Chapter 14
Life Cycle Pattern Study of Malicious Codes

June Wei
University of West Florida, USA

Randall C. Reid
University of West Florida, USA

Hongmei Zhang
University of South Carolina, USA

ABSTRACT

This chapter investigates the patterns of malicious code attacks based on monthly data of the top 10 virus shares from 1998 to 2005. Three parameters were identified for study, overall pattern of the attack, the number reentries into the top 10 most prevalent attacks, and the maximum percentage share. The dataset was validated by comparing it to an independent dataset that measured the same parameters for a subset of the period of the primary dataset. The effects of malicious code that started before or disappeared outside the collection period were found to not have a significant effect. A multivariate regression analysis showed that the number of entries and the maximum share had a strong relationship with the visible life span. Multivariate cluster analysis was conducted on the reentry parameters and yielded six virus clusters classifications. The high impact viruses, 43 of the 230, are identified and further grouped.

INTRODUCTION

Computer and information security breaches have been a serious threat to the information technology (IT) industry (McClure, 2001; Whitman, 2003). This threat has resulted in dramatic financial losses. Despite the continued efforts of government and industry toward the defense against malicious codes, both the number of attacks and the resulting financial losses continue to increase (Gordon, Loeb, & Richardson, 2006). The CIS/FBI survey estimated the amount of loss due to virus contamination was $15,691,460 (Gordon et al., 2006).

The term "virus" is often used generically to refer to viruses, worms and other forms of malicious code (malware). Sophos defines a virus as "a computer program that copies itself" (www.sophos.com). A virus requires a host program and will not infect a computer until the host program has been run. In this chapter, the generic term virus refers to virus code, worms, Trojans, and all other forms of malware.

While exact numbers are not available, it is estimated that there are over 100,000 viruses in existence in today's computer information systems. Sophos (www.sophos.com) reported that there were 15,907 new malware threats identified during 2005 alone. Virus attacks shows a strong positive correlation with its costs, and denial of service, unauthorized access, and net abuse (Li, Wei, Lai, & Koong, 2004).

Problems and Objectives

Kephardt and White (1991) proposed a theoretical model using an epidemiological model of infectious diseases to study computer viruses. Kephart, Chess and White (1993) defined the epidemiological approach as "characterizing viral invasions at the macro level – has led to some insights and tools that may help society to cope better with the threat (and which may aid the study of biological viruses, too)." This chapter uses a macro level analysis of the life cycle of viruses to help to develop an understanding of how they behave in the environment. An extended discussion of the epidemiological approach to virus analysis can be found in Serazzi and Zanero (2003). A qualitative understanding of the epidemiology of computer viruses has been developed (White, 1998) and a quantitative analysis of the evolving attacks patterns and exploits used by viruses (Coulthard & Vuori, 2002) has been done.

An early approach to defining a virus's life cycle is the Internet Worm Propagation Data Model (IWPDM) (Mcalerney, 1999) which defined the life cycle as having four phases, starting with an activation phase and ending with a death phase. Between these two phases are a series of hibernation and reactivation phases. These would comprise the visible portion of the virus's life cycle. The complete life cycle would include a development phase prior to the activation phase and an epilogue phase where the author(s) are, hopefully, apprehended, tried in criminal court and possibly incarcerated. The development and epilogue phase are not considered part of the visible portion of the virus life cycle due to incomplete and inaccurate data and the lack or relevance to the understanding of the behavior of the virus while it is in the wild. The emergence/disappearance rates and the duration of the active attack period of viruses are not well documented. This emergence, disappearance, and the duration of active infection comprise the visible portion of virus life cycle.

A large number of viruses are developed in a laboratory setting as a proof of concept or as a test of a possible exploit. These are never released into environment so there is no need to directly control them and they often have a very abbreviated and hidden life cycle. In many cases, a patch is developed and released to remove the exploit prior to any attacks taking place. The focus of this chapter is only on those viruses in the general computing environment or "in the wild." The reasons for limiting the focus to those viruses "in the wild" is the abbreviated and often hidden life cycle of the virus while it is in the laboratory. Additionally, it is desirable from an analytical perspective to identify those viruses that passed a real life test as to its virulence and persistence. The ability to compare viruses with varying levels of virulence and persistence should aid in the identification of the characteristics that were the source of its virulence and persistence. Once the mechanism of the attack is understood, hopefully a defense can be developed.

The viruses have been the leading attack sources for the duration of this study and are predicted to remain as the leading attack sources