Chapter XIII
CAPTCHAs: Differentiating between Human and Bots

Deapesh Misra
VeriSign iDefense Security Intelligence Services, USA

ABSTRACT

The Internet has established firm deep roots in our day to day life. It has brought many revolutionary changes in the way we do things. One important consequence has been the way it has replaced human to human contact. This has also presented us with a new issue which is the requirement for differentiating between real humans and automated programs on the Internet. Such automated programs are usually written with a malicious intent. CAPTCHAs play an important role in solving this problem by presenting users with tests which only humans can solve. This chapter looks into the need, the history, and the different kinds of CAPTCHAs that researchers have come up with to deal with the security implications of automated bots pretending to be humans. Various schemes are compared and contrasted with each other, the impact of CAPTCHAs on Internet users is discussed, and to conclude, the various possible attacks are discussed. The author hopes that the chapter will not only introduce this interesting field to the reader in its entirety, but also simulate thought on new schemes.

INTRODUCTION

Human interactive proofs (HIPs) are schemes which require some kind of interaction from a human user that is tough for a program to simulate. "Completely automated public Turing test to tell computers and humans apart" (CAPTCHAs) are a class of HIPs which are tests that are so designed that humans can easily pass them while automated programs have a very tough time in passing them. Thus, such tests try to prevent malicious automated programs from accessing Web services which are meant to be used by human users only.

Differences in the capabilities between humans and computer programs, which can be tested and evaluated over the Internet, are made use of to create a CAPTCHA. Generally, hard "artificial intelligence" (AI) problems are turned into

CAPTCHAs. Usually such tests utilize schemes which exploit the differences in the cognitive capabilities between humans and computers, for instance, exploiting the difference between humans and computer programs in understanding distorted text.

Necessity

As the Internet grows into our daily lives and removes human to human interaction by considerable leaps and bounds, the necessity to identify whether the entity on the other side of Internet is really a human being or an intelligent program has gained immense importance. Many e-commerce businesses which cater to such a growing population of human users on the Web have business models in which the primary assumption is that humans are the users of the service. Automated programs are increasingly able to perform many tasks on the Web just like a human user. In many cases, these automated bots are to be denied access to the service. In all such scenarios CAPTCHAs play the role of the guard which keeps the bots from accessing the services.

Some of the immediate scenarios wherein there is a necessity of segregating the human and the non-human user are as follows:

- Online polls
- Preventing spammers from getting free mail IDs
- Preventing chat bots from irritating people in chat rooms with advertisements
- Preventing automated dictionary attacks in password systems (Pinkas & Sander, 2002)
- Preventing unruly search engine bots from indexing sites
- Preventing unethical pricing practices in e-commerce
- Preventing inflating/deflating rankings in online recommender systems
- Preventing spam in blog comments
- Preventing game bots from playing online games
- Preventing DDoS attacks (Gligor, 2005)
- Preventing automated worm propagation (e.g., Santy Worm, Provos, McClain, & Wang, 2006)

While these were some of the current reasons for the deployment of CAPTCHAs, as e-commerce grows and as the Internet replaces human to human interaction, new scenarios requiring CAPTCHAs will emerge.

History

The earliest attempt and perhaps the longest continuing one, is a classic example of trying to fool the automated programs which try to harvest mail IDs on the Web. This is the custom of putting out mail IDs on the Web with the "@" symbol replaced by "at" and by other such variations. Some variants are:

- Mail_id(AT)mail_provider(DOT)com
- Mail_id@mail_providZr.nZt (Replace Z with E)

instead of mail_id@mail_provider.com. This practice called "address/mail munging" is still prevalent and has been able to withstand attacks from basic automated scripts which try to harvest mail IDs.

Moni Naor (Naor, 1996) and the researchers at Georgia Tech (Xu, Lipton, & Essa, 2000; Xu, Lipton, Essa, & Sung, 2001) were one of the earliest contributors to the field of CAPTCHAs. The earliest attempt of using a CAPTCHA on the Internet was by Altavista in 1997 and was to prevent Web-bots from abusing the free URL submission utility. This was a word based CAPTCHA in which the user had to recognize the distorted word. In 2000, Yahoo was in need of some mechanism to prevent bots from joining the chat rooms and directing the chat room users to advertisements.