B2 Quantifying Error Rates of File Type Classification Methods

Vassil Roussev, PhD, University of New Orleans, Computer Science, 2000 Lakeshore Drive, 311 Mathematics Building, New Orleans, LA 70148*

After attending this presentation, attendees will gain an appreciation of the limitations of current file type classification techniques. Attendees will also gain a methodological insight on how to critically examine published work and will learn of new techniques and tools that aid their work.

This presentation will impact the forensic science community by demonstrating how to practically approach the problem of "examining validation and expelling incompetence" for an important area in the field. It is believed that this basic approach readily extends beyond the specific area and has much wider methodological implications.

The problem of identifying the file type of a sample of data arises as part of basic digital forensic processing, such as data recovery and reconstruction, as well as network forensics. Looking ahead, its importance is expected to grow further with the adoption of forensics triage and statistical sampling techniques, which will be increasingly needed to deal (in part) with the rapid growth of target data.

This presentation will discuss both methodological issues not addressed by current work and provide a case study to illustrate the points. Specifically, it will be argued that the current formulation of the file type classification problem is inherently flawed for large classes of file types, such as *pdf* and *doc/docx*, and cannot possibly yield informative error rate estimates. Therefore, the problem is re-formulated as two separate problems—primary data format classification and compound data format classification that are independent of each other.

It is also argued that existing studies have been flawed both methodologically and statistically, as the volume of data studied is woefully inadequate to draw any reliable conclusions. This presentation will demonstrate that for some popular storage formats, such as *deflate*- coded data, the problem of classifying it cannot be based on current statistical approaches and a deeper, specialized analysis, including expectations of the content of the uncompressed data, is a hard requirement.

Finally, a case study will be discussed in which classification techniques were evaluated for some popular primary data formats, such as *jpeg* and *mp3*, and quantify their reliability as a function of the sample size. The reliability of compound format detection for *pdf* and *zip/docx* formats will be evaluated and a new analytic technique will be demonstrated that can distinguish *deflate*-compressed data from other types of high-entropy data.

Digital Forensics, Error Rate Estimation, File Type Classification

** Presenting Author*