

Virus Removal Cost (VRC) Metric

Kuangnan Chang Bobby C. Adkins
Department of Computer Science
Eastern Kentucky University
{kuangnan.chang, bobby_adkins13}@eku.edu

ABSTRACT

This paper proposes a software metric for estimating the cost of removing a virus from infected machines. The metric, Virus Removal Cost (VRC), provides an objective estimate of the cost based on two factors: the characteristics of a virus and the computer skills of persons who try to remove the virus from their machine. The metric can give us an insight into the cost of removing a virus, and, hence, produce more reliable data that reduce the deviation of the estimate of the overall damage cost caused by the virus.

KEYWORDS

Virus removal, Software metric, Regression analysis, Logest function

1. INTRODUCTION

It has been noticed that there is a knowledge gap between judges and computer scientists, in which the justice system lacks the knowledge for dealing with cybercrime and attacks on the information security infrastructure [4]. One essential way to help equipping judges and law enforcement to deal with cybercrimes is to study digital forensics, which concerns the discipline studying the analysis and use of electronic evidence in law enforcement and judicial process. Developing appropriate metrics to measure the seriousness of a cybcrime is an important topic in digital forensics. The data produced by the metrics should be objective and understandable, which can be used as evidences or reliable references in the court [2,4].

A metric for measuring the cost of removing a virus from all infected machines is proposed in this paper. The metric, named Virus Removal Cost (VRC), provides an objective estimate of the cost based on two factors. Based on our observation, the characteristics of a virus and the computer skills of persons who try to remove the virus from their machine decide the cost of the virus removal. Using the

past studies and datasets, we can use regression analysis to produce a formula to explain the trend of the data in the past, and the relationships between these two factors and the data. The metric focuses on the virus removal cost only, not the cost of overall damage caused by the virus. The virus removal cost is a part of the overall damage cost. The metric can give us an insight into the cost, and, hence, produce a more reliable data that reduce the deviation of the estimate of the overall cost.

The rest of this paper is organized as follows: In section 2 the proposed VRC metric is discussed in detail, including the equation of the metric, and the terms used in the equation. Section 3 illustrates the use of the metric with an example. The cost of removing the Melissa virus from one million infected machines is estimated to be ten million dollars. Finally, section 4 concludes the paper.

2. THE PROPOSED METHOD

We first determine the terms that are used in the proposed VRC metric in section 2.1. The formal determination of the metric is described in section 2.2.

2.1. DETERMINING THE EQUATIONS

Based on our observation, the time needed to remove a virus is based on two factors, the computer skill of the person removing the virus, and the difficulty of removing the virus [1,2,7]. Let variables x_1 and x_2 represent the skill factor and the difficulty factor respectively.

There are many ways to determine a person's computer skill. According to [1], people's computer skill can be classified into five categories, starting from the Excellent level to the Poor level. Figure 1 shows the distribution of computer skills for men and women. In our method, a value from 1 to 5 is assigned to the skill level x_1 . So a person with excellent computer skill will have a value of 1 and a person with poor computer skill will have a value of 5.

	Men	Women
	%	
Access to computer	69	66
Self-rated computer ability		
Excellent	15	8
Very good	19	22
Good	28	31
Fair	24	23
Poor	15	16
Internet use in past 12 months	56	50
General technology use index: ¹	3.8	3.5
Average years of using computer	7.5	7.1

1. See "What you should know about this study" for definition.
Source: Statistics Canada, General Social Survey, 2000.

Figure 1. Distribution of computer skills [1]

For the difficulty level x_2 , we use the definition of the difficulty of removing a virus by Symantec [7] and classify the difficulty level into three categories: Difficult, Moderate, and Easy. The definition given by Symantec is as follows:

Removal

Measures the skill level required to remove the threat from a given computer. Removal sometimes involves deleting files and modifying registry entries. The three levels are Difficult/High (requires an experienced technician), Moderate/Medium (requires some expertise), and Easy/Low (requires little or no expertise).[7]

In our method, the difficulty level x_2 can have values 1 through 3. If a virus can be removed easily, then we assign value 1 to x_2 . If it is very difficult to remove the virus, then we assign 3 to x_2 .

The combinations of variables x_1 and x_2 is 15. For each combination, we pick at least three sets of 100 numbers randomly with normal distribution to represent the times that are needed for people with x_1 level computer skill to remove a virus that is at x_2 level. 50 groups of numbers are generated and the average time for each group is shown in Appendix A.

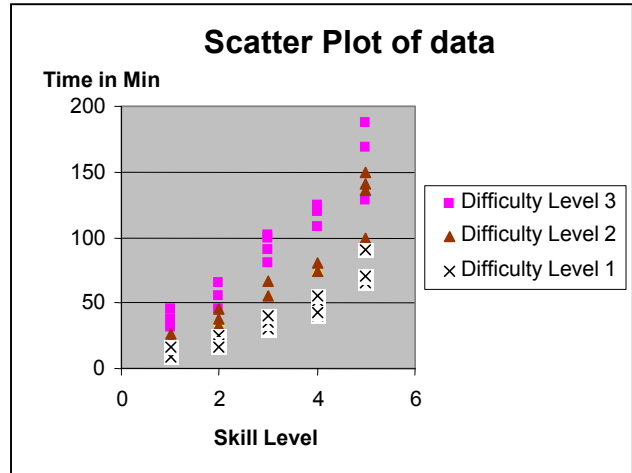


Figure 2. Scatter plot of dataset.

Figure 2 shows the data in Appendix A with a scatter plot. The data shows a slight curve to it, which suggests an exponential regression analysis to find an appropriate equation to fit the data best. By applying the Logest function [5,6] to the variables skill (x_1) and difficulty (x_2), the function returns three values that we will refer to as b , m_1 , and m_2 . The exponential equation that the function has fit to the data is below.

$$RTime = b(m_1^{x_1})(m_2^{x_2})$$

The values $b = 5.697677$, $m_1 = 1.508947$, and $m_2 = 1.631908$ were returned through exponential regression analysis of our data. Hence, our basic equation for determining the time (in minutes) a person at a certain skill level uses to remove a virus at a certain difficulty level as determined by our data would be:

$$RTime = 5.697677 (1.508947^{x_1})(1.631908^{x_2})$$

The equation suggests that if a person at skill level 3 needs to remove a virus from their computer that has a difficulty removal level of 2, it will take about 52 minutes for that person to remove the virus from his computer.

$$RTime = 5.697677(1.508947^3)(1.631908^2) = 52.1328$$

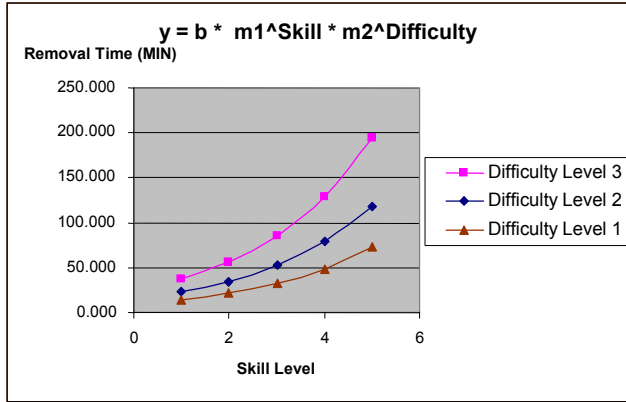


Figure 3. Exponential analysis of dataset.

The graph of our fitted equation through exponential regression is shown in Figure 3. The figure represents a removal difficulty level. Notice how closely it matches the shape of our data in the scatter plot of Figure 2.

Next, we determine how many people fall into different computer skill levels. We borrow the data in Figure 1 in our study. The data is based on a Canadian study done in 2000 where women and men were asked how they rated themselves with a computer (Self-rated computer ability). We assume that the U.S. is close to the values of the Canadians. The results in the figure are in percentages. We take this data and formulate Table 1 with the combined Men and Women values as percentages (p).

Table 1. Combined men & women computer skills.

	Men	Women	Total	P
Excellent	15	8	23	11.44%
Very Good	19	22	41	20.40%
Good	28	31	59	29.35%
Fair	24	23	47	23.38%
Poor	15	16	31	15.42%
Total	101	100	201	100.00%

The graph of p has an approximately normal distribution, shown in Figure 4. We will use this value of p in our main formula to determine the cost of removing a virus.

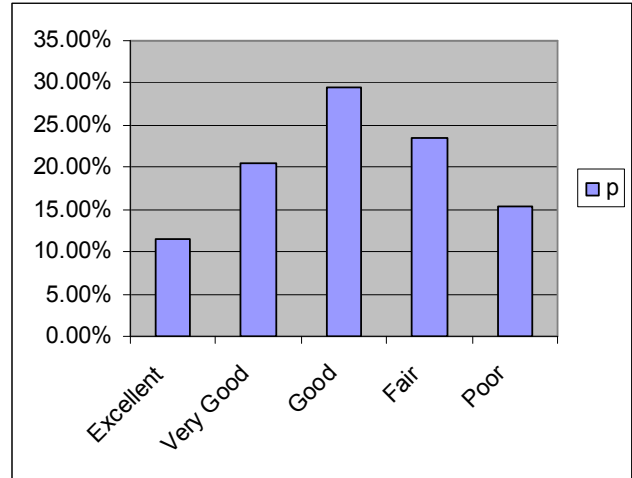


Figure 4. Distribution of p .

For the number of machines that have been infected by a virus, it is easy to find a good estimate by searching the Internet for such a number [3,7]. Let z be the estimated number of infected machines. If we take p multiplied by z , this should give us an estimate of how many computers are infected with the virus by users of each skill level. For example, if 1000 computers (z) have been infected then we can multiply that by p of the people who have a self-rated computer skill ability of excellent.

$$\text{Infected} = z \times p = 1000(.1144) = 114.4$$

Let k be the average hourly earning for all the persons in these skill levels. We can use k along with the number of computers infected from that skill level to get how much it costs for a person to work on the virus for one hour, which is defined as follows:

$$\text{PerHourCos } t = z \times p \times k$$

According to the U.S. Bureau of Labor Statistics, the average hourly earning for the United States workers is about 16 dollars in 2005 [8]. Based on this value, we can estimate the cost for all the individuals with excellent computer skills to remove the virus from 1000 infected machine in the above example as (if it takes one hour to remove the virus):

$$\begin{aligned} \text{PerHourCost} &= z \times p \times k = 1000(.1144)(16) \\ &= 1830.4 \end{aligned}$$

So it costs approximately \$1830.4 per hour for the people with excellent computer skills to remove the virus from

their computer systems. Combine this with our regression analysis formula and we can get a good estimate of how much it costs to completely remove a specific virus from all computer systems.

2.2. PUTTING IT ALL TOGETHER

For each skill level, the cost for removing a virus is defined as:

$$PerHourCost \times RTime$$

Applying the equations determined in the previous section, the cost is extended as:

$$z \times p \times k \times \frac{b(m_1^{x_1})(m_2^{x_2})}{60}$$

Since there are five different computer skill levels, our major metric, Virus Removal Cost (VRC), is defined as:

$$VRC_1 = \sum_{i=1}^5 z \times p_i \times k \times \frac{b(m_1^{x_{1i}})(m_2^{x_2})}{60}$$

The metric determines the estimated total cost of removing a virus from all infected computer systems by persons with different computer skill levels. After pulling the values that will remain constant outside of the summation resulting, the above equation becomes:

$$VRC_1 = \frac{z \times k \times b \times m_2^{x_2}}{60} \sum_{i=1}^5 p_i \times m_1^{x_{1i}}$$

This is the generalized form of the VRC metric. It will work with any dataset as long as the dataset has an exponentially shaped scatter plot. However, users will need to do exponential regression analysis to get their b , m_1 , and m_2 values. For our random dataset, the values $b = 5.697677$, $m_1 = 1.508947$, and $m_2 = 1.631908$.

In the equation, the average hourly earning k can have only one single value. We can have a more accurate estimate by studying different average hourly earning for the persons with different computer skills. When considering this factor, the above metric can be modified as follows.

$$VRC_2 = \frac{z \times b \times m_2^{x_2}}{60} \sum_{i=1}^5 p_i \times k_i \times m_1^{x_{1i}}$$

This doesn't take into account region, which could change the pay rate somewhat. The values of these k_i terms will be various when different regions are considered. A new

survey for determining the values of these terms in the equation is suggested when applying the metric in a different area.

3. AN EXAMPLE

In year 2000, a virus named "Love Bug" spread quickly across Europe the United States. The virus originates in an email entitled "I love you." Once the attachment is opened, the virus sends copies of same email to all the persons that are listed in the user's address book. A similar virus named Melissa has infected about one million computers in the United States [3].

We use this case and our random dataset to illustrate how to estimate the cost for removing such a virus from the infected computer with the proposed metric. According to Symantec, Love Bug is easy to be removed. Hence, the term x_2 in the metric is set to 1. The term z is set to 1,000,000, which represents the number of the infected machines. Again, from our random dataset, the values $b = 5.697677$, $m_1 = 1.508947$, and $m_2 = 1.631908$. We set k to 16, according to the average hourly earning data provided by the U.S. Bureau of Labor Statistics. By using VCR_1 , we can estimate the cost as follows.

$$VRC_1 = \frac{1000000 \times 16 \times 5.697677 \times 1.631908^1}{60} \times [.1144 \times 1.508947^1 + .2040 \times 1.508947^2 + .2935 \times 1.508947^3 + .2338 \times 1.508947^4 + .1542 \times 1.508947^5] \approx 10,076,424$$

Hence, the estimated total cost of removing the Melissa virus from all computers is about ten million dollars. Note that the Melissa virus has clogged whole networks in the United States once, and caused about eighty million dollars in damage [3]. The cost is much higher than our estimate. The difference between these two numbers is that our metric estimates only the cost for removing the virus from infected machines. It does not count any other damages that the virus can bring to business or enterprises. The role of the metric is to give a more objective and detailed estimate, which may be considered one part of the overall damage a virus caused.

4. CONCLUSION

We have proposed a metric, VRC, for estimating the cost of removing a virus from all infected computers. The metric considers two factors, which we believe are two most important influences in the cost: the computer skills of the

persons whose computers have been infected by the virus, and the difficulty for removing the virus based on the characteristics of the virus. The computer skills are classified into five levels: excellent, very good, good, fair, and poor. The difficulty of a virus may be one of three levels: easy, moderate, and difficult. With these two factors and our dataset, VCR is derived defined from the exponential regression analysis on the data.

The equation of the metric described in section 2.2 is in its general form. We have determined the terms used in the equation of the metric in section 2.1. Some of these terms are constants, and they can be various values when the metric is applied in different regions. For example, the average hourly earning rates are different in different regions or countries. These constants can be more accurate when a wide survey is finished. This survey remains our future work.

The metric gives an insight into the cost of virus removal. The estimate made by the metric can reduce the deviation of the overall damage cost caused by a virus. Hence, a more objective and reliable data can be reached. We hope this metric in a way can help in law enforcement and judicial process.

REFERENCES

- [1] Heather Dryburgh, "Learning Computer Skills," Canadian Social Trends, Statistics Canada, No. 11-008, pp. 20-24, Spring, 2002.
- [2] Hanno Langweg and Einar Snekkenes, "A Classification of Malicious Software Attacks," Proceedings of 23rd IEEE International Conference on Performance, Computing, and Communications, pp. 827-832, 2004.
- [3] LYCOS Wired News, *Love Bug Virus Running Amok*, http://www.wirednews.com/news/technology/0,1282,36113,00.html?tw=wn_story_related, May 4, 2000.
- [4] Michael Losavio, "Non-technical Manipulation of Digital Data: Legal, Ethical and Social Issues for Computing, Judicial Process & Digital Forensics," IFIP Working Group 11.9 on Digital Forensics, International Conference on Digital Forensics, February 2005.
- [5] Microsoft Office Online, *Definition of Logest Function*, <http://office.microsoft.com/en-us/assistance/HP052091591033.aspx>.
- [6] Rodrigo Vivanco and Nick J Pizzi, "Identifying effective software metrics using genetic algorithms," Proceedings of the IEEE Canadian Conference on Electrical and Computer Engineering, pp. 1305-1308, Montreal, Canada, May 4-7, 2003.
- [7] Symantec, *Glossary Definition of Removal*, <http://securityresponse.symantec.com/avcenter/refa.html#removal>.
- [8] U.S. Department of Labor-Bureau of Labor Statistics, *Comparison of Average Hourly Earning*, <ftp://ftp.bls.gov/pub/suppl/empstat/compes.txt>.

Appendix A. Hypothetical Time for Virus Removal

Skill Level	Difficulty Level	Average Removal (Min)
1	1	10
1	2	25
1	3	38
1	1	15
1	2	26
1	3	45
1	1	9
1	2	22
1	3	31
1	1	16
2	1	25
2	2	36
2	3	55
2	1	19
2	2	45
2	3	66
2	1	16
2	2	38
2	3	45
2	2	33
3	1	30
3	2	55
3	3	100
3	1	35
3	2	67
3	3	80
3	1	40
3	2	55
3	3	102
3	3	91
4	1	40
4	2	75
4	3	108
4	1	51
4	2	81
4	3	120
4	1	55
4	2	74
4	3	124
4	1	43
5	1	65
5	2	100
5	3	128
5	1	71
5	2	136
5	3	188
5	1	90
5	2	150
5	3	168
5	2	141