## B6 Smart Unpacking: Methods for Characterization and Extraction of Embedded Content

*Benjamin Long, BBA*, NIST, 100 Bureau Drive, Stop 8970, Gaithersburg, MD 20899*

After attending this presentation, attendees will learn about the theoretical and practical frameworks, being developed for the NSRL, to characterize, extract, and measure embedded digital objects using mathematically-based techniques.

This presentation will impact the forensic community by presenting frameworks for addressing the content analysis, data extraction, and measurement of these embedded digital objects.

This work presupposes that digital content can be characterized and classified according to mathematical properties and structures. Discussed is ongoing work as well as how this can form a foundation for validation and measurement of the structures discovered.

This talk describes what Smart Unpacking Research is with respect to the National Software Reference Library (NSRL) project. The NSRL is a project of the National Institute of Standards and Technology (NIST) for collecting and providing identifying information about known files. The NSRL project's goal is to unpack as many files as possible. This research was originally conceived for addressing a particular scenario that occurs frequently in NSRL operations – the need to extract files from compound files (files containing other files) that have no corresponding off-the-shelf unpacker. In addition to this primary scenario, the results of this work may be applied in more general ways. The unpacking methods being developed in this ongoing research are derived by means of modeling the patterns in relevant files using mathematical and other modeling techniques. The objective of this talk is to present the current status of this work and its more general relevance to related problems in the computer forensics domain.

Specifically, this work uses Pattern Theory to develop high quality models for patterns of interest in files, file formats, as well as specific types of content. These models describe how certain patterns are formed and allow us to develop algorithms and techniques for recognizing patterns in certain types of files, file formats, and file content. These algorithms are then implemented using a software framework of parsers to extract files. The parsing framework may also provide measurements to help assess the completeness and quality of such file-extraction operations both for these derived unpackers, as well as, for off-the-shelf unpackers.

Also discusses are how such techniques might be applied to other challenges of general relevance in computer forensics. Generalized versions of this work will be most relevant to one of two tasks: (1) improving understanding of file format and content, as well as, (2) enhanced file carving techniques to extract digital objects out of their digital context.

The focus of the current work is not to reveal the content or structure in encrypted or compressed patterns, but simply to identify data that might contain embedded compressed or encrypted information. Once identified, such data can be *extracted* as objects for further processing (e.g., decompression or decryption).

**Mathematical Content Analysis, Data Measurement, Content Validation**

 ** Presenting Author*