Epidemiological Simulation of a Nonlinear Computer Network Laboratories Using Kermack-Mckendrick Model

Las Johansen B. Caluza
Leyte Normal University, Philippines
lasjohansencaluza@lnu.edu.ph

Abstract **- Computer viruses, malware, worms, spyware, and the like have shown a significant impact in the digital world today. It has shown its power to destroy computer systems and hardware and most of the time used for hacking, spying, and other cyber crimes. With this scenarios, it empirical to unravel the story behind computer network laboratories using Kermack-McKendrick model through simulation of a nonlinear network using NetLogo v.6.0 in an actual setting happened in a Philippine teacher education institution. Results revealed the impact of the computer viruses and the difference between a free antivirus and a licensed software antivirus in the computer network laboratories. Finally, with this initiatives, the university decided to purchase a licensed software antivirus as it is deemed necessary to protect the end-user from the computer viruses and the equipment itself for long term use, savings, and helping the environment as a whole.**

Keywords **- Simulations and Modeling, Computer Viruses, Kermack-McKendrick Model, Nonlinear Network, Computer Security, Quantitative, Philippines**

I. INTRODUCTION

Over the last few years, publications appeared on the subject of computer viruses and malware were very few. Works such as technical details that were necessary to understand and effectively defend against computer viruses (Szor, 2005) were some of those. The attacks caused by viruses and malware mostly came from the internet (Graham-Cumming, 2006; as cited by Kondakci, 2008). Moreover, these computer viruses and malware nowadays were being used in cyber crimes like hacking using the internet as a deployment tool considering that millions of internet users are online every day. Based on history, computer viruses arose in the 80's (Chen, Hattaf, & Sun, 2015) and had made a significant impact on the computer network and the internet. Among those was the infamous iloveyou virus that causes billions of dollars in loss and damages (Ravi, Raghunathan, Kocher, & Hattangady, 2004). In the current years, the rapid development and utilization of hardware and software technologies and the popularity of computer network and the internet (e.i. Social media) had made a significant threat and getting more severe (Chen, Hattaf, & Sun, 2015) in terms of predicted results when these viruses and malware attacks and activated in the computer systems. Only recently, however, researchers began to study the patterns of connectivity and the activities made by these viruses and malware, the ability to handle perturbation or attack, and the ability of the ecosystem to handle disturbances (Lloyd & May, 2001).

II. MODEL

A. SIR Model

A SIR model is an epidemiological model that calculates the hypothetical number of people infected with a contagious illness in a closed population over time. The name of this class of models derives from the fact that they involve coupled equations relating the number of susceptible people *S(t)*, the number of

individuals infected *I(t)*, and some individuals who have recovered *R(t)*. One of the simplest SIR models is the Kermack-McKendrick model (Weisstein, 2008).

B. Kermack-McKendrick Model

The Kermack-McKendrick model (Weisstein, 2009) is a SIR model (S-Susceptible, I-Infected, R-Resistance) for the number of people infected with a spreadable illness in a closed population over time. It was predicted to elucidate the fast rise and fall in the number of infected patients observed in epidemics such as the plague (London 1665-1666, Bombay 1906) and cholera (London 1865) (cited by Weisstein, 2017). The fixed population size was adopted (i.e., no births, deaths due to disease, or deaths by natural causes), the incubation period of the infectious agent is instantaneous, and the duration of infectivity is same as the length of the disease. It also assumes an entirely homogeneous population with no age, spatial, or social structure.

III. METHOD

The experimental and descriptive design was utilized in this research to test the variability of a computer virus infect other computers in a nonlinear computer network established in a university computer laboratory. First, this research conducted an investigative approach to observing, analyzing, and understanding the infrastructure of the network topology design implemented in the five (5) computer laboratories in the university. Students were using these computer laboratories across the disciplines I the university. Second, the gathered information was used as a basis for SIR simulation in a nonlinear computer network prior, during and after the virus and malware attack. NetLogo 5.2.1 version (Wilensky, 1999) was utilized using the existing model on the SIR Model for Computer Viruses and Malware in the experiment. This model demonstrates the spread of a virus through a network. Although the model is somewhat abstract, one interpretation is that each node represents a computer, and we are modeling the progress of a computer virus (or worm) through this network. Each node may be in one of three states: susceptible, infected, or resistant. In the academic literature, such a model sometimes referred to as a SIR model for epidemics (Stonedahl and Wilensky, 2008). In continuation, the Number of Nodes represents the total number of computers in the laboratories; Average Nodes Degree represents the number of computers that are connected or has the network permissions; Initial Outbreak Size is an estimated number of computers that are computer virus infected. While the Virus Spread Chance set at 2.5% is a percentage probability of a computer to be infected, by default as suggested by Stonedahl and Wilensky. Furthermore, Virus Check Frequency represents the time step increment (tick), this is used to do nodes check whether they were infected. While Recovery Chance was used identifying when the virus has been detected and has the probability to be removed, then the Gain Resistance Chance is a node recover, from the virus, then there is a probability that the node will become resistant to this virus in the future.

This research employed the SIR Model using the Kermack-McKendrick Model. The model consists of a system of three coupled nonlinear ordinary differential equations:

1) $d\,S/d\,t$
2) $d\,I/d\,t$
3) $d\,R/d\,t$

Where t is time, $S(t)$ is the number of susceptible nodes, $I(t)$ is the number of nodes infected, $R(t)$ is the number of nodes which have recovered/resistant and developed immunity to the infection, β is the infection rate, and γ is the recovery rate.

The key value governing the time evolution of these equations is the so-called epidemiological threshold,

4) $R_0 = \frac{\beta s}{\gamma}$

Note that the choice of the notation R_0 is a bit unfortunate. Since it has nothing to do with R. R_0 is defined as the number of secondary

infections caused by a single primary infection; in other words, it determines the number of nodes infected by contact with a single infected node before its death or recovery.

When $R_0<1$, each node who contracts the disease will infect fewer than one node before dying or recovering, so the outbreak will peter out ($d\ I \mid d\ t < 0$). When $R_0<1$, each node which gets the virus will infect more than one node so that the epidemic will spread $(d\ I \mid d\ t < 0)$. R_0 is probably the single most significant quantity in epidemiology. Note that the result $R_0 = \beta\ S/\gamma$ derived above, applies only to the basic Kermack-McKendrick model, with alternative SIR models having different formulas for $d\ I \mid d\ t$ and hence for R_0.

IV. NETLOGO DESIGN AND SET-UP

A. Input Variables and Description

- **Number-of-Nodes** is the total number of computers in a network. In this case, there are 185 computers (nodes) in the computer laboratories.

- **Average-Node-Degree** is the number of servers in the computer laboratories. The use of these servers is for file sharing purposes only.

- **Initial-Outbreak-Size** is the number of computers assumed infected by viruses and or malware through either connecting the internet using emails, downloading, using social networking sites, and using secondary devices like flash drive, CDs, and others.

- **Virus-Spread-Chance** is the probability of infecting susceptible computers and neighbors (nodes). Virus-Spread-Chance is an assumed value in percentage.

- **Virus-Check-Frequency** the frequency of the computers checked whether a virus infects them.

- **Tick** is the time step each infected node attempts to infect all of its neighbors. One (1) tick is equivalent to 1 milliseconds. Tick1 is the time step of Free Antivirus Software while Tick2 is the time step for Licensed Antivirus Software.

- **Recovery-Chance** is the number of times the user conduct virus scanning and applying quarantine and deleting the viruses and or malware. Recovery-Chance is an assumed value in some scanning, quarantine, and deleting viruses and malware.

- **Gain-Resistance-Chance** is an assumed value in percentage if a node does recover, there is some probability that it will become resistant to this virus in the future.

B. Output Variables and Descriptions

- **Susceptible** these are nodes (computers) that are likely to be influenced or infected by a virus. Susceptible_1 were the values incurred based on the parameters defined for a Free Antivirus Software (FAS), while Susceptible_2 were the values incurred based on the parameters defined for a Licensed Antivirus Software (LAS).

- **Infected** these are nodes (computers) that were infected by a computer virus.

- **Resistant** these are nodes (computers that were infected and were able to resist from a computer virus. Resistant_1 were the values incurred in the given parameters defined for FAS, while Resistant_2 were the values incurred based on the parameters defined for LAS.

V. NETLOGO MODEL SETTINGS

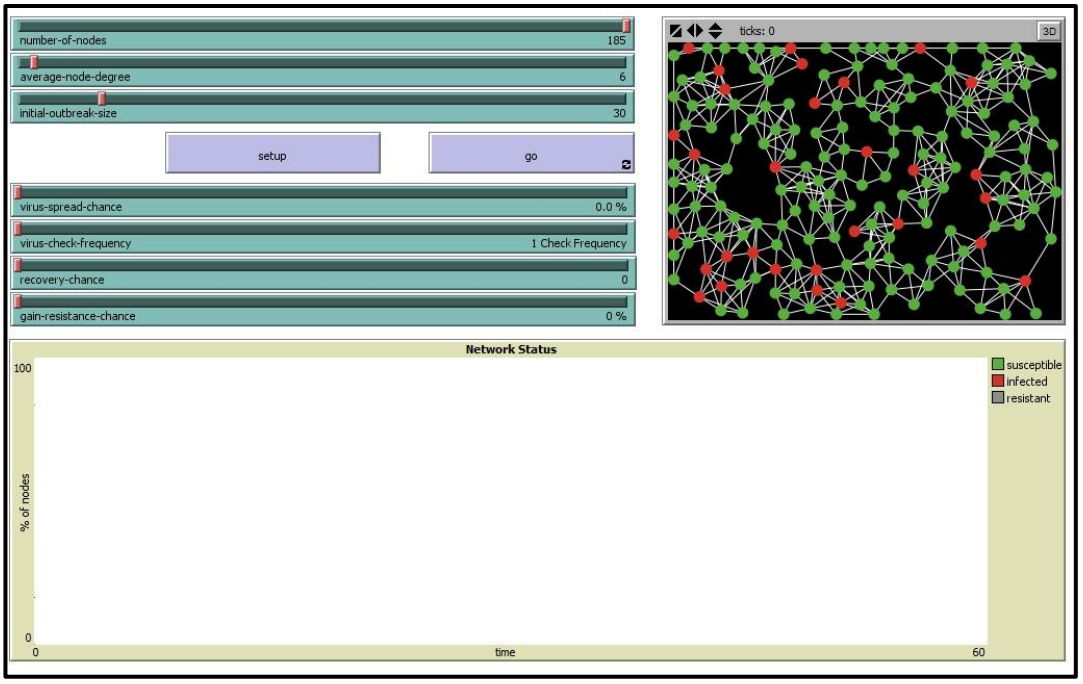

Fig. 1 NetLogo 3D Real World Representation of Nonlinear Computer Network

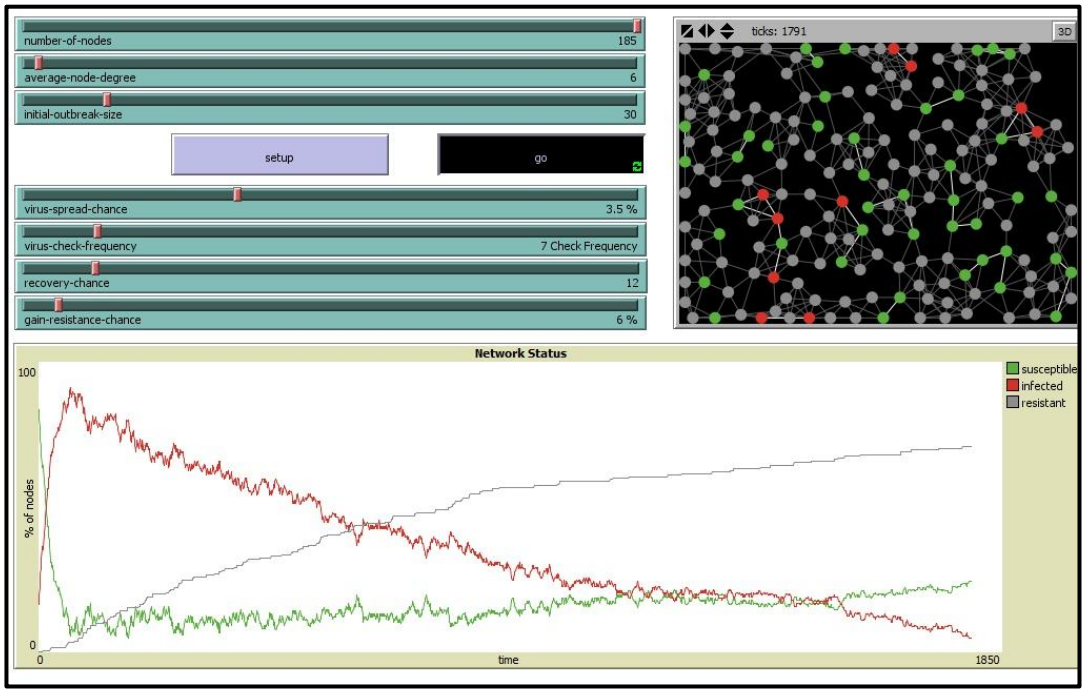

Fig. 2 Sample Simulation Output of Real World Nonlinear Computer Network

VI. RESULTS AND DISCUSSIONS

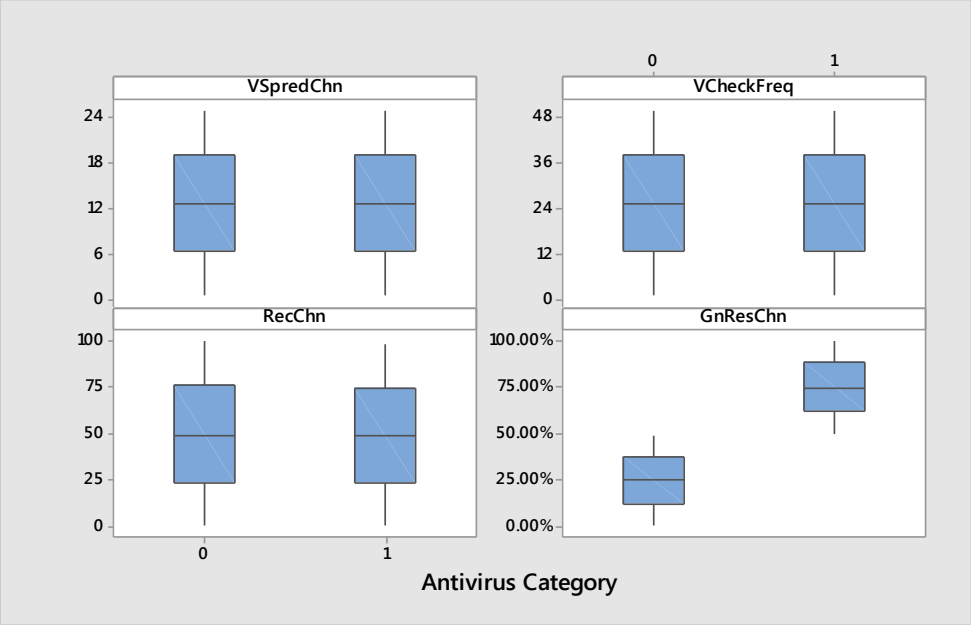

Fig. 3 Boxplots of Input Variable

Fig. 3, illustrates the difference between a FAS and LAS based on the virus spread chance, virus check frequency, resistance chance, and gain resistance chance. As observed, the gain resistance chance of a LAS has higher Gain Resistance Chance than FAS.

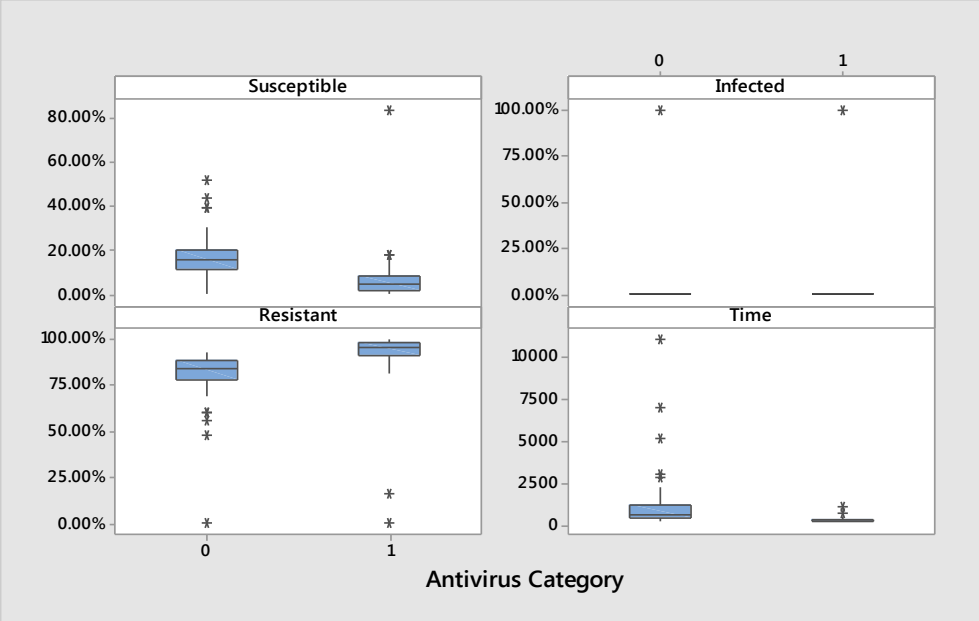

Fig. 4 Boxplots of Output Variables

Based on the results from the simulation, Fig. 4 shows the difference between FAS and LAS based on Susceptibility, Infected, and Resistant (SIR). Regarding Susceptibility, FAS is more susceptible than LAS after the simulation while, regarding Infected Nodes, both at one time, in the beginning, were infected. Moreover, LAS group is more resistant to computer viruses than the FAS group. Finally, LAS group processes faster than the FAS group regarding performance time evaluation.

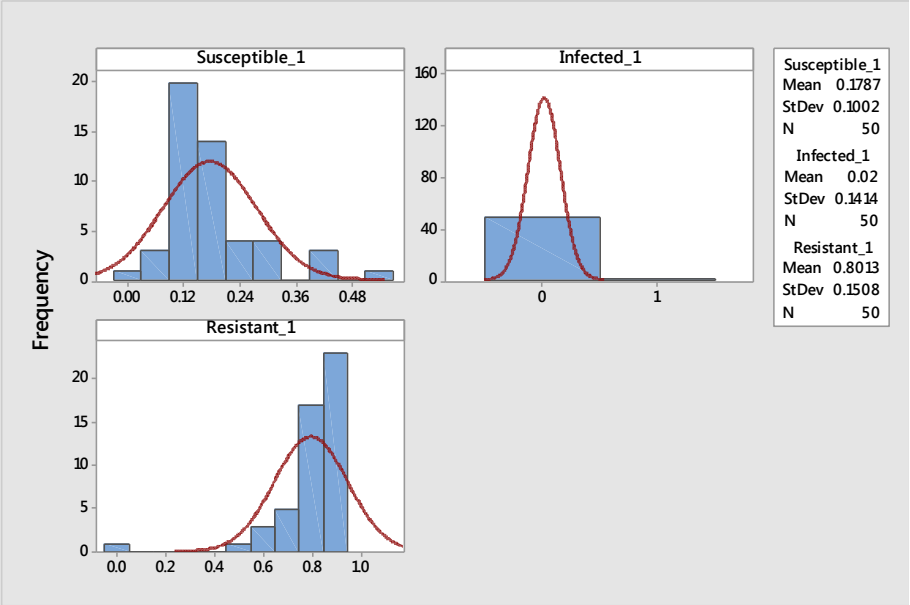

Fig. 5 Histogram of Free Antivirus Software on SIR

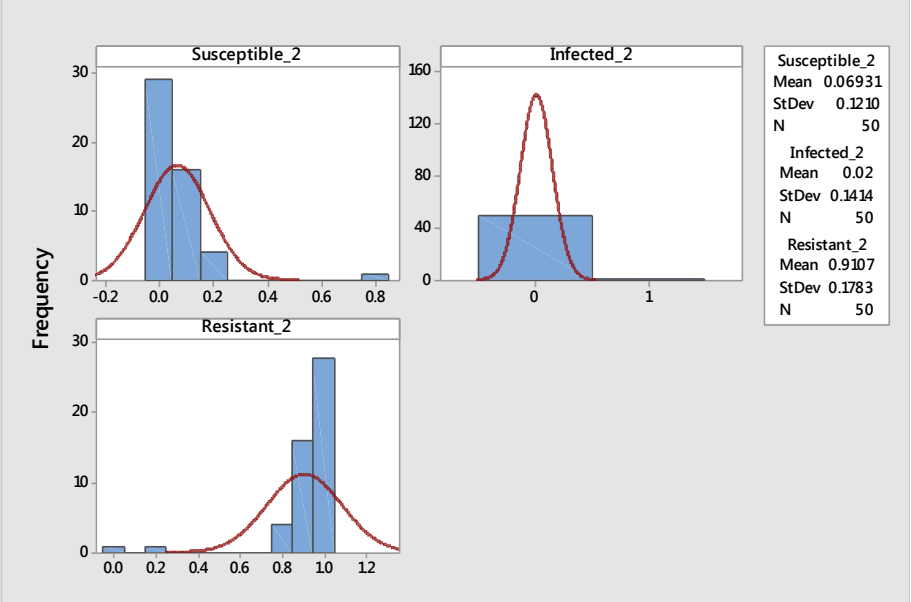

Fig. 6 Histogram of Licensed Antivirus Software on SIR

TABLE I
PAIRED SAMPLES STATISTICS – TIME SPENT IN PROCESSING

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Tick1	1230.2000	50	1883.51654	266.36946
	Tick2	283.0600	50	161.29848	22.81105

Table I shows, the Paired Samples Statistics Mean on Time Spent in Processing (Tick) with 1230.2 for Tick1 and 283.06 for Tick2 from 50 samples. The result implies that *Tick1* spent much time in the simulation than *Tick2*.

TABLE II
PAIRED SAMPLES CORRELATIONS – TIME SPENT IN PROCESSING

		N	Correlation	Sig.
Pair 1	Tick1 & Tick2	50	.814	.000

As shown in Table II, the paired samples correlations on time spent in processing indicate that *Tick1* and *Tick2* scores are significantly positively correlated (*r=0.814*). As a result, a strong association was established between the two variables.

TABLE III
PAIRED SAMPLES TEST – TIME SPENT IN PROCESSING

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1 Tick1 – Tick2	947.14000	1754.76798	248.16167	448.44046	1445.83954	3.817	49	.000

Based on Table III, the average difference between the *Tick1* and *Tick2* is 947.14 with Std. Dev. of 1754.76 and Std. Error Mean of 248.16. On the other hand, *Tick1* and *Tick2* scores were vigorously and positively correlated (*r=0.814, p<0.000*) and there was a very highly significant average difference between *Tick1* and *Tick2* scores (t_{49} *= 3.817, p<0.000*). On average, *Tick1* scores were 947.14 milliseconds higher than *Tick2* Scores (95% CI [448.44, 1445.83]).

TABLE IV
PAIRED SAMPLES STATISTICS – SIR

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	Susceptible_1	.1784	50	.09982	.01412
	Susceptible_2	.0696	50	.12144	.01717
Pair 2	Infected_1	.0200[a]	50	.14142	.02000
	Infected_2	.0200[a]	50	.14142	.02000
Pair 3	Resistant_1	.8018	50	.15101	.02136
	Resistant_2	.9106	50	.17857	.02525

a. The correlation and *t* cannot be computed because the standard error of the difference is 0.

Table IV shows, the paired samples statistics of the SIR results that the correlation and the t value cannot be computed of the *Infected_1* and *Infected_2* because the standard error of the difference is 0. Looking into the simulation results, it shows that at the start of the simulation of both groups; the Virus Spread Chance is at 0.50, Virus Check Frequency is at 1, Recovery Chance is at 0, and the Gain Resistance Chance is at 0 % (for free antivirus) and 50 % (for a fee antivirus) resulting to 100% Infection to all nodes (computers). Moreover, the rest of the values after the first simulation showed 0% Infections.

TABLE V
PAIRED SAMPLES CORRELATIONS – SIR

		N	Correlation	Sig.
Pair 1	Susceptible_1 & Susceptible_2	50	.438	.001
Pair 3	Resistant_1 & Resistant_2	50	.751	.000

The paired sample correlations in SIR results as shown in Table V, revealed that *Susceptible_1* and *Susceptible_2* (*r= 0.438*) and *Resistant_1* and *Resistant_2* (*r=0.751*) were significantly positively correlated.

TABLE VI
PAIRED SAMPLES TEST – SIR

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference Lower	Upper			
Pair 1	Susceptible_1 - Susceptible_2	.10880	.11871	.01679	.07506	.14254	6.481	49	.000
Pair 3	Resistant_1 - Resistant_2	-.10880	.11905	.01684	-.14263	-.07497	-6.462	49	.000

The paired sample test in SIR (table VI) shows that *Susceptible_1* and *Susceptible_2* scores were average and positively correlated (*r=0.438, p<0.001*). Moreover, there was a significant average difference between *Susceptible_1* and *Susceptible_2* scores (t_{49}*=6.48, p<0.000*). Meanwhile, *Susceptible_1* scores were 0.10880 points higher that *Susceptible_2* scores (95% CI [0.075, 0.142]). *Resistant_1* and *Resistant_2* scores were strongly and positively correlated (*r=0.751, p<0.000*). While there was a negative significant average difference between *Resistant_1* and *Resistant_2* scores (t_{49}*=-6.46, p<0.000*). Finally, *Resistant_1* scores were -0.10880 points lower that *Resistant_2* score (95% CI [-0.142, -0.074]).

VII. CONCLUSION

Free Antivirus Software processing time is slower in determining SIR than Licensed Antivirus Software. While regarding SIR, the Free Antivirus Software has higher Susceptibility to computer viruses than Licensed Antivirus Software. Moreover, both groups were infected with computer viruses at one time at the beginning of the simulation revealed. Finally, it revealed in the groups (FAS & LAS) that both were able to resist from computer viruses but Licensed Antivirus Software has a higher value in resisting computer viruses. Moreover, the use of simulation and modeling software helps further understand the capabilities of different software like computer antiviruses and the like to give decision support to administrators in real life situation. Finally, the researchers were able to enlighten possible research area in simulations and modeling in an actual situation by looking into in-depth computer security simulation of the same area.

REFERENCES

(Arranged in the order of citation in the same fashion as the case of Footnotes.)

[1] Chen, L., Hattaf, K., and Sun, J. (2015). "Optimal Control of a Delayed SLBS Computer Virus Model". Physica A: Statistical Mechanics and its Applications, 427, pp. 244-250.

[2] Graham-Cumming, J. (2006). "Malware Prevalence Report". <https://www.virusbulletin.com/>.
[3] Kondakci, S. (2008). "Epidemic State Analysis of Computers under Malware Attacks". Simulation Modelling Practice and Theory, 16(5), pp. 571-584.
[4] Lloyd, A.L. and May, R.M. (2001). "How Viruses Spread among Computers and People". Science, 292(5520), pp. 1316-1317.
[5] Ravi, S., Raghunathan, A., Kocher, P., and Hattangady, S. (2004). "Security in Embedded Systems: Design Challenges". ACM Transactions on Embedded Computing Systems (TECS), 3(3), pp. 461-491.
[6] Weisstein, E.W. (2017). "Kermack-McKendrick Model". From MathWorld--A Wolfram Web, <http://mathworld.wolfram.com/Kermack-McKendrickModel.html>.
[7] Weisstein, E.W. (2009). "Kermack-McKendrick Model". MathWorld--A Wolfram Web Resource, <http://mathworld.wolfram.com/Kermack-McKendrickModel.html>.
[8] Weisstein, E.W. (2008). "SIR Model". MathWorld-A Wolfram Web Resource, <http://mathworld.wolfram.com/SIRModel.html>.
[9] Wilensky, U. (1999). "NetLogo". <http://ccl.northwestern.edu/netlogo/>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.
[10] Stonedahl, F. and Wilensky, U. (2008). "NetLogo Virus on a Network Model". <http://ccl.northwestern.edu/netlogo/models/VirusonaNetwork>. Center for Connected Learning and Computer-Based Modeling, Northwestern University, Evanston, IL.
[11] Szor, P. (2005). "The Art of Computer Virus Research and Defense". Pearson Education.