June 2012

IMPERVA®

# Hacker Intelligence Initiative, Monthly Trend Report #11

## A CAPTCHA in the Rye

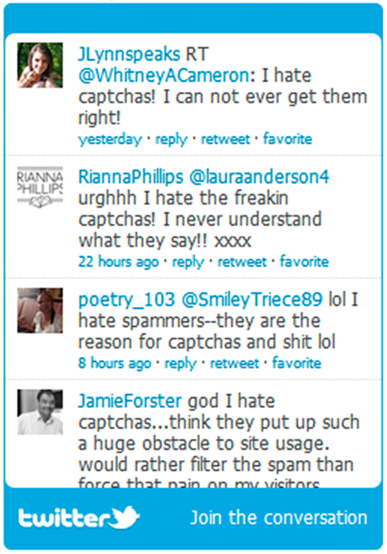
***ADC Monthly Web Attacks Analysis, June 2012***

## 1. Overview

*In J.D. Salinger's The Catcher in the Rye, main character Holden Caulfield is a very confused teenager. One way he deals with his problem? He views virtually anyone he meets or knows as a "phony." This attitude, among other factors, contributes to Holden's dysfunctional nature.*

*A CAPTCHA, or Completely Automated Public Turing test to tell Computers and Humans Apart, is a common security measures used to distinguish between humans and a "phony." Ideally, a CAPTCHA should distinguish human users from automated browsing applications, preventing automated tools from abusing online services. Hackers have deployed numerous methods to bypass CAPTCHAs, such as outsourcing readers to India or creating CAPTCHA games that reward users with pornographic images. Like Holden, the line between real and phony isn't clear, forcing security professionals to present CAPTCHAs sub optimally. The result can be summarized in this Twitter feed:*

twitter speaks out!



JLynnspeaks RT @WhitneyACameron: I hate captchas! I can not ever get them right!
yesterday · reply · retweet · favorite

RiannaPhillips @lauraanderson4 urghhh I hate the freakin captchas! I never understand what they say!! xxxx
22 hours ago · reply · retweet · favorite

poetry_103 @SmileyTriece89 lol I hate spammers--they are the reason for captchas and shit lol
8 hours ago · reply · retweet · favorite

JamieForster god I hate captchas...think they put up such a huge obstacle to site usage. would rather filter the spam than force that pain on my visitors

*But there is good news. Several innovative CAPTCHA methods have appeared recently – including game-based CAPTCHAs. In this report, we review the use of CAPTCHAs as a security mechanism against malicious automation. We report and analyze four case studies and provide recommendations on ways to implement CAPTCHAs as an integrated part of a security strategy. Specifically, security teams should:*

- *Use novel CAPTCHA methods that make the CAPTCHA into something enjoyable, like a mini-game.*
- *Minimize the number of CAPTCHA challenges that legitimate users encounter. The idea is to present a CAPTCHA only when users exhibit suspicious behavior. To detect such, the site should use the other automation detection mechanisms.*

## 2. What is a CAPTCHA

Due to the pervasiveness of bots, Web applications must try to identify when interacting with an actual human user from an automated tool. Automated tools can be used for many malicious purposes, like scraping, spamming, and application-level DoS attacks. More specifically, attackers may use automated tools to post comments in blogs and forums, create fake accounts, retrieve mailing lists, and advertise products.

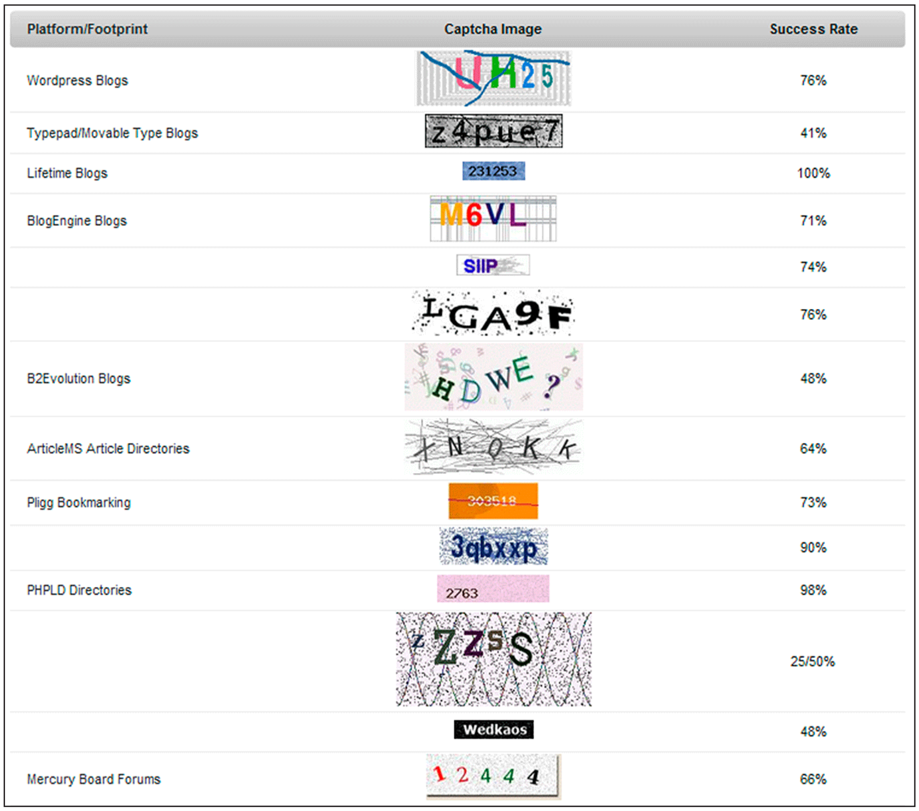
The CAPTCHA (an acronym of "Completely Automated Public Turing test to tell Computers and Humans Apart") is a common security measure used today against automated attacks. In general, a CAPTCHA is a test intended to distinguish human users from automated browsing applications, and thus prevent automated tools from abusing online services. CAPTCHAs do so by asking users to perform a task that is quick and simple for humans and, ideally, impossible for automated software. For example, CAPTCHA implementations assume it's easy for a human and difficult for a computer to recognize textual content in a noisy image or, likewise, recognize spoken words in a noisy audio recording. Thus, hackers who wish to use automated tools for their malicious purposes must break (solve) or bypass the CAPTCHA mechanism in order to be permitted into the site.

### 2.1 CAPTCHA Solving Mechanisms

Given that the CAPTCHA is implemented correctly and cannot be bypassed in the code level, there are two main approaches to solving massive amounts of CAPTCHAs. The first is by using automated tools that can break the CAPTCHA, like Optical Character Recognition (OCR) and Machine Learning technologies. The other involves outsourcing the CAPTCHA to a third-party human solver. Each approach has its benefits, limitations, and costs.

#### 2.1.1 Computer-based tools

A fair amount of effort is put into developing pattern recognition algorithms that could identify distorted text just as well as do humans. Such algorithms exist today, with varying success rates, and keep improving constantly. Yahoo's EZ-Gimpy CAPTCHA was broken as early as 2003, and other CAPTCHAs used by popular sites were soon to follow. PWNtcha, for example, can break CAPTCHAs of 12 different vendors, 9 of them with more than 90% success. A more contemporary tool, CAPTCHA Sniper, can solve CAPTCHAs by 50 vendors, with a success rate between 27-100% depending on the vendor.[1] CAPTCHA Sniper is intended to integrate with automated comment spamming tools like ScrapeBox, so the goal of solving large amounts of CAPTCHAs is quite clear in this case.



| Platform/Footprint | Captcha Image | Success Rate |
|---|---|---|
| Wordpress Blogs | UH25 | 76% |
| Typepad/Movable Type Blogs | z4pue7 | 41% |
| Lifetime Blogs | 231253 | 100% |
| BlogEngine Blogs | M6VL | 71% |
|  | SIIP | 74% |
|  | LGA9F | 76% |
| B2Evolution Blogs | HDWE? | 48% |
| ArticleMS Article Directories | XNQKK | 64% |
| Pligg Bookmarking | 303518 | 73% |
|  | 3qbxxp | 90% |
| PHPLD Directories | 2763 | 98% |
|  | zZZsS | 25/50% |
|  | Wedkaos | 48% |
| Mercury Board Forums | 12444 | 66% |

*Figure 1: A partial list of the success ratios of the CAPTCHA Sniper tool, for different CAPTCHA services.*

[1] See here: http://caca.zoy.org/wiki/PWNtcha and here: http://www.CAPTCHAsniper.com/

Due to the high success ratios, CAPTCHA providers must constantly monitor and improve their CAPTCHAs to avoid breaking. They do this, for example, by adding more distortions to the background or twisting the characters further.

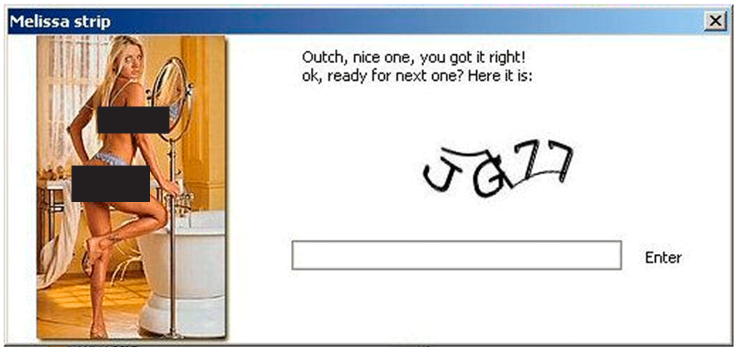
In this arms race, the solving mechanisms keep getting better while the CAPTCHAs keep getting tougher, often becoming too difficult even for humans. Users might find too difficult CAPTCHAs annoying, so site managers need to maintain a delicate balance between the site's security and the user experience.

### 2.1.2 Human-based outsourced CAPTCHA solving

While some attackers put their efforts into developing technologies to break CAPTCHAs, others came up with a creative way around that problem: why invest a great deal of money and resources in developing complex algorithms when across the globe there is an infinite pool of individuals who would solve CAPTCHAs for you? This way is a lot cheaper and doesn't require any complicated technological skills; a spammer can simply buy a service from a company that has hundreds of CAPTCHA-solving employees.

Many online services today outsource work to freelance workers from around the world, especially from developing countries where Internet is accessible and people are willing to work cheap. It is not uncommon today to encounter advertisements in job sites, offering freelance CAPTCHA solving or looking for such solvers. In this case, there's a direct connection between the service provider and the client. In other cases, there's a third party involved. Services like DeCaptcher recruit CAPTCHA solvers from around the world and offer CAPTCHA-solving services as a retailer. Having many employees allows 24-7 service guarantee while handling massive amounts of CAPTCHAs in very little time. At current rates, CAPTCHA solvers get $1-$3 dollars for solving thousands of CAPTCHAs, and are often rewarded (or penalized) according to their speed and achieved percent of accurate responses. All a spammer has to do is buy a package. One provider, "Bypass CAPTCHA" charges $14 for 1,000 CAPTCHAs. Another provider, "Death by CAPTCHA" charges $1.39 for 1,000 CAPTCHAs. Whatever the service, if the CAPTCHAs are broken, the doors are left wide open for a high volume of spam.

Human CAPTCHA solvers can also be rewarded in ways other than money. A clever way to use human abilities to solve CAPTCHAs is demonstrated by different sites that offer free porn as an incentive. Instead of paying for a subscription, the user browsing the site gets every now and then a pop-up containing a CAPTCHA, which he is required to solve in order to keep enjoying the site or be allowed to see more content. Another way is to implement the CAPTCHA is as an integrated part of a game, as was done in this example[2]:



Melissa strip

Outch, nice one, you got it right!
ok, ready for next one? Here it is:

This way, the participants don't even have to know they are helping a hacker or a spammer to overcome security measurements. If they don't know they are doing it, they don't expect to get paid for it…

Human-based CAPTCHA solving services pose a serious threat to Web security and challenge the whole concept of CAPTCHAs. They were originally intended to distinguish humans from computers, but now automated software is using actual humans to cheat the test and pass as humans. This challenges even the most innovative efforts to create better CAPTCHAs, some of which will be described in section 2.3.

[2] Picture originally published here: http://www.mobileinc.co.uk/2010/10/spammers-get-innovative-and-use-porn-game-to-beat-CAPTCHA/

## 3. Considerations in implementing CAPTCHAs

### 3.1 Home grown VS existing implementation

An important question that needs to be addressed is which CAPTCHA implementation mechanism to use in order to minimize the chances of it being broken. One approach is to create your own proprietary CAPTCHA, and hope that no one would take the effort to break it. In the case of spamming, attackers usually want to post their advertisement wherever and as much as they can. They target the entire Web equally, so, if a site happens to be a tough nut to crack, they might just skip it. Creating your own home-grown CAPTCHA is a costly and time-consuming process, with the risk of ending up with an inefficient solution. Furthermore, regardless of how strong the CAPTCHA implementation may be upon introduction to the world, its effectiveness drops as automated tools evolve and become more sophisticated. This risk demands close monitoring of the breaking tools and adaptation of the CAPTCHA to keep it unsolvable for automated tools.

The other approach to integrating CAPTCHA in a Web application is to use an existing, widely used service, like *reCAPTCHA*.[3] Such a service has a battery of supporting developers who keep improving and adapting it. The service is not necessarily stronger than any home-brewed CAPTCHA, but enjoys the benefit of its being widely spread: when a CAPTCHA service provider identifies it is being bypassed in one Web application, it can update at once all the Websites using the same service. Thus, a site enjoys protection against new tools before it is even attacked by them. The downside here is that the prevalence of the CAPTCHA service in many sites might also make it a target for strenuous cracking attempts. Once it's broken, all the sites using it are exposed.

### 3.2 Visual CAPTCHA vs. Audio CAPTCHA

The vast majority of sites today implement a visual CAPTCHA, in which a string of letters is presented over a noisy background. The characters can be distorted, in different colors, sizes, orientations or fonts, to make the work of separating them from the background and from each other more difficult.

Similarly, an audio CAPTCHA contains a recording of voiced words, letters, or numbers in a noisy background. Here, too, the user is required to decipher the text from the noise.

One of the reasons to use an audio CAPTCHA in addition to the more common, visual CAPTCHA, is to make the Internet more accessible to people suffering from visual impairment. This is specifically important in the case of governmental sites, providing every-day services otherwise inaccessible. One of the few government agencies in Brazil that implements an audio CAPTCHA was chosen as the target site in an experiment designed to check blind and partially blind users' ability to navigate through the site and complete certain tasks.[4] Sadly, most of the participants couldn't understand and complete the CAPTCHA correctly without assistance – either they couldn't understand the voiced numbers or exceeded the timeout permitted by the site. In addition, another study found that audio CAPTCHAs were more difficult and time consuming for non-blind users than visual CAPTCHAs.[5] As audio CAPTCHAs are generally easier for breaking by machine learning algorithms[6], this raises the question whether their implementation is worthwhile – while they expose the site to greater risk at the expense of usability.

A study[7] that analyzed the security strength of audio CAPTCHAs from three popular sites using machine-learning algorithms managed to achieve up to 71% correct solutions. With that in mind, it is important to note that the current human pass rate for some audio CAPTCHAs is also around 70%. Thus, this task hardly differentiates a human from a machine. The authors suggest several parameters that can strengthen an audio CAPTCHA against automated breaking attempts:

[3] http://www.google.com/recaptcha

[4] http://www2.dbd.puc-rio.br/pergamum/tesesabertas/0711270_09_cap_04.pdf

[5] Bigham, J. P. and Cavender, A. C. Evaluating Existing Audio CAPTCHAs and an Interface Optimized for Non-Visual Use. In Proc. of the SIGCHI Conf. on Human Factors in Computing Systems (CHI '09), Boston, Massachusetts, 2009

[6] Elie Bursztein, Romain Beauxis, Hristo Peskov, Daniele Perito, Celine Fabry. "The Failure of Noise-Based Non-Continuous Audio CAPTCHAs". IEEE Symposium on Security and Policy, pp.19-31, 22, May, 2011.

[7] Tam, J., Simsa, V., Hyde, S., and Von Ahn, L. 2009. Breaking audio CAPTCHAs. Adv. Neu. Inform. Process. Syst. 21, 1625–1632
Available here: http://www.CAPTCHA.net/Breaking_Audio_CAPTCHAs.pdf

1. **Vocabulary size**: The larger the vocabulary, the harder it is to learn.
2. **Speaker voices**: Using multiple voices makes learning more difficult.
3. **Background noise**: The background noise should be very similar to the voiced text, e.g., distorted human speech, as random or artificial noise has very different characteristics than human voice. Similarity between the message and the background makes the stage of separating the words from the background more difficult.

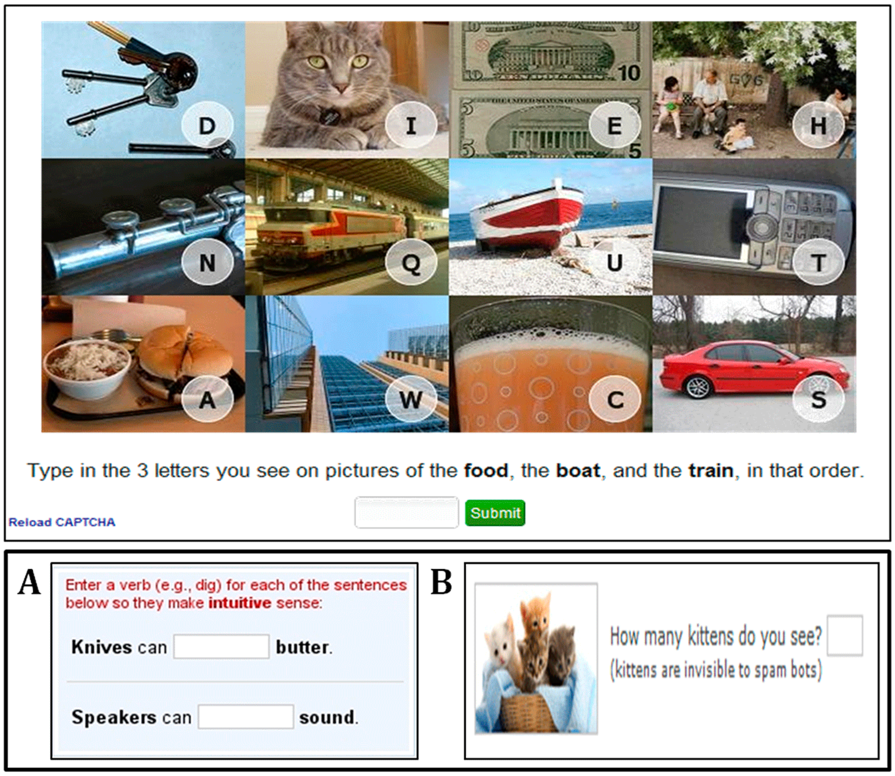
### 3.3 Novel Approaches to CAPTCHAs

One problem that might occur when CAPTCHAs become more and more sophisticated is that they might become too difficult to use even for the average human. Users might find such CAPTCHAs annoying and avoid sites that use them. For this reason, site administrators need to maintain a delicate balance between the site's security and the user experience.

One approach to minimize this nuisance is to prompt CAPTCHA challenges only after suspicious activity was recognized, or to adjust the difficulty of the CAPTCHA according to the attributes of each user's observed behavior. For example, the nuCAPTCHA service gives legitimate users easy CAPTCHAs, and gives suspicious users progressively more difficult CAPTCHAs. This maximizes usability for legitimate users while providing high security against automated access.[8] Services like Mollom scan user-generated content for suspicious amounts of links and other spam indicators, and prompt CAPTCHA only to suspicious users. This way, most bots, but only 4% of human users, are bothered with solving the challenge.[9]

Another approach is to treat CAPTCHAs as something that can be fun and interesting for the user, like a game, a riddle, or even advertisement content. The additional advantage of such CAPTCHAs is that they are also supposedly better at distinguishing humans from bots. One example is the "Semantic CAPTCHAs" that rely on higher human cognitive functions that are hard to implement in software. These may require complex image processing, understanding of context, or solving riddles. For example, CAPTCHAs by Confident CAPTCHA[10] (see Figure 2) require the user to identify a word, interpret its meaning, and select the correct picture. The pictures vary in light, shape, angle, and color, so that the word "boat" in different trials may be connected to many different pictures of boats. Better yet, words like "food" refer to broader concepts that can be portrayed by many different pictures. "Food" can just as well be a hamburger, a banana, or a salad, which all have very different characteristics as pictures. These are all very easy for humans to understand and categorize but not so for a computer. Another type of CAPTCHAs asks the user to rotate a picture until it has a "normal" orientation. Others use tasks that don't require any image processing skills, but understanding of context- like sentence completion tasks. CAPTCHAs that rely on language and context understanding may also raise the bar for some human solving services, as they require a certain level of proficiency in the language. Assuming that the majority of human solvers are from third-world countries, tasks involving language understanding may be more difficult in comparison to mere character recognition. This was nicely demonstrated in a study that tested the geolocation of solvers by sending CAPTCHAs in which the solution was the sequence of numerals corresponding to the words that indicate these numerals in a certain language (e.g., the solution for "one, two, three" was "123"). For this challenge, the average accuracies of the tested services dropped significantly from around 90% to 39.9% correct responses, thus showing that semantic CAPTCHAs not only pose a challenge for automated solving tools, but also to human ones.[11]

[8] http://www.nucaptcha.com/features/security-features
[9] http://mollom.com/files/mollom-technical-whitepaper.pdf
[10] http://www.confidenttechnologies.com/demos/CAPTCHA-demo
[11] http://static.usenix.org/events/sec10/tech/full_papers/Motoyama.pdf



Type in the 3 letters you see on pictures of the **food**, the **boat**, and the **train**, in that order.

Enter a verb (e.g., dig) for each of the sentences below so they make **intuitive** sense:

**Knives** can ______ **butter**.

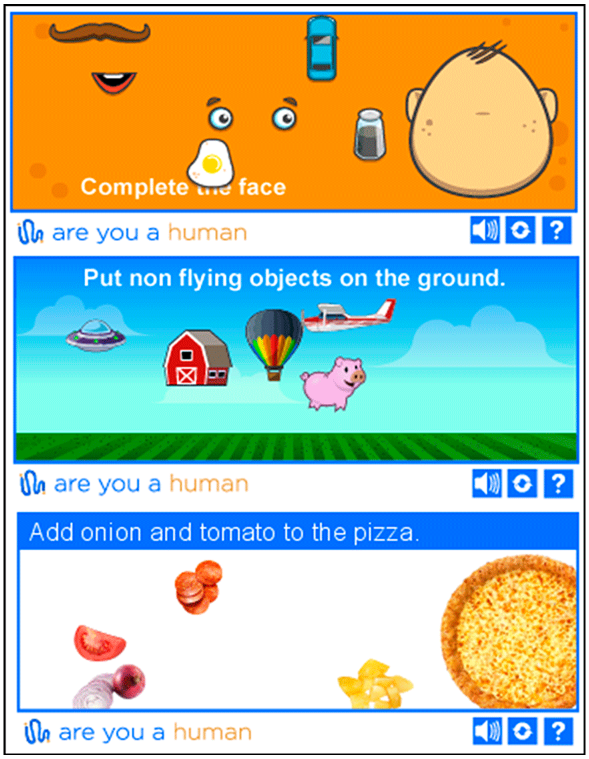
**Speakers** can ______ **sound**.

How many kittens do you see? □

(kittens are invisible to spam bots)

*Figure 2: Examples of novel CAPTCHAs. Top banner: CAPTCHA solution by Confident CAPTCHA. Following the instructions is very easy for a human user but not trivial at all for an automated tool. Bottom banner (A) Egglue semantic CAPTCHA, which relies on understanding of context and completing sentences. Bottom banner (B) another example of CAPTCHAs requiring both image processing and understanding of instructions.*

Another trend is using animated CAPTCHAs. The animations can look like a regular textual CAPTCHA, only with the characters and/or the background constantly moving, rotating, and changing colors. Others take the animation one step further to create an interactive platform, where the CAPTCHA is in itself a game.



Complete the face

Put non flying objects on the ground.

Add onion and tomato to the pizza.

*Figure 3: Examples where the user needs to perform an interactive task that requires both understanding of context and image processing. The first challenge is to understand the instructions, then to correctly identify the moving objects, and finally put them in the right place, not a trivial task for automated software.*

These CAPTCHAs are still relatively rarely used, and the area is still predominated by the textual CAPTCHAs. The novel alternatives are expected to become more abundant with time.

All the efforts to create more sophisticated CAPTCHAs target the abilities in which humans still perform significantly better than computers. However, these efforts are helpless against malicious automation in cases where the CAPTCHA is sent to a human solver. Even a test that can perfectly distinguish human from bot (a perfect Turing test) can be bypassed by another human, holding the door open for the bot. For this reason, even the best CAPTCHA should be taken with caution and preferably be used with other security measures.

## 4. Case Studies

In this section we describe case studies of CAPTCHA bypassing in the wild. The targets were all government Websites, providing different bureaucratic services.

### 4.1 Automation Identifiers

The next table summarized the automation characteristics suggested by Imperva in our Automation HII report, as were observed in each of the case studies.[12] It can be seen that while in all of these attacks the attacker used an innocent User-Agent, the requests lacked important browser-like headers, like Accept Language and Accept Charset. The absence of such headers normally indicates that the source of the requests is not a browser but some other tool interacting with the Web application. In addition, the requests rate in each of the cases is significantly higher than is expected from a human user.

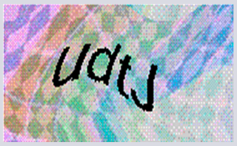
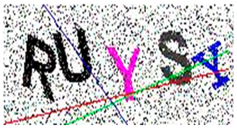
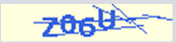
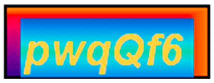
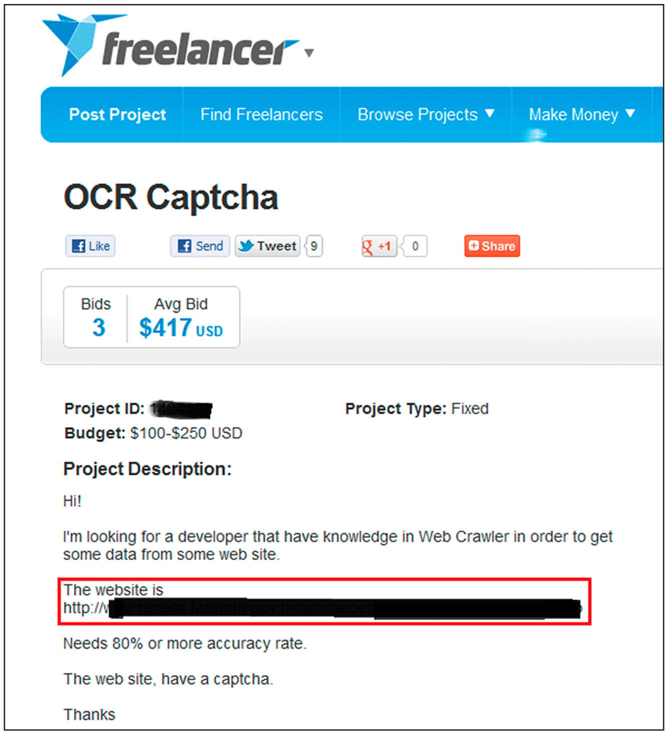
| Website | Automation Characteristics | | | CAPTCHA |
|---|---|---|---|---|
| | User-Agent | Accept headers in request | Rate (requests per minute) | |
| Government Agency #1 | Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.1.8) Gecko/20071022 Ubuntu/7.10 (gutsy) Firefox/2.0.0.8 | Accept Encoding | 20 |  |
| Government Agency #2 | Mozilla/5.0 (Windows NT 6.1; WOW64; rv:6.0.2) Gecko/20100101 Firefox/6.0.2 | Accept | 60 |  |
| Government Agency #3 | Mozilla/5.0 (X11; U; Linux i686; en-US; rv:1.8.1.8) Gecko/20071022 Ubuntu/7.10 (gutsy) Firefox/2.0.0.8 | Accept Encoding | 17 |  |
| A Bank | None | Accept | 13 |  |

*Table 1: Automation characteristics of the observed cases.*

### 4.2 Government Agency #1

This Brazilian government agency administrates taxes and provides tax-related certificates. Besides providing a platform for tax payment, the website is used to check if a person's CPF (the Brazilian social security number) is up to date, and all the taxes properly paid. Documents from this agency are required as a pre-condition to many other services in Brazil, like issuing a passport or getting a loan. The same information is available for companies, whose unique identifier is the CNPJ. As the only way to check CNPJ or CPF status is online, many companies that need to use this database for their customers (e.g. loan companies, banks, accounting offices, and law firms) need to go online and do so **manually**. To make the process more efficient, these companies purchase different software that can bypass the CAPTCHA verification and run multiple queries at a time (see Figure 4 for an actual ad looking for such software). The target agency is aware of this and implements other security measures to fight those automated accesses, like a rate limit, source IP limit, and some other limitations to avoid a denial of service and automated access.

[12] http://www.imperva.com/docs/HII_Automation_of_Attacks.pdf



OCR Captcha

| Bids | Avg Bid |
|---|---|
| 3 | $417 USD |

**Project ID:** [redacted] **Project Type:** Fixed
**Budget:** $100-$250 USD

**Project Description:**

Hi!

I'm looking for a developer that have knowledge in Web Crawler in order to get some data from some web site.

The website is
http://[redacted]

Needs 80% or more accuracy rate.

The web site, have a captcha.

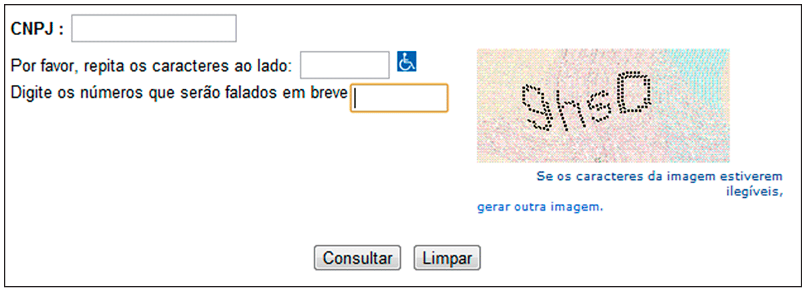
Thanks

*Figure 4: An ad looking for a freelancer who is willing to develop a CAPTCHA bypassing crawler, specifically for this Agency's site.*

### 4.2.1 CAPTCHA characteristics

The Website implements two kinds of CAPTCHAs, audio and visual.

In the audio CAPTCHA, two male voices speak simultaneously, and the user has to type only the numbers that are spoken. The visual CAPTCHA has a wide vocabulary and uses different backgrounds, fonts, and texture. For example, in the CAPTCHA in Figure 5, the letters are made of separated dots, which are not trivial to cluster together to a letter shape by an algorithm.



CNPJ :

Por favor, repita os caracteres ao lado:

Digite os números que serão falados em breve

Se os caracteres da imagem estiverem ilegíveis,
gerar outra imagem.

*Figure 5: The CAPTCHA interface provided by Agency #1.*

### 4.2.2 Attack description

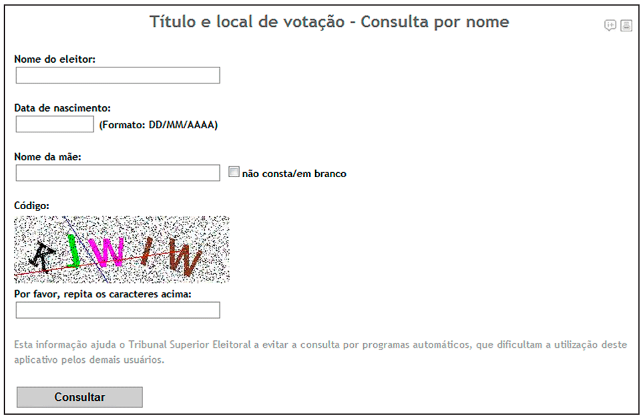
We've seen many high-rate requests for the CAPTCHA URL of this Website. This activity is ongoing and has lasted for at least three months. We see equal amounts of requests for the audio and visual CAPTCHAs.

As can be seen from Table 1, the attacker used a common browser-like user agent. Despite that, the requests lacked proper Accept headers and were sent in a very high rate. Taken together, these parameters suggest that the continuous requests for the CAPTCHA challenges are the result of an automated tool. It is plausible to assume that the tool has some success in breaking the CAPTCHA and getting the desired content. This is not surprising after seeing the ad in Figure 4, looking specifically for such a tool.

### 4.3 Government Agency #2

This agency is in charge of the voting process in Brazil. They issue a kind of "Voting ID" with a citizen's name, mother's name, and voting place. They also issue certificates required in various bureaucratic procedures.

#### 4.3.1 CAPTCHA Characteristics



Título e local de votação - Consulta por nome

Nome do eleitor:

Data de nascimento:
(Formato: DD/MM/AAAA)

Nome da mãe:
□ não consta/em branco

Código:

Por favor, repita os caracteres acima:

Esta informação ajuda o Tribunal Superior Eleitoral a evitar a consulta por programas automáticos, que dificultam a utilização deste aplicativo pelos demais usuários.

*Figure 6: a form from Government Agency #2 website, used to get voter's information*

The CAPTCHA used here uses different letter colors, different orientation of background lines, and changing letter alignment. On the other hand, the letters are always in upper case and there are always five of them, which allows only a limited vocabulary.

#### 4.3.2 Attack description

We witnessed around 6,000 requests to this website. About 40% of the requests were to different query pages at the site, which present the user with registration forms. The forms include a CAPTCHA and thus redirect the user to a specific CAPTCHA URL.

At the observed attack, the user first approached one of the three forms and was given a session ID. After being redirected to the CAPTCHA image, the attacker asked for the CAPTCHA image repeatedly. Every few requests for the image, the attacker managed to solve one and then sent an HTTP POST request with the form parameters: the voter's name, mother's name, and date of birth. As the values used in the forms included real names and dates, the attacker had to use some predefined list or get these data somewhere.

Analyzing the requests according to URL revealed that there were at least three times more requests for the CAPTCHA URL than to any other URL at the site. This is definitely not the pattern of normal use, as a human user would ask for the form, fill it with the CAPTCHA, and go on browsing the site or leave. Here we see repeating requests for the CAPTCHA URL, which implies that the user had a specific interest in the CAPTCHA images or was not capable of solving them. Either scenario, coupled with the high rate of requests, strongly suggests an automated attack.

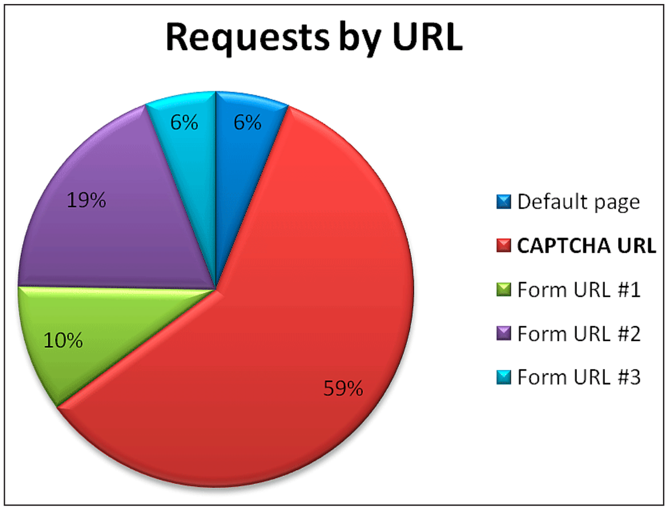


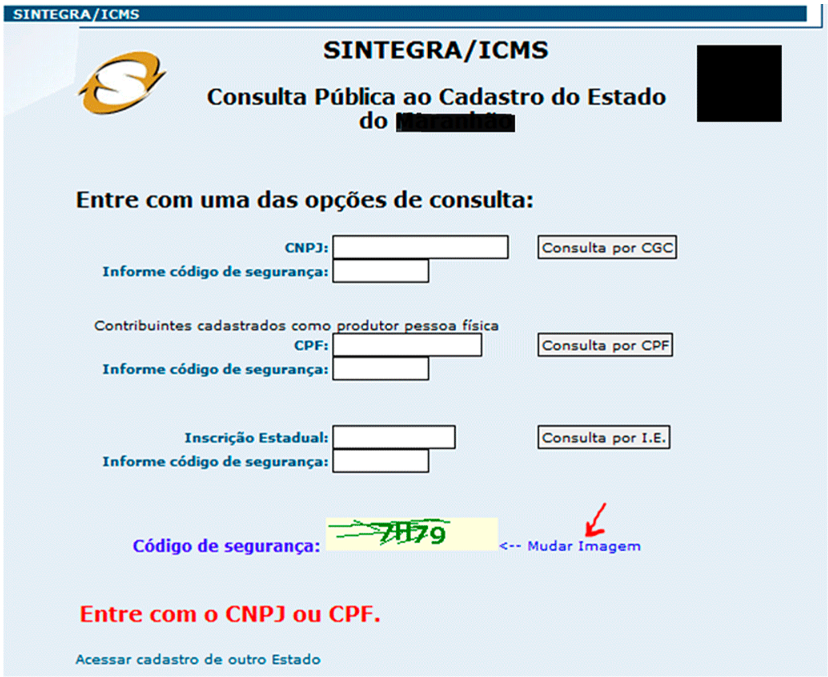
*Figure 7: The requested URLs in Government Agency #2.*

### 4.4 Government Agency #3

The target host was the Ministry of Finance of one of the Brazilian states. The site is used mainly for online tax payment.

#### 4.4.1 CAPTCHA characteristics

In all samples, all of the letters were the exact same color, which was also identical to the background lines. Although this may make the separation of letters from background (and from each other) more difficult, the constant font, letter size, and alignment might make the automated CAPTCHA-breaker's work relatively easy. Also, the CAPTCHA length is relatively short – only four characters.



SINTEGRA/ICMS

**SINTEGRA/ICMS**

**Consulta Pública ao Cadastro do Estado do [redacted]**

[redacted]

**Entre com uma das opções de consulta:**

**CNPJ:** ______ [Consulta por CGC]
**Informe código de segurança:** ______

Contribuintes cadastrados como produtor pessoa física
**CPF:** ______ [Consulta por CPF]
**Informe código de segurança:** ______

**Inscrição Estadual:** ______ [Consulta por I.E.]
**Informe código de segurança:** ______

**Código de segurança:** <-- Mudar Imagem

**Entre com o CNPJ ou CPF.**

Acessar cadastro de outro Estado

*Figure 8: A form from Agency #3 website with a CAPTCHA*

### 4.4.2 Attack Description

The access to the site began with several requests for URLs that contain forms in which the user can enter either CNPJ or CPF (see previous section regarding Government Agency #1 for more information about these identifiers) to get information about his tax status. Some of these requests were POST requests and included the query number and the mentioned identifiers as HTTP parameters. After that, we've seen periods of high-rate access to the CAPTCHA URL, which is equivalent to pressing the "change image" button in the form (see red arrow in Figure 8). So it seems that the attack tool "pressed" this button to gather more and more CAPTCHA images, perhaps to create an offline dictionary for future use.

## 4.5 A Bank

This is the official site of one of the central banks in Argentina. Among other services, the site offers a system for issuing a status report of taxes, debts, loans, and credit payments.

### 4.5.1 CAPTCHA characteristics

This is probably the weakest CAPTCHA of the bunch. It seems like the CAPTCHA-generator randomly chooses two backgrounds (template color and orientation), the characters, and the font color. In our sample, the combinations seemed random, so that if the font color is by chance very similar to the background, the CAPTCHA would be difficult, and otherwise it would be extremely easy for any OCR tool to solve.

### 4.5.2 Attack Description

The requests appear in repetitions of three: request for the form, request for a CAPTCHA challenge, and a POST with the user's details. These three stages were repeated in a rate of about 13 requests per minute.

It seems that, like the previous cases, the activity we see may also be an automation of a legitimate process. Financial institutions often check potential customers' status in the bank, and, to do so, might use automated processes.

## 5. Summary and Conclusions

- **CAPTCHA is not a silver bullet against automation**. CAPTCHAs are a good, but imperfect, attempt to identify automated access to a web application, as opposed to normal human use. One of its inherent flaws today is that it can be easily bypassed by outsourcing it to human solvers for a very low cost. When the CAPTCHA is solved for the attacker by other humans, it doesn't matter how good it is at distinguishing humans from machines. Therefore, a CAPTCHA alone is not enough to guarantee the security and the content quality of the site.
- **CAPTCHA should be supplemented with other anti automation measures**. To improve the efficiency of existing CAPTCHA mechanisms they should be integrated with other automation detection measures.
  - **Traffic-based automation detection**. Examples of such are rate and HTTP headers based detection. These measures are discussed in depth in our previous report on Automation.
  - **Normal behavior analysis**. A site can create a profile of legitimate use, in terms of rate of access, requested URLs distribution, traffic volume, Geo-location (IP from Ukraine trying to post in a Japanese blog, etc.).
  - **Content analysis**. For example, counting the amount of links any blog comment contains, looking for irrelevant keywords (Viagra, Rolex, online-poker) or profanity.
  - **Blacklists and reputation**. Any traffic originating from a suspicious IP should be treated with caution, if not blocked.
- **CAPTCHA security should be balanced with positive user experience**. Although CAPTCHAs can be useful for the sites' security, they must be balanced with the importance of keeping a *positive user experience*. Users who face difficult or annoying CAPTCHAs will eventually leave the site. This can be avoided by:
  - Using novel CAPTCHA methods (see section 2.3) that make the CAPTCHA into something enjoyable, like a mini-game.
  - Minimizing the number of CAPTCHA challenges that legitimate users encounter. The idea is to present a CAPTCHA only when users exhibit suspicious behavior. To detect such, the site should use the other automation detection mechanism discussed previously.

## Hacker Intelligence Initiative Overview

The Imperva Hacker Intelligence Initiative goes inside the cyber-underground and provides analysis of the trending hacking techniques and interesting attack campaigns from the past month. A part of Imperva's Application Defense Center research arm, the Hacker Intelligence Initiative (HII), is focused on tracking the latest trends in attacks, Web application security and cyber-crime business models with the goal of improving security controls and risk management processes.

**Imperva**
3400 Bridge Parkway, Suite 200
Redwood City, CA 94065

Tel: +1-650-345-9000
Fax: +1-650-345-9004
www.imperva.com