A Novel Framework for PDF-Based Stegomalware Detection Using Kolmogorov Complexity

Alexis, Sebastian (School: Northwood High School)

PDF-based stegomalware is a growing cybersecurity threat that disguises malicious payloads within the tree structure of PDF documents, exploiting features like embedded JavaScript, metadata, and actions while maintaining the outward appearance of a legitimate file. Traditional detection methods, both heuristic and signature-based, fail to identify stegomalware due to the external similarities between clean and infected documents. This project presents a novel method for detecting PDF-based stegomalware using Kolmogorov complexity. Kolmogorov complexity measures the shortest possible program to describe an object and is incredibly useful for differentiating between benign javascript, which repeats frequently, and malicious javascript, which is much more random and concentrated. However, Kolmogorov complexity cannot be computed directly due to the halting problem. This project indirectly calculates it by compressing the object, and concatenating that with a decompressor script to form a self-extracting archive, with the total size of that code serving as an upper bound of its Kolmogorov complexity. By calculating the complexity of key components within a PDF's internal tree structure and comparing them to baselines derived from known clean documents, malicious modifications can be identified by detecting deviations from expected values. Initial testing achieved a 100% true positive rate but an 86.2% false positive rate when analyzing the entire PDF as a whole. Refining the approach to analyze specific PDF components such as javascript or open actions helped to shift focus on suspicious changes, and ignore benign variations like file size or text content, which improved results to a 97.8% true positive rate and a 3.7% false positive rate.

Awards Won: