Breaking ReCAPTCHA

Security, Privacy and Cryptography
9 December 2009
Michael Lissner

> If on the other hand your poll should have anything in it that is potentially lulzworthy in our sense of humor you are not safe.
>
> Even if you put in the measurements our good friend Ben gives you here, our strength lies in our numbers and our numbers are wast.
>
> - Anonymous [1, 2]

Two tasks that modern computers have not yet been able to do consistently and efficiently are optically recognizing images of characters (OCR) and differentiating between humans and computers. reCAPTCHA is a system that attempts to solve both of these problems. When used, the reCAPTCHA system presents two images consisting mostly of letters and/or numbers. One of the images contains a string the value of which the system knows, but the system does not know the value of the letters and numbers in the other image. Users are asked to type in the value of both images. If the user inputs the correct value for the known image, it is assumed that the user is a human, and it is assumed that their input for the second (unknown) image is correct as well.

At this point, the system has determined that the user is a human, but for it to achieve OCR, several additional users must confirm the value of the unknown image. Thus, they too are presented with two images – a different known image, and the same unknown image. If they respond correctly to the known image, again, it is assumed they

[1] A post on the reCAPTCHA mailing list made by the hacker group Anonymous in response to how to hack-proof the reCAPTCHA system.

[2] Anonymous, “Time.com Hack of reCAPTCHA - reCAPTCHA | Google Groups,” *reCAPTCHA*, April 29, 2009, http://groups.google.com/group/recaptcha/msg/0be81b0edfd6102d.

responded correctly for the unknown image. If their input for the unknown image is the same as the previous user's, the system assumes that the value that has been put in by these people is valid, and thus learns the value of the letters in the image (achieving OCR).

Of these two goals, the second is likely the more critical, as determining whether something is a human or computer is an important task for palliating the scourge that is spam. Unfortunately, though, there are many ways for spammers (and other attackers) to defeat this goal, which I describe in the remainder of this paper.

One recent attack that has been made on CAPTCHAs in general is to use client-side Javascript to parse the contents of an image, and input the value. A Greasemonkey script was recently made that automated this attack for several websites that utilized rather basic CAPTCHA systems.[3] A shortcoming of the system however is the rather simple approach it takes to decoding the contents of an image.

Where that system leaves off, the hackers at 4chan have picked up. In an effort to hack the Time.com site, they created a system that pulled reCAPTCHA images off the Time.com site, and then used OCR to analyze the images.[4] This attempt failed though because reCAPTCHA images are chosen for their ability to defeat OCR systems, and are then further marred before being displayed to users.[5] Some recent methods of identifying the parts of images that have been manipulated may be of use in defeating defeat this aspect of reCAPTCHA.[6] Using these systems, it may be possible to remove any additional marring that is added to the already challenging words, thus giving a hacker the same challenge that was originally posted to the OCR machines. By pulling hundreds or thousands of reCAPTCHA challenges form a site, removing any marring that has been

[3] John Resig, "John Resig - OCR and Neural Nets in JavaScript," Blog, *OCR and Neural Nets in JavaScript*, http://ejohn.org/blog/ocr-and-neural-nets-in-javascript/.

[4] Paul Lamere, "moot wins, Time Inc. loses « Music Machinery," *Music Machinery*, April 27, 2009, http://musicmachinery.com/2009/04/27/moot-wins-time-inc-loses/.

[5] "reCAPTCHA Security," *Security*, http://recaptcha.net/security.html.

[6] Neal Krawetz, "Body By Victoria - Secure Computing: Sec-C," *Body by Victoria*, November 2, 2009, http://www.hackerfactor.com/blog/index.php?/archives/322-Body-By-Victoria.html.

added to them, and then performing OCR on the images, it may be possible and cost-effective to defeat the system.

There are two additional layers of security that should be mentioned here. First, reCAPTCHA tracks IP addresses and blocks ones that appear to be illegitimate. Second, after a user fails a reCAPTCHA challenge, the system presents the user with two words for which it knows the value, thus doubling the difficulty of the challenge. Both of these problems could likely be defeated by IP spoofing (i.e. presenting a different IP to the system for each query that is made).

Another approach that could defeat the reCAPTCHA system is to attempt to poison it by providing the same value for all unknown texts several thousand times, thus teaching the system that a large percentage of the images in the system correspond to the same word. This attack was also attempted by the 4chan group, however the scale of the reCAPTCHA system has gotten so large that to successfully poison the system would take hundreds of thousands of entries per day, likely too many to make such an effort worthwhile.

An additional approach that has been proposed to defeat reCAPTCHAs is to attack their audio component, which is provided for users that are visually impaired. Like the visual challenges, the audio challenge presents the user with a challenge, and then requests that they type it in. Also like the visual challenges, the audio challenge uses humans to decode something that computers are believed to be incapable of decoding. In this case, old radio streams are used as a source corpus. One paper has described how this area may be a vulnerability, and explains how to analyze the contents of the audio stream, decode the sound, and then present the results.[7] This may be an economical approach to defeating reCAPTCHAs. Similar to the added visual marring that is used on the visual images, audio marring is added to the audio streams, and so this too would have to be removed before the stream could be analyzed properly.

[7] Jennifer Tam et al., "Breaking Audio CAPTCHAs," *Advances in Neural Information Processing Systems* 21 (2008).

There are two remaining approaches to defeating the reCAPTCHA system. The first is to create a man in the middle approach, which pulls the images from one site, presents them to a human in another site, and then fools the human into entering the values in the images. Coupled with an existing method of sending spam and socially engineering people, this could be a very economical approach to defeating reCAPTCHAs.

The final approach for defeating reCAPTCHAs is to simply not defeat them. Instead of using a computer and a complicated scheme to input the values into the system, simply use a human to input the value of the image. This approach was used successfully by the Anonymous hacking group to input approximately 200,000 votes into a Time.com poll, thus defeating the purpose of the reCAPTCHA system.[8] Coupled with cheap labor, and/or Amazon's Mechanical Turk system of "artificial artificial intelligence,"[9] this may in fact be the most efficient method of defeating the system, since humans are able to solve challenges at a rate of approximately 30/minute.[10]

[8] Lamere, "moot wins, Time Inc. loses « Music Machinery."

[9] Amazon.com, "Amazon Mechanical Turk - Welcome," *Mechanical Turk is a marketplace for work.*, https://www.mturk.com/mturk/welcome.

[10] Lamere, "moot wins, Time Inc. loses « Music Machinery."