symantec.

Phishing Attacks In and Around April through September 2006

Symantec Corporation

Version 1.0

Author:

Zulfikar Ramzan
Symantec Security Response
Advanced Threat Research

30th November 2006

Executive summary

This paper describes phishing trends from April 2006 through September 2006 inclusive. Our aim is to perform the kind of careful analysis done in the *Symantec Internet Security Threat Report*, but with a focus on phishing. We looked at data that originated from both the Symantec Brightmail AntiSpam and Norton Confidential systems. We were specifically interested in the overall magnitude of phishing, the geographic breakdowns of phishing servers, and the segmentation of brands spoofed in a phishing attack. Some of the more noteworthy observations include the following:

➢ The analyzed data supports seasonal and "weekend" type effects in terms of observed phishing data. In particular, there is a drop in phishing activity during the summer months, and also on Sundays and Mondays.

➢ The number of unique brands being spoofed did not steadily increase during the period studied; this observation may indicate that phishers are narrowing their focus on which brands to spoof. The analyzed data also shows that, on average, the number of recipients for a given unique phishing email is growing, which seems to imply that targeted phishing campaigns are outweighed by more scattered ones.

➢ For brands associated with specific American states, Florida was the most targeted in terms of unique phishing attacks (followed by California, New York, Wyoming, and Michigan); the correlation coefficient between percentage of phishing attacks targeting a specific state and the state's elderly population is +0.61 (and this correlation is stronger than what one obtains by considering the state's overall population, average per-capita income, or number of affluent counties).

➢ Nine of the top ten spoofed brands are in the financial sector. Six of these brands are based in the United States, and four are headquartered in the United Kingdom.

➢ The United States also hosts the largest percentage of phishing servers (46 percent), followed by Germany, Korea, the United Kingdom, and China.

This paper details the statistics we computed, the methodology used to derive these statistics and the limitations of our methodology.

1 Introduction

The last few years has seen a rise in the frequency with which people have conducted meaningful transactions online; from making simple purchases to paying bills to banking – and even to getting a mortgage or car loan or paying their taxes. This rise in online transactions has unfortunately been accompanied by a rise in attacks. Phishing attacks, which are the focus of this paper, typically stem from a malicious email that victims receive effectively convincing them to divulge sensitive information – which can then be later used to the detriment of the victim.

In many ways, phishing is an evolutionary threat – a natural analogue of various confidence games (for example, ones involving telephone solicitation) that existed in the bricks and mortar world. However, with the advent of the online world, phishing becomes a bigger threat for several reasons. First, it's relatively easy to automate a phishing attack – every step can be carried out online, and little human involvement is necessary (in fact, one can even purchase ready-made kits for carrying out attacks). From the phisher's perspective, his investment is low, but the returns could be quite high. Second, the likelihood of success is potentially higher; i.e., it is very easy for people to "mess up". Nowadays, you can divulge your data literally in seconds. Furthermore, your money can vanish from your account literally seconds later. Finally, with the increase in online transactions, there is bound to be one phishing attack attempt that is sufficiently believable (because the victim might really believe that a particular email really applies to him or her).

Phishing is a problem for several other reasons:

➢ It costs consumers real money.

➢ Organizations whose brands have been spoofed in a phishing attack have to bear the support costs; e.g., dealing with frantic customers whose accounts have been emptied or who are wondering about a suspicious email they have received.

➢ Many organizations depend almost exclusively on the online medium to carry out their business; these organizations could potentially suffer if individuals are skittish and stop carrying out transactions online.

➢ Many organizations use email to reach their customers. If customers start to think legitimate emails are in fact phishing emails, then they will start to ignore them, and these organizations will lose out on the benefits of email as a low-cost (and generally convenient) communications channel.

This paper strives to give the reader a better understanding of this emerging threat and its magnitude. We examine phishing data taken from Symantec's Brightmail AntiSpam System (from April 1, 2006 to September 30, 2006) and Norton Confidential server (from June 1, 2006 to September 30, 2006). Note that the tenth edition of Symantec's Internet Security Threat Report focuses on data from January 1, 2006 to June 30, 2006. Therefore, the present paper contains previously unpublished analysis on data from the third quarter of 2006.

We looked at the following high-level areas:

➢ Aggregate number of unique phishing emails sent and blocked

➢ Geographic locations of phishing servers

➢ Trends in attack targeting – specifically, the number of unique brands spoofed as well as the average number of victims who receive a given phishing email

➢ Brands that are spoofed in phishing attacks (according to industry segmentation)

➢ Geographic segmentation of spoofed brands

➢ Characteristics of domestic local brands (i.e., brands that are specific to a given US state)

The analyzed data supports the following observations for the periods that were studied:

➢ There is a drop in phishing activity during the summer months, and also in general on Sundays and Mondays; these constitute seasonal and weekend-type effects.

➢ Phishers are narrowing their focus on which brands to spoof; in particular, the number of unique brands targeted stayed nearly flat between July and August, and actually went down in September.

➢ The effects of scattered phishing attacks outweighs the effects of targeted phishing attacks; on average the number of recipients for a given (unique) phishing email has been growing.

➢ Of the top ten brands spoofed in a phishing attack, nine are in the financial sector; of these, six brands are based in the United States, and four are headquartered in the United Kingdom.

➢ Nearly half of all phishing severs (46 percent) are hosted in the United States; rounding out the top five are Germany (8.51 percent), followed by Korea (3.69 percent), the United Kingdom (2.75 percent), and China (2.53 percent).

➢ With respect to spoofed brands that are specific to a given US State, Florida was the most targeted; rounding out the top five are California, New York, Wyoming, and Michigan.

➢ The correlation coefficient between percentage of phishing attacks targeting a specific state and the state's elderly population is +0.61. This correlation is stronger than what one

obtains by considering the state's overall population (+0.58), average per-capita income (+0.19), or number of affluent counties (+0.44).

This paper is organized as follows:

➢ Section 2 describes the data sources we used as well as any limitations on these sources.

➢ Section 3 describes aggregate phishing statistics taken from the Symantec Brightmail AntiSpam System Data.

➢ Section 4 describes the geographic location of phishing servers taken from the Symantec Norton Confidential data.

➢ Section 5 describes trends in attack targeting; in particular that the number of unique brands does not seem to be growing month-to-month (from the Symantec Norton Confidential data), and that on average emails are not being sent to more targeted sources (from the Symantec Brightmail AntiSpam data).

➢ Section 6 categorizes the brands spoofed in phishing attacks according to industry segmentation.

➢ Section 7 considers the geographic segmentation of these brands.

➢ Section 8 examines the characteristics of data related to local brands (i.e., brands that are local to a given US state).

2   Data sources

We gathered phishing data from the following sources: the Symantec Brightmail AntiSpam System and the Symantec Norton Confidential system.

For the Brightmail AntiSpam System, we considered data taken from April 1, 2006 to September 30, 2006. For the Norton Confidential System, our data covered the June 1, 2006 to September 30, 2006 period. Note that the tenth edition of the *Symantec Internet Security Threat Report* covers the period from January 1, 2006 to June 30, 2006. Therefore, the present paper contains previously unpublished analysis of data from the July 1, 2006 to September 30, 2006 period.

Symantec's Brightmail AntiSpam System is a prevalent antispam offering. It collects unsolicited spam emails through several means. First, Brightmail uses two million decoy email accounts. Second, Brightmail is used by a number of major Internet Service Providers and free email account providers. As a result, on the order of one third of all email sent around the world is processed by Brightmail. Brightmail is able to detect unsolicited emails through a combination of heuristics, human analyst determination, email fingerprinting, and intelligence provided from partners and customers. Brightmail categorizes the unsolicited emails that appear to be phishing attempts.

Brightmail uses sensors to record the total number of unique phishing emails per day as well as the total number of blocked phishing attempts per day. Note that a given unique email may be sent to multiple recipients and blocked at each one; therefore the number of blocked phishing attempts is always at least as large as the number of unique phishing attempts. Also, note that there may be multiple unique emails that point users to the same phishing URL.

The second data source is Symantec's Norton Confidential anti-phishing technology (which is utilized in several Symantec products, such as Norton Confidential and Norton Internet Security 2007). Symantec's Norton Confidential is a consumer-oriented product that aims to provide transaction security. Two important components of Norton Confidential are its anti-phishing and anti-crimeware engines. On the back end, the Norton Confidential server collects phishing URLs through several sources including, but not necessarily limited to, the following:

➢ A number of feeds including those from the Symantec Phish Report Network; the Phish Report Network feed includes data provided by various contributors.

- ➢ Actual customers who browse to a phishing site on one of the products that leverages the Norton Confidential anti-phishing technology, including Symantec Norton Internet Security 2007 and Norton Confidential.
- ➢ An online reporting mechanism for people who wish to report phishing sites.

Through a number of heuristics, as well as human analyst input, Norton Confidential can both identify phishing sites and tag each phishing URL with the brand that is being spoofed in the attack.

The phishing data analyzed in this paper reflects what we specifically know about. As with any real-world data, there are natural biases that occur. We discuss these biases in the appendix, but stress that any conclusions one draws from the statistics we present must take these biases into consideration.

3   Aggregate Phishing Statistics

According to the tenth edition of the *Symantec Internet Security Threat Report*, from January 2006 through June 2006, the Brightmail system blocked 1.3 billion phishing attempts and recognized 157,477 unique phishing emails. Since then, during the July 2006 through September 2006 time period, Brightmail blocked an additional 790 million phishing attempts and recognized 85,106 more unique phishing emails. See TABLE 1. Assuming a sustained rate, it appears that the number of blocked attempts and unique phishing emails for the second half of 2006 is on target to surpass the totals reported from the first half of the year.

We broke down the Brightmail data by month. We found an interesting seasonal effect in that the number of unique phishing emails dropped during the summer months. In particular, the number of unique phishing emails dipped from May (28,573) to June (24,819). While the numbers started climbing again in July (25,987) and August (27,995), it is not until September (31,124) that the number rises again to the pre-summer levels. See TABLE 2 and FIGURE 1. The mean number of phishing emails sent per day during the April 1, 2006 to September 30, 2006 period was 905. The median was 831 and the standard deviation was 406 with a kurtosis of 8.8 and a skewness of 2.2. The mean blocked phishing attempts per day was 7,487,465. The median was 7,521,050 and the standard deviation was 2,263,181 with a kurtosis 0.22 and a skewness of 0.42. TABLE 3 summarizes these descriptive statistics.

Period	Blocked phishing Attempts	Unique Phishing Emails
Jan 06 - Jun 06	1.3 Billion	157,477
Jul 06 - Sep 06	790 Million	85,106

TABLE 1: BLOCKED AND UNIQUE PHISHING ATTEMPTS (SOURCE: BRIGHTMAIL DATA)

Month	Blocked phishing Attempts	Unique Phishing Emails
April	149,168,677	27,149
May	192,300,932	28,573
June	239,169,184	24,819
July	290,059,522	25,987
August	251,066,039	27,995
September	248,441,740	31,124

TABLE 2: BLOCKED AND UNIQUE PHISHING ATTEMPTS (SOURCE: BRIGHTMAIL DATA)

Statistic	Blocked Phishing Attempts	Unique Phishing Emails
Mean	7,487,465	905.2
Median	7,521,050	831
Standard Deviation	2,263,181	406.2
Kurtosis	0.22	8.8
Skewness	0.42	2.2

Table 3: Descriptive statistics for Apr 1, 2006 to Sep 30, 2006 phishing data (Source: Brightmail data)
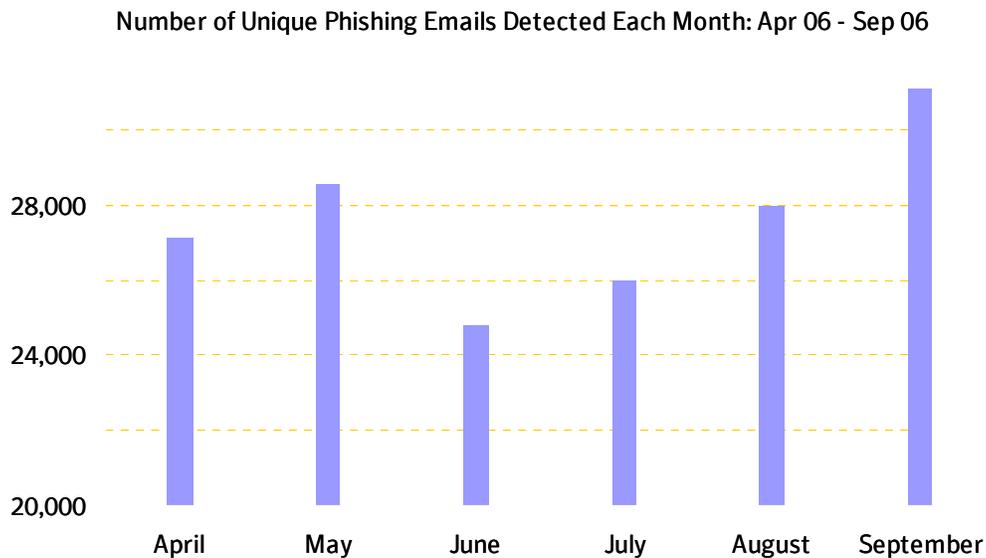
Number of Unique Phishing Emails Detected Each Month: Apr 06 - Sep 06

Figure 1: Unique phishing emails by month (source: Brightmail data)

We also broke this data down by the day of the week. For the purposes of taking aggregate statistics, our data does not include the final day of the period (Sep 30) to ensure that each day of the week is represented the same number of times (26 times). Here, the data supports some interesting day-of-week effects. In particular, the number of unique phishing emails is down dramatically on Sundays and Mondays. The average number of unique phishing emails on Sundays (16,933 unique) and Mondays (16,173 unique) drops by nearly 36 percent and the average number of blocked attempts drops by almost 8 percent when compared with the average for the remaining days of the week. Starting on Tuesday, the number of unique phishing emails goes up to 25,702 and slowly increases until its peak on Thursday when it reaches 28,537. On Friday, the number drops to 24,907 and stays nearly the same on Saturday before plummeting on Sunday.

Day of Week	Blocked Phishing Attempts	Unique Phishing Emails
Sunday	181,614,938	16,933
Monday	185,037,189	16,713
Tuesday	199,475,152	25,702

Wednesday	200,792,165	27,432
Thursday	196,233,775	28,537
Friday	199,918,785	24,907
Saturday	198,750,130	24,549

Table 4: Day of week breakdown - Apr 1 to Sep 29, 2006 (Source: Brightmail data)

4   Locations of phishing servers

We are interested in geographically segmenting the location of phishing servers. We determined this information as follows. We went through the URLs from the Symantec Norton Confidential data and used the NSLOOKUP service to obtain the corresponding IP address (in some cases, the URL itself contained an IP address, so we did not need to perform a reverse look-up). We discarded any URL for which we could not reliably obtain the IP address. Then, we ran these IP addresses through services that determine the (physical) geographic location corresponding to the IP address. In addition, these services assign a confidence score (from 0 to 5) for each response. We discarded any response with a score lower than 4. The result is that we determined the geographic locations for a more than 10,000 phishing servers. Note that the physical location of a phishing server may differ from the physical location of the people who are responsible for carrying out the actual phishing attack. Also, note that there are some limitations in our approach.

The most notable limitation is that we performed geographic look-ups after the phishing attacks occurred. So, for those phishing URLs that do not contain an IP address, there is a chance that the URL might have changed IP addresses (and locations) between the time the actual attack was carried out and the time that we performed the look-up. To partially compensate for this limitation, we have restricted our reporting to the country level (because even if a server changed physical locations, it is less likely to have moved from one country to another). Nearly half of the phishing servers are located is the United States (46.11 percent). Germany (8.51 percent), Korea (3.69 percent), the United Kingdom (2.75 percent), China (2.53 percent), and Taiwan (2.51 percent) round out the next five. See Table 5 for a breakdown of top 25 countries hosting phishing servers.

Country	Percentage
United States	46.11 percent
Germany	8.51 percent
Republic of Korea	3.69 percent
United Kingdom	2.75 percent
China	2.53 percent
Taiwan	2.51 percent
Canada	2.46 percent
Netherlands	2.29 percent
France	2.21 percent
Japan	2.03 percent
Russian Federation	1.72 percent
Hong Kong	1.58 percent
Italy	1.42 percent
India	1.20 percent
Spain	1.16 percent
Denmark	1.15 percent
Thailand	1.14 percent
Brazil	1.11 percent
Australia	1.05 percent
Poland	1.02 percent
Romania	0.89 percent
Switzerland	0.74 percent
Czech Republic	0.69 percent
Belgium	0.62 percent
Argentina	0.53 percent

Table 5: Top 25 countries that host phishing servers (source: Norton Confidential data)

5   Trends in attack targeting: unique brands and email reach

This section explores data that relates to the extent to which attacks are getting targeted. In particular, for the periods studied, our data does not support the hypothesis that attackers are going after more and more specialized targets. Also, for the periods studied, our data indicates that targeted phishing campaigns are outweighed by more scattered ones.

Let us consider unique brands first. From the June through September 2006 period, the Symantec Norton Confidential System recorded 154 distinct brands that were spoofed in a phishing attack. Of these 154 brands, 93 of them were spoofed in a phishing attack that occured during June; this number jumped to 109 in July, dipped to 108 in August, and dipped again in September to 101.   (A brand is counted in the tally of each month that it is spoofed.)

The number of unique brands that were spoofed per month does not seem to be rising steadily as one might expect. In fact, the number of unique brands that were spoofed declined in August and again in September. The decline from July to August is very slight (only one brand), whereas the decline from August to September is 6.5 percent, which is more pronounced.   See TABLE 6 and FIGURE 2.

The analyzed data represents only known phishing sites (which only encompasses a subset of the phishing sites that exist). Also, it is more challenging to gather data about attacks that target lesser-known brands because the corresponding phishing email might only go to a smaller number of people (and diminish the likelihood that the site comes to our attention). So, while our data clearly does not support the hypothesis that phishers are targeting more brands, the data may not be robust enough to contradict this hypothesis.
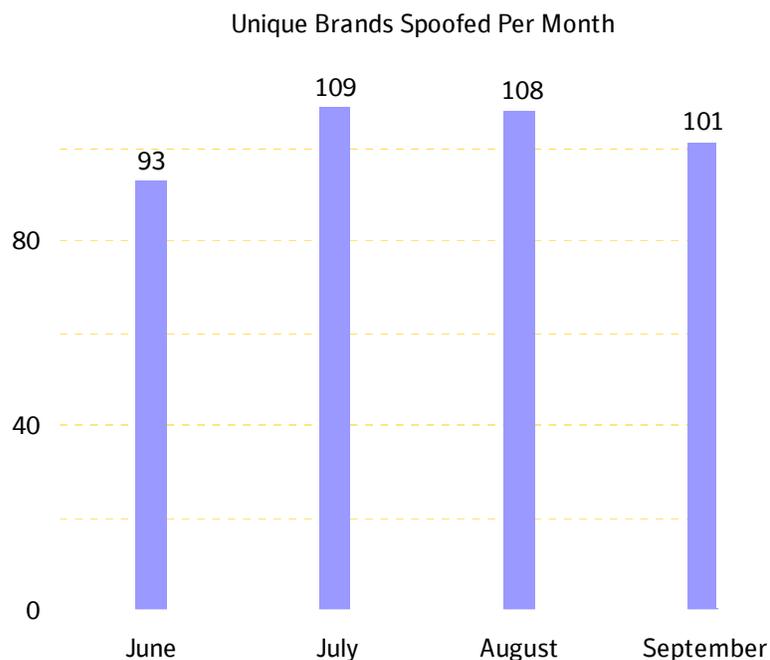

FIGURE 2: UNIQUE BRANDS PHISHED PER MONTH (SOURCE NORTON CONFIDENTIAL DATA)

Month	Unique Brands Phished
June	93
July	109
August	108
September	101

The next area we investigated is whether phishing email campaigns, as a whole, are becoming more targeted. To measure this effect, we looked at the ratio between the number of blocked phishing attempts and unique phishing emails from the Brightmail data. This ratio gives an indication of how many times a given phishing email is being sent out to targets. A high ratio indicates that, on average a given email is being sent to many individuals (i.e., attacks are scattered). A low ratio indicates that, on average, emails are being sent to fewer people (i.e., attacks are more targeted). It is important to note that a single scattered attack may outweigh the effects of multiple targeted attacks in this ratio.

From the Brightmail data we computed this ratio, for each day between April 1, 2006 and September 30, 2006. The mean ratio per day is 9,608 with a standard deviation of 4,696, kurtosis of 2.94 and skewness of 1.30. The minimum is 1,400 and the maximum is 31,210. We ran a regression to compute a trend line and found it has a positive slope. See Figure 3. So, a given email is, on average, being sent to more people. Note that the analysis here is based on the aggregate daily totals. So, it is possible that the number of targeted attacks has increased, but the effect could not be observed because of a highly scattered attack. The wide range and high standard deviation further support the assertion that some attacks are highly scattered. Note also that the R^2 value is small, but because the goal is to see the trend rather than be predictive, we believe the R^2 value has less significance (the primary objective is to see whether the linear fit that gives the smallest error has positive slope, which appears to be the case).

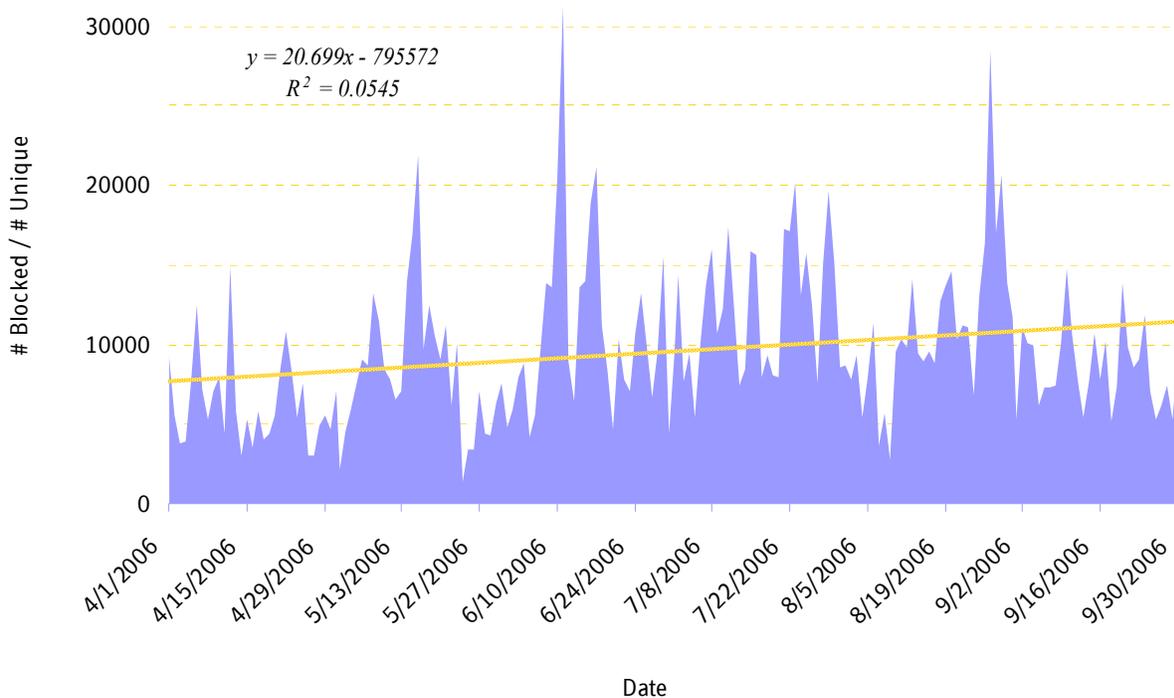$$y = 20.699x - 795572$$
$$R^2 = 0.0545$$

Figure 3: The ratio of blocked attempts to unique phishing attempts over time for capturing aggregate scattering level of phishing campaigns and trend line (source: Brightmail data).

6   Spoofed brand industry segmentation

This section analyzes the industry segmentation of the brands spoofed in a phishing attack. We divided the spoofed brands into the following categories:

➢ **Financial:** sites associated with online banking, brokerage, lending, and similar financial services or sites that directly support such a brand;

➢ **Service Provider:** sites that provide some common internet-related services including one or more of the following: internet access, email accounts, or information portals;

➢ **General Retail:** sites that are associated with the sale of merchandise online;

➢ **Computer Hardware:** sites that are associated almost exclusively with the sale of computer hardware and peripherals;

➢ **Government:** sites whose common URL ends in the .gov extension;

➢ **Social Networking:** sites whose exclusive purpose is to facilitate connection, collaboration, and communication among members resulting, possibly, in the formation of online communities;

➢ **Certificate Authority:** sites whose purpose is to issue digital certificates for the purposes of enabling PKI-leveraging services such as Secure Sockets Layer (SSL) communication.

We then went through the Norton Confidential data and ranked each spoofed brand by the number of unique phishing URLs associated with that brand for the period from June 2006 through September 2006. It turned out, not too surprisingly, that nine of the top ten spoofed brands are in the financial sector.

The first online service provider came in at the 11th spot. Respectively, the first government, social networking, certificate authority and hardware sites came in at 14, 41, 88, and 107. See TABLE 7. In terms of the overall data picture, the financial sector represented almost 84 percent of spoofed brands; retail came in second at 5.19 percent, and the remaining sectors were all below 5 percent. See TABLE 8, and FIGURE 4. Again these numbers are not surprising because phishers are motivated by economic interests and therefore are more likely to go after financially-oriented brands. For example, even the Government sites that were spoofed were financially oriented.

Sector	Highest ranking spoofed brand
Service Provider	11
Government	14
Social Networking	41
Certificate Authority	88
Hardware	107

TABLE 7: HIGHEST RANKING SPOOFED BRAND BY SECTOR (SOURCE: NORTON CONFIDENTIAL DATA)

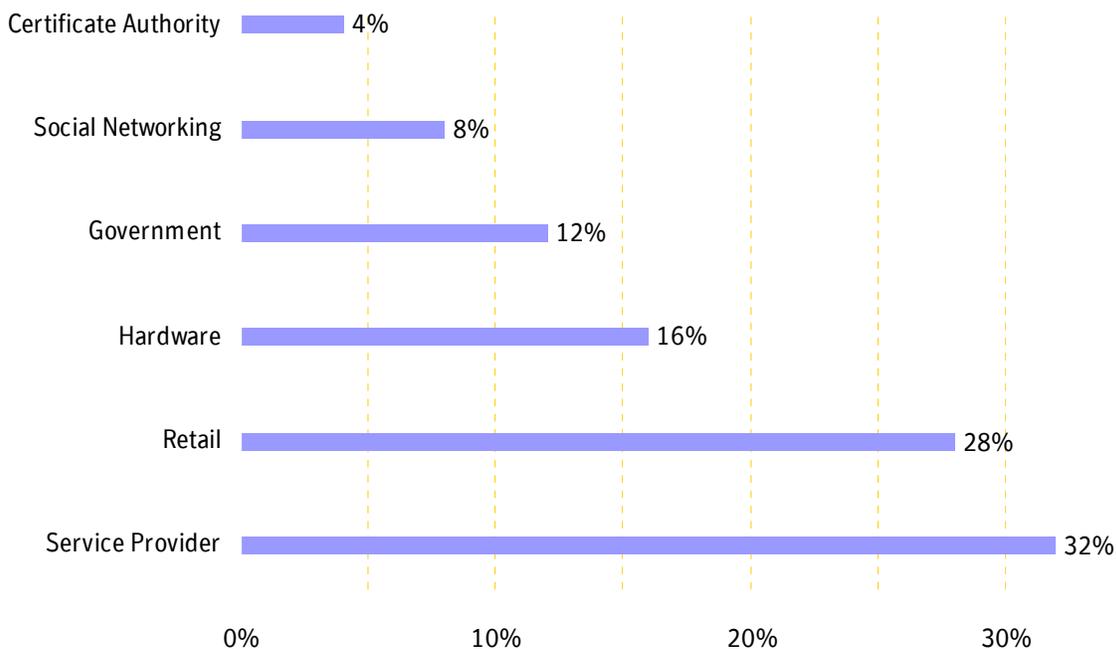| Sector	# Spoofed Brands	Percentage
Financial	129	83.77 percent
Service Provider	8	5.19 percent
Retail	7	4.55 percent
Hardware	4	2.60 percent
Government	3	1.95 percent
Social Networking	2	1.30 percent
Certificate Authority	1	0.65 percent

FIGURE 4: INDUSTRY BREAKDOWN OF SPOOFED BRANDS OUTSIDE FINANCIAL SECTOR (SOURCE: NORTON CONFIDENTIAL DATA).

7 Geographic Segmentation of Spoofed Brands

We segmented the spoofed brands according to geographic location. We went through the 154 spoofed brands and manually assigned a country to each brand based on where the corporate headquarters of that brand was located (keep in mind that a company may do business in multiple countries even if it is headquartered in just one). Of the top ten spoofed brands, six are based in the United States. The remaining four are headquartered in the United Kingdom. We remark that several of these United States brands actually have extensive global reach, so while they are headquartered in the United States, it would incorrect to label them as exclusively associated with the United States.

Overall, we found that 72 percent of the 154 spoofed brands are based in the United States. The remaining 28 percent are from outside the United States. Of the 28 percent that are not from the United States, the United Kingdom has largest number of spoofed brands (nine brands), followed by Canada (seven brands), Spain (six brands), Germany (six brands), and Australia (four brands). The remaining countries have one or two spoofed brands. Of the phishing sites targeting brands outside the United States, the United Kingdom had the overwhelming majority with slightly over 69.31 percent. Canada came in a distant second with 9.29 percent of non-US phishing sites, followed by Germany (7.76 percent) and Spain (4.43 percent). See TABLE 9.

As expected, the number of spoofed brands correlated roughly with the overall percentage of actual phishing sites attacking those brands (correlation coefficient 0.74). So, we found that countries with more spoofed brands also had more phishing URLs associated with those brands. Note that while this

observation is expected, it is not always true. For example, it is plausible that there could be many spoofed brands, but where each brand is only associated with a small number of actual phishing sites. In our data, Australia had 4 spoofed brands and roughly 2 percent of non-US phish sites. Mexico had half as many spoofed brands, but nearly twice as many phishing sites.

Although we mentioned it already, we would like to re-emphasize that our data is taken from specific sources and one has to consider the distribution induced by that source when drawing conclusions from the statistics we present.

Country	# Spoofed Brands	Percent of phishing
Australia	4	2.04 percent
Canada	7	9.29 percent
Dubai	1	0.06 percent
Germany	5	7.76 percent
Ireland	1	0.43 percent
Isle of Man	1	0.14 percent
Italy	2	1.50 percent
Mexico	2	3.97 percent
Netherlands	1	0.02 percent
South Africa	1	0.89 percent
Spain	6	4.43 percent
Sweden	1	0.18 percent
United Kingdom	9	69.31 percent

TABLE 9: NUMBER OF SPOOFED BRANDS ASSOCIATED WITH A GIVEN COUNTRY AND THE PERCENTAGE OF PHISHING SITES ASSOCIATED WITH THOSE BRANDS (SOURCE: NORTON CONFIDENTIAL DATA)

8 Local banking brands

Among all the spoofed brands from the June through September 2006 Norton Confidential data, we examined the distribution of those that represent local US banks; for example, credit unions that are local to a specific state. For this purpose, we considered a brand to be local if all the branch locations were in a specific state (or in states that directly bordered that state).

A total of 42 local banking brands across 23 states were spoofed in phishing attacks from June through September 2006. As one might expect, the number of spoofed brands and phishing URLs seem to be positively correlated with state population. We computed the actual correlation coefficient between individual state populations (from the 2000 US Census numbers) and the percentage of phishing sites, to obtain a correlation coefficient of +0.58. To make the comparison fair, states that did not have any local spoofed brands were included at 0 percent.

However, there are a few states that have a disproportionate number of spoof sites. The most noticeable is Florida, which has 3 spoofed brands and 14.24 percent of phishing sites - more than any other state. That is, of those phishing attacks that target a local bank brand, 14.24 percent of them targeted a Florida bank. One might venture to guess that phishers have targeted Florida due to the high elderly population in Florida (because this demographic is often targeted in scams). The other interesting outliers are Colorado (3 spoofed brands, 7.95 percent of spoof sites), Indiana (2 spoofed

brands, 5.96 percent of spoof sites), Wisconsin (3 spoofed brands, 5.30 percent of spoof sites), and Wyoming (1 spoofed brand, 9.60 percent of spoof sites).   Of these, Wyoming has the largest phishing site to spoofed brand ratio. See Table 11 and Figure 5.

We more formally measured the correlation between elderly population and the percentage of spoof sites local to a given state using the 2000 Census numbers for population older than 65. The corresponding correlation coefficient is +0.61 which indicates a slightly stronger positive correlation than just the population metric alone (it is important to keep in mind, though, that the population of elderly is positively correlated with overall population; the correlation coefficient is 0.97).

Next, we measured the correlation between the percentage of spoof sites and the average per-capita income of a state (from the 2005 census). The correlation coefficient was only +0.19.   In fact, five of the top ten states in per capita income did not have any phishing attacks associated with local brands (Connecticut, New Jersey, Maryland, New Hampshire, and Delaware).   One possible reason for this trend is that per capita income represents an average. Some states may exhibit a division of wealth with some portions of the population being especially affluent. The 2000 census numbers indicate the top 100 counties nationwide in per capita income. For each state, we computed the number of counties in that state that are in the top 100.   Correlating this number with the percentage of phishing sites yields a coefficient of +0.44. See Table 10.

Factor	Correlation with percent phishing sites
Population	+0.58
Elderly Population	+0.61
Mean Per Capita Income	+0.19
Number of Affluent Counties	+0.44

Table 10: Correlation between the percentage of phishing sites spoofing a state-specific brand and other characteristics (source: Norton Confidential data, US Census data)

State	# Local brands	% Phishing sites
Alaska	1	3.64%
Arizona	1	0.66%
Arkansas	1	0.33%
California	5	10.60%
Colorado	3	7.95%
Florida	3	14.24%
Georgia	1	0.66%
Hawaii	1	0.66%
Illinois	1	0.66%
Indiana	2	5.96%
Maine	1	2.32%
Massachusetts	1	0.99%
Michigan	3	8.94%
Minnesota	1	0.66%
New York	3	9.60%
Oregon	1	4.97%
Pennsylvania	2	0.33%
Texas	4	6.62%
Utah	1	1.32%
Virginia	1	0.33%
Washington	1	3.64%
Wisconsin	3	5.30%
Wyoming	1	9.60%

TABLE 11: NUMBER OF LOCAL BRANDS SPOOFED AND PERCENTAGE OF PHISHING ATTACKS ON THESE BRANDS BROKEN DOWN BY STATE (SOURCE: NORTON CONFIDENTIAL DATA)
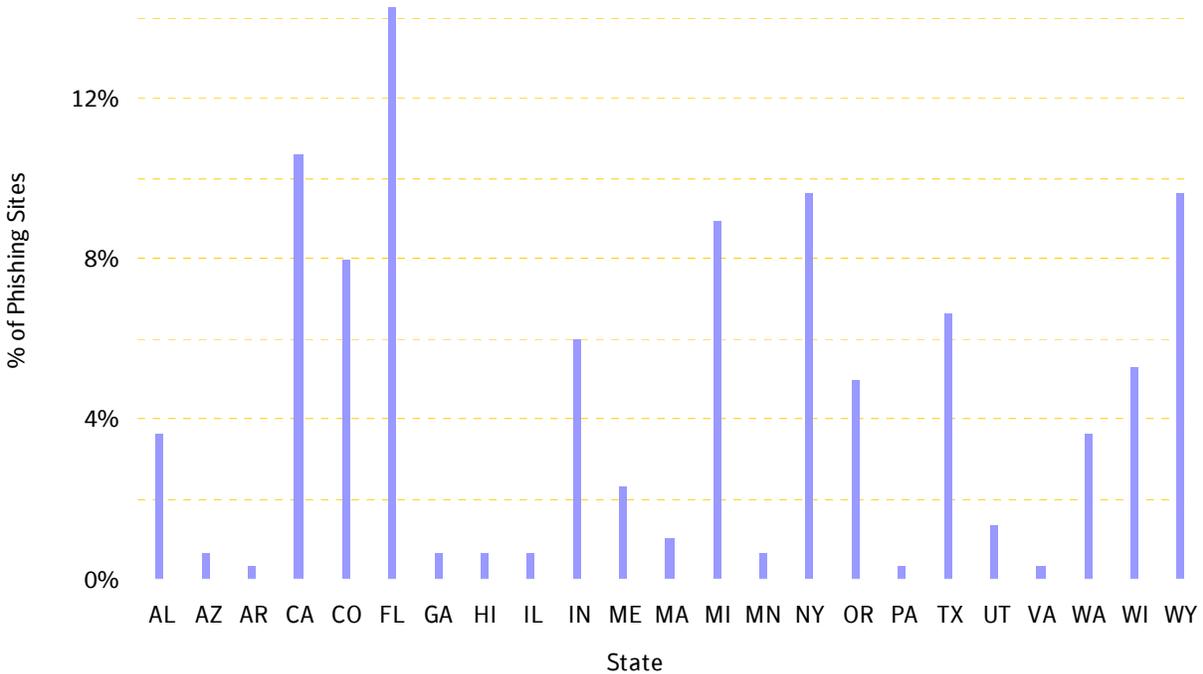
Figure 5: distribution of phishing attacks spoofing state-specific local brands (source: Norton Confidential data).

9 Conclusions

In this paper, we looked at phishing data that was taken from the Symantec Brightmail AntiSpam System from April 1, 2006 to September 30, 2006 and from the Symantec Norton Confidential server from June 1, 2006 to September 30, 2006. We analyzed this data and reported on characteristics that we thought were particularly noteworthy. In particular, we looked at aggregate phishing statistics, geographic locations of phishing servers, the industry segmentation of the brands being spoofed, the geographic segmentation of these brands, and the properties of local brands. We observed the following:

➢ Our data supports seasonal and "weekend" type effects in overall phishing. Specifically, phishing activity drops during the summer months, and also on Sundays and Mondays.

➢ Our data did not support the conventional wisdom that phishers are targeting more unique brands each month. Specifically, the number of spoofed brands remained essentially flat between July and August, and decreased in September. Also, on the whole, phishing emails do not appear to be getting more targeted. The trend is that, on average, a given unique phishing email is being sent to more individuals.

➢ Of the domestic banking brands that are affiliated with a given state, Florida was the most targeted in terms of unique phishing attacks. The next four states were California, New York, Wyoming, and Michigan.

➢ The correlation coefficient between the percentage of phishing attacks specific to a given state and the elderly population of that state was 0.61. This correlation is stronger than that compared to overall population (0.58), per capita income (0.19), and number of affluent counties (0.44).

- ➢ Of the top ten spoofed brands in our data set, nine are in the financial sector. Six of these are brands are based in the United States, and the remaining four in the United Kingdom.

- ➢ The United States also hosts the largest percentage of phishing servers (46 percent), followed by Germany, Korea, the United Kingdom, and China.

Perhaps the most ominous conclusions one can draw from this paper are the following:

- ➢ Phishing is becoming a bigger problem in terms of sheer numbers;

- ➢ While the data analysis provides a better understanding of this threat, there is still a considerable amount that remains to be learned about phishing.

We hope that the lessons we have learned from this paper can not only guide us in future analyses of phishing data, but also improve our overall understanding of the problem so that we can continue to improve the countermeasures we develop.

Acknowledgements

Appendix: Data biases

This paper analyzes the data that Symantec collects. As is the case with almost any real-world data there are bound to be some biases; we discuss those here.

First, let us consider the Symantec Brightmail Antispam system. Generally speaking, the analysis done on unsolicited mails is rigorous and we believe this analysis leads to a high accuracy rate across the board for a variety of email samples. At the same time, the Brightmail System also benefits from intelligence that is provided by Symantec partners and customers. As a result of this additional input, our classification abilities on phishing emails that spoof partner and customer brands are correspondingly higher.

Second, the Symantec Norton Confidential system receives feeds from various partners who have made efforts to report on sites that spoof their brands. Also, the Symantec Norton Confidential system receives input data from an online reporting mechanism as well as client machines that have installed either the Symantec Norton Confidential software or the Symantec Norton Internet Security 2007 software. These sources are more likely to capture widespread phishing attacks than they are to see targeted attacks that are aimed at a very small population. Also, they are more likely to capture attacks that target the demographics of the installed base. The system can and does capture small-scale attacks; for example, section 8 discusses attacks on smaller, localized brands. Our main point, however, is that large-scale attacks are, by definition, more noticeable and hence more likely to be captured. On the other hand, this bias is partially offset by the numerous other data feeds that the system uses.

Finally, any study of phishing statistics can only analyze the data from known phishing sites. There are undoubtedly phishing attacks that will escape the net we cast. While we hope that these sites are few and far between, we are unaware of any scientifically rigorous way of determining how many attacks we missed and how representative our sample is. With these limitations in mind, we would be hesitant to make firm over-sweeping generalizations about trends in phishing attacks; instead we would prefer to view our data as either supporting or not supporting various hypotheses.

symantec.