DOES PREVALENCE MATTER? RANKING ANTI-MALWARE PRODUCTS BY POTENTIAL VICTIM IMPACT

Holly Stewart & Joe Blackbird
Microsoft, USA

Peter Stelzhammer, Philippe Rödlach & Andreas Clementi
AV-Comparatives, Austria

Email hollyst@microsoft.com;
p.stelzhammer@av-comparatives.org

ABSTRACT

Most anti-malware tests count a 'miss' equally. If one sample out of 100 is missed, the score for that set is 99 per cent, regardless of the sample missed. But should all samples be treated equally? Should vendors receive a lower test score when they miss samples that have victimized more people? Should vendors receive an equal score if they miss the same number of low-prevalence samples, rather than the high-prevalence ones?

Even if you agree with the principle that not all misses are the same, how would you factor in polymorphism where a particular sample may impact only one victim, but the malware family impacts millions? How is a sample measured if there is no record of the sample or the family in the wild at all?

In this paper, we will take you through several prevalence-weighted models using real-world data from hundreds of millions of computers. We will show how the prevalence-weighted models compare to the standard method of scoring sample detection. We'll discuss each model's benefits, deficits, and the lessons learned along the way.

INTRODUCTION TO ANTI-MALWARE TEST SCORING MODELS

Most of today's malware file-detection tests follow a fairly standard methodology:

1. They select a set of samples that are representative of the kind of malware they want to test.
2. They test each product's detection capability against each sample.
3. They assign a score based on the percentage of files that was detected or missed.

So, if there are 10 samples in the test, and a vendor misses one of them, then they receive a score of 90%. Each sample is treated equally.

The premise behind prevalence-weighted testing is that not all samples are treated equally because some are more prevalent, and have the potential to impact more people than others. Geography is also an important consideration, because some malware might affect only one region of the world and might not be detected by every vendor if they don't have a strong customer base in that region. However, one can argue that all vendors should detect prevalent malware, and the more prevalent malware is, the more important it is for a vendor to detect it. So, the prevalence-weighted model is designed to factor the malware prevalence into a vendor's final test score.

Figure 1 shows a simplified model of just 10 samples. In the ecosystem (real world), the first three samples were much more prevalent than the other samples in the test. With a standard sample-centric score, each miss would represent 10% of the test. However, in the real world, 60% of people encountering malware were likely to encounter sample #1. Should sample #1 and sample #10 (affecting 1/6,566 or 0.015% of people) count the same in the test? The standard method of testing using the sample-weighted impact would treat them the same whereas a perfectly modelled prevalence-weighted test would score them according to their real-world impact.

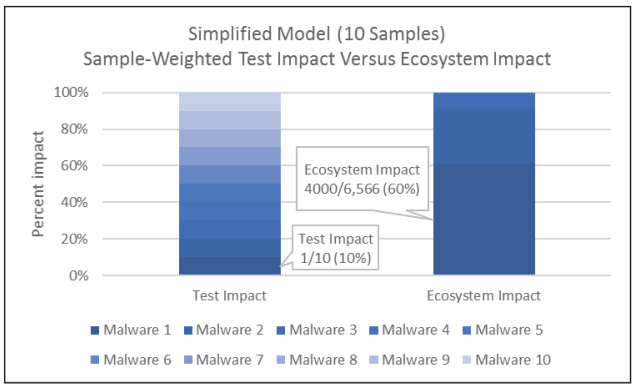

Figure 1: Simplified model.

To create a prevalence-weighted model, you must define what prevalence means and decide how to measure it. In this paper, prevalence is defined as the number of times a distinct computer encountered malware measured by a specific file (file prevalence) or any files related to a specific malware family (family prevalence). So, if many computers reported encountering a specific malware file, then that file is considered to have high prevalence. Similarly, if many computers reported encountering a specific malware family, then that family is considered to be a high prevalence family.

In a perfect test, the entire ecosystem of malware files and their prevalence would be known and available for testing. As Figure 2 shows, malware detection can be tested and scored directly to prevalence using the number of people that encountered that malware in the ecosystem during the test period.

In reality, there are hundreds of thousands of new pieces of malware, exploits and unwanted software appearing every day, and many of them are only seen at one time during their short lifespan. Although many of the malware families in the ecosystem are very prevalent and reuse some of the same malware code for infections, many other samples are polymorphic or targeted and will never be experienced by anyone other than the victim themselves.

Figure 3 shows how samples chosen for the March 2015 *AV-Comparatives* File-Detection Test [1] compare to the distribution of malware in the ecosystem. *AV-Comparatives* selected samples based on sample availability and family prevalence to match the ecosystem as closely as possible. Figure 3 shows how the traditional method of scoring tests (based on number of samples missed in this test set) compares to the actual distribution in the ecosystem.

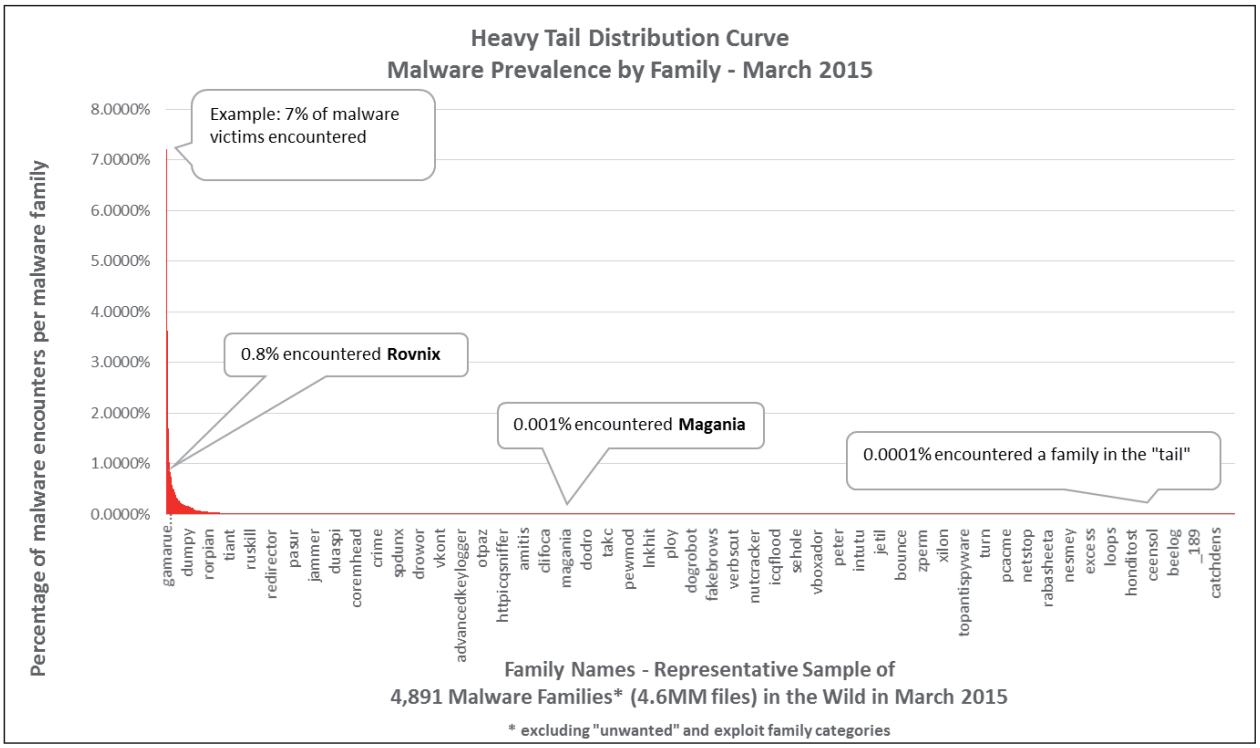

Figure 2: Heavy tail distribution curve.

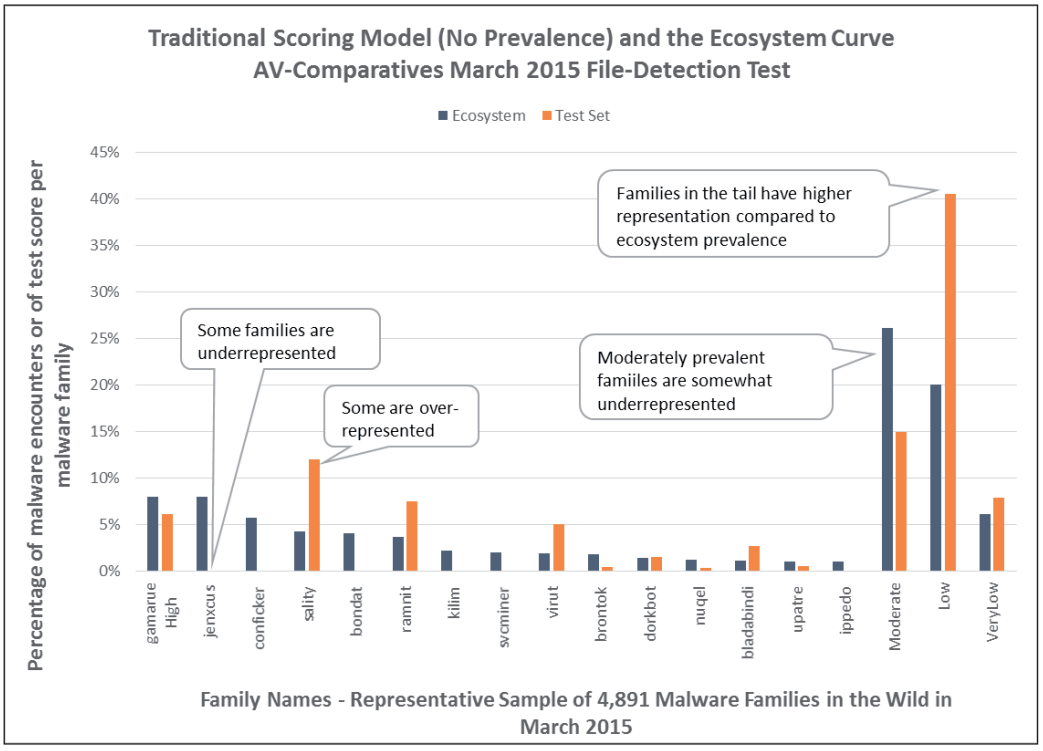

Figure 3: Traditional scoring model compared with the actual distribution in the ecosystem.

What complicates the sample selection criteria for the test set is the availability of recent samples that are PE (portable executable) files. These files must also belong to indisputable malware families. Indisputable malware families are those that must always match a vendor's detection criteria. Samples that fall into categories that may be disputed by anti-malware vendors are often referred to as unwanted, or potentially unwanted, and include families such as adware, software bundlers, etc.

The number of new samples available that match that criteria is limited. For example, although Jenxcus was one of the most prevalent families in March 2015, the most prevalent component was not a PE file, it was a script component. In fact, *Microsoft* only saw 12 new PE files for

Jenxcus leading up to the test. *AV-Comparatives* sourced more than that, but it demonstrates the scarcity of samples, especially for certain families that propagate through non-PE components.

In this imperfect world with a daily churn of hundreds of thousands of new malware files a day, combined with limited access to those files, testers must rely on sampled versions of the real world with a bias toward malware that lends itself to being collected and tested. This is no easy task. To solve this imperfect situation, as often happens in statistics, we must try to create a model that best represents the real world.

OVERVIEW OF PREVALENCE-WEIGHTED MODELS

In this paper, we have analysed four different prevalence-weighted models:

- A model that incorporates file prevalence only.
- Two models that incorporate file *and* family prevalence – one that favours family prevalence and another that favours file prevalence.
- A final model that incorporates file and family prevalence, and the position of the family in the ecosystem.

File prevalence model

The file prevalence model is straightforward. For all the files in the test, take the prevalence of the file in the ecosystem and use that prevalence to weight the test score.

Table 1 describes the benefits and drawbacks of the file prevalence model, while Figure 4 gives a comparison of model to the ecosystem.

Conclusion

Unless a tester can select files that perfectly represent the ecosystem in terms of prevalence and family distribution, a model that only calculates the prevalence of the tested files cannot represent the ecosystem fairly.

Benefits	Method of measurement is simple to explain.
Drawbacks	Many malware families have files that are only seen at one time in the wild (polymorphic, etc.), but in total are very prolific. Counting a sample from those families with a prevalence of one doesn't properly represent the family unless the tester perfectly selected the right number of polymorphic samples to match the prevalence of the family in the ecosystem.
	For polymorphic files, there might be no real telemetry on the file. Solution: count any files without telemetry as having an impact of one computer.

Table 1: Benefits and drawbacks of the file prevalence model.

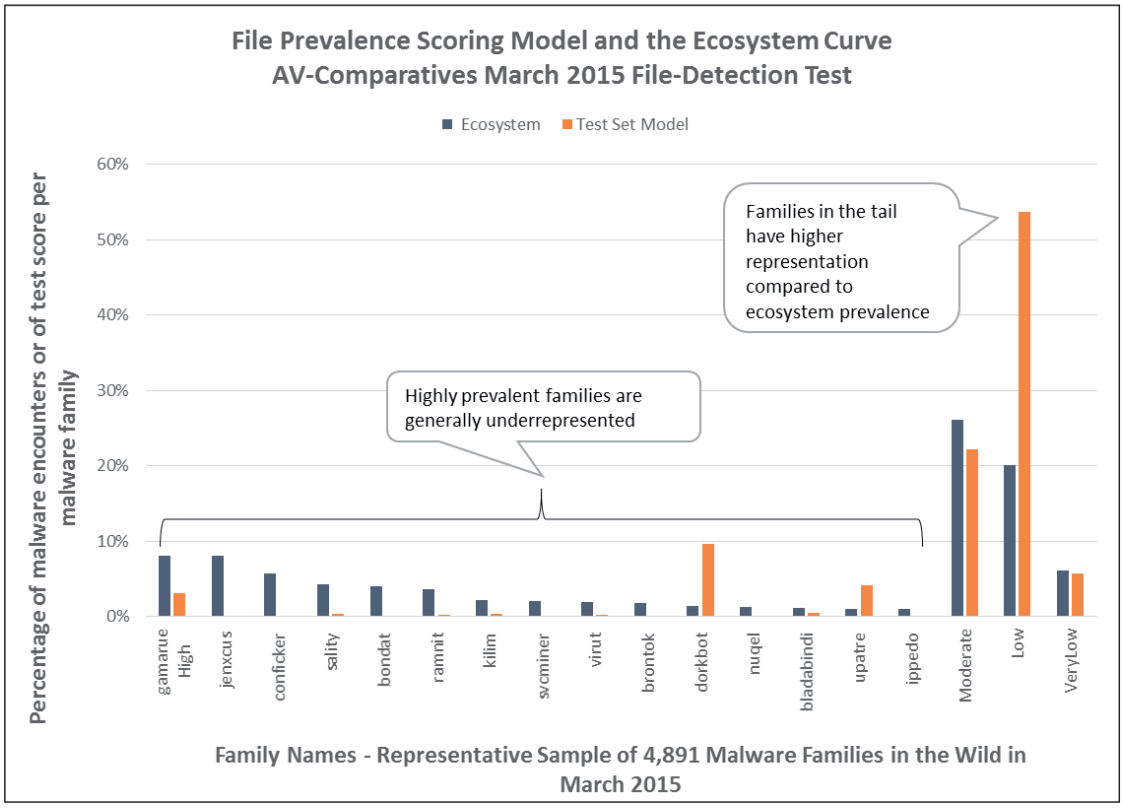

Figure 4: Comparison of file prevalence model to the ecosystem.

File and family prevalence – family priority

One way to incorporate family prevalence into a prevalence-weighted model is to use the samples tested in each family to equate to the prevalence of that family in the ecosystem.

For example, if Zeus (or Zbot) represented 20% of the number of computers reporting a malware encounter during the testing period, then the Zeus samples (no matter how many tested) should equate to 20% of the test score. This model force-fits all samples in the test set to the exact ecosystem prevalence of the family, which results in a perfect representation of the ecosystem as long as:

- All families, or a statistically significant selection of families are represented in the test.
- Families that are in the test have a statistically representative set of samples.

Table 2 lists the benefits and drawbacks of the family-weighted model, while Figure 5 shows a comparison of model to the ecosystem.

In some cases, testers may only have a small number of samples per family or may not have any samples to represent a family at all.

Conclusion

Unless a tester can select files that represent a statistically significant number of families and enough samples within each of those families to properly represent the family's prevalence distribution, this model will not represent the ecosystem fairly.

Benefits	Force-fits the misses to the real-world ecosystem curve.
Drawbacks	Some families reported by anti-malware engines are generics and not true malware families. Model correction: put generic families in a separate category equal to the average family prevalence associated with real malware families.
	Some families in the test set may have too few samples for adequate family representation. The tester must choose a perfect test set for every family in the ecosystem. Model correction: remove untested family categories, such as adware, bundlers, exploits, etc.

Table 2: Benefits and drawbacks of the family-weighted, family prevalence priority model.

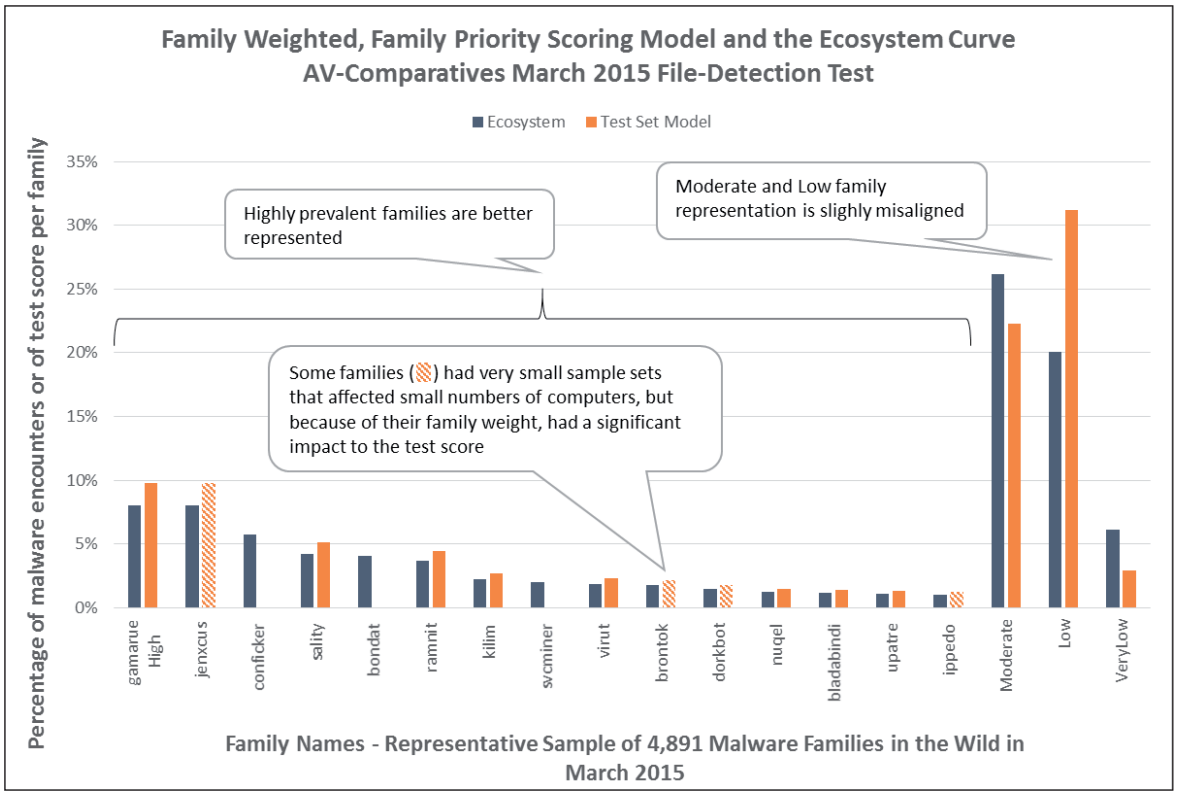

Figure 5: Comparison of family-weighted, family prevalence priority model to the ecosystem.

Benefits	Solves the issue of requiring a perfect test set for each family.
Drawbacks	The distribution has no forcing function to ensure it maps to the ecosystem. For example, if the tester only has a large number of files in tail (low prevalence) families, then the combined weight of many files selected from the tail can still override the misses related to the most prevalent families.

Table 3: Benefits and drawbacks of the family-weighted, file prevalence priority model.

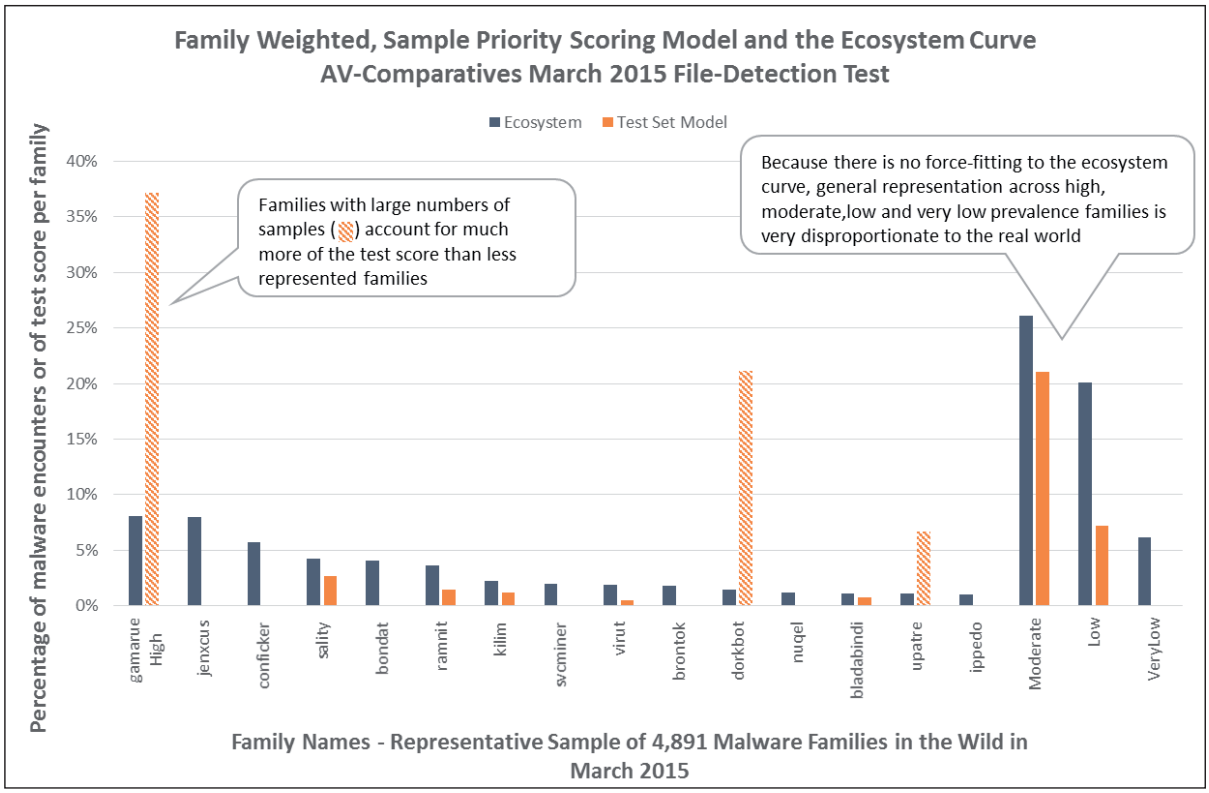

Figure 6: Comparison of family-weighted, file prevalence priority model to the ecosystem.

File and family prevalence – file prevalence priority

Rather than force-fitting samples in a test set to match the ecosystem prevalence of a family, another model uses the prevalence of each family to upgrade and downgrade the importance of a sample when calculating the miss score. So, for example, if Family A affected 1,000 computers and Family B affected 100, a sample from each family affecting equal numbers of computers, for this example, let's say 10 computers, would have differing weights in the test set. The sample from Family A would be 10 times as impactful to the score as the sample affecting the same number of computers from Family B.

Table 3 lists the benefits and drawbacks of the file prevalence model, while Figure 6 shows a comparison of model to the ecosystem.

Because the samples are not force-fitted to the ecosystem curve, the test score can be disproportionate if the samples for each family in the test are not proportional to their respective prevalence or if the prevalence of the samples selected are not representative of the real world.

Conclusion

Unless a tester can select files that perfectly represent high, moderate, and low prevalence families, this model will not represent the ecosystem fairly because it does not force-fit the sample selection to the prevalence of the ecosystem.

Hybrid model

The hybrid model incorporates file and family prevalence in addition to the position of the family in the ecosystem.

If we combine the flexibility of the family file prevalence methodology with another method that force-fits the sample selection to the ecosystem, then you achieve the best of both worlds. The tester has the freedom to select some or many samples from a representative set of families, but at the same time, you ensure that the resulting score will match the prevalence of malware in the ecosystem.

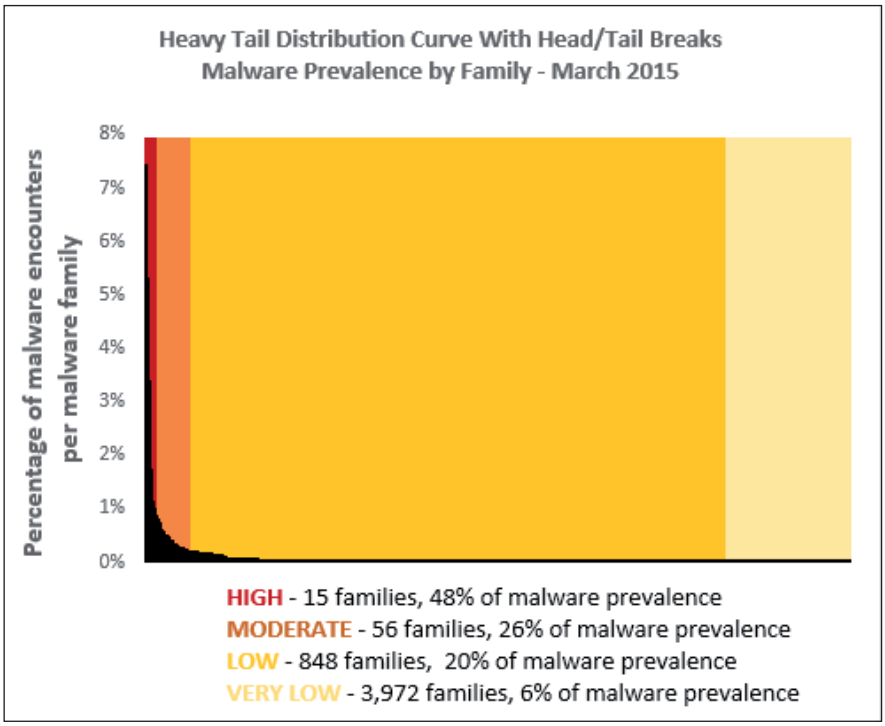

Figure 7: The number of families from March 2015 that were distributed into the partitions.

Benefits	Force fits a collection of families to the ecosystem curve. Doesn't require perfect sample selection for every family.
Drawbacks	Complicated to explain and calculate.

Table 4: Benefits and drawbacks of the hybrid model.

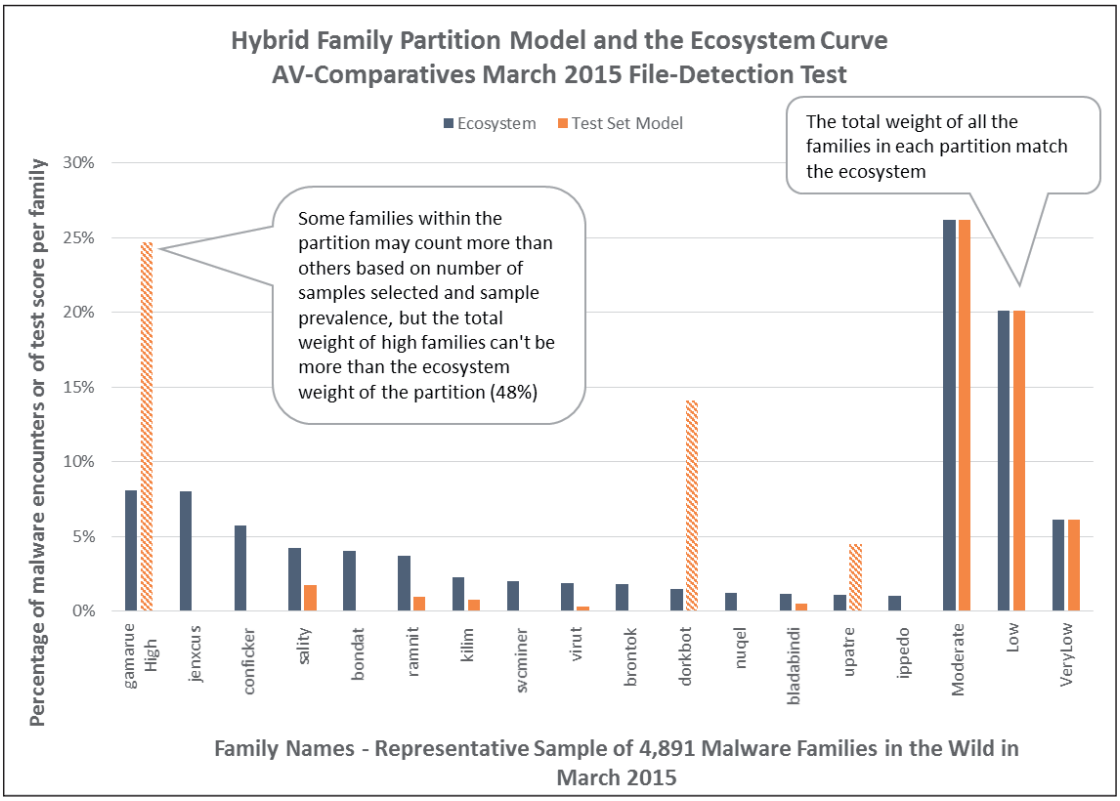

Figure 8: Comparison of hybrid model to the ecosystem.

Instead of force-fitting to the really narrow category of malware family, we investigated a method of creating partitions of families based on their prevalence. The standard method we chose was the head/tail breaks method [2], which is useful for partitioning heavy-tailed distributions, such as the prevalence distribution of malware families. The method partitions a distribution by taking the average, in this case, of family prevalence to create a head and a tail partition, then we continue splitting the head of the distribution to create four prevalence partitions: high, moderate, low, and very low. Figure 7 shows how many families from March 2015 were distributed into these partitions.

Table 4 lists the benefits and drawbacks of the hybrid model, while Figure 8 shows a comparison of the model to the ecosystem.

After separating the families into their respective partitions, we use the family and sample prevalence related to each file to raise and lower the relative importance of a miss within each partition, but force-fit the total impact of the misses into partitions representing high, moderate, low and very low (tail) sections for the final test score calculation.

In this model, the samples are force-fitted to the ecosystem curve in each partition. The tester has the freedom to choose a selection of families and high and low prevalent samples within those families. The one constraint is that the tester needs to choose a statistically significant number of families and samples within the partition, but this is much easier to achieve for a partition than for all families within the ecosystem as in the family prevalence priority model.

Conclusion

Although no model is perfect, this model provides a means to allow some flexibility in choosing samples for a test. It also ensures that the resulting test score fits to the real-world prevalence in the ecosystem.

RESULTS

Score comparisons

We compared the test scores of 17 of the vendors in the *AV-Comparatives* March 2015 File Detection test [1]. Five out of 17 vendors moved up or down by three or more places. Three of them had significantly lower test scores and two had significantly higher test scores.

Table 5 compares the two test rankings according to the model used, and shows the highest and lowest scores in the test. The 'Movement' column shows places gained or lost (in parentheses) if the alternative model is used.

Table 5 shows that the difference between the highest and lowest scores is much smaller using the alternative model. However, the top three products in the traditional model were still amongst the top five in the prevalence model, and the bottom three similarly in the bottom five. For most vendors, there is a high correlation between the ranking in one model and the ranking in the other.

The two vendors that did significantly better using the alternative model had a high number of misses in low and very low families. When the prevalence of the families and samples was factored in, their scores increased considerably.

	Vendor ranking – traditional model	Vendor ranking – prevalence model	Movement
	1	1	-
	2	2	-
	3	5	(2)
	4	8	(4)
	5	3	2
	6	7	(1)
	7	11	(4)
	8	4	4
	9	10	(1)
	10	6	4
	11	9	2
	12	14	(2)
	13	12	1
	14	17	(3)
	15	13	2
	16	15	1
	17	16	1
Highest	99.96%	99.99%	
Lowest	86.26%	98.83%	

Table 5: Comparison of the two test rankings according to the model used.

	Global vendor ranking and regional detection score																
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Brazil																	
Canada																	
China																	
Colombia																	
Egypt																	
France																	
Germany																	
India																	
Indonesia																	
Italy																	
Korea																	
Mexico																	
Pakistan																	
Philippines																	
Russia																	
Spain																	
Thailand																	
Turkey																	
Ukraine																	
UK																	
US																	
Vietnam																	

Table 6: Global vendor ranking and regional detection score.

The three vendors that did significantly worse using the alternative model all had more misses than their formerly closely ranked peers in highly prevalent families, such as Kilim, Gamarue, Jenxcus and Bladabindi. These misses drove their test scores down.

Another point of comparison is that, in general, the test scores for the prevalence-weighted model were higher than those for the sample-weighted model with less diversity (vendor scores were closer together). For a tester, this result might make it more difficult to highlight a distinction between detection quality.

Country score comparisons

The prevalence-weighted model is especially relevant when comparing detection rates for specific geolocations (geolocation refers to the client/potential victim's PC). To calculate the score, the family and partition weights were assessed for each country/region and applied to each vendor's misses. The vendors who scored highly in the overall prevalence-weighted test (using global numbers) also did well in most regions, and vendors scoring at the bottom of the test also did worse than most other vendors in localized regions.

However, some vendors in the top five had very localized results – performing very well in certain regions and not so well in others, such as Brazil, China, Columbia, Egypt and Korea. Other vendors performing in the middle of the global test had especially high scores for certain regions, such as Canada, Indonesia, Russia, Ukraine and the US. These differences indicate that there is some market bias for detection. Table 6 shows global vendor ranking and regional detection scores.

LESSONS LEARNED AND NEXT STEPS

Modelling the ecosystem based on incomplete sample sets is not straightforward. The comparison of the models has shown that file prevalence is not useful without the context of family prevalence and other malware families in the ecosystem. A model that both allows some flexibility in the sample selection and fits the samples in the test set to the ecosystem curve is the most accurate model we've identified so far and shows very significant differences between vendors especially when locality prevalence is factored into the score.

Reliance on a single vendor's telemetry, especially if that telemetry was localized or from a small customer base, can skew the results for one or more vendors in a test. To really make this model work and avoid bias, the industry needs more vendors to submit telemetry data and that vendor telemetry data needs to use consistent reporting methodology (distinct machines, family prevalence, common timeframes, and locality-specific data).

The Anti-Malware Testing Standards Organization [3] Real-Time Threat List (RTTL) initiative is working towards a

common platform for sharing telemetry on files, which could be expanded to encompass this additional context needed for a prevalence-weighted test. Vendors should be given incentive to share because sharing their telemetry (like their samples) will help ensure that files and families affecting their customers are properly represented, thus improving their own test score.

If we can expand the RTTL to provide this additional context and encourage vendors to share, we can work towards a better model that ensures testers have a balanced telemetry set from which to select samples and calculate test scores. This new data will result in more accurate test scores and a more informed public that can make better choices about protection products.

REFERENCES

[1] http://www.av-comparatives.org/wp-content/uploads/2015/04/avc_fdt_201503_en.pdf.

[2] Wikipedia. http://en.wikipedia.org/wiki/Head/tail_Breaks.

[3] http://www.amtso.org/.

[4] Clementi, A.; Stelzhammer, P.; Colon Osorio, F. C. Global and local prevalence weighting of missed attack sample impacts for endpoint security product comparative detection testing. MALWARE 2014: 35–42.