# MessageLabs Accuracy Project

A Report by Professor R A Walton,
Information Security Group,
Royal Holloway, University of London

6th July 2009

Royal Holloway
University of London

Contents



References


[1] Accuracy Project Overview, MessageLabs, 20 April 2009
[2] Commercial Product Comparison Design, MessageLabs, 26 February 2009
[3] Commercial Product Comparison Specification, MessageLabs, 26 February 2009
[4] Commercial Product Comparison Results, MessageLabs, 20 April 2009
[5] Competitor Comparison Design, MessageLabs, 16 April 2009
[6] Competitor Comparison Specification, MessageLabs, 16 April 2009
[7] Competitor Comparison Results, MessageLabs, 29 April 2009
[8] Information Security Breaches Survey 2008 (ISBS08)


## A Introduction

1. The accuracy project [1] has been undertaken by MessageLabs to compare the security risk reduction capabilities of its managed e-mail service compared to alternative antivirus (AV) solutions available in the marketplace. This report is concerned with two such comparisons, the commercial product comparison, [2], [3], [4] and the competitor comparison, [5], [6], [7]. The purpose of the report is to give an independent opinion of the validity of the design of the experiments, the models and assumptions underlying the analysis, the analysis of the results, and the conclusions drawn. The author has not examined the identification of any specific e-mail messages as clean or as containing malware, but accepts the statistics provided by the experiments at face value.

2. This examination has been conducted on the basis of the written reports [1-7], oral descriptions and explanations by MessageLabs staff and their consultant, John Leach, the table of results as provided in Excel spreadsheets and specific responses to questions, comment and requests for additional available data during the course of the examination. An additional level of quality control of this examination has been provided by Professor K Paterson of the Information Security Group at Royal Holloway.

3. The experiments were concerned with actual malware that is capable of causing harm and being spread through e-mail. The experiments counted putative instances of e-mails containing malware sorting them into those that have been blocked by both MessageLabs and the alternative AV solution, (commercial product or competitor), those that have been blocked only by the MessageLabs service,

those that have been blocked only by the alternative solution[1], and those that have been missed by all solutions[2]. As well as blocking malware, the products and services also block other unwanted e-mail such as Spam, Phishing e-mails and damaged malware[3]. All such e-mails were omitted from the malware counts because this project is solely concerned with the risks of a customer suffering a directly harmful security incident arising from e-mail borne malware. The e-mails counted were those containing harmful malicious code (hereafter referred to as HMC). In this paper an e-mail containing HMC may be described as infected, similarly a customer receiving such an e-mail will be regarded as having suffered an infection (for example this language is used in paragraph 12).

4. There is a number of possible sources of HMC from which an organisation might suffer an infection of HMC. e-mail is not the only method by which HMC is spread. However e-mail borne infection is the most likely danger for a security-aware organisation and is the one requiring specialist processing to manage the risks while retaining the ability to conduct normal business. Other than e-mail, the main sources of infection are:

- Web browsing on infected sites
- Direct connection by hostile computer (e.g. Bob Morris Jnr's infamous internet worm)
- Download (e.g. over the internet)
- Upload (e.g. from portable media).

The incidence of the second of these is generally lower than for infected e-mails, and is more easily controlled by perimeter defences such as firewalls. The remaining three can all be addressed through policy and good management practices. This leaves e-mail as the one source of infection that is allowed access to an organisation's computers as part of necessary business and thus is likely to be the overwhelming source of infection leading to major security incidents.

5. The data collected in the experiments was analysed by MessageLabs and calculations made of the probability of various typical organisations receiving HMC infected e-mail, and the expected number of such e-mails in the cases where they are customers of the services being compared. Further calculations were then made to estimate the probabilities of suffering a major incident and the expected cost of such incidents. These costs then provide a measure of the relative risks to which the various organisations are exposed. The risk calculations are based on estimates of the probability that an e-mail received by a customer (having passed through the filtering products or service) is infected and the assumption that this may be fairly represented by the ratio of the number of infected e-mails passed by the product or service to the total number of e-mails passed by the service during the experiment. The reasonableness of such estimates depends both on the stability of the

[1] This only applies in the case of the Competitor comparison because use of the commercial products forms part of the MessageLabs service.

[2] This can only be estimated since by definition if MessageLabs had been able to identify such e-mails they would have been blocked.

[3] i.e. attempts at including malicious code which for some reason would not actually work and thus not have the capability of causing damage.

distribution of infected e-mails and on experiment covering a sufficiently large and representative sample of e-mails. This will be discussed in greater detail in section C, paragraphs 16-18.

6. The nature of the experimental data is such that the results should be valid for the risks faced by most organisations arising from generic e-mail borne malware. The conclusions may not apply to risks arising from specifically crafted malware delivered in the highly targeted way that might apply to that small number of organisations that are especially attractive to well-resourced and highly capable hostile entities[4].

## B Summary of Conclusions

7. This section presents the conclusions of this examination in terms of the purpose as stated in paragraph 1. More detailed discussion of specific points is undertaken in later sections of this report.

8. Experimental Design: the design of both experiments was appropriate for the purposes required. The commercial product comparison is the more straightforward and the suitability of the experimental design is clear in this case. The assumption is made in [2], section 7.2 that no user of a single commercial AV product could possibly achieve greater effectiveness than that of MessageLabs' optimised combination of two major products. As stated, this is unprovable. However, it is a fact that the two products used are individually market leaders and the combination used will outperform either of them individually and provides an example of best practice application of commercial products. The probability estimates are all snapshots of a specific period and the specific values will not necessarily be the same if calculated from observations from different periods. The estimate for the commercial products based on the 14 day period used in the experiment is reasonably settled within the period and agrees with the long-term average over several months. However there is considerable variation from month to month over the longer term and although the estimate can be taken as typical, and the comparative position appears sound, extrapolation beyond the period in question is only valid subject to the assumption that the calculated probabilities are typical. The competitor comparison is achieved by a different form of experiment, involving less direct measurement. Instead of using a source of genuine customer data the two services were given a more artificial and less voluminous[5] e-mail feed based on honeypot sources available to MessageLabs. The process then measured the relative performance of the two services (MessageLabs and the competitor) in the detection of HMC. The results were then converted into probabilities for the performance on genuine customer data by reference to the MessageLabs performance in the earlier experiment. This approach is valid and it does not matter that the Honeypot data might not be typical when considering the validity of the calculated probabilities. Of more direct concern

[4] The experiments were designed explicitly to exclude such targeted malware so that the results would be generally applicable to all organisations.

[5] The commercial product comparison involved a total e-mail feed averaging over 100 million e-mails per day with data from several months available, yielding a daily average of around 85,000 infected e-mails, whereas the honeypot delivered a total of about 20,000 infected e-mails over a 76 day period.

is that the shorter time-period and smaller sample means that there is less confidence that the calculated averages are truly representative. Although the relative performance between the two services seems clear. Ideally yet more samples should be collected to gain extra confidence in the specific results. See section C, paragraph 16 for more detail.

9. Effect of False Positives (FP): for any e-mail screening process, in the classification of the e-mails (HMC, spam etc) there is a trade-off between the performance in terms of false positives versus false negatives. It is therefore in theory possible that performance comparisons based on false negatives (as measured in these experiments) are skewed because of differing performance regarding false positives. Firstly for the commercial product comparison it might have been that detection thresholds were adjusted to ensure that doubtful e-mails were processed through the more sophisticated later processing of Skeptic. In fact this was not the case. The products were used directly as they would be by any customer and there was no scope for adjustments. Furthermore, for the most part, the products are based on signature detection methods and FP detections are random in nature. Obviously the FP rate for the MessageLabs service is higher than for the commercial products (since any FP from the commercial products will also be an FP for the full service) but the overall rate is within limits believed to be acceptable to customers and there is no evidence that the superior False Negative performance of the overall MessageLabs service is accounted for by this factor. The same holds true for the competitor comparison (although, of course, in this case there is no specific connection between FPs from the two services).

10. Estimation of False Negatives: for both comparisons it is necessary to estimate the number of false negatives of the full process (those HMC e-mails that are missed by all service components and thus passed on to customers). This is done by making estimates based on MessageLabs internal research and customer feedback. There are some issues concerning the accuracy of this estimate but MessageLabs have adopted a conservative approach that is likely to result in an overestimate (perhaps considerably so) of the number of these false negatives and thus an overestimate of the risk to its customers - thereby understating the relative advantage of using their service. See section C, paragraph 15 for more details.

11. Calculations of probabilities etc: the various formulae used and calculations made from the data are correct. Hence the results obtained are accurate subject to the underlying assumptions. See section C, paragraph 16 for more details

12. Underlying model and assumptions: the underlying model is appropriate and the assumptions are reasonable. The relevant assumptions made are:

- The appropriate measure for the proportion of HMC e-mail is the ratio of the number of HMC e-mails not blocked by the AV service to the number of supposedly clean e-mails sent to a customer. (i.e. spam and other unwanted e-mails do not feature in the ratio) - see section C, paragraph 16.
- This proportion is sufficiently ‘typical’ that it can be viewed as a probability of a received e-mail being infected.

- The number of false negatives passed to a customer from MessageLabs is as estimated - see paragraph 8 and section C paragraph 15.
- The probability that an infection will lead to a major incident varies between organisations but is likely to lie in the range 2-5%.
- The figures derived from [8], provide appropriate estimates of the cost of a major incident.

The first two of these are crucial to the underlying model, as far as extrapolation to overall risk is concerned. The implication is that the exposure of an organisation to HMC is proportional to its received e-mail activity. There are certainly factors that support such a model but there are also other factors involved. There is also an inherent difficulty in the second of these bullets, which limits the precision of the measure of risk in financial terms but does not invalidate the comparisons. Otherwise, the least supported of the assumptions is the 2-5% probability of an infection leading to a major incident. Because of this MessageLabs have sought to reconcile their findings with the published survey results of major security incidents. This reconciliation is appropriate as a validation exercise but the inclusion of a further factor in the calculations to force equality is purely a conservative response to avoid the potential overstatement of the risks and possible financial losses. The factor of about 8.5 is well within the range that might be expected given the uncertainty of the various assumptions involved and is equivalent to reducing the probability of an infected e-mail leading to a major incident by a factor of 8.5. The inclusion of this factor does not affect the relative performance of the various AV solutions.

13. Project conclusions: the following conclusions drawn from the project results are fully justified:

a. The MessageLabs service offers significantly superior performance both to the commercial products and to the competitor service, resulting in a measurably reduced risk in terms of the expected cost of major security incidents.
b. The competitor service performs significantly better than the commercial products.
c. The probability estimates for HMC e-mail being received by customers following processing by the commercial products[6] appear to be typical but the underlying variability is significant, implying that the measurement of the risk in £ is only an indication of the order of magnitude of the expected cost. Otherwise, the risk measurement depends on the estimate of the probability that an infection will lead to an incident. Where there is doubt, assumptions have been conservative and the quoted results are likely to be underestimates.
d. The estimates for the competitor are less precise than for the commercial products but nevertheless do have validity. The variability observed during the longer period of the commercial product experiments indicates a need for caution when extending conclusions beyond the specific period of the experiment. Hence the confidence in the extrapolated risk calculations is

[6] i.e. the false negatives from the processing

lower than for the commercial products. However, a real financial risk is indicated, while the relative superiority of the MessageLabs performance is very marked and can be asserted with confidence.

e. The MessageLabs service misses very few instances of HMC emails. Consequently the precise figures for the risk to MessageLabs customers are not that relevant. What can be inferred is that the residual risk for MessageLabs customers is very low.

## C Detailed Discussion

14. In the following discussion the notation developed in the project reports [1-7] will be used, making use of some of the quantities defined there.

For the Commercial Product Comparison:

N41 = the number of HMC e-mails blocked by the commercial products.
N61 = the number of HMC e-mails blocked by Skeptic.
N7FN = the number of HMC e-mails let through by Skeptic.
N9 = the number of e-mails passed to the customer following MessageLabs processing.
N9FN = the number of HMC e-mails passed to the customer following MessageLabs processing.

For the Competitor Comparison:

Fig 1 taken from [5], section 7 defines the regions A - E. Abusing notation I shall also use A-E to denote the number of e-mails in each region.

N41 = HMC e-mail identified as such by both MessageLabs and the competitor.
N42M = HMC e-mail identified as such by MessageLabs but not by the competitor.
N42P = HMC e-mail identified as such by the competitor but not by MessageLabs.
N51M = HMC e-mail from N42P identified as spam by MessageLabs.
N51P = HMC e-mail from N42M identified as spam by the competitor.
N52 = HMC e-mail identified by both services as spam.
N61(M & P) = HMC e-mail not identified either as HMC or spam but which it is assumed would have been blocked as spam had the service had access to the original source data.

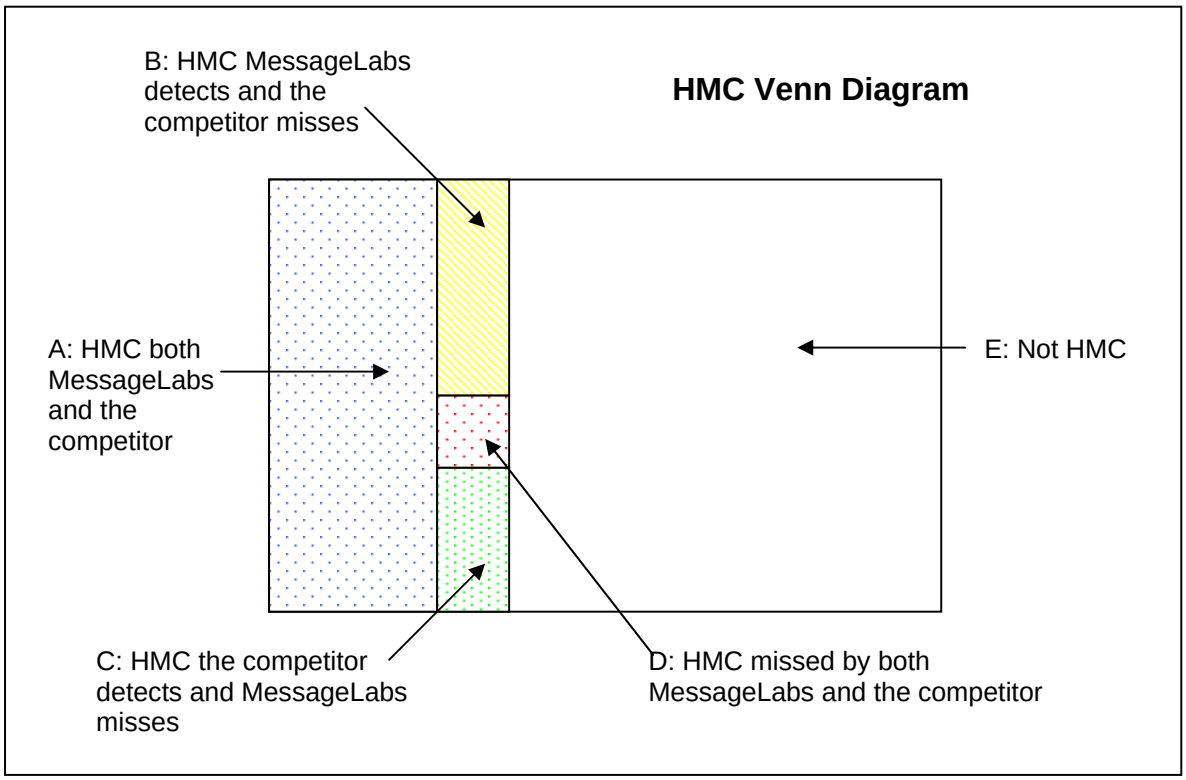


Fig 1

15. Estimation of false negatives: the quantity N9FN is required for calculating the residual risk for MessageLabs customers. As mentioned in section B, paragraph 8 this is estimated from the reports generated from customer feedback and from retrospective examination of those e-mails that remain unblocked after processing by Skeptic (including, but not limited to, those that are later blocked by the final filtering stages - principally the antispam engine, Cynic). There is every incentive for customers to provide MessageLabs with feedback (and it should certainly be expected for any HMC that caused an incident). Although the actual risk to customers only comes from those e-mails that pass through all of the processing, the figure used in the computation of risk is taken from those that pass through Skeptic - i.e. N7FN. Even so this figure is extremely small. For the 14 day October 2008 comparison the estimate per day was only 6.7 compared to the total number of e-mails processed being in the region of 100 million per day. However statistics for other periods suggest that, while always relatively low, the figure varies considerably. This is to be expected. A much longer period is needed to understand this variation. Because even the worst figures are still extremely low in comparative terms, it turns out this variability is not of great importance to the conclusions of this project as will be discussed in paragraph 17.

16. Calculation of probability of HMC: the way that the gathered data is used is to estimate the incidence of HMC emails that would be received by a customer as a proportion of his legitimate (i.e. ‘wanted’) e-mail feed. This figure is obtained using averages calculated by MessageLabs from the actual externally originated e-mail quantities by their variously sized customer organisations (typically this average is in the range 7-15 per user per day). Thus the appropriate ratio to use is that of the number of HMC emails to the total number of clean and wanted e-mails (clean ham),

for the product comparison this is: (N61+N7FN)/(N9-N9FN) for the e-mail passed by the commercial products and N9FN/(N9-N9FN) for the e-mail passed to the customer following the full MessageLabs service. In this study N9 is of the order of 60 million and the previous paragraph shows that N9FN is tiny and hence can be ignored in the denominator. The ratios used to represent the probabilities of an e-mail containing HMC are (N61+N7FN)/N9 and N7FN/N9. For the competitor comparison a more indirect calculation was used because of the differing nature of the email feeds. Referring to Fig 1, from the data collected, the sizes of A, B and C can be calculated directly, but D cannot.

A = N41 + N51M + N51P + N52 + N61P + N61M
B = N42M - N51P - N61P
C = N42P - N51M -N61M

D is obtained from the estimate for N7FN from the commercial product comparison scaled appropriately to allow for the different quantity of e-mails. The ratios (B+D)/(A+B+C+D) and (C+D)/(A+B+C+D) then give the proportions of the HMC total that pass the two AV services. These are combined with the results of the more voluminous commercial product comparison to arrive at the required ratios of undetected HMC to wanted e-mail.

17. Commercial product comparison - sample size: the e-mail quantities used in this comparison are huge. This indicates high confidence in the figures for N41, N61 and N9 for any particular period of several days duration[7]. The data shows short-term stability but medium-term variability for the incidence of HMC (N41 and N61). This gives high confidence in the range of values observed for the proportion of HMC to legitimate e-mails, and the probability of HMC e-mail blocked and missed by the commercial products covering the typical variation that might be encountered. Additionally, the fact that the average obtained from the 14 day experimental period is essentially the same as that for the full 7 month period for which statistics have been collected suggests that this is a good representative value on which to base the risk calculations. On the other hand, the effectiveness of the MessageLabs process is so close to 100% that all that can be said with confidence is that the probability of MessageLabs missing HMC is very small even when using the worst figures observed in any period.

18. Competitor comparison - additional considerations: as mentioned in paragraph 8 the lower e-mail volumes used in this comparison mean that there is less confidence in the representative nature of the figures. Also data covering other time periods might give further insight into the potential variability of the results - although it would be surprising if the figures were more stable over time than the others discussed here. For this reason all the specific results are best seen as indicative of an ‘order of magnitude’ and, rather than quoting them as a single figure, a likely range should be quoted. Cross-calibration with the product comparison gives some added confidence in the results and the relative performance between the two services is clear. Furthermore, the number of HMC e-mails missed by the competitor

[7] The 14 day period used for the full test is sufficient for this purpose, anything from 10 days upwards would be satisfactory.

is sufficiently large that there is higher relative confidence in the calculated probabilities for the competitor service than for the MessageLabs service. Nevertheless, there is some inherent variability arising from the specific assumption of the corresponding performance of the MessageLabs service. The probability for the competitor service is calculated from the data and the MessageLabs probability in the form

$$P = \frac{HMCmissedbycompetitor}{N9} = \frac{HMCAll}{N9} \times \frac{B + D}{A + B + C + D}$$

The value of HMCAll/N9, the ratio of the total volume of HMC e-mail to the clean wanted e-mail, is taken from the commercial product data. D is also estimated from the commercial product data from C + D, appropriately scaled, deriving from N7FN. The range of possible values for these quantities implies a range for the competitor performance. This range features in recommendation 2 in section D, paragraph 20. There was some concern expressed that the honeypot source might provide a higher proportion of HMC e-mails than typical customer feeds. This fear does not seem to have been realised in practice, for although the observed proportion of HMC e-mails was about 10% higher than in the product comparison this is well within the expected random variation and provides no evidence that the underlying proportions are different.

19. Risks of major incidents: while the calculations of the expected number of infected e-mails are directly supported by the data collected, further calculation of the number of incidents and expected costs depends on assumptions that depend on the professional judgement of the MessageLabs experts. There are three components to the calculations that require justification.

a. The estimate of the chance of an infection leading to a major incident.
b. The cost of a major incident.
c. The scaling factor used in typical examples in an attempt to reconcile the results with the incident reporting according to the ISBS08 [8].

The first of these is dependent on the specific conditions pertaining within an organisation and is best estimated by the organisation itself. The illustrative figures used in the MessageLabs comparisons are based on the professional judgement of their experts and are likely to be reasonable, although there is no independent evidence available for cross-confirmation. The estimates range from 5% for a small company (100 users) down to 2% for an enterprise company (3000 users). The second estimate is based on figures reported in [8] and has been taken as 1/4 of the quoted maximum costs of major incidents. This appears to be eminently reasonable. Unsurprisingly (given the suspected degree of under-reporting of incidents reflected in the survey [8]) the calculation based on the data and the above assumption suggests a rather more frequent occurrence of such incidents than reported in [8]. To reconcile this disparity and ensure that the case for the MessageLabs service is not overstated the representative calculations include a further scaling factor (of about 8.5) which matches the calculated number of incidents for a medium size company with the reported figures. Although this is not obviously justified, the consequence is to understate the potential risks, perhaps considerably, and hence underplay the advantage to be obtained from using the MessageLabs service in absolute terms.

These assumptions only affect the absolute values associated with the risks and not the relative performance of the various AV solutions.

D Further comments and recommendations

20. The main uncertainty derives from the variability of the incidence and treatment of HMC at different times, which means that a single probability estimate is not really adequate for the risk calculations. In particular the calculations for the full MessageLabs service are inherently very imprecise. This is because it performs so well that the incidence of missed HMC e-mail is so low that an average value cannot be established with confidence. Data has been collected beyond the initial 14 day period at first envisaged in setting up the experiment. The additional data has given insight into the variability and the calculations are based on this greater volume of data. Two recommendations suggest themselves.

Recommendation 1: instead of relying on a single estimate for the risk calculations, in all cases the calculations should be given not only for a 'typical' average value but also for the endpoints of a range allowing higher confidence that the appropriate risk lies within that range.

Recommendation 2: to add to the confidence in the estimates, the data gathering should continue.

21. There is also slightly greater uncertainty about the competitor comparison than is desirable. The most useful additional confidence could come from an extended comparison to investigate the variability over time.

Recommendation 3: the competitor comparison should be continued to confirm the incidence of missed HMC by the competitor's service.

22. Finally it would be useful to try to confirm the reasonableness of the assumptions underlying the calculation of the risk arising from major incidents. To some extent this will be dependent on others providing better information. However, there is also additional modelling work that might be done under the aegis of MessageLabs, supported by their data sources. The model here relies on a number of assumptions (see paragraph 12) and involves several stages. The direct measurements (in both experiments) give counts for the total number of infected e-mails detected. The counts are supplemented by an estimate of the number of such e-mails that that go undetected by MessageLabs. These figures are used to compare the effectiveness of the detection services in terms of the proportion of the total HMC e-mails that remain undetected. Suppose the values are:

H = total number of infected e-mails detected (by anyone).
F = number of infected e-mails detected by no-one.
X, Y, Z = number of infected emails detected by at least one service but undetected by, respectively, the commercial products, the competitor and MessageLabs. The proportions that to be calculated are:

$(X + F)/(H + F)$ ; $(Y + F)/(H + F)$ and $(Z + F)/(H + F)$.[8]

The calculations based on the figures are fine for comparing the effectiveness of the services at detecting infected e-mail. However, more is required to measure the risk faced by customers. There are three crucial steps to the model.

a. The first step is to calculate the probability that a random e-mail received by a customer is infected. This calculation was accomplished directly for the commercial products experiment as the ratio between the number of undetected e-mails to the clean wanted e-mails (see paragraphs 12 and 16). The indirect calculation for the competitor comparison illustrates what is going on rather better. The expression given in paragraph 18 shows in the form $P = E Q$, where E is the exposure to infected e-mail and Q is the probability that an infected e-mail has remained undetected. The experiment has delivered the value of Q. E is down to the model. Further study might give more insight into the way that malware is distributed and thus of how different organisations are exposed to it. The model used is that the exposure is directly proportional to the number of genuine e-mails received.

b. The second step is to include factors expressing the likelihood of a received infected e-mail giving rise to a major incident. Other than by using professional judgement (educated guess), MessageLabs is not in a position to improve this part of the model. Improvements will only come from other organisations being prepared to co-operate in appropriate research.

c. The third step is to plug in estimates for the cost of a major incident. This is very much an area for external research requiring honest reporting by victims.

Recommendation 4: MessageLabs should encourage further research on the modelling assumptions and, in particular, investigate where it might contribute to improved understanding of the first step (a above).

[8] While the value of F is essential for the calculation of the MessageLabs performance, in practice F is demonstrably very much smaller than H or X and for the commercial products the proportion is essentially X/H and the actual value of the estimate F does not really affect the result of the calculation. However in the competitor experiment the observed values were approximately, H = 20000, Y = 420, Z = 0, which makes the value of F of crucial importance - the key comparison being of F with (420 + F). The estimates for F from the commercial product comparison put it at around 20.