SENSEI: Scraper for ENhanced AnalySis to Evaluate Illicit Trends

Daniel De Pascale[1(✉)], Giuseppe Cascavilla[2], Damian A. Tamburri[2], and Willem-Jan Van Den Heuvel[1]

[1] Tilburg University - Jheronimus Academy of Data Science, 's-Hertogenbosch, The Netherlands
d.de.pascale@tue.nl

[2] TU/e - Eindhoven University of Technology, Eindhoven, The Netherlands

Abstract. Over the last years, we faced an exponential growth of illegal online market services in the Dark Web, making it easier than ever before of acquiring illicit goods online via a simple service interaction. To study and understand this emerging illegal services economy, we developed a trend analysis and (dark-)web services monitoring tool: SENSEI, which stands for 'Scraper for ENhanced analySis to Evaluate Illicit trends'. SENSEI extracts specific service transaction trends and analyses the human behaviours behind, to produce symmetric insights on specific service transaction habits from both customers and vendors on the Dark Web. Moreover, a trend analysis tool is provided to discover and typify relationships among different criminal activities and hence provide evidence and support investigation activities and Law Enforcement Agencies (LEAs) detecting criminal operations.

Keywords: Dark Web · Trend analysis · LEAs · Illicit goods

1 Introduction

Nowadays, the number of illegal drugs market services is rapidly increasing [3] in the unaccessible area of services and internet computing, also known as, the *Dark Web*. The Dark Web [2] plays a key role in this exponential growth due to its concealed nature, part of the web not indexed by search engines and hence chosen by criminals to build illegal service transactions.

The goal of this work is to build a service transaction investigation platform called SENSEI to help Law Enforcement Agencies (LEAs) analyze big data coming from Dark Web services (e.g., monitoring of drugs movements accross the world). The framework provides a collection of tools for big data analysis to extract valuable insights for cybercrime investigations. The framework's features include but are not limited to trend analysis of specific temporal snapshots, network analysis of vendors, and comparison of trends between different countries to evaluate the movement of illicit goods across the world. Moreover, the platform allows narrowing the field, acting on a specific time range to provide more specific information. To the best of our knowledge, SENSEI is the

J. Troya et al. (Eds.): ICSOC 2022 Workshops, LNCS 13821, pp. 374–383, 2023.
https://doi.org/10.1007/978-3-031-26507-5_36

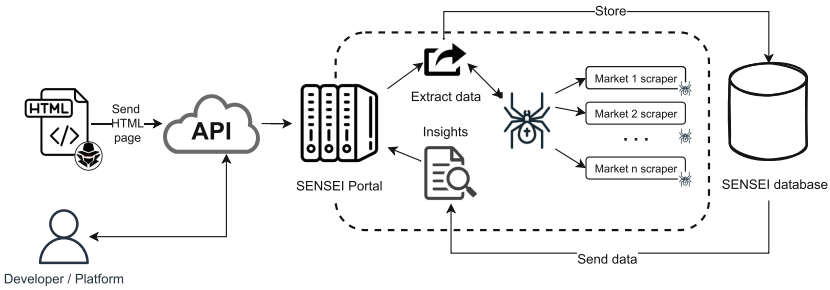

Fig. 1. Overview of the SENSEI SOA architecture, following the classical storing/retrieving cycle.

first open-source proof of concept platform for big data analysis on crime and user behavior activities from the Dark Web. Major big data analysis platforms like those offered by technologies such as Palantir[1] which result in expensive licensing and are often limited in their analytic and architectural extensibility capacity. Conversely, SENSEI relays solely on open-source technologies to provide an accessible analysis platform.

2 SENSEI Solution Architecture

SENSEI platform relies on a Service-Oriented Architecture (SOA) to store and retrieve insights from Dark Web pages and on a web platform to show them, instrumenting each analysis with explanatory narratives that explain the results.

SENSEI SOA allows interaction with the end-user by implementing a RESTful API architecture. The architecture enacts endpoints clustered into two main categories: *Load services*, to load new HTML files (or a zip HTML file) and store the information into a database (DB) and *Get services*, to collect and group the information stored from the database to visualize the analysis through the GUI (Sect. A in the online appendix [1] for further details). The input source of the architecture are HTML pages retrieved using a Dark Web crawler designed by us, working on the TOR darknet. Currently, the tool supports three Dark Web markets: Agartha, Cannazon, and Dark Market. Nevertheless, the platform eases the integration of new marketplaces by adding a python script with the scraping rules without changing the architecture. Each scraper is based on the HTML tag structure in order to extract the information. Hence, if the HTML structure of a Dark Web page changes, the scraper set of rules must be updated accordingly.

Lastly, the SOA architecture relies on a MySQL database to store the scraped data. The database is structured into three tables: *vendor*, *product* and *review*. We adopt a MySQL database system to increase the platform's performances by considerably speeding up all the writing and loading queries. Unlike

[1] https://www.palantir.com/.

other databases, such as NoSQL or Graph databases, MySQL optimizes high-performance joins across multiple indexes tables [5] (e.g., join between vendor and review to retrieve vendor's info given a review). Introducing a solid DBMS, the system reduces the server's workload while searching all the requested features for our final Trend Analysis tool in a fast, tractable and secure manner [4].

On top of the architecture, we built a web platform to show the insights provided by SENSEI API. The API receives the information needed, converts it into a JSON format, and sends it to the back-end platform, ready to be visualized in the web-based GUI. Based on the different types of analysis available, to visualize them in the front-end, the back-end platform provides 25 services grouped into 6 categories: *country*, *drug*, *market*, *vendor*, *trend analysis* and *general*.

3 Trend Analysis Services

The trend analysis platform provides multiple views divided into six different pages. Each page provides a specific type of analysis and insight (major details of each service are available in Sect. B in the online appendix [1]).

The home page shows an overview of the information retrieved after scraping the HTML pages loaded in the SENSEI SOA platform. It contains five general analysis tabs. *General analysis overview* shows the total markets available for analysis, the number of vendors, the number of products, and reviews collected through all the scraped markets. *Number of sales for each country* tab includes analysis displayed as a geo-map. It is possible to highlight a country to receive specific information related to vendors and the number of sales. *Top 4 countries* is a home page section that highlights countries based on the number of sales. The home page provides an overview of the seven most active vendors using the number of products sold in the *top vendors* tab. For a more general analysis, we provide raw data for all the countries in the *countries insights* tab. It does not organize the data to show any insight. Conversely, they are displayed in a table to be examined and reviewed. Lastly, the *trend analysis snapshot* tab shows the analysis regarding the products on sale in a specified period and time. The value of each day is the sum of all the products on sale in all countries involved.

The trend analysis page provides a line chart for three categories: drugs, markets, and countries. The platform performs the trend analysis by price, the number of products, or the number of vendors. In addition, the platform allows picking a specific date, showing the trend analysis by year (trend analysis of 2021) or month (trend analysis of January 2021). Along with the trend analysis page, showing the trend of a specific category, the platform provides an additional page, namely *Trend Analysis Comparison* page, where the end user can see the trend analysis of two different categories and compare them. The comparative analysis provides information on users' behavior and the type of drugs consumed. Moreover, the analysis is not only related to products. Conversely, it also gives an overview of the money involved in comparing the total revenue of products and vendors between the two countries. The platform allows the comparison of two categories: drugs, filtering by markets, or markets, filtering by drugs.

The tree-map page provides detailed information regarding vendors and the number of items sold. The tree-map visualization helps to have a glance an overall idea about the most active vendors with higher profits. The financial information about vendors plays a vital role during an investigation to better understand the investigation subject. Moreover, by clicking on a specific vendor, it is possible to know the number of items on sale per drug category.

The vendor's search page enables finding specific vendors and all the related information organized on the same page. The tool then provides two views: general vendor's and specific vendor's information. The general vendor's information provides the vendor's id, the name, the market or the markets where is active, the origin country specified in the marketplace and the delivery available places. When the user selects a vendor, the tool provides a side panel with the vendor's details, such as the vendor's score, email, phone number, Wickr's username [6], the country delivery place, and the number of snapshots done. For example, if the tool receives five dumps with the same vendor, the snapshot value is five.

The last page provides an interactive network map to build relationships among vendors and markets. The graph eases the identification of vendors in different markets providing all the interconnection. The analysis provides a graphical visualization of the available network among vendors and markets and contributes to understanding the marketplace size based on the number of vendors and the extent of a vendor's business among different markets. Moreover, the network map is interactive, meaning the investigator can have more information on a specific vendor by simply clicking on a vendor node. An investigator can retrieve information from the interactive network map, like the number of markets where the vendor is active and the number of products sold in each market.

4 Conclusions and Future Work

Our Trend Analysis tool aims to support LEAs during cybercrime investigations. The overall framework encompasses two main platform with their respective architectures, the SENSEI SOA platform, used to feed it with data from different investigations and the SENSEI platform, with a web-based GUI to show insights and trends in real-time. We outlined the framework architecture and platform components, from the trend analysis to the interactive market-vendor graph, describing the key design decisions and assumptions. The Trend Analysis platform has been built on top of RESTful API architecture, continuously fed with data.

Currently, SENSEI lacks of an exhaustive experimentation because we do not have enough data to feed the tool. Our priority is to gather data from several Dark Web marketplaces to validate our work. We provide a tutorial of SENSEI at the following link: https://youtu.be/aaT4J0Yd9lQ.

Requirement

In the following, we provide installation information and the interaction process between the SENSEI platform and the SENSEI SOA architecture.

Installation

The SENSEI architecture lays on Docker to build and run its components. Docker-compose is used to create an interconnected infrastructure, where the database can communicate with the SENSEI SOA tool and then with the SENSEI platform. The usage of a docker-compose eases the configuration settings, the building process, and the execution process. Indeed, to build and execute the entire framework, it is sufficient to run, the first time, from the main project folder, the following commands:

```
sudo docker network create anita-network
```

to create the network where the SENSEI framework operates and this command:

```
sudo docker-compose up --build
```

to build and execute the framework. The platform is available in the GitHub repository: https://github.com/danieldp92/SENSEI.

A Services

A.1 Load Services

Table 1. Original dataset. The attribute name is an identifier. Instead age, gender and postcode are quasi-identifiers.

Service name	Description
Delete dumps	Delete all market's dumps for each market passed as input from the local folder
Upload dump	Upload into local folder and scrape data to store into the DB
Delete market dumps	Delete all market's dumps (local folder and DB)

Load services manage the loading of datasets and the extraction of data from HTML pages passed as input. Table 1 shows the endpoint list of all LOAD services developed.

The *upload dump* endpoint reads the HTML page received and extract information to store into the database. As shown in Fig. 1, the tool has as many scrapers as the number of dark markets taken into account. This work takes into account three dark markets: Agartha, Cannazon and Dark Market.

The *delete dumps* is used to delete one or more markets scraped in the past. It is possible to set parameters as 'timestamp' and 'market name'.

A.2 Get Services

Get services allow the end-user to effectively retrieve information stored during the loading process. The information retrieved by the GET services includes

Table 2. Get services.

Service name	Description
Market services	
Get markets	This module provides different services for the integration and the editing of new marketplaces
Get dumps	Get all dumps stored for each market
Get market's dumps	Get all market's dumps
Vendor services	
Get vendors	Get the list of all vendors stored in the system (only the name of the vendor and the marketplace to which it refers)
Get vendor	Get the vendor in a specific marketplace. This service provides detail information of a vendor
Get vendor's products	Get the products of a vendor in a specific marketplace (only the name of each product sold by a vendor)
Get vendor's product	Get the product of a vendor in a specific marketplace. This service provides detail information of the product
Product services	
Get products	Get the list of all products stored in the system (only the name of the product and the marketplace to which it refers)
Get product	Get the product in a specific marketplace. This service provides detail information of a product
Graph analysis by vendor's name	Provides the graph analysis of a vendor, showing the relation with all marketplaces available
Graph analysis services	
Graph analysis by vendor's pgp key	Provides the graph analysis of a vendor, showing the relation with all marketplaces available, based on its pgp key

information like vendors and the product list and data selection to build the graph analysis by vendor's name or PGP key. The services provided in this macro category can be grouped into four main categories:

- Market services: services used to get information regarding marketplaces, such as details of a specific marketplace stored, timestamp, size of the dump, and the number of pages.
- Vendor services: services used to retrieve information about vendors.
- Product services: services used to retrieve information about products.
- Graph analysis services: services used to show the connections of a vendor among marketplaces, leveraging the vendor's username and the vendor's PGP key.

In Table 2 we provided a detailed list of all the services under the GET category. In addition, we listed all the service names and a short description for each of them to explain their purpose.

B Trend Analysis Platform

B.1 Home Page

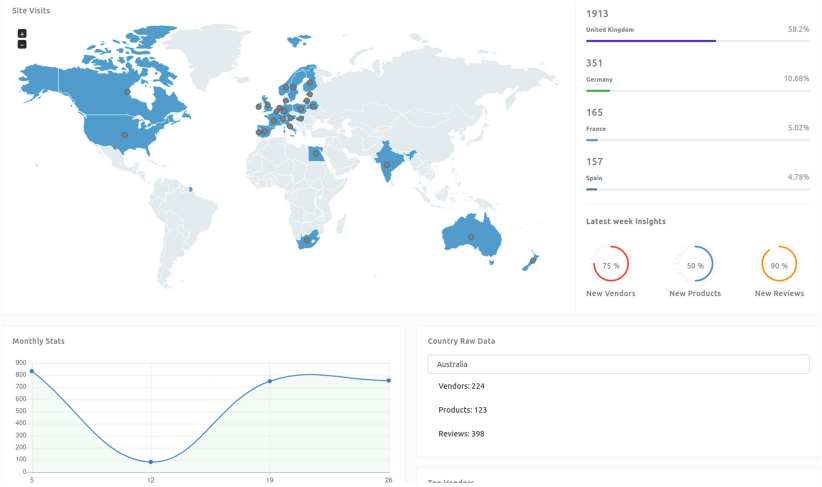

Fig. 2. SENSEI platform screen: home. It contains general info about the top vendors, a trend analysis of the last month recorder, general info for each country and a map containing the number of products sold in every country.

Table 3. Platform home services.

Service name	Description
GET /insights/	Retrieve the following insights: number of markets, number of vendors, number of products, number of reviews
GET /country/sales/	For each country, get the number of products on sale in the last month
GET /country/top-sales/	Get the top 4 countries with the highest number of products on sale
GET /top-vendors/	Retrieve the top vendors of the last month
GET /country/list/	Return all the countries recorded by the platform
GET /country/rawdata/	Return the raw data for each country. Raw data: “number of products”, “number of vendors”, “number of reviews”
GET /ta/sales/last-month/	Retrieve the sales of the last month

B.2 Trend Analysis Page

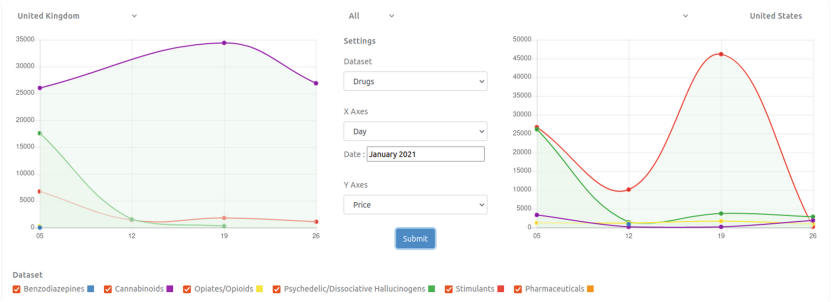

Fig. 3. SENSEI platform screen: trend analysis. Comparison between two countries based on the amount of drug, counted as the total price, in all the markets.

Table 4. Platform trend analysis services.

Service name	Description
GET /ta/drugs/	Get the trend analysis of all drugs in the marketplace
GET /ta/markets/	Get the trend analysis of all markets in the marketplace
GET /ta/countries/	Get the trend analysis of all countries in the marketplace

Table 5. Platform trend analysis comparison services.

Service name	Description
GET /ta/drugs/	Get the trend analysis of all drugs in the marketplace
GET /ta/markets/	Get the trend analysis of all markets in the marketplace
GET /ta/countries/	Get the trend analysis of all countries in the marketplace
GET /country/list/	Get the list of all the countries stored
GET /market/list/	Get the list of all the markets stored
GET /drug/list/	Get the list of all the drugs stored

B.3 Vendor's Tree-Map Page

In Table 6 we provide the list of services used to build the tree-map analysis. In the first row of the table, we have `GET /vendor/treemap/n-products`, the service is in charge of retrieving the number of products on sale per each vendor. Next, the service `GET /market/list/` is used to show the market list as a treemap's filtering option. The last service, `GET /vendor/treemap/vendor-name` extracts further details about the number of products on sale per each drug of a specific vendor.

Table 6. Platform treemap services.

Service name	Description
GET /vendor/treemap/n-products/	Get the total products on sale for each vendor
GET /market/list/	Get the list of all the markets stored
GET /vendor/treemap/{vendor-name}	Get the total products on sale for each drug's category of a vendor

B.4 Vendor's Search Page

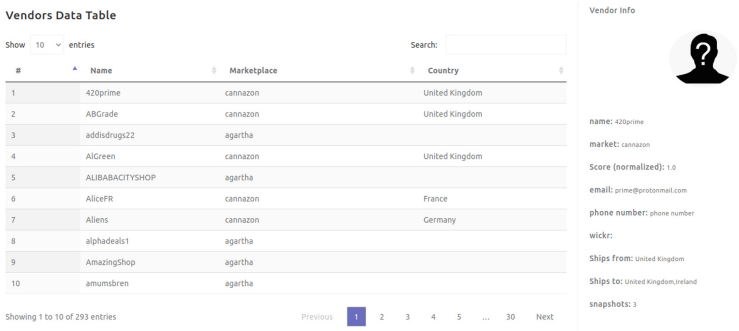

Fig. 4. SENSEI platform screen: vendors general info with detailed information regarding a specific vendor.

In Table 7 are shown the two services used in the vendor search page. The first service is in charge of retrieving the general data of all the vendors. While, the second one `GET /vendor/info/vendor-name` retrieves the information of a specific vendor.

Table 7. Platform search vendor services.

Service name	Description
GET /vendor/search/	Get the general insights (name, market, country) of all the vendors
GET /vendor/info/{vendor-name}	Get the vendor's info

B.5 Vendor-Market Graph Analysis

Table 8 shows the services used to build the interactive graph. First, `GET /market/n-products/` is used to estimate the impact of the vendor in a market, retrieving the number of products sold. Next, `GET /market/graph/` service

provides the list of all the vendors from a specific market. Last, the service `GET /market/graph/vendor/` provides additional information, like the number of markets connected to a vendor and the number of products for each market.

Table 8. Platform interactive graph services.

Service name	Description
GET /market/n-products/	Get the number of products for each market
GET /market/graph/	Graph info. Tt contains the mapping between vendors and markets
GET /market/graph/vendor/	Retrieve the number of markets where the vendor is active and the number of products on sale for each market

References

1. Appendix: Sensei (2022). https://doi.org/10.6084/m9.figshare.21131557.v2
2. Chen, H.: Dark web: exploring and mining the dark side of the web. In: 2011 European Intelligence and Security Informatics Conference, pp. 1–2. IEEE (2011)
3. EMCDDA: European drug report (2021). https://doi.org/10.2810/18539
4. Foster, E.C., Godbole, S.: Overview of MySQL. In: Database Systems. Apress, Berkeley, CA (2016). https://doi.org/10.1007/978-1-4842-1191-5_24
5. Győrödi, C., et al.: A comparative study: MongoDB vs. MySQL. In: 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), pp. 1–6. IEEE (2015)
6. Mehrotra, T., Mehtre, B.M.: Forensic analysis of Wickr application on android devices. In: 2013 IEEE International Conference on Computational Intelligence and Computing Research, pp. 1–6. IEEE (2013)