Economics of Cyber Security Individual Report

Ramon Houtsma

December 31, 2018

`https://github.com/RamonH93/eos_proj/`

Abstract

This paper considers the security issue of phishing. The geographic distribution of phishing sites over the past 10 years is investigated, to see whether the findings in literature 10 years still hold. We wanted to find out whether one continent was targeted significantly more by phishing sites than the others. By comparing the means of the number of phishing sites per 1000 citizens of countries in the continent with the rest of the world, we can find whether the mean is significantly higher using a t-test. We find that the mean is significantly higher for North America only. This result is in line with literature, and can be explained because websites in North America reach a worldwide audience.

1 Introduction

This paper considers the security issue of phishing. Even though phishing has been a problem for over 10 years, it still costs society a lot to this day, because personal data are stolen or financial losses are suffered because people are deceived by legitimate looking phishing websites. The study is performed using a dataset that contains phishing website records from 2007 until 2017 (see fig. 1).

The structure of this paper is as follows: first we will perform a literature review in section 2 to provide context for the current state of the research topic. Next, in section 3 the research question and hypotheses are defined. In section 4 the methodology of the research is describe, and in section 5 the results are discussed. Finally, limitations are recognized and we finish with a conclusion.

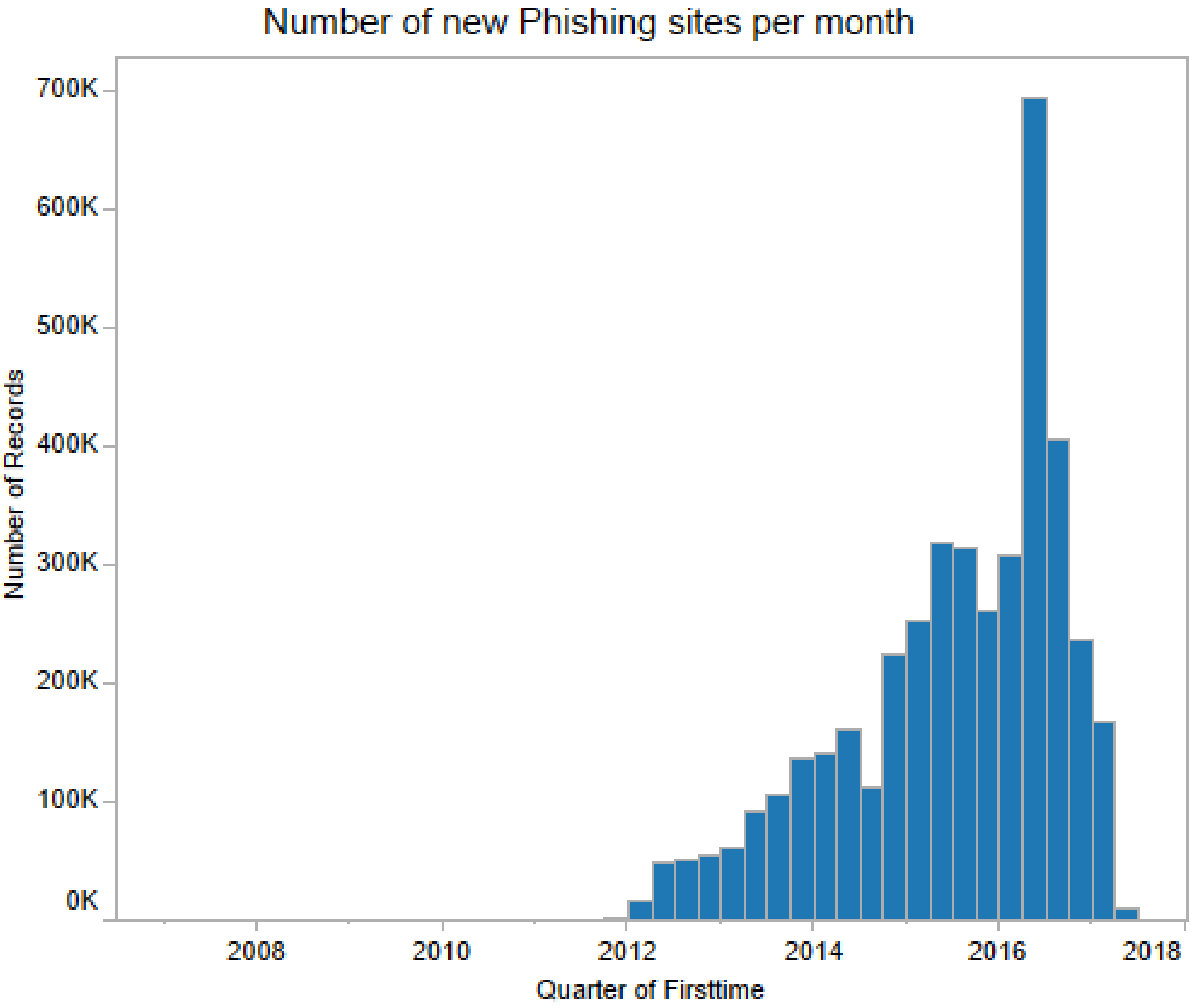

Figure 1: New phishing sites per quarter recorded in the dataset.

2 Literature Review

Phishing works because a large amount of users (23%) do not look at cues such as the address bar, leading to incorrect choices 40% of the time[1]. Phishing remains a problem to this day because even the most sophisticated detection algorithms only correctly classify up to 90% of the phishing websites[2]. Detecting phishing sites using moderators is also a possibility, such as is the case with PhishTank[3] where users submit phishing websites, however results vary heavily based on the number of moderators and their accuracy[4]. For a company, hiring such a team might be expensive as well. It appears that the effectiveness of training workers on phishing awareness is not significantly effective either[5]. We can conclude that although progress is made in phishing detection and prevention, it remains a problem to this day. Therefore it might be interesting to investigate where the phishing attacks are aimed at. In 2007 Garera et. al.[6] found that 70.3% of the phishing websites in their dataset were hosted in the United States. It is worth investigating what the geographical distribution of phishing sites is now, ten years later.

3 Aim

3.1 Objective

Based on the literature and the data available, a study was conducted to find out if there are any differences between continents for how often companies in that continent are targeted by phishing websites. With this research we can find out whether the geographical distribution of phishing sites found in 2007 by Garera et.al.[6] still holds in 2017.

3.2 Research Question

The research question is formulated as follows:

Are there one or more continents that are targeted by phishing websites significantly more than the rest of the world?

3.3 Hypothesis

The research question leads to six null hypotheses and alternative hypotheses for each of the six continents:

Null hypotheses h_0: There are equal number of phishing sites per citizen in *[continent]* as in the rest of the world.

Alternative hypotheses h_1: There are not equal number of phishing sites per citizen in *[continent]* as in the rest of the world.

4 Methodology

In order to answer the research question, the aim was to reject the null hypothesis. This could be researched by comparing two populations: North America and the rest of the world. Specifically, the number of phishing sites targeted at countries within these continents are compared. The number of phishing sites are normalized for the population of the country. Now, the means of the normalized number of phishing sites within the populations could be compared. The populations are independant because the subjects (countries) are only used once, i.e. they are unique for each sample.

The *Independant Samples t-test* is the statistical test of choice for this case, because the means two independant samples are compared. Hypothesis h_0 can be rejected within if the p-value $\leq \alpha$. The *Levene's Test for Equality of Variances* is used to determine whether equal variances can be assumed for the t-test or not.

5 Results

Python[7] and SPSS[8] were used to apply the statistical tests. Python was used to mangle and combine the data sources. The resulting table was used as input for SPSS, where the Levene's Tests for Equality of Variances and Independant Samples t-tests were conducted.

We are comparing the means of the number of phishing sites per 1000 citizens per subject within a continent. Population data was used from *GeoNames*[9]. For each of the following cases, hypothesis h_0 can be rejected within 95% confidence interval if the p-value $\leq \alpha$, with $\alpha = 0.05$.

5.1 North America versus Rest of the World

The hypothesis that were tested are formulated as such:

h_0=There are equal number of phishing sites per citizen in North America as in the rest of the world.
h_1=There is a significant difference in number of phishing sites per citizen in North America as in the rest of the world.

The descriptive data for this test are found in table 1. The results of the statistical tests can be reviewed in table 2. The Levene's Test for Equality of Variances is significant: $0,000 \leq 0,05$, therefore equality of variances can be assumed. With equality of variances assumed, the results of the Independant Samples t-test is significant as well with $0,004 \leq 0,05$. Therefore we can reject h_0 and accept h_1, and conclude that the mean of number of phishing sites per 1000 citizens in North America is significantly different from the rest of the world. From the descriptive data we can infer that there are in fact a lot more phishing sites per 1000 citizens, almost 10 times as many. This can be explained by the fact that most of the continent is English, and that websites targeting companies in that country are also used by foreigners.

	NA_WW	N	Mean	Std. Deviation	Std. Error Mean
Phishing sites per 1000 citizens	Rest of World	162	1,0530	2,763	,217
	North America	30	9,837	38,179	6,971

Table 1: Descriptive data for comparison of North America with rest of the world.

	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
Equal variances assumed	32,763	,000	-2,921	190	,004
Equal variances not assumed			-1,260	29,056	,218

Table 2: Levene's Test and Independant Samples t-test for comparing means of number of phishing sites in North America with the rest of the world.

5.2 Comparison of other continents

The same tests as in section 5.1 were conducted for South America, Europe, Africa, Asia and Oceania. The results are summed up in table 3. For these continents the p-value > 0.05, which meant that h_0 could not be rejected. This means that these continents are not targeted significantly more or less by phishing sites than other continents.

	Levene's Test for Equality of Variances		t-test for Equality of Means		
	F	Sig.	t	df	Sig. (2-tailed)
South Africa	,606	,437	1,639	189,739	,103
Europe	1,514	,220	,557	158,539	,579
Africa	1,234	,268	1,447	188,200	,15
Asia	1,875	,173	1,450	154,984	,149
Oceania	,296	,587	1,205	108,924	,231

Table 3: Statistical test results of comparisons of continents with the rest of the world.

6 Limitations

This research is based on the assumption that phishing sites targeting companies within a country only affect citizens of that country. However in this day and age this is a weak assumption, especially when .com or .net domains are used worldwide. Therefore it is not recommended to conduct further research.

7 Conclusions

In this paper we conducted research on the current geographical distribution of phishing websites. We compared the mean of the number of phishing sites per 1000 citizens for all countries within a continent, with the mean of the rest of the world. With the results we can answer our research question:

Are there one or more continents that are targeted by phishing websites significantly more than the rest of the world?

North America is targeted significantly more statistically by phishing websites than the rest of the world. This can be explained by the fact that phishing sites reach more citizens by focusing on companies within North America, because their websites are also used a lot by citizens worldwide. This result supports the current literature, because in 2007 Garera et. al.[6] found that the United States was targeted the most as well. We can conclude that their findings still hold in 2017.

References

[1] R. Dhamija, J. D. Tygar, and M. Hearst, "Why Phishing Works," ser. CHI '06, New York, NY, USA: ACM, 2006, pp. 581–590, ISBN: 978-1-59593-372-0. DOI: `10.1145/1124772.1124861`. [Online]. Available: `http://doi.acm.org/10.1145/1124772.1124861`.

[2] C. Whittaker, B. Ryner, and M. Nazif, "Large-Scale Automatic Classication of Phishing Pages," en, p. 14,

[3] *PhishTank — Join the fight against phishing.* [Online]. Available: `https://www.phishtank.com/`.

[4] G. Gupta and J. Pieprzyk, "Socio-technological phishing prevention," *Information Security Technical Report*, Social Networking Threats, vol. 16, no. 2, pp. 67–73, May 2011, ISSN: 1363-4127. DOI: `10.1016/j.istr.2011.09.003`. [Online]. Available: `http://www.sciencedirect.com/science/article/pii/S1363412711000549`.

[5] D. D. Caputo, S. L. Pfleeger, J. D. Freeman, and M. E. Johnson, "Going Spear Phishing: Exploring Embedded Training and Awareness," *IEEE Security Privacy*, vol. 12, no. 1, pp. 28–38, Jan. 2014, ISSN: 1540-7993. DOI: `10.1109/MSP.2013.106`.

[6] S. Garera, N. Provos, M. Chew, and A. D. Rubin, "A Framework for Detection and Measurement of Phishing Attacks," ser. WORM '07, New York, NY, USA: ACM, 2007, pp. 1–8, ISBN: 978-1-59593-886-2. DOI: `10.1145/1314389.1314391`. [Online]. Available: `http://doi.acm.org/10.1145/1314389.1314391`.

[7] *Welcome to Python.org*, en. [Online]. Available: `https://www.python.org/`.

[8] *IBM SPSS Software*, en-us, Dec. 2018. [Online]. Available: `https://www.ibm.com/analytics/spss-statistics-software`.

[9] M. Wick, *GeoNames*, Dec. 2018. [Online]. Available: `https://www.geonames.org/`.