# Towards Image-Based Dark Vendor Profiling

An Analysis of Image Metadata and Image Hashing in Dark Web Marketplaces

Susan Jeziorowski
sjeziorow42@students.tntech.edu
Department of Computer Science
Tennessee Technological University
Cookeville, Tennessee

Muhammad Ismail
mismail@tntech.edu
Department of Computer Science
Tennessee Technological University
Cookeville, Tennessee

Ambareen Siraj
asiraj@tntech.edu
Department of Computer Science
Tennessee Technological University
Cookeville, Tennessee

## ABSTRACT

Anonymity networks, such as Tor, facilitate the hosting of hidden online marketplaces where *dark vendors* are able to anonymously trade paraphernalia such as drugs, weapons, and hacking services. Effective dark marketplace analysis and dark vendor profiling techniques support dark web investigations and help to identify and locate these perpetrators. Existing automated techniques are text-based, leaving non-textual artifacts, such as images, out of consideration. Though image data can further improve investigative analysis, there are two primary challenges associated with dark web image analysis: (a) ethical concerns over the presence of child exploitation imagery in illegal markets, and (b) the computational overhead needed to download, analyze, and store image content. In this research, we investigate and address the aforementioned challenges to enable dark marketplace image analysis. Namely, we examine image metadata and explore several image hashing techniques to represent image content, allowing us to collect image-based intelligence and identify reused images among dark marketplaces while preventing exposure to illegal content and decreasing computational overhead. Our study reveals that approximately 75% of dark marketplace listings include image data, indicating the importance of considering image content for investigative analysis. Additionally, 2% of considered images were found to contain metadata and approximately 50% of image hashes were repeated among marketplace listings, suggesting the presence of easily obtainable incriminating evidence and frequency of image reuse among dark vendors. Finally, through an image hash analysis, we demonstrate the effectiveness of using image hashing to identify similar images between dark marketplaces.

## CCS CONCEPTS

• **Applied computing → Evidence collection, storage and analysis**.

## KEYWORDS

dark web, Tor, dark marketplace, metadata, image hashing

*IWSPA '20, March 18, 2020, New Orleans, LA, USA*
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7115-5/20/3…$15.00
https://doi.org/10.1145/3375708.3380311

**ACM Reference Format:**
Susan Jeziorowski, Muhammad Ismail, and Ambareen Siraj. 2020. Towards Image-Based Dark Vendor Profiling: An Analysis of Image Metadata and Image Hashing in Dark Web Marketplaces. In *Sixth International Workshop on Security and Privacy Analytics (IWSPA '20), March 18, 2020, New Orleans, LA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3375708.3380311

## 1 INTRODUCTION

Anonymity networks, such as Tor, The Onion Router, have grown increasingly popular among web users who want to conceal their online identity and activities. Though Tor and other popular anonymity networks promote our human right to privacy, they provide an avenue for criminals to conduct illegal activities online without fear of consequences. Specifically, Tor allows for the hosting of hidden online marketplaces where *dark vendors* are able to anonymously trade paraphernalia such as drugs, weapons, and hacking services. Due to Tor's hidden service infrastructure, owners, vendors, and users of these marketplaces are difficult to identify and locate.

In typical web browsing sessions, various types of information about a user can be collected and stored including, but not limited to, location, searches conducted, browsed sites, social networks, banking data, email addresses, online behavior, and more. Thus, in typical surface web settings, law enforcement and intelligence agencies can leverage open source intelligence to investigate cyber criminals. However, utilization of web anonymity techniques by criminals makes such open source intelligence neither readily available nor easily obtainable. Exposing anonymous activity is integral in locating and prosecuting cyber criminals, making dark web analytic techniques essential for investigators.

An important capability of dark web analytics is Dark Vendor Profiling (DVP), i.e. the collection and analysis of a dark vendor's characteristics for the purpose of establishing incriminating evidence against them and de-anonymizing their identity. Examples of such characteristics include vendor names, products they sell, countries they ship goods from, marketplaces they participate in, and alias accounts they control among others. The vast majority of the related work has been conducted using only text-based data scraped from dark forums and marketplaces. For example, many of the works rely on fingerprinting users' *writing styles* for the purpose of identifying their aliases. Thus, existing studies are missing some important hidden service artifacts which could lead to more incriminating evidence, especially for dark marketplaces that are image-based, rather than text-based. However, considering images in profiling is challenging due to several factors, namely, the availability of complete datasets, computational overhead, and relevant ethical considerations. In this paper, we aim to address these challenges to pave the road for efficient image-based DVP techniques.

In general, previous studies have avoided downloading image content due to ethical concerns, such as unintentionally accessing child pornography and other paraphernalia. In our study, we explore methods to *represent* image content without the need to view, download, or store images for DVP. Specifically, we examine the *metadata* of images and evaluate the effectiveness of storing *hashes* of dark marketplace images, rather than the image content. This avoids unintentional exposure to child obscenity and, at the same time, saves computational resources. The nature of hashing allows for large amounts of data to be represented in short character streams. Thus, it is a natural candidate for representing large amounts of image data without the need to store the actual image content. Classic hashing algorithms are also designed such that any small change in data results in a major change in the hash. Consequently, this work also examines several types of *image hashing* algorithms, which allow similar images to generate the same hash, even if resized, cropped, or filtered. In particular, our study analyzes images from 47 Tor-based dark marketplaces and compares four image hashing techniques to determine the most effective means to identify similar images within and between dark marketplaces, leading to DVP.

In addition, our study aims to support alias attribution, i.e. the correlation of several accounts belonging to the same vendor, using image-based data. According to Black Widow [12], a cyber intelligence gathering framework for dark web applications, there is substantial overlap between actors across dark forums, even if the forums are not based in the same language. Therefore, it is reasonable to suspect a similar overlap exists between dark marketplaces as well. Presumably, images have the potential to the identification of dark vendor aliases, since it is likely a vendor would use the same images to sell their product if they were participating on several dark marketplaces. Furthermore, images could assist in the development of incriminating evidence if they contain metadata concerning the image's author, date and time of creation, location, camera make and model, and more.

Since very few reliable datasets exist for the purpose of dark web analysis research, developing complete, reusable data is undeniably one of the largest roadblocks. For the purpose of this study, we started with a publicly available Darknet Market (DNM) Archive consisting of data scraped from 89 dark marketplaces from 2013-2015 [6]. Despite the author's warning of potential incompleteness of each crawl, the 1.6TB dataset has been used by several researchers in a variety of work. In an attempt to address the incompleteness issue, another goal of our study is to identify the most complete and useful set of dark marketplaces from the DNM Archive to support future dark marketplace analysis research.

In summary, this research directly supports intelligence and law enforcement communities' investigative efforts in DVP by offering the following contributions: we demonstrate how more effective DVP can be achieved by including image data in the analysis of dark marketplaces; we determine the most effective image hashing technique for the identification of images repeatedly used by vendors, leading to dark vendor alias attribution; and we identify the most complete and useful set of dark marketplaces from the well known DNM Archive.

The rest of this paper is organized as follows. Section 2 describes the anonymity techniques that allow cyber criminals to conduct illegal activity online. Section 3 explores related dark web analysis studies conducted and discusses their limitations. The methodology and experimental results of this work are discussed in Sections 4 and 5 respectively, followed by a discussion of limitations and future work in Section 6. Finally, Section 7 concludes this research.

## 2 BACKGROUND

In this section, Tor and its hidden service infrastructure is explained to illustrate the challenges in identifying owners, vendors, and users of dark marketplaces when anonymity techniques help to conceal user behavior, location, and other potentially incriminating data.

### 2.1 Tor: The Anonymous Network

Anonymity networks, which are known as *overlay networks*, use software solutions deployed on top of existing infrastructure, i.e. the Internet, to map virtual links between clients and services for the creation of new virtualized network infrastructures [10]. By far, the most prevalent anonymity network is The Onion Router, **Tor**, developed by The Tor Project, Inc. and initially released in 2002 [7]. Typical Tor connections are based on *circuits* of three relay nodes - namely, an entry node, a middle node, and an exit node. When preparing a stream of data to be sent down a circuit, a user will encrypt their data three times, using each relay's public key once. As the data is passed from the entry node, to the middle node, and to the exit node, a layer of encryption is removed at each hop. Finally, once the data has reached the circuit's exit node, the data is fully decrypted and passed to the destination node. This scheme allows anonymity for the user not only by performing several rounds of encryption but also by ensuring each node is only aware of its neighboring nodes in a circuit, i.e. no node is aware of the overall end-to-end communication, ensuring clients and services are never directly connected.

Additionally, when the Tor browser is used, little to no remnants of internet activity can be forensically recovered from the device. Specifically, forensic analysis may verify whether or not the Tor browser was installed on a client computer, but not if and when it was used, nor what it was used for [13]. It is important to note that the connection between a user and an anonymity network is **not** hidden in Tor. However, the user's location and the content of the communications within the network remain concealed. Most often, the user's traffic is delivered on shared bandwidth, making it even more difficult to distinguish between individual connections.

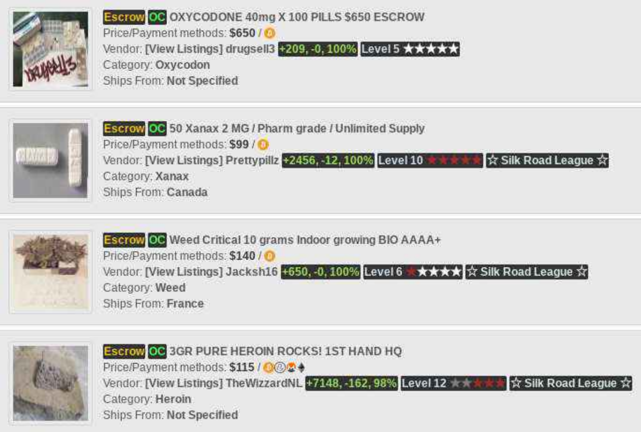
### 2.2 Hidden Services

Tor's most distinctive feature is its ability to provide *hidden services*, each of which are hosted with `.onion` addresses [7]. Tor hidden services enable users to host anonymous, theoretically untraceable websites by implementing additional security measures. This feature enables dark marketplaces to be hosted and dark vendors to conduct criminal activity. Unlike typical Tor network connections, connections to hidden service involve additional interactions with Introduction Points and Rendezvous Points and result in six total relay nodes: one entry, one middle, and one exit node for both the client and the service [10, 11].

With double-sided anonymity, both users and service providers are able to mask their identities, thereby disabling either party's ability to discover the other party's true location. For the anonymous community, this is a very attractive web hosting solution. In fact, it is estimated that 70,000-100,000 hidden services are running on the Tor network at any time. This statistic (and many others like it) are reported by the Tor Project [5] and accessible on their metrics portal [2].

As previously mentioned, the location of a hidden service is *theoretically* untraceable. However, many studies have challenged the design of the hidden service system in an attempt to de-anonymize their user base and owners. These studies will be further discussed in Section 3. Likewise, there are many cases where, leveraging *user error*, law enforcement has been able to successfully locate a criminal hidden server and prosecute the owner of the service subsequently.

One of the most notable such cases was that of the Silk Road anonymous marketplace take-down executed by the FBI and Europol in 2013. The Silk Road was a multi-million U.S dollar dark marketplace specialized in narcotics and controlled substances. Ultimately, the owner of this marketplace was identified by an FBI agent who was able to expose the owner's email address and full name. Despite the successful take-down, newer versions of the Silk Road became available through other hidden service operators, as shown in Figure 1. In fact, dozens, if not hundreds of similar dark marketplaces, such as the one in Figure 2, are available today, enabling the sale of paraphernalia and demonstrating the need for effective methodologies for targeting both dark marketplace administrators and vendors.


Escrow OC OXYCODONE 40mg X 100 PILLS $650 ESCROW
Price/Payment methods: $650 /
Vendor: [View Listings] drugsell3 +209, -0, 100% Level 5 ★★★★★
Category: Oxycodon
Ships From: Not Specified

Escrow OC 50 Xanax 2 MG / Pharm grade / Unlimited Supply
Price/Payment methods: $99 /
Vendor: [View Listings] Prettypillz +2456, -12, 100% Level 10 ★★★★★ ☆ Silk Road League ☆
Category: Xanax
Ships From: Canada

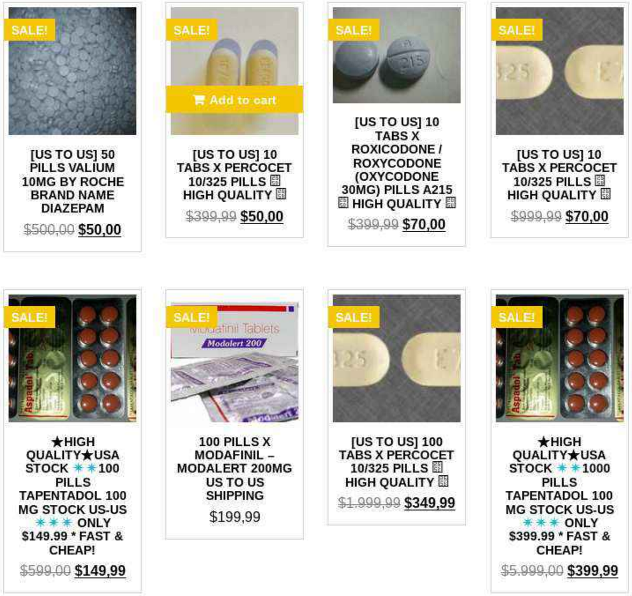
Escrow OC Weed Critical 10 grams Indoor growing BIO AAAA+
Price/Payment methods: $140 /
Vendor: [View Listings] Jacksh16 +650, -0, 100% Level 6 ★★★★★ ☆ Silk Road League ☆
Category: Weed
Ships From: France

Escrow OC 3GR PURE HEROIN ROCKS! 1ST HAND HQ
Price/Payment methods: $115 /
Vendor: [View Listings] TheWizzardNL +7148, -162, 98% Level 12 ★★★★★ ☆ Silk Road League ☆
Category: Heroin
Ships From: Not Specified

**Figure 1: Screenshot of Silk Road marketplace listings in October 2019 from silkroad7rn2puhj.onion.**

## 3 RELATED WORK

This section discusses various studies related to dark marketplace analysis and introduces the core concepts of each work along with how they relate to ours.



[US TO US] 50 PILLS VALIUM 10MG BY ROCHE BRAND NAME DIAZEPAM
~~$500,00~~ $50,00

[US TO US] 10 TABS X PERCOCET 10/325 PILLS ▯ HIGH QUALITY ▯
~~$399,99~~ $50,00

[US TO US] 10 TABS X ROXICODONE / ROXYCODONE (OXYCODONE 30MG) PILLS A215 ▯ HIGH QUALITY ▯
~~$399,99~~ $70,00

[US TO US] 10 TABS X PERCOCET 10/325 PILLS ▯ HIGH QUALITY ▯
~~$999,99~~ $70,00

★HIGH QUALITY★USA STOCK ✳✳100 PILLS TAPENTADOL 100 MG STOCK US-US ✳✳✳ ONLY $149.99 * FAST & CHEAP!
~~$599,00~~ $149,99

100 PILLS X MODAFINIL – MODALERT 200MG US TO US SHIPPING
$199,99

[US TO US] 100 TABS X PERCOCET 10/325 PILLS ▯ HIGH QUALITY ▯
~~$1.999,99~~ $349,99

★HIGH QUALITY★USA STOCK ✳✳1000 PILLS TAPENTADOL 100 MG STOCK US-US ✳✳✳ ONLY $399.99 * FAST & CHEAP!
~~$5.999,00~~ $399,99

**Figure 2: Screenshot of Quality King prescription pill listings in November 2019 from quality2ui4uooym.onion.**

### 3.1 Marketplace Analysis

Most current research analyzes hidden services in general, not specifically dark marketplaces. In this study, we focus on authorship analysis, which is two-fold. One part is *user attribution*, where the goal of an investigator is either to de-anonymize a given user based on their browsing behavior, traffic, or semantic styles. The other is *alias attribution*, where distinct online profiles are correlated between communities. This domain of work determines whether a user in one dark platform is that same user acting in another forum, marketplace, or other platform.

Although identities are protected in anonymous environments like Tor, users may leave traces of their textual identities in writing styles and through the nature of their participation in dark forum and marketplaces. Furthermore, they may be characterized by the number of forums, marketplaces, chatrooms and other users they are associated with, or by the time of day when they are active in anonymous network environments. To achieve user attribution, work in [8] considers scenarios where a set of suspects is known, and the task is to determine which of the suspects are responsible for the web activities related to a cybercrime. Relying on knowledge of a set of potential suspects, none of these works attempt to attribute a user based on dark marketplace activity data.

When individuals operate under a number of different accounts, they are considered to have *aliases*. Attributing these aliases is an important aspect in performing authorship analysis because it may assist in the identification of a user and potential prosecution of a cybercriminal. In an attempt to achieve alias attribution, Spitters et al. describe a methodology in which they analyze user profiles based on topic-independent features, such as length of text and words, use of function words, interpunction and shallow syntactic patterns, along with time-based features and character n-grams

[15]. Ultimately, this analysis is used to compute similarities between pseudo users and attribute aliases. The researchers achieved sufficient precision for small sets of forums and users, but underachieved in terms of recall, resulting in 25% and 45% in pseudo user set sizes of 177 and 25 respectively. Also, this model is not likely to be successful in marketplace settings where writing styles are less distinctive.

Another study [16] discusses the limitation of stylometric analysis in dark marketplaces. The success of text-based analysis, where users post rich and diverse text, is not possible in dark marketplaces since there is sparse text data available and the text usually reflects the product type instead of the vendor's writing style. Thus, the researchers proposed to fingerprint vendors by their *photographic* style instead, building a classifier to identify vendors with multiple accounts across several dark marketplaces based on high level features like object, scene, background, camera angle, and others. By considering image data, Wang et al. developed a highly accurate model to correlate vendor accounts. However, neither image metadata nor text-based data was considered in their classification, which might have improved their model's accuracy. Furthermore, their work relied on the content of the images. This requires substantial computing resources, which may not always be available in practice, especially in smaller, local investigative organizations.

A recent research was conducted using crypto-currency data to analyze dark marketplaces for buyer and seller de-anonymization[14]. In this work, Sima considers a single dark marketplace, *Valhalla*, and de-anonymizes users by analyzing scraped data against publicly available bitcoin data. This work resulted in an application that associated dark identities with related bitcoin addresses and transactions. However, since this research was based on a single dark marketplace, it does not address the specific dark vendor profiling objective of alias attribution.

As discussed, the existing work in marketplace analysis differ from our work as they are based on either text-based data, image*content*, or crypto-currency data, whereas our work focuses on image metadata and image hashing for alias attribution across various dark marketplaces.

### 3.2 Image Analysis

Researchers in [9] used the same DNM Archive as used in this study to evaluate the prevalence of metadata by collecting all jpeg image files from the archive and extracting their metadata. In total, authors Lisker and Rose observed 223,471 unique photos from the *entire* DNM Archive, 229 of which contained GPS information, defining unique photos by the image file name. Since the authors did not consider any other file formats other than jpeg, data that was embedded in base64 encoding and data that was saved in PNG format was excluded from analysis. Also, in some cases, images can share the same filename despite being present under different listings and marketplaces. Thus, it is likely that many images were disregarded that were not actual duplicates, but perceived as such. To alleviate these problems, our study defined duplicate images by the listings they belonged to, rather than their titles, and considered all image formats, resulting in 297,922 unique images from a *fraction* of the DNM Archive, 5,944 of which contained metadata. The results of the metadata analysis in our work will be presented in Section 5.

Another contribution of our study is a determination of the best image hashing technique for DVP applications. Image hashes have been considered and compared beforehand to evaluate robustness against image modifications such as changes in brightness, contrast, scaling, and more [4], resulting in a determination of how often image modifications resulted in different hashes compared to that of the original source image. Such studies report Perceptual Hashing, PHASH, as the most robust hashing algorithm. In contrast, the work described in this paper aims to determine the accuracy in image matching based on image hashing without purposefully modifying images. Additionally, this paper considers dark marketplace image data specifically. Our study concludes that PHASH is the most effective hashing algorithm for DVP as well, which will be further detailed in Section 5.

## 4 METHODOLOGY

In this section, the research approach is discussed in steps. All code supporting this work was written in Python version 2.7.15.

### 4.1 Dark Marketplace Dataset

The dataset pulled from the DNM Archive consists of 34.4 GB of zipped directories containing HTML, JavaScript, and styling files along with images. After extracting the zipped directories, 47 of the 74 download directories (totaling 667.6 GB) were determined to be useful for the DVP database (DVP DB). The remaining 27 directories (115 GB) were not processed into the DVP DB due to a number of factors including lack of image data or inconsistent HTML formatting which made organizing web scraped data into the DVP DB time consuming and futile. This is a limitation to this work since the DNM Archive used for this study was not considered in its entirety.

Each marketplace directory was then individually inspected for HTML formatting and directory hierarchy. Unfortunately, almost all marketplaces were scraped differently - some had photos saved into directories while others used base64 encoding in HTML, some organized listings into separate directories while others contained all listings under one directory, and so on. Each marketplace directory had to be processed with unique Python scripts. Furthermore, the amount of unique listings collected varied immensely between the marketplaces. For example, the Agora directory contained 116,858 unique listings, while the Bloomsfield directory contained only 8. Nonetheless, each marketplace directory was parsed to collect the same data per listing: *product ID, marketplace, product name, vendor name, scrape date,* and *image path.*

For base64 encoded images that were not originally stored in the directory, the Python scripts created a new PNG file in the current directory based on the base64 encoded data. Each image was further processed to check for metadata within the photos and hashed with each of the four different image hashing schemes considered in this study. Namely, the Python implementations of Average Hash (AHASH), Difference Hash (DHASH), Perception Hash (PHASH), and Wavelet Hash (WHASH) were considered for image hash analysis [1].

Any images or image paths that caused errors in processing, such as *incorrect file extension* and *file not found*, were not included in the DVP DB and therefore not considered for metadata and hash analysis.

## 4.2 Image Hash Analysis Approach

To compare the image hashing techniques, a hash analysis was conducted. The goal was to determine which hashing technique was most accurate in calculating identical hashes for similar or equivalent dark marketplace images. Presumably, if a hash provides enough data to effectively represent images for the purpose of image matching, then the hash would be sufficient for investigating vendor aliases, rather than requiring an image in its entirety.

To evaluate the accuracy of this approach, images were first grouped by hash and hash type, such that if there were 10 images that produced the same AHASH value, then they were considered to belong to a single group. For each group, a Structural Similarity Index Metric (SSIM) [3] was calculated between each possible pair of images belonging to the said group. The SSIM is a metric that determines the percentage of similarity between photos. The more alike two photos are, the closer their SSIM value will be to 1.0. The more different two photos are, the closer their SSIM value will be to 0. Therefore, by incorporating SSIM calculations in the hash analysis, we were able to quantify the level of accuracy in using image hashes for image matching in dark marketplaces.

To explain further, let us consider a group of 10 images to be a part of group *X* sharing the AHASH, *abc*. For the hash analysis, each image in group X was paired with each other, such that SSIM values were calculated between unique image pairs $X_1X_2$, $X_1X_3$, $X_1X_4$, etc. Thus, the 10 images from group X resulted in 45 image pairs and 45 SSIM values. The 45 SSIM values would then be averaged and stored into a database such that a single entry in the database would contain data regarding the hash, the hash type, the average SSIM, and the number of pairs considered in calculating the average SSIM for a particular group.

Finally, after calculating average SSIM values for each group of unique hashes and hash types, an overall *weighted* average SSIM was calculated for each of the four hash types. The weighted average SSIM provides each groups' SSIM value a weight determined by the number of image pairs used to calculate it. This way, groups with a large number of images more heavily influenced the overall weighted averaged SSIM compared to groups with a small number of images. Figure 3 summarizes the workflow of image-based analysis conducted in this study.

## 4.3 DVP Database

The DVP database (DVP DB) is a MySQL database organized into three tables, the first of which holds main listing data, such as listing ID, product ID, marketplace, product name, vendor, and scrape date. The second contains all image data, i.e. the image path, any metadata found, and the four hashes of the image. The final table is used to store all hash analysis data, i.e. the hash, hash type, number of images with said hash, the number of image pairs used in calculating the average SSIM, the number of pairs *not* used in the calculating of the average SSIM due to error, and the average SSIM for each hash. DVP DB contains no duplicate listings and was cleaned of any erroneous entries, such as having empty values for either product ID, marketplace, product name, vendor name, or scrape date.

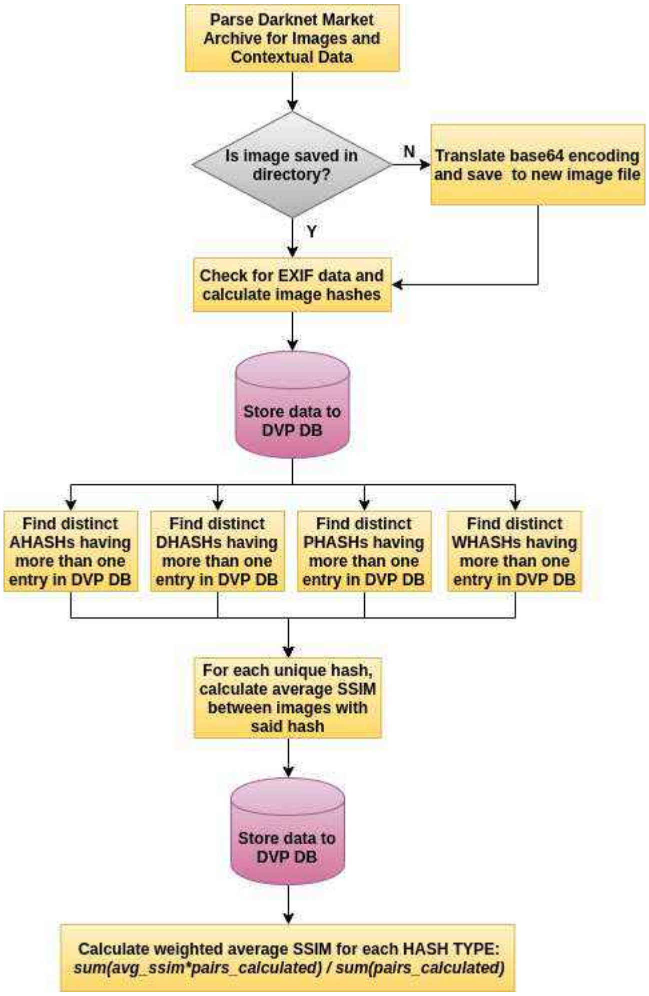

Parse Darknet Market Archive for Images and Contextual Data
Is image saved in directory?
N
Translate base64 encoding and save to new image file
Y
Check for EXIF data and calculate image hashes
Store data to DVP DB
Find distinct AHASHs having more than one entry in DVP DB
Find distinct DHASHs having more than one entry in DVP DB
Find distinct PHASHs having more than one entry in DVP DB
Find distinct WHASHs having more than one entry in DVP DB
For each unique hash, calculate average SSIM between images with said hash
Store data to DVP DB
Calculate weighted average SSIM for each HASH TYPE:
sum(avg_ssim*pairs_calculated) / sum(pairs_calculated)

**Figure 3: Workflow Summary of Image-Based Analysis for DVP.**

# 5 EXPERIMENTAL RESULTS

Overall, the DVP DB consisted of data from 47 marketplaces with 400,741 product listings, 297,922 images, and 10,712 unique vendor names from 391 scrape dates. This shows that the majority of dark marketplace listings incorporate image data (*approximately 75%*). The following section discusses the results for the three contributions of this research based on the experimentation with the DVP DB.

## 5.1 Top Dark Marketplaces

To determine the most meaningful marketplaces present in the DMN Archives, first each of the 47 marketplaces in DVP DB were ranked against each other using the following measures: number of listing entries, number of vendors, number of images, number of images with metadata, number of images with GPS coordinate data, proportion of images with metadata, and proportion of images with GPS coordinate data. Then, an average rank was calculated for each of the marketplaces using the seven aforementioned metrics, which denoted the overall significance of each marketplace in comparison with the others. Among all the dark marketplaces under study, *Agora* is found to be the most informative, with 116,858 entries and 64,535 images, 3.6% of which contained metadata. The remaining marketplaces are listed in Table 1 in order of their significance as determined by this study. This table can effectively serve as a reference for the most effective dark marketplaces present in the DMN Archives for future DVP research.

In addition, we also analyzed vendor names to determine the frequency of vendor names being shared across dark marketplaces. Interestingly, we found that names were frequently repeated across platforms, as shown in Table 2. While it is possible that individual vendors coincidentally shared the same name in separate marketplaces, this analysis supports the more likely presumption of dark marketplace overlap by multi-market vendors. Furthermore, this analysis supports the idea that image data used in conjunction with textual data can lead to more accurate DVP by verifying whether a repeated vendor name is simply a coincidence or a probable alias.

## 5.2 Image Metadata

Beside image content, image files also hold information relevant to the image's production such as data on camera settings, camera brand and model, time of creation, GPS location, image creator and more, which can be embedded into image files. Evidently, such image metadata has the potential to present investigators with a plethora of identifiable evidence, which may lead to the de-anonymization of dark vendors.

In our case, of the 297,922 image files stored in DVP DB, 5,944 were found to contain some metadata (2.0% of all images). Table 3 summarizes the top marketplaces containing the highest proportion of images with metadata. Interestingly, 37 of the 47 marketplaces contained 0 images with metadata, suggesting those sites purposefully scraped images of their metadata upon uploading to the site as a precaution.

DVP DB images were further analyzed for the presence of GPS data embedded in images which could greatly assist investigators in identifying the physical locations of dark vendors. Of all DVP DB images, 828 were found to contain GPS data specifically (0.28% of all images). Though only few images contained metadata and GPS location data, collecting such information can aid in dark vendor profiling and locating. For example, in the case that a dark vendor under alias A uploads a photo to marketplace A, where metadata is automatically scraped, and the same vendor under alias B uploads the same photo to marketplace B, where metadata is **not** automatically scraped, correlating alias A and alias B through DVP could generate evidence (such as physical location) against alias A that would not have been generated had the correlation not been made. Therefore, despite the presence of metadata being limited, its significance and impact can be extended when considered in conjunction with alias attribution.

## 5.3 Image Hashing

The last goal of this study is to find an effective way to represent image data so that images listed may be used to correlate vendors without having to save or view image content. The DVP DB resulted in 76,261 unique AHASHes, 85,651 DHASHes, 80,321 PHASHes, and 75,114 WHASHes. Out of these, over 40,000 images per hash type were repeated at least once in the database. By running a hash analysis on DVP DB images, PHASH was determined as the most effective solution for image hashing in dark web applications. The complete results of the hash analysis are listed below in Table 4 in order of weighted SSIM. Again, the weighted average SSIM takes into account the number of image pairs considered when calculating average SSIM values per unique hash, and is therefore a more accurate calculation of hash type reliability.

# 6 LIMITATIONS AND FUTURE WORK

The main limitation within this work is in the DVP DB. While parsing the DNM Archives, any anomalous files which caused processing errors for unknown reasons or did not match the expected HTML formatting were passed over and not included in image metadata analysis and hash analysis. In addition, the amount of data provided by each of the DNM Archive marketplace datasets varies due to inconsistent scrape dates and listing quantities. As a future work, a dark web crawler can be developed which scrapes marketplace listings more frequently and systematically such that data acquired is better balanced and more representative of existing dark marketplaces in the wild and their vendors.

We have several immediate research plans to follow up. The top 30 DNM Archive marketplaces identified in this study will be used to parse text-based data in addition to image-based data for the purpose of designing machine learning based classifying techniques for alias attribution. Also, PHASH's of images will be used to represent image content rather than saving the image content itself, thereby avoiding both (a) legal issues caused by downloading and possessing exploitative imagery and (b) storage and processing overhead.

# 7 CONCLUSION

Analyzing dark marketplaces is an imperative part of cyber criminal investigation and prosecution. However, due to the anonymous nature of the Tor network and hidden services, dark marketplace analysis is non-trivial. This research considers the task of alias attribution between dark vendors in dark marketplaces, i.e. Dark Vendor Profiling. Accurately determining alias vendors conducting business in multiple marketplaces will aid dark web investigations and lead to the de-anonymization of anonymous sellers of paraphernalia.

Whereas previous research relied on text-based content for alias attribution across hidden services, this research examines the availability and significance of image data for dark vendor profiling in Tor-based dark marketplaces specifically. Namely, this work determined the most informative dark marketplaces available from the public Darknet Market Archive dataset, evaluated the presence of metadata in dark marketplace images, and analyzed four image hashing techniques, leading to identifying Perceptual Hashing to be the most accurate technique for matching similar images

| Marketplace | # Listings | # Vendors | # Images | # W/ Meta | # W/ GPS | % W/ Meta | % W/ GPS |
|---|---|---|---|---|---|---|---|
| Agora | 116,858 | 3,154 | 64,535 | 2,292 | 214 | 3.55% | 0.33% |
| Blackbank Market | 12,852 | 905 | 10,565 | 2,086 | 470 | 19.74% | 4.45% |
| Evolution | 89,208 | 3,922 | 69,492 | 749 | 62 | 0.88% | 0.07% |
| Alphabay | 88,722 | 1,446 | 79,060 | 0 | 0 | 0% | 0% |
| Pandora | 15,223 | 516 | 15,066 | 11 | 0 | 0.07% | 0% |
| Tor Escrow | 958 | 185 | 866 | 241 | 43 | 27.83% | 4.97% |
| Abraxas | 16,641 | 432 | 11,979 | 0 | 0 | 0% | 0% |
| Tor Market | 1,502 | 200 | 817 | 230 | 14 | 28.15% | 1.71% |
| Cloudnine | 10,952 | 1,088 | 10,070 | 0 | 0 | 0% | 0% |
| Dream Market | 7,251 | 398 | 6,385 | 0 | 0 | 0% | 0% |
| Cryptomarket | 4,422 | 411 | 3,941 | 0 | 0 | 0% | 0% |
| Middle Earth | 6,650 | 359 | 6,167 | 0 | 0 | 0% | 0% |
| Andromeda | 3,054 | 237 | 2,947 | 0 | 0 | 0% | 0% |
| Bluesky | 2,400 | 213 | 2,089 | 0 | 0 | 0% | 0% |
| Oxygen | 2,212 | 257 | 2,012 | 0 | 0 | 0% | 0% |
| Freebay | 507 | 175 | 417 | 97 | 6 | 23.26% | 1.44% |
| Hydra | 2,282 | 166 | 2,240 | 0 | 0 | 0% | 0% |
| Cannabis Road 2 | 1,537 | 155 | 1,442 | 0 | 0 | 0% | 0% |
| Area51 | 489 | 74 | 479 | 89 | 6 | 18.58% | 1.25% |
| East India Company | 1,429 | 143 | 1,232 | 0 | 0 | 0% | 0% |
| The Real Deal | 981 | 82 | 873 | 0 | 0 | 0% | 0% |
| Black Services | 639 | 167 | 621 | 0 | 0 | 0% | 0% |
| The Marketplace | 823 | 124 | 584 | 0 | 0 | 0% | 0% |
| Amazon Dark | 199 | 41 | 190 | 57 | 5 | 30% | 2.63% |
| Haven | 741 | 74 | 704 | 0 | 0 | 0% | 0% |
| Darkbay | 538 | 124 | 533 | 0 | 0 | 0% | 0% |
| Cannabis Road 3 | 318 | 95 | 258 | 3 | 0 | 1.16% | 0% |
| Panacea | 461 | 21 | 459 | 0 | 0 | 0% | 0% |
| Silkstreet | 35 | 14 | 33 | 11 | 7 | 33.33% | 21.21% |
| Freemarket | 169 | 6 | 167 | 52 | 1 | 31.14% | 0.6% |
| Poseidon | 427 | 17 | 427 | 0 | 0 | 0% | 0% |
| Torbazaar | 383 | 27 | 332 | 0 | 0 | 0% | 0% |
| Tochka | 197 | 29 | 192 | 0 | 0 | 0% | 0% |
| 1776 | 171 | 37 | 170 | 0 | 0 | 0% | 0% |
| Darknet Heroes | 207 | 28 | 153 | 0 | 0 | 0% | 0% |
| Deepzon | 56 | 6 | 55 | 19 | 0 | 34.55% | 0% |
| Dogeroad | 112 | 28 | 100 | 0 | 0 | 0% | 0% |
| Underground Market | 143 | 20 | 112 | 0 | 0 | 0% | 0% |
| The Majestic Garden | 88 | 16 | 63 | 0 | 0 | 0% | 0% |
| Horizon | 44 | 11 | 44 | 2 | 0 | 4.55% | 0% |
| Cantina | 21 | 9 | 19 | 5 | 0 | 26.32% | 0% |
| Topix2 | 34 | 24 | 24 | 0 | 0 | 0% | 0% |
| Sheep | 8048 | 370 | 0 | 0 | 0 | 0% | 0% |
| Bloomsfield | 8 | 3 | 8 | 0 | 0 | 0% | 0% |
| White Rabbit | 313 | 62 | 0 | 0 | 0 | 0% | 0% |
| Kiss | 415 | 10 | 0 | 0 | 0 | 0% | 0% |
| Greyroad | 21 | 9 | 0 | 0 | 0 | 0% | 0% |

**Table 1: DVP DB Marketplaces listed in order of significance in DVP DB by calculating an average rank over seven characteristics: number of listing entries, number of vendors, number of images, number of images with metadata, number of images with GPS coordinate data, proportion of images with metadata, and proportion of images with GPS coordinate data.**

| Vendor Name | # Marketplaces | # Listings |
|---|---|---|
| mikehamer | 16 | 227 |
| bcdirect | 16 | 652 |
| blackhand | 14 | 1,227 |
| idealpills | 13 | 1,351 |
| theblossom | 13 | 180 |

**Table 2: Top five vendor names in DVP DB based on the number of distinct marketplace appearances.**

| Marketplace | % Images w/ Meta | % Images w/ GPS |
|---|---|---|
| deepzon | 34.55% | 0% |
| silkstreet | 33.33% | 21.21% |
| freemarket | 31.14% | 0.60% |
| amazondark | 30.0% | 2.632% |
| tormarket | 28.15% | 1.71% |

**Table 3: Top five marketplaces in DVP DB based on the proportion of images containing metadata.**

| Image Hash Type | Weighted Avg SSIM | Avg SSIM |
|---|---|---|
| PHASH | 0.991 | 0.986 |
| DHASH | 0.987 | 0.989 |
| AHASH | 0.881 | 0.976 |
| WHASH | 0.660 | 0.975 |

**Table 4: Hash analysis results for average, difference, perceptual, and wavelet hashing in order of weighted average SSIM values.**

between dark marketplace listings. This work helps future dark vendor research by identifying not only the list of marketplaces best suited for experimentation, but also identifying the image hashing technique best suited for dark web scraping.

This work supports our efforts toward multi-modal Dark Vendor Profiling using machine learning based classification techniques where text, image, and behavioral data will be considered for improved results.

## REFERENCES

[1] [n.d.]. ImageHash · PyPI. https://pypi.org/project/ImageHash/. (Accessed on 12/04/2019).

[2] [n.d.]. Onion Services – Tor Metrics. https://metrics.torproject.org/hidserv-dir-onions-seen.html. (Accessed on 12/04/2019).

[3] [n.d.]. SSIM: Structural Similarity Index | imatest. http://www.imatest.com/docs/ssim/. (Accessed on 12/11/2019).

[4] [n.d.]. Testing different image hash functions. https://content-blockchain.org/research/testing-different-image-hash-functions/

[5] [n.d.]. Tor Project | Anonymity Online. https://www.torproject.org/. (Accessed on 12/04/2019).

[6] Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Delyan Kratunov Sohhlz, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark Net Market archives, 2011-2015. https://www.gwern.net/DNM-archives. https://www.gwern.net/DNM-archives Accessed: 2019-10-21.

[7] Roger Dingledine, Nick Mathewson, and Paul Syverson. 2004. Tor: The Second-Generation Onion Router. In *IN PROCEEDINGS OF THE 13 TH USENIX SECURITY SYMPOSIUM*.

[8] Xiaoxi Fan, Kam-Pui Chow, and Fei Xu. 2014. Web user profiling and tracking based on behavior analysis. (Jan 2014). https://doi.org/10.5353/th_b5731085

[9] Paul Lisker and Michael Rose. 2017. Illuminating the Dark Web. https://medium.com/@roselisker/illuminating-the-dark-web-d088a9c80240

[10] Joao Marques. 2018. Tor: Hidden Service Intelligence Extraction.

[11] The Tor Project. [n.d.]. *Tor: Onion Service Protocol*. https://2019.www.torproject.org/docs/onion-services.html.en

[12] M. Schäfer, M. Fuchs, M. Strohmeier, M. Engel, M. Liechti, and V. Lenders. 2019. BlackWidow: Monitoring the Dark Web for Cyber Security Information. In *2019 11th International Conference on Cyber Conflict (CyCon)*, Vol. 900. 1–21. https://doi.org/10.23919/CYCON.2019.8756845

[13] Brett Shavers and John Bair. 2016. *Hiding Behind the Keyboard*. Chapter 2.

[14] Tomas Sima. 2018. *Darknet market analysis and user de-anonymization*. Master's thesis. Masaryk University Faculty of Informatics. An optional note.

[15] M. Spitters, F. Klaver, G. Koot, and M. v. Staalduinen. 2015. Authorship Analysis on Dark Marketplace Forums. In *2015 European Intelligence and Security Informatics Conference*. 1–8. https://doi.org/10.1109/EISIC.2015.47

[16] Xiangwen Wang, Peng Peng, Chun Wang, and Gang Wang. 2018. You Are Your Photographs: Detecting Multiple Identities of Vendors in the Darknet Marketplaces. In *Proceedings of the 2018 on Asia Conference on Computer and Communications Security (ASIACCS '18)*. ACM, New York, NY, USA, 431–442. https://doi.org/10.1145/3196494.3196529