \sum_{i=1}^{N} P_i = N \times \left(\frac{\sum_{i=1}^{N} P_i}{N} \right) = N \times P_{ave}$$

<div align="center">Equation 1   Equation 2   Equation 3   Equation 4</div>

Figure 3, The derivation of Equation 4, from Equation 1.

Equation 1 calculates the **expected value** for third-party data breach using a fundamental property of probability theory: **Linearity of Expectation**, where E_{cum} is the expected cumulative value for a reportable third-party data breach, N is the number of third-parties and P_i is the probability for an individual third-party, i. Following are the derivation steps:
- Equation 1 can be multiplied and divided by N to obtain Equation 2,
- Equation 2 can be reorganized to obtain Equation 3, where the expression in brackets is the definition of the average probability across N third parties.
- Equation 4 substitutes P_{ave} for the expression in brackets.

Third-Party Breach is a Systemic Risk

In this section we will analyze Equation 4 for what it shows us about the nature of third-party risk and the effectiveness of the current approach for managing this risk. To simplify our discussion, we will assume that we are speaking only about third parties that could expose very large amounts of data and that risk is therefore proportional to the probability.

The current approach is to only perform *due diligence* on third parties—individually. *Due diligence* consists of a **risk assessment**, for example a questionnaire or an attestation report, followed by some kind of risk mitigation performed on the individual third party.

After mitigation is performed, cybersecurity practitioners like to say that some nonzero **residual risk**[14] remains. Let's refer to this residual risk as **residual probability** since most mitigation is about reducing probability. This **residual probability** is represented mathematically by the individual P_i values in Equation 1, which are averaged together to calculate the value for P_{ave} in Equation 4.

P_{ave} therefore reflects the **average residual probability** (or average residual risk) across all of the **individual** third parties that have been assessed. The **Central Limit Theorem** and *the **Law of Large***

---

[14] Some cybersecurity practitioners may regard controls as being 100% effective. Indeed, the average residual probability is very very small (0.066%, see Estimating P_{ave} below)—but it is not zero. It is this tiny residual probability that builds up over tens to hundreds of third-parties, as we explain below.

Numbers[15] tells us that P_{ave} should rapidly converge to a **constant value** as more third parties are added to the organization.

But Equation 4 shows us that the overall risk from third parties does not converge as P_{ave} converges. Instead there is an additional **systemic risk** that arises from the number of third parties N. **Systemic risks** arise from the complexity of the systems and in this case the complexity comes from the sheer number of third parties. Or, as viewed from the mathematically equivalent Equation 1, systemic risk arises from the **accumulation of individual probabilities**. As more third parties are added, the probability (or Expected value E_{cum}) continues to increase as N increases.

Let's conduct a thought experiment to better understand this **systemic risk**. Imagine that we are adding third parties one by one to an organization. We are sharing large amounts of unencrypted and unobfuscated data with each of the third parties and only third parties which are vetted through our **due diligence** will be added. Imagine we somehow know the **residual probability** P_i of each third party for causing a third-party data breach after our **due diligence** and that we can calculate P_{ave} from the individual P_i values as we add each third party to our organization.

With our first third party, P_{ave} is simply the **residual probability** after our **due diligence** which is represented by P_1 . The value for E_{cum} which is our probability is therefore 1 x P_1. As we add our second third party, P_{ave} becomes an average of the **residual probabilities** P_1 and P_2, but our E_{cum} now becomes two times as large as the average of P_1 and P_2. As we add more third parties, we know from the **Central Limit Theorem** and the **Law of Large Numbers** that P_{ave} will rapidly converge to an unchanging value, because it is an average of random numbers. But Equation 4 also shows us that the probability continues to increase with each new third party. Because P_{ave} remains unchanged, the formula essentially becomes N times a *constant*. As a result, the risk doubles as the number of third parties increases from ten to twenty and doubles again as the number of third parties increases from twenty to forty. The following examples use the $\overline{P}_{ave}$ value from Figure 5 below to demonstrate the increase in risk.

---

[15] In Probability Theory, the Central Limit Theorem (CLT) tells us that we should expect the values for P_i to follow a normal or log normal distribution (bell-curve), exactly like the curves that we find empirically and which are shown in Figure 5 below. The **Law of Large Numbers** states that as we add P_i to our average, this average should converge to the mean value of our normal or log normal distribution and remain unchanged. Even if *due diligence* is very effective, the individual P_i values (although very small) will not be zero, and we should still expect a P_{ave} value that should converge and remain unchanged as an organization adds more third parties.

Emergent risk for third-party breach

The **emergent risk** from the **number** of third parties

Average **residual risk** after *due diligence.* Converges to a constant.

$$E_{cum} = N \times P_{ave}$$

Risk increases in proportion to the **number** of third parties

E_{cum} <u>doubles</u> as the number of third parties increases from ten to twenty:

- **10** x 0.066% = **0.66%**
- **20** x 0.066% = **1.3%**

↓ *double*

E_{cum} <u>doubles again</u> as the number of third parties increases from twenty to forty:

- **20** x 0.066% = **1.3%**
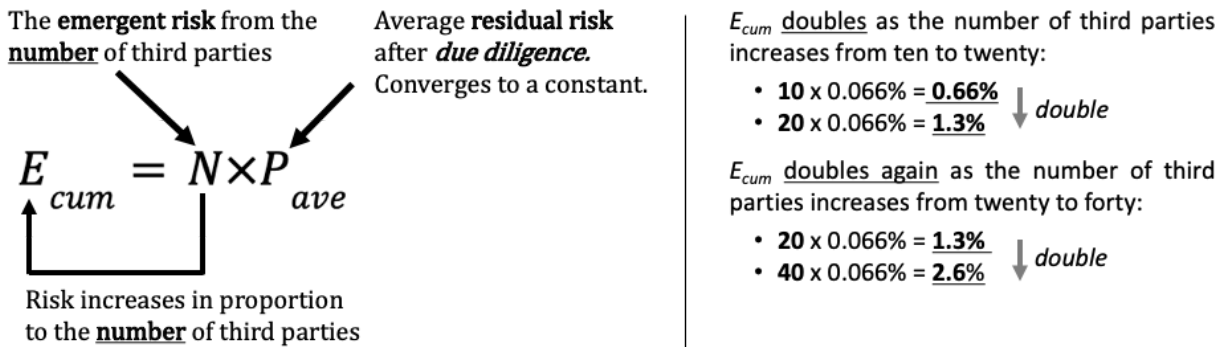- **40** x 0.066% = **2.6%**

↓ *double*

Figure 4, Equation 4 annotated.

Therefore Equation 4 shows us that, by itself, **due diligence** of **individual third parties** fails to address the **emerging risk**. The emerging risk is many times larger than the **average residual risk** from individual third parties. One can see from the [Reference Table of Probabilities](#) section, that probability begins to become a significant business concern as the number of third parties with large amounts of data exceeds **one hundred** (7% or once in 14-years with a fifty-fifty chance in 7-years) and that a large breach is nearly assured when the number of third parties exceeds **two hundred** (14% or once in 7-years with a fifty-fifty chance in 3.5-years).

Probability versus Expected Value

In cybersecurity, we say that *risk* is the product of *impact* and *probability*. In this paper, *probability* will be determined objectively by using **Probability Theory**. The *probabilities* and *frequency* that we will calculate will be in real world units and will help business leaders, partners, and the public understand how often they should expect a data breach. *Risk per-se*, will not be calculated. Instead risk will be addressed by organizing third parties into groups based upon the amount of data they can expose (see *[Data Organization](#)*) then calculating *probabilities* for these groups.

In **Probability Theory** the term *probability* has a precise meaning and readers that know **Probability Theory** will object to the misuse of this term. As explained in the section *[Third-Party Breaches are Count Outcomes,](#)* the correct term to use instead of *probability* is *expected value.* In this section we will explain the difference between *probability* and the more correct term *expected value*. For cybersecurity people the difference is not important, at the same time they should not confuse the term *expected value* with the term *risk* or *impact*.

In **Probability Theory**, *probability* is a number between zero and 1. In other words, it can never be greater than 100%. In the case of third-party data breaches, because an organization can experience multiple third-party data breaches within a year (i.e. more than 100%), we should use the term **expected value** instead of **probability**. The difference between the value for **probability** and the **expected value** is accounted for by outcomes that overlap or are not disjoint. In the case of third-party data breaches, the overlapping outcomes are the multiple third-party data breaches which occasionally occur in a single year for an organization. When **expected value** is small, for example below 10%, overlapping events are rare and the difference between **expected value** and **probability** is immeasurably small. In fact, the difference is likely much smaller than the error from accurately determining the number of third parties that can actually expose data. When **expected value** is large, for example 50%, it is likely that an organization will occasionally experience more than one third-party data breach within a given year and the difference between **probability** and **expected value** will be significant. When **expected value** is large, it is recommended to use the term **expected value** rather than **probability** when presenting results to an organization's business leadership. Some in business leadership may have studied **Probability Theory** and understand that multiple events can happen within a single year.

Estimating P_{ave}

If one can know the individual probabilities for each third party as in Equation 1, then plotting these probabilities as if they were *scores on an exam*, can bring insights. This is called a *probability distribution* and we have created such plots in Figure 5, using a regression model that is based upon the InfoSec and IT Audit staffing levels within the third parties[16]. The curves in Figure 5 look like the *bell curves* that one would find with exam scores. Figure 5 shows these curves for a range of organizations of different sizes, across multiple industries and countries, with varying numbers of third-parties, and overseen by a variety of regulatory agencies from multiple countries. Surprisingly, the curves overlap despite the large range in differences between organizations. This overlap allows us to make some general observations.

One observation is that the average annual probability for third-parties is the same across multiple third-party lists: $\overline{P}_{ave}$ = 0.066% (1,515-years[17]) with a standard deviation of just 0.027%. Note that we have used the notation $\overline{P}_{ave}$ (read as *P-average-bar*[18]) to indicate the average of P_{ave} values across

---

[16] The model was based upon six factors including a number of employees with certain cybersecurity, audit and IT certifications. See VivoSecurity Inc. for more details.
[17] Frequency can be calculated by inverting probability: 1 ÷ 0.00066 = 1,515.
[18] Bar is the horizontal line drawn above the *P* which in statistics indicates an average

probability distributions, and the standard deviation is the deviation between P_{ave} values. The similarity of the curves in this graph, despite the large differences in organizations, suggests most organizations can use $\overline{P}_{ave}$ together with Equation 4 to calculate their probability for experiencing a third-party data breach.

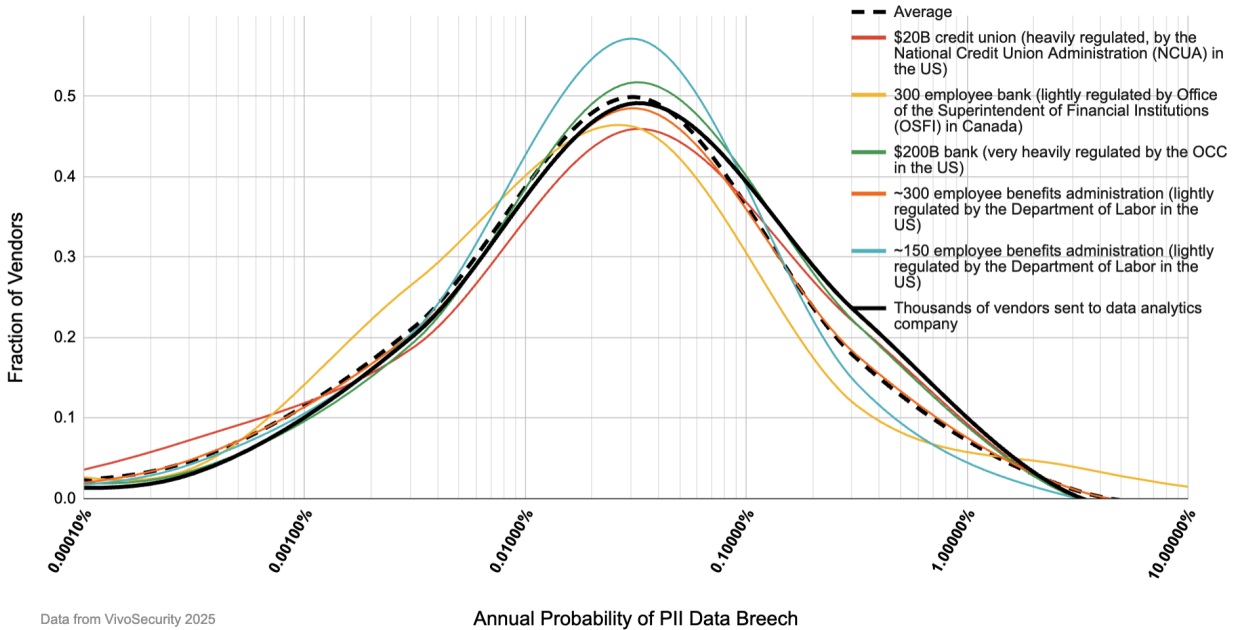Data from VivoSecurity 2025   Annual Probability of PII Data Breech

Figure 5, Third-party data breach probability distributions for six organizations
These organizations were chosen to represent a large range in industries, countries, company sizes, numbers of third parties and diverse regulatory regimes.
- The horizontal axis indicates a third party's annual probability for causing a third-party data breach.
- The vertical axis indicates the fraction of third parties for each representative organization's population of third parties.
- Each curve indicates the distributions of annual probabilities for third parties for the indicated organization.

The graph shows that curves follow a log-normal distribution (or a bell-curve, similar to the distribution of grades on an exam) over a large range in probabilities from once in one million (0.0001%) to once in one hundred (1.0%). Most interesting is that the curves overlap, with similar medians and averages P_{ave}, despite representing such a large range in organizations. We find an average annual P_{ave} of 0.066% with a standard deviation of 0.027%, which we indicate with the symbol $\overline{P}_{ave}$ and which was calculated by calculating a P_{ave} for each organization's curve, then averaging the P_{ave} values together. The standard deviation is a measure of the variability among the P_{ave} values. With such a large range of organizations represented, it is likely that the third-party probability distribution for your organization is similar, with a value for P_{ave} within the standard deviation. Data is from VivoSecurity, 2025.

Testing P_{ave} with Your Enterprise

Estimating a value for P_{ave} does not require an accurate regression model such as the one we used to generate the curves in Figure 5. Some organizations have enough third parties that they can

independently calculate a value for P_{ave} based upon a history of third-party breaches for their organization. There is value in performing this calculation for your organization since it provides an independent check and gives management additional confidence in results from Equation 4.

We recommend using the largest pool of third parties that must report to your organization when they have experienced a breach that exposed your data. The following is an example calculation of P_{ave}.

How to Estimate P_{ave} from Past Data Breaches

An organization has 600 third parties that have some amount of the organization's data and that are contractually required to report a breach of the organization's data. The organization has experienced two third-party breaches over the past 7-years during which time they have had 600 third parties. Solving Equation 4 for P_{ave}, one obtains $P_{ave} = E_{cum}/N$. Note that E_{cum} is 2 breaches in 7-years (28.6%):

P_{ave} = (2 breaches / 7 years) / 600 vendors = 0.047%

Note that with this example, the result is within a standard deviation of P_{ave} value we report in Figure 5.

Even organizations that have not experienced a third-party data breach can make observations about the magnitude of the value for P_{ave}. In the following example, an organization can at least set an upper limit for the value of P_{ave}.

How to Estimate P_{ave} with No Past Data Breaches

An organization has 100 third parties that have some amount of the organization's data and that are contractually required to report a breach of the organization's data. The organization has not experienced any third-party breaches over the past 10-years during which they have had 100 third parties. In the following calculation, we will calculate a *pseudo* P_{ave} assuming that one breach did occur during the ten-years and we will know that the actual P_{ave} is likely below this value.

$P_{ave} < pseudo\ P_{ave}$ = (1 breaches / 10 years) / 100 vendors = 0.1%

Answers to Frequent Questions and Comments

Our organization already performs *due diligence*, why do I need to calculate expected frequency?

Due diligence focuses on the individual third party. The expected frequency for a third-party breach comes from the number of third parties (see [Third-Party Breach is a systemic Risk](#)) and there is nothing intrinsic about *due diligence* that limits the number of third parties.

The risk for a third party breach is a weak-link problem. I am culling this weak-link via due diligence.

For each third party that has passed your scrutiny, the probability may be very small but it is not zero. In fact, the $\overline{P}_{ave}$ that we find is a very small 0.066% (see [*Estimating P_{ave}*](#)) or a once in 1,500-year expected frequency. Trained cybersecurity practitioners understand this and use the term *residual risk* for this tiny probability. But probabilities add (see Equation-1) and at [some point this tiny residual risk will become a problem](#). If you don't measure this accumulation of risk you will not know when you have arrived at that point.

Risk is the product of probability and impact. I don't see impact in Equation 4.

Equation-4 only calculates expected value (or probability when expected value is small). One can address risk by applying Equation-4 to subsets of third parties based upon the amount of data they could expose. See the section [*Data Organization*](#).

Equation 4, which is $E_{cum} = N$ x P_{ave} does not reflect the mitigation I have performed on a third party.

Equation 4 was derived from <u>and is equivalent</u> to Equation 1, which is a sum of individual probabilities. Equation 4 is telling you that your mitigation efforts become [averaged together and are captured in the value for P_{ave}](#). We find the value for [P_{ave} to be very small, so your efforts are good](#), but equation 4 is also telling you that the number of third parties makes up a large portion of the risk.

Why should I consider the number of third parties since I have no control over this?

Because Equation-4 which is $E_{cum} = N$ x P_{ave} shows us that the number of third parties is a major part of the risk. Even if you cannot manage this risk, [business leadership is depending on you to honestly report the risk](#).

My organization is already aware that risk is from the number of third parties. Why calculate it?

Failing to calculate this risk hurts your [organization's competitiveness](#). Without measuring the risk, business leaders will not know when they have used too many third parties or if they can use more. A thinking attributed to Peter Drucker: If you cannot measure it, you cannot manage it.

I don't want to know the probability since I cannot do anything about it.

Actually, there are many ways to manage cumulative third-party data breach risk, once you have measured it, including encrypting data, obfuscating data, reducing data, consolidating third parties, ensuring data is securely purged post engagement.

Glossary

Achievable Frequency   A data breach frequency that can be achieved with a cost that is acceptable to the business. The cost is in terms of a limitation on the use of third parties, the cost to encrypt or obfuscate data and the cost of ensuring that data is purged post-engagement with a third party.

Acceptable Frequency   A data breach frequency that is acceptable to an organization's customers, with the understanding that risk also accumulates for your customers (see Judge Acceptability).

Count outcomes   A predicted outcome that we predict using probability theory. Expected Value Predictions for count outcomes. Expected values are similar to probability outcomes when expected values are small.

Cumulative Risk   The risk from sharing data with a large number of third parties. We know from the Linearity of Expectations that probabilities add.

Systemic Risk   A risk that comes about from the complexity of a system. In the case of third-party data breach risk, this is the cumulative risk from sharing data with a large number of third parties.

Frequency   The average frequency in years between data breaches. Calculated as the inverse of Expected Value.

Nonpublic PII data   Personally Identifiable Information that is not public information and that would therefore trigger federal and state reporting requirements. This might include PHI (protected health information), CHD (card holder data) and PFI (protected financial information).

Probability   Predictions for one time events. A number from zero to one. When probabilities are small, they have values similar to expected values.

P_{ave} The average annual probability for a third-party to cause a third-party data breach within a list of third parties. We find this value to be similar across third-party lists.

$\overline{P}_{ave}$ The average of P_{ave} values across many third-party lists. We find a value of 0.066% with a standard deviation of 0.027%.

Reasonable Assurance   In this paper we mean a probability that is so low that an organization can assume that a breach will not happen. The authors regard a probability below 0.5% (200-years) as reasonably assured not to happen.

Third Party   An organization you do business with and which might cause a third-party data breach for your organization.

Third-Party data breach   A breach of your organization's data that is caused by a third party.

Disclaimer

It is important to understand that even with a small probability, for example, 1% or 100-years, data breaches still occur. Among 100 organizations all with a 1% probability for a third-party breach, there will be one organization every year on average that will experience a third-party breach.

Permission to Copy and Distribute

The authors give permission to copy, present, distribute and include this document in other reports, as long as the document is present in whole and not in part. The risk of a third-party data breach is one of the most significant cybersecurity risks for most organizations and it is our desire in writing this paper to help organizations recognize and manage this risk.