D39 Smart Unpacking Research: Using Mathematics to Unpack More

Benjamin Long, BA, National Institute of Standards and Technology,*
100 Bureau Drive, Stop 8970, Gaithersburg, MD 20899

After attending this presentation, attendees will learn new methods, developed by author, to elicit as well as validate structure and content using mathematically-based techniques. Attendees will not only learn the details of the methods, but also their development and use to date for accomplishing the objectives of the NSRL – to provide more file- identifying information and to validate the accuracy and completeness with which that information has been extracted.

This presentation will impact the forensic community by introducing new methods for analyzing and thinking about the issues of content analysis, data extraction, and measurement of these operations.

This work presupposes that digital content can be characterized and classified according to mathematical properties and structures. How this can form the basis for new kinds of analyses as well as a foundation for validation and measurement of the structures discovered will be discussed.

The idea of unpacking content from within another structure is a very general notion. It encapsulates a large portion of the activity in computer forensics to date.

The National Software Reference Library (NSRL) Project was formed to reduce the workload of investigators as they sought to separate what files constituted evidence of user activities and what files did not. The NSRL provides databases of file-identifying information (FII) for the purposes of reducing the number files that must be investigated, among other things. A large portion of the work involved in providing this data is performed by unpackers – tools and methods utilized to extract embedded files from compound files such as archives, compressed files, and so forth. Extracting such files increases the amount of file-specific information that may be provided.

The rate of appearance of new compound structures often exceeds that of corresponding unpacking methods. This is largely due to the fact that most unpacking is performed by pre-written tools that understand these structures and can extract their contents for processing. This problem also reflects a more general problem in forensics: accurate comprehension and/or extraction of embedded content in a timely manner. This problem is often addressed in a largely manual, time-consuming manner by those who create unpackers. In addition to the time lag introduced by these methods, there are few, if any, methods for validating the accuracy or completeness of their unpacking functions. Thus, users are often left to settle for whatever is provided to them.

Smart unpacking research was born to address these issues in a new way. The problems are addressed mathematically by identifying and locating the invariant meta-patterns of digital content. This allows the characterization and extraction of embedded content without necessarily requiring pre-written unpackers. These methods are also utilized to form measurements as to the completeness and accuracy of a given unpacking method for a given compound file or meta-structure.

This research, although new, has yielded some very promising results that suggest not only the soundness of the concept but perhaps a new approach to these problems in general. This talk presents the findings to date and demonstrates their use in practice.

Mathematical Content Analysis, Data Measurement, Content Validation