

Phishing Attacks: Analyzing Trends in 2006

Zulfikar Ramzan
Symantec, Inc.
Mountain View, CA 94043

Candid Wüest
Symantec, Inc.
Zurich, Switzerland 8050

Abstract

This paper analyzes trends seen in phishing attacks throughout 2006 based on real-world data obtained through Symantec's phishing data collection fabric. We examine both the prevalence and breakdown of phishing web sites as well as the frequency and breakdown of phishing emails. Beyond just the extent of data collected, our study differs from previously published studies in this area in two regards:

- We discuss the data collection methodology (together with its limitations and biases) so that readers are better positioned to place the results in the appropriate context;
- We perform a fine-grained analysis considering seasonal & day-of-week effects, geographic distinctions, brand segmentations, and geographic/population targets.

We found a number of intriguing properties of phishing attacks. These include seasonal and day-of-week fluctuations in activity and fluctuations related to what brands are being spoofed. We also determined the industries, regions, languages, and population segments that appear to be targeted in these attacks.

1. Introduction

PHISHING EMAILS. Spam is a nuisance, causing unwanted Internet traffic and cluttering individuals' mailboxes. Estimates suggest that 59% of all email is spam [1]. Among these emails, one worrisome class are *phishing emails*; these emails appear to come from legitimate institutions and lure victims into divulging sensitive information. Typically, this involves asking the victim to click on a hyperlink which takes them to a fraudulent web site that *spoofs* a legitimate one; the web site usually includes a web form where sensitive information is requested.¹

¹While clicking on a hyperlink is the traditional way to lure a victim, some phishing attacks involve asking the victim to call a telephone number that the phisher controls [4]. Also, while email is the traditional vector for dissemination of phishing messages, we have observed instances involving Instant Messaging programs and also SMS messages on mobile phones.

Phishing emails can cause serious financial harm to those who fall for the lure. Furthermore, they are problematic for the institutions whose brands are being spoofed for three reasons. First, their brand is tarnished. Second, the institutions often bear the victims' financial burden (or they may implicitly pass the costs back to customers through higher fees). The institutions also find themselves in a dilemma since they may not wish to keep a particular customer who was victimized by phishing fearing recidivism; at the same time, they might not wish to ostracize customers for fear of a public relations backlash. Finally, end users might start ignoring the institutions' legitimate emails – thereby disabling email as an inexpensive form of communication from the institution to its clients. This paper examines phishing in more detail, drawing on an extensive collection of real-world phishing email and web site data.

OUR CONTRIBUTIONS. Our goal is to better understand the nature of the phishing threat. We step toward that by analyzing trends in phishing attacks, specifically for data collected during 2006. The following observations, which we describe in detail below, came up during our investigations:

- Phishing activity declines considerably on weekends with more than a 20% dip in unique phishing emails and a nearly 5% dip in number of blocked emails compared to weekdays;
- Phishing activity seems to increase considerably during times when people are pre-occupied with other events and perhaps likely to let their guard down – in particular, there was a ~30% increase (compared to the yearly averages) in the number of phishing messages Symantec blocked during the following events: the Superbowl, the FIFA World Cup final, Christmas, and New Years;
- Spoofed brands exhibit Pareto behavior with a small number of brands being spoofed in a substantial percentage of phish sites;
- Only 57 out of 343 brands are consistently spoofed each month. The month-to-month turnover among spoofed brands is substantial (~30%);
- While phishers² mainly spoof financial brands, other target sectors include commerce, community, govern-

²Throughout the paper, we refer to phishers as a collective entity; however, we only do so for convenience. There is evidence to suggest organized elements in phishing campaigns

ment, job boards, certificate authorities, and greeting cards;

- While US-based brands bear the brunt of phishing attacks, phishers are going after geographically diverse targets across 31 regions and 16 languages;
- A substantial number (122) of local banks have brands that are spoofed. The geographic locale of these banks seems to imply that phishers are going after specific population sectors like the elderly and students, who are traditional fraud targets.

DIFFERENTIATION FROM PRIOR WORK. There have been previous published analyses that consider phishing; e.g., the APWG trend reports [2] and the PhishTank statistics [3]. This paper differs from previous works in one or more of three ways:

1. **Methodology:** We provide a description of our data collection methodology (together with its limitations and biases) so that our results can be placed in the appropriate context;
2. **Data:** Our data collection fabric is among the most extensive, if not the most extensive, available. Therefore, our analysis accounts for more phishing emails and web sites, within the time periods studied, than other analyses. Also, the data is vetted at several levels to ensure its accuracy;
3. **Analysis:** In addition to summary statistics, we examine various breakdowns (such as day-of-week & seasonal fluctuations, geographic & industry segmentation of brands, and population targets). We also provide some informed speculation concerning these trends.

Finally, we remark that phishing appears to be very much an evolving threat. Therefore, we believe there is always value in examining recent trends.

ORGANIZATION. Section 2 describes our data sources and its biases. Section 3 analyzes phishing emails, considering not only their frequency, but also seasonal and day-of-week fluctuations. Section 4 analyzes the brands spoofed in phishing attacks, examining turnover, industry, geographic, and target victim population trends. Section 5 concludes.

2. Data Collection Methodology

We gathered phishing data from the Symantec Brightmail AntiSpam System and the Symantec Norton Confidential system. For the Brightmail AntiSpam System, we considered data taken from Jan 1, 2006 to Dec 31, 2006. For the Norton Confidential System, our data covered Jun 1, 2006 to Dec 31, 2006.

BRIGHTMAIL DATA. Symantec’s Brightmail AntiSpam System is a prevalent antispam offering. It collects unsolicited spam emails through several means. First, Brightmail uses over two million decoy email accounts. Second, Brightmail is used by a number of major Internet Service Providers and specific groups of phishers. However, different phishing groups may act independently, and one group’s motivations and tactics may differ from another’s.

and free email account providers. As a result, on the order of twenty-five percent 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 sub-categorizes unsolicited emails that appear to be phishing attempts. Brightmail uses sensors to record both the total number of unique phishing emails per day and 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 unique messages is a lower bound on the number of blocked phishing attempts. Also, note that there may be multiple unique emails that point users to the same phishing web site.

NORTON CONFIDENTIAL DATA. The second data source we employ is Symantec’s Norton Confidential anti-phishing server (which is utilized in several Symantec products, such as Norton Confidential and Norton Internet Security 2007). Symantec’s Norton Confidential is a transaction security product that, among other things, offers phishing protection. 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 [5]; the Phish Report Network feed itself includes data provided by various contributors. These contributors comprise companies who are aware of different web sites spoofing their own brands (as well as companies who themselves aggregate intelligence on phishing web sites).
- Actual customers who browse to phishing sites on products that use 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. Because the data is vetted at multiple levels, we can ensure that it has high integrity.

DATA BIASES. 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. 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, Brightmail benefits from intelligence that is provided by Symantec partners and customers leading to improved classification rates on phishing emails that spoof these brands. 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 4 discusses attacks we observed on 122 smaller, localized brands. Our main point, however, is that large-scale attacks are, by definition, more noticeable and hence more likely to be captured. At the same time, this bias is partially offset by the numerous other data feeds that the system uses. Finally, any study of phishing trends can only analyze data from known phishing sites. There are undoubtedly phishing attacks that our collection fabric will not capture. 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 phishing attack trends; instead we would prefer to view our data as either supporting or not supporting various hypotheses.

Having said that, we remark that any analysis of real-world data on phishing is bound to have similar biases as well. However, we believe that the extensive nature of our data sources help partially offset these biases.

3. Phishing Email Statistics

This section describes statistics related to both unique and blocked phishing emails recorded by Brightmail. The first part of the section gives aggregate statistics and the second part gives a temporal breakdown.

AGGREGATE STATISTICS. In 2006 Symantec’s Brightmail system blocked 2,848,531,611 phishing emails. Of these, 323,725 were unique phishing messages. On average, therefore, in 2006 there were 7.8 million blocked phishing attempts and 887 unique phishing messages each day. To get a better trend overview, we broke the data down by month and calculated the average per day for each month (Table 1). We found an interesting seasonal effect in that the number of blocked phishing emails dropped toward April – from 9,253,646 blocked phishing attempts on an average January day to nearly half of it with 4,972,289 per average day in April. Starting from May the numbers rose again to the normal level. In contrast, the average number of unique phishing emails stayed roughly flat during this time period.

One explanation for this drop might be the discovery and termination of a large bot-network. As bot-infected machines are often used to send out phishing emails, terminating a network would degrade normal attack output (which explains why the number of unique phishing messages was stable, but the throughput was lower than average).

TEMPORAL EMAIL BREAKDOWN. To gain insight into time-based phishing trends, we broke the data down by the day of the week, taking all 365 days of 2006 into account. We then computed the average number of blocked phishing attempts and unique phishing messages on each weekday. Figures 1 & 2 illustrate the results. On weekends (Saturday & Sunday) 20.7% fewer phishing messages were sent compared to the overall daily average of 887 unique phishing messages in 2006. The number of blocked phishing attempts was smaller

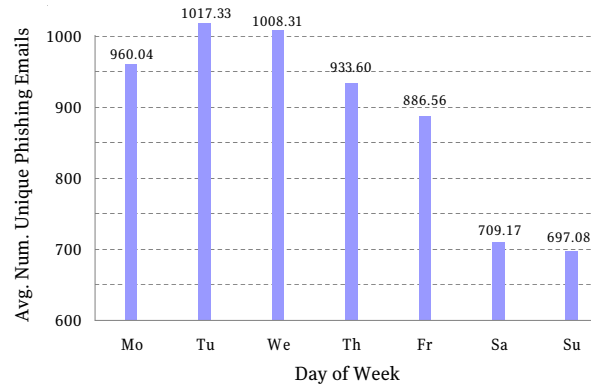


Figure 1: 2006: Avg. unique phishing messages per weekday.

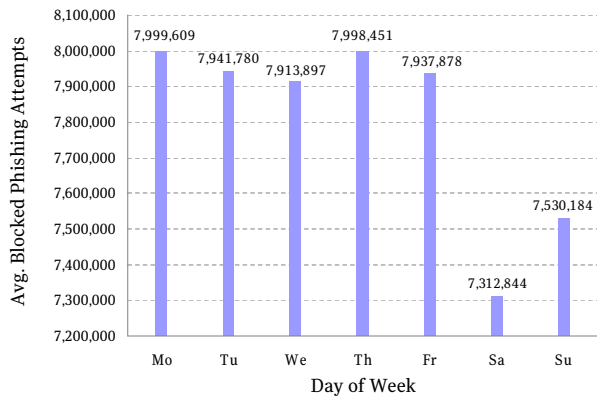


Figure 2: 2006: Avg. blocked phishing attempts per weekday.

	UNIQUE EMAILS	BLOCKED ATTEMPTS	AVG. UNIQUE EMAILS PER DAY	AVG. BLOCKED EMAILS PER DAY
JAN	20,020	286,863,040	645.8	9,253,646.5
FEB	26,215	216,116,532	936.3	7,718,447.6
MAR	30,701	217,958,268	990.4	7,030,911.9
APR	27,149	149,168,677	905.0	4,972,289.2
MAY	28,573	192,300,932	921.7	6,203,255.9
JUN	24,819	245,367,786	827.3	7,972,306.1
JUL	25,987	290,059,522	838.3	9,356,758.8
AUG	28,007	251,066,039	903.5	8,098,904.5
SEP	31,175	248,441,740	1039.2	8,281,391.3
OCT	22,881	261,450,869	738.1	8,433,899.0
NOV	30,048	241,847,259	1,001.6	8,062,475.3
DEC	28,150	254,062,549	908.1	8,195,566.1

Table 1: Monthly phishing email activity and averages

by 4.9% on weekends compared to the average of 7.8 million attempts per day.

This decline on Saturday and Sunday might suggest that attackers are resting on weekends, and perhaps treat their phishing activities as a normal full-time job. Another explanation is that phishing campaigns are short lived and therefore most effective for the attacker when people receive and read the emails soon after they were sent, which is not necessarily true on weekends. Therefore to maximize their success, phishers might time their attacks on working days. On the other hand, phishers might have *more* incentive to attack on weekends since it will be harder to shut down their phishing sites (given that many smaller web hosting services might have limited support on the weekends, and may be unable to shut down live phishing sites hosted on their server). This last point supports the resting-on-weekends hypothesis.

In addition to increased phishing activity on weekdays, during 2006 we also observed increased activity during big events or festive days like Christmas and New Years. These events might potentially make people more susceptible to social engineering attacks because they may be otherwise occupied and their guards (and natural baseline suspicions) could be lowered. On a related note, phishing sites set up during these times might prove more challenging to take down as the (unsuspecting) owner of the web server on which the site is hosted might be on vacation (or otherwise occupied) and unable to respond to requests. In 2006 there was a 33% increase (compared to the yearly average) in blocked phishing attempts around the week of the Super Bowl final, which took place on the Feb 5, 2006; In the week of the FIFA world cup final on July 9, 2006, there was a 140% increase over the yearly average (after already having been high for the beginning of the competition). During Christmas and New Years the number of blocked phishing messages peaked to 28.5% above the average. Some events are only of local importance and are therefore only seen in a restricted region. Diverse smaller events cumulatively added an effect to the global attack volume over the year or had regional effects, resulting in a minor rise or fall. While the data indicates spikes during specific periods of time, it would be premature to conclude that the phishers’ choice to attack during these was intentional. See Figure 3.

We note that the volume of phishing attacks could depend further on various other factors like the release of new security products and availability of vulnerabilities and patches.

After AntiPhishing vendors release new heuristic detections that catch most attacks, the attacker will need some time to adapt and invent new methods to possibly bypass those security measurements. Appearance of easy-to-use phishing tool kits in the underground market might cause a spike in sent phishing emails, once they are widely used. Also, phishers might tailor their target choice based on the existence of an effective cash-out mechanism. Furthermore, the number of existing bot-infected machines influences the attack statistics as well, as they are often used to send the phishing email messages.

Having looked at breakdowns of phishing emails, we will turn our attention to the analysis of brands spoofed in phishing attacks.

4. Spoofed Brand Analysis

This section analyzes the brands spoofed in phishing attacks – in other words, the companies that phishers are targeting. For confidentiality reasons, we will not reveal specific brand names, but will describe the brands in more general terms.

We used Norton Confidential server data collected from June through December 2006 in this analysis. From this data we were able to derive which brands are targeted, and for each brand how many phishing sites were set up to spoof it. We then manually went through each brand and assigned geographic locale as well as industry segmentation labels to them. We examined aggregate statistics like how many brands are spoofed, which of these seem to be core brands, and how these numbers change month-to-month. We also analyzed industry and geographic segmentation to better understand who phishers are targeting. As part of this effort, we also examined local brands – that is brands that are local to specific regions (e.g., credit unions) to see which regions are more heavily targeted by phishers.

AGGREGATE BRAND STATISTICS. During the analysis period, phishers spoofed a total of 343 brands. These brands exhibited Pareto-type properties in that a small number of brands accounts for a large number of actual phishing sites. For each brand, we determined the number of phishing web sites spoofing that brand (and sorted the brands according to this number); figure 4 shows the resulting graph (note that the y -axis is log scaled). It’s clear that the top few

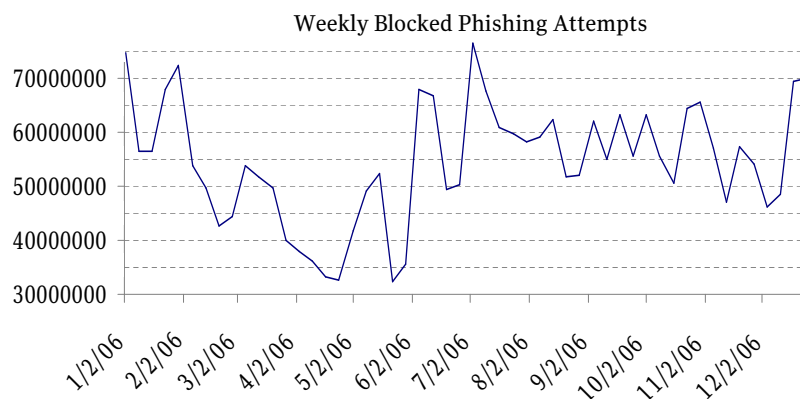


Figure 3: Weekly blocked phishing attempts in 2006

brands account for a large portion of phishing web sites. In fact, the top 10 spoofed brands account for nearly 85% of phishing web sites we encountered in the wild.

Of the 343 brands, only 57 are *core brands* that are spoofed each month. These core brands were determined by identifying seven lists of brands – one for each month in our data collection (Jun through Dec) in which a new web site spoofing that brand was reported. The core brands, then, comprised the intersection of these lists. While many core brands are among the most frequently spoofed brands, the list does not coincide. In particular, among the top 57 spoofed brands, only 47 are in the core. The highest overall non-core spoofed brand appeared 9th among the list of most frequently spoofed brands. The least frequently spoofed core brand ranked 112th out of 343 among the most frequently spoofed brands. For this core brand, only 12 phishing sites were set up to spoof it. One new site was reported in each of June, July, August, September, and November; three sites were reported in October and two sites were reported in December. These numbers suggest that phishers do not always take a scatter-shot approach in their attack attempts. Instead, for specific targets, they prefer methodical smaller-scaled approaches, albeit at a consistent pace.

Next, we looked at how many brands are targeted in a given month and how brand composition changes month-to-month (in other words, the *turnover* in brand choice). Table 2 summarizes the results. Table 2 can be interpreted as follows. In June, we saw 128 distinct brands being spoofed in phishing attacks. Going into July, phishers stopped spoofing 36 of these brands and launched phishing sites spoofing 58 different brands. Consequently, in July, there were 150 distinct brands spoofed in phishing attacks ($150 = 128 - 36 + 58$). These numbers illustrate considerable turnover in spoofed brands. Approximately 27% to 35% of the brands were dropped from the previous month and the number of new brands comprised approximately 30% to 38% of that month’s total.

INDUSTRY SEGMENTATION. We now consider the different industry segments associated with spoofed brands. We identified seven website categories:

- **Financial:** sites associated with online banking, brokerage, credit card companies, loans, insurance, and similar financial services (or sites that directly support such a brand);

	DROPPED	NEW	TOTAL
JUN	-	-	128
JUL	36	58	150
AUG	53	45	142
SEP	50	41	133
OCT	37	49	145
NOV	41	64	168
DEC	53	64	179

Table 2: Month-to-month brand turnover

SECTOR	BRANDS	%SITES
FINANCIAL	287	63.5549
COMMERCE	23	33.6305
COMMUNITY	26	2.4444
GOVERNMENT	4	0.3355
JOB BOARD	1	0.0289
CERT. AUTHORITY	1	0.0042
GREETING CARD	1	0.0016

Table 3: Industry breakdown of spoofed brands and phishing web sites.

- **Commerce:** sites that are associated with the sale of merchandise online;
- **Community:** sites that provide common Internet-related services including one or more of the following: Internet access, email accounts, social networking or information portals;
- **Government:** sites whose common URL ends in .gov (followed, if applicable, by a country code);
- **Job Board:** sites that provide an exchange medium where employers make job postings and prospective job seekers reply to them;
- **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;
- **Greeting Card:** sites that offer free (or paid) online greeting cards.

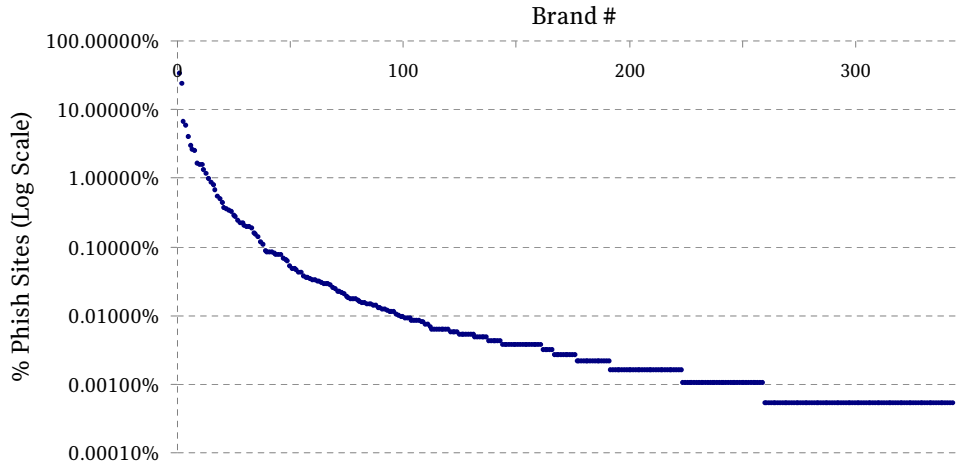


Figure 4: Pareto properties of spoofed brands: sites spoofing the leftmost brands form the bulk of phish sites.

See table 3 for a summary of the industry breakdown. As one might expect from table 3, phishers target financial brands the most since they offer the clearest money-making opportunities; in fact, eight of the top ten spoofed brands are in the financial category. They account for 287 brands (83% of brands spoofed). However, financial brands are only spoofed in roughly 64% of phish sites. Commerce is the next most popular category; phishers most likely spoof such sites since they also offer a clear path to monetization. While the remaining categories account for nearly 10% of brands spoofed, they account for less than 3% of phishing sites. This trend indicates that many phishers invest their efforts in small-scale attacks on many brands.

We also examined the question of whether retail-oriented brands are spoofed more during the holiday shopping season. We manually identified those brands in the commerce sector that are US-based and might generally see an increase during this time period. Of these brands, we found the number of sites that spoofed the top 6 (most frequently phished) from Nov 18, 2006 through Dec 31, 2006 (44 days). Next, we computed the number of sites we would have *expected* to see spoofing these brands if attacks were spread out uniformly across the 214 days in our data set (i.e., we multiplied the number of spoof sites for each brand by $\frac{44}{214}$). We computed the relative gain (or loss) of the actual number compared to the expected one. The results were 33.6%, 65.3%, -10.8%, 224%, 386.4%. In all but one case there was a substantive increase. Brands that are not spoofed often will see substantial changes in part to there being fewer sites spoofing these brands – meaning that any extra site contributes more to the percentage. The sixth brand on our list was *only* spoofed during the holiday shopping season. Still, it’s actually not clear that phishers are heavily targeting retailers primarily because of the holidays. In particular, we repeated the above exercise across all brands (not just retail) and found an 88.6% increase in phishing sites over what we might expect. So, while there is an increase in retail phishing, it’s not clear that the increase is due to the shopping season or is an artifact of the overall increase in phishing activity.

GEOGRAPHIC SEGMENTATION. We segmented the spoofed brands according to geographic location. We went through

the 343 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; two are headquartered in the United Kingdom; one is headquartered in Germany; and one is headquartered in Australia. 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 be misleading to think of them as being exclusively associated with the United States. Table 4 summarizes our results.

Even though less than 72% of spoofed brands are US-centric, over 76% of spoofed sites are. At the other end of the spectrum, the UK is associated with under 6% of the brands but over 13% of the sites. Table 4 also illustrates that phishing attacks spoof geographically widespread brands across numerous languages. In total, brands associated with 31 geographic regions were spoofed in a phishing attack from June through December. The corresponding web sites are written in 16 languages: English, German, Spanish, Italian, Danish, Arabic, Portuguese, Chinese, Japanese, Greek, French, Malay, Korean, Dutch, Hungarian, and Turkish.

LOCAL BRANDS. Among all spoofed brands from the Jun through Dec 2006, we examined the distribution of those that represent *local* banks in the United States; for example, credit unions that are local to a specific US state or region. We considered two classes of local banks. The first class, which we refer to as **MULTI-STATE** was limited to those that have operations in more than five US states. Of the 343 brands in our collection, 122 satisfied this condition. The second class, which we refer to as **SINGLE-STATE** was further restricted to banks that operate in only one US state; 89 of the 122 **MULTI-STATE** banks satisfied this condition. First, note that local US brands already account for nearly 36% ($= \frac{122}{343}$) of *all brands* spoofed under the period of consideration. This measure supports the hypothesis that phishers are going after smaller institutions in more targeted attacks. Going further, this measure seems to indicate that phishers have access to localized email databases that allow

REGION	BRANDS	%SITES
UNIT. STATES	246	76.085324%
UNIT. KINGDOM	19	13.205546%
GERMANY	11	4.117644%
AUSTRALIA	8	3.285505%
CANADA	8	2.250713%
SPAIN	7	0.320255%
WEST INDIES	1	0.218404%
MEXICO	2	0.158028%
IRELAND	2	0.133352%
ITALY	6	0.088201%
SOUTH AFRICA	1	0.035701%
DENMARK	1	0.033076%
CURACAO	1	0.017850%
ISLE OF MAN	1	0.012075%
UNIT. ARAB EMIRATES	2	0.009975%
BRAZIL	2	0.005250%
CHINA	5	0.003675%
JAPAN	3	0.003150%
NEW ZEALAND	1	0.003150%
GREECE	3	0.002625%
FRANCE	2	0.002100%
MALAYSIA	1	0.002100%
KOREA	2	0.001050%
BELGIUM	1	0.001050%
PORTUGAL	1	0.001050%
CENTRAL AMERICA	1	0.000525%
HUNGARY	1	0.000525%
JORDAN	1	0.000525%
NETHERLANDS	1	0.000525%
PHILIPPINES	1	0.000525%
TURKEY	1	0.000525%

Table 4: Geographic focus of spoofed brands and corresponding phish sites.

MULTI-STATE			
STATE	%SITES	STATE	BRANDS
IL	12.09%	FL	12
CA	12.02%	CA	10
NY	11.95%	IL	10
WI	11.17%	TX	10
FL	10.18%	MI	9
NV	8.77%	WI	8
NJ	8.70%	NY	7
MI	8.56%	WA	7
PA	7.99%	CO	6
GA	7.71%	IN	6

Table 5: The 10 US states with the most spoofed brands and phishing activity, according to local banks that serve at most five states including the ranked state (MULTI-STATE case).

SINGLE-STATE			
STATE	%SITES	STATE	BRANDS
PA	11.27%	FL	8
WA	10.15%	TX	8
FL	9.46%	MI	6
WY	8.07%	CA	5
OH	7.79%	NY	5
MI	6.82%	WI	5
CA	4.87%	WA	4
GA	4.73%	CO	3
TX	4.59%	MA	3
CO	3.76%	OR	3

Table 6: The 10 US states with the most spoofed brands and phishing activity, according to local banks that only serve the ranked state (SINGLE-STATE)

them to target specific populations.

For the first class (MULTI-STATE) of local banks, table 5 provides the top 10 US states in terms of both the percentage of phishing sites and number of brands local to that state (a version of table 5 with results for all 50 states is available in the full version of the paper, but is omitted because of space restrictions). Interpreting table 5, of all phishing sites for banks operating in at most five US states, 12.09% of them spoofed a local brand that does business in Illinois. Note that when totaled over all 50 states, the percentages exceed 100% since a phish site is counted multiple times (once for each state the corresponding spoofed brand does business in). Table 6 provides results for the SINGLE-STATE class of local banks.

The only state in the top 5 on all counts is Florida. Florida has the highest per-capita elderly population in the United States [7]. This population segment has been targeted in numerous types of offline fraud activity [6]. According to the 2000 US Census, the top nine states ordered by population age 65 or over are California, Florida, New York, Texas, Pennsylvania, Ohio, Illinois, Michigan, and New Jersey. These 9 states comprise half of the 18 states that appear in the tables 5 and 6. To more rigorously analyze a correlation between local phishing targets and other characteristics of US states, we downloaded US 2000 Census data for the following metrics: overall population (per state) in general;

	MULTI-STATE		SINGLE-STATE	
	SITES	BRANDS	SITES	BRANDS
POPULATION	0.68	0.76	0.49	0.69
ELDERLY	0.71	0.79	0.55	0.71
INCOME	0.27	0.23	0.11	0.21
COUNTIES	0.46	0.53	0.33	0.52

Table 7: Table of correlations between phishing activity and other state characteristics

population (per state) older than 65 years, and mean per capita income (per state).

This census data was the most recent we could find as of the time our research was conducted. We then computed correlation coefficients between the census data and the percentage of phishing sites and spoofed brands for both the MULTI-STATE and SINGLE-STATE scenarios. The results are summarized in Table 7. Note that across all categories, the strongest correlation is with the elderly population, exceeding the correlation with regard to overall population and mean per-capita income. We did not expect the mean per-capita income correlation to be so low, as one might expect phishers to target states whose population has a higher income. One plausible explanation is that mean per-capita income alone is not always indicative of the size of the affluent population, especially if most individuals in the state have incomes centered close to the mean. From the 2005 US Census, we obtained data on the top 100 affluent counties in the United States. For each state on our list, we computed the number of these counties located in that state, and computed the correlation coefficients between this metric and our phishing data. Table 7 contains the results. Here we notice a stronger correlation to phishing activity focused on local brands; however the correlation is still not as strong as what we associate with the elderly.

Another population segment traditionally targeted by scam artists are students. We found some evidence that phishers might be targeting this segment as well. In particular, many credit unions and local banks exist primarily to serve a particular college or university. Among the 89 banks local to a particular state, we identified 8 brands (that constituted 12.5% of phish sites targeting the SINGLE-STATE category) specifically catering to local colleges or universities.

While this analysis provides some evidence to support a hypothesis that the elderly and student populations are being targeted by phishers, there may be other considerations. For example, a phisher might have obtained access to a specific email database local to a specific region (that would include email addresses for a bank’s customers or a university’s students) and could be leveraging that information. In such cases, the attack target might be based on more opportunistic considerations than a pre-meditated plan.

5. Conclusions

This paper examined and analyzed real-world phishing data from 2006. Our aim was to analyze more data more thoroughly, and with greater context than previous efforts. We presented several intriguing characteristics of phishing trends. These included fluctuations in activity on the weekends and during major events. Also, we studied fluctuations related to what brands are being spoofed. We determined

the industries, regions, languages, and population segments that appear to be targeted in these attacks. While not all the results are surprising, we know of no previous study that attempts to quantify them. In order to develop countermeasures to emerging threats like phishing, it is important to understand the nature of the problem by analyzing real-world data. This paper attempts to take a preliminary step in this direction and we hope to further these efforts with more substantial data sets in the future.

6. Acknowledgements

We would like to thank Joseph Blackbird, Dave Cole, Oliver Friedrichs, Marc Fossi, Jim Hoagland, Elias Levy, Dylan Morss, Sainarayan Nambiar, Prabhat Singh and Dean Turner for their help, either through illuminating discussions, providing and reasoning about raw data, or suggesting feedback on early drafts of this paper.

7. References

- [1] Symantec Internet Security Threat Report, Volume XI. *Symantec Corporation*, March 2007.
- [2] Anti-phishing Working Group, Phishing Activity Trends Report for December 2006. Available from http://www.antiphishing.org/reports/apwg_report_december_2006.pdf.
- [3] PhishTank, Phishing Activity statistics February 2007 <http://www.phishtank.com/stats/2007/02/>.
- [4] Zulfikar Ramzan. “Phishing Phone Numbers.” *Symantec Security Response Web Log*, 18-May-2006. http://www.symantec.com/enterprise/security_response/weblog/2006/05/phishing_phone_numbers.html.
- [5] Symantec Phish Report Network, <http://www.phishreport.net>.
- [6] Consumer Fraud Research Group. “Investor Fraud Study Final Report.” *NASD Investor Education Foundation*. http://www.nasdfoundation.org/WISE_Investor_Fraud_Study_Final_Report.pdf
- [7] U.S. Census Bureau. “Census 2000 Summary File 1 – United States.” Washington, D.C., 2001.