УДК 004.056.5

ОБНАРУЖЕНИЕ УТЕЧЕК КОНФИДЕНЦИАЛЬНЫХ ДАННЫХ В ТЕЛЕМЕТРИЧЕСКИХ HTTP-ЗАПРОСАХ С ИСПОЛЬЗОВАНИЕМ МЕТОДОВ МАШИННОГО ОБУЧЕНИЯ

Новиков В. Д.[1]
Научный руководитель – инженер ФБИТ Савков С. В.[1]
[1]Университет ИТМО
uliyara007@gmail.com

Введение

В современных веб-приложениях повсеместно используются механизмы сбора аналитики для анализа поведения пользователей, контроля бизнес-процессов и иных задач [1]. При этом в большинстве случаев используются сторонние провайдеры телеметрии, что приводит к передаче данных на неподконтрольные владельцу приложения внешние аналитические платформы, и повышает риск утечки конфиденциальной информации. Практика показывает, что в телеметрические запросы могут попадать чувствительные данные, включая токены доступа и персональные данные пользователей [2]. Но существующие инструменты, используемые при аудитах безопасности для анализа трафика веб-приложений, ориентированы на ручной анализ и не обеспечивают автоматизированного выявления телеметрических запросов и поиска в них утечек. Таким образом, актуальной является задача автоматизации выявления утечек конфиденциальных данных в телеметрическом трафике веб-приложений.

Основная часть

В работе рассматривается задача выявления утечек конфиденциальных данных в телеметрических HTTP-запросах, под которыми понимается передача чувствительной информации на сторонние домены в составе параметров и тела запроса. Предложен двухэтапный подход к анализу трафика. На первом этапе выполняется автоматическое выделение телеметрических и потенциально телеметрических запросов из общего массива перехваченного HTTP-трафика на основе набора признаков: ключевые слова в URL, параметрах и теле запроса, а также особенности доменного имени получателя. Этап предназначен для сокращения объема анализируемых данных и повышения производительности обработки.

Использование исключительно регулярных выражений и эвристических правил для поиска конфиденциальных данных имеет ограничения, связанные с высокой вариативностью форматов токенов и идентификаторов. В связи с этим на втором этапе применяется подход на основе методов машинного обучения. Из каждого запроса извлекаются значения его параметров, путь и тело. Для представления строк используется векторизация символьными n-граммами TF-IDF [3]. Детектирование чувствительных значений выполняется линейным классификатором, который основан на методе опорных векторов, с вероятностной калибровкой, что позволяет ранжировать запросы по степени риска.

Реализован программный прототип, обеспечивающий полный цикл обработки экспорта HTTP-трафика из инструмента Burp Suite [4]. Для оценки качества используется размеченный, синтетический набор данных, сформированный путем

внедрения типовых конфиденциальных значений в реальные телеметрические запросы на основе сценариев, описанных в работе Leaky Forms [2].

Выводы

Предложен и реализован двухэтапный метод выявления утечек конфиденциальных данных в телеметрическом HTTP-трафике. Показана применимость символьных TF-IDF признаков и линейного метода опорных векторов для анализа строковых значений. Разработанный подход может использоваться при аудитах безопасности веб-приложений и автоматизированном анализе сетевого трафика. Дальнейшее развитие связано с расширением обучающей выборки и уточнением признаков телеметрии.

Литература

1. HTTP Archive. Third Parties // Web Almanac [Электронный ресурс]. – Режим доступа: https://almanac.httparchive.org/en/2025/third-parties (дата обращения 10.02.2026).
2. Starov O., Nikiforakis N. Leaky Forms: A Study of Email and Password Exfiltration Before Form Submission // USENIX Security Symposium. – 2022 [Электронный ресурс]. – Режим доступа: https://github.com/leaky-forms/leaky-forms (дата обращения 12.02.2026).
3. Методы извлечения признаков из текстовых документов // КиберЛенинка [Электронный ресурс]. – Режим доступа: https://cyberleninka.ru/article/n/metody-izvlecheniya-priznakov-iz-tekstovyh-dokumentov/viewer (дата обращения 11.02.2026).
4. PortSwigger Ltd. Burp Suite Community Edition [Электронный ресурс]. – Режим доступа: https://portswigger.net/burp/communitydownload (дата обращения 16.02.2026).