The Cynicism of Modern Cybercrime: Automating the Analysis of Surface Web Marketplaces

Nikolaos Lykousas  
Data Centric  
Bucharest, Romania  
nlykousas@unipi.gr

Vasilios Koutsokostas  
Department of Informatics  
University of Piraeus  
Piraeus, Greece  
basilhskoutsokostas@gmail.com

Fran Casino  
Department of Computer Engineering & Mathematics  
Universitat Rovira i Virgili  
Tarragona, Spain  
franciscojose.casino@urv.cat

Constantinos Patsakis  
Department of Informatics  
University of Piraeus  
Piraeus, Greece  
kptask@unipi.gr

Abstract—Cybercrime is continuously growing in numbers and becoming more sophisticated. Currently, there are various monetisation and money laundering methods, creating a huge, underground economy worldwide. A clear indicator of these activities is online marketplaces which allow cybercriminals to trade their stolen assets and services. While traditionally, these marketplaces are available through the dark web, several of them have emerged in the surface web.

In this work, we perform a longitudinal analysis of a surface web marketplace. The information was collected through targeted web scraping that allowed us to identify hundreds of merchants’ profiles for the most widely used surface web marketplaces. In this regard, we discuss the products traded in these markets, their prices, their availability, and the exchange currency. This analysis is performed in an automated way through a machine learning-based pipeline, allowing us to quickly and accurately extract the needed information. The outcomes of our analysis evince that illegal practices are leveraged in surface marketplaces and that there are no effective mechanisms towards their takedown at the time of writing.

Index Terms—Cybercrime, Marketplaces, Data Leaks, Machine Learning, Topic Modelling

I. INTRODUCTION

With the continuous digitisation of procedures, services, and products, crime has shifted towards the same direction. As a result, almost every aspect of modern crime is facilitated by digital means, and consequently, almost every criminal investigation involves some sort of digital evidence. The above are the primary reasons why cybercrime has evolved into a multi-billion underground economy. Its economic impact is devastating [1], with the FBI estimating the losses to $3.5 billion only within the USA [2]. In fact, this continuous rise is so threatening that it has become the second most-concerning risk for global commerce over the next decade, according to the World Economic Forum [3]. Notably, the recent COVID-19 pandemic resulted in a huge spike in cybercrime activities [4, 5], making the problem even more thorny. It is indisputable that the introduction of cryptocurrencies has significantly facilitated illegal money flows. The transactions of many cryptocurrencies offer high inherent privacy guarantees; e.g., Monero, ZCash, or they are difficult to be traced due to the wallet anonymity; e.g., Bitcoins. The rise of the dark web has also boosted the underground economy, providing perpetrators with more advanced privacy guarantees.

While dark web markets are still the key stakeholders, when it comes to illegal trading, several surface web marketplaces have been repeatedly reported for trading leaked data [2]. A very interesting characteristic, in this case, is that despite the trading of illicit products, these marketplaces have a very open form, e.g., they do not require any registration to access them, and the ‘loot’ that is traded is advertised openly across the web.

A. Motivation and Contribution

Delinquent behaviour is unarguably a characteristic that can be observed in every human society. The extent of this behaviour, in terms of how many people exhibit it, and the harm that is caused by it define the ethics of society and its limits. However, the ethics on the Internet follow different rules [5].

Moreover, the dark web is considered the default place on the Internet in which such behaviours and actions flourish. Nonetheless, they are promoted in closed circles so that there is some “control” on who can access this information and retain the anonymity of the perpetrators. However, should this information be openly disseminated in public channels, it implies that the promoted behaviour is widely practised and is considered a norm by some groups.

In the past few years, there has been a significant increase in reported data leaks, online extortion schemes and credential trading. One of our initial research questions was whether such

1https://www.europol.europa.eu/newsroom/news/covid-19-sparks-upward-trend-in-cybercrime  
2https://www.forbes.com/sites/zakdoffman/2020/04/20/facebook-users-beware-hackers-just-sold-267-million-of-your-profiles-for-540/  
https://www.zdnet.com/article/hacker-selling-data-of-538-million-weibo-users/
actions are so widely performed that they can be observed on the surface web. In this regard, we wanted to check whether the perpetrators were using platforms of the surface web to advertise their “loot” and the existence of marketplaces in the surface web. Currently, there are several such marketplaces operating with similar functionality; however, this work is mainly focused on Shoppy (https://shoppy.gg/), which appears to have the most users and products at the time of writing. Nonetheless, similar illicit trends have been found in the rest of the surface web marketplaces. The goal of this work is to provide an overview of what is actually being sold in such a marketplace and leverage methods (e.g. machine learning) to automatically determine which are the illegal products and the main organisations affected. The main limitation in the automation of such a task is the lack of text. These sellers do not need to add a lot of text about what they trade in these marketplaces, and on many occasions, there are typos, abbreviations, and slang, posing even more issues in the analysis of the derived text. Further to the analysis of the traded products, we discuss the modus operandi of the sellers and some insight regarding the pricing of “big leaked data.”

The rest of this work is structured as follows. First, we provide an overview of the related work and a brief discussion of these marketplaces. Then, we detail our data collection methodology. In Section IV, we analyse the collected data to extract actual knowledge out of the short descriptions of the shops in an automated way. Finally, the article concludes by summarising our contributions and highlighting some ideas for future work.

II. RELATED WORK

In recent years, there have been multiple incidents of massive data breaches affecting a broad spectrum of online services and service providers, including retailers, payment processors, and government entities [6]. Malicious actors gain internal access to sensitive data sources and then acquire millions of credit and debit card details, user credentials, as well as sensitive data which can be used to identify individuals uniquely. The sheer quantity of data that can be acquired has given rise to a burgeoning market for actors who sell the information they obtain, e.g. hacking and other forms of data theft, to other users. Participants in these illegally acquired data markets leverage various communication and networking methods, enabling them to form communities and interaction mechanisms freely. The most prevalent forms of such marketplaces, as identified in the literature, are Internet forums and Internet-Relay-Chat (IRC) channels [7], [8]. In particular, forums have been shown to comprise the principal medium for cybercriminals to network, form communities, and operate online stolen data markets, despite numerous successful infiltrations by law enforcement agencies [9]. To a large extent, these marketplaces reside in the dark web, commonly behind Tor [10], and are referred to as “Darknet markets”. Darknet markets are popular among criminals since they enable them to anonymously trade illegal goods and services, extending well beyond stolen data. The latter was discussed by Thomas et al. [11] by pointing out the complex value chain of the underground market economy at scale. These marketplaces comprise the essential pillars of this global-scale cybercrime economy and thus have become the key information source for investigating the cybercriminal ecosystem. An extensive body of literature has explored the darknet marketplaces [12], the involved stakeholders and their communication patterns [13], and their modus operandi [14].

Of particular interest to researchers are the marketplaces dedicated to the sale of stolen personal and financial information, known as “carding forums”, where cybercriminals sell the artefacts of large-scale data breaches, often containing stolen financial information [7], [15], [16]. Subsequently, the compromised credit and debit card information enables malicious actors to commit crimes such as identity theft, financial fraud, and, most importantly, online money laundering [6], [17]. Moreover, due to their illicit and underground nature, carding forums and marketplaces are characterised by unique trading dynamics between vendors and sellers since the quality of merchandise and the identities of traders are unknown to potential buyers. In this regard, a number of works focus on untangling the mechanics of transactions in this particular kind of underground marketplaces [18], [19], [20], [21].

Identifying key players is essential when investigating emergent threats and developing efficient disruption strategies [22], in particular, considering the fact that members of such communities are characterised by cross-forum posting activity, which can be used to identify user roles based on the type of posts and their frequencies. The indicators of trustworthiness and reputation of a seller play a pivotal role in the sale of illicit services and stolen data through underground hacking forums and markets, as users are more likely to conduct business with sellers who hold reputable standing. As such, it has been shown that reputation status can be used to identify prominent players in illicit online marketplaces [18].

Apart from the complexity of its dynamics, the ecosystem of illegal online markets is characterised by an equally wide range of offerings, services and products relevant to various illicit topics, such as underground drug economies, data breaches, and cyber warfare. For instance, the work proposed in [23] focuses on malicious assets traded in hacker forums, such as hacking tools, rootkits and exploits, from which cyber threat intelligence can be distilled, and analyses the extracted data for predicting and mitigating cyber attacks. In [24], the authors provide both quantitative and qualitative categorisation of offerings in 17 different marketplaces. Their findings indicate the existence of both highly specialised products with respect to particular vendors and markets, as well as the cross-listing of products on multiple sites and nearly identical products for sale by multiple vendors. Nonetheless, the most prevalent categories are related to stolen credentials and information, extending beyond financial accounts. In this direction, Madarie et al. [25] examined how, a diverse set of outlets such as stolen credentials, are disseminated by malicious actors as “account dumps”. Their analysis revealed that
the illicit dissemination of stolen account credentials covers a broad spectrum of online services, and highlighted Pastebin\(^3\) as one of the main sites used to spread the information. Pastebin is a paste website intended for sharing plain text snippets, which is radically different compared to the typical marketplaces and forums due to the complete lack of structure. Interestingly enough, while stolen combinations of usernames and passwords for various online services were posted, data thieves used advertisements (embedded within the dumps) to establish communication with potential customers seeking access to financial or more sensitive data.

### III. Methodology

After thorough literature research, we observed a literature gap focusing on surface web marketplaces and an analysis of some of their activities, including automated methodologies for the extraction and processing of information to discover potentially malicious behaviour. First, we explored two well-known hacking-related forums, namely blackhatworld\(^4\) and cracked\(^5\) looking for indications of the emergence of marketplaces supporting illegal activities, similar to the deep web forums/marketplaces. In this regard, we observed that Shoppy\(^6\) was widely used in such platforms to monetise some of the reported activities. In fact, such activities were advertised. Therefore, we established a methodology to analyse what types of activities and products were being sold in Shoppy. Shoppy is a shop hosting service that provides the opportunity to individual vendors to sell their products, allows payments in different forms, and a set of APIs to, e.g., advertise one’s products in forums etc. A crucial difference between Shoppy and the underground marketplaces studied in the literature is that the former does not offer a centralised listing of the sold vendors and products. Each vendor obtains a unique URL where they can host their shop without providing any means for a user to look for similar shops of products offered by different vendors, a common feature in e-commerce platforms [26]. The decentralised architecture of Shoppy hinders the extraction of knowledge, and thus, our proposed data collection methodology specifically aims to discover shops associated with illicit offerings and services, given the context established by focusing on hacking-related forums.

The methodology that we adopted to address this challenge consists of several steps, as depicted in Figure 1. First, we crawled the blackhatworld and cracked forums, collecting usernames, as well as references to Shoppy accounts in post signatures. Given the size of these two communities, we specifically focused our crawling only on the “Marketplace” forums. To his end, we adopted the architecture of the Structure-driven Incremental Forum crawler (SiFo) [27], which enabled us to crawl data from the aforementioned forums. Nevertheless, we did not leverage user accounts that could potentially allow us to access even more content, restricted to authenticated users [28]. Next, we examined the extent to which the collected usernames and Shoppy account data could be correlated with existing shops in the Shoppy ecosystem. The data collection process lasted from March to April of 2020. We collected a total of 68,045 usernames, and Shoppy links from forum post signatures, 2,906 of which were linked to existing Shoppy shops at the time of crawling. The results are summarised in Table I. Notably, a large fraction of the links to Shoppy accounts found in post signatures, did not resolve to existing shops, indicating that accounts in Shoppy may be banned, deleted, or renamed.

![Fig. 1: Workflow of the steps followed in our research.](image)

| Source               | #    | Valid |
|----------------------|------|-------|
| blackhatworld - usernames | 24,658 | 827   |
| blackhatworld - signatures | 660   | 359   |
| cracked.to - usernames  | 41,890 | 1,230 |
| cracked.to - signatures | 837   | 490   |
| **Total (unique)**  | **64,726** | **2,906** |

**TABLE I: Collected usernames and Shoppy links.**

With the collected data, we used the open Shoppy API to retrieve all the information associated with these shops, including products, prices, and their corresponding metadata to create a curated dataset.

### IV. Data Exploration

In the following sections, we explore the Shoppy data in different steps. First, we provide a quantitative review of the collected dataset. Next, we detail our topic modelling approach and, finally, we leverage an exploratory analysis of a subset of the surface data.

#### A. Shoppy in Numbers

In this section, we provide a quantitative analysis of the collected Shoppy shops and advertised products, as well as highlight the particular behaviours of vendors. In total, our dataset contains 64,726 products advertised by 2,906 vendors. Shoppy provides vendors with the ability to categorise their products as accounts, services or files. The distribution of product categories in our dataset is provided in Table II. “Account” is the default category, which evidently dominates the other two by a large margin.

---

1https://pastebin.com/
2https://www.blackhatworld.com/
3https://cracked.to/
4https://shoppy.gg/
μ lognormal distribution (CDFs) of the number of items per shop (Fig 2a) and the product prices, in USD (Fig 2b). We can observe that, while around 40% of the shops have less than ten items listed, there exist shops with thousands of items. As seen in Fig 2b, the price distribution of products is remarkably well described by a lognormal distribution ($\mu = 1.6, \sigma = 1.58$), highlighting that the prices of approximately 62% of the products fall within a small range comprised between 1 to 10$. Moreover, the median price of Shopp products is 5 USD, and, as observed in our dataset, the prices can reach up to 10,000 USD.

To get a deeper insight into how different types of products are priced, focusing on possible outliers priced well above the median of 5 USD, we bin the products based on their price and we calculate the fractions of each product type in each bin in Figure 3. We can observe that while the lowest price bin is dominated by accounts, the fractions of services per bin follow a consistently increasing trend as the prices increase. In contrast, the relative representation of accounts is inversely proportional to the price, ultimately making services the predominant product category (approx. 70% of total) for the last bin reflecting the highest-priced offerings ($\geq 500\$). The fractions of the file type products, which as previously shown comprise only a small fraction of the total offerings, are generally sustained, accounting for less than 10% of the products in each bin. It is worth noting that our initial observation related to the use of default categories is reflected in Figure 3, which shows that, for instance, account products are well represented within all the range of possible prices. The latter behaviour seems quite unrealistic in a real and competitive market scenario and is further supported by the experiments performed in the next sections.

To investigate the high-priced services dominating the upper price bracket, we manually examined them and provided some illustrative examples in Table III. Evidently, these items are false products and rather contain information such as merchants’ terms of service, notes regarding provided shop feedback, support information, and links to Discord servers and Telegram channels maintained by the merchants. This behaviour has been highlighted in recent literature by arguing that the unregulated and anonymous nature of platforms such as Telegram and Discord, makes them the perfect habitats for scammers and cybercriminals [29], [30].

### B. Topic Modelling

In this section, we analyse the Shopp stores and elaborate a topic-based characterisation of the offered products by analysing their titles. To this end, we consider a statistical model, namely “topic model”, which is a method well suited to the study of high-level relationships between text documents. Specifically, we leverage Latent Dirichlet Allocation (LDA), a generative probabilistic model proposed in [31]. It comprises an endogenous NLP technique, which as highlighted in [32] “involves the use of machine-learning techniques to perform semantic analysis of a corpus by building structures that approximate concepts from a large set of documents” without relying on any external knowledge base. LDA, as the name implies, is a latent variable model in which each document is modelled as a finite mixture over an underlying set of topics. Each of these topics is characterised by a distribution over item properties (e.g. words). LDA assumes that these properties are exchangeable (i.e. ordering of words is ignored, as in many other “bag of words” approaches in text modelling) and that the properties of each document are observable (e.g. the words in each document are known). The word distribution for each topic and the topic distribution for each document are unobserved; they are learned from the data.

Since LDA is an unsupervised topic modelling method, there is no direct measure to identify the optimal number of topics to include in a model. In this sense, LDA assigns documents to different clusters of topics with certain probabilities (i.e. the number of clusters is defined with an integer number $k$ provided by the user), where these probabilities depend on the occurrence of words which are assumed to co-occur in documents belonging to the same topic (Dirichlet prior assumption). This exemplifies the main idea behind all unsupervised topic models, that language is organised by latent dimensions that actors may not even be aware of [33]. Researchers have recommended various approaches to establish the optimal $k$ (e.g. [34], [35], [36], [37], [38]). These approaches provide a good range of possible $k$ values that are mathematically plausible. However, according to [39], when topic modelling is used to identify themes and assist in interpretation (like in the present study), rather than to predict a knowable state or quantity, there is no statistical test for the optimal number of topics or the quality of a solution. A simple way to evaluate topic models is to look at the qualities of each topic and discern whether they are reasonable [33]. To the best of our knowledge, the topic coherence measure with the largest correlation to human interpretability is the $C_v$.
score defined in [37], which we also adopt in this study to establish the optimal number of topics.

In our setting, we consider the aggregate titles of the offered products in each of the 2906 shops in our dataset as a document. For training LDA models on the generated documents, we employed the implementation provided by Machine Learning for Language Toolkit (MALLEI)\(^7\). To obtain the most coherent topic model for our data, we considered the number of topics \((k)\) within the range from 5 to 50 with a step of 5 and trained the LDA models with 1,000 Gibbs sampling iterations and priors \(\alpha = 5/k, \beta = 0.01\). For each trained model, we compute the \(C_v(k)\) metric. This metric combines the indirect cosine measure with the normalised pointwise mutual information (PMI) and the boolean sliding window technique, to determine the number of optimal topic classes according to data distribution [40]. According to Figure 4, the value yielding the highest \(C_v\) corresponds to \(C_v(20) = 0.621\) and thus, we set the number of topics \(k\) to 20.

\(^7\)http://mallet.cs.umass.edu/
presented in [41].

To provide an insight on the products sold by the shops classified in each topic, Table V includes some indicative examples per topic, with respect to the number of topic-relevant terms contained in their titles.

The latter further allows us to characterise each one of the learned topics in a qualitative manner. Topics #1 and #2 describe “premium” accounts for a variety of online services and software products including streaming and VPN services. Topic #3 describes accounts associated with popular restaurants and fast food companies. Topic #4 reflects accounts associated with in-game items and collectibles for the popular online game Fortnite. This topic is found to be dominant in most shops, in comparison to the other topics, with 466 occurrences (i.e. 16% of all shops). Although selling game accounts can be perceived as an innocuous activity, provided the context of our data collection, these selling activities could be linked with money laundering schemes, based on the idea of converting stolen money to virtual currencies which are used to purchase in-game items [42], [43]. Topic #5 focuses on OpenBullet configurations. OpenBullet is a brute-forcing tool used for performing credential stuffing attacks against online services [44], which are described by configuration files “configs”, offering features such as checking multiple credentials simultaneously (advertised by metrics such as CPM, standing for “Checks Per Minute”) and bypassing rate-limiting. Topic #6 contains several classes associated with a broad spectrum of products ranging from game accounts to hacking and reconnaissance tools such as dorks. Topic #7 includes mainly subscriptions to various sports and video streaming services. Topic #8 highlights accounts, hacking tools and in-game items for the popular video game Minecraft. Topic #9 models the false products previously described (cf Table III), containing information regarding vendor’s terms of service and links to external Discord servers, Telegram channels, and etc. Topic #10 includes product licences and keys for a variety of software packages, games and operating systems. Topic #11 describes subscription plans for streaming services, similar to Topic #7. Topic #12 involves accounts for the popular game League of Legends. Topic #13 describes selling leaked user data from security breaches, in the form of combo lists, i.e. combinations of usernames/emails and passwords [45], which can be used for compromising accounts with the same credentials in other services, by means of credential stuffing attacks, as seen in Topic #5. Topic #14 involves mostly guides and e-books regarding carding and other methods of financial fraud. Topic #15 contains discount codes and accounts containing redeemable credits for various online shops and e-commerce platforms. Topic #16 is closely related to topics #14 and #15 and includes vouchers for online purchases in various venues as well as methods to perform a fraud or to scam sellers. Topic #17 mainly includes subscriptions for online services and products with a focus on mobile apps. Topic #18 is related to serial numbers for computer peripherals such as monitors, keyboards, etc. Topic #19 provides assorted “random” accounts for various social media and sites. Finally, Topic #20 is related to products such as redeemable gift cards, mainly for restaurants and food suppliers.

C. Use case analysis of surface data

In this section, we focus on the Topics #5 and #13, which as highlighted above, model products related to cybercriminal activities such as selling breached credential dumps and using tools for automating the compromise of accounts in different online services. To this end, we leverage the term-salience metric defined in [46], which given the set of representative terms per topic, ranks them according to their distinctiveness, i.e. how informative a specific term is for determining the generating topic, versus a randomly-selected term. Subsequently, we select the top-3 most salient terms for topics #5 (config, openbullet, capture) and #13 (combo, database, records), and we use them to query product titles, in order to identify the most prevalent products modelled by these topics. For Topic #13, we additionally include the term db which is a common abbreviation for the term database.

As previously reported (Table V), Topic #13 models leaked data from online data breaches, which are sold in the form of username/email and password combinations, along with other personal information. Such listings usually advertise the number of breached records, as well as the source of the leak. In Table VI we present some of the largest account dumps found in our Shoppy dataset, along with their prices. Indeed, we discovered that popular password breach checker platforms, such as https://haveibeenpwned.com, list the majority of the account database dumps sold on Shoppy. Moreover, this could explain the relatively low price tag for leaks, including up to millions of records, as the respective breaches have already been made public.

In Table VII, we list some illustrative products with titles including at least one of the selected salient terms for Topic #5. We observe that these products represent configurations for software such as OpenBullet8, BlackBullet9, SilverBullet10, Storm11. As previously stated, such tools can be used to automate credential stuffing attacks [47], versus various online services, as shown from the product titles. Sellers of such ‘configs’ often advertise features such as CPM (checks per minute) and capturing functionality offered, i.e. the ability to capture specific information associated with a compromised account, such as saved credit cards and payment methods, reward points, etc. Moreover, since OpenBullet and its variants are very configurable, they allow for modules which automate the use of proxies, can bypass CAPTCHAs etc.

From the above, we can largely infer the modus operandi of the account sellers of Shoppy and other cybercriminal markets. One is able to purchase massive quantities of breached credentials, and by exploiting the password reuse behaviour

8https://github.com/openbullet/openbullet
9https://redskylalliance.org/sindustry/blackbullet-credential-stuffing
10https://github.com/mohamm4dx/SilverBullet
11https://www.netacea.com/blog/storm-cracker-tool/
exhibited by many users [48], she could compromise users accounts with same credentials in other online services by using credential stuffing tools with different configurations. Moreover, depending on the credentials, the attacker could use them to penetrate systems as part of a more advanced attack, e.g. an APT group performing a targeted attack. Finally, it should be stressed that some of these advertised logs may be tokens that can be used to bypass two-factor authentication and/or compromise business email accounts.

V. DISCUSSION AND FINAL REMARKS

There are several conclusions that can be extracted from the analysis and the outcomes obtained in the previous sections. First and foremost, we found evidence of malicious activities which are usually taking place in the dark web, yet this time arising on the surface web. In this sense, the cynicism of malicious actors, who are perpetrating these activities, is covered by a lack of methodologies and takedown mechanisms, due to several factors such as, e.g. the decentralised nature of the marketplaces and the delays caused by inherent bureaucratic issues from cross-jurisdiction [49]. To the best of our knowledge, this is the first work that provides a solid and automated methodology to find, quantify and classify in a comprehensive way such activities. Nevertheless, despite the promising analysis leveraged in this article, malicious actors always find a way to circumvent analysis, since, e.g. they only use such platforms as a contact point, redirecting all of their activities to other external channels such as Telegram or Discord. Moreover, the use of technologies such as IPFS can augment the possibilities and resilience of such malicious practices [50], [51]. As a result, there are several challenges that may also arise regarding the collection of digital evidence [52].

Another dimension to be yet explored is the underlying connection between the activities reported in this article and further criminal campaigns. Therefore, despite the fact that most of the sold products can be classified as ‘soft’ cybercrime (i.e. passwords, credit card credentials, personal data) they can pose significant damage to individuals and businesses, and they may be just the tip of the iceberg. More concretely, money laundering and the financing of other, probably more dangerous activities, can be just happening in front of our eyes [53], [17], [54].

As previously stated, there exist several challenges for the analysis and takedown of the illegal activities being hosted on such platforms. The decentralised nature of, e.g. Shoppy avoids crawling mechanisms that could be used to collect all the stored information. Moreover, Shoppy is a resilient

| Topic | Key Terms | Documents |
|-------|-----------|-----------|
| 4     | skin, fortnite, trooper, black, knight, renegade, galaxy, raider, ghostrider, psn, linkable, aerial, random, assault, season, stack | 466 |
| 2     | spotify, premium, netflix, nordvpn, vpn, family, huhlu, pornhub, upgrade, nord, crunchyroll, owner, expire, bulk, uhld, disney, grammarly, screen, student, support | 344 |
| 9     | discord, month, paypal, service, bte, day, buy, week, server, term, read, bot, youtube, twitch, contact, apple, cheap, token, rust, music | 219 |
| 14    | method, amazon, free, balance, paypal, store, acc, guide, money, carding, make, google, ebook, check, work, ebay, refund, script, bank, link | 205 |
| 8     | access, full, minecraft, nfa, minecon, cape, sfa, hypixel, acces, optifine, semi, rank, roblox, unmigrated, vip, mpv, ufa, letter, elite, ninja | 197 |
| 10    | key, lifetime, steam, origin, game, uplay, window, license, edition, office, pro, battlefield, standard, global, year, crypto, ffa, deluxe, microsoft, home | 178 |
| 5     | config, capture, openbullet, checker, fast, cpm, full, ultra, package, proxy, api, onetap, instagram, aimware, master, mycanal, view, cfg, update, real | 153 |
| 11    | premium, full, disney, hbo, live, monthly, pass, plan, yearly, commercial, showtime, ad, nba, gold, annual, starz, tidal, fitbit, directv, espn | 135 |
| 13    | private, day, combo, pack, email, mail, pass, combolist, usa, fresh, access, valid, hit, list, user, domain, mixed, shopping, guarantee, database | 130 |
| 3     | account, random, gta, shop, dork, crack, site, move, box, source, red, shoppy, mixed, high, target, game, million, keyword, turbo | 128 |
| 1     | premium, year, subscription, vpn, auto, lifetime, renewal, adult, security, site, brazzers, monthly, device, avast, pro, membership, pornportal, unlimited, kaspersky, renew, method, follower, include, depop, credit, balance, subway, point, free, deliveroo, footasylum, pizza, voucher, refundable, wowcher, tesco, studio, attach, guide, disney | 119 |
| 7     | good, order, android, web, ios, pack, iptv, basic, complete, learn, facebook, website, deezer, test, video, theme, virtual, country, datum, tool | 96 |
| 12    | account, random, level, skin, champion, inactivate, euw, legend, champs, verify, league, valorant, unverified, lvl, eune, active, unverified, lol, region, verify | 77 |
| 7     | warranty, month, sport, sling, year, orange, blue, pro, dazn, extra, package, viaplay, hbo, pass, usa, nfl, directv, total, disney, xfinity | 77 |
| 15    | account, code, reward, pin, coupon, hellowordstock, stock, hotel, pizza, visa, starbucks, ulta, discount, invite, online, american, include, usd, optimum, canada | 73 |
| 20    | card, gift, giftcard, grill, restaurant, pizza, bar, juice, jamba, kitchen, pin, city, read, noodle, fat, deli, bbq, cafe, event, company | 68 |
| 19    | account, porn, paypal, bitcoin, random, pocentity, instagram, twitter, verify, playstation, unverified, uber, ride, automatic, cyclegear, apple, follower, perfect, lifescelester, likes | 65 |
| 18    | serial, number, gaming, logitech, mouse, wireless, steelseries, pro, monitor, lenovo, inch, model, keyboard, series, gen, notebook, dell, headset, prime, imei | 64 |

TABLE IV: Different topic classes and their corresponding key terms, sorted according to the number of documents found.
| Topic | Sample Products | Topic | Sample Products |
|-------|-----------------|-------|-----------------|
| 1     | Avast Premier Premium Security 2 Year 1 Device | 11    | Hulu — Premium — Plan - No Commercials, ShowTime, Live TV, HBO, Cinemax, STARZ, Entertainment Add-on. |
| 2     | Netflix Premium Accounts (UHD) | 12    | EUNE Verified Inactive (IRON) Level 30 Account Random Champs & Skins |
| 1     | HMA VPN — PREMIUM — AUTO RENEWAL — MONTHLY/YEARLY — KEY/ACCOUNT | 13    | EUW Level 30+ Verified [Silver Kayle] Account Random Champs & Skins |
| 1     | Spotify Premium Account + [Auto-renewal] | 14    | Instagram Random Account |
| 1     | Nord Vpn Premium — Expire - 2021 | 15    | EUW Level 30+ Verified [Silver Kayle] Account Random Champs & Skins |
| 2     | Netflix Premium Accounts (UHD) | 16    | Deliveroo Refundable- £30.00- £34.99 -48HR (Method Included) |
| 2     | Spotify Family Owner Premium | 17    | SkillShare.com Premium - 3 months warranty [Android/iOS/Web] |
| 3     | Jersey Mike’s Free Regular Sub - Wrap - Tub USA | 18    | Deezer [Android/iOS/Web] - 1 year warranty |
| 3     | Dominos 2 Free Pizza (USA) | 19    | Logitech PRO Wireless Gaming Mouse (Serial Number) |
| 3     | Buffalo Wild Wings (USA) (1500-2000pts) | 20    | Round Table Pizza $40 Giftcard + PIN |
| 4     | SPECIAL OFFER RENEGADE RAIDER+160SKINS | 10    | Disney+ — Bundle Monthly Plan Hulu, ESPN+ |
| 4     | Fortnite I Renegade Raider + Recon Expert + Black Knight | 11    | NBA League Pass Premium — Monthly Subscription — 1 Week Warranty |
| 4     | Fortnite account with 3 EPIC SKIN skull trooper + ikonik + the ace | 12    | EUNE Verified Inactive (IRON) Level 30 Account Random Champs & Skins |
| 5     | CONFIG INTERMARCHE FR API FULL CAPTURE ULTRA FAST CPM (socks4/5) | 13    | Instagram Random Account |
| 5     | Subway Config + Full Capture for OpenBullet (Fast CPM) | 14    | SeaSnap |
169

ACKNOWLEDGEMENTS

This work was supported by the European Commission under the Horizon Europe Programme, as part of the projects HEROES (Grant Agreement no. 101021801), and LAZARUS (Grant Agreement no. 101070303). This work was also supported by the European Commission (call ISFP-2020-AGTERFIN) as part of the CTC Project (Grant Agreement no. 830929). Fran Casino was supported by the Government of Catalonia with the Beatriu de Pinós programme (Grant No. 2020 BP 00035), and by AGAUR with the 2021-SGR-00111 project. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

REFERENCES

[1] D. S. Thomas, “Cybercrime losses: An examination of us manufacturing and the total economy,” 2020.
[2] I. C. C. C. (IC3), “2019 internet crime report,” https://pdf.ic3.gov/2019 IC3Report.pdf, 2019.
[3] W. E. Forum, “Wild wide web consequences of digital fragmentation,” https://reports.weforum.org/global-risks-report-2020/wild-wide-web/, 2020.
[4] H. S. Lalite, L. A. Shepherd, J. R. Nurse, A. Erola, G. Epiphanio, C. Maple, and X. Bellekens, “Cyber security in the age of covid-19: A timeline and analysis of cyber-crime and cyber-attacks during the pandemic,” Computers & Security, vol. 105, p. 102248, 2021.
[5] R. Spinello, Cyberethics: Morality and law in cyberspace. Jones & Bartlett Learning, 2010.
[6] D. Décary-Hérou and A. Leppiainen, “Criminals and signals: An assessment of criminal performance in the carding underworld,” Security Journal, vol. 29, no. 3, pp. 442–460, 2016.
[7] K. K. Peretti, “Data breaches: what the underground world of carding reveals,” Santa Clara Computer & High Tech. J., vol. 25, p. 375, 2008.

AUTHOR: FRAN CASINO

TABLE VI: Known breaches sold through Shoppy, as identified by Topic #13’s most salient terms.

claim. Finally, to raise awareness and highlight the relevance of our findings, we discussed the implications of our research, the current challenges and limitations, and proposed some measures to overcome them. Future work will focus on exploring similar marketplaces and trying to find correlations between different platforms in an automated way. Moreover, we plan to analyse the possible links between the activities leveraged in such marketplaces and cryptocurrencies, as well as other widely used financial platforms. Finally, measures to mitigate the impact of credential stuffing will be considered.

This work was supported by the European Commission under the Horizon Europe Programme, as part of the projects HEROES (Grant Agreement no. 101021801), and LAZARUS (Grant Agreement no. 101070303). This work was also supported by the European Commission (call ISFP-2020-AGTERFIN) as part of the CTC Project (Grant Agreement no. 830929). Fran Casino was supported by the Government of Catalonia with the Beatriu de Pinós programme (Grant No. 2020 BP 00035), and by AGAUR with the 2021-SGR-00111 project. The content of this article does not reflect the official opinion of the European Union. Responsibility for the information and views expressed therein lies entirely with the authors.

REFERENCES

[1] D. S. Thomas, “Cybercrime losses: An examination of us manufacturing and the total economy,” 2020.
[2] I. C. C. C. (IC3), “2019 internet crime report,” https://pdf.ic3.gov/2019 IC3Report.pdf, 2019.
[3] W. E. Forum, “Wild wide web consequences of digital fragmen-

TABLE VI: Known breaches sold through Shoppy, as identified by Topic #13’s most salient terms.

| Title | # Records | Price |
|-------|----------|-------|
| Combo List — 528M Yahoo.com | 528,000,000 | 400 |
| Combo List 376M Hotmail.com | 376,000,000 | 200 |
| Facebook - 267 Million Records Breach [FULL DB] | 267,000,000 | 500 |
| Combo List 258M Gmail.com | 258,000,000 | 250 |
| Zynga - 213 Million Records | 213,000,000 | 250 |
| Dubsmash Full Database — 162 million — hashed | 162,000,000 | 120 |
| Dubsmash - 162 Million Records (FULL SQL DB) | 162,000,000 | 35 |
| MyFitness Pal - 4.6GB (144 Million Records) | 144,000,000 | 25 |
| Xiaomi - 144 Million Records | 144,000,000 | 25 |
| Canvas - 137 Million Records | 137,000,000 | 80 |
| 111 MILLION-RECORD PEMBLANCL USA DATABASE COMBO LIST | 111,000,000 | 30 |
| MyHeritage - 92.2 Million Records | 92,200,000 | 75 |
| Houze - 49 Million Records | 49,000,000 | 300 |
| Facebook DB - 45 Million | 45,000,000 | 265 |
| Chegg - 29 Million Records (Dehashed) | 29,000,000 | 50 |
| Evony - Multiplayer Game - 28.7 Million + 13.8 Million Records | 28,700,000 | 24 |
| HauteLook - 28 Million Records (Full DB) | 28,000,000 | 100 |
| 24 Million LUMINATI PROXY DATABASE HQ [Email/Pass] Combo | 24,000,000 | 40 |
| YouNow - 18.2 Million Records | 18,200,000 | 50 |
| 8tracks - 18 Million Records | 18,000,000 | 44 |
| 500PX - Full DB [14.9 million records] | 14,900,000 | 250 |
| Dubsmash — 12 million lines — Private Combo | 12,000,000 | 25 |
| CouponMom / Armor Games - 11 Million Records | 11,000,000 | 41 |
| Cafepress *NEW* 11 Million Records | 11,000,000 | 41 |
| Bitly - 9.3 Million Records | 9,300,000 | 41 |
| BlankMediaGames - 7.6 Million Records Breach | 7,600,000 | 41 |
| GAMESTOP.COM Database — 7.6M UHQ LINES — 100K Splits — Mail:Pass — Good for EVERYTHING | 7,600,000 | 10 |
| StockX - Full 6.8 Million Records DB | 6,800,000 | 65 |
| SNAPCHAT.COM — DATABASE — LEAK — 4.6M LINES — PHONE NUMBERS, USERNAMES | 6,000,000 | 2 |
| 5.7M Facebook Profiles w/Email | 5,700,000 | 20 |
| Stronghold Kingdoms - 5.1 Million Records | 5,100,000 | 50 |
| 5M BITLY DATABASE HQ COMBOLIST (NetfliX, Hulu, Spotify, PSN, & More) | 5,000,000 | 20 |
| 4.6 Million Snapshat.com Databases Private Combos | 4,600,000 | 20 |
| Game Salau [Dehashed] - 1.8 Million Records | 1,800,000 | 20 |
| PRIVAT HQ 800 Yahoo.com USA Combolist | 800,000 | 20 |
| Hackers.nl (Dutch prostitution forum) - 291K Records | 291,000 | 50 |
| 240k Sbglobal.net Domain HQ private Combolist 100% | 240,000 | 8 |
| 225k Icloud.com Domain HQ Private Combolist 100% HQ | 225,000 | 12 |
| 211k Cox.net Domain HQ Combolist 100% Private Base | 211,000 | 12 |
| Contmama debashed db 209k mail:pass: exclusive naughtyamerica.com [PORN] Mail:Pass 114K Database | 209,000 | 700 |
| 110k Sbglobal.net Domain HQ private Combolist 100% | 110,000 | 8 |
| 100K ONLY YAHOO COMBOLIST - EMAIL-PASS | 100,000 | 3 |
| 100K ONLY HOTMAIL COMBOLIST - EMAIL-PASS | 100,000 | 3 |
| 54k Cox.net Valid Mail Access Private Combolist | 51,800 | 12 |
| 39k Cox.net Domain HQ Private Combolist | 39,000 | 7 |

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
TABLE VII: Indicative products modelled by Topic #5 in descending price order.

| Title                                                                 | Price |
|----------------------------------------------------------------------|-------|
| Custom Config for Openbullet                                         | 999   |
| New Nectar Card capture Site #2 config for Blackbullet                | 450   |
| nintendo captchaeless open bullet config                              | 250   |
| Spotify — Config [OpenBullet]                                         | 200   |
| Skout Dating Site OpenBullet config [Anon]                            | 100   |
| Badoo.com OpenBullet config (Fixed on 30.1.2020)                       | 100   |
| Rennaughty & Naughtdate Openbullet Configs(2 configs)                 | 100   |
| NINTENDO SWITCH CONFIG [WITH FULL GAME CAPTURE PAYMENT METHOD AND BALANCE] | 100   |
| Luminati.io OB config                                                 | 50    |
| [OPENBULLET] COOP UK CONFIG WITH FULL CAPTURE                         | 50    |
| CONFIG MYCANAL API FULL CAPTURE + CHECK MAIL ACCESS ULTRA FAST CPM    | 40    |
| PSN Captchaeless API (50K CPM) Config — Full Capture                 | 40    |
| Bitc.com BlackBullet Confuc with CAPTURE                              | 35    |
| Custom Config ( We code your Contig, Web applications, Scrapers, Bruteforcers, Everything related ) | 20    |
| [OPENBULLET] KrippKreme Config — With Detailed Captures              | 20    |
| [OPENBULLET] Grubhub Config — CAPTURES CC, PAYPAL, AND GC BALANCE     | 15    |
| NordvPN Config + Expiration Capture                                   | 15    |
| [OPENBULLET] Papa John’s Config — CAPTURES POINTS                     | 10    |
| CodeCademy API Checker — +2.6k CPM — Capture: Pro                    | 15    |
| Apple Valid Emails Checker By OPENBULLET GROUP                        | 15    |
| [CCShop] streetcc.pw . *lof for Open Bulletin [Capture Balance]       | 15    |
| Facebook Config Capture (check if its ad count or not) Very Fast      | 15    |
| [OB] Config ICams.com With Capture                                   | 10    |
| [OPENBULLET] McDonalds USA Conf — CAPTURES CC                         | 10    |
| SkinHub — +8k CPM — Captures: [Balance,RelBalance,Country,TotalWithdrawals,TotalDeposits] | 10    |
| [OPENBULLET] Wiki Mining BTC Config + Capture                         | 5     |
| [OB] PoArt With Capture Followers                                     | 5     |
| [openbullet] Shupc — [High CPM] Orders + Cards + Rewards Program Capture | 5     |
| [STORM] CONFIG NETFLIX + FULL CAPTURE [WORKING] & FAST                | 3     |
| [STORM] Hostconfig + Capture (Fast)                                   | 2     |

[8] V. Benjamin, W. Li, T. Holt, and H. Chen, “Exploring threats and vulnerabilities in hacker web: Forums, irc and carding shops,” in 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE. IEEE, 2015, pp. 85–90.
[9] M. Yip, N. Shadbolt, and C. Webber, “Why forums? an empirical analysis into the facilitating factors of carding forums,” in Proceedings of the 5th Annual ACM Web Science Conference. ACM, 2013, pp. 453–462.
[10] R. Dingledeine, N. Mathewson, and P. Syverson, “Tor: The second-generation onion router,” Naval Research Lab Washington DC, Tech. Rep., 2004.
[11] K. Thomas, D. Y. Huang, D. Y. Wang, E. Bursztein, C. Grier, T. Holt, C. Kruegel, D. McCoy, S. Savage, and G. Vigna, “Framing dependencies introduced by underground commoditization,” in 44th Annual Workshop on the Economics of Information Security, WEIS 2015, Delft, The Netherlands, 22-23 June, 2015.
[12] F. Whenger, “The dark net: Self-regulation dynamics of illegal online markets for identities and related services,” in 2011 European Intelligence and Security Informatics Conference, IEEE. IEEE, 2011, pp. 209–213.
[13] A. Hutchings and T. J. Holt, “A crime script analysis of the online stolen data market,” British Journal of Criminology, vol. 55, no. 3, pp. 596–614, 2015.
[14] X. Wang, P. Peng, C. Wang, and G. Wang, “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, J. Kim, G. Ahn, S. Kim, Y. Kim, J. López, and T. Kim, Eds. ACM, 2018, pp. 431–442.
[15] W. Li, H. Chen, and J. F. Nunamaker Jr, “Identifying and profiling key sellers in cyber carding community: Asecure text mining system,” Journal of Management Information Systems, vol. 33, no. 4, pp. 1059–1086, 2016.
[16] A. Haslebacher, J. Onolapolo, and G. Stringhini, “All your cards are belong to us: Understanding online carding forums,” in 2017 APWG Symposium on Electronic Crime Research (eCrime), IEEE. IEEE, 2017, pp. 41–51.
[17] A. Mikhaylov and R. Frank, “Cards, money and two hacking forums: An analysis of online money laundering schemes,” in 2016 European Intelligence and Security Informatics Conference (EISIC), IEEE. IEEE, 2016, pp. 80–83.
[18] M. Yip, C. Webber, and N. Shadbolt, “Trust among cybercriminals? carding forums, uncertainty and implications for policing,” Policing and Society, vol. 23, no. 4, pp. 516–539, 2013.
[19] A. Haslebacher, “Understanding online carding forums: On products, prices and sellers,” Ph.D. dissertation, Citeseer, 2015.
[20] T. J. Holt, O. Smirnova, and A. Hutchings, “Examining signals of trust in criminal markets online,” Journal of Cybersecurity, vol. 2, no. 2, pp. 137–145, 2016.
[21] S. Sundaresan, D. McCoy, S. Afroz, and V. Paxson, “Profiling underground merchants based on network behavior,” in 2016 APWG Symposium on Electronic Crime Research (eCrime), IEEE. IEEE, 2016, pp. 1–9.
[22] A. J. Park, R. Frank, A. Mikhaylov, and M. Thomson, “Hackers hedging bets: A cross-community analysis of three online hacking forums,” in 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), IEEE. IEEE, 2018, pp. 798–805.
[23] S. Santani, R. Chinn, H. Chen, and J. F. N. Jr., “Exploring emerging hacker assets and key hackers for proactive cyber threat intelligence,” Journal of Management Information Systems, vol. 34, no. 4, pp. 1023–1053, 2017.
[24] E. Marin, A. Diab, and P. Shakarian, “Product offerings in malicious hacker markets,” in 2016 IEEE Conference on Intelligence and Security Informatics (ISI), IEEE. Sep. 2016, pp. 187–189.
[25] R. Madarie, S. Ruiet, W. Steenbeek, and E. Kleemans, “Stolen account credentials: an empirical comparison of online dissemination on different platforms,” Journal of Crime and Justice, vol. 42, no. 5, pp. 551–568, 2019.
[26] J. B. Schafer, J. A. Konstan, and J. Riedl, “E-commerce recommendation applications,” Data mining and knowledge discovery, vol. 5, no. 1-2, pp. 115–153, 2001.
[27] M. Pavkovic and J. Protic, “Sinfo—structure-driven incremental forum crawler that optimizes user-generated content retrieval,” IEEE Access, vol. 7, pp. 126941–126961, 2019.
