

PDF Forensics and Attack Analysis: Development of a Unified Investigation Tool

Michele Merico (mat. 332058)

Supervisor:
Prof. Andrea Atzeni

Co-supervisor:
Prof. Paolo Dal Checco

December 2025

Problem Addressed

In recent years, the number of cyber attacks delivered through PDF files has risen sharply. Due to its ubiquity and flexibility, the PDF format is frequently abused as an attack vector for malware, phishing, obfuscation techniques, embedded malicious payloads or tampered documents. At the same time, PDFs often constitute crucial digital evidence in forensic investigations, especially when received as attachments in standard email or Certified Email (PEC).

Despite the availability of multiple PDF analysis tools, the current ecosystem suffers from several critical limitations:

- Most tools perform only **static** analysis of the PDF structure, without correlating the document with its transmission metadata (email/PEC headers, timestamps, certificates).
- Many tools do not guarantee **forensic soundness**: they do not isolate the evidence, do not record the execution environment and do not ensure reproducibility of results.
- Almost no existing tool performs reliable **digital signature verification**, including detection of modifications after signing or objects excluded from signature coverage.
- Analysis of **embedded files** (attachments internal to the PDF, compressed payloads, secondary documents) is largely unsupported.
- Almost no existing tool guarantees **easy reproducibility**: a large portion of experiments in digital forensics and security research fail due to undocumented or non-standardized analysis environments. Experts often face inconsistencies caused by manual configurations, dependency mismatches

or lack of precise records of the execution setup. As a consequence, identical inputs may produce different outputs across systems or analysts. This creates significant reliability issues for forensic work, making it essential to have a solution that ensures deterministic behaviour: the same input must always lead to the same output.

These shortcomings make investigations more error-prone, reduce reliability in legal contexts and hinder analysts, especially non-experts, from safely examining potentially malicious PDFs.

Proposed Solution

To address these limitations, the thesis introduces **foredf**, an integrated and forensically sound tool for PDF analysis. Built on top of the open-source **peepdf** framework, **foredf** extends its core capabilities and embeds the entire workflow inside a controlled Docker environment.

The solution introduces three main advancements:

1. Unified and automated acquisition workflow

- Automatic downloading of emails and PEC messages through an Email Fetcher module.
- Preservation of all transmission metadata (headers, PEC receipts, timestamps).
- Secure extraction of PDF attachments while maintaining chain-of-custody.

2. Extended PDF forensic capabilities

- Full object-level inspection inherited from **peepdf**, including detection of suspicious structures, JavaScript, launch actions, incremental updates, and obfuscation techniques.
- Extension of **peepdf** for the analysis of **embedded files**, introducing additional capabilities such as hashing, entropy analysis, MIME identification and static malware scanning through YARA.
- Additional **peepdf** module for **digital signature verification**, enabling the identification of signed objects, validation of PKCS#7 signatures, detection of tampered or unsigned objects in incremental updates, and graphical visualization of covered objects along with related signature metadata and validity information.

3. Forensic soundness and reproducibility

- Execution in a Docker container to isolate evidence and avoid accidental modification.

- Deterministic analysis environment guaranteeing reproducible results.
- Automatic generation of human-readable preliminary basic reports summarizing metadata, structural findings, signatures and embedded content.

Together, these features create a comprehensive solution that bridges the gap between static PDF scanners, malware sandboxes and forensic suites.

Results Achieved

The tool was evaluated on a diverse set of real-world PDFs, including files extracted from emails and PEC communications. The tests covered malicious documents, partially or fraudulently signed PDFs, files containing embedded content and PDFs crafted using known attack techniques.

The main results can be summarized as follows:

- **Successful integration of email/PEC acquisition:** **foredf** reliably fetches messages, preserves metadata and extracts attachments while maintaining forensic integrity.
- **Accurate structural analysis:** the tool correctly identifies objects, embedded files, incremental updates, embedded JavaScript, suspicious actions and obfuscation patterns.
- **Reliable digital signature verification:** the tool detects partial or forged signatures, unsigned incremental updates and objects added after signing, capabilities missing in standard PDF viewers and most analysis tools.
- **Secure handling of malicious PDFs:** containerization prevents any interaction between potentially dangerous documents and the analyst's system.
- **Improved usability for non-experts:** **foredf** enables users with moderate technical skills to understand risks, authenticity and integrity of PDF documents without the introduction of any risk for the user's system as it analyzes PDFs in a containerized environment.
- **Enhanced support for forensic investigators:** generates a basic report including email metadata, user interactions, enhanced peepdf analysis results and sender/hash verification, providing a foundation for more formal forensic reporting and chain-of-custody tracking.

Overall, the results demonstrate that **foredf** provides a **unified, reproducible and forensically sound workflow** for analyzing PDFs, whether they are standalone files or attachments embedded in email or PEC communications. With further development, the tool has the potential to become a valuable asset in both operational cybersecurity environments and complex forensic investigations.