



Article

# The Spatial Analysis of the Malicious Uniform Resource Locators (URLs): 2016 Dataset Case Study

Raid W. Amin <sup>1</sup>, Hakki Erhan Sevil <sup>2,\*</sup> , Salih Kocak <sup>3</sup>  and Guillermo Francia III <sup>4</sup> and Philip Hoover <sup>1</sup>

<sup>1</sup> Mathematics & Statistics, University of West Florida, Pensacola, FL 32514, USA; [ramin@uwf.edu](mailto:ramin@uwf.edu) (R.W.A.); [pwh4@students.uwf.edu](mailto:pwh4@students.uwf.edu) (P.H.)

<sup>2</sup> Intelligent Systems & Robotics, University of West Florida, Pensacola, FL 32514, USA

<sup>3</sup> Construction Management, University of West Florida, Pensacola, FL 32514, USA; [skocak@uwf.edu](mailto:skocak@uwf.edu)

<sup>4</sup> Center for Cybersecurity, University of West Florida, Pensacola, FL 32514, USA; [gfranciaiii@uwf.edu](mailto:gfranciaiii@uwf.edu)

\* Correspondence: [hsevil@uwf.edu](mailto:hsevil@uwf.edu)

**Abstract:** In this study, we aimed to identify spatial clusters of countries with high rates of cyber attacks directed at other countries. The cyber attack dataset was obtained from Canadian Institute for Cybersecurity, with over 110,000 Uniform Resource Locators (URLs), which were classified into one of 5 categories: benign, phishing, malware, spam, or defacement. The disease surveillance software SaTScan<sup>TM</sup> was used to perform a spatial analysis of the country of origin for each cyber attack. It allowed the identification of spatial and space-time clusters of locations with unusually high counts or rates of cyber attacks. Number of internet users per country obtained from the 2016 CIA World Factbook was used as the population baseline for computing rates and Poisson analysis in SaTScan<sup>TM</sup>. The clusters were tested for significance with a Monte Carlo study within SaTScan<sup>TM</sup>, where any cluster with  $p < 0.05$  was designated as a significant cyber attack cluster. Results using the rate of the different types of malicious URL cyber attacks are presented in this paper. This novel approach of studying cyber attacks from a spatial perspective provides an invaluable relative risk assessment for each type of cyber attack that originated from a particular country.

**Keywords:** cyber attack; spatial analysis; Uniform Resource Locators (URLs); phishing; malware; spam; defacement



**Citation:** Amin, R.W.; Sevil, H.E.; Kocak, S.; Francia, G., III; Hoover, P. The Spatial Analysis of the Malicious Uniform Resource Locators (URLs): 2016 Dataset Case Study. *Information* **2021**, *12*, 2. <https://dx.doi.org/10.3390/info12010002>

Received: 16 November 2020

Accepted: 16 December 2020

Published: 22 December 2020

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Copyright:** © 2020 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The use of internet has been showing a continuous growth in recent years as we become more dependent on computer networks and infrastructure in the “connected” digital age [1]. This, however, leads to an increase of cyber attacks, potential platform for fraud, and vulnerability for identity theft [2]. As a result, it is crucial for security systems to be up-to-date against scams, malware (malicious software), spam, and phishing attacks [3]. Although one solution can be focusing on detection [4] and classification [5] of the cyber attacks, it will not be sufficient enough considering the fact that these attacks happen globally [6]. There is an evident need for deeper understanding of cyber attacks in terms of spatial analysis [7]. That can provide not only unique perspective regarding origins-victimizations of cyber attacks but also potential risks for regions being the target for cyber attacks [8].

One of the most important subjects in cybersecurity is the detection of cyber attacks to prevent possible threats and disruption. In the literature, there are various techniques and approaches presented for detection and identification of cyber attacks, including K-mean clustering [9], statistical classifier [10], adversary threat modeling [11], Bayesian networks [12], non-attribution based anomaly detection [13], Schmitt analysis [14], bio-inspired immunological metaphors [15], Simulated annealing [16], genetic algorithm [17], fuzzy logic [18], deep neural networks [19], and machine learning [20]. Considering spatial analysis of cyber attacks, there are a few studies in the literature worth mentioning. Merien

et al. presented an entropy-based model to represent cyber attacks as path from a source to target [21]. Their results lead to a pattern in cyber attack origins that could be used to categorize internet attacks [21]. To characterize attack patterns, Chen et al. introduced predictability measures using probability matrix [22]. They showed the correlation between large-scale cyber attacks and their predictabilities [22]. Geographic Internet Protocol (GEO-IP) based analysis was presented by Hu et al. [23]. In their study, GEO-IP was used to detect the location of cyber attack origin, and then, advanced spatial statistical analysis was used to explore cyber attack patterns [23]. Lin et al. presented lexical analysis approach for detecting malicious Uniform Resource Locators (URLs) [24]. According to the results in big data, 90% of malicious URLs could be detected using two-step filtering on the data. Another URL classification based approach was introduced by Feroz and Mengel [25]. They used clusters for URLs to divide them into various categories, which was used for predictions of cyber attacks by the classification model [25]. Du and Yang followed a slightly different approach for detecting and classification of cyber attacks: grouping sources as coordinated teams [26]. Their approach provided an additional perspective to high impact attacks compared to trivial statistical approaches [26]. Furthermore, Koike et al. focused on visualization of cyber threats using 2-D IP address matrix [27]. Their claim was that presented visualization framework could enhance detection algorithms in, for instance, worm propagation models [27].

Although all these studies are presented in the literature, little focus has been given on spatial analysis of cyber attacks, especially malicious URLs, in terms of physical geo-location of the attacks. Most of the studies in the literature presented cyber locations with IP addresses, which makes advanced analysis challenging if additional covariates are to be added to find complex relations between attacks and other variables, instead of geo-locations. Moreover, previous studies looked into just trivial source-target relations in cyber attacks, where risk analysis could be performed with state-of-the-art approach, such as applying epidemiology perspective. Contributions of this study can be summarized as to find answers to following questions:

- Which nations are the top for origins, by count, of cyber attacks?
- Which nations are the top for origins, by relative risk, of cyber attacks?
- Do the spatial hotspots for cyber attack origins differ from the spatial relative risk hotspots for cyber attack?

The main differentiator of this paper from that of the annual reports on malware published by commercial entities is on the in-depth analyses, both visual and text, that are provided. While the annual reports offer cursory description of the data gathered, this paper dives deep into the number and offers meaningful insight on the state of cyber attacks delineated by geographical locations.

The rest of the paper is organized as follows: Section 2 describes the background of the data used, as well as software and pre-data analysis details, while the methodology used is explained in Section 3. Section 4 provides information about analysis design, results, and discussions. In the final section (Section 5), conclusions and future directions are presented.

## 2. Background

### 2.1. Data Collection

In this study, we utilize URL data from 2016. A URL is the fundamental network identification for any resource connected to the web. A URL consists of five parts: Scheme, Subdomain, Second-level Domain, Top-level Domain, and Subdirectory. The data consists of 5 different URL category information, namely (i) benign, which is safe websites with normal services, (ii) phishing, which is a website performs the act of attempting to get information, such as usernames, passwords, and credit card details, by masquerading as a trustworthy entity in an electronic communication, (iii) malware, which is created by attackers to disrupt computer operation, gather sensitive information, or gain access to private computer systems, (iv) spam, which is the act of spreading unsolicited and

unrelated content, and (v) defacement, which is an exploitation of the techniques to alter the content of web pages by suspicious user.

We used cyber attack data obtained from the Canadian Institute for Cybersecurity (CIC) [28], and the details of the dataset can be found in Reference [8]. Mamun et al. developed a lexical analysis to detect and categorize malicious URLs. The data came from disparate internet sources from web crawlers, web spam dataset repository (from the Web Laboratory of the University of Milan), open source OpenPhish website (a widely recognized global website for phishing information), the DNS-BH website of malware collections, a random selection of URLs from a website with a list of defaced URLs. The fact that these data came from multiple sources, properly categorized, and publicly available repositories that are open for inspection and close scrutiny provides ample evidence to support the credibility of the dataset. The results were a categorization of over 110,000 URLs from year 2016. We used the resulting categorized data to perform a spatial analysis of the country of origin for each type of attack. One limitation of this study is that the authors did not attempt to identify or correct any errors in identification or categorization of potentially malicious URL made in the original data. The data from the CIC is publicly available, and the authors downloaded a dataset titled “URL dataset (ISCX-URL-2016)” [28].

The URL dataset (ISCX-URL-2016) contains over 35,000 URLs identified as benign, and those URLs were not included in this paper’s spatial analysis due to fact they are safe websites with normal services. Because our study is focused on cyber attacks, we deliberately eliminated that 35,000 benign URLs so as not to introduce data that could lend to some misinterpretation. Further, in order to normalize the data and put the study in light of internet activities, we inject the population and the estimated number of internet users for that year as bases for the model analysis. Additionally, there were approximately 12,000 URLs categorized as spam, 10,000 URLs categorized as phishing, 11,500 URLs categorized as malware, and 45,500 URLs categorized as defacement [28]. In the covariate analysis, we used population of countries, and population and estimated number of internet users per country were obtained from the 2016 CIA World Factbook, which is publicly available [29]. Three countries were missing values for internet users (Afghanistan, Ascension, and South Sudan) and these values were filled using wiki online sources and compared with others countries to cross check the ratio of population to internet users’ similarity. The number of internet users per country was used as the population baseline for computing rates and for the Poisson model analysis in SaTScan<sup>TM</sup> [30]. Based on the raw data, a ranking by country of origin for each type of cyber attack is presented in Table 1. It should be noted that the countries are listed in rank order based on the raw number of cyber attacks data in the dataset [28]; that is why not all countries are listed in the table. Further, the identified clusters in the following sections may include one or more countries.

## 2.2. Software and Tools Used

In our analysis, we used SaTScan<sup>TM</sup> software, which is a free, open source software that analyzes spatial, temporal, and space-time data using probability models in statistics. It is designed for any of the following interrelated purposes:

- Perform geographical surveillance of disease, to detect spatial or space-time disease clusters, as well as to see if they are statistically significant.
- Test whether a disease is randomly distributed over space, over time, or over space and time.
- Evaluate the statistical significance of disease cluster alarms.
- Perform prospective real-time or time-periodic disease surveillance for the early detection of disease outbreaks.

**Table 1.** Rank order by country of raw number of cyber attacks.

Rank	Defacement	Malware	Phishing	Spam *	Total Number of Cyber Attacks
1	United States (18047)	United States (5945)	United States (3168)	United Kingdom (5898)	United States (28742)
2	Germany (15532)	China (2053)	China (439)	United States (1582)	Germany (15914)
3	Netherlands (4683)	Hong Kong (406)	Germany (197)	Ireland (1249)	United Kingdom (9299)
4	Italy (3466)	South Korea (271)	France (178)	Germany (110)	Netherlands (4771)
5	United Kingdom (3280)	Canada (201)	Hong Kong (117)	France (66)	China (3746)
6	Russia (2054)	Brunei (143)	United Kingdom (113)	Netherlands (12)	Italy (3575)
7	Brazil (2003)	Russia (135)	Australia (109)	Finland (8)	Russia (2264)
8	France (1907)	Germany (75)	South Africa (103)	Italy (4)	France (2157)
9	Australia (1748)	Poland (60)	Italy (83)		Ireland (2140)
10	Spain (1570)	Italy (22)	Ireland (81)		Brazil (2042)
11	China (1254)	Cayman Islands (20)	Russia (75)		Australia (1860)
12	Denmark (1026)	South Africa (18)	Poland (74)		Spain (1621)
13	Poland (955)	Philippines (11)	Canada (67)		Hong Kong (1214)
14	Czech Republic (862)	Taiwan (11)	Netherlands (67)		Poland (1089)
15	Switzerland (821)	Netherlands (9)	Singapore (57)		Denmark (1041)
16	Ireland (809)	Spain (9)	Spain (42)		Canada (953)
17	Hong Kong (691)	United Kingdom (8)	Brazil (39)		Czech Republic (878)
18	Canada (685)	Thailand (8)	South Korea (36)		Switzerland (841)
19	Sweden (587)	Singapore (7)	Turkey (35)		Sweden (603)
20	Ukraine (565)	France (6)	Thailand (34)		Portugal (582)

\* Spam-type attacks were only recorded from 8 countries.

Although SaTScan<sup>TM</sup> can be used for temporal and space-time data, as well, we did only the purely spatial analysis due to the data being for one year. We chose SaTScan<sup>TM</sup> for the surveillance due to the use of statistical modeling with probability models in which hypothesis testing is used with  $p$ -Values to identify significant clusters with small chances of existing due to random causes. While elliptically shaped clusters could have been used here, the results are very similar when using circular windows to identify clusters. It is understood that clusters may include countries with low cyber attack counts, but the overall counts within a cluster will be unusually high. It is possible to re-analyze each identified cluster again with SaTScan<sup>TM</sup> to identify smaller sized clusters. Below is the Poisson model's likelihood function used in our analyses, which is proportional to

$$\left(\frac{n}{E}\right)^n \left(\frac{N-n}{N-E}\right)^{N-n} I(n > E), \quad (1)$$

where  $n$  is the number of cyber attack counts within the scan window,  $N$  is the total number of internet users in the population, and  $E$  is the expected cyber attack counts under the null hypothesis. There are several other cluster analysis algorithms in the literature, such as DBSCAN, and there many available libraries on the internet, in different programming languages, for instance Python, R, or MATLAB. It is understood that no cluster analysis software is better than all other cluster analysis software packages. Each major cluster analysis software package has some advantages for specific applications. We selected SaTScan<sup>TM</sup> as it allows, unlike DBSCAN, statistical significance to be detected for each cluster.

While SaTScan<sup>TM</sup> has been widely used with epidemiology data, it is perfectly fine to use this surveillance software with data from applications that have nothing to do with epidemiology. The statistical modeling used in this software is based on probability distributions, such as Normal distribution or Poisson distribution, or a nonparametric modeling can be chosen. In our cyber attacks study, we used counts of cyber attacks as the random variable of interest. The likelihood function of the Poisson distribution includes the number of internet users. In this context, the “disease” was considered as the different types of malicious URL cyber attacks. We used SaTScan<sup>TM</sup> to perform a purely spatial, Poisson statistical analysis to identify significant clusters of cyber attack by type and country of origin. Settings in SaTScan<sup>TM</sup> were used to search for high-rate clusters only, with a restriction of at least 3 cases per potential cluster. Further, high rate clusters were restricted to have a relative risk greater than or equal to 1.2 and were reported only the most likely clusters using a hierarchical clustering. These are countries with cyber attacks that are at least 20% higher than in the rest of the world. It should be noted that this paper does not claim to use the idea of the spatial spread of pathogen model to model the behavior of cyber attacks. Although the SaTScan<sup>TM</sup> software is used primarily for performing space-time disease cluster analyses and testing whether a disease is randomly distributed over time, we used its capability to perform purely spatial Poisson statistical to identify significant clusters of cyber attacks.

Additionally, we used ArcMap which is a licensed software tool from ArcGIS that is used to represent geographic information as a collection of layers and other elements in a map. ArcMap was used to display the clusters found from SaTScan<sup>TM</sup> and for two types of heat maps. The first type of heat map displays the different malicious URL cyber attacks by country based on rate (attacks per internet users). The type of heat map displayed relative risk and clusters obtained from SaTScan<sup>TM</sup>. A 2015 Tiger Shapefile from ArcGIS was used as a base map that was then manipulated with joined data to present heat maps by country and clusters when appropriate.

### 2.3. Data Manipulation

Before spatial analysis, we, first, processed the data to extract geo-locations. We used URL information to map them to a physical geographic location. Further, we performed normalization on the data. The values of the covariates were adjusted for population (where appropriate) and normalized by the Blom method [31]. This was done so that

variation would be preserved within each variable, but the variation between variables would be minimized. Normalizing prevents variables that are measured with larger units from having a disproportionately larger impact on the model. Normalizing also lessens the effect of outliers. Finally, for each country and type of malicious URL cyber attacks, a rate was calculated by dividing the count of each type of attack by the country's internet users. The number of internet users was also used as the population file in SaTScan<sup>TM</sup>.

### 3. Methodology

The disease surveillance software, SaTScan<sup>TM</sup>, was used to identify significant clusters of high cyber attacks around the world. Specifically, SaTScan<sup>TM</sup> was used to perform a spatial scan for clusters of high counts of the different types of malicious URL cyber attacks. It allows user to choose among various models for analysis; the Poisson model was selected for cyber attacks study since it is more appropriate for (rare) count data. The malicious URL cyber attack counts by country were used as the SaTScan<sup>TM</sup> variables of interest in the analysis. Country ID codes were used to relate the counts to relevant country, population, latitude, and longitude. The analysis was performed by “purely spatial” method (vs. temporal) using hierarchical clusters, with no geographical overlap (meaning that clusters will be reported only if they do not overlap with a previously reported cluster), and with the clusters required to have a relative risk of at least 1.2 and contain at least 3 counties. Please refer to the SaTScan<sup>TM</sup> documentation for more details about these settings [30].

SaTScan<sup>TM</sup> performs a cluster analysis as follows: A dynamic geographic unit, or “moving window,” is systematically scanned across the contiguous states and compared to expected and observed variable counts (Recall that we are using the discrete Poisson model to analyze counts.). For this study, each country serves as the initial geographic unit used by SaTScan<sup>TM</sup> as a potential cluster center. For a cluster, a circle of varying size (from 0 up to 3500 km) is analyzed around each county with higher than expected counts. The maximum size of a cluster was a circle containing 25% of the population at risk not to exceed 3500 km. The 3500 km upper limit was used to keep clusters from extending to the Polar regions which caused unusual patterns when transferred to a flat projection map. The null hypothesis is that the number of cases in each area is proportional to its population size; the alternative hypothesis is that there is an elevated risk within the window.

Clusters are reported for those circles where the number of observed are “much” greater than the expected values. To identify clusters, a likelihood function is maximized across all locations. The cluster with the maximum (largest) likelihood function indicates the cluster which is the least likely to have occurred by chance. This cluster is identified as Cluster 1. Once Cluster 1 has been calculated, secondary clusters are calculated and ranked by their likelihood ratio test statistic in decreasing order after Cluster 1. Clusters can be statistically significant or not, based on a *p*-Value obtained from SaTScan<sup>TM</sup> via Monte Carlo hypothesis testing. For this study, we reported only clusters that were significant at the  $\alpha = 0.01$  significance level. SaTScan<sup>TM</sup> provides detailed information about each cluster, some of which are listed below.

- Location IDs: the geographic center and a list of countries that belong to the cluster.
- Population: the number of internet users in each cluster.
- Observed/expected: the observed number of cases within the cluster divided by the expected number of cases within the cluster (under the null hypothesis that risk is the same inside and outside the cluster). Put another way, this is the estimated risk within the cluster divided by the estimated risk for the study region as a whole.
- Relative risk: the estimated risk within the cluster divided by the estimated risk outside the cluster. It is calculated as the observed divided by the expected within the cluster divided by the observed divided by the expected outside the cluster.
- *p*-Value: the probability of obtaining the observed (or a greater) number of cases in a cluster if the risk were the same as it is outside the cluster.

Heat maps were used to visualize geographic data patterns, especially in conjunction with cluster maps. Since our research objective was to study the origination of malicious



URL cyber attacks, we created heat maps by country for both the rate of attack and relative risk. The clusters obtained from SaTScan<sup>TM</sup> were overlaid on these heat maps with ArcMap.

#### 4. Results and Discussion

Using the data and methodology described above, obtained results are presented in this section. Results using the rate of the different types of malicious URL cyber attacks are presented first, followed by results presenting relative risk. Both sets of results are similar.

##### 4.1. Type of Cyber Attacks by Rate

The following data are presented as normalized rate data for each type of cyber attack. The corresponding differences are rates and map colors:

- “Red” represents normalized rates at or above 1.18 standard deviations above the mean.
- “Orange” represents normalized rates 0.38 to 1.17 standard deviations above the mean.
- “Yellow” represent normalized rates centered on the mean (plus or minus 0.37).
- “Light Green” represents normalized rates 0.38 to 1.17 standard deviations below the mean.
- “Dark Green” represents normalized rates more than 1.18 standard deviations below the mean.

##### 4.1.1. Defacement

First, we look into the defacement cyber attack type in our analysis. According to the results depicted in Figures 1–3 and given in Table 2, the highest rate of defacement attacks originate in Europe. Each rate is defined as (cyber attack counts)/(internet users count). All of the world’s highest normalized rates (“red” countries) are present in Europe, as well as most of the countries with normalized rates (“orange” countries) above the mean. Outside of Europe, Turkmenistan, Australia, and the United States are the only countries with above average rates of defacement type malicious URL attacks. Turkmenistan is the only country outside of Europe to be placed in the top 10 countries when ordered by normalized rates. It is also worth noting the importance of using rate data. If simple count data had been used, 8 of 10 countries in the table would have been replaced.

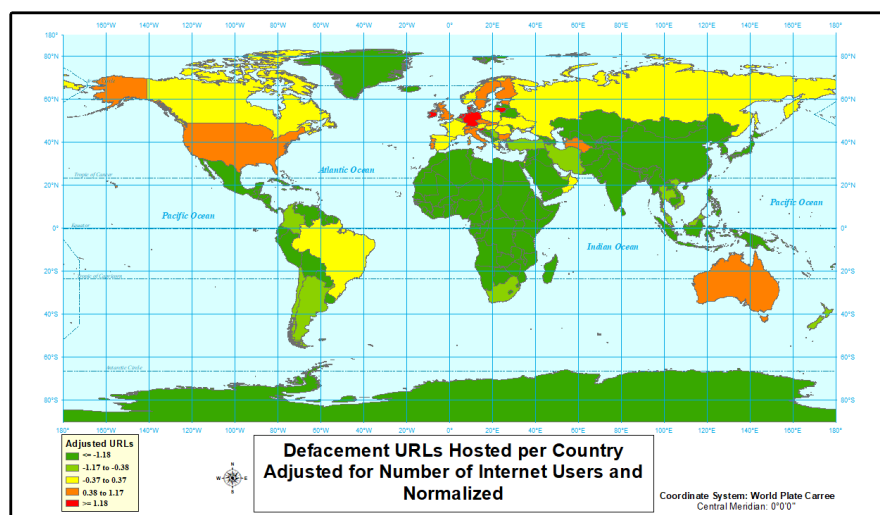


Figure 1. Defacement Uniform Resource Locators (URLs) hosted per country.

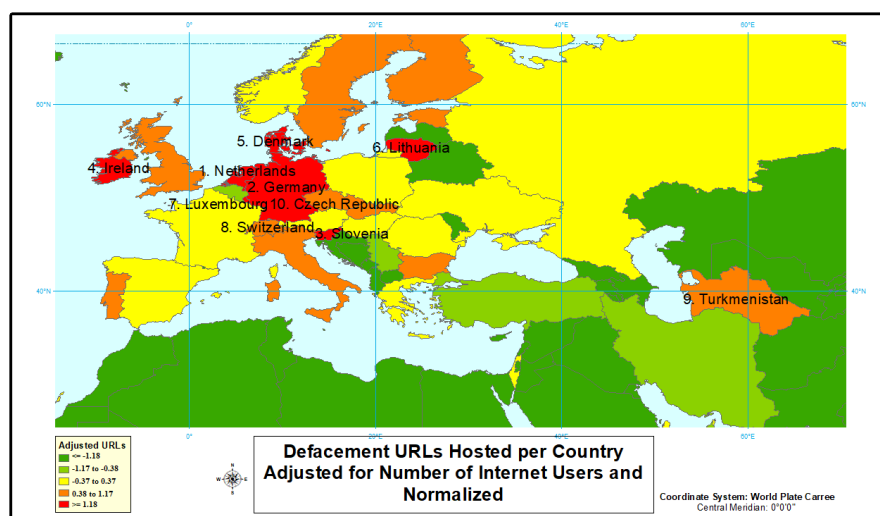


Figure 2. Defacement URLs hosted per country—Europe cluster.

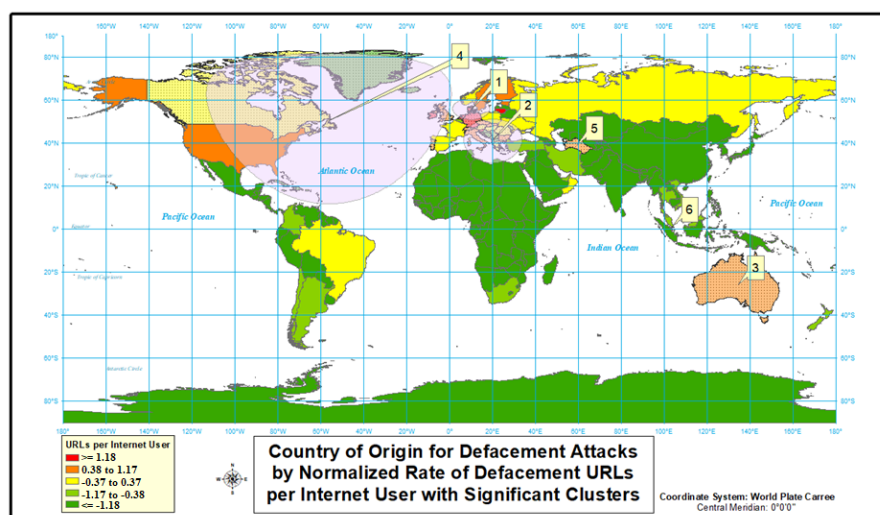


Figure 3. Country of origin for defacement attacks.

Table 2. Rank order by country of defacement cyber attacks.

Country	Internet Users	Population	URLs	Adjusted URLs	Normalized Adjusted URLs	URLs Rank	Adjusted URLs Rank
Netherlands	15,385,203	16,981,285	4683	30.43833741	2.279717741	3	1
Germany	72,365,643	82,193,77	15,532	21.46322392	1.88950996	2	2
Slovenia	1,493,382	2074.205	315	21.09306259	1.669478315	29	3
Ireland	4,069,432	4695.79	809	19.87992427	1.509301626	16	4
Denmark	5,424,169	5711.346	1026	18.91533984	1.380538791	12	5
Lithuania	2,122,884	2889.555	397	18.70097471	1.271305716	28	6
Luxembourg	567,698	579.266	95	16.73424955	1.17543931	43	7
Switzerland	7,312,744	8379.915	821	11.22697581	1.08930797	15	8
Turkmenistan	951,925	5662.371	102	10.71512987	1.010580622	40	9
Czech Republic	8,141,303	10,618.868	862	10.58798573	0.937665596	14	10



#### 4.1.2. Malware

Second, our analysis focus on malware type cyber attacks. The data show a much greater diversity for the normalized rate of malware type attacks than occurred with defacement (Figures 4 and 5 and Table 3). North America accounts for all of the highest normalized rates except for Brunei. The Cayman Islands, British Virgin Islands, and Brunei are hard to depict in the figure due their relative small size and the scale of the map. Even though normalized rates are used for rankings, seven countries would still be placed in the top 10 based on count data alone.

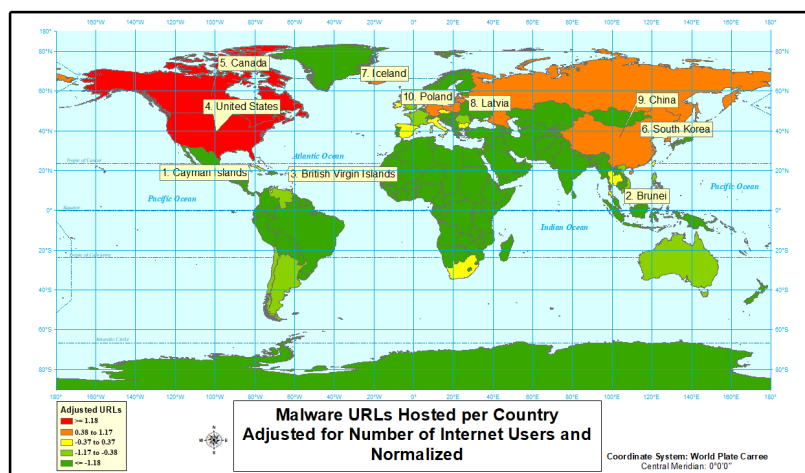


Figure 4. Malware URLs hosted per country.

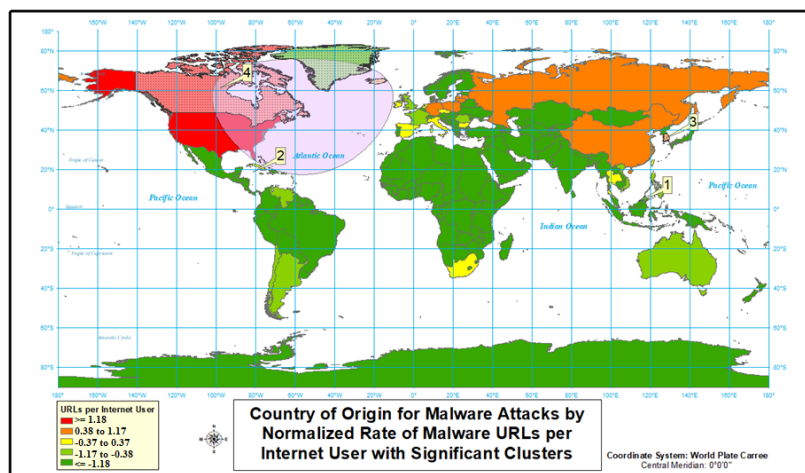


Figure 5. Country of origin for malware attacks.

Table 3. Rank order by country of malware cyber attacks.

Country	Internet Users	Population	URLs	Adjusted URLs	Normalized Adjusted URLs	URLs Rank	Adjusted URLs Rank
Cayman Islands	45,242	62,564	20	44.20671058	2.146324055	10	1
Brunei	410,800	419,791	143	34.81012658	1.734651522	5	2
British Virgin Islands	14,600	29,355	3	20.54794521	1.499446023	29	3
United States	246,809,221	323,015,992	5945	2.408743067	1.326381463	1	4
Canada	31,770,034	36,382,942	201	0.632671655	1.185882285	4	5
South Korea	44,153,000	50,983,446	271	0.613774828	1.06555643	3	6
Iceland	329,967	332,209	2	0.606121218	0.958958568	34	7
Latvia	1570374	1974,265	6	0.382074589	0.862277135	22	8
China	736,789,960	1,421,292,894	2053	0.278641148	0.773053962	2	9
Poland	28,237,820	37,989,218	60	0.212480992	0.689599853	8	10

#### 4.1.3. Phishing

Third, we look into phishing cyber attacks. Even more than malware, the country of origin for phishing-type attacks is diverse, and there is no geographic relationship when using normalized rate (Figures 6 and 7 and Table 4). Four of the ten highest rate countries would also be present in a list of top 10 countries for the number of phishing-type attacks. Nine of ten countries listed for phishing-type attacks are in the highest rate category (“red”); this is highest percentage of any type of attack.

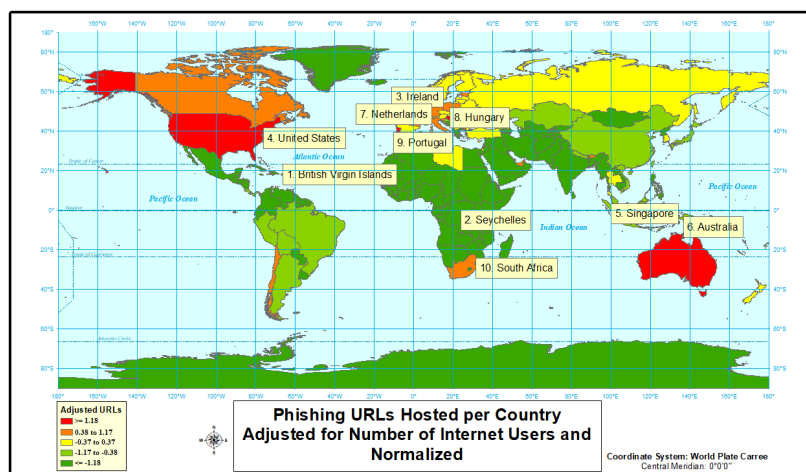


Figure 6. Phishing URLs hosted per country.

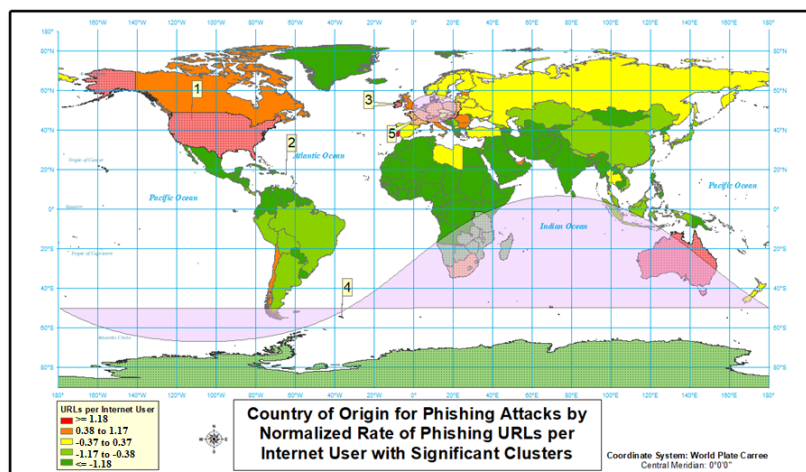


Figure 7. Country of origin for phishing attacks.

Table 4. Rank order by country of phishing cyber attacks.

Country	Internet Users	Population	URLs	Adjusted URLs	Normalized Adjusted URLs	URLs Rank	Adjusted URLs Rank
British Virgin Islands	14,600	29,355	37	253.4246575	2.385307593	17	1
Seychelles	52,664	95,711	3	5.696490962	2.010591708	52	2
Ireland	4,069,432	4,695,79	81	1.990449773	1.8011937	9	3
United States	246,809,221	323,015,992	3168	1.283582513	1.64983783	1	4
Singapore	4,683,200	5,653,625	57	1.217116502	1.528937803	14	5
Australia	20,288,409	24,262,71	109	0.537252576	1.426987084	6	6
Netherlands	15,385,203	16,981,285	67	0.435483367	1.338027808	13	7
Hungary	7,826,695	9,752,97	31	0.396080338	1.258554954	21	8
Portugal	7,629,560	10,325,54	27	0.353886725	1.186322712	23	9
South Africa	29,322,380	56,207,649	103	0.35126753	1.119800855	7	10

#### 4.1.4. Spam

Finally, we analyze spam-type cyber attack in the data. The results lead to the fact that all of the highest rate countries are located in Europe or the United States (Figures 8 and 9 and Table 5). It should be noted that cluster analysis was not performed for spam attacks (Figure 9), as we only had data for 8 countries. For the first time in our analysis, there is only one country (Ireland) that has the highest level of normalized rate, while two other countries show slightly elevated rates. Moreover, again the first time, the list of top rates contains countries with a below average normalized rate. Spam-type attacks when analyzed by rate appear to be confined to only 3 countries with higher than average rates.

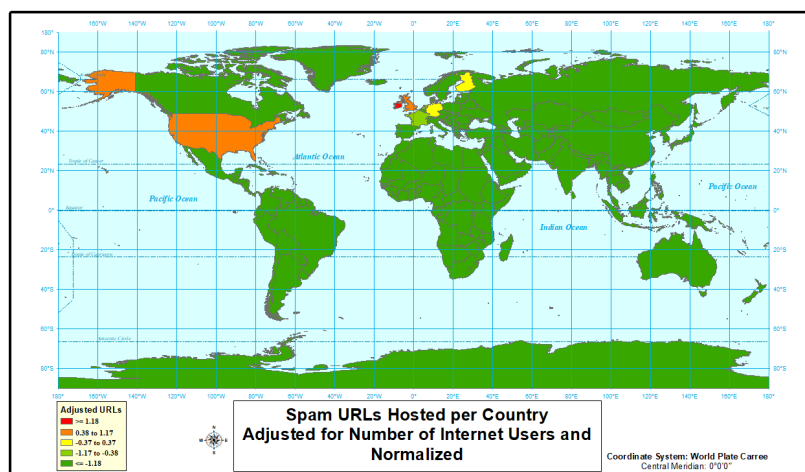


Figure 8. Spam URLs hosted per country.

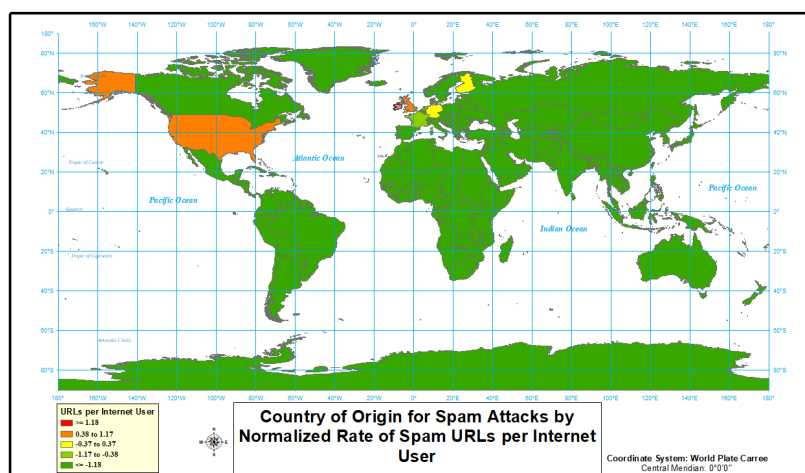


Figure 9. Country of origin for spam attacks.

Table 5. Rank order by country of spam cyber attacks.

Country	Internet Users	Population	URLs	Adjusted URLs	Normalized Adjusted URLs	URLs Rank	Adjusted URLs Rank
Ireland	4,069,432	4695.79	1249	30.69224403	1.43420016	3	1
United Kingdom	61,064,454	66,297.944	5898	9.658646911	0.852495034	1	2
United States	246,809,221	323,015.992	1582	0.640980914	0.472789121	2	3
Finland	4,822,132	5497.714	8	0.165901722	0.152505974	7	4
Germany	72,365,643	82,193.77	110	0.152005835	−0.15250597	4	5
France	57,226,585	64,667.59	66	0.115331013	−0.47278912	5	6
Netherlands	15,385,203	16,981.285	12	0.077997021	−0.85249503	6	7
Italy	38,025,661	60,663.068	4	0.010519212	−1.43420016	8	8

#### 4.2. Cyber Attacks by Relative Risk with Clusters

To further the analysis, we look into the relative risk of cyber attacks with cluster. The resulting maps and tables for relative risk analysis are given in this section. The relative risk values, along with the clusters, were obtained by SaTScan<sup>TM</sup> for each type of cyber attack. Relative risk can be interpreted as “x times more likely than expected” to have a cyber attack originate from that country. Since SaTScan<sup>TM</sup> calculates different expected counts for each type of attack, the relative risk and break points will be different for each type of cyber attack. The color scheme is explained in the legend on each map.

##### 4.2.1. Defacement

According to the results given in Tables 6 and 7, the highest relative risk of defacement attacks originate in Europe. The highest risk cluster includes almost all of Europe (32 countries) has a relative risk 11.2 times greater than expected based on internet users and attack counts. The next five clusters all have relative risks between 2.4 and 5.4. While Russia and Oman were declared as clusters, the *p*-Value for those clusters indicates they should not be included. The clusters and relative risk align well with those found using normalized rate.

**Table 6.** Rank order for country origin for defacement attack in terms of relative risk.

Cluster	Location(s)	Observed Cases/Expected Cases	Cluster Relative Risk	Cluster's <i>p</i> -Value	Cluster's Internet Users
1	Andorra; Austria; Belgium; Channel Islands; Croatia; Czech Republic; Denmark; Estonia; Faroe Islands; Finland; France; Germany; Hungary; Iceland; Ireland; Italy; Latvia; Liechtenstein; Lithuania; Luxembourg; Isle of Man; Monaco; Netherlands; Norway; Poland; Portugal; San Marino; Slovakia; Slovenia; Spain; Sweden; Switzerland; United Kingdom	5.58	11.21	$1 \times 10^{-17}$	398,355,738
2	United States	4.27	5.40	$1 \times 10^{-17}$	243,004,928
3	Australia	4.79	4.89	$1 \times 10^{-17}$	20,996,948
4	Hong Kong	6.14	6.19	$1 \times 10^{-17}$	6,477,174
5	Turkmenistan	4.88	4.88	$1 \times 10^{-17}$	1,203,254
6	Singapore	2.42	2.43	$1 \times 10^{-17}$	4,774,486
7	Russia	1.07	1.07	0.426	110,423,808
8	Oman	1.28	1.28	0.999	3,591,884

**Table 7.** Country rank for defacement relative risk with number internet users.

Country	Internet Users	Defacement Relative Risk
Netherlands	15,826,558	18.17
Germany	69,371,542	16.26
Ireland	3,968,882	11.85
Slovenia	1,636,340	11.12
Denmark	5,545,717	10.79
Lithuania	2,242,873	10.24
Luxembourg	566,696	9.66
Hong Kong	6,477,174	6.19
Switzerland	7,852,818	6.07
Czech Republic	8,359,173	5.99

##### 4.2.2. Malware

Subsequently, we analyze malware with relative risk. Malware results show two countries with a relative risk over 150 times greater than expected; the Cayman Islands and Brunei (Figure 10 and Tables 8 and 9). These two countries were also top 2 in normalized rates of malware type cyber attack. Two other countries, U.S. and Hong Kong, were the ones with higher than expected relative risk, with each at approximately 27 times more likely to originate a malware attack. Four countries were approximately twice as likely

to have malware attacks originate (“yellow” countries). After this, the relative risk falls rapidly to below 1.0, which indicates a lower likelihood than expected.

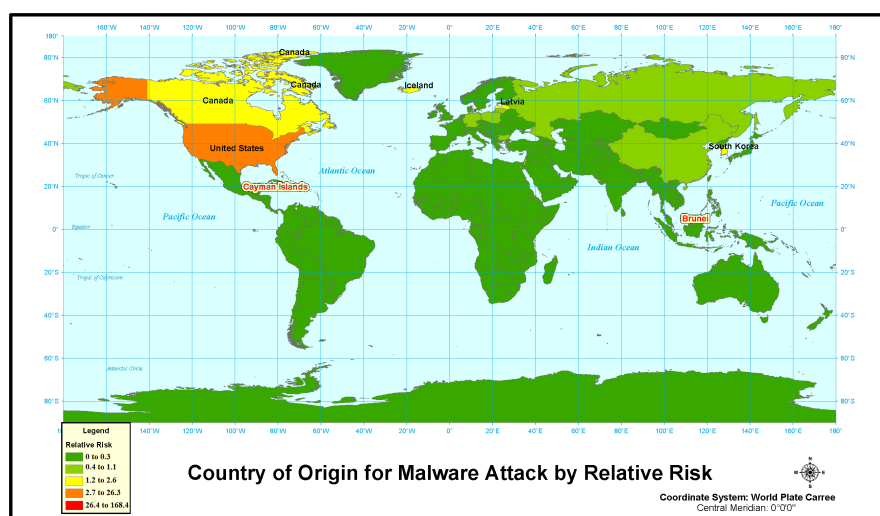


Figure 10. Country of origin for malware attack by relative risk.

Table 8. Rank order for country origin for malware attack in terms of relative risk.

Cluster	Location(s)	Observed Cases/ Expected Cases	Cluster Relative Risk	Cluster's $p$ -Value	Cluster's Internet Users
1	United States	10.42	26.277	$1 \times 10^{-17}$	243,004,928
2	Hong Kong	26.71	27.86	$1 \times 10^{-17}$	6,477,174
3	Brunei	152.99	155.32	$1 \times 10^{-17}$	398,256
4	South Korea	2.38	2.42	$1 \times 10^{-17}$	48,485,256

Table 9. Country rank for malware relative risk with number internet users.

Country	Internet Users	Malware Counts	Malware Relative Risk
Cayman Islands	50,721	20	168.36
Brunei	398,256	143	155.32
Hong Kong	6,477,174	406	27.86
United States	243,004,931	5945	26.28
Iceland	326,429	2	2.61
Canada	33,726,987	201	2.57
South Korea	48,485,257	271	2.42
Latvia	1,605,472	6	1.59
China	831,461,020	2053	1.07
Poland	28,868,007	60	0.88

#### 4.2.3. Phishing

We then analyze phishing attacks with relative risk, and results show some differences between relative risks and rates. France and Hong Kong replaced South Africa and The British Virgin Islands on the list of top 10 countries as origin of a phishing attack (Figure 11 and Tables 10 and 11). There is one large cluster that covers 23 countries in Europe, while the other six clusters are all single country clusters. The United States is the most likely cluster and has the highest relative risk (21 times more likely than expected) after the Seychelles (39 times more likely than expected).

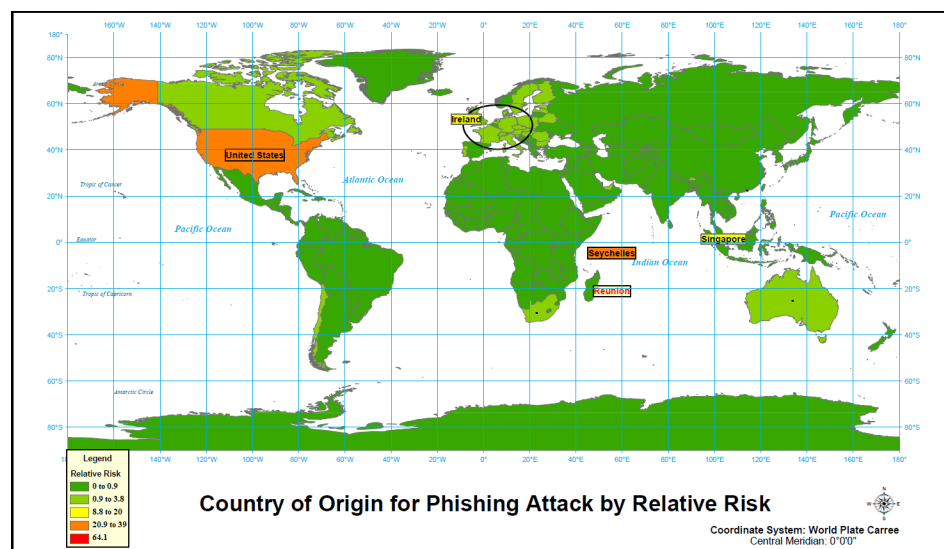


Figure 11. Country of origin for phishing attack by relative risk.

Table 10. Rank order for country origin for phishing attack in terms of relative risk.

Cluster	Location(s)	Observed Cases/ Expected Cases	Cluster Relative Risk	Cluster's <i>p</i> -Value	Cluster's Internet Users
1	United States	9.52	20.94	$1 \times 10^{-17}$	243,004,928
2	Andorra; Austria; Belgium; Channel Islands; Croatia; Czech Republic; Denmark; France; Germany; Hungary; Ireland; Italy; Liechtenstein; Luxembourg; Isle of Man; Monaco; Netherlands; Poland; San Marino; Slovakia; Slovenia; Switzerland; United Kingdom	2.01	2.20	$1 \times 10^{-17}$	326,542,990
3	Hong Kong	13.19	13.45	$1 \times 10^{-17}$	6,477,174
4	Singapore	8.72	8.80	$1 \times 10^{-17}$	4,774,486
5	Australia	3.79	3.85	$1 \times 10^{-17}$	20,996,948
6	South Africa	2.38	2.41	$1.03 \times 10^{-11}$	31,571,836
7	Seychelles	38.95	38.97	0.015	56,249

Table 11. Country rank for phishing relative risk with number internet users.

Country	Internet Users	Phishing Count	Phishing Relative Risk
Seychelles	56,249	3	38.97
United States	243,004,931	3168	20.94
Ireland	3,968,882	81	15.11
Hong Kong	6,477,174	117	13.45
Singapore	4,774,486	57	8.80
Australia	20,996,949	109	3.85
Netherlands	15,826,558	67	3.12
Hungary	7,485,404	31	3.04
Portugal	7,619,216	27	2.60
France	52,057,410	178	2.55

#### 4.2.4. Spam

The last analysis with relative risk is for spam attack-type. The data for spam-type cyber attacks only contains observations for 8 countries. Just as when analyzed by normalized rate, three countries, Ireland, UK and U.S., account for the only countries where spam attacks originate at a higher than expected rate (Figure 12 and Tables 12 and 13). There is one cluster includes Ireland, UK, and Isle of Man. This cluster has a relative risk nearly 240 times greater than expected. The individual countries of Ireland (165 times) and UK (123 times) have relative risks an order of magnitude greater than any other country. The U.S. is the only other country with a relative risk greater than expected at 3.4 times. From the



maps and data, it is apparent that spam attacks are not as common as other forms of cyber attack, but the origination is limited to small number of countries.

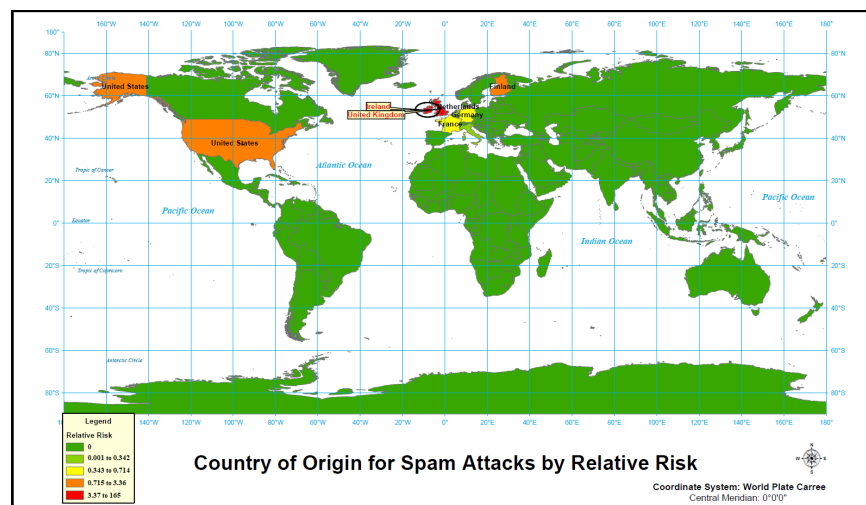


Figure 12. Country of origin for spam attack by relative risk.

Table 12. Rank order for country origin for spam attack in terms of relative risk.

Cluster	Location(s)	Observed Cases/ Expected Cases	Cluster Relative Risk	Cluster's $p$ -Value	Cluster's Internet Users
1	Ireland; Isle of Man; United Kingdom	48.5	238.85	$1 \times 10^{-17}$	66,699,998
2	United States	2.9	3.36	$1 \times 10^{-17}$	243,004,928

Table 13. Country rank for spam relative risk with number internet users.

Country	Internet Users	Spam Count	Spam Relative Risk
Ireland	3,968,882	1249	165.34
United Kingdom	62,731,115	5898	123.34
United States	243,004,931	1582	3.36
Finland	4,808,850	8	0.75
Germany	69,371,542	110	0.71
France	52,057,410	66	0.57
Netherlands	15,826,558	12	0.34
Italy	37,186,461	4	0.05

#### 4.2.5. Total

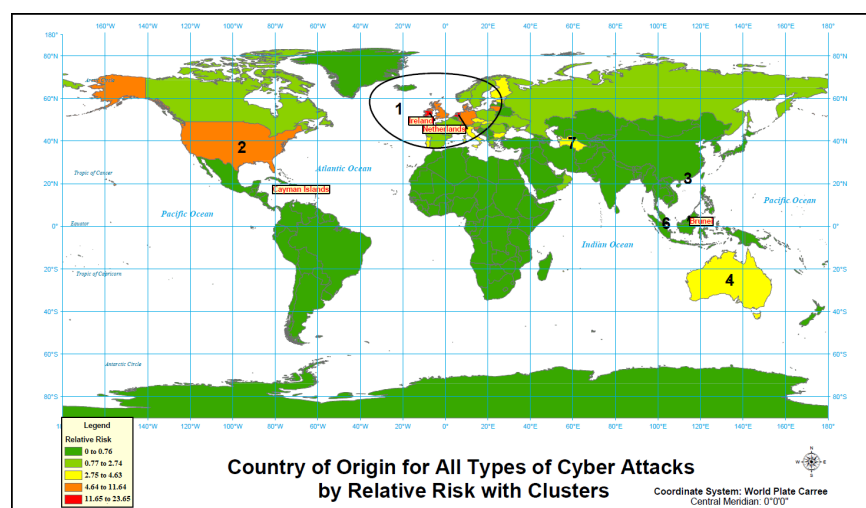
Finally, we run analysis for countries including all cyber attack types combined. Figure 13 and Table 14 and 15 are a summary and categorization of all types of cyber attacks combined and analyzed as a group. The count data is simply the total of all 4 individual types of cyber attacks. The SaTScan<sup>TM</sup> analysis and heatmap followed the same procedures.

**Table 14.** Rank order for country origin for all attack types in terms of relative risk.

Cluster	Location(s)	Observed Cases/ Expected Cases	Cluster Relative Risk	Cluster's <i>p</i> -Value	Cluster's Internet Users
1	Andorra; Austria; Belgium; Channel Islands; Croatia; Czech Republic; Denmark; Estonia; Faroe Islands; Finland; France; Germany; Hungary; Iceland; Ireland; Italy; Latvia; Liechtenstein; Lithuania; Luxembourg; Isle of Man; Monaco; Netherlands; Norway; Poland; Portugal; San Marino; Slovakia; Slovenia; Spain; Sweden; Switzerland; United Kingdom	5.09	9.2	$1 \times 10^{-17}$	398,355,738
2	United States	5.07	6.86	$1 \times 10^{-17}$	243,004,928
3	Hong Kong	8.04	8.13	$1 \times 10^{-17}$	6,477,174
4	Australia	3.8	3.86	$1 \times 10^{-17}$	20,996,948
5	Brunei	15.4	15.43	$1 \times 10^{-17}$	398,256
6	Singapore	2.38	2.39	$1 \times 10^{-17}$	4,774,486
7	Turkmenistan	3.64	3.64	$1 \times 10^{-17}$	1,203,254

**Table 15.** Country rank for all attack types relative risk with number internet users.

Country	Internet Users	Total Attack Count	Total Attack Relative Risk
Ireland	3,968,882	2140	23.65
Cayman Islands	50,721	20	16.92
Brunei	398,256	143	15.43
Netherlands	15,826,558	4771	13.57
Germany	69,371,542	15,914	11.64
Slovenia	1,636,340	320	8.41
Hong Kong	6,477,174	1214	8.13
Denmark	5,545,717	1041	8.13
Lithuania	2,242,873	402	7.72
Luxembourg	566,696	96	7.27

**Figure 13.** Country of origin for all attack types by relative risk.

## 5. Conclusions and Future Directions

In this study, our goal is to analyze cyber attack data obtained from Canadian Institute for Cybersecurity (CIC) and to identify significant clusters of different cyber attack types based on country of origin by looking into Uniform Resource Locators (URLs). The data from CIC contains over 110,000 URLs with 4 cyber attack types: phishing, malware, spam, or defacement. We perform spatial analysis using SaTScan<sup>TM</sup>, along with number of internet users per country. We present cluster analysis results in two categories, cyber attack type by rate per country, and cyber attacks clusters by relative risk. Our results not

only provide geo-physical representation of the cyber attacks but also novel perspective for these attacks as “hotspots” for cyber attack origins. To close, we provide this summary of the important contributions of this work:

- provide the feasibility of visual analytics as a cybersecurity tool;
- enable the realization that cybersecurity data analysis could be approached using multiple perspectives;
- provide the base framework for more advanced and enhanced spatial cluster analytics tools; and
- provide the recognition of the need for reliable data that can be used for analytics.

This study presents our initial work on the application of spatial analysis to cybersecurity. We recognize the abundant research frontier ahead of us and plan to pursue the following directions:

1. introduce the technical capacity of a country as an independent variable;
2. investigate the propensity of a country to defend and/or offensively react to a cyber attack;
3. perform a longitudinal study on the same data with the objective unraveling trends in risk, cyber defense, and cyber attacks; and
4. develop a formal model of the cyber attacks similar to the established epidemiology models.

**Author Contributions:** Supervision, project administration, R.W.A.; methodology, validation, formal analysis, data curation, P.H.; writing—original draft preparation, H.E.S.; writing—review and editing, S.K., G.F.III. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Darling, M.; Heileman, G.; Gressel, G.; Ashok, A.; Poornachandran, P. A lexical approach for classifying malicious URLs. In Proceedings of the IEEE 2015 International Conference on High Performance Computing & Simulation (HPCS), Amsterdam, The Netherlands, 20–24 July 2015; pp. 195–202.
2. Lallie, H.S.; Shepherd, L.A.; Nurse, J.R.; Erola, A.; Epiphaniou, G.; Maple, C.; Bellekens, X. Cyber security in the age of COVID-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic. *arXiv* **2020**, arXiv:2006.11929.
3. Abdalrahman, G.A.; Varol, H. Defending Against Cyber-Attacks on the Internet of Things. In Proceedings of the IEEE 2019 7th International Symposium on Digital Forensics and Security (ISDFS), Barcelos, Portugal, 10–12 June 2019; pp. 1–6.
4. Pivarníková, M.; Sokol, P.; Bajtoš, T. Early-Stage Detection of Cyber Attacks. *Information* **2020**, *11*, 560. [CrossRef]
5. Hu, C.; Yan, J.; Wang, C. Advanced cyber-physical attack classification with extreme gradient boosting for smart transmission grids. In Proceedings of the 2019 IEEE Power & Energy Society General Meeting (PESGM), Atlanta, GA, USA, 4–8 August 2019; pp. 1–5.
6. Doynikova, E.; Novikova, E.; Kotenko, I. Attacker Behaviour Forecasting Using Methods of Intelligent Data Analysis: A Comparative Review and Prospects. *Information* **2020**, *11*, 168. [CrossRef]
7. Yao, Y.; Su, L.; Lu, Z.; Liu, B. STDeepGraph: Spatial-Temporal Deep Learning on Communication Graphs for Long-Term Network Attack Detection. In Proceedings of the 2019 18th IEEE International Conference On Trust, Security and Privacy in Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE), Rotorua, New Zealand, 5–8 August 2019; pp. 120–127.
8. Mamun, M.S.I.; Rathore, M.A.; Lashkari, A.H.; Stakhanova, N.; Ghorbani, A.A. Detecting malicious urls using lexical analysis. In *International Conference on Network and System Security*; Springer: Berlin, Germany, 2016; pp. 467–482.
9. Bloedorn, E.; Christiansen, A.D.; Hill, W.; Skorupka, C.; Talbot, L.M.; Tivel, J. *Data Mining for Network Intrusion Detection: How to Get Started*; Technical Report; Citeseer, 2001. Available online: [https://www.mitre.org/sites/default/files/pdf/bloedorn\\_datamining.pdf](https://www.mitre.org/sites/default/files/pdf/bloedorn_datamining.pdf) (accessed on 21 December 2020).
10. Kim, D.W.; Yan, P.; Zhang, J. Detecting fake anti-virus software distribution webpages. *Comput. Secur.* **2015**, *49*, 95–106. [CrossRef]
11. Burmester, M.; Magkos, E.; Chrissikopoulos, V. Modeling security in cyber-physical systems. *Int. J. Crit. Infrastruct. Prot.* **2012**, *5*, 118–126. [CrossRef]
12. Xie, P.; Li, J.H.; Ou, X.; Liu, P.; Levy, R. Using Bayesian networks for cyber security analysis. In Proceedings of the 2010 IEEE/IFIP International Conference on Dependable Systems & Networks (DSN), Chicago, IL, USA, 28 June–1 July 2010; pp. 211–220.

13. Bou-Harb, E.; Debbabi, M.; Assi, C. A systematic approach for detecting and clustering distributed cyber scanning. *Comput. Netw.* **2013**, *57*, 3826–3839. [[CrossRef](#)]
14. Michael, J.B.; Wingfield, T.C.; Wijesekera, D. Measured responses to cyber attacks using Schmitt analysis: A case study of attack scenarios for a software-intensive system. In Proceedings of the 27th Annual International Computer Software and Applications Conference, Dallas, TX, USA, 3–6 November 2003; pp. 622–626.
15. Dasgupta, D. Immuno-inspired autonomic system for cyber defense. *Inf. Secur. Tech. Rep.* **2007**, *12*, 235–241. [[CrossRef](#)]
16. Staniford, S.; Hoagland, J.A.; McAlerney, J.M. Practical automated detection of stealthy portscans. *J. Comput. Secur.* **2002**, *10*, 105–136. [[CrossRef](#)]
17. Neri, F. Mining TCP/IP traffic for network intrusion detection by using a distributed genetic algorithm. In *European Conference on Machine Learning*; Springer: Berlin, Germany, 2000; pp. 313–322.
18. Ahmad, S.; Baig, Z. Fuzzy-based optimization for effective detection of smart grid cyber-attacks. *Int. J. Smart Grid Clean Energy* **2012**, *1*, 15–21. [[CrossRef](#)]
19. Bapiyev, I.M.; Aitchanov, B.H.; Tereikovskiy, I.A.; Tereikovska, L.A.; Korchenko, A.A. Deep neural networks in cyber attack detection systems. *Int. J. Civ. Eng. Technol. (IJCIET)* **2017**, *8*, 1086–1092.
20. Karimipour, H.; Dehghantanha, A.; Parizi, R.M.; Choo, K.K.R.; Leung, H. A deep and scalable unsupervised machine learning system for cyber-attack detection in large-scale smart grids. *IEEE Access* **2019**, *7*, 80778–80788. [[CrossRef](#)]
21. Mérien, T.; Bellekens, X.; Brosset, D.; Claramunt, C. A spatio-temporal entropy-based approach for the analysis of cyber attacks (demo paper). In Proceedings of the 26th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, Seattle, WA, USA, 6–9 November 2018; pp. 564–567.
22. Chen, Y.Z.; Huang, Z.G.; Xu, S.; Lai, Y.C. Spatiotemporal patterns and predictability of cyberattacks. *PLoS ONE* **2015**, *10*, e0124472. [[CrossRef](#)] [[PubMed](#)]
23. Hu, Z.; Baynard, C.W.; Hu, H.; Fazio, M. GIS mapping and spatial analysis of cybersecurity attacks on a florida university. In Proceedings of the IEEE 2015 23rd International Conference on Geoinformatics, Wuhan, China, 19–21 June 2015; pp. 1–5.
24. Lin, M.S.; Chiu, C.Y.; Lee, Y.J.; Pao, H.K. Malicious URL filtering—A big data application. In Proceedings of the 2013 IEEE International Conference on Big Data, Santa Clara, CA, USA, 6–9 October 2013; pp. 589–596.
25. Feroz, M.N.; Mengel, S. Phishing URL detection using URL ranking. In Proceedings of the 2015 IEEE international Congress on Big Data, Santa Clara, CA, USA, 29 October–1 November 2015; pp. 635–638.
26. Du, H.; Yang, S.J. Temporal and spatial analyses for large-scale cyber attacks. In *Handbook of Computational Approaches to Counterterrorism*; Springer: Berlin, Germany, 2013; pp. 559–578.
27. Koike, H.; Ohno, K.; Koizumi, K. Visualizing cyber attacks using IP matrix. In Proceedings of the IEEE Workshop on Visualization for Computer Security (VizSEC 05), Minneapolis, MN, USA, 26 October 2005; pp. 91–98.
28. Canadian Institute for Cybersecurity. URL Dataset (ISCX-URL2016). Available online: <https://www.unb.ca/cic/datasets/url-2016.html> (accessed on 28 April 2020).
29. Central Intelligence Agency. The World FactBook. Available online: <https://www.cia.gov/library/publications/the-world-factbook/fields/204rank.html> (accessed on 28 April 2020).
30. Kulldorff, M. *SaTScan—Software for the Spatial, Temporal, and Space-Time Scan Statistics*; Harvard Medical School and Harvard PilgrimHealth Care: Boston, MA, USA, 2015.
31. Altman, D.G. *Practical Statistics for Medical Research*; Chapman and Hall: London, UK, 1991.