National-level risk assessment: A multi-country study of malware infections

Fanny Lalonde Lévesque[1], José M. Fernandez[1], Anil Somayaji[2], and Dennis Batchelder[3] [4]

École Polytechnique de Montréal[1], Carleton University[2], Microsoft Corporation[3], AppEsteem[4]

Abstract. The security of computers is a function of both their inherent vulnerability and the environment in which they operate. Much as with the public health of human populations, the "public health" of computer populations can be studied in terms of what factors influence their security. Using data collected from Microsoft Windows Malicious Software Removal Tool (MSRT) running on more than one billion machines, we conduct a multi-country analysis of malware infections and measures of economic development, educational achievement, Internet infrastructure, and cybersecurity preparedness. We find that while increases in these factors is often correlated with reduced infection rates, their significance and magnitude vary considerably. In contrast to past work, these variations suggest that policy interventions, such as efforts to increase the quality of home Internet connections, are likely to decrease infection rates in only some circumstances.

Keywords: risk factor, malware, ecological study, public policy, cybersecurity, population health

1 Introduction

The susceptibility of computers to malware infections is known to be affected by technological factors (e.g. their hardware, operating system, and applications) [22, 24, 10] and human factors (e.g. computer expertise, risk aversion) [22, 31, 37]. With human health, however, we know that factors such as economic development, geographic location, and the aggregate health choices all influence the health of individuals. For example, while individuals can take steps, such as using mosquito nets and insect repellent, to avoid catching malaria, the biggest factor influencing whether you may get malaria is simply where you happen to live. If the mosquitoes in your area happen not to carry malaria then you are safer from it —even if you take no other protective steps. Similarly, if authorities in the area you live in take steps to reduce the number of mosquitoes, your

15th Workshop on the Economics of Information Security (WEIS) 2016
University of California at Berkeley, June 13-14

risk of malarial infection also goes down, all without any changes in individual susceptibility or individual behavior. Our question here, then, is can we identify analogous factors that could be changed through national policies, such as the prevalence of mosquitoes or vaccination rates, that would improve the security level of entire computer populations?

For instance, while it may be intuitive that wealthier nations perform better in cybersecurity, or that nations with higher Internet connectivity are more susceptible to cybercrime, it is essential to validate those hypotheses and understand their causes. Many studies, mostly from antimalware vendors, security experts, or networking providers, report on geographical patterns and trends in malware infections without investigating the factors behind those variations. So far, only few studies have examined how national factors (e.g. income, education, Internet penetration) correlate with cross-country differences in malware infections [17, 27, 38, 8]. However, there is no overwhelming consensus in the literature on which factors are the best predictors of malware infections at the national-level. Moreover, those studies offer little or no discussion on potential underlying causes for their findings. Consequently, lack of consensus and scarcity of evidence represent a serious challenge for cybersecurity policy making. In order to support good, evidence-based policy making, we need to conduct empirical studies on large and representative populations of computer systems that will provide understanding of the causes of malware infections.

Fortunately observations of large computer populations is now feasible due to telemetry systems embedded into commonly-used security software. While these systems were originally developed for quality assurance, they can also be used to study the patterns associated with malware infection and other security violations. Security software telemetry data thus allows us to adopt a *population health* approach. Formally, population health refers to "the health outcomes of a group of individuals, including the distribution of such outcomes within the group" [20]—the population, in our setting, being computer systems. Similarly, a large body of work has also looked at how *public health* may serve as a model for cybersecurity [35, 34, 36, 30, 39, 11]. Much as with health, epidemiological techniques can then be applied to security to investigate factors and conditions that affect the heath status of computer systems in order to develop cybersecurity policies and strategies that reduce the risk of security compromise.

There have been a few previous epidemiological studies that used security telemetry data to identify risk factors related to malware infections [41, 44]. While these past studies have identified technical and behavioral factors related to individuals and organizations, they were not designed to identify risk-modifying factors at the national-level. Moreover, interventions focused on individuals or organizations are unlikely to succeed if the environmental condition in which they are delivered are not supportive. Therefore, there is a need to understand the multi-level risk factors leading to malware infection, including both *proximal*, *intermediate* and *distal* factors, as the latter two are often determinants of the risk factors. While proximal factors act directly or almost directly on the cause of infection, distal factors are further back in the causal chain and

act via a number of intermediate factors. As both ecological along with individual and behavioral determinants play an important roles in the development and prevention of malware infection, it becomes important to conceptualize the problem within multiple levels of influence.

Commonly used within population health research, *ecological studies* can be designed to identify risk factors at higher levels. In such studies, populations are defined by temporal (tracking a population over an extended period or time) or spacial (comparing populations in different geographic locations) units and compared on their prevalence or incidence of disease. This type of observational study is particularly useful for generating and testing hypotheses on potential risk factors, whether distal, intermediate, or proximal. From there, other epidemiological or laboratory approaches can be used to test the causality, if any.

In this paper we report on a multi-country ecological study of risk factors related to malware infections. Country infection rate is computed using large-scale telemetry data from millions of systems running Microsoft Malicious Software Removal Tool (MSRT), a malware cleaner utility that scans Windows systems for infections of specific malicious software. We investigate association of factors related to economics, education, technology, and cybersecurity on malware infection rate by country. We develop regression models for the prevalence of malware infection to identify and quantify the relative importance of those risk factors and how their effects vary between countries with different socio-economic status. In summary, our main contributions are:

1. We present a multi-country ecological study of malware infection risk factors, based on a large sample of unprotected hosts (100 million).
2. We investigate how malware infections at the national-level correlate with factors related to economy, education, technology and cybersecurity, including some previously unstudied factors like antivirus penetration, Global cybersecurity index, etc.
3. We develop a regression model for the prevalence of malware infection that identifies and quantifies the relative importance of those factors and how their effect vary between countries with different socio-economic status.
4. We identify potential risk-modifying factors that can be influenced by cybersecurity policies.

The remainder of the paper is organised as follows. Section 2 presents a review of previous studies and Section 3 describes the study in terms of data collection and analysis. In Section 4 we present the results in terms of national-level risk factors for malware infection. We discuss our observations and study limitations in Section 5. We conclude and discuss potential implications of our findings in Section 7.

2 Previous studies

We present a short review of past work focusing on the link between national factors and malicious attacks at a cross-country level. In what follows we dis-

tinguish our study with prior research in terms of datasets, study design, and analysis methodology.

Some researchers have focused on the impact of national cybersecurity policies on malicious attacks at the country level. Ivan *et al.* [32] adapted the event study methodology from research in financial economics to study the impact of government enforcement and economic opportunities on information security attacks in the US. They found limited evidence that domestic enforcement deters attacks within the country. Microsoft also sought to understand whether certain policies can measurably reduce cyberrisks at the national level [21]. They conducted a descriptive analysis and found that countries adopting or implementing certains policies, like the London Action Plan (LAP) or the Europe Convention on Cyberbrime (ECC), may contribute to reduce the risk of malware infection. Overall, those studies more or less all found that national cybersecurity policies may contribute to reduce the risk of malicious attacks.

Garg *et al.* [17] performed a cross-country empirical analysis to investigate how macroeconomic factors grounded in traditional theories of crime offline relate to the rate of machines acting as spambots. Factors related to the availability of machines, guardianship, economic deprivation, legal framework and governance were investigated. Results suggested that higher Internet adoption, measured by the total number of fixed broadband Internet subscribers, is related with a higher percentage of spambots while countries with higher secure Internet servers (per million people) were associated with a lower percentage of spambots.

In another study, Microsoft did a cross-country analysis of different social and economic policy indicators to predict the rate of malware infections within countries [8]. They used 2013 data from MSRT and defined the infection rates as the number of computers cleaned for every 1,000 executions of MSRT. Their predictive model identified 11 factors related to digital access, institutional stability and economic development. Countries with above-average development across those areas were expected to see greater improvement in cybersecurity. Although the authors included in their study a broad set of national factors, their statistical analysis was predictive, and not explanatory. That is, the purpose of their statistical model was to predict the rate of malware infections, which is different from causal explanation. In opposition, our study used explanatory modeling for testing potential causal factors behind international differences in malware infections.

Only few explanatory research have investigated the effects of multiple factors in terms of economics, technology, and cybersecurity on malware prevalence at the national level. Mezzour *et al.* [27] performed an empirical study to understand how the average malware encounters rate of home users vary internationally. Using 2009-2011 telemetry data from the Symantec's Worldwide Intelligence Network Environment (WINE) [13], the authors empirically test the validity of specific factors related to computing and monetary ressources (i.e. GDP per capita, Internet bandwidth, ICT development index), cybersecurity expertise (as measured by cybersecurity research and the existence of cyberse-

curity institutions), international relations, computer piracy and web browsing. They found that high piracy rates was the main factor associated with high malware encounters especially in countries with low computing resources.

Subrahmanian *et al.* [38] also leveraged the WINE telemetry data from host machines protected by Symantec's antivirus products. They computed the average number of infection attempts per host of a given country as a proxy of its level of cyber-vulnerability. In an attempt to explain international differences in attack frequencies, they performed a multivariate analysis including macroeconomic factors (per capita GDP, Internet penetration, software piracy) and host-based features aggregated at the country-level (total number of binaries installed, fraction of downloaded binaries, of unsigned binaries, and of low-prevalence binaries). Overall, they found per capita GDP and fraction of downloaded binaries to be significant predictors; countries with low economic wealth (as measured by per capita GDP) and high fraction of downloaded binaries were more vulnerable. In contrast to [27], they found software piracy to be non-significant, suggesting that its effect may be more a function of other variables, such as per capita GDP, than a direct cause of cyber-vulnerability. In comparison to [27] and [38], our research is distinct in three important ways. First, the sample population is different; they studied protected host machines (from Symantec) of home users, and we focus on unprotected host machines including both home and corporate users. Second, their dependent variable was the average malware encountered by computer in each country, while we are interested in countries' malware infection rate. Three, our work accounts for a broad set of national factors; neither [27] or [38] investigated factors related to both economy, education, technology, and cybersecurity readiness.

The key way our research differs from past work is in how we designed our study and performed our analysis. While past studies have focused on the identification of national factors [17, 27, 38] or the development of a predictive model [8], our research goes beyond previous work as we also quantify the relative importance of the studied factors. We also evaluate how the direction and magnitude of those factors vary between countries with different socio-economic development levels, while all research previously cited is limited to a global analysis. Moreover, most of the papers cited above offer little discussion of how their results in terms of national factors should be interpreted, for instance, whether they should be seen as direct or indirect effect or whether they are confounded by other factors. Finally, compared to previous work, our study is grounded in traditional epidemiological techniques.

3 Study design and methods

A multi-country ecological study was conducted in order to identify which national factors are the best predictors of malware prevalence across countries. This type of observational study was selected as it is often used to identify factors on health when the outcome is averaged for the population in geographical or temporal units. The main advantage is that it allows to study variables that

cannot be measured at the individual level or that may have a different effect at the individual and population level. Such variables, called *ecological factors*, can be classified as *aggregate*, *environmental* or *global variables*, depending on what they measure. Aggregate factors are data based on individuals aggregated at the population level. Environmental factors relate to the characteristics of the environment in which people live. Although they are measured at the population level, they can also be measured at the individual level. Global factors are variables computed from groups, organizations, or places for which there is no analogue at the individual level. While ecological studies are convenient to test multiple hypothesis at the same time, special care should be taken to select the appropriate sample size and sampling method, limit potential bias and effect of chance, and control for potential *confounding variables* —undesirable factors that may influence the results and threaten the internal validity of the study.

The data was collected by Microsoft Malicious Software Removal Tool (MSRT), a malware cleaner utility that scans computers for infections of specific, prevalent malicious software and helps remove these infections [1]. MSRT is delivered and runs every month on more than one billion machines through Windows Update as well as being available as a separate download from Microsoft. Upon its execution, MSRT also calls the Windows Security Center (WSC) API to collect information about the protection state of computers, such as the antivirus (AV) actively protecting the machine and its signature status. Such information is then reported by MSRT to Microsoft for a random sample of 10% of the machines. The data used in this analysis was —monthly— collected from June to September 2014 on computers running Windows XP, Vista, 7, 8 and 8.1., which represents 100+ million computers.

3.1 Data collection

The dependent variable under consideration is the rate of malware infection by country for unprotected computer systems, which represent approximately 10% of the 100+ million computers. The rate of malware infections was computed based on the proportion of unprotected systems that reported at least one infection over the 4 months. Systems were considered unprotected if they had no AV product enabled on their machine. We chose to focus on unprotected systems so as to avoid the bias other AV software would potentially introduce into rates of infection. Moreover, it allows us to focus on malware infections, rather than malware encounters. As far as we know, this is the largest study on malware infections on unprotected systems. Furthermore, Internet Protocol (IP) geolocation was used to identify the country associated with each user report.

Independent variables were selected based on two criteria, i.e. 1) they were plausible risk factors, and 2) they constituted factors that might be possibly reduced by intervention at the country level. We selected 15 factors (see Table 1) to cover the socio-economic and technological reality of countries, as well as their level of cybersecurity. A detailed description of factors considered in the current study is presented in the following text and in Appendix A.

Table 1: Country-level factors

Model	Description	Year Source
Economy	GDP	2013 WB
	GDP-PPP	2013 WB
Education	Mean years schooling	2013 UNDP
Technology	%Households with computer	2013 ITU-D
	%Households with Internet	2013 ITU-D
	Fixed Internet subscriptions (per 100 people)	2012 ITU-D
	Fixed broadband subscriptions (per 100 people)	2013 ITU-D
	Fixed (wired) broadband speed	2013 ITU-D
	%Fixed broadband subscriptions (256kbit/s - 2Mbit/s)	2012 ITU-D
	%Fixed broadband subscriptions (2Mbit/s - 10Mbit/s)	2012 ITU-D
	%Fixed broadband subscriptions (above 10Mbit/s)	2012 ITU-D
	International Internet bandwidth (per million people)	2013 ITU-D
Cybersecurity	Secure Internet servers (per million people)	2013 WB
	%Protected	2014 MSRT
	Global cybersecurity index	2014 ITU-D

Economic performance We used the Gross Domestic Product (GDP) and the Gross Domestic Product per capita by purchasing power parity (GDP-PPP) from the World Bank (WB) [40], as indicators of the economy. While GDP measures the wealth within a country, GDP-PPP embeds a measurement of income inequality across countries. The direction in which those variables may play is difficult to predict. On the one hand, the economy of countries could influence their resources and opportunities to make choices that could protect their population [17]. On the other hand, higher monetary resources may cause an increase in malicious attacks, as many malware have a monetary goal [27]. Those factors are global variables and were considered control variables in the analysis as they may be markers for variables we cannot measure nor control.

Education As a measure of the level of education, we used the mean years of schooling (MYS) from the United Nations Development Programme (UNDP) [42], which represents the average number of years of education received by adults aged 25 and older. This variable may account for a possible direct effect of aggregate education and for indirect effects, that are not captured by other factors, such as user behaviour, information technology (IT) literacy, and cybersecurity awareness. We expect that education will be negatively associated with infection rates, as it may affect, among others, users' ability to understand IT information, and follow guidelines for their online safety. This factor is an aggregate variable and was considered as a control variable for the purpose of the analysis.

Technology Technological factors were selected from the International Telecommunication Union Development Sector (ITU-D) [3] to capture both the quantity and quality of information and communications technology (ICT). The quan-

tity was evaluated in terms of the percentage of households with a computer, the percentage of households with Internet access, the number of fixed Internet subscriptions (per 100 people) and the number of fixed broadband subscriptions (per 100 people). While factors related to technology quantity are aggregate variables, we are interested in their environmental effect. For example, countries with large population of computers and Internet users could be more subject to malicious attacks as they may have more potentially vulnerable machines [17, 27].

The indicators for the quality were selected to measure both the broadband speed and the bandwidth. For broadband speed, we used the fixed (wired) broadband speed (Mbit/s) (FBS), which refers to the advertised maximum theoretical download speed; it does not refer to the actual speed delivered. We also used the percentage of fixed broadband subscriptions for different speed categories: advertised downstream speed between 256 kbit/s and less than 2Mbit/s (%FB(256-2)), between 2 Mbit/s and less than 10Mbit/s (%FB(2-10)), and greater than or equal to 10 Mbit/s (%FB(10+)). To evaluate the bandwidth, we used the international Internet bandwidth (IIB), which refers to the total used capacity of international Internet bandwidth, in megabits per second. It measures the sum of used capacity of all Internet exchanges offering international bandwidth. We divided the IIB by the country's total population and multiplied by 1 million to obtain the bandwidth by 1 million inhabitants. The direction of the association between technology quality and malware infection rates is difficult to say in advance. As one could argue that technology quality may affect users' ability to protect themselves (i.e. having an up-to-date system or performing signature updates for anti-malware products), it may also contribute to increased cybercrime (remote attacks, spam distribution, software piracy, etc). The factors related to the technology quality were considered aggregate variables.

Cybersecurity Factors were selected to capture both private and individual investment in security, and national cybersecurity development. Similar to Garg *et al.* [17], we used the number of secure Internet servers (per million people) from the WB as a proxy for private investments in security. For the investment at the individual level, we used the percentage of users protected by antivirus product. This last factor was obtained from MSRT and is based on the percentage of systems that have at least one antivirus product actively running with up-to-date signatures during the 4-month period. To evaluate the level of cybersecurity development of countries, we used the Global cybersecurity index (GCI) [26]. This index was developed by an ITU-ABI research joint project to rank the cybersecurity capabilities of nation states within five categories: legal measures, technical measures, organizational measures, capacity building and cooperation. We except all variables to have a negative association with the rate of malware infections. Factors related to cybersecurity at the private and individual level were both considered aggregate variables, while the GCI was a global variable.

3.2 Statistical methods

The goals of the statistical analysis were 1) to estimate variations in the prevalence of malware infection across countries, 2) to quantify the relative importance of national factors in this variation, and 3) to study the relationship between specific factors and malware infection rates.

While our data set is large overall, for some countries our sampled population is too small to allow for proper analysis. In order to determine the minimum representative sample size for each country, we performed a power analysis to identify the minimum number of computer system reports required. We used a two-tailed one proportion Chi-Square test with a desired power of 90% and a level of significance of 1%. The minimum sample size computed was 37 149 system reports by country, which was rounded to 38 000. We then excluded all countries that had less than 38 000 reports over the 4 months, reducing our sample from 241 to 187 countries.

We implemented a general linear regression model —a specific generalized linear model. First, we ensured that the relationship between the dependent and independent variables was linear, and applied when required a log transformation to the independent variables in order to meet the linearity assumption. All factors were log transformed, except mean years of schooling, %Households with computer, %Households with Internet, %Protected and Global cybersecurity index. Descriptive statistics of the factors before and after the transformation are presented in Table 2 and Table 3 respectively. The mean allows to measure the central tendency of the data and the standard deviation measures how concentrated the data are around the mean; the more concentrated, the smaller the standard deviation (SD).

Table 2: Descriptive statistics (whithout transformation)

Factor	Mean	SD
%Infected	0.22	0.13
GDP	4.82e+11	1.64e+12
GDP-PPP	1.91e+04	2.03e+04
Mean years of schooling	8.29	1.29
%Households with computer	45.46	30.45
%Households with Internet	41.79	30.38
Fixed Internet subscriptions (per 100 people)	14.12	12.67
Fixed broadband subscriptions (per 100 people)	12.93	12.97
Fixed broadband speed	4.48	8.61
%Fixed broadband subscriptions (256kbit/s - 2Mbit/s)	0.35	0.35
%Fixed broadband subscriptions (2Mbit/s - 10Mbit/s)	0.34	0.24
%Fixed broadband subscriptions (above 10Mbit/s)	0.34	0.33
International Internet bandwidth (per million people)	2.63e+05	1.34e+06
Secure Internet servers (per million people)	4.16e+02	9.53e+02
%Protected	0.20	0.09
Global cybersecurity index	0.33	0.22

Table 3: Descriptive statistics after log-transformation

Factor	Mean	SD
%Infected	0.22	0.13
GDP*	10.83	0.88
GDP-PPP*	4.03	0.51
Mean years schooling	8.28	2.92
%Households with computer	45.46	30.45
%Households with Internet	41.79	30.38
Fixed Internet subscriptions (per 100 people)*	0.76	0.80
Fixed broadband subscriptions (per 100 people)*	0.60	0.97
Fixed broadband speed*	0.18	0.61
%Fixed broadband subscriptions (256kbit/s - 2Mbit/s)*	-0.87	0.78
%Fixed broadband subscriptions (2Mbit/s - 10Mbit/s)*	5.10	1.22
%Fixed broadband subscriptions (above 10Mbit/s)*	-0.87	0.85
International Internet bandwidth (per million people)*	4.55	0.97
Secure Internet servers (per million people)*	1.62	1.15
%Protected	0.20	0.09
Global cybersecurity index	0.33	0.22

*Variables have been log-transformed.

In order to identify and assess the unique impact of each factor, we looked for multicollinearity —strong correlations between the independent variables—, as it can reduce the amount of information available to evaluate the effect of the factors. The presence of multicollinearity was investigated by computing the variance inflation factor (VIF), which estimates how much the variance of a coefficient is inflated because of linear dependence with other variables. A low value (VIF < 5) implies that the variable is uncorrelated with all the other variables [5, 45]. To the opposite, a high value is a sign of multicollinearity. We excluded GDP-PPP (VIF > 10) and retained GDP as an indicator of economic performance. Secure Internet servers (per million people) was also found to be highly multicorrelated (VIF > 10) with other variables. This factor was excluded, while we retained the Global cybersecurity index and the percentage of users protected by antivirus product as indicators of cybersecurity. All factors related to technology quantity were excluded as they all presented high multicollinearity (VIF > 80). The remaining nine factors all had VIF values under five.

To further evaluate if a linear regression model is appropriate for the data, we performed a graphical analysis of the residuals —the difference between the observed value of the dependent variable and the expected value. The goals of the analysis were to examine if the residuals 1) have a constant variance, 2) have a mean of 0, and 3) are normally distributed. Results of the residual analysis (see Appendix B) suggested that a linear regression model is adequate. China was also identified as an outlier according to our regression model —it had one of the lowest infection rates (2%), while the regression model predicted a value of 23%. This low infection rate is consistent with recent observations and reports from Microsoft [28, 29]. However, research conducted by Microsoft suggested

that these low infection rates, as measured by MSRT, may not reflect the threat landscape in China [33]. Moreover, as many systems in China use third-party software for update and patch management instead of Windows Updates, those systems are more likely to be fully patched and protected in ways that can't be measured by MSRT. Based on those potential bias, along with the residual analysis, we decided to exclude China from our regression model, reducing our sample to 186 countries.

4 Results

The five less infected countries were Aland Islands (1.4%), Japan (3%), Cayman Islands (3.4%), Finland (3.6%) and Liechtenstien (3.7%), while the five most infected countries were Ethiopia (63.8%), Iraq (54.5%), Pakistan (54.2%), Yemen (51.8%) and Sudan (51.2%). As illustrated in Figure 1, Africa and South Asia had the highest infection rates while North America and Europe had the lowest.

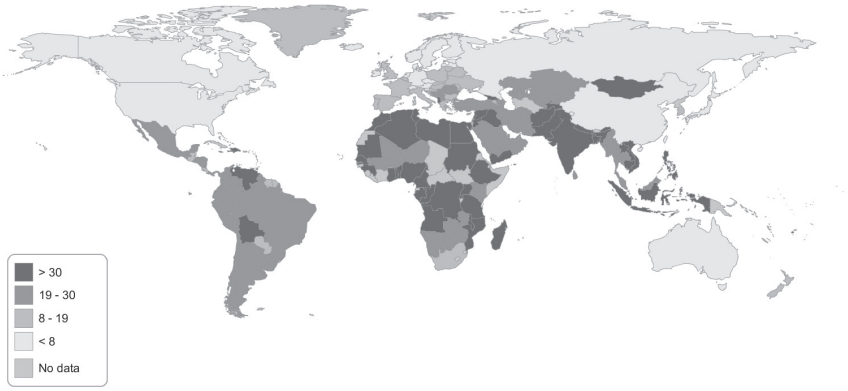


Fig. 1: Global map of malware infection rates

To better understand the geographical variations, we correlated the infection rate of the countries with factors measuring economic performance, education, technology and cybersecurity using a general linear regression model. Finally, the same analysis was conducted after categorizing countries by their socio-economic status.

4.1 Global model

In order to study how each factor individually relates to the dependent variable, we computed the Pearson correlation coefficients between the infection rate and the nine country-level factors (see Table 4). The value r, the correlation coefficient, represents the strength of the relationship between the variables. The

value ranges between -1 and 1, with a value of 0 indicating that there is no linear correlation between the variables. As not all factors were available for the 186 countries, we reported the sample size (N) used for each factor. The p-value was also computed to measure the significance of the results. A low p-value (such as 0.01) means that there is a 1 in 100 chance that we would have obtained the same results if the variables were not correlated. For the purpose of the analysis, we considered a correlation to be significant if the p-value was lower than 0.05.

Table 4: Pearson correlation coefficients between infection rate and country-level factors

Factor	r	N	p-value
GDP-log	-0.37	127	1.86e-05
Mean years schooling	-0.75	145	1.25e-27
Fixed broadband speed-log	-0.57	148	1.03e-13
%FB (256kbit/s - 2Mbit/s)-log	0.53	71	2.03e-06
%FB (2Mbit/s - 10Mbit/s)-log	-0.38	79	4.38e-04
%FB (above 10Mbit/s)-log	-0.72	65	1.21e-11
IIB (per million people)-log	-0.69	147	5.37e-22
%Protected	-0.83	186	0.00e-01
Global Cybersecurity Index	-0.38	154	1.02e-06

Table 4 shows that all factors are highly significantly (p-value < 0.001) correlated with the infection rate. Except for the variable %Fixed broadband subscriptions (256kbit/s - 2Mbit/s) that has a positive correlation, all other variables were found to have negative association with the infection rate.

Although the Pearson correlation coefficients provide insights on the dependence between two variables, it is very difficult to draw conclusions about the effect of one single factor on the dependent variable. We therefore conducted a multiple general linear regression to estimate the effect of each factor while controlling for the other factors that simultaneously affect the dependent variable. Detailed results of the regression are presented in Table 5.

For each factor, the standardized regression coefficient β and its associated standard error (Std. Error) were computed. The p-value, which is interpreted as an indicator of the significance of the results, was also computed: a low p-value indicates that the null hypothesis can be rejected with high confidence, and that the variable is relevant in the regression model. We also provided the t-value of each factor, which provides insight on the direction (positive or negative) and magnitude of the effect. The number of countries (N) used for each regression model is also provided. As not all factors were available for the 186 countries, we applied a casewise deletion method, also known as listwise deletion, to handle missing data. With this method, observations that have missing values in at least one factor are removed from the analysis. While such an approach reduces the number of countries, it has the advantage of keeping each studied variable with exactly the same number of observations.

Table 5: Global multiple general linear regression results (N=50 countries)

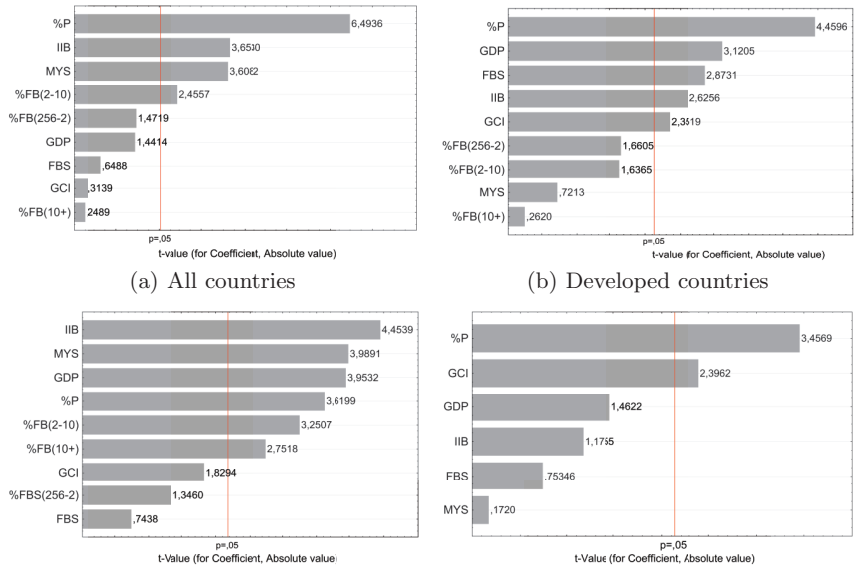
Factor	β	Std. Error	t-value	p-value
GDP-log	-0.14	0.10	-1.44	0.16
Mean years schooling	-0.31	0.08	-3.61	8.66e-04***
Fixed broadband speed-log	-0.05	0.07	-0.65	0.52
%FB (256kbit/s - 2Mbit/s)-log	-0.13	0.09	-1.47	0.15
%FB (2Mbit/s - 10Mbit/s)-log	0.25	0.10	2.46	1.86e-02*
%FB (above 10Mbit/s)-log	0.03	0.12	0.25	0.80
IIB (per million people)-log	-0.28	0.07	-3.65	7.61e-04***
%Protected	-0.65	0.10	-6.49	1.07e-07***
Global Cybersecurity Index	0.02	0.07	0.31	0.75
	R^2 adjusted	0.86		
	F-statistic	34.30		
	Degree of freedom	9		
	Df (residuals)	39		
	p-value	7.77e-16		

*Statistically significant at 0.05 level; **Statistically significant at 0.01 level; ***Statistically significant at 0.001 level.

From the regression (see Table 5), we can see that the main factors of malware infections are %Protected, international Internet bandwidth, mean years of schooling, and the percentage of fixed broadband subscriptions between 2Mbit/s and 10Mbits/s. As expected, %Protected and mean years of schooling are negatively correlated with the dependent variable. Our results also support that the quality of technology in a country may have an important effect on the rate of malware infections. Bandwidth was found to present a strong negative relationship with the infection rate while broadband speed, as measured by %FB(2-10), presents a weak positive association with the infection rate. Surprisingly, GDP and the GCI were not found to be significant after controlling for the other factors, as opposed to the results from the Pearson correlations (see Table 4). One plausible explanation is that technology quality, education and users' investment in security are channel variables between GDP, the GCI and malware infections. This would imply, for example, that GDP *per se* is not a significant factor for malware infections.

We used a Pareto chart (see Figure 2(a)) to visualize the relative importance of the nine country-level factors on the infection rate. The chart displays the absolute value of the effects (as measured by the t-value) and draws a reference line; any factor that extends beyond this line has a statistically significant impact (p-value < 0.05) on the infection rate. The main factor appears to be users' investment in security, as measured by %Protected. Bandwidth and education were found to be equivalent in their effect on the dependent variable, followed by %FB(2-10).

To evaluate the regression model we used the adjusted R^2, also known as the coefficient of determination. This number can be interpreted as how well the regression model can explain the variance of the dependent variable. In general,



(a) All countries (b) Developed countries

(c) Newly industrialized countries (d) Developing countries

Fig. 2: Pareto charts by socio-economic status

models with values over 80% are considered strong and models with values over 90% very strong. Overall, our regression model offers a strong prediction ability with an adjusted R^2 of 86%. This indicates that the nine country-level factors selected can explain 86% of the infection rate, a value that is quite high in regard to the literature known to the authors.

4.2 Model by socio-economic status

To investigate whether our previous findings apply in countries with different socio-economic development levels, we repeated all analyses after categorizing countries based on their 2013 Human Development Index (HDI) [25]. Overall, 45 countries were considered as developed (HDI >= 0.8), 74 as newly industrialized (0.8 > HDI >= 0.55), and 26 as developing (0.55 > HDI), which give us a sample of 145 countries.

As Figure 3 illustrates, there is an important variation in the malware infection rates between each category. Developed countries had the lowest infection rates, ranging from 2.9% to 26.8%, with an average of 10.4% (SD=0.06, 95% CI= 0.05-0.07). They were followed by newly industrialized countries, which had infection rates between 6.6% and 54.5% with an average of 26.3% (SD=0.10, 95% CI=0.08-0.12). The highest levels of malware infections were in developing countries, varying from 23.5% to 63.8%, with an average of 38.1% (SD=0.10, 95% CI=0.08-0.14).

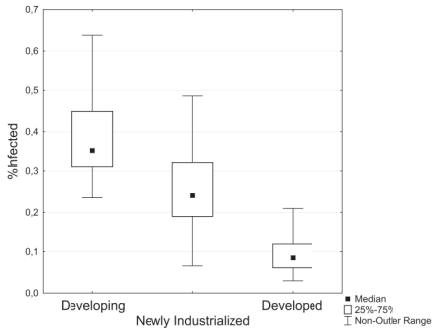


Fig. 3: Box plot of infection rates by socio-economic status

As for our previous analysis, we first computed the Pearson correlation coeffcients to investigate any potential associations between the nine country-level factors and infection rate (see Table 6). We further conducted a multiple general linear regression (see Appendix C for detailed results) by stratifying countries based on their socio-economic development to disentangle the individual effect of each factor.

Table 6: Pearson correlation coefficient between infection rate and country-level factors

	Developed			Newly industrialized			Developing		
Factor	r	N	p-value	r	N	p-value	r	N	p-value
GDP	-0.19	35	0.26	0.08	64	0.53	0.35	24	0.10
MYS	-0.68	45	2.10e-07	-0.46	74	4.26e-05	-0.21	26	0.31
FBS	-0.49	43	9.66e-04	-0.17	68	0.16	0.02	24	0.94
%FB(256-2)	0.37	33	0.03	-0.01	30	0.93	0.37	5	0.54
%FB(2-10)	-0.13	36	0.45	-0.06	33	0.76	0.73	4	0.22
%FB(10+)	-0.68	36	6.08e-06	-0.29	24	0.18	-	2	-
IIB	-0.21	43	0.18	-0.40	70	4.91e-04	0.02	25	0.93
%P	-0.81	45	1.29e-11	-0.65	74	3.08e-10	-0.54	26	4.67e-03
GCI	-0.37	44	0.01	0.13	72	0.26	-0.23	26	0.25

Developed countries The results from the Pearson correlation in Table 6 show that mean years of schooling, fixed broadband speed, fixed broadband subscriptions (above 10Mbit/s), %Protected and the Global cybersecurity index are significantly negatively associated with the infection rate. To the opposite, only the percentage of fixed broadband subscriptions (256kbit/s-2Mbit/s) has a significant positive association with the infection rate.

Five variables (%P, GDP, FBS, IIB, GCI) were identified to be potential risk and protective factors by the regression model for developed countries (see Appendix C). Factors related to cybersecurity (%P, GCI) had, as expected, a negative relationship with the infection rates, which is similar to the results of the global analysis. The quality of technology, in terms of bandwidth and speed, was also found to be negatively correlated with the dependent variable. To the opposite, economic performance (GDP) had a negative sign in the regression model. Surprisingly, mean years of schooling was not found to be a significant factor for developed countries. The insignificance of mean years of schooling can be explained by the higher education in developed countries and less variation of this variable.

Similar to the global model, users' investment in security has the stronger impact on the dependent variable (see Figure 2(b)). The second factor appears to be GDP, followed by technology quality (FBS, IIB), and the GCI. Overall, the regression model for developed countries offers a strong predictive ability as it can explain 89% of the variance of the infection rate with the nine country-level factors.

Newly industrialized countries The results of the Pearson correlation (see Table 6) for newly industrialized countries show that only mean years of schooling, international Internet bandwidth and %Protected are significantly correlated with the infection rate. All factors present a negative association, meaning they could be potential protective factors.

From the regression (see Appendix C), six variables (IIB, MYS, GDP, %P, %FB(2-10), %FB(10+)) were identified to be statistically significant. The results for cybersecurity and education are consistent with our previous findings; they both have a negative association with the dependent variable. While bandwidth presents a negative correlation, broadband speed shows a positive correlation, as opposed to what we previously found. Finally, economic performance (GDP) was negatively associated with the rate of malware infections.

As shown in the Pareto chart (Figure 2(c)), the most important factor for newly industrialized countries seems to be bandwidth. Education (MYS) and economic performance (GDP) follow with a similar impact. Users' investment in security (%P) and broadband speed (%FB(2-10), %FB(10+)) are the factors with the smallest effect. Overall, the regression model for newly industrialized countries was able to explain 79% of the variance of the infection rate with the nine country-level factors.

Developing countries Based on the results from the Pearson correlation coefficients (see Table 6), only %Protected was found to have a significant correlation with the infection rate. Countries with a higher protection coverage (%P) were associated with fewer malware infection rates.

Before conducting the regression we excluded the factors related to fixed broadband subscriptions as data were missing for many developing countries.

The results of the regression in Appendix C show that two variables, %Protected and the Global cybersecurity index, were found to be significant in the model. Those variables are both related to cybersecurity and have a negative sign, which means they could be potential protective factors of malware infections. Factors related to education (MYS), economic performance (GDP), and technology quality (IIB, FBS) were not found to be significant for developing countries. This could be explained by lower level of education, economy, and technology for developing countries, resulting in less variation in these variables.

The relative importance of the factors can be visualized by the Pareto chart presented in Figure 2(d); users' investment in security has the stronger impact, followed by the GCI. In the end, the regression model for developing countries can explain 41% of the variance of the infection rate with all six factors. This means that other factors, beyond education (MYS), economic performance (GDP), and technology quality, explain the rate of malware infections for developing countries.

5 Interpretation

Overall, we found that factors related to economic performance, education, cybersecurity, and Internet connection quality are correlated with the prevalence of malware infection in unprotected hosts. As we discuss below, however, these variables interact in some surprising ways. We also compare our results to those reported in prior studies where possible, and highlight instances in which our findings corroborate or refute theirs.

Economic performance While there is some correlation between economic activity (as measured by GDP) and lowered unprotected host infection rates (as measured by Pearson correlation), it appears this relationship is not significant in the global model after controlling for the other variables. Similarly, Garg *et al.* [17] found no significant association between economic activity (GDP-PPP) and the percentage of total number of spambots, while controlling for other macroeconomic factors. This seems plausible as here factors such as education, technology quality, and cybersecurity investment should explain most of the variance, whereas GDP *per se* should only play a minor role.

When we stratified countries based on their socio-economic status, GDP appeared to be a risk factor for developed countries. This is consistent with the hypothesis that increased GDP means increased incentives for cybercrime, as there are more individuals and organizations with significant wealth to steal from. However, the relationship appeared to be negative for newly industrialized countries; higher economic activity was correlated with reduced malware infection rates. This change of direction may be explained by a potential non-linear association; malware infections decrease as economy grows until a turning point where it rises with econonic performance, independently of other risk factors.

A first hypothesis for this relationship could be that GDP acts as a marker for technology quantity, as we removed this factor from our analysis —it was highly

correlated with other variables. This would imply that increased ICT adoption in newly industrialized countries could be associated with reduced malware infections, with those countries being more technologically developed and more resilient to malicious attacks than developing nations. However, the effect would be the opposite for developed countries: higher ICT adoption would contribute to increase malware infections, as there are more potential machines to steal from or to exploit for malicious activities (e.g. remote attacks, spam distribution). This explanation can be tested by examining the partial correlation between GDP and malware infections. In contrast to bivariate correlation, partial correlation allows one to measure the association between two variables while controlling for the effect of other factors. We first computed the correlations while controlling for education, technology quality, and cybersecurity. As expected, the association was positive for developed countries (r=0.54, N=25, p-value=2.1e-02) and negative for newly industrialized countries (r=-0.67, N=22, p-value=6.0e-03). We then added the percentage of households with Internet (%HouseholdInternet) to account for ICT penetration. Results show that the associations are still statistically significant for developed countries (r=0.55, N=25, p-value=2.2e-02) and newly industrialized countries (r=-0.67, N=22, p-value=9.0e-03), suggesting that ICT penetration cannot explain the non-linear relationship between GDP and malware infections.

A second possibility is that software piracy acts as an intermediate factor between GDP and malware infections. This hypothesis is plausible as software piracy has often been associated with increased risk of malware infections [27, 4]. As economic activity increases, adoption of legal software should also rise [18, 6]. In contrast, higher economic activity for developed countries could be associated with higher software piracy. While this may be counter-intuitive, Fischel *et al.* [16] found evidence that software piracy is positively correlated with income for West European and North American countries. To investigate this relationship, we computed the partial correlations between GDP and malware infections while controlling for education, technology quality, cybersecurity, and software piracy. This last factor was collected from the Business Software Alliance and represents the national ratio of the number of unlicensed software units installed to the total number of software units installed for 2011 [9]. Results show that the associations still hold for developed countries (r=0.63, N=25, p-value=7.0e-03) and newly industrialized countries (r=-0.71, N=18, p-value=2.1e-02). This suggests that software piracy may not account for the relationship between GDP and malware infections.

A third explanation could be the distribution of the different versions of Windows (e.g. XP, Vista, 7, 8 and 8.1), i.e., how the operating system (OS) market of a country is shaped could influence his rate of malware infections. To examine this possibility we looked at the distribution of the OS market for unprotected hosts between developed and newly industrialized countries (see Table 8 in Appendix D). Rates were similar for XP, Vista, and 8.1, but different for 7 and 8. We therefore decided to include the prevalence of Windows 8 (%Windows8), as it is the platform with the highest difference between developed (Mean=10%,

SD=4%) and newly industrialized countries (Mean=17%, SD=7%). Partial correlations were computed with education, technology quality, cybersecurity, and prevalence of Windows 8 for unprotected hosts as control variables. This time, both the associations for developed (r=0.26, N=25, p-value=2.6e-01) and newly industrialized countries (r=-0.36, N=22, p-value=2.0e-01) were found to be not statistically significant. This suggests that the OS market distribution may be an intermediate factor between GDP and malware infections. Although our analysis provides empirical support for this explanation, it is necessary to develop and test new theories that can account for the causes of this relationship.

Overall, results suggest that GDP *per se* is not a significant factor of malware infections. Rather, economic performance would act as a distal factor via multiple intermediate variables (e.g. technology quality, OS market distribution) that were captured in our analysis.

Education Education seems to be more consistently associated with reduced malware infection rates. As expected, mean years of schooling was negatively correlated with malware infections in the global analysis. When we stratified countries by socio-economic status, education was only significant for newly industrialized countries. Similarly, Microsoft [8] found that countries with high education, as measured by the literacy rate, are less likely to be infected by malware.

Overall, our analysis suggests that education is a significant distal factor of malware infections. This could imply that education is involved in the causal chain via a number of intermediate factors (e.g. IT literacy, cybersecurity awareness) that were not captured by our analysis. Another potential explanation is that mass education, as measured by mean years of schooling, has a direct aggregate effect at the population level. Testing those hypotheses would require the collection of more specific data on potential intermediate factors, both at the population and individual level. From there, additional studies could be designed to disentangle the aggregate effect of education, if any, from its indirect effect on malware infections.

Technology The quality of a country's technological infrastructure does seem to be correlated with reduced malware infections. Increased international Internet bandwidth and high fixed broadband speed were both associated with reduced unprotected host infection rates when looking at the bivariate correlations. After controlling for economic development, education level, and cybersecurity investment, bandwidth was found to be a protective factor, regardless of the socio-economic development level. The effect of broadband speed in terms of direction and magnitude, however, turned out to depend of the socio-economic status. While higher broadband speed (FBS) was negatively correlated with malware infections for developed countries, higher proportions of moderate (%FB(2-10)) and high speed fixed broadband (%FB(10+)) were actually positively correlated with infections for newly industrialized countries.

One explanation for this inconsistent relationship between the quality of Internet connectivity and infection rates is that while better bandwidth makes it

easier to keep systems updated, faster connections make it easier for attackers to exploit large populations of unmaintained systems. To partially investigate this hypothesis, we first looked for associations between measures of system status and bandwidth. We computed from MSRT the percentage of users that had out of date AV signatures during the 4-month period and the percentage of users who performed their Windows updates every month during the study. As expected, the first measure (%Out-of-date AVs) was negatively correlated with bandwidth (r=-0.59, N=186, p-value=4.47e-15) while the second measure (%Up-to-date Systems) was positively correlated (r=0.75, N=186, p-value=3.13e-28). This suggests that bandwidth could be a protective factor for malware infections though various measures of system status as intermediate variables. However, testing the second part of the hypothesis —that faster connection makes it easier to infect large populations of vulnerable computers— would require conducting large-scale studies of malware propagation based on epidemiological models.

Overall, these findings provide evidence that bandwidth could be a protective factor that contributes to decreased risk of malware infections via multiple intermediate variables related to system status. Moreover, results suggest that fast broadband connections are associated with reduced malware infections in only some circumstances. Further studies are required to determine the exact nature of the causal relationship, if any.

Cybersecurity Individual investment in security (as measured by %Protected) appeared to have a strong negative correlation with malware infections for all countries, regardless of their socio-economic status. Intuitively, countries with higher percentage of users protected by antivirus products were found to have lower unprotected host infection rates. A first explanation for this observation is that antivirus product penetration acts as a marker for other variables that were not captured by our analysis. For example, usage of antivirus products may be related to individual risk-taking behavior —users who tend to underestimate cybersecurity-related risk may tend to unprotect their computer. Hence, AV penetration could be a marker for risk-attitude towards cybersecurity at the population level. One potential way to investigate this hypothesis would be to correlate AV market penetration with individual risk-taking behavior in other specific contexts, such as finances, sports and leisure, health, career, and car driving [12]. As a first attempt, we correlated %Protected with tobacco consumption. We used 2012 male smoking prevalence among persons aged 15 years and over from the World Health Organization [2] as an aggregate measure of risk attitude in the domain of health [19, 12, 15]. We first performed a bivariate correlation using Pearson correlation coefficient to investigate any linear association between the two variables: results indicate that the association is not significant (r=-0.019, N=104, p-value=0.845). As smoking is a function of various determinants (e.g. education, income, social support) beyond risk attitude, we also performed a partial correlation. We used the mean years of schooling (MYS) and GDP as markers of socio-economic status. This time, results of the partial correlation reveal a weak negative association between %Protected and

smoking prevalence (r=-0.22, N=92, p-value=0.036), suggesting that AV penetration may be a marker of risk-taking behavior towards cybersecurity at the population level. Although our analysis provides limited empirical support, validation of this hypothesis would require to conduct either country level studies based on aggregated measures of cybersecurity risk-aversion or large-scale user studies. A second but tenuous possibility is that unprotected systems benefit indirectly from the protection of other —protected— systems. This is similar to the free-rider effect in economics, where non-paying individuals can benefit from the goods, resources or services of others, even though they did not pay for them. Unprotected hosts would then benefit from a "AV herd immunity" effect from systems protected by antivirus products. Even though prior work [7, 23] have provided some empirical evidence for this explanation, proper validation should be achieved by conducting further epidemiological studies designed for the purpose.

The Global cybersecurity index was found to be a weak protective factor for developed and developing countries —its effect was not significant in the global model and for newly industrialized countries. In comparison, previous work [21, 32, 43] provided limited empirical evidence of the effect of national policies on cybersecurity. However, those results can't be directly compared to our research, as previous studies used various proxy variables to evaluate the impact of cybersecurity policies. Overall, our results tend to confirm that investment in cybersecurity at the national level, as measured by the GCI, is associated with reduced unprotected host infections. From there, further studies could be conducted to understand the individual contribution of each component of the GCI (legal measures, technical measures, organizational measures, capacity building or cooperation) and help design better evidence-based cybersecurity policies.

6 Study limitations

This study and its conclusions are subject to a number of limitations and potential bias. First, there is an inherent limitation to our results because our sample population is drawn from Windows systems running MSRT; thus, it does not provide insight into Windows systems that do not run Windows Update, and it does not give insight into the infection rates on non-Windows systems such as MacOS and Unix-based OSes. Furthermore, the analysis was limited to personal computers (e.g. desktop and laptop) meaning that the factors identified may differ significantly on mobile devices and tablets. However, given that there are more than one billion computers regularly running MSRT, patterns discovered in this population are important on their own, whether or not they are representative of patterns in other computational contexts.

Another significant limitation is that the detections from MSRT are only for a subset of malware families. While these families may represent some of the most significant malware families on Windows, they are not representative of the entire threat landscape, and so MSRT reported infection rates will be different from the overall malware infection rate. Nevertheless, given the significance

of MSRT-targeted malware these infection rates are also of inherent interest. Detections by MSRT are also dynamic and fluctuate over time [8]. To partially compensate this volatility, we used the *period prevalence* of malware infections, that is the prevalence during a specific period of time. While period prevalence may be a better measure than averaged prevalence, our measurement may still be subject to temporal variation, as is often the case for security data [14]. Moreover, malware infection rates reported in this study may not be representative of other time frames.

As this was an observational study at the population level, we only intended to identify correlations to generate hypotheses on the causes of malware infections. We did not attempt to infer causation. While ecological studies can be used to identify potential factors based on aggregate variables, care must be taken to avoid the risk of ecological fallacy —an error in the interpretation of the results when conclusions are inappropriately made about individuals based on aggregated data. The fallacy assumes that individual members of a group all have the average characteristics of the group. Another limitation of ecological studies is their susceptibility to confounding. Both economic and education factors have been considered control variables in our study. We cannot ensure, however, that our results are not affected by other unknown extraneous variables. Although we included a broad set of national-level factors in our study, there may be other plausible predictors of malware infections that were not captured though our analysis. It would be interesting in future work to consider additional factors, such as culture, demographics, technology quality, or private investment in security, as the latter two were excluded due to high multicollinearity.

7 Conclusion and policy implications

We presented the results of the first ecological study applied to computer security designed to identify national-level malware infection risk factors. We found relationships with economic performance, education, Internet connectivity, and cybersecurity that have not been previously reported, particularly in how their relationships with infection rates are not simple correlations. We also explored in detail the potential underlying causes between the studied national-level factors and malware infections.

While our work corroborates some findings in earlier research, our results suggest that GDP *per se* is *not* a significant factor of malware infections. Rather, economic performance could be a distal factor acting through multiple intermediate variables, such as technology quality, OS market distribution, or education. The later, as measured by mean years of schooling, was also found to be a protective distal factor of malware infections. However, the question of whether it is a direct aggregate effect or an indirect effect should be investigated in further studies. We also found evidence that bandwidth acts as a protective factor of malware infections via multiple variables of system status. Interestingly, results suggest that Internet connection quality, as measured by broadband speed, may be a protective factor in only some circumstances. While high broadband

speed was associated with reduced malware infections for developed countries, its effect was the opposite for newly industrialized countries. The percentage of AV-protected machines and the Global cybersecurity index were also found to be significant protective factors. This suggest that investment in cybersecurity, both at the individual and national level, could contribute to reduce the risk of malware infections. Finally, our work shows that risk and protective factors may not have the same effect and relative importance in countries with different socio-economic status.

More interestingly, our findings have potential policy implications. For example, education was identified as a major protective factor —countries with higher level of education had lower malware infection rates. Although education was measured at the population level, this may suggest that governments should prioritize investments in user education. Such efforts could focus, among others, on the promotion of safe computer behavior, like installing an AV product, or keeping applications, software and OS up-to-date. However, although user education may foster the adoption of safe computer behavior, it is possible that many risky computer behaviors, particularly in developing and newly industrialized countries, are also determined by income. For example, users in such countries may face a trade-off between buying a legitimate software or downloading a pirated software and saving money. Prior understanding of how risky computer behavior is determined by a lack of cybersecurity risk awareness and the costs of adopting safety measures and behaviors would therefore be essential in the success (or failure) of such interventions.

Similarly, technology quality (as measured by bandwidth) was also identified as a protective factor. While users can install free AV products, they will not be fully effective if their signature databases cannot be updated as a result of poor Internet connection. This could also suggest that governments should invest in ICT infrastructure. On the other hand, investing in better ICT infrastructure on the basis of risk reduction alone might not represent a sufficiently great value proposition for developing and newly industrialized countries. Moreover, we found evidence that higher broadband speed was associated with higher malware infection rates for newly industrialized countries, while its effect was the opposite for developed countries. Hence, interventions proven to be successful in developed countries might not be effective (or even possible due to budget constraints) in newly industrialized and developing countries.

In light of this discussion, we believe that policy interventions, whether technical, legal, or educational, might prove ineffective if they do not take into account the socio-economic circumstances of populations and individuals. As shown by our findings, it is important to consider the socio-economic status of countries in future risk analysis of security threats and consequent evidence-based decision making. Moreover, the relative effect of protective factors were found to differ depending on the socio-economic context. This suggests a prioritisation of efforts by policy makers, where stronger protective factors should be leveraged first. For example, for newly industrialized countries it would seem that increasing bandwidth availability would have stronger effect than increas-

ing AV usage. In contrast, the opposite is true for developed countries. In both cases, however, this prioritisation of effort must also take into account the relative cost-effectiveness of such counter-measures, e.g. can a similar effect be more effectively obtained by investing the same amount of resources to address one factor vs. the other. Assessing the cost-effectiveness of such counter-measures is beyond the scope of our study.

This work also demonstrates that rigorous ecological studies can be used to identify risk factors for malware infections at the population level. We believe a population health approach could provide a skeleton from which security threats can be researched and for which appropriate national-level interventions can be developed. It is important that further research be conducted to assess the multi-level risk factors of malware infections, in order to verify some of the hypotheses we have advanced and establish sound causation. Since explaining individual cases requires that we consider both underlying causes of infection in the population and individual circumstances, research into the different levels of risks should be seen as complementary. From there, cybersecurity policies could be designed to reduce the prevalence of malware considering both individual and ecological influences. We hope this work illustrates the merits of future large-scale ecological studies applied to computer security.

8 Acknowledgements

The authors would like to thank the Microsoft Malware Protection Center (MMPC) for granting us access to the MSRT telemetry data and for supporting this work. First author would also like to thank Thierry Lavoie for his useful comments and suggestions on designing the study in the paper. Finally, we would like to thank our anonymous reviewers who provided many helpful comments on the paper.

References

1. Malware Families Cleaned by the Malicious Software Removal Tool. `http://www.microsoft.com/security/pc-security/malware-families.aspx`.
2. Tobacco use Data by country. `http://apps.who.int/gho/data/node.main.65`.
3. World Telecommunication/ICT Indicators database. `http://www.itu.int/en/ITU-D/Statistics/Pages/publications/wtid.aspx`.
4. Unlicensed Software and Cybersecurity Threats. `http://globalstudy.bsa.org/2013/Malware/study_malware_en.pdf`, 2013.
5. M. O. Akinwande, H. G. Dikko, and A. Samson. Variance inflation factor: As a condition for the inclusion of suppressor variable (s) in regression analysis. *Open Journal of Statistics*, 5(07):754, 2015.
6. A. R. Andrés. Software piracy and income inequality. *Applied Economics Letters*, 13(2):101–105, 2006.
7. J. Blackbird and B. Pfeifer. The global impact of anti-malware protection state on infection rates. In *23th Virus Bulletin International Conference*, 2013.
8. D. Burt, P. Nicholas, K. Sullivan, and T. Scoles. The Cybersecurity Risk Paradox: Impact of social, economic, and technological factors on rates of malware. Technical report, Microsoft, 2014.

9. Business Software Alliance. 2011 BSA Global Software Piracy Study. Technical report.
10. Y. Carlinet, L. M, H. Dbar, and Y. Gourhant. Analysis of computer infection risk factors based on customer network usage. In *Second International Conference on Emerging Security Information, Systems and Technologies (SECURWARE'08)*, pages 317–325., 2008.
11. S. Charney. Collective defense: Applying public health models to the internet. *white paper (Redmond, Wash.: Microsoft Corporation, 2010), http://www. microsoft. com/security/internethealth*, 2010.
12. T. Dohmen, A. Falk, D. Huffman, U. Sunde, J. Schupp, and G. G. Wagner. Individual risk attitudes: Measurement, determinants, and behavioral consequences. *Journal of the European Economic Association*, 9(3):522–550, 2011.
13. T. Dumitras. Field data available at symantec research labs: The worldwide intelligence network environment (wine).
14. B. Edwards, S. Hofmeyr, S. Forrest, and M. van Eeten. Analyzing and modeling longitudinal security data: Promise and pitfalls. In *Proceedings of the 31st Annual Computer Security Applications Conference*, pages 391–400. ACM, 2015.
15. R. M. Feinberg. Risk aversion, risk, and the duration of unemployment. *The Review of Economics and Statistics*, pages 264–271, 1977.
16. J. A. Fischer and A. Rodriquez Andrés. Is software piracy a middle class crime? investigating the inequality-piracy channel. *University of St. Gallen Economics Discussion Paper*, (2005-18), 2005.
17. V. Garg, T. Koster, and L. J. Camp. Cross-country analysis of spambots. *EURASIP Journal on Information Security*, 2013(1):3, 2013.
18. R. K. Goel and M. A. Nelson. Determinants of software piracy: economics, institutions, and technology. *The Journal of Technology Transfer*, 34(6):637–658, 2009.
19. J. Hersch and W. K. Viscusi. Cigarette smoking, seatbelt use, and differences in wage-risk tradeoffs. *Journal of Human Resources*, pages 202–227, 1990.
20. D. Kindig and G. Stoddart. What is population health? *American Journal of Public Health*, 93(3):380–383, 2003.
21. A. Kleiner, P. Nicholas, and K. Sullivan. Linking Cybersecurity Policy and Performance. Technical report, Microsoft Trustworthy Computing, 2013.
22. F. Lalonde Levesque, J. Nsiempba, J. M. Fernandez, S. Chiasson, and A. Somayaji. A clinical study of risk factors related to malware infections. In *Proceedings of the 2013 ACM SIGSAC Conference on Computer & Communications Security*, pages 97–108. ACM, 2013.
23. F. Lalonde Levesque, A. Somayaji, D. Batchelder, and J. M. Fernandez. Measuring the health of antivirus ecosystems. In *Malicious and Unwanted Software: The Americas (MALWARE), 2015 10th International Conference on*. IEEE, 2015.
24. G. Maier, A. Feldmann, V. Paxson, R. Sommer, and M. Vallentin. An assessment of overt malicious activity manifest in residential networks. In *Detection of Intrusions and Malware, and Vulnerability Assessment*, pages 144–163. Springer, 2011.
25. K. Malik and Jespersen. Human Development Report 2013. Technical report, United Nations Development Programme, 2013.
26. M. Menting. Global Cybersecurity Index. Technical report, ABI Research, 2014.
27. G. Mezzour, K. M. Carley, and L. R. Carley. An empirical study of global malware encounters. In *Proceedings of the 2015 Symposium and Bootcamp on the Science of Security*, page 8. ACM, 2015.
28. Microsoft Corporation. Microsoft Security Intelligence Report Volume 17. Technical report, 2014.

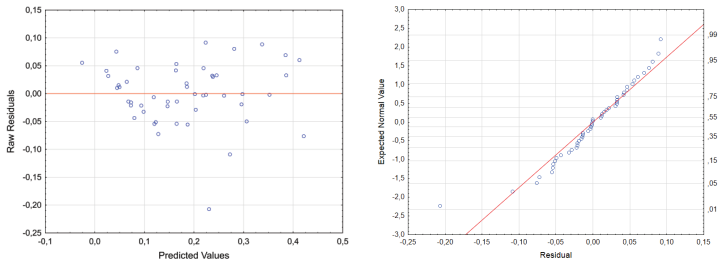
29. Microsoft Corporation. Microsoft Security Intelligence Report Volume 18. Technical report, 2014.
30. D. K. Mulligan and F. B. Schneider. Doctrine for cybersecurity. *Daedalus*, 140(4):70–92, 2011.
31. K. Onarlioglu, U. O. Yilmaz, E. Kirda, and D. Balzarotti. Insights into user behavior in dealing with internet attacks. In *NDSS*, 2012.
32. I. P. Png, C.-Y. Wang, and Q.-H. Wang. The deterrent and displacement effects of information security enforcement: International evidence. *Journal of Management Information Systems*, 25(2):125–144, 2008.
33. T. Rains. The threat landscape in china: A paradox. `https://blogs.microsoft.com/cybertrust/2013/03/11/the-threat-landscape-in-china-a-paradox/`, 2013.
34. M. Rice, J. Butts, R. Miller, and S. Shenoi. Applying public health strategies to the protection of cyberspace. *International Journal of Critical Infrastructure Protection*, 3(3):118–127, 2010.
35. B. Rowe, M. Halpern, and T. Lentz. Is a public health framework the cure for cyber security? *CrossTalk*, 25(6):30–38, 2012.
36. J. Rowe, K. Levitt, and M. Hogarth. Towards the realization of a public health model for shared secure cyber-space. In *Proceedings of the 2013 New Security Paradigms Workshop*. ACM, 2013.
37. S. Sheng, M. Holbrook, P. Kumaraguru, L. F. Cranor, and J. Downs. Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions. In *ACM Conference on Human Factors in Computing Systems (CHI)*, pages 373–382, 2010.
38. V. Subrahmanian, M. Ovelgonne, T. Dumitras, and A. Prakash. The global cyber-vulnerability report, 2016.
39. K. Sullivan. The internet health model for cybersecurity. 2012.
40. The World Bank. World Development Indicators. `http://data.worldbank.org/products/wdi`.
41. O. Thonnard, L. Bilge, A. Kashyap, and M. Lee. Are you at risk? profiling organizations and individuals subject to targeted attacks.
42. United Nations Development Programme. Human Development Repoers. `http://hdr.undp.org/en/data`.
43. M. van Eeten, J. M. Bauer, H. Asghari, S. Tabatabaie, and D. Rand. The role of internet service providers in botnet mitigation: An empirical analysis based on spam data. Technical report, OECD Publishing, 2010.
44. T.-F. Yen, V. Heorhiadi, A. Oprea, M. K. Reiter, and A. Juels. An epidemiological study of malware encounters in a large enterprise. In *Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security*, pages 1117–1130. ACM, 2014.
45. H. Zainodin, G. Khuneswari, A. Noraini, and F. Haider. Selected model systematic sequence via variance inflationary factor. *International Journal of Applied Physics and Mathematics*, 5(2):105, 2015.

A Description of country-level factors

Factor	Definition
GDP per capita	GDP converted from domestic currencies to U.S. dollars using single year official exchange rates.
GDP per capita by purchasing power parity	GDP per capita based on purchasing power parity.
Mean years of schooling	Average number of years of education received by people ages 25 and older, converted from education attainment levels using official durations of each level.
%Households with computer	Percentage of households with computer.
%Households with Internet	Percentage of households with Internet.
Fixed (wired) Internet subscriptions (per 100 inhabitants)	Number of active fixed (wired) Internet subscriptions at speed less than 256 kbits/s and the total fixed (wired) broadband subscriptions.
Fixed (wired) broadband subscriptions (per 100 inhabitants)	Number of fixed (wired) broadband subscriptions with access over wireline networks. Wireless broadband is not included.
Fixed (wired) broadband speed	Refers to the advertised maximum theoretical download speed, and not speeds guaranteed to users associated with a fixed (wired) broadband Internet monthly subscriptions. It does not refer to the actual speed delivered.
Fixed broadband subscriptions between 256 kbits/s and less than 2 Mbits/s	Percentage of Internet broadband subscriptions with advertised downstream speed equal to 256 kbits/s and less than 2 Mbits/s.
Fixed broadband subscriptions between 2 Mbits/s and less than 10 Mbits/s	Percentage of Internet broadband subscriptions with advertised downstream speed equal to 2 Mbits/s and less than 10 Mbits/s.
Fixed broadband subscriptions above 10 Mbits/s	Percentage of Internet broadband subscriptions with advertised downstream speed equal to, or greater than 10 Mbits/s.
International Internet bandwidth	Total used capacity of international Internet bandwidth, in megabits per second. Measures the sum of used capacity of all Internet exchanges offering international bandwidth.
Secure Internet servers (per 1 million people)	Number of secure Internet servers using encryption technology in Internet transactions.
%Protected	Refers to the percentage of users that have at least one antimalware product actively running with up-to-date signatures.
Global cybersecurity index	Index of level of cybersecurity development in terms of legal measures, technical measures, organizational measures, capacity building and cooperation.

B Residual analysis

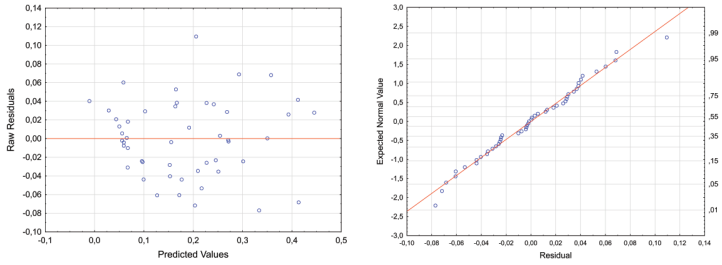
We first plotted the raw residuals versus the predicted values to examine if the raw residuals have a constant variance, and a mean of 0. As depicted in Figure 4(a), the fitted line plot shows that the mean is 0 and that the assumption of equal variance does not seem to be violated. The plot of the expected normal value versus the raw residual was also examined to see if the residuals follow a normal distribution. Visual inspection of the plot (see Figure 4(b)) suggests that the residuals follow a straight line, meaning that a linear regression model is adequate.



(a) Raw residual vs predicted values (b) Expected normal value vs raw residual

Fig. 4: Residual analysis

The graphical analysis also suggested that one observation (China) may be an outlier. We computed for each country the associated standardized residual (also known as the studentized residual) to identify potential outliers. In general, an absolute value larger than 3 indicates that the observation is an outlier. Results of the analysis (3.46) confirmed that China is an outlier according to our regression model. We then performed another residual analysis excluding China (see Figure 5(a) and Figure 5(b)), which also suggested that a linear regression model is appropriate for the data.



(a) Raw residual vs predicted values (b) Expected normal value vs raw residual

Fig. 5: Residual analysis without China

C Multiple general linear regression results

Table 7: Multiple general linear regression results by socio-economic status

Factor	Developed (N=25)				Newly industrialized (N=22)				Developing (N=22)			
	β	Std. Error	t-value	p-value	β	Std. Error	t-value	p-value	β	Std. Error	t-value	p-value
GDP-log	0.47	0.15	3.12	7.01e-03***	-0.91	0.23	-3.95	1.92e-03***	0.29	0.20	1.46	0.16
Mean years schooling	-0.08	0.11	-0.72	0.48	-0.71	0.17	-3.99	1.79e-03***	0.03	0.19	0.17	0.86
Fixed broadband speed-log	-0.24	0.08	-2.87	0.01**	0.12	0.16	0.74	0.47	0.16	0.21	0.75	0.46
%FB (256kbit/s - 2Mbit/s)-log	-0.19	0.11	-1.66	0.12	0.29	0.22	1.34	0.20	-	-	-	-
%FB (2Mbit/s - 10Mbit/s)-log	-0.28	0.17	-1.63	0.22	0.67	0.21	3.25	6.95e-03***	-	-	-	-
%FB (above 10Mbit/s)-log	-0.03	0.13	-1.02	-0.26	0.69	0.25	2.75	1.75e-02**	-	-	-	-
IIB (per million people)-log	-0.31	0.11	-2.62	0.01**	-0.74	0.17	-4.45	7.87e-04***	0.28	0.24	1.17	0.26
%Protected	-0.80	0.17	-4.45	4.59e-04***	-0.51	0.14	-3.62	3.51e-03***	-0.72	0.0.21	-3.46	3.52e-03***
Global Cybersecurity Index	-0.21	0.09	-2.35	0.03*	0.30	0.17	1.83	0.09	-0.51	0.21	-2.40	0.03*
	R^2 adjusted			0.89	R^2 adjusted			0.79	R^2 adjusted			0.41
	F-statistic			22.88	F-statistic			9.73	F-statistic			3.39
	Df			9	Df			9	Df			6
	Df (residuals)			15	Df (residuals)			21	Df (residuals)			15
	p-value			3.41e-07	p-value			2.82e-04	p-value			0.02

*Statistically significant at 0.05 level; **Statistically significant at 0.01 level; ***Statistically significant at 0.001 level.

D Windows versions statistics

Table 8: Distribution of Windows versions by socio-economic status

	Developed		Newly industrialized		Developing	
Version	Mean	SD	Mean	SD	Mean	SD
Windows XP	0.18	0.06	0.14	0.08	0.14	0.06
Windows Vista	0.08	0.03	0.03	0.02	0.03	0.01
Windows 7	0.54	0.04	0.57	0.08	0.53	0.07
Windows 8	0.10	0.04	0.17	0.07	0.23	0.05
Windows 8.1	0.09	0.06	0.08	0.03	0.07	0.02