Securing Data Integrity of CSV

A Novel Watermarking Approach for Tamper Detection

In an increasingly inter-connected world, ensuring the integrity and authenticity of shared information is more crucial than ever. As digital assets are exchanged across platforms, the need for interoperable data formats is evident. Among plain text data, the Comma-Separated Values (CSV) format prevails to be one of the most used data formats, widely employed for processing and transfer due to its simplicity and compatibility. However, due to its simplistic nature, CSV files lack inherent security measures, exposing data consumers to risks related to data integrity. To address this issue, researchers from Fujitsu and Fraunhofer ISST have introduced a novel approach: a fragile watermarking technique that employs invisible line-ending control characters to verify the integrity of CSV files. This blog post explores the essential elements of their research and its implications for data security.

The Challenge: Data Integrity of CSVs

The CSV format is one of the most used text data formats, particularly in governmental and organisational settings, e.g., to publish open data. Despite their popularity, these files are vulnerable to tampering since they do not offer any mechanisms for verifying the integrity of their contents. Recent studies have highlighted data integrity attacks as a significant concern, with potential consequences ranging from financial losses to severe operational disruptions.

Traditional methods of ensuring data integrity, such as cryptography, might not always be feasible, as additional metadata or signature files need to be shipped and stored separately. This significantly increases costs, as existing systems require extensive modifications. Therefore, watermarking — especially fragile watermarking — presents itself as a low-cost solution to address the concern of CSV integrity. Fragile watermarks are, opposed to regular robust watermarks, aimed at tamper detection of the cover medium. However, while several fragile text watermarking techniques for tamper detection already exist, many are either visible to the human eye or incompatible with the CSV structure, leading to a need for a more effective solution.

The Solution: Fragile Watermarking

In order to address this gap, a novel fragile watermarking approach is presented that embeds a digital signature within the CSV text using invisible line-ending control characters. Applying this method ensures that any alterations to the file will disrupt the watermark, allowing data consumers to detect tampering. Additionally, as no changes to the values and the data format are made, this solution can easily be applied to existing systems without major changes.

The proposed solution works by embedding a piece of information into a CSV text using a combination of carriage return (CR) and line feed (LF) line-ending control characters, commonly found in text encoding standards such as ASCII or Unicode. By mapping binary values (0 and 1) to CRLF and LF, respectively, it is possible to embed any byte-encodable information into the CSV format without altering the visible data. This is due to the fact that tools and platforms are capable of handling and recognising both types of line-endings. Accordingly, existing systems do not require any modification to adapt to the introduced changes, and regular users cannot tell any difference.

To illustrate the embedding process, consider the following simple CSV file with three rows of data. It is worth noting that the line-endings have been made visible for the sake of this example. Usually, they are not visible to the human eye.

```
1  Name, Age, Country CR LF
2  Alice, 30, Germany CR LF
3  Bob, 25, Japan CR LF
```

To embed the bitstring '101' in this CSV file, the embedding scheme first checks for enough rows to fit the given information. Since each line represents one bit, the total capacity is limited by the number of entries present in the CSV file. In this case, with the number of rows exactly matching the length of the binary string, the capacity is sufficient, allowing the embedding scheme to proceed. Accordingly, the information is embedded by replacing the existing line endings for the i-th row based on whether the i-th position of the information to be embedded holds a 0 or 1. The resulting CSV file looks like the following, again, with line-endings made visible:

```
1  Name, Age, Country LF
2  Alice, 30, Germany CR LF
3  Bob, 25, Japan LF
```

In order to extract the embedded information, the watermarked CSV simply needs to be iterated once again. That is, if a CRLF is found, a 0 is added to the extracted bitstring; if an LF is found, a 1 is added. After all rows are processed, the extracted bitstring will reveal the embedded original information. This way, a provider is able to embed a digital signature directly into the CSV, enabling the consumer to verify the origin and authenticate the integrity.

The watermark's fragility, i.e. its ability to detect any tampering, is based on two factors. First, the digital signature enables the consumer to check if the text has been altered before it reaches them. Digital signatures are a well-proven method to do so, commonly employed for tamper detection. Second, line endings can be affected by normalisation and formatting changes made by various tools on different platforms. This leads to the embedded information being fundamentally fragile in nature. Therefore, both the signature and the watermark itself are sensitive to changes, prompting consumers to refrain from using any CSV files they cannot verify.

Practical Implications and Future Directions

The implications of this research are significant for industries or use-cases relying on CSV files for data sharing and processing. By adopting the proposed fragile watermarking scheme, organisations are able to enhance their data integrity measures without compromising the user experience or requiring significant changes to existing systems, as the embedded fragile watermark detects any changes made to the CSV file. Additionally, using this approach, it is possible to embed virtually any information into a CSV file, given that it is byte-encodable. Therefore, non-critical meta-data or copyright notes could also be included without affecting existing systems or user experience.

However, some limitations do have to be considered, such as limited capacity due to its row-based approach and the fragility towards potentially desired normalisations. Therefore, future research should explore developing a semi-fragile watermarking technique that allows for common formatting and normalisation procedures. Lastly, it should be explored if this concept can be utilised for other texts that also feature numerous line-endings.

In conclusion, the proposed fragile watermarking represents a promising advancement in securing CSV files. By embedding invisible signatures within the data itself, the critical issue of data integrity from a data consumer perspective is tackled. As data continues to grow exponentially, solutions like this will be essential in ensuring that the information can be relied on and remains trustworthy and authentic.

Authors

Fraunhofer ISST
Florian Zimmer

Fujitsu Research
Janosch Haber
Mayuko Kaneko