AUTOMATED FORENSICS

Today's forensic tools are interactive programs designed to conduct investigations and prepare courtroom testimony. Although widely used, they have two important limitations: they cannot be used by personnel who have no forensic training; and they do not scale to handle the massive amounts of information collected by today's intelligence operations. **Automated forensics will make both possible.**

But neither vendors nor academic researchers are building automated tools. Vendors are spending their R&D dollars on improving performance, improving user interfaces, and supporting new kinds of file types and data formats. Research into automation is too risky: an automation *product* would need to integrate facility with non-English languages, hypothesis generation and validation, knowledge representation, error recovery, and new kinds of reporting technologies.

A few researchers are investigating new algorithms that could lead to automation, but most researchers are hampered by a lack of data to develop their techniques, lack of standardized corpora that can be used to scientifically compare results, and lack of a guidance—which kinds of forensic problems should we pursue? Furthermore, because most researchers test their tools on their own hard drives and network traffic, they already know what their tools are going to find. This is not an environment where deep, meaningful breakthroughs are likely to occur.

Working with large heterogeneous datasets, some researchers have generated important results. Building the MIT/Harvard Used Drive Corpus, with 750 images from around the world, led me to develop a range of automated techniques for determining the previous owner of a hard drive; automatically finding sensitive financial information; and automatically determining which drives in the Corpus had previously been used by the same organization (Cross-Drive Analysis). I am now using this corpus to develop new file carving techniques that can automatically recover data.

Large data sets can also be used to validate tools or show the feasibility of new ideas. For example, Simon Byers at AT&T showed that significant data could be automatically gleaned from Microsoft Word files published on the web, but to do so he had to download over 100,000 documents and subject them to automated document exploitation. Much of his research effort was spent in corpus development, not forensic research.

Funding organizations could kick-start a whole new approach to forensics—large-scale automated forensics—by establishing forensic capability goals and then validating techniques and tools with a series of challenges. Research would further be accelerated by the creation and distribution of large-scale corpora of disk images, packet archives, document archives, log files, and other kinds of information.

Simson L. Garfinkel, March 10, 2007
Associate Professor, Naval Postgraduate School, slgarfin@nps.edu