Chapter XIII
Benchmarking Steganalysis

Andrew D. Ker
Oxford University Computing Laboratory, UK

ABSTRACT

This chapter discusses how to evaluate the effectiveness of steganalysis techniques. In the steganalysis literature, numerous different methods are used to measure detection accuracy, with different authors using incompatible benchmarks. Thus it is difficult to make a fair comparison of competing steganalysis methods. This chapter argues that some of the choices for steganalysis benchmarks are demonstrably poor, either in statistical foundation or by over-valuing irrelevant areas of the performance envelope. Good choices of benchmark are highlighted, and simple statistical techniques demonstrated for evaluating the significance of observed performance differences. It is hoped that this chapter will make practitioners and steganalysis researchers better able to evaluate the quality of steganography detection methods.

INTRODUCTION

Steganography is the study of the concealment of information. Typically this means embedding a covert payload in an item of digital media such as an image, audio file, or video, but steganographic methods have now been proposed for a wide range of cover objects. Successful steganography means that nobody other than the intended recipient can even detect the existence of the embedded payload, let alone decode it, in which case other information security measures are for naught.

This motivates the competing field of *steganalysis:* to determine whether digital media objects contain a covert payload or not. It seems that every steganographic embedding scheme is sooner or later followed by publication of steganalysis techniques for attacking it. It is then vital to evaluate the ability of these methods to meet their detection aims, and particularly to compare the efficacy of competing steganalysis algorithms. Unfortunately, although there is copious work proposing methods for steganalysis, the literature barely considers the question of how to measure

their reliability; we observe poor practice and inconsistent benchmarks. We shall argue that some of the currently used benchmarks are statistically flawed, while others over-value detectors with weak practical detection power. Better choices exist. Furthermore, in the case of comparison of two steganalysis methods, there is rarely a consideration of the statistical significance of an observed difference: this could lead to flawed conclusions.

The chapter is not concerned with the creation of better steganalysis methods, nor with detailed benchmarking of current steganalysis methods, but with ways to measure steganalysis performance. It is aimed at practitioners who need to evaluate a particular steganalysis method and researchers who want to compare a new steganalysis proposal with those in competing literature. We will clarify the precise aims of steganalysis, separating those methods that detect payload from those that attempt to measure it, and survey some commonly used steganalysis benchmarks, pointing out weaknesses in some of the popular choices. We then suggest some good choices for benchmarks, and give simple statistical techniques to decide whether the evidence in a particular batch of experiments is sufficient to conclude a significant improvement. Throughout the chapter we will illustrate the techniques by comparing two competing steganalysis algorithms.

BACKGROUND

The terminology of steganography and steganalysis is now settled: the covert payload is embedded into a *cover object* producing a *stego-object*. Details of the *stego-system* (the embedding and extraction methods) are not relevant to this chapter, but it is generally assumed that the sender and recipient share knowledge of an embedding key, and that the recipient does not have access to the original cover object. The communicating parties' enemy is the *steganalyst* (often referred to as a *Warden*) and this is the role we are taking in this work, assuming that we are given *steganalysis* methods which try to determine whether an object is an innocent cover or a payload-carrying stego-object. Usually, different embedding methods and cover media types are attacked by specific steganalysis methods.

The literature contains a vast array of steganalysis techniques, for a range of embedding methods in a variety of cover media: the folklore method of replacing least significant bits (LSBs) in digital images is attacked in increasingly sophisticated ways in Westfeld and Pfitzmann (1999); Fridrich, Goljan, and Du (2001); Dumitrescu, Wu, and Wang (2003); Fridrich and Goljan (2004); Lu, Luo, Tang, and Shen (2004); and Ker (2005b, 2007a); replacement of multiple bit planes is attacked in Yu, Tan, and Wang (2005) and Ker (2007b); an alternative LSB method that avoids the previous steganalysis, described in Sharp (2001), is attacked in Harmsen and Pearlman (2003); Ker (2005a); and Fridrich, Goljan, and Holotyak (2006); the steganography software OutGuess embeds in JPEG images (Provos, 2001) and is attacked in Fridrich, Goljan, and Hogea (2002b); another popular JPEG embedding method is known as F5 (Westfeld, 2001) and this is detected by methods including Fridrich, Goljan, and Hogea (2002a); Harmsen and Pearlman (2004); Fridrich (2004); Shi, Chen, and Chen (in press); and Pevný and Fridrich (2007). Steganalysis is also possible in domains other than digital images: simple additive-noise embedding in video is attacked in Budhia, Kundur, and Zourntos (2006), and a method for embedding in MP3 audio files (Petitcolas, 1998) is attacked in Westfeld (2002). These references are not exhaustive and some others, including those proposing steganographic embedding schemes and those giving methods for their steganalysis, can be found under Additional Reading at the end of this chapter.

There does exist a concept of perfect, undetectable, steganography (Cachin, 2004) but it is difficult to practice in real cover media. Some