Similarity in Anonymity
A Dark Web Network Analysis

Csanád Végh and Niels Boonstra

JADS

Abstract. Dark web marketplaces provide a platform for illicit trade to thrive under a veil of anonymity, making it difficult to uncover connections between vendors. This paper introduces a tool that leverages similarities in product listings, such as pricing, strain types, shipping destinations, and textual descriptions of cannabis products, to identify possible links between sellers. We combined methods, including TF-IDF, Jaccard similarity, and Wasserstein distance, to build a vendor network based on these listing features. To evaluate our approach, we focused on vendors with nearly identical aliases and showed that our similarity metric could distinguish them from unrelated vendors with statistical significance. The end result is a web-based application that allows users to explore the vendor network interactively, detect communities, and thereby support investigative efforts by law enforcement.

Keywords: Dark web · Vendor alias detection · Product similarity · Network analysis · Entity resolution · Cannabis marketplaces · TF-IDF.

1 Introduction

The anonymity that the dark web provides poses challenges for law enforcement seeking to understand the inner workings and dynamics of illicit trade. It is difficult to link different vendor accounts that may be interconnected, or even belong to the same individual. By identifying these connections, we could enhance investigations, and provide valuable insights on the structure and dynamics of these networks. In this paper, we present a tool that uses similarity in listings of weed and hash products to infer potential connections between vendors on the dark web. We assume that shared connections correspond to shared linguistic traits, pricing, destinations, etc. across listings.

Even though dark web vendors excel at anonymizing and making themselves difficult to track, they may still exhibit patterns in how they describe, price, or present their products. These patterns may serve essentially as the behavioral fingerprints of the vendors.

Our primary objective is to develop a tool that can link similar product listings, resulting in a hypothesized network of vendors. We also aim to evaluate the validity of our method through testing, and provide an accessible tool that can enhance the work of the law enforcement in unveiling dark web markets. In the remainder of the paper, related works are reviewed in alias resolution and similarity analyses in section 2. Section 4 outlines the data collection process, the techniques used to compute similarities, and the approach for constructing the vendor network. In section 5, we present a way to assess whether the similarity metrics effectively distinguish vendor pairs likely to be aliases. We describe the implementation and functionality of the tool developed in section 6. Finally, we conclude.

2 Background and Related Work

This section reviews studies on dark web marketplaces and explores techniques for detecting vendor aliases through metadata and textual analysis.

2.1 Dark Web Marketplaces and Illicit Trade

Dark web marketplaces are online platforms operating on anonymizing networks like Tor, facilitating the trade of illicit goods and services. These marketplaces offer anonymity to both buyers and sellers. Christin [4] conducted a comprehensive analysis of Silk Road, one of the earliest and most prominent dark web marketplaces. The study revealed that the majority of listings were for narcotics, with cannabis being the most prevalent. Vendors on Silk Road exhibited varying lifespans, with the majority of sellers disappearing within roughly three months of their arrival.

Building upon this, Soska and Christin [12] examined sixteen different marketplaces over a two-year period. Their longitudinal study highlighted the resilience of the dark web marketplace ecosystem, noting that despite law enforcement interventions and marketplace shutdowns, the overall market adapted and continued to thrive. They observed that vendors often migrated to new platforms, maintaining their operations and customer bases.

Dittus [5] explored the geographical aspects of darknet drug markets, particularly focusing on the 'last-mile' delivery of drugs like cannabis and cocaine. Their findings indicated that most vendors were located in consumer countries rather than traditional producer nations. This suggests that darknet markets primarily contribute to the distribution of drugs within end-user regions.

2.2 Alias Detection and Entity Resolution

To survive in this dark, ever-changing landscape, Vendors often use multiple aliases across markets or over time. The next section explores how these identities can be linked through alias detection and entity resolution techniques. Various methodologies have been employed for alias detection, including **Stylometric Analysis**, **Temporal Analysis**, **Network Analysis**, **Behavioral Pattern Recognition**, **Cross-Platform Correlation**, and **Technical Fingerprinting**. In this study, we concentrate on two complementary approaches: **Stylometric Analysis** and **Network Analysis**.

Stylometric Analysis involves examining writing patterns, vocabulary choices, and grammatical structures to identify consistent linguistic fingerprints across different posts or communications. Ekambaranathan [6] presented an unsupervised method for linking user pseudonyms based on stylometry, demonstrating high accuracy in tracking user migrations across darknet forums. This was done by comparing the similarity of cosine vectors between each user. Similarly, VendorLink [11] utilized BERT-based NLP models to analyze vendor advertisements across multiple darknet markets, successfully identifying vendor migrations and potential aliases. Al Nabki [2] expanded this line of work by combining textual features and semantic similarity measures with darknet crawling to uncover cross-market vendor relationships.

Network Analysis examines communication patterns, transaction flows, and social connections between different accounts. Fonseca dos Reis [9] analyzed Bitcoin transaction networks linked to dark web marketplaces, identifying key players and highlighting the role of 'multihomers'—users operating on multiple marketplaces concurrently. Their findings show the importance of network structures to track users amongst different marketplaces. Although not a traditional network analysis, Kumar [8] introduced eDarkFind, an unsupervised model that links vendor accounts across markets by comparing multiple types of information, such as writing style, product descriptions, shipping locations, and drug types. By combining these features, their method creates a kind of similarity space where related accounts are close together. While they do not build an actual graph, this approach captures relationships between vendors in a way that resembles a network structure. Their model achieved 98% accuracy in matching accounts using feature embeddings derived from BERT and domain-specific language models.

Building upon these insights, our approach integrates stylometric and network analysis techniques to enhance alias detection. While prior work has explored stylometric features or transaction networks independently, there is a lack of integrated approaches that analyze similarities between vendors in a network. Additionally, existing tools often lack interactivity or configurability for investigative use. This project aims to construct a robust method for identifying vendor networks on these anonymous marketplaces by combining stylometric features with relational data.

3 Research Question

To what extent can similarity in product listings be used to detect alias relationships and vendor communities in dark web cannabis marketplaces?

3.1 Subquestions

To structure our investigation and develop a focused empirical approach. The following subquestions were created:

1. **How do different product listing attributes and their assigned weights affect the structure of similarity networks?**
   This question explores how variations in attributes and weights change the resulting vendor networks and the inferred relationships between vendors. Different choices of parameters can all result in valid networks depending on the context.
2. **How can we find matches where we know with high confidence that two vendor aliases refer to the same entity?**
   This subquestion addresses the challenge of ground-truth validation. In this project, we need to identify anonymous users and it is vital to know if an alias has been found or not.
3. **Can statistical differences in similarity scores reliably distinguish likely alias pairs from unrelated vendors?**
   This evaluates the effectiveness of our similarity metric using statistical tests to determine whether behavioral patterns captured through listings can meaningfully separate true alias pairs from the general population of vendors.
4. **How can we design a tool that is usable by both data scientists and non-technical investigative professionals?**
   This subquestion informs the development of the front-end interface and the configurability of our tool. We aim to support both technical users who may want to experiment with feature weights and thresholds, and investigative users seeking actionable insights without programming knowledge.

These subquestions are central to our research and tool design, and more importantly, the tool enables users to answer them directly. Users can test which features matter most, explore potential alias matches, and examine similarity patterns. The tool supports both technical and non-technical users through adjustable parameters and clear visualizations. It is not just a solution, but a platform for investigation.

4 Methodology

To answer the research question, we first need to collect relevant data from dark web cannabis marketplaces. Then, we need to define how to compare vendors to determine whether they are similar. This section explains how we gathered the data and how we measured similarity between vendors to uncover potential connections.

4.1 Crawling and scraping the data

This project started with developing a sraping and crawling pipeline. The system uses Python's `requests` library to send HTTP/1.1 requests and routes all traffic through the Tor network using a local SOCKS5 proxy (`127.0.0.1:9050`) to access `.onion` sites. A randomized user-agent is attached to each request using the `fake_useragent` library to prevent detection and rate-limiting by simulating traffic from various browsers. Additionally, authenticated session cookies are loaded manually, which is necessary to bypass login forms and CAPTCHA challenges that otherwise block automated access.

Each darknet marketplace was structured differently, so custom scraping logic had to be implemented for each of the six websites. While the underlying HTML varied significantly between sites, all marketplaces displayed a core set of attributes:

- **Name**: Product description
- **Title**: Short identifier
- **Origin**: Shipping location
- **Seller**: Vendor alias
- **Destination**: Countries or regions to which the product can be shipped

Because these elements were reliably present in product listings but not on vendor profile pages we focused exclusively on scraping product listings. Listings also provided a more standardized structure, making them more suitable for cross-market comparison.

The crawler iterated through listing pages and saved the HTML locally. A `BeautifulSoup` parser then extracted the relevant data. Additional post-processing included standardizing the quantity of cannabis products: the product name field was parsed using regular expressions to extract numeric quantities and convert all units (e.g., pounds, ounces, kilograms) to a common unit (grams) to allow for price normalization and quantitative comparison.

In total, six dark web marketplaces were scraped, yielding 4,116 unique cannabis product listings. Although some listings also included additional data, this was not included in the core dataset since their structure varied significantly and would only be useful for building vendor networks inside one website.

4.2 Data exploration and processing

Data Cleaning and Preprocessing The dataset was first cleaned by dropping irrelevant columns and removing duplicate entries. During this step, extra symbols or formatting artifacts that appeared uniquely on certain marketplaces were also removed. This created a cleaner and more uniform base for analysis. One of the websites displayed both a low and high price range for some listings. To handle this, each of these entries was split into two separate rows, one for each price. This allowed for consistent price comparisons across all listings.

Strain Name Standardization via Fuzzy Matching Strain names were extracted using fuzzy matching against a known list of cannabis strains. Fuzzy matching finds the closest valid match based on string similarity. If a confident match was found, the strain was accepted and further metadata could be added. If not, the entry was marked as unknown or dropped. This approach was more reliable than using language models like DarkBERT or OpenAI's general-purpose models, which often returned hallucinated or non-existent strain names. Fuzzy matching provided a necessary confidence threshold and ensured only real, verifiable strains were included.

Currency Normalization and Feature Enrichment Prices were then converted to EUR using a fixed exchange rate dictionary. This made all listings directly comparable, regardless of the currency they were originally listed in. Afterward, additional strain metadata was added by merging with an external file, strains.csv. This allowed each strain to be enriched with information like effects or types. Finally, a price-per-gram column was calculated for standardization. A one-hot encoding column was also added to distinguish between listings for weed and for hash. This helped structure the dataset for later filtering, grouping, and analysis.

Final attributes The above preprocessing steps resulted in the following data to be stored about the scraped products:

- name of the product as it appears on the website
- destination (the countries/continents that the product is shipped to)
- the alias of the user that sells the product
- price of the product converted to euros
- weight of one item of the product
- strain and type of the product
- the rating of the sold weed type
- effects associated with the product
- flavor of the product

- description of the type of weed sold
- whether the product is weed or hash
- the website that the product data was scraped from
- the origin country that the product is sold from
- the number of items sold
- rating of the seller

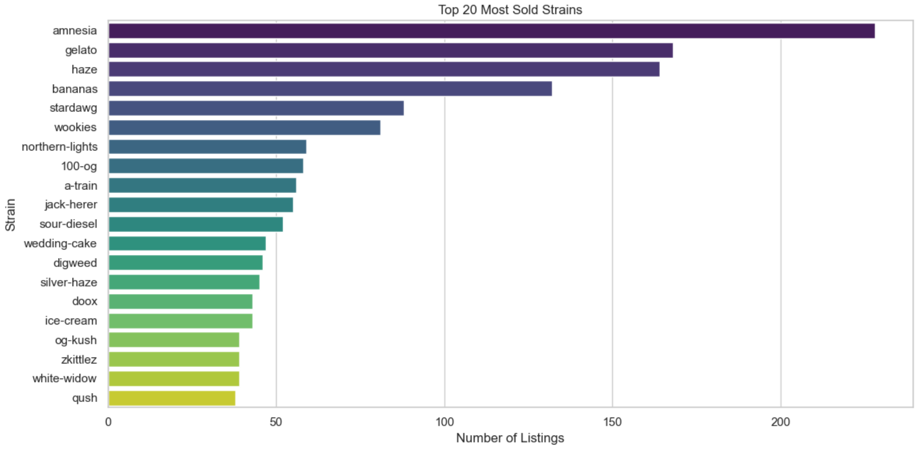


Fig. 1. Top 20 most sold strains of cannabis

Figure 1 shows the 20 cannabis strains most frequently listed across the dark web markets included in the dataset. The strain **Amnesia** appears nearly 200 times in total. This means around **1 in every 20 listings features Amnesia**. Vendors may group related strains under this label, which helps market familiar names even if the products differ genetically.

After Amnesia, popular strains include **Gelato**, **Haze**, and **Bananas**, which also rank highly in legal and illicit markets. The list features classic strains like *Northern Lights*, *White Widow*, and *Jack Herer*. It also includes trendy varieties such as *Wedding Cake*, *Stardawg*, and *Zkittlez*, which are known for high demand. This pattern suggests that vendors cater to customer expectations using well-known strain names. Brand familiarity likely plays a major role in buyer decisions. These naming trends reflect a growing similarity between darknet product marketing and that of legal cannabis outlets.

Figure 2 illustrates how cannabis products move from source countries to destination regions, based on marketplace listing data. Leading source countries include **Germany**, the **United Kingdom**, **Spain**, and the **United States**. Major destinations are **Europe**, **Worldwide**, and specific countries such as France, UK, and Germany itself.

Germany appears prominently as a source, largely because many listings come from the *TorZon* market. This platform features a high number of German

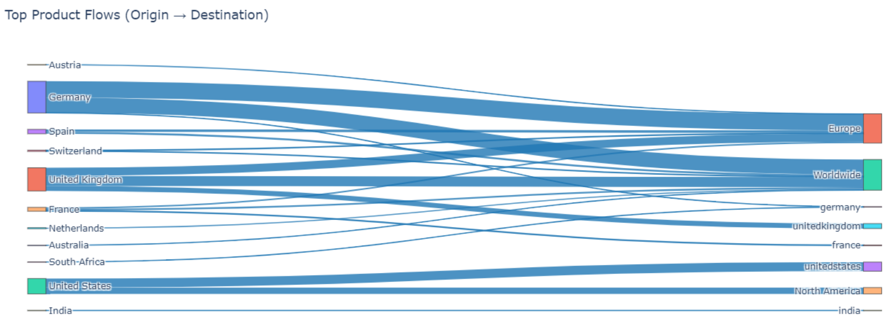


Fig. 2. Slinky diagram from origin to destination

vendors, making up about 85% of all sellers in the dataset. As a result, the visual data overrepresents Germany as a source. This reflects the composition of the data more than the full global market.

The United States is another major source and may be growing in influence. Cannabis is now legal in many U.S. states, making it easier for vendors to access supply. This change also lowers risk and improves product consistency. Many listings mention *"Cali weed"*, referring to cannabis from California. This phrase suggests higher quality and has become a marketing term. The strong reputation of California products adds to their appeal. Together, these factors show how changes in law, reputation, and market bias shape cannabis distribution on the dark web.

Figure 3 shows the distribution of cannabis prices per gram from dark web marketplace listings. Most prices are concentrated between approximately €5 and €12, with a clear peak around €8, indicating this is the typical price range for individual cannabis products. The distribution is right-skewed, meaning while most listings fall in the lower price range, a few listings are priced much higher, sometimes exceeding €40 per gram. These higher prices often correspond to premium products, pre-rolled joints, or special vendor services, and in some cases may indicate potential scams.

The presence of prices close to zero likely reflects free samples vendors offer to attract buyers or build trust. Additionally, listings selling larger quantities generally have a lower price per gram, consistent with typical bulk pricing strategies. The kernel density estimate smooths the data to highlight these overall trends and confirms the main concentration of prices in the lower range, with a long tail to the right, representing market variability.

Figure 4 presents the top 30 cannabis vendors based on the number of product listings across six scraped darknet marketplaces. The top five vendors include DieseGute, apeman420, QueenOfCannabis, Dreamjar, and Maling47. All of them are active on Torzon, which represents the largest share of listings in the dataset. Within this group, Dreamjar and Maurelius have fewer listings compared to their peers, which is unexpected given their overall prominence. However, both

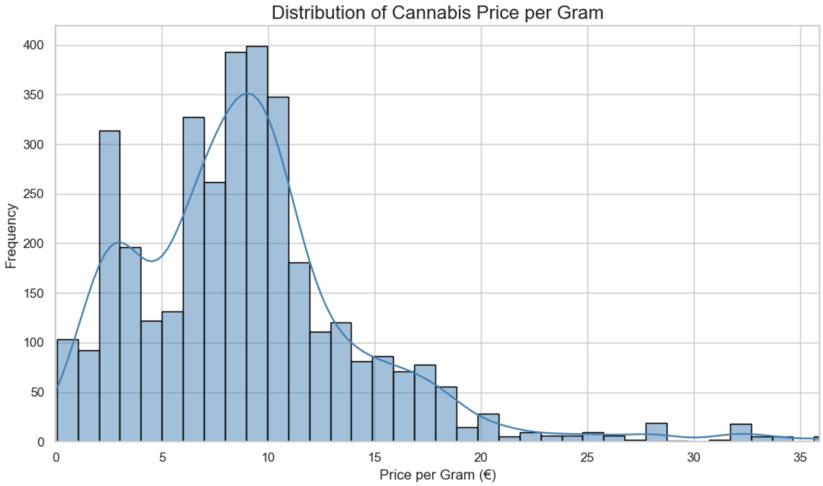


Fig. 3. Histogram for prices per gram

vendors are also heavily active on another marketplace called DarkMatter. In total, 190 unique vendors were identified, with 26 operating across multiple marketplaces. This suggests that while most vendors focus on a single platform, a smaller group adopts a cross-market strategy to broaden their reach.

After exploring the data, we decided to use the attributes for the similarity comparisons that can be most indicative of relation of sellers. For this, we used destination, price (adjusted by the weight), strain, effects, flavors, and the product text. The products were grouped by their sellers. This resulted in sets of values for each attribute from the products related to the same seller, as shown in Figure 5.

4.3 Computing similarities

To compute the similarity between any two sellers, a number of different similarity metrics were chosen depending on the data type. For categorical data, we used Jaccard similarity, for numerical data, we used Wasserstein distance, and for the text analysis, we used cosine similarity of TF-IDF vectors.

Jaccard similarity The similarity of two sets can be computed as the ratio of the common elements and all elements [7]. So, for example, to compute the Jaccard similarity of two sellers based on the flavors of the products, we would look at the number of flavors that they both offer, and divide that by all of the flavors that they collectively offer. With this, we can compute the similarity between any two sellers i and j using the following formula:

$$S_{ij} = J(A_i, A_j) = \frac{|A_i \cap A_j|}{|A_i \cup A_j|}$$

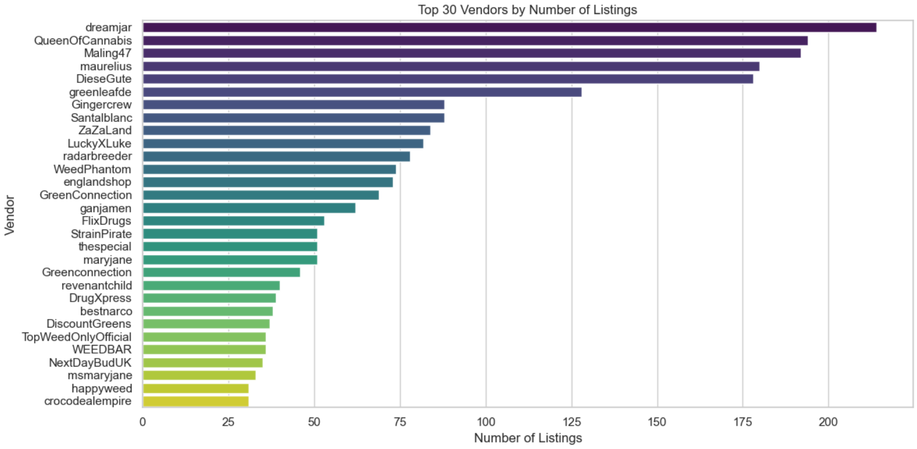
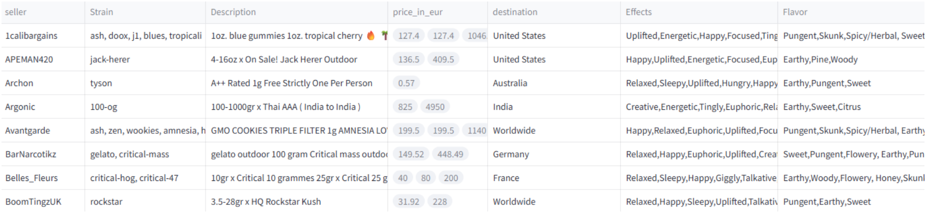


Fig. 4. Vendors with most listings



seller	Strain	Description	price_in_eur	destination	Effects	Flavor
1calibargains	ash, doox, j1, blues, tropicali	1oz. blue gummies 1oz. tropical cherry 🔥 🌴	127.4 127.4 1046	United States	Uplifted,Energetic,Happy,Focused,Ting	Pungent,Skunk,Spicy/Herbal, Sweet
APEMAN420	jack-herer	4-16oz x On Sale! Jack Herer Outdoor	136.5 409.5	United States	Happy,Uplifted,Energetic,Focused,Eup	Earthy,Pine,Woody
Archon	tyson	A++ Rated 1g Free Strictly One Per Person	0.57	Australia	Relaxed,Sleepy,Uplifted,Hungry,Happy	Earthy,Pungent,Sweet
Argonic	100-og	100-1000gr x Thai AAA (India to India)	825 4950	India	Creative,Energetic,Tingly,Euphoric,Rel	Earthy,Sweet,Citrus
Avantgarde	ash, zen, wookies, amnesia, h	GMO COOKIES TRIPLE FILTER 1g AMNESIA LO	199.5 199.5 1140	Worldwide	Happy,Relaxed,Euphoric,Uplifted,Focu	Pungent,Skunk,Spicy/Herbal, Earthy
BarNarcotikz	gelato, critical-mass	gelato outdoor 100 gram Critical mass outdo	149.52 448.49	Germany	Relaxed,Happy,Euphoric,Uplifted,Crea	Sweet,Pungent,Flowery, Earthy,Pun
Belles_Fleurs	critical-hog, critical-47	10gr x Critical 10 grammes 25gr x Critical 25 g	40 80 200	France	Relaxed,Sleepy,Happy,Giggly,Talkative	Earthy,Woody,Flowery, Honey,Skunl
BoomTingzUK	rockstar	3.5-28gr x HQ Rockstar Kush	31.92 228	Worldwide	Relaxed,Happy,Sleepy,Uplifted,Talkativ	Pungent,Earthy,Sweet

Fig. 5. The data after grouping by seller

TF-IDF similarity To measure the similarity of two texts, linguists often use TF-IDF [10]. TF refers to term frequency, that is the amount of times a term appears in a given text. IDF refers to inverse document frequency, that is the inverse of the amount of times a term appears in the whole corpus. Using this, more frequent terms like "the" or "and" would not contribute a lot towards similarity since they appear in all texts. However, if a term does not appear frequently in the corpus, but is frequent in two selected texts, that means that the two texts are indeed similar. By computing the TF-IDF scores for each seller's product texts, we get one TF-IDF vector per seller. We can measure the similarity between two sellers i and j given their TF-IDF vectors using cosine similarity.

$$S_{ij} = \frac{\boldsymbol{v}_i \cdot \boldsymbol{v}_j}{\|\boldsymbol{v}_i\| \cdot \|\boldsymbol{v}_j\|}$$

Wasserstein distance Assume you compare two real-valued price distributions, represented as sorted samples or empirical cumulative distributions. μ_i, μ_j are univariate distributions of prices for products sold by seller i and j. The first-order Wasserstein distance W_1 between them is:

$$W_1(\mu_i, \mu_j) = \int_0^1 \left| F_i^{-1}(t) - F_j^{-1}(t) \right| \, dt$$

Where $F_i^{-1}(t)$ is a quantile function (inverse CDF) of the price distribution μ_i, and the integral computes the average difference between the quantiles across the distributions [3]. The following formula computes the similarity based on prices between two sellers i and j:

$$S_{ij} = 1 - \frac{W_1(\mu_i, \mu_j)}{C}$$

Where C is a normalization constant to bring the value into the $[0, 1]$ range.

Overall similarity From the above formulae, we calculate an individual similarity matrix of sellers per each attribute. We want to aggregate these matrices. For this, we need the weight that each attribute should play in the computation of the similarity. These weights can be given by the user in the tool.

Let the weight vector be:

$$\boldsymbol{w} = \begin{bmatrix} w_{strain} \\ w_{description} \\ w_{prices} \\ w_{destination} \\ w_{effects} \\ w_{flavors} \end{bmatrix}$$

and the similarity matrix vector be:

$$\boldsymbol{S} = \begin{bmatrix} S_{\text{strain}} \\ S_{\text{description}} \\ S_{\text{prices}} \\ S_{\text{destination}} \\ S_{\text{effects}} \\ S_{\text{flavors}} \end{bmatrix}$$

Then the combined similarity matrix is:

$$S_{\text{combined}} = \boldsymbol{w}^\top \cdot \boldsymbol{S}$$

Finally, the normalized similarity matrix is:

$$S_{\text{final}} = \frac{S_{\text{combined}} - \min(S_{\text{combined}})}{\max(S_{\text{combined}}) - \min(S_{\text{combined}})}$$

4.4 Creating a network

To investigate a network of potentially related vendors, we can use the similarity scores to indicate ties. One could create a fully connected network where the weights between any two nodes (sellers) would be the similarity between their postings. However, visually this fully connected network would not significantly aid the user's understanding about network dynamics. Such a network would be useful for further investigations of endogenous and exogenous effects on tie strengths.

Instead of a fully-connected weighted graph, we proposed an undirected binary graph through the introduction of a threshold λ. In the similarity matrix, we convert each similarity score between two sellers i and j to 1 if the score is above the threshold, and 0 otherwise. This will yield the adjacency matrix of our final network. The following formula gives the adjacency matrix from the normalized overall similarity matrix.

$$A_{ij} = \begin{cases} 1, & \text{if } S_{ij} \geq \lambda \\ 0, & \text{otherwise} \end{cases} \quad \text{for all } i \neq j, \quad \text{and} \quad A_{ii} = 0$$

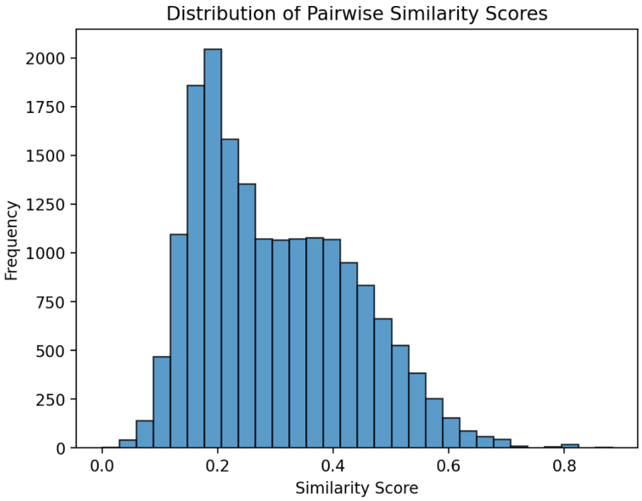


Fig. 6. Distribution of similarities with $w_{strain} = 0.2$, $w_{description} = 0.3$, $w_{prices} = 0.1$, $w_{destination} = 0.1$, $w_{effects} = 0.15$, $w_{flavors} = 0.15$.

The threshold λ defines the network structure and we can choose different values of λ for different purposes. Figure 6 shows the calculated similarity scores for a chosen vector of weights. Choosing a higher threshold would mean that only few ties are kept, and the network becomes very sparse. In this case, ties would potentially mean that two connected nodes are just two different aliases of the same vendor, or the connected sellers work together in very close proximity. On the other hand, choosing a lower threshold will make the network denser, and can reveal information about larger communities sharing a supplier, or having similar customer bases.

4.5 Community detection

To identify clusters of related vendors in the network, we use greedy modularity community detection [1]. The greedy algorithm starts by placing each node in its own community. Then, at each step, it merges the pair of communities that results in the greatest increase (or smallest decrease) in modularity. This continues until no further improvement is possible. This method is fast and effective for large networks, making it suitable for our purposes.

4.6 Workflow

The entire workflow of the methodology is shown in Figure 7.

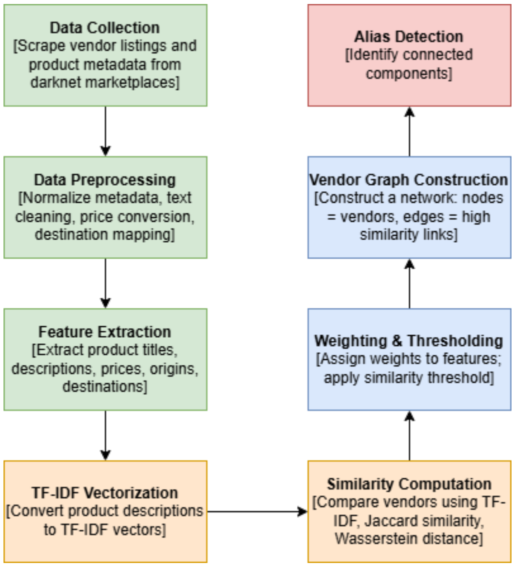

Data Collection
[Scrape vendor listings and product metadata from darknet marketplaces]
Data Preprocessing
[Normalize metadata, text cleaning, price conversion, destination mapping]
Feature Extraction
[Extract product titles, descriptions, prices, origins, destinations]
TF-IDF Vectorization
[Convert product descriptions to TF-IDF vectors]
Similarity Computation
[Compare vendors using TF-IDF, Jaccard similarity, Wasserstein distance]
Weighting & Thresholding
[Assign weights to features; apply similarity threshold]
Vendor Graph Construction
[Construct a network: nodes = vendors, edges = high similarity links]
Alias Detection
[Identify connected components]

Fig. 7. The workflow of acquiring and processing the data, computing the similarities, and constructing the network.

5 Validating the tool

Validating the tool is very challenging since we do not have access to ground truth data. We cannot be sure that any formed tie actually represents a connection in real life, or whether two seller aliases with very high similarity in listings is actually the same person. The similarities could just be by chance. Therefore we need some way to validate our methodology.

5.1 Experimental setup

In order to validate the tool, we need to find sellers that we know are related, and look at the similarity between their postings. There are some sellers in the scraped data with very similar aliases, they only differ in one added letter, or one capitalized letter. For example:

- DarkAngus - Darkangus
- DieseGute - DieseGute_
- TopCat - TopCatt
- GreenConnection - Greenconnection

We assume that these very similar aliases belong to the same person. From here, we can run a statistical analysis to compare the similarity scores that belong to these pairs to scores between random pairs. We also analyze the listings beforehand to ensure that similarities are not because the postings with the similar aliases are identical from different sites of the dark web.

The experimental group contains the similarity scores between 9 pairs of sellers chosen as potentially identical people, examples listed above. The control group contains all other seller pairs with similarity scores. The weight vector chosen for the experiment is the following:

$$\boldsymbol{w} = \begin{bmatrix} w_{strain} \\ w_{description} \\ w_{prices} \\ w_{destination} \\ w_{effects} \\ w_{flavors} \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.3 \\ 0.1 \\ 0.1 \\ 0.15 \\ 0.15 \end{bmatrix}$$

We hypothesize that the similarity between vendors with near-identical aliases is significantly higher compared to pairs of vendors with different aliases.

5.2 Results

A Wilcoxon rank-sum test was conducted to compare similarity scores between two groups of seller pairs: an experimental group consisting of hypothesized matched sellers (different usernames representing the same entity), and a control group comprising all other seller pairs. The results indicated a statistically significant difference in similarity scores between the two groups ($W = 131{,}537$, $p = 0.01072$), providing evidence that the experimental pairs exhibit greater similarity than seller pairs with unknown relations. The distribution of similarity scores in the two groups is shown in Figure 8. These findings support the hypothesis that the similarity in aliases corresponds to similarity in listings. This was done to validate our tool by showing that real-life relations between vendors can indeed be captured by our similarity calculation.

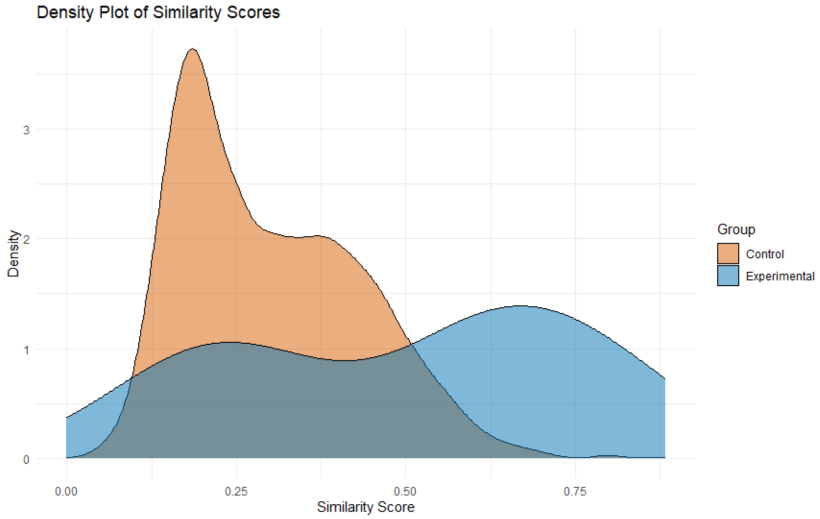


Fig. 8. Distribution of similar-alias similarities and non-similar-alias similarities.

6 The tool

The tool is a Streamlit web application that is designed to help users analyze and explore a network of related vendors in the black market. The tool can be accessed here, and the code can be accessed via GitHub. Below, we provide detail about what is possible with the tool.

6.1 Data upload

Users can upload a csv file containing data about the products scraped from the dark web, shown in Figure 9. At this point, we expect the data to have the columns with the specific names that correspond to the one we used in our data. However, in the future this can be made dynamic, allowing to explore similarities between any attributes.

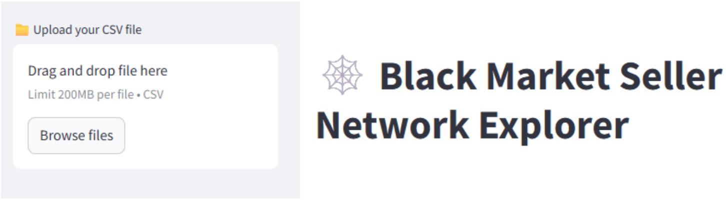

Upload your CSV file
Drag and drop file here
Limit 200MB per file • CSV
Browse files
Black Market Seller Network Explorer

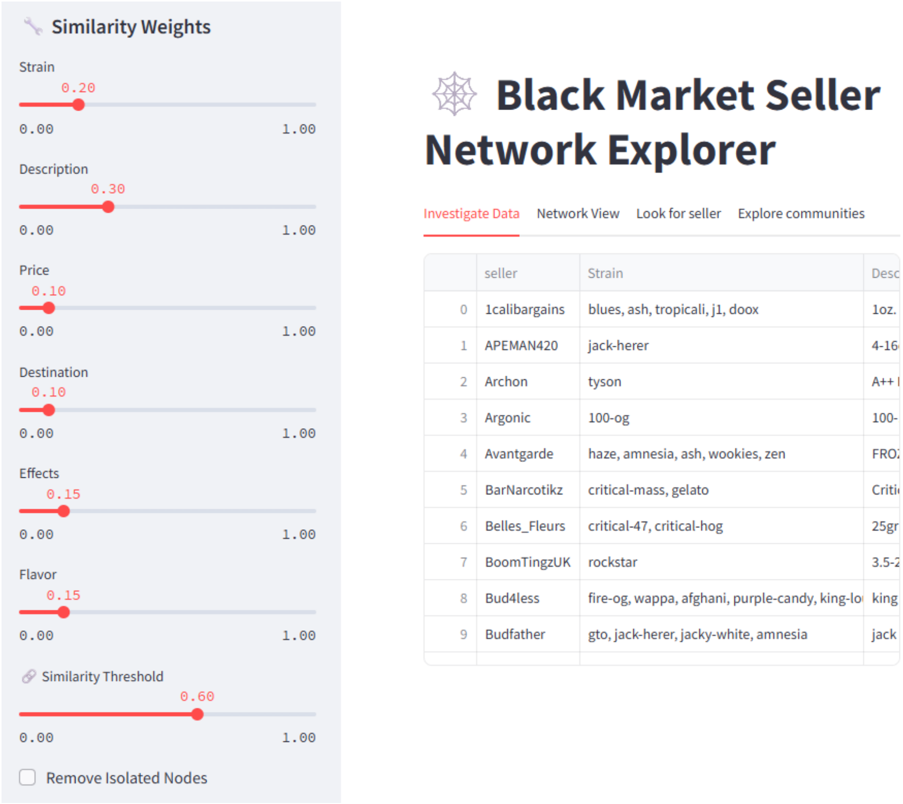
Fig. 9. The data upload function.

6.2 Similarity weights and threshold

As mentioned before, the tool can have a wide range of use cases depending on which similarities the user wants to focus on. Furthermore, they can create sparse or dense networks given what they set as the similarity threshold.

In some cases, one might focus on the isolates in the graph, to give an idea about the independence of certain actors, but most of the time these nodes do not carry useful information about the dynamics in the network. For this reason, there is an option to remove the isolated nodes from the network.

These functions are shown in Figure 10.



Similarity Weights

Strain 0.20 (0.00 – 1.00)

Description 0.30 (0.00 – 1.00)

Price 0.10 (0.00 – 1.00)

Destination 0.10 (0.00 – 1.00)

Effects 0.15 (0.00 – 1.00)

Flavor 0.15 (0.00 – 1.00)

Similarity Threshold 0.60 (0.00 – 1.00)

□ Remove Isolated Nodes

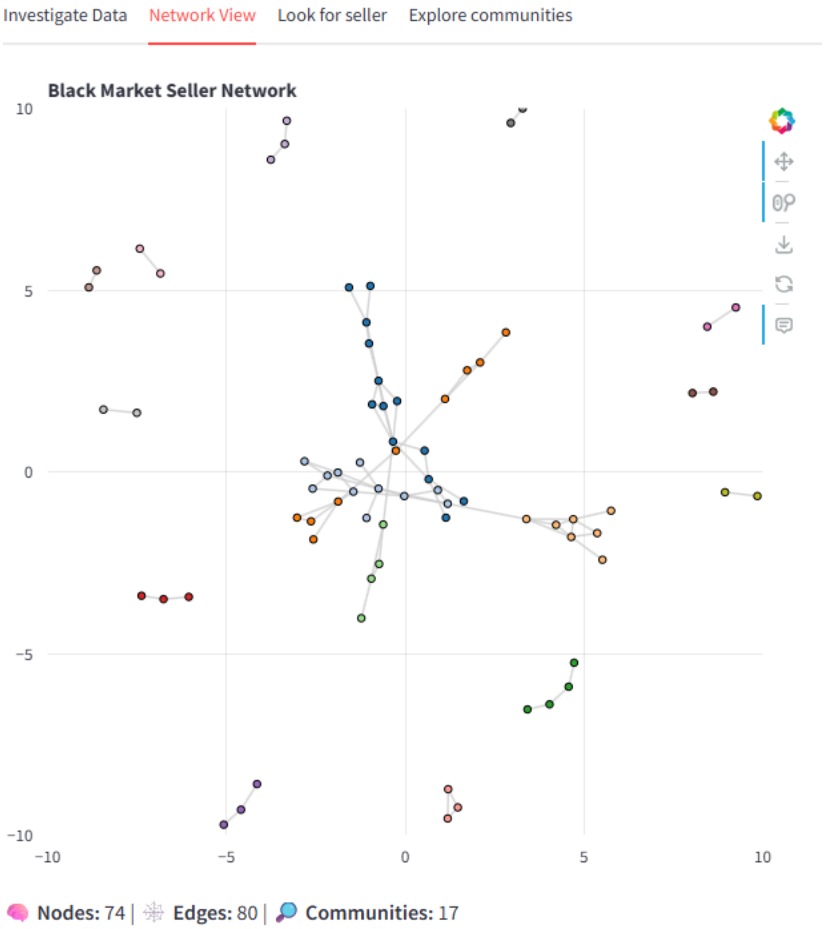
Black Market Seller Network Explorer

	seller	Strain	Desc
0	1calibargains	blues, ash, tropicali, j1, doox	1oz.
1	APEMAN420	jack-herer	4-16
2	Archon	tyson	A++
3	Argonic	100-og	100-
4	Avantgarde	haze, amnesia, ash, wookies, zen	FRO
5	BarNarcotikz	critical-mass, gelato	Criti
6	Belles_Fleurs	critical-47, critical-hog	25gr
7	BoomTingzUK	rockstar	3.5-2
8	Bud4less	fire-og, wappa, afghani, purple-candy, king-lo	king
9	Budfather	gto, jack-herer, jacky-white, amnesia	jack

Fig. 10. Setting parameters.

6.3 Table and network view

After uploading the data and setting the parameters, the user can explore the data. The data is shown in a table where the products are grouped by seller, and each column contains a set of product attributes that belong to a certain seller. Under 'Network View', the user can explore the network of connected sellers, shown in Figure 11.

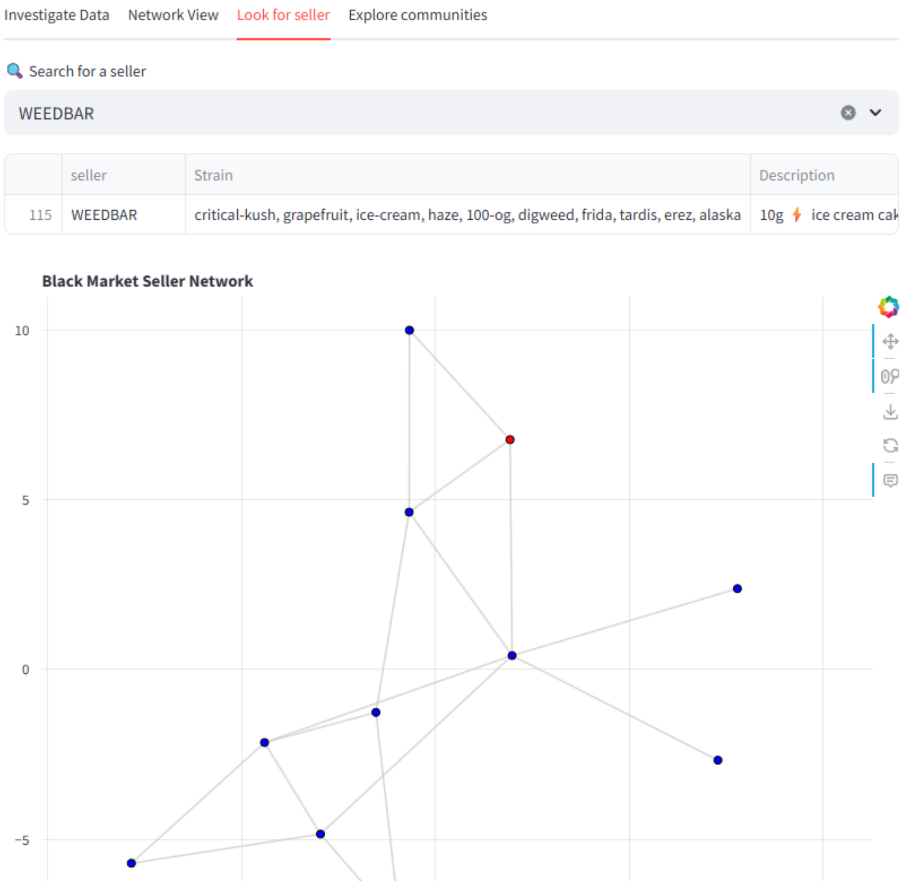


Nodes: 74 | Edges: 80 | Communities: 17

Fig. 11. The data in network view.

6.4 Searching specific vendors

The user also has the option to search for the alias of a seller. After choosing the seller, the tool shows the community that the specific seller belongs to, highlighting the chosen node. Furthermore, the user can also look within the identified communities. Figure 12 shows this function of the tool.



WEEDBAR

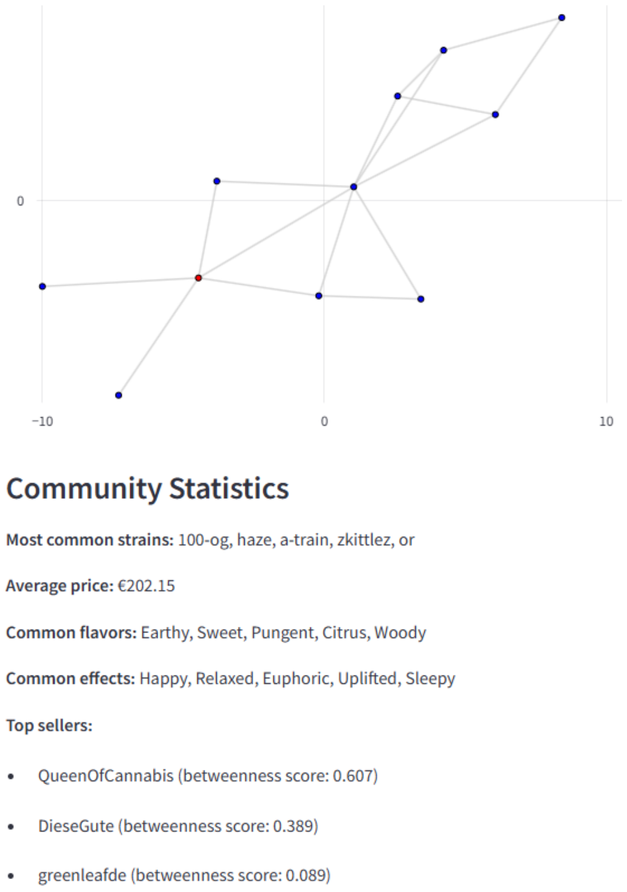
	seller	Strain	Description
115	WEEDBAR	critical-kush, grapefruit, ice-cream, haze, 100-og, digweed, frida, tardis, erez, alaska	10g ⚡ ice cream cak

Fig. 12. Searching for sellers function.

This function can be especially useful for law enforcement. For instance, through an operation the police gains access to the digital infrastructure of a specific vendor, *BestWeed* a not-careful-enough weak-link. Using the tool, the police can upload the product data belonging to this seller, even extending on the already existing scraped data from the dark web. Then the tool identifies similar sellers based on similarities in listings. Being a weak-link, *BestWeed* is probably just a leaf-node in the network. However, now we can identify the community that this seller belongs to, and we can identify the central hub of the community, *TheBotanist*. *TheBotanist* is likely to be a distributor. Focusing resources on this seller, or the interactions of *BestWeed* with this distributor, the police can uncover a broader network of illicit trade that extends beyond their original target.

6.5 Example workflow

To demonstrate the potential of the tool, let's imagine an example situation. The police have spotted an increase in black market cannabis products that appear in Netherlands. These products are very similar in their composition and packaging to ones previously found in different countries across Europe. Knowing this, they can upload data about dark web cannabis listings on the website, and set the similarity weights and threshold accordingly. Since they are looking for a specific type of product, they assume that connected sellers will have similarities in their products' strain, effect, flavor, description, and pricing. As the products ship to many different countries, the destinations could vary. The police can set the parameters and threshold accordingly.



Community Statistics

Most common strains: 100-og, haze, a-train, zkittlez, or

Average price: €202.15

Common flavors: Earthy, Sweet, Pungent, Citrus, Woody

Common effects: Happy, Relaxed, Euphoric, Uplifted, Sleepy

Top sellers:

- QueenOfCannabis (betweenness score: 0.607)
- DieseGute (betweenness score: 0.389)
- greenleafde (betweenness score: 0.089)

Fig. 13. The information about the community from the example, as shown in the tool.

Now under *Explore communities*, they can see the detected communities in the network. By choosing one, they can investigate the most common strains, flavors, effects, and pricing within the community. This allows them to look for the community which aligns with the products that they were looking for. They find that *QueenOfCannabis* has the highest betweenness score, and hence is likely the central hub of the community. They also get a list of the 11 aliases on the dark web markets that they can suspect are part of this illicit trade. This community information is depicted in Figure 13.

7 Discussion & Conclusion

This project set out to answer the question: *To what extent can similarity in product listings be used to detect alias relationships and vendor communities in dark web cannabis marketplaces?*

We combined TF-IDF, Jaccard similarity, and Wasserstein distance into a single similarity matrix, which we used to build a vendor network. This enabled a flexible, user-driven tool for exploring potential relationships between illicit sellers.

Our experimental validation showed that vendors with similar aliases tended to have higher similarity scores, which supported our assumption that the product listings can serve as behavioral fingerprints. This suggests that linking vendors based on similarity in listings can be promising for alias detection in the dark web.

Our tool offers an interactive method for analysts and law enforcement agents to explore seller connections. The customization of parameters allows for flexible use depending on the context of the investigation. One can focus on identifying individuals with several different aliases, or look at broader communities to detect supply chains and distributors. By combining network analysis with stylometric analysis, this work aims to provide a deeper understanding of the relational structures that exist within the anonymous landscape of the dark web.

7.1 Limitations

The validation of the tool can hardly be trusted empirically. It already relies on our assumption that vendors with nearly identical aliases belong to the same vendor. This is because we have no ground-truth data which could validate our methods. This could only be done while the tool is used in the field, or by accessing data about past police investigations.

Another limitation of the tool is that listings may be coincidentally similar, identifying non-existent links between sellers. For example, not interconnected sellers could still use the same price ranges or similar descriptions because they aim to sell to the same market. This also means that sellers can intentionally disrupt similarities to fake their behavioral fingerprints.

Furthermore, the dataset we used is limited to six marketplaces focused primarily on cannabis-related listings. Expanding the dataset to include more websites could improve robustness and allow temporal analysis to capture evolving vendor behaviors over time. Also, we concentrated on cannabis because vendors might share suppliers and products, which makes similarity detection more plausible. It was assumed that matches between highly distinct product categories, such as sellers of guns and sellers of weed gummies, are less likely and harder to identify accurately.

7.2 Future work

In the future, the tool could be expanded to other product categories with limited effort, since it contains hardly any contextually specific functionality to cannabis products. Extending the dataset beyond cannabis to include all types of listings would enable a more comprehensive detection of vendor relationships across diverse markets. Furthermore, one could also integrate a similarity analysis of product pictures or user profile pictures, or comparisons of other profile metadata into the analysis.

Different similarity frameworks and distance metrics could be tested to improve matching accuracy and robustness. Most importantly, testing the tool on larger, more reliable datasets of known alias relations should be a priority in further development. Additionally, continuous monitoring of marketplaces over time could allow temporal analyses to detect changes in vendor behavior and relationships dynamically.

References

1. Community detection, https://noesis.ikor.org/wiki/algorithms/community-detection
2. Al Nabki, M., Fidalgo, E., Alegre, E., de la Mata, J.: Classifying illegal activities on tor network based on web textual contents (2020), https://aclanthology.org/E17-1004/
3. Chilamkurthy, K.: Wasserstein Distance, contraction Mapping, and Modern RL Theory (12 2021), https://kowshikchilamkurthy.medium.com/wasserstein-distance-contraction-mapping-and-modern-rl-theory-93ef740ae867
4. Christin, N.: Traveling the silk road: A measurement analysis of a large anonymous online marketplace. In: Proceedings of the 22nd international conference on World Wide Web (2013), https://www.andrew.cmu.edu/user/nicolasc/publications/TR-CMU-CyLab-12-018.pdf
5. Dittus, M., Quattrone, G., Capra, L.: Platform criminalism: The 'last-mile' geography of the darknet market supply chain. Social Science Computer Review (2018). https://doi.org/https://arxiv.org/abs/1712.10068
6. Ekambaranathan, A., Gangopadhyay, A., Mukherjee, A.: Using stylometric analysis to link user pseudonyms on darknet forums. In: Proceedings of the International Conference on Social Computing, Behavioral-Cultural Modeling, and Prediction (2018), https://essay.utwente.nl/75908/1/Ekambaranathan_MA_EEMCS.pdf
7. Karabiber, F.: Jaccard similarity, https://www.learndatasci.com/glossary/jaccard-similarity/
8. Kumar, R., Yadav, S., Daniulaityte, R., Lamy, F., Thirunarayan, K., Lokala, U., Sheth, A.: edarkfind: Unsupervised multi-view learning for sybil account detection

(2020), https://www.researchgate.net/publication/341129252_eDarkFind_Unsupervised_Multi-view_Learning_for_Sybil_Account_Detection

9. Fonseca dos Reis, E., Teytelboym, A., ElBahrawy, A., De Loizaga, I., Baronchelli, A.: Identifying key players in dark web marketplaces through bitcoin transaction networks (2024), https://www.nature.com/articles/s41598-023-50409-5
10. Salton, G., Buckley, C.: Term-weighting approaches in automatic text retrieval. Information Processing Management **24**(5), 513–523 (1 1988). https://doi.org/10.1016/0306-4573(88)90021-0
11. Saxena, V., Rethmeier, N., Van Dijck, G., Spanakis, G.: Vendorlink: An nlp approach for identifying and linking vendor migrants and potential aliases on darknet markets (2023), https://aclanthology.org/2023.acl-long.481/
12. Soska, K., Christin, N.: Measuring the longitudinal evolution of the online anonymous marketplace ecosystem (2015), https://www.usenix.org/system/files/sec15-paper-soska-updated_v2.pdf