Phishing by Form: The Abuse of Form Sites

Hugo Gonzalez
Universidad Politecnica de San Luis Potosi
hugo.glez@gmail.com

Kara Nance
University of Alaska Fairbanks
klnance@alaska.edu

Jose Nazario
Arbor Networks
jose@arbor.net -mail

Abstract

The evolution of phishing methods has resulted in a plethora of new tools and techniques to coerce users into providing credentials, generally for nefarious purposes. This paper discusses the relatively recent emergence of an evolutionary phishing technique called phishing by form that relies on the abuse of online forms to elicit information from the target population. We evaluate a phishing corpus of emails and over a year's worth of phishing URLs to investigate the methodology, history, spread, origins, and life cycle as well as identifying directions for future research in this area. Our analysis finds that these hosted sites represent less than 1% of all phishing URLs, appear to have shorter active lifetimes, and focus mainly on email account credential theft. We also provide defensive recommendations for these free application sites and users.

1. Introduction

All Phishing provides a means to social engineer a target into voluntarily providing identity attributes or other private information. In most cases the intent is to use these attributes as a component of identity theft for other nefarious purposes. As users become increasingly aware of the phishing methods commonly employed, the effectiveness of the individual methods may decrease. In addition, data about phishing sites is now gathered by multiple organizations, including for-profit security companies and community sites [1]. These organizations work hard to shut down phishing sites, thus incentivizing the attacker to work harder to maintain a functional credential harvesting site. The result is that phishing approaches must continually evolve in order to entice users to surrender their credentials.

In the past several years we have seen the appearance and rise of phishing emails that use free sites on the web to capture credentials via forms. This abuse mirrors the appearance of these free form creation sites, often used to conduct surveys and similar data captures. Their use in phishing enables the attacker to have a stable credential capture site. One downside that attackers face is that the sites do not enable themselves to be "skinned" to appear to be arbitrary websites very well. This limits the number of potential target institutions that can be impersonated by the phishing site.

The outline of the remainder of this paper is as follows. In Section 2 we discuss the background of the phishing problem and introduce the use of these form sites in particular. In Section 3 we raise several questions about this change in tactics that we seek to answer in this paper. In Section 4 we discuss how these phishing campaigns operate and how we are investigating them. In Section 5 we present our results and analysis, and conclude in Section 6 with some directions for future research..

2. Background

Phishing can be an efficient means to obtain information voluntarily from an individual. It is trivial to send a large number of identical non-personalized messages to a target population and wait for a response. The response percentage does not need to be very high for the venture to be viewed as successful. While many early email phishing attempts, especially those evolving from the Nigerian 419 approaches, were easy to detect by many user groups (largely due to spelling errors, grammar issues, etc.), the evolution of the methods have resulted in messages that are more challenging to detect as phishing techniques.

As characteristics associated with phishing are identified and user populations are educated, phishing techniques continue to evolve to remain effective. Images replace text in messages to evade spam filters. Spelling and professionalism in the messages coupled

with the use of trusted authority names and logos further confuse users. Many now use the combination of authority and immediacy to obtain the desired response from the unsuspecting target. Authority generally comes in the form of masquerading as a trusted entity with whom the user has a relationship. It could be a financial institution, system administrator, service provider or other relationship that a user might feel is critical to maintaining life style choices. The second attribute of the messages is immediacy. They generally call for an immediate response with an associated penalty (such as cancellation of service, financial loss, etc.) or with a reward for an immediate response. The immediacy of the required response, the apparent authority of the request, and the perceived value associated with the service motivate users to respond before they have thoroughly thought through the situation. This is what makes many current phishing techniques successful.

Niche phishing techniques have evolved leading to the invention of new terminologies including spear phishing, whaling, "smishing" (phishing over SMS), in-session phishing, spy phishing, "vishing" (phishing over voice calls) and a plethora of other terms. Recently, the authors have observed a shift towards the use of web form services and online office suites as a means of harvesting credentials. These collection sites are typically free, with most not requiring any validation of the user's account to collect responses. Many of these websites can be themed to add the appearance of a legitimate site..

3. Problem Statement

As new approaches to phishing are identified, important questions about the techniques merit investigation. These questions can help guide approaches to mitigating these threats. Some solutions may involve technology, such as filtering [2, 3], blacklisting [4], or whitelisting [5]. Other approaches may use education and awareness to alleviate the threat [6]. While the success rates of various approaches are widely varied, they do result in the evolution of phishing techniques. As we observe and investigate new phishing threats, we start by asking the following questions:

1. What is the methodology?
2. When was this first observed?
3. How widespread does it appear to be?
4. How many distinct groups appear to be using these methods?
5. What is the apparent lifetime of these websites?
6. How successful is this method? (How many victims fall for it?)
7. What kind of data is being collected?

While the definitive answers to all of the questions are not easy to obtain, we can analyze the captured messages to begin to answer some of the questions and contribute to the body of knowledge. The most important is the methodology or modus operandi (MO), as that defines the phishing method and distinguishes it from other approaches..

4. Methodology

The lure emails, which motivate the user to visit the phishing site and provide their credentials, follow the typical phish lure methods of immediacy, generally scaring the user into action to prevent a loss of account access. For this type of phishing, where the information is collected on a form site, one common lure theme is an email account that is over the user's quota. The message states that if the user does not re-validate their credentials, they will lose access. An example of such a lure message is shown in Figure 1.

The landing pages, where the user is directed upon reacting to a phishing lure, are generally constructed to provide a simple form to harvest the user's account credentials: email address as needed, login name, and password. In some cases these are altered to appear less like a simple form tool and more like a legitimate response from the actual site the user believes they are validating against, thus increasing the authoritative posture of the message. One such screen is shown in Figure 2, demonstrating a themed Google Spreadsheet phishing site for the email lure shown in Figure 1.

To analyze the questions posed in Section 3 as applied to this new phishing by form methodology, we used two data sources. The first is an update of the PhishingCorpus developed by one of the authors [7]. The PhishingCorpus is an collection of UNIX mbox-format files of phishing messages that have been hand-screened. Messages include various kinds of phishing targets and approaches, and give insights into the types of material being sought. They are also useful to estimate how many different organizations are actively phishing using these methods, assuming changes in their lure methods between groups. For this study we analyzed a PhishingCorpus of 2240 messages from 7 August 2007, to 26 July 2011, covering nearly 4 years. In this corpus we found 82 phishing emails (lures, with links) that lead to these free form sites to capture a user's credentials.

The second data set we used is the PhishTank archive of URLs [8], specifically from July 2010 until August 2011. PhishTank is a community phishing alert web service where users can submit suspected phishing websites, confirm the phish site, and track the site lifetimes. These URLs were harvested every hour and stored and can be used to estimate phishing site lifetimes. Approximately 151,500 confirmed phish URLs were collected in this time period.

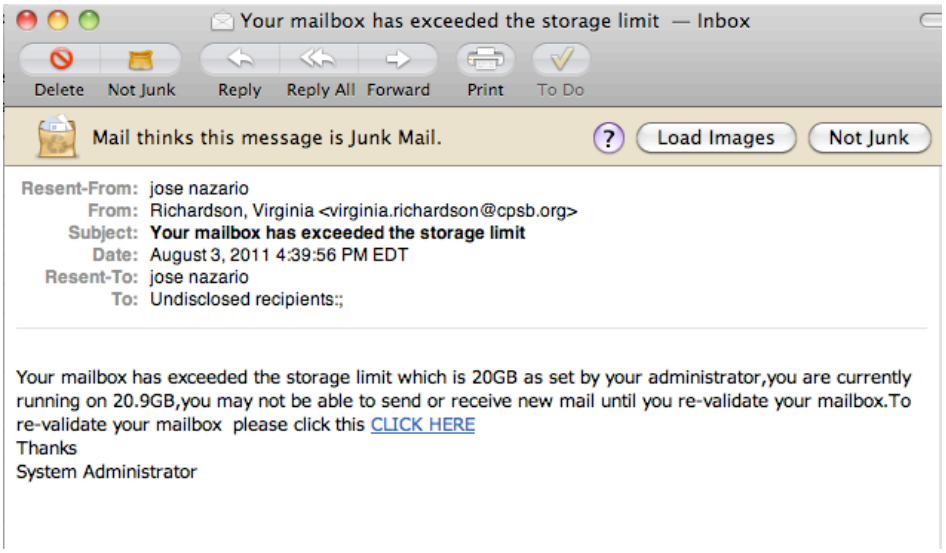

Figure 1: An email lure for a "mailbox over quota" theme. The link in this example goes to a Google Spreadsheets page which collects the information.

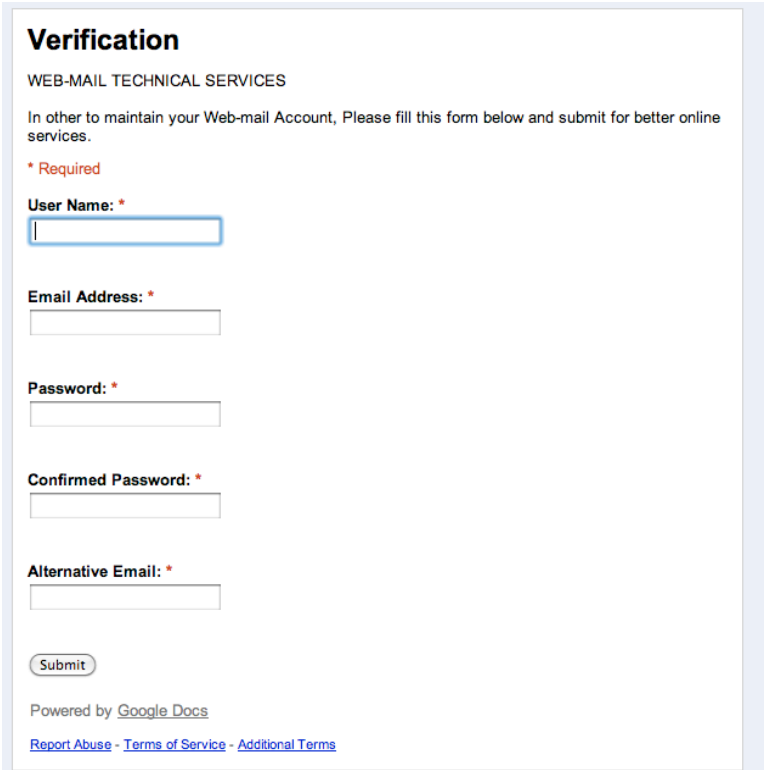

Figure 2: Google Spreadsheets landing page for an email account phish. The attacker in this case has done some basic work to theme the page to not look like a typical spreadsheet page.

5. Results

While the general phishing characteristics of authority and immediacy were described above and can be applied to most current phishing techniques, observations that support the application of these characteristics in the phishing by forms messages were found. The subject lines, and their frequencies as observed in the PhishingCorpus, are shown below in Table 1:

Table 1: Subject lines and their frequencies in phishing emails (lures) that link to free form sites. Capitalization has been normalized to lowercase.

Frequency	Subject Line (Normalized)
24	system administrator
22	your mailbox has exceeded the storage limit
4	your mailbox has exceeded
3	webmail account verification update ! ! !
3	quota/limit 23gb
3	quota limit
3	mail abuse <28373772>
3	mail abuse +44813185462
3	helpdesk
3	fwd: quota/limit 23gb
3	dear account user,
3	=?windows-1256?q?keeping_track_of_your_usage=fe?=
2	webmail technical support unit
1	web mail help desk.
1	account upgrade and deletion announcement
1	=?iso-8859-1?q?copyright_=a9_2011_webmail?=

The most frequently occurring subject lines used immediacy, authority, or a combination of the above to entice the user to action.

To answer the second question regarding when this was first observed, we used the PhishingCorpus described above to examine when the first form site abuse occurred in phishing. The first recorded instance we see was on Monday, December 14, 2009 at approximately 21:00. This message was a spam email that told the reader that the associated email inbox had exceeded the storage quota of 20GB, and that he must re-validate by authenticating or risk losing email access. The URL was hosted on the "eformit.com" site, a free form hosting site abused with some regularity as shown in Table 2 (below).

Since then, a number of new sites have been abused in a similar way. The most prominent such site being abused is the Google Spreadsheets website, part of Google Documents. The first Google Spreadsheet phish messaged appears in the PhishingCorpus on Thursday, March 10, 2011, at approximately 17:00. Since then we have seen many more Google Spreadsheet links from phishing emails that follow the same pattern and form, typically informing users that their email accounts are to be deactivated unless they respond. According to analysis of the messages in the PhishingCorpus, this form-based phishing has become increasingly popular since this time, with much of the growth coming in 2011. We observed more URLs using these methods to capture credentials in this timeframe.

The third question concerns how widespread this evolutionary technique is at this time. It appears that phishing sites that use free form websites are relatively scarce at this point. From the PhishTank corpus of URLs above, we found that 209 matched the pattern (see the domains in Table 2), for a frequency rate of 0.13% across all of the phishing URLs we tracked in this time period. When we examine the PhishingCorpus messages, we see that the frequency of these phishing emails is approximate 3.7%, much higher than in the phishing URLs seen in PhishTank. Inspection of the URLs over time reveals that these sites are being abused across the board, as well. The attackers who use these sites are not moving from one site to the next, suggesting that they are not encountering any screening of their activities on these free form websites that would block their use in phishing.

Questions 4 and 5 concern the origins and lifetime of the sites. Seventeen sites hosting forms abused in phishing were identified by manual inspection. The URL distribution across these domain names is uneven, as shown in Table 2. Also shown in Table 2 are the lifetimes of the URLs seen on these free form websites. The average lifetime of these sites across all domains in Table 2 is 15 days 22 hours. By comparison, the lifetime of all phishing sites in the time period captured by PhishTank range from less than an hour to over 500 days, with an average of 13 days 11 hours.

6. Analysis and Discussion

The above data and analysis enables us to assess how effective these phishing attacks may be that are hosted on these sites, and the risks these sites may face in the future.

Two of the most important considerations we suspect that phishers have for their harvesting site choices, we believe, are the possible longevity of a phishing website, and the security of their captured

credentials. While we cannot evaluate the utility of using a free form website service on the latter idea, the former appears to be worse for these phishing campaigns. When we look at phishing lifetimes for these form-based sites, we see that the average lifetime of all phishing sites is longer, on average, than this subset under study. We believe, therefore, the main considerations for phishers using these form sites are ease of construction and server availability and not harvest site lifetime.

Table 2: Free form sites abused for phishing attacks analyzed in this study and the number of unique phishing URLs seen per-domain from March 2010 to July 2011.

Domain	URLs	Minimum lifetime	Maximum lifetime	Average lifetime
addaform.com	51	4 days 17:19:17	292 days 19:18:35	133 days 19:05:42
spreadsheets.google.com	41	<1:00	420 days 15:22:35	6 days 23:25:12
formbuddy.com	36	1 day 21:10:32	248 days 02:31:45	52 days 12:27:12
123contactform.com	23	<1:00	190 days 07:41:02	21 days 15:09:57
eformit.com	17	3 days 04:22:40	130 days 07:45:47	19 days 01:04:34
formchamp.com	11	1 day 23:46:59	49 days 02:49:45	12 days 18:24:09
icebrrg.com	8	1 day 02:11:33	362 days 08:01:23	95 days 10:37:20
creator.zoho.com	6	2 days 03:32:48	13 days 00:24:08	4 days 19:25:22
.formstack.com	3	11:37:45	23 days 09:03:55	11 days 11:21:02
www.icebrrg.com	3	3 days 02:12:04	71 days 07:49:18	26 days 06:38:54
secureform.com	3	1 day 14:51:40	25 days 11:36:51	16 days 14:22:01
.tfaforms.com	2	1 day 09:08:09	6 days 02:08:15	3 days 17:38:12
forms3.createforms.com	1	242 days 18:31:35	242 days 18:31:35	242 days 18:31:35
forms5.createforms.com	1	82 days 02:24:53	82 days 02:24:53	82 days 02:24:53
create-form.com	1	5 days 14:33:24	5 days 14:33:24	5 days 14:33:24
formkid.com	1	30 days 12:17:56	30 days 12:17:56	30 days 12:17:56
sureforms.net	1	252 days 13:59:07	252 days 13:59:07	252 days 13:59:07

Secondary considerations may be the reputation of the domains used, which are typically benign and not purpose-built for phishing. We have witnessed this "trust riding" before where open redirectors from established web properties lead to a phishing site. In the case of these hosted form sites, this directly affects the difficulty of blacklisting the sites, as the domain itself cannot be blacklisted but instead the entire URL must be blacklisted. If this kind of phishing attack becomes more popular, it could put size pressure on phishing URL blacklist operators.

This method of executing the credential harvesting step in a phishing attack appears to have evolved to evade detection by automated means. Existing phishing detection algorithms, such as those described by Prakash *et al.* [9] and the large-scale system described by Whittaker *et al.* [10], would miss the phishing features in these sites. The features used in PhishNet include hostname similarity, directory structures present in the URL, and query string substitution to detect a phishing attack. However, while these phishing sites use query strings to load a specific form, they do not have arbitrary hostnames or directory structures available to the attacker. The Google phishing classification system uses similar features, including the hostname (or IP address presence), the directory structure, and page contents such as an IFRAME. However, some of the assumptions about the presence of a brand in the URL are absent in these types of phishing attacks. Furthermore, while these pages may contain JavaScript for dynamic behaviors (for example Google spreadsheet pages), this JavaScript code is shared with non-phishing pages. From a reputation standpoint, the infrastructure and components used by these phishing attacks are shared with typically benign content and so benefits from this scoring.

Detection of these phishing sites would require a re-evaluation of the algorithms to detect and qualify phishing attacks. Existing assumptions about how the user is fooled into trusting the site is their institution fail in this scenario. We find that page contents are a useful guide to detect phishing in these situations, coupled to analysis of the spammed URLs in the lure messages.

7. Future Work

Clearly one of the open questions from our study is "how many victims fall for these attacks?" Because we have not been working with these site operators, we do not have any of the collected data and cannot evaluate how successful these types of phishing campaigns are compared to other kinds of phishing. To answer that, we would need to work closely with the form site operators to inspect the results gathered from these forms.

The final open question that our data collection and analysis does not facilitate us to answer, is a definite accounting of the kinds of data being captured by the attackers. To do this, we would need to harvest a number of these sites while they are live and perform analysis on their fields. We can infer from the subject lines observed in the PhishingCorpus (see Table 1) that these attacks largely focus on collecting email account information. This may be by design, however. One major limitation that attackers face is that the form sites themselves do not enable the page to be arbitrarily themed with graphics and interactive code. This makes it difficult to reasonably impersonate a major business that is a traditional phishing target, for example a financial institution. This may limit the damage to the population of users, but for affected users the takeover of their email account may be a stepping-stone to larger attacks including identity theft..

8. Recommendations

For end users at risk of having their credentials harvested, anti-phishing mechanisms such as browser plugins to check URLs and safe browsing/data sharing guidelines can be helpful, but may be limited. The browser plugins are fed by anti-phishing data sources. If these data sources are not looking for the kinds of attacks being perpetrated, then the associated URLs will remain undetected and the end user will remain unprotected. While detection limitations may be due to limited distribution of the phishing emails, akin to a "spear phishing" campaign, or missing checks for the semantic analyzers in the phishing URL discovery tools, this remains a potential issue. Users are thus best protected by defending themselves, through education, safe browsing habits not sharing confidential information to untrusted, unverified sites that may not actually be what they claim to represent.

For the general hosting sites and services, we see room for improvement with added controls. Based upon the phishing URL lifetime analysis shown in Table 2, it appears that controls for form content may be lacking, either both automated and manual screening. Without further experimentation to determine how well abuse reports work or the difficulties encountered in setting up a credential harvesting page on one of these sites, we cannot say what approaches will work best to combat this abuse. This is an obvious next step in this research, as these sites are proliferating and the risk of abuse is rising as well.

Still, generic security "best practices" recommendations for any site offering free services can be made. Clearly these form sites should monitor for signs of abuse, be prepared to respond to these incidents quickly, and limit the off-loading of captured credentials by the attacker. It would be beneficial for these form sites to consume external data data feeds from abuse monitoring sites such as PhishTank and others, as well as their own internal abuse report handling with a clearly visible link for visitors to report such attacks..

9. Acknowledgments

This work was facilitated by the Honeynet Project (www.honeynet.org), a leading international security research organization, dedicated to investigating the latest attacks and developing open source security tools to improve Internet security.

10. References

[1] Christian Ludl, Sean Mcallister, Engin Kirda, and Christopher Kruegel. 2007. On the Effectiveness of Techniques to Detect Phishing Sites. In Proceedings of the 4th international conference on Detection of Intrusions and Malware, and Vulnerability Assessment (DIMVA '07), Bernhard Hä;mmerli and Robin Sommer (Eds.). Springer-Verlag, Berlin, Heidelberg, 20-39. DOI=10.1007/978-3-540-73614-1_2 http://dx.doi.org/10.1007/978-3-540-73614-1_2

[2] André Bergholz, Jan De Beer, Sebastian Glahn, Marie-Francine Moens, Gerhard Paab;, and Siehyun Strobel. 2010. New filtering approaches for phishing email. J. Computer Security. 18, 1 (January 2010), 7-35.

[3] M. Dolores Del Castillo, Angel Iglesias, and J. Ignacio Serrano. 2007. An integrated approach to filtering phishing e-mails. In Proceedings of the 11th international conference on Computer aided systems theory (EUROCAST'07), Roberto Moreno Diaz, Franz Pichler, and Alexis Quesada Arencibi (Eds.). Springer-Verlag, Berlin, Heidelberg, 321-328.

[4] Mohsen Sharifi and Seyed Hossein Siadati. 2008. A phishing sites blacklist generator. In Proceedings of the 2008 IEEE/ACS International Conference on Computer Systems and Applications (AICCSA '08). IEEE Computer Society, Washington, DC, USA, 840-843. DOI=10.1109/AICCSA.2008.4493625 http://dx.doi.org/10.1109/AICCSA.2008.4493625

[5] Ye Cao, Weili Han, and Yueran Le. 2008. Anti-phishing based on automated individual white-list. In Proceedings of the 4th ACM workshop on Digital identity management (DIM '08). ACM, New York, NY, USA, 51-60. DOI=10.1145/1456424.1456434 http://doi.acm.org/10.1145/1456424.1456434

[6] Ponnurangam Kumaraguru, Yong Rhee, Steve Sheng, Sharique Hasan, Alessandro Acquisti, Lorrie Faith Cranor, and Jason Hong. 2007. Getting users to pay attention to anti-phishing education: evaluation of retention and transfer. In Proceedings of the anti-phishing working groups 2nd annual eCrime researchers summit (eCrime '07). ACM, New York, NY, USA, 70-81. DOI=10.1145/1299015.1299022 http://doi.acm.org/10.1145/1299015.1299022

[7] PhishingCorpus, 2011, Jose Nazario. Available at http://monkey.org/~jose/wiki/doku.php?id=PhishingCorpus.

[8] PhishTank, 2011. Available online at http://www.phishtank.com/.

[9] Prakash, P. and Kumar, M. and Kompella, R.R. and Gupta, M., 2010. Phishnet: predictive blacklisting to detect phishing attacks. In Proeedings of IEEE INFOCOM 2010.

[10] Whittaker, C. and Ryner, B. and Nazif, M., 2010. Large-scale automatic classification of phishing pages. In Proceedings of the 17th NDSS Conference.