A Methodology for Large-Scale Identification of Related Accounts in Underground Forums

José Cabrero-Holgueras, Sergio Pastrana

Universidad Carlos III de Madrid, Avenida de la Universidad 30, Leganes, 28911 ,Madrid, Spain

Abstract

Underground forums allow users to interact with communities focused on illicit activities. They serve as an entry point for actors interested in deviant and criminal topics. Due to the pseudo-anonymity provided, they have become improvised marketplaces for trading illegal products and services, including those used to conduct cyberattacks. Thus, these forums are an important data source for threat intelligence analysts and law enforcement. The use of multiple accounts is forbidden in most forums since these are mostly used for malicious purposes. Still, this is a common practice. Being able to identify an actor or gang behind multiple accounts allows for proper attribution in online investigations, and also to design intervention mechanisms for illegal activities. Existing solutions for multi-account detection either require ground truth data to conduct supervised classification or use manual approaches. In this work, we propose a methodology for the large-scale identification of related accounts in underground forums. These accounts are similar according to the distinctive content posted, and thus are likely to belong to the same actor or group. The methodology applies to various domains and leverages distinctive artefacts and personal information left online by the users. We provide experimental results on a large dataset comprising more than 1.1M user accounts from 15 different forums. We show how this methodology, combined with existing approaches commonly used in social media forensics, can assist with and improve online investigations.

Keywords: Social Media Forensics, Underground Forums, Large-Scale Measurement, Related Accounts, Cybercrime

1. Introduction

Nowadays, the Internet is present in many areas of our society. Such digital transformation also allows for illegal and illicit activities, which frequently flourish in online underground communities. Some of the factors that favor these activities are the anonymity, the difficulty for prosecution, the high-reward to low-investment ratio, and the simplicity to acquire and use tools to carry on malicious actions [1, 2]. In this regard, underground forums and markets play a key role, as they are one of the main platforms where these communities interact [3, 4].

Underground communities can be divided into two main categories, depending on the types of goods and services traded, the anonymity factor, and the difficulty to reach and join them [5]. On the one hand, markets and forums on the Dark Web specialize in the exchange and discussion related to illegal goods (e.g., drugs, weapons, or child abuse material). Operators and users of these communities wary of their privacy, and thus these are often accessed through anonymous networks or *darknets* such as Tor or I2P. Some of these forums are private and can only be accessed by invitation [6]. On the other hand, other underground forums operate in the regular, surface web (i.e., they are indexed by popular search engines), and are easier to access. In these communities, the illegality of the materials is relatively lower and usually disguised. For example, malware known as *RAT* (which can either stand for "Remote Access Tool" or "Remote Access Trojan"), which is usually packed as binary trojans, is offered as an auxiliary administration tool that allows users to control their computers remotely [7, 8]. Also, *stresser* services, allegedly used to test the network resistance to large traffic loads, are used to perform Distributed Denial of Service Attacks (DDoS) [9]. Access to these communities is open to anyone, though sometimes it requires a registered account [10, 11]. Also, members in these forums might be less concerned about their privacy, and indeed sometimes they post personal credentials, like their personal e-mail addresses or Skype handles [12]. Still, some of these forums have been linked to high-profile cyberattacks, like the DDoS attacks on October 2016 against a major DNS provider that affected large companies like Twitter, Amazon or the New York Times [13, 14].

Motivation. According to their Terms of Service, most underground forums forbid the use of multiple accounts, unless they are used to dispute the banning of the original, in which case the link between both original and alternate accounts must be explicit. However, users can create and use more than one account to take advantage of each other. For example, they could use those accounts to game the reputation system[1], [18]. Also, by increasing and decreasing the price of goods, product

[1] In general, trust among users is highly dependent on reputation. Given the lack of ruling authority, this reputation system has a distributed model where each user can increase or decrease the reputation of others, and users are typically more inclined to acquire goods from trusted sellers [15, 16, 17].

sellers influence the market causing competitors to leave the marketplace [17]. Another motivation to create additional accounts is to diversify their activities (e.g., by opening accounts on multiple forums, being used to offer the same products or services). This way they can keep their market operative even if the accounts are banned, or the reputation in some forum suffers a degradation. For those forums that relax their terms with respect to the use of multiple-accounts, this fraudulent activity is also prosecuted and forbidden. Finally, from the prosecution point of view, especially if the activities involve illegal actions, users might create multiple accounts to thwart law enforcement investigations and prevent attribution [19].

Various works in the literature deal with the detection of Sybil attacks in social media [20, 17, 21, 22]. In these attacks, miscreants create and use multiple accounts to commit malicious actions by automatic means (i.e., bots). These accounts can be distinguished by patterns in their social relations [20] or metadata [23]. Detection of bot-based related accounts assumes that they interact with each other in isolated clusters, with few links to regular-user clusters in the social network [20]. However, the pursued goal in this research is to detect highly interacting accounts typically operated by humans, where the assumptions for bot-based detection do not hold.

Previous works have proved the efficacy of using text analysis for multi-account detection in social media [18, 24, 25]. However, these methods have two main drawbacks. First, they rely on ground truth data to fine-tune the detection methods or to conduct supervised classification. Second, they suffer from scalability issues. For example, the efficacy and effectiveness of Doppelgänger-finder (a state-of-the-art stylometry analysis tool [18]) decreases when more than 50 pairs of accounts are analyzed simultaneously [24].

Due to the aforementioned limitations in the detection of related accounts, current efforts use simple strategies (e.g., relying on previously known information [4], or assuming that users use the same pseudonym across forums [26]). These strategies do not work for users willing to hide their double identity [18]. Moreover, analyzing data from underground forums involves processing hundreds of thousands of accounts, out of which only a few might be of interest.

As a motivating example, the operator of Silk Road (a now-defunct dark market) was de-anonymized once he posted his personal email in a bitcoin forum, using the same pseudonym as in a chat room where a few months after he was advertising the market [27]. While this was paramount to start collecting further evidence and to prosecute the suspect, the analyst (a tax investigator) claimed that "*he had spent much of the weekend [..] scrolling through arcane chat rooms and old blog posts*". Thus, providing automatic tools capable of processing such large datasets is of great benefit to assist analysts during their investigations (e.g., to detect actors that post the same information using different pseudonyms and online accounts).

Contributions. In this work, we present a methodology to identify *related accounts* in underground forums at scale. The scope of our methodology is for accounts that post the same *distinctive* features. These features are infrequent in the posts made by these accounts and can be used to characterize the actors. The selection of features relies on heuristics and expert knowledge related to credentials and characteristic information publicly posted by users, such as Skype users, emails, or IP addresses. Our methodology first applies a feature extraction process to automatically gather and sanitize features from the raw posts written in the forums. Second, it performs data reduction by removing features appearing with high frequency or which are meaningless. Finally, it leverages the coincidences in the remaining features to compute a similarity score for each pair of users. Concretely, we propose a new metric called *Multi-Feature Similarity (MultFS)* score (§3.4), which takes into account both the number of coincidences between pair of accounts and the rarity of these in the dataset.

We have applied our methodology to a large dataset composed by data from different forums. We extract pairs of users who share relations between them, both in the same forum and across different forums. Related accounts do not necessarily belong to the same person, and in Section 4 we show how the proposed methodology can be combined with existing techniques to assist in online investigations for authorship attribution. Due to the lack of ground truth for validating our results, we apply stylometry analysis to a subset of accounts and also conduct manual analysis. Finally, we provide some case studies resulted from our manual validation, which shows the potential benefits of the proposed methodology to study the use of multiple accounts in underground forums. In summary, we present the following contributions.

1. We propose a methodology to derive relationships between forum users at scale. The methodology relies on information publicly posted by users and that, either alone or in combination, characterize them uniquely. For example, such information include IP addresses, email addresses, cryptocurrency wallets, or Skype handles (§3.1).
2. We define novel metrics to compare pairs of accounts using all extracted features. These metrics consider the rarity (i.e., how prevalent a feature is in the entire dataset) and uniqueness (i.e., how relevant a feature is for a given pair of accounts) of the features. To deal with large datasets, we propose a technique to compare users efficiently by applying vectorization techniques (§3.3). We then propose the MultFS score, which aggregates the various metrics into a single value (§3.4).
3. We apply the methodology on a large dataset of more than 56M posts made by 1.1M accounts in 15 different underground forums (§4). We first analyze the performance speed-up of our methods against a baseline approach. Second, we combine our methodology with Social Network Analysis and looking at username similarities to investigate whether pairs of accounts belong to the same user or not. Third, we validate our results by conducting stylometry analysis on a subset of pairs of related accounts and by manually analyzing some of these. Finally, we describe case studies on some interesting pairs of accounts, which undercover different relationships between users, like commercial partnerships for selling proxies, or accounts being victims of *doxing* attacks, where

the user behind the account is de-anonymized and its personal information (e.g., home address, family details, etc.,) are exposed publicly [28].

We conclude with the discussions of the limitations, potential applications, and ethical issues in §5. Finally, to foster research on this area and to allow for reproducibility, we release open-source the prototype used in our experiments.[2]

2. Background and Related Work

In the last years underground forums have become a place of interest for researchers for several reasons. First, they are an interesting source to study cybercrime activities, such as software or services used for cyberattacks [3, 29, 6, 30, 1], or the social aspects of the offenders [16, 12, 31]. Additionally, some actors prosecuted by law enforcement were active members of these forums [4]. For example, it has been demonstrated that information related to attacks on critical infrastructure was shared in online forums before the actual attacks took place [29]. Various economic and social factors influence the activities in these forums [6]. For example, various authors have analyzed the role of trust and reputation systems in online communities [16, 15]. Additionally, manual analysis by individual security researchers unmasked actors in underground forums related to banking malware [19], or the authors behind the well-known Mirai botnet [14]. While these analyses are insightful and can be used to understand the type of activities carried out in underground communities, the methods do not scale and are hardly applicable to other cases. Moreover, the lack of ground truth about users and the limitations of acquiring data from underground forums, thwart large-scale cybercrime research [10, 31].

Previous works have addressed the detection of bots in social media [20, 32, 17, 21, 22]. However, the problem addressed in our work is different since we do not assume that related accounts are (only) from bots, but from actual humans, and thus they might resemble actual human behavior. Stylometry analysis has been previously used to identify social media accounts operated by the same user [18, 24, 25, 33]. The seminal work by Afroz et al. proposed Doppelgänger-finder, a stylometry analysis tool focused on detection of duplicate accounts from the same user in underground forums [18]. This tool provides a probability score of each pair of accounts being from the same user, and it has been successfully applied to other domains such as Blogs or Social Networks [24]. One limitation of this technique is that it does not scale well to large datasets (e.g., it is computationally expensive for more than 50 accounts [24]), and also requires ground truth to establish the thresholds. Zhang et al. combined stylometry analysis with image analysis and 'Attributed features' (i.e., PGP keys, username, and contact information) to detect multiple accounts from the same vendors in Darknet Markets [33] and applied to medium-size datasets of 5.4k vendor accounts. Also, the authors relied on contact information from the profile pages. However, users willing to hide their double identities are likely to change such information. Tai et al. used similar information and a Machine Learning classifier to detect duplicate vendor accounts in adversarial settings, i.e., considering that the users behind the accounts attempt to evade their linkage [34] While these approaches are similar to ours, authors rely on personal information commonly found in markets (e.g., pictures or PGP keys) but not necessarily in forums.

Tsikerdekis et al. used non-verbal features to detect multi-accounts in Wikipedia [35]. The approach relied on metadata left by the users, like the number of revisions done, the bytes added/removed or the time elapsed between these. As the author state, one of the challenges in this approach is the identification of non-verbal variables that need to be considered, and that these might require changing the implementation or design of the proposed method (which is only adapted to Wikipedia). In our work, we have designed the feature selection and extraction as an independent process from the rest of the methodology, thus being more flexible to adapt to other online media sources.

Other works have also relied on characteristic features posted online by users to detect accounts in social media. Sundaresan et al. extracted Skype handles from public posts and translated those to their actual IP address to characterize the location of users in underground forums [12]. Gharibshah et al. presented a cross-correlation between the IP addresses that users post and the database from VirusTotal to understand and characterize malicious users [36]. Egele et al. modeled message characteristics on Social Networks to detect compromised (hacked) accounts [37]. Similarly, Mariconti et al. presented the reuse of usernames and identifiers in several well-known internet forums for certain malicious activities [38]. These works have motivated our initial election of features to link accounts.

Finally, the high amount of data extracted from online communities (e.g., social networks or forums) hinders manual analysis. Edwards et al. surveyed the usage of automatic data mining and machine learning models for law enforcement, noting that one of the key problems is the lack of reliable datasets and ground truth [39]. Similarly, approaches such as the one exposed by Nunes et al. provide extensive use of large-scale analytics to understand potential threats [30]. Overdorf et al. used Machine Learning models to guess private relationships between users [40]. However, authors used leaked datasets, which are not always available and, its use raises ethical concerns [41].

3. Methodology

The proposed methodology consists of four main steps (see Figure 1). First it performs feature extraction and selection from the raw content, tracking the features posted by each user account. In this step, the investigator must select the features to extract, and to create the corresponding methods (e.g., regular expressions) to extract these from the data. In §3.1 we detail the selection process conducted in our analysis. Second, to improve performance and reduce noise, the methodology allows to automatically conduct data reduction (e.g., removing accounts

[2] `https://github.com/anonymous-png/MultFS.git`

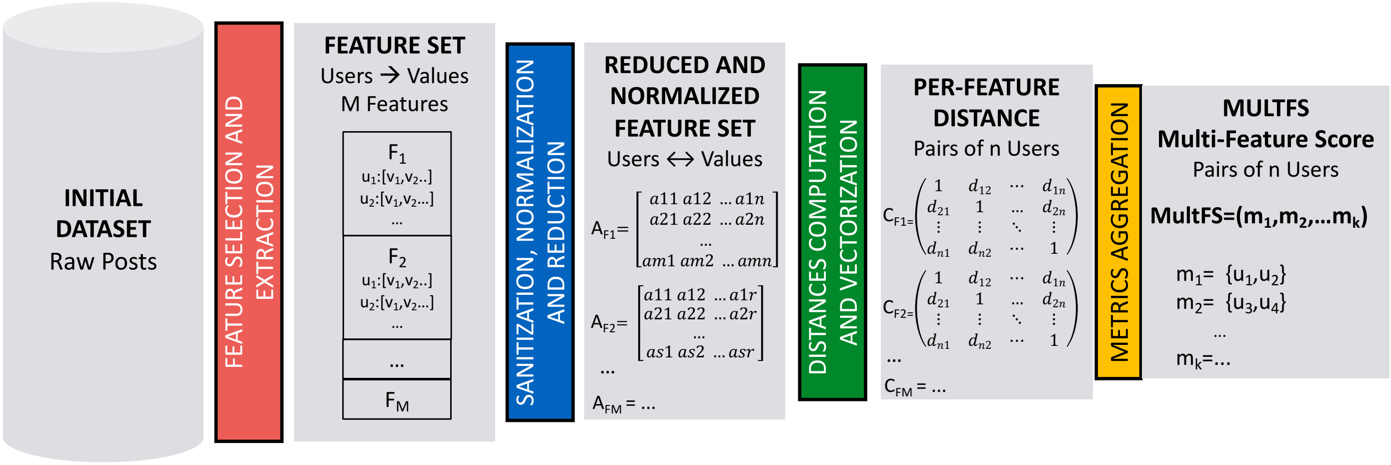

Figure 1: Scheme of the methodology used to extract pairs of similar user accounts.

that do not provide enough information to analyze similarities) and normalization (§3.2). Third, for each feature, and each pair of accounts, a distance metric is computed that considers: i) the relevance of each of the feature values (i.e., number of times that a pair of users share the same value), and ii) the relevance of each of these values for each pair of accounts, which represents 'how unique' these values are for the accounts and in the dataset. (§3.3). Finally, it aggregates the similarity metrics of the various features into a single similarity score dubbed *Multi-Feature Similarity*, or *MultFS*, which indicates the actual similarity of each pair of accounts (§3.4). The higher the score, the more similar these accounts are. A key aspect of this methodology is that it prevents computing expensive operations and has been designed as a flexible framework that can be executed in other communities to discover related accounts. We note that our methodology is fully automated, which facilitates its application in real settings. An online investigator would only need to adjust the feature extraction step, i.e., to select which features need to be extracted, and to prepare regular expressions to extract them from the text (e.g., forum posts). One these features are extracted and linked to the users that shared them, the rest can be applied in a straightforward manner.

3.1. Feature Selection and Extraction

Data contained in underground forums are highly heterogeneous and unstructured. The first step of the methodology is to choose and extract meaningful features from the raw content, employing regular expressions. Since the focus of the methodology is to find related accounts, the features selected must characterize almost uniquely the account from which it was posted, e.g., user credentials, like emails or Skype handles, and other identifiable information, like IPs or BTC wallet addresses. The feature selection is a process that should be adapted to each particular scenario. Concretely in this work, we extract the following features: links (in form of URLs), Skype handles, email addresses, IP addresses, and cryptocurrency wallets (concretely, Bitcoin identifiers, or BTC). The rationale behind these is that in many cases, the analyzed forums are used as actual marketplaces. In those, BTC addresses represent the financial information, and, emails or Skype are employed as personal information for contact with sellers. We note that all the previous features are language-agnostic, meaning that they are useful to relate accounts independently of the language used. Additionally, we extract all the *trigrams* (i.e., a sequence of 3 characters) from the content posted by each user, which allows us to account for the writing stylometry.

The goal in this step is to create a mapping of users to the values of these features. As described in §3.3, the metrics used to score account similarities rely on the coincidences and the relevance of the values posted for each of the selected features. Below, we motivate our choice and describe the extraction process the features used in our work. We note that the proposed methodology can be extended with additional features and requires specific knowledge of the target dataset.[3]

3.1.1. Links

Underground forums have become improvised marketplaces and are used to promote and advertise various services and products. Typically, forum users post links to external hosting services, e.g., to share pictures or videos. Forum users also post links to share information from other personal sites (e.g., blogs or social networks). Additionally, sellers provide links to their products in third-party online shops, or even to their e-commerce websites. Moreover, to make their product appealing, most sellers provide banners in the form of pictures or videos, which in turn are links to external sites. Therefore, links (as URLs) can associate related accounts. However, while links represent the behavior of users, there are certain links that, due to being widely used, are not valid to relate two accounts (e.g., referring other users to the Kali Linux download page). Based on this premise, in Section 3.2, we carry out several sanitization and cleanup processes that eliminate these values.

[3] For example, investigation on forums related to video-game hack and cheats could include Steam accounts, a popular video game distribution service.

The extraction of links is based on regular expressions We note that users might use anti-analysis techniques in which they obfuscate the links to avoid the detection of those links, e.g., by replacing *'.'* characters with the *'dot'* keyword or the *'t'* by *'x'* in HTTP. While we have not covered these evasion techniques in our implementation, these could be easily integrated into our methodology, e.g., by using heuristics to normalize the URLs.

3.1.2. Skype

Skype identifiers are commonly used in underground forums. Typically, these are used in two contexts: to negotiate the trading (e.g., the price or conditions) or to establish external relationships (e.g., for partnership or out-of-forum discussions). In the former case, it is a common practice for sellers to provide a Skype account for interested users, which allows them to contact and solve doubts in a more direct manner. In the latter, we observe users posting their accounts to create or join groups with people about specific topics, or to engage with conversations in topics of their interest. In both cases, two accounts posting the same Skype handle imply that users behind these accounts are the same or at least belong to the same group, and thus are of interest for our study.

To extract Skype accounts, we follow instructions specified in [42] to generate a Skype regular expression parser. Usernames contain 6 to 32 characters, which can be letters (lowercase), numbers, commas, periods, dashes, and underscores. Additionally, in case that the account is created from a Non-Microsoft mail service, the username contains the *"live:"* prefix. Unfortunately, applying these rules to the forum data returns a high volume of matching strings which are false positives. To prevent such situation, we only consider the structure *"Skype: <username>"* (where "<username"> is obtained by means of the regular expression). This is the most common way in which members write their Skype accounts (e.g., "contact me at `Skype: foo`"), and indeed this approach was applied in previous work with underground forums [12]. The drawback is that some identifiers are not extracted (thus reducing the coverage of our measurement). This limitation is partially overcome due to the extraction of other features. Indeed, although this approach covers a subset of the total Skype accounts present, we err on the side of reducing false positives.

3.1.3. Email

Until very recently, emails were the primary means for personal Internet communications. Nowadays, nearly all online services that require registration rely on emails to identify users, and thus added to a communication platform, emails have become an all-purpose digital identity.

In the case of underground forums, we have identified two main reasons by which emails are shared. On the one hand, members write their emails to get in touch with other users, as a personal identifier. On the other hand, we have observed lists of leaked email accounts, which typically include the email and the associated (sometimes hashed) password.

Since emails identify users uniquely, we consider this as an important feature to extract. We use a regular expression to extract all the existing emails that were posted by users. Concretely, the regular expression looks for a sequence of alphanumeric characters, including the dot and plus sign, followed by the address sign (@) and another sequence of alphanumeric characters with at least one dot. Thus, our extraction includes addresses using sub-aliases by means of the '+' sign.

3.1.4. IP

IP addresses unequivocally identify a computer on the Internet. The exchange of these addresses is typical in underground forums. For example, forum users could post IP to their own hosted services, like gaming servers, bastion hosts, or Virtual Private Networks.

Even with new servers adopting IPv6, by mid 2021 IPv4 is still the most prevalent on the internet, with over 94% use according to [43]. Moreover, underground forums contain data posted in the past, dating back various years, where IPv6 was not even available. Thus, in our experimentation, we focus our extraction on IPv4 addresses. These follow a common format, which we encode in a regular expression to extract them from the forum posts.

3.1.5. Bitcoin addresses

Another characteristic feature covered in our work is the usage of Bitcoin (BTC) addresses. Like any other Blockchain implementation, Bitcoin works by generating two complementary keys (public and private) which are represented by Base56 strings. The public key of a user is in turn used to send and receive payments. During the last few years, Bitcoin has become a common online payment method. Moreover, it has been shown the preferred virtual currency used by underground communities [10, 44], for example to ask for ransomware payments [45] or as the main payment method for trading illicit goods and services [46, 47]. This is due to the sense of anonymity provided [48]. Indeed, cryptocurrencies that make payments untraceable (such as Monero) have become an important medium for secure money laundering and have become a key part of the cybercrime ecosystem [49]. We note that the usage of cryptocurrencies is not necessarily a sign of illegal activities. However, these are frequently shared in underground forums, for example, to ask for donations for a service or product given for free (e.g., a tutorial).

While our methodology allows us to include any currency, we have focused on Bitcoin, since it dominates the cryptocurrency landscape. In this case, the addresses follow a common structure. Addresses start with the numbers 1 or 3 and are followed by a string of 26 to 35 characters [50]. Accordingly, we use a regular expression that encodes this format and extracts the BTC addresses from the posts.

3.1.6. Trigrams

When an user writes in an internet forum, it inherently leaves a personal footprint. The language, expressions, and grammar mistakes characterize users and their way of expressing things. Furthermore, in underground forums it is common to observe a high prevalence of slang. Slang is the specific language developed by the interactions of individuals discussing a particular topic. The use of slang permits identifying areas of interest and further characterizes the user. However, it hinders the use of off-the-shelf NLP tools that have been proven effective in other areas [51]. Accordingly, in this work, we extract the n-grams (with n=3) from users' messages to characterize these different uses of language. While we do not aim at conducting further stylometry analysis, these characteristics serve to find similarities between users based on their use of the language. Since this feature depends on the language being used, in our experimentation we have only applied it to English-speaking forums.

3.2. Data Sanitization, Normalization and Reduction

Once the different features related to users are extracted, the next step is to preprocess the dataset. First, the data is normalized. Concretely, if the extracted features are not case-sensitive (e.g., email or domains), the values can be transformed into the lower case. Indeed, we have observed users mixing the upper and lower case indistinctly, and the objective is to remove any possible redundancy and normalize the values. Second, we conduct data sanitization and reduction, to remove instances not providing enough information to identify account relationships. To speed up the process, we combine sparse matrices and map structures (i.e., dictionaries) for auxiliary purposes. Sparse matrices have the advantage that they reduce a lot the storage in memory of big amounts of data while dictionaries enhance the data access complexity.

For each feature $f \in F$ (note, that we explicitly remove it from mathematical notation), we create two indexed dictionaries to map each user to a user identifier $D_{user} : u_i \mapsto i$ and each value to a value identifier $D_{value} : v_j \mapsto j$. Then, we represent users and values in a matrix $A_f = \{a_{i,j} | u_i \in U, v_j \in V\}$ in which each row i corresponds with user u_i and each column j corresponds with value v_j. Thus, the value $a_{i,j}$ is the number of times that user u_i have posted value v_j in the forum. Thus, removal of users and values only requires removing their corresponding row or column respectively. This step is only performed at the end to speed up the removal, thus, the different rows and columns are marked to be removed. This implies just one rescaling and re-indexing of the matrix and permits faster indexing and optimization of storage space in memory. Moreover, the use of these data structures allows for efficient sanitization since we can speed up the search.

To reduce the dataset, we remove unnecessary values. First, we remove all the values that are only shared by a single account since these cannot be used to create relationships with other accounts. This is the first process since it requires little computation effort and a large amount of information can be removed. Second, we also conduct ad-hoc reduction for links, IP addresses and trigrams using custom heuristics. We note that these heuristics can be tuned specifically for each particular scenario. Concretely, for our experimentation we proceed as follows:

1. Links. We remove *internal* links. We consider a link to be internal when it contains the domain of the forum where it was posted (e.g., users referring to another thread in the forum). Furthermore, to reduce false positives, we also remove host-only links i.e., those that do not provide a path in the URL, e.g., *www.mainsite.com*). It might increase the risk of having false negatives, e.g., in cases where two accounts share a link to an external owned host. Since the proposed methodology uses more features, we chose to err on the side of minimizing false positives.
2. IP addresses. We use a whitelist of reserved addresses, i.e., addresses whose usage is specific to cases such as local area networks (e.g., 192.168.0.0/16), local-hosts (e.g., 127.0.0.1), or masks (e.g., 255.255.0.0). For example, many tutorials use local IP addresses to explain how to set up local environments. Thus, we remove these IP addresses since they do not imply any relationship between users. Additionally, some users provide a large list of IP addresses, e.g., to share proxies. These are not valid to characterize accounts. Thus, we filter out IP addresses that are shared in large lists (i.e., having 30 or more distinct addresses).
3. Trigrams. We remove trigrams that contain non-ASCII characters. This is due to various users using special characters to represent emojis or other graphical ideograms in text. In these cases, trigrams are not representative of a single user, and two different accounts might be well using the same ideogram and not being related. While it may impact the feature performance, we reduce the number of false positives. Additionally, given the variety of features, the methodology remains unaffected.

Finally, we repeat the removal of those values that appear only once (i.e., they are shared by a single account). Then, we remove accounts that do not have any associated values.

As a result of this process, we obtain a cleaned and normalized dataset of users mapped to the values. This way, we reduce the processing time required in the next steps.

3.3. Distance Computation and Vectorization

Once the dataset is reduced and normalized, the next step is to compute the similarity metrics for each pair of users. The simplest approach would be to perform one-to-one comparisons of the features for each pair of users (this is referred to as 'baseline' algorithm in §4.2). This solution is not scalable, even after the dataset is sanitized and reduced. For example, a dataset of 10k users (e.g., a medium-size forum) would require the evaluation of nearly 50M pairs of users.

Processing large datasets requires efficient data structures to perform computations in RAM and to optimize the use of the disk. Thus, sparse and persistent matrices are used. Indeed, it is needed to optimize the computation of each of the pairs (i.e., comparing each pair of users to find the intersection of values shared by them), to reduce the performance required to compute a huge number of pair combinations.

To efficiently compare pairs of users, we conduct all the computations over matrices. Concretely, we leverage the indexed data structures presented in §3.2 to create user-to-feature matrices that can be efficiently stored and processed. For the comparison of the users in pairs, we define metrics that satisfy three main requirements:

R1 Any pair of users with no relationship should acquire the lowest score. This means that, if users U_a and U_b have not posted any common information, their score must be 0.

R2 The higher the number of coincidences for a given feature, the higher the score. This means that if two users U_a and U_b post the same information various times (e.g., the same email and the same BTC address in various posts), then they should receive a high score.

R3 The score of the pair should increase further if values that coincide are *rare* (i.e., they have a low prevalence in the dataset). For example, an ordinary IP is `8.8.8.8`, the DNS server offered by Google. Thus, the posting of this value by two authors (for the feature IP) should have a null or negligible increase in their mutual score.

Next, we describe the metrics used to analyze the similarity of users. In a nutshell, for each pair of users in the dataset ($u_i, u_j \in U_f$), and for each of the extracted features ($f \in F$), we calculate a Per-Feature distance metric that reflects how many values of f are shared by u_i and u_j, and how unique these values are. Here, uniqueness refers to how common or rare a feature value is across all the users (for example, `www.google.com` is a common value for the links across all users, whereas a particular URL referring to a personal resource shared in Pastebin would be more unique or particular to certain users). Accordingly, we obtain $|F|$ different metrics (6 in our implementation) for each pair of users. Then, these metrics are aggregated into a unique metric dubbed *Multi-Feature Similarity* (MultFS) metric, which fulfills the aforementioned requirements (§3.4).

Per-Feature distance metric. This metric is obtained from a particular variation of the Term Frequency - Inverse Document Frequency (tf-idf) metric used in Information Retrieval to reflect the importance of a given word to a document in a set of documents [52]. In this case, we define the Feature Frequency (FF) as described in Equation 1. This frequency characterizes the part that a feature value v_j represents in the whole data ($\sum_{t=1}^{V} a_{i,t}$) of the user u_i (i.e., it tells the importance of a feature value for a user). It also enhances features that are characteristic of one user, e.g., frequently-posted IPs or emails.

We also define the inverse user frequency (IUF), described in Equation 2, which highlights feature values v_j that are less common across the different users U.

The FF metric can be calculated with complexity $O(n * U)$ and the IUF metric with $O(n * V)$, that is, one operation per feature value. The usage of sparse matrices and operations over vectors allows us to speed up the computation.

$$FF_{v_j,u_i} = \frac{a_{i,j}}{\sum_{t=1}^{V} a_{i,t}} \tag{1}$$

$$IUF_{v_j} = log(\frac{|U|}{\sum_{i=1}^{U} d_{i,v}}) : d_{i,v} = \begin{cases} 1 & u_i \mapsto v \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

Once we have computed the FF for each user u_i and value v_j and the IUF for each value, we can combine them into a single metric by multiplying $FF_{u_i,v_j} * IUF_{v_j}$. This product equally considers the importance of each value for each user, i.e., how often the value is shared by the user, and the uniqueness of such value in the whole dataset, i.e. how 'rare' the value is for all the users. This results in a matrix $B_f = \{b_{i,j} = FF_{u_i,v_j} * IUF_{v_j} | u_i \in U, v_j \in V\}$.

At this point, we have achieved requirements R2 and R3. To combine the users in pairs, we compute the scalar product of the row vectors of the matrix B_f each with the rest. This last part is the most computationally expensive since we have to perform half of the matrix multiplication to obtain the relations of all pairs, counting for a total of $O(U^2)$ combinations. At this point, we make use of efficient vectorized computation and conversions of matrices to achieve cache efficiency and multithreading. It permits obtaining a new symmetric matrix C_f of dimensions $U \times U$ where the value of entry (u_i, u_j) is the similarity between two users. In the case that they share no feature, the scalar product remains 0, thus achieving requirement R1.

At the end of this process, each pair of users is assigned with $|F|$ scores (i.e, one distance metric for each of the extracted features). In the next section, we define the *Multi-Feature Similarity* (MultFS) metric, which combines these metrics into a single value.

3.4. The Multi-Feature Similarity (MultFS) metric

In the last step of the proposed methodology, we need to combine each of the pairwise similarities into one single metric. The purpose of the MultFS score is to combine the metrics from the different features equivalently. Due to the different scales of the features, we normalize all the possible values into the [0.0,1.0] interval. We use the min-max feature scaling method, shown in Equation 3, where *max* is the maximum value for each feature, and *min* is always zero.

$$norm(v) = \frac{v - min}{max - min} \tag{3}$$

To combine the different scores into a single metric, we propose an aggregation technique that considers different weightings for each feature, since these might have different meanings, depending on the scenario. For example, in the case of underground forums, most pairs likely share some trigrams, while it is less likely to find pairs of accounts sharing the same Skype identifier. Additionally, given that certain features are more common, it is likely to have some features where the standard deviation of the metrics are much bigger than others. Accordingly, we scale the contribution of each feature to the MultFS formula depending on its importance (i.e., the more unique a feature is, the more important, and vice versa). For this, we apply the IUF equation with a slight modification that prevents a feature shared by all pairs that get zeroed. It prevents removing feature values that are common to all users, for example, trigrams. Accordingly, we define the *Soft IUF (SIUF)* function (described in Equation 4), which consistently adds one in the formula to avoid zeroing elements. The result in our case is that most common terms get reduced, but they are still relevant whereas, less used terms, such as Skype, are highlighted if present. We denote the set of pairs as $U \times U$, the normalized pair similarity for feature f as $P_f = \{f_{f,i,j} | f_{f,i,j} = norm(c_{i,j})\}$ (recall that the matrix C_f is the resulting per feature pair similarity matrix).

$$SIUF_f = log(1 + \frac{|U \times U|}{\sum_{t=1}^{|U \times U|} e_{t,f}}) : e_{t,f} = \begin{cases} 1 & (u_i, u_j) \mapsto f \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

$$MultFS_{UxU} = \sum_{f=0}^{|F|} (P_f * SIUF_f) \quad (5)$$

Finally, we consider the sum of the different per-feature similarities to generate a MultFS per pair (Equation 5).

4. Analysis of underground forums

This section presents experimental results of the application of the proposed methodology to a large corpus of data collected from various underground forums (the dataset is described in §4.1). First, we analyze the performance gained due to the optimizations implemented to speed-up the comparison of accounts from large datasets (§4.2). Then, we conduct different analyses to validate the results of the methodology and to show how it helps with online investigations to detect multi-accounts(§4.3). Finally, we present case studies resulted from manual analysis of a subset of interesting accounts. (§4.4).

4.1. Dataset

The gathering of data is the first step and one of the most challenging to conduct online investigations [39]. For this project, we use the CrimeBB dataset [10]. This dataset was provided by the Cambridge Cybercrime Center under a legal agreement (see §5). The dataset contains the data scrapped from various underground forums, including both English and Russian forums. Topics in these forums include various deviant topics, such as computer hacking [4], video-game hacks and cheats [53] or social engineering techniques [54]. The dataset contains around 56M posts written by 1.1M accounts in 15 different forums. The size of the dataset motivates the application of methodologies for automatic analysis that are focused on performance efficiency.

4.2. Performance Analysis

One of the main goals of the proposed methodology is to scale to datasets having a large set of accounts, and thus we have designed a vectorized approach to optimize resources. To quantify the improvements, we compare our approach to a baseline algorithm. Such an algorithm gets the intersection of the set of values shared by each pair of users and computes the scores from the resulting set, i.e. it processes pairs of users at once. For the comparison, we generate a test benchmark that is executed both by the baseline algorithm and by our vectorized implementation.

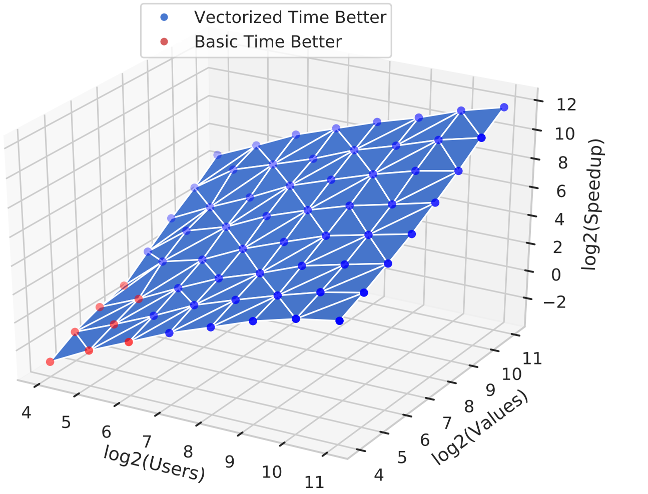

Figure 2: Vectorized Algorithm Speedup based on number of Users and Values.

We have experimented with various synthetic datasets, varying the number of users (U) and overall feature values (V). Then, we randomly assign to each user $u \in U$ a subset of $v \in V$ (which represents the sharing of features values by user $u \mapsto v$). We execute both algorithms in an Intel Xeon E5-2683 v3 @ 2.00GHz with 56 cores and 64 GB of RAM running CentOS8.

Figure 2 shows the comparison of two algorithms, in terms of execution time for the various tests. We observe that the improvement of the vectorized approach over the basic algorithm increases both with the number of values and users. In settings where the number of users and values is low, (i.e. less than 2^6), there is little improvement, and even the vectorized approach performs worst (red dots in the figure). This is due to the delay imposed by the generation of the in-memory structures. However, the speedup gain increases exponentially as the number of values and users increases. This quantitative analysis confirms that our approach scales well for a large number of values.

The previous results were based on test samples. To calculate the time saved on a real dataset, we analyze the execution time of both algorithms on CrimeBB. Since it is unfeasible to execute the basic implementation on the entire dataset, we apply linear regressions from the results obtained in the test sample. The results are presented in Table 1, including the total number of values extracted (V) for each feature and the total number of users (U), after the data reduction process. In all cases, the speedup gained is above 10^4 seconds, but we observe that for larger sets of values, e.g. Links, the vectorized approach performs 10^8 times better than the baseline. This implies that, in the case of links, processing 524k values extracted for 202k users takes around 24 minutes to complete in our approach, while in a baseline approach this would be infeasible (i.e., more than 7.7k years), since the number of combinations and the time to compute each pair increases factorially (i.e., adding the n+1 user would imply computing n+1 more pairs).

4.3. Detection of Multi-accounts

Multi-accounters (also called *doppelgangers* [55] or *sock-puppets* [35]) are users that open more than one account to ob-

Table 1: Algorithm execution time (seconds) estimated after dataset cleanup for the vector and basic approaches.

F	U	V	Vector	Basic	Speedup
IP	15 277	12 801	3e+01	7e+06	2.3e+05
Link	201 799	524 566	1.4e+03	2.4e+11	1.7e+08
Trigram	113 987	8391	1.7e+02	2.1e+08	1.3e+06
Skype	15 268	6659	2.3e+01	2.7e+06	1.2e+05
Email	28 848	18 333	6.3e+01	4.2e+07	6.6e+05
BTC	4894	3576	6.4e+00	1.2e+05	1.8e+04
Total	380 073	574 326	1.7e+03	2.4e+11	1.4e+08

tain some benefit, usually with illicit goals. Having multiple accounts is prohibited in most online communities, and when detected, these are banned. In this section, we leverage the information obtained from our methodology together with two other sources of information to detect such accounts: i) data related to the social network, and ii) the analysis of similar pseudonyms used by the accounts. The main goal is to evaluate the soundness and validity of the proposed methodology and to show how it can help during online investigations when combined with other forensic approaches to analyze actors of interest. We then conduct manual validation on a subset of 100 pairs of accounts to validate the detection. Finally, we apply stylometry analysis on a subset of accounts to further investigate whether they belong to the same user or not. This way, we can compare the benefits of using our methodology to improve existing approaches to detect multi-accounts.

Graph Analysis. The MultFS score indicates the similarity of each pair of accounts, based on the artifacts left online in the forums. However, two accounts might be well posting the same artifacts, buy referring to each other (or even a third person), e.g., to report scammers or to recommend other's services. Thus, we refine the identification process employing Social Network Analysis. Concretely, we proceed as follows. First, we build two different graphs:

1. $G_{mfs} = \{N, E\}$. A unique undirected graph where nodes are forum accounts and edges represent that the two accounts have been related by the MultFS. Accordingly, edges are weighted by the MultFS score of the accounts linked. Formally, each edge $e_{ij} \in E$ connects two nodes $n_i, n_j \in N$ and is weighted by $MultFS_{n_i,n_j}$.
2. $G_{SN}(F) = \{N, E\}$. For each underground forum F, we build a directed graph representing the Social Network. We follow the same approach made in previous works [4, 54]. Concretely, each node in the graph represents a forum account, and an edge $e_{ij} \in E$ from node n_i to n_j ($\forall n_i, n_j \in N$) is weighted by the number of responses of user n_i to a post previously made by n_j.

Second, we calculate the connected components of G_{mfs} to get the different sets of users that have been grouped according to the MultFS. The size of these components indicates the number of accounts grouped. It is reasonable to expect that most of these components will be of reduced sizes, e.g., the same user or group managing 2 or 3 accounts. Indeed, in our experiments we have obtained a total of 5 372 different groups, out of which the majority are of size 2, i.e., pairs of users (4 494, 83.6%) and size 3 (698, 13%). Then, there are 173 groups of 4 users (3.22%), 3 groups of 5 users (0.06%). Finally, there are 4 groups of size 7,9,15, and 22 users each (0.08%).

Finally, we combine information from each of the graphs. For each pair of accounts linked in G_{mfs}, provided that these two accounts are from the same forum F, we check if they are connected in $G_{SN}(F)$, and if so, we get the number of interactions between them. The goal of this step is to enrich the analysis by filtering out users that are strongly connected in the Social Network, i.e., one is actively responding to the other (and/or vice versa) with high frequency. Two accounts are strongly connected if they have more than N interactions (i.e., responses from one to the other), with the value N being dependant on the total number of interactions made by the accounts under investigation. The rationale behind this is that, if two accounts belong to the same user, it is likely that these accounts would not interact with each other as frequently as they interact with other accounts.

Username similarity. The graph analysis only considers the interactions of user accounts that belong to the same forum. Thus, we analyze the similarity of the nicknames used by pairs of accounts from different forums. We use the Jaro-Winkler distance, which provides an edit distance that considers the size of the words being compared (with 1 indicating that the words are the same) as well as common prefixes used in these [56]. We chose this metric due to it being faster than others, and it is optimized for comparing small strings as in the case of usernames [57, 58]. Before computing the distances, we transform each username to lowercase. By manually inspecting the Jaro-Winkler distances for 672 usernames, we establish a conservative threshold of 85% to consider two accounts being related in our dataset. This way, we include usernames that are the same, as well as usernames that do have small modifications, such as character replacement related to leet ("l33t") language or additions in form of suffixes such as '2' or 'unbanned'.

Validation of results. To validate the results of our methodology, we manually analyze pairs of accounts that are only related to each other (i.e. they form a connected component of size 2), and that fulfill one of these two requirements:

1. If the accounts are from the same forum, they have less than 5 interactions between them. The rationale is that, if two accounts are from the same user or gang, they won't interact with each other.
2. If the accounts are from different forums, they are related due to their username similarity. As described before, this implies that their Jaro-Winkler distance is over 0.85 (out of 1).

Using these heuristics, we obtained a total of 3 716 pairs of accounts. From these, we selected the top 200 according to their MultFS score, and conducted conducted manual validation to verify whether the accounts are from the same user or not. This validation consisted of the reading of the posts where the various artifacts (e.g. Skype usernames or IP addresses) have been posted. We consider two accounts belong to the same user if there is clear evidence of it, i.e. the user provides strong evidence of being the owner of the identifier posted by the two accounts (e.g. *"Add me on Skype: xxxxx"* or *"send me an email to xxxxx"*). Overall, we found that 60% of the pairs were from the same user. In some cases (15%), we observe that the accounts were wrongly related. Most commonly, this is due to one account reporting the other as a scammer, or one account quoting verbatim the content of the other account. We found also threads where a user was being *doxed* (see §4.4), and thus its personal information was being re-posted by various actors. Finally, some accounts (25%) were providing the same identifiers, but we were not able to find strong evidence of these being the same user. Overall, the validation showed that 3 out of 5 of the accounts that are linked using our methodology belong to the same user. If we ignore those which are unknown, we observe a False Positive rate of 20%, which means that an online investigator can expect that 80% of the accounts being linked by our method actually belong to the same actor. In Section 5 we discuss about the limitations with respect to False Negatives in our methodology.

Stylometry analysis. To further validate our methodology, we use Doppelgänger-finder, a state-of-the-art stylometry analysis tool, which was first used to detect multi-accounts in underground forums [18] and later to analyze cross-domain accounts on different social media sites [24]. The use of this tool on large datasets is limited, e.g., dealing with more than 50 accounts is computationally expensive and does not increase the accuracy of the classifier [24]. Accordingly, the validation works as follows. First, we select the list of 100 pair-of-accounts with higher MultFS, i.e., those that are most related and that also are not connected in the social network graph or have a similar username (see above). Since the validation method relies on natural language analysis, we exclude accounts from Russian forums. Second, we group the posts of each account in documents of around 500 words and filter out pairs where one of the accounts has less than 4k words [24]. This resulted in 13 pairs (26 accounts) being analyzed. Third, we used JStylo [59] to extract the same set of features from the documents originally used by Afroz et al. [18]. Finally, we applied the open-source version of Doppelgänger-finder to obtain the pairwise probabilities for each of the accounts.

Results from Doppelgänger-finder can be analyzed either using thresholds (which consider two accounts are from the same user if their joined probability is above a predefined threshold) [18] or considering accounts with higher probabilities [24], i.e. to get the account that looks more similar to each other. The former requires ground truth on the dataset to establish thresholds. Thus, we follow the latter approach. For each account, we get an ordered list based on how similar the other accounts are (i.e., ordered by their joint stylometry probability). Accordingly, for each of the 13 pairs, we analyze the ranks of each partner from the list.

Table 2 shows this comparison (DGG is the Doppelgänger-finder probability). D1 (D2) is the position of user 2 (1) in the Doppelgänger-finder list of user 1 (2). Column 2 (DF?) indicates whether the two accounts are from the same forum. We have manually analyzed all the pairs and labeled them as either the accounts being from the same user (S) or false positives (F) (see column 'Type' in Table 2). We observe that out of the 13 analyzed pairs, only one is a False Positive. This pair has been related by their MultFS due to one of the accounts quoting verbatim the content of the other account in the reply, including the contact details. Most of the pairs have both members having each other in the first position of their ranks, i.e. they have the highest stylometry similarity and are thus most related to each other. All these cases are pairs of accounts from the same user. Only 2 pairs (#6 and #13) do not have similar stylometry. One of them is the False Positive discussed before. The other pair is undoubtedly from the same user, but we have observed that the language used is complex, containing several grammatical errors and typos, and also extended use of jargon. Under these circumstances, analysis based on Natural Language Processing has limitations [51]. Thus, methods that do not rely on NLP (or a combination of these), like the one proposed here, are a potential direction to improve investigations on online underground communities. Two pairs contain accounts from different forums.

In general, accounts that are linked using our methodology also have similar stylometry, but this is not always the case. Moreover, applying stylometry analysis with large datasets is limited since it requires high computational resources and also a minimum amount of text to do the analysis [18, 24], and indeed we were able to run the tool only over 13% of the pairs selected for manual validation. It suggests that a hybrid approach is a promising technique to deal with large-scale datasets and to identify accounts from the same user. During our experimentation, we have detected various accounts from Russian forums and English forums that are closely related. While we have not gone through further validation, due to our lack of understanding of Russian, the features for which they are related indicate a high likelihood of these belonging to the same user. Again, in such a scenario, language-dependent techniques such as NLP would fail to link accounts. Since we mostly rely on features that are independent of the language, our method allows identifying such accounts.

The proposed validation has two main limitations. First, due to the lack of labeled data, it requires manual validation, which is error-prone. We have partially addressed this limitation by relying on strong evidence, i.e., checking that the information posted is claimed as being from the user who posts. Second, our validation does not allow us to quantify at scale the number of false positives and false negatives. Again, this process would require a labeled dataset to actually account for the number of mislabeled pairs. We note that the scope of the methodology is to assist during online investigations, by automatically linking related accounts. Thus, we next provide some case studies on it could be used for such investigations.

Pair	SF?	DGG	MultFS	D1	D2	SU?
1	✓	0.114	135.762	1	1	✓
2	✓	0.024	121.143	1	1	✓
3	✓	0.011	136.690	1	1	✓
4	✓	0.050	143.939	1	1	✓
5	✓	0.068	162.113	1	1	✓
6	✓	0.002	124.744	14	4	✓
7	✓	0.056	133.872	1	1	✓
8	✓	0.005	149.433	1	1	✓
9	✗	0.052	141.755	1	1	✓
10	✓	0.028	200.000	1	1	✓
11	✓	0.052	125.098	1	1	✓
12	✗	0.198	108.237	1	1	✓
13	✓	0.001	128.520	12	10	✗

Table 2: Analysis of the Doppelgänger Finder (DGG) and MultFS similarities. SF? indicates whether the accounts are from the same forum (✓) or not (✗). D1 and D2 are the positions of each other account in the DGG ranking. SU? indicates whether the accounts are from the same user (✓) or not (✗), done by manual inspection

4.4. *Case Studies*

As we have mentioned before, the purpose of our methodology is to identify related accounts. In this section, we analyze some interesting cases of related accounts that do not necessarily belong to the same user. As presented in §4.3, our methodology can be combined with other approaches (like SNA or stylometry analysis) to further identify which of these related accounts belong to the same user. We leave for future work the analysis about the use of multiple accounts by the same user.

4.4.1. *Proxy Sellers*

We have identified a group of accounts sharing the same IP addresses and links. These accounts are selling proxies and use multiple forums. Three of them were selling the same products (i.e., a single seller having multiple accounts or various members of an organization), mostly proxies.[4] Their posts mainly consisted of large lists of IP addresses, together with the description of where these IP addresses were geographically located. Additionally, each IP contained a link to an external page, where the actual trading occurs. In one of the forums, the activity carried out by the accounts was similar. By longitudinal analysis of the posts and threads made by these accounts on a single forum, we have seen that the group started their business with an initial account. After a couple of months, they created two other accounts. Indeed, the first account ceased its activity soon after the two accounts were created. Both accounts ceased activity on the same day and made their last post at similar times. Regarding the posting hours, we manually verified that the average hours at which the users posted followed similar patterns.

4.4.2. *Doxing*

Trading in underground forums relies on trust [16, 15]. In some cases, transactions end up in users giving their money to sellers which do not respond and steal their money (i.e., scammers). A common approach to revenge scammers is through *doxing* [28], where scammed users post personal or identifiable information about the scammers. One of the accounts is accusing the other of having multiple accounts, and for such purpose, he or she posted various contact details. This case shows that a potential application of our approach would be to detect doxing in underground forums since this practice commonly requires posting personal information. Indeed, various of the accounts that were wrongly identified as being from the same user (see §4.3 was due to one of them was doxing the other, and thus posting the same identifiers.

4.4.3. *Bots*

While our methodology has not been designed as a bot detection system, in certain cases it can capture the usage of automated tools or bots in pairs of accounts. We have detected various pairs of accounts using similar or the same posts to advertise some products or services or to increment the traffic or views in certain threads. They were advertising various third-party services, with links to such services. These accounts have different names, but were registered at the same time, and had their last post the same day. Thus, these were most probably created and operated by automated means (i.e., bots). Thus, these are spam-bots. Moreover, we have verified that these accounts have been banned and no longer exist in the online forum.

4.4.4. *Post Copies/Plagiarism*

Some pairs were related due to the members copying posts. These were both highly-related, since the information posted is the same, and this information included links and contact emails. In one of the cases, the information that our system considered to relate a variety of accounts refers to a tutorial on a *stresser* service setup (used to create DDoS attacks), which included various configuration parameters that were unique for such service. After analyzing other posts, they seem to be different users that copied the tutorial from the same source. In another case, the accounts were related because they were copying a tutorial related to email spam. This tutorial included emails and IP addresses and was classified as related. These situations result in false positives for our system. Still, we consider these accounts are somehow related (i.e., they are posting the same tutorial), and thus might be of interest to investigate specific activities. We recall that our methodology is not intended to indicate the reason by which two accounts are related, but to detect such relationships at scale. In the particular case of accounts posting copies, the intention can vary from sharing some useful information (e.g., to gain reputation), to promote others' products or services by reselling, or to gain notoriety by popping up their post. Indeed, without analyzing further context (e.g., the replies), it is challenging to understand whether accounts posting the same content are indeed from the same user (e.g.,

[4] Note that we ignore whether the infrastructure for these proxies was hacked or stolen, or it comes from licit means.

users diversifying their activity across forums), or one copying content from the other.

5. Discussions

This work presents a methodology to improve investigations of cybercrime activities sustained by online social media sources. In this section, we discuss the limitations, potential applications, and associated ethical issues.

Limitations. We manually select the features used for the comparison of accounts, which are based on existing works and our expertise related to the analysis of underground forums. Nevertheless, this set can be modified or extended when adapted to other social media platforms. We assume that users either do not care about their accounts being linked, or they make mistakes. If they do not share any of these features or are cautious to use different ones, then our methodology would fail to detect them, turning into false negatives. However, previous investigations have shown that this information is not always hidden, even by high profile actors [27, 19]. If these identifiers are present in the dataset, then our methodology would link the accounts together. Moreover, the main contribution of the proposed methodology is that it helps to quickly process a large dataset and link together accounts that are related. These relations are in many cases due to being from the same actor (see §4.3). Also, some users might use techniques to bypass automatic scrapings, such as changing 'http' by 'hxxp' or the @ symbol for 'at' Our current implementation does not deal with all possible cases, though these can be easily included by adding heuristics or improving the regular expressions used for data extraction. Similarly, the feature extraction might include values that do not belong to the corresponding feature, e.g., text resembling BTC addresses or IPs due to these having the same format. We have partially overcome this limitation by sanitizing and reducing the dataset as explained in §3.2. It is important to note that certain accounts might get a higher degree of similarity due to plagiarism, i.e., a user that copies the contents of another user. We note that our methodology allows linking these two accounts together since it might be of interest in certain investigations. As explained before, our validation was done by manually inspecting the posts of a limited subset of pairs of users where the accounts have shared the same artifact. This requires a certain amount of manual effort and does not scale, so it does not allow to quantify the accuracy of the process. Thus, the proposed approach must be considered as an auxiliary tool to help manual investigators, which at the end must collect proper evidence that must be used in court, a process that is challenging employing fully-automated tools. Also, as discussed before, a limitation of our methodology is that it only works when two accounts provide similar information. In case this information is missing, then the analysis must uniquely rely on other methods, like stylometry analysis or social network analysis. Nevertheless, the benefits of the proposed methodology outweigh its limitations, and can be used together with other approaches. State of the art tools are not able to deal with large datasets, and simple pattern extraction does not work due to the need to process and link together a large number of user accounts with their features.

Potential applications. The proposed methodology can be beneficial in various domains. Firstly, it can assist law enforcement and cyber-intelligence practitioners to quickly get interesting accounts out of a pool of members, or to identify accounts with stronger links to a known offender. Another potential application would be for forum administrators. Certain forums do prohibit the possession of multiple accounts, and identifying related accounts may increase the security of these forums, as well as removing certain account behaviors such as spamming or *botting*. Additionally, this methodology can be incorporated to improve existing analysis approaches of social media data. For example, it can be used to reduce the number of users being analyzed in resource-consuming processes like stylometry analysis [55]. Also, it can improve social network analysis, e.g., by grouping accounts belonging to the same actor, or by creating new links between nodes based on their MultFS similarity. Finally, our methodology can be used to generate a ground-truth for testing supervised algorithms used to capture similarities between pairs of users by other means. It can be adapted to various domains, provided that the online investigator(s) properly select and extract the features and provide a mapping of these features to the users that shared them. Then, the rest can be applied directly in an automated way.

Ethics. This project deals with a dataset collected by the Cambridge Cybercrime Centre which was shared with us under a legal agreement. We comply with the terms and restrictions of the data usage stated on such agreement and follow standard guidelines in computer science research [60, 41]. Concretely, we only use the data for the research exposed in this document. Even though the main purpose of this project is to identify related accounts, we do not aim at identifying the individuals behind these accounts. Thus, we never process information out of what these users have shared publicly in the forums, even if such analysis would have been useful to refine our measurements. For example, we do not aim at geolocating the IPs or to further explore the links. Moreover, to reduce the potential harm caused to these individuals, we do not publish or disclose any personally identifiable information posted in the forums. Also, to reduce the likelihood that the identities of the users could be leaked, we were careful with the presentation of our results (e.g. not providing further details on the case studies). The data is treated with due precautions, stored encrypted in one of our servers, and with access restricted only to the authors of this work. Finally, to preserve justice and fairness, we do not arbitrarily target specific groups based on any non-technical factors such as social, racial, or religious issues.

Reproducibility. Finally, to foster reproducibility, we make our code publicly available. [5] Note that, in order to reproduce our results, researchers must contact the Cambridge Cybercrime Centre and request access to the CrimeBB dataset used in our experiments. [6]

[5] https://github.com/jcabrero/multfs_public

[6] www.cambridgecybercrime.uk

6. Conclusions

Online media sources, such as underground forums and markets, are a valuable source of information for security practitioners and law enforcement. Most of these are openly accessible and only require users to register an account. Users can thus make use of various accounts, for example, to hinder law enforcement investigations or to influence the market. In this work, we have presented a methodology for the identification of accounts that are related to each other. The methodology relies on characteristic artifacts publicly posted by users (e.g. Skype handles or email addresses) and is able to compute similarities of a pair of users even in forums of different languages. The methodology is designed to analyze online identities at scale, in reduced time frames, and thus can deal with large datasets. We conduct our experimentation with a dataset of more than 56M posts from underground forums and show how our methodology can be combined with existing approaches to assist in online investigations. The proposed methodology is flexible and can be adapted to the analysis of other online media sources.

Acknowledgements

We thank the anonymous reviewers for their valuable feedback. We also thank the Cambridge Cybercrime Centre, for sharing the CrimeBB dataset.

This work was partially supported by CERN openlab, the CERN Doctoral Student Programme, the Spanish grants ODIO (PID2019-111429RB-C21 and PID2019-111429RB) and the Region of Madrid grant CYNAMON-CM (P2018/TCS-4566), co-financed by European Structural Funds ESF and FEDER, and Excellence Program EPUC3M17. The opinions, findings, and conclusions or recommendations expressed are those of the authors and do not necessarily reflect those of any of the funders.

References

[1] R. Van Wegberg, S. Tajalizadehkhoob, K. Soska, U. Akyazi, C. H. Ganan, B. Klievink, N. Christin, M. Van Eeten, Plug and prey? measuring the commoditization of cybercrime via online anonymous markets, in: Proceedings of the 27th USENIX Security Symposium, 2018, pp. 1009–1026.

[2] A. Hutchings, Crime from the keyboard: Organised cybercrime, co-offending, initiation and knowledge transmission, Crime, Law & Social Change 62 (1) (2014) 1–20.

[3] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, G. M. Voelker, An analysis of underground forums, in: Proceedings of the ACM SIGCOMM conference on Internet Measurement Conference, 2011, pp. 71–80.

[4] S. Pastrana, A. Hutchings, A. Caines, P. Buttery, Characterizing eve: Analysing cybercrime actors in a large underground forum, in: International symposium on Research in Attacks, Intrusions, and Defenses (RAID), Springer, 2018, pp. 207–227.

[5] J. Lusthaus, Beneath the dark web: Excavating the layers of cybercrime's underground economy, in: 2019 IEEE European symposium on security and privacy workshops (EuroS&PW), IEEE, 2019, pp. 474–480.

[6] L. Allodi, Economic factors of vulnerability trade and exploitation, in: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security, ACM, 2017, pp. 1483–1499.

[7] M. Rezaeirad, B. Farinholt, H. Dharmdasani, P. Pearce, K. Levchenko, D. McCoy, Schrödinger's rat: Profiling the stakeholders in the remote access trojan ecosystem, in: 27th USENIX Security Symposium (USENIX Security 18), 2018, pp. 1043–1060.

[8] V. Valeros, S. Garcia, Growth and commoditization of remote access trojans, in: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2020, pp. 454–462.

[9] A. Noroozian, M. Korczyński, C. H. Gañan, D. Makita, K. Yoshioka, M. van Eeten, Who gets the boot? Analyzing victimization by DDoS-as-a-service, in: International Symposium on Research in Attacks, Intrusions, and Defenses, 2016, pp. 368–389.

[10] S. Pastrana, D. R. Thomas, A. Hutchings, R. Clayton, Crimebb: Enabling cybercrime research on underground forums at scale, in: Proceedings of the 2018 World Wide Web Conference, WWW '18, 2018, pp. 1845–1854. doi:10.1145/3178876.3186178.
URL https://doi.org/10.1145/3178876.3186178

[11] K. Turk, S. Pastrana, B. Collier, A tight scrape: methodological approaches to cybercrime research data collection in adversarial environments, in: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS&PW), IEEE, 2020, pp. 428–437.

[12] S. Sundaresan, D. McCoy, S. Afroz, V. Paxson, Profiling underground merchants based on network behavior, in: APWG Symposium on Electronic Crime Research (eCrime), IEEE, 2016, pp. 1–9.

[13] Christopher Heine, 3 Waves of Cyberattacks Brought Down Twitter, Spotify, Pinterest, NYT and Other Sites, adweek.com/digital/. Last visited 17th December 2020 (october 2016).

[14] B. Krebs, Who is Anna-Senpai, the Mirai worm author? (January 2017).
URL https://krebsonsecurity.com/2017/01/who-is-anna-senpai-the-mirai-worm-author/

[15] B. Dupont, A.-M. Côté, C. Savine, D. Décary-Hétu, The ecology of trust among hackers, Global Crime 17 (2) (2016) 129–151.

[16] S. Afroz, V. Garg, D. McCoy, R. Greenstadt, Honor among thieves: A common's analysis of cybercrime economies, in: eCrime Researchers Summit, IEEE, 2013, pp. 1–11.

[17] A. Mell, et al., Reputation in the market for stolen data, Tech. rep., University of Oxford, Department of Economics (2012).

[18] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, D. McCoy, Doppelgänger finder: Taking stylometry to the underground, in: 2014 IEEE Symposium on Security and Privacy, IEEE, 2014, pp. 212–226.

[19] B. Krebs, Who is Marcus Hutchins? (September 2017).
URL https://krebsonsecurity.com/2017/09/who-is-marcus-hutchins/

[20] G. Danezis, P. Mittal, Sybilinfer: Detecting sybil nodes using social networks., in: NDSS, San Diego, CA, 2009, pp. 1–15.

[21] Q. Cao, M. Sirivianos, X. Yang, T. Pregueiro, Aiding the detection of fake accounts in large scale social online services, in: Proceedings of the 9th USENIX conference on Networked Systems Design and Implementation (NDSI), USENIX Association, 2012, pp. 15–15.

[22] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, M. Tesconi, The paradigm-shift of social spambots: Evidence, theories, and tools for the arms race, in: Proceedings of the 26th international conference on World Wide Web companion, International World Wide Web Conferences Steering Committee, 2017, pp. 963–972.

[23] B. Perez, M. Musolesi, G. Stringhini, You are your metadata: Identification and obfuscation of social media users using metadata information, in: Proceedings of the 12th International AAAI Conference on Web and Social Media, 2018, pp. 241–250.

[24] R. Overdorf, R. Greenstadt, Blogs, twitter feeds, and reddit comments: Cross-domain authorship attribution, Proceedings on Privacy Enhancing Technologies 2016 (3) (2016) 155–171.

[25] A. Rocha, W. J. Scheirer, C. W. Forstall, T. Cavalcante, A. Theophilo, B. Shen, A. R. Carvalho, E. Stamatatos, Authorship attribution for social media forensics, IEEE Transactions on Information Forensics and Security 12 (1) (2016) 5–33.

[26] R. Frank, M. Thomson, A. Mikhaylov, A. Park, Putting all eggs in a single basket: A cross-community analysis of 12 hacking forums, in: IEEE International Conference on Intelligence and Security Informatics (ISI), IEEE, 2018, pp. 136–141. doi:10.1109/ISI.2018.8587322.

[27] N. Popper, The tax sleuth who took down a drug lord (December 2015).
URL https://perma.cc/V6UK-DZ5M

[28] P. Snyder, P. Doerfler, C. Kanich, D. McCoy, Fifteen minutes of unwanted fame: Detecting and characterizing doxing, in: Proceedings of the 2017 Internet Measurement Conference, ACM, 2017, pp. 432–444.

[29] M. Macdonald, R. Frank, J. Mei, B. Monk, Identifying digital threats in a hacker web forum, in: International Conference on Advances in Social

Networks Analysis and Mining, IEEE/ACM, 2015, pp. 926–933.

[30] E. Nunes, A. Diab, A. Gunn, E. Marin, V. Mishra, V. Paliath, J. Robertson, J. Shakarian, A. Thart, P. Shakarian, Darknet and deepnet mining for proactive cybersecurity threat intelligence, in: Conference on Intelligence and Security Informatics (ISI), IEEE, 2016, pp. 7–12.

[31] V. Benjamin, S. Samtani, H. Chen, Conducting large-scale analyses of underground hacker communities, Cybercrime Through an Interdisciplinary Lens 26 (2016) 56.

[32] Z. Yang, C. Wilson, X. Wang, T. Gao, B. Y. Zhao, Y. Dai, Uncovering social network sybils in the wild, ACM Transactions on Knowledge Discovery from Data (TKDD) 8 (1) (2014) 2.

[33] Y. Zhang, Y. Fan, W. Song, S. Hou, Y. Ye, X. Li, L. Zhao, C. Shi, J. Wang, Q. Xiong, Your style your identity: Leveraging writing and photography styles for drug trafficker identification in darknet markets over attributed heterogeneous information network, in: The World Wide Web Conference, ACM, 2019, pp. 3448–3454.

[34] X. H. Tai, K. Soska, N. Christin, Adversarial matching of dark net market vendor accounts, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery ’|&’ Data Mining, KDD ’19, Association for Computing Machinery, New York, NY, USA, 2019, p. 1871–1880. `doi:10.1145/3292500.3330763`.
URL `https://doi.org/10.1145/3292500.3330763`

[35] M. Tsikerdekis, S. Zeadally, Multiple account identity deception detection in social media using nonverbal behavior, IEEE Transactions on Information Forensics and Security 9 (8) (2014) 1311–1321.

[36] J. Gharibshah, T. C. Li, M. S. Vanrell, A. Castro, K. Pelechrinis, E. E. Papalexakis, M. Faloutsos, Inferip: Extracting actionable information from security discussion forums, in: International Conference on Advances in Social Networks Analysis and Mining, IEEE/ACM, 2017, pp. 301–304.

[37] M. Egele, G. Stringhini, C. Krügel, G. Vigna, Towards detecting compromised accounts on social networks, IEEE Transactions on Dependable and Secure Computing 14 (2017) 447–460.

[38] E. Mariconti, J. Onaolapo, S. S. Ahmad, N. Nikiforou, M. Egele, N. Nikiforakis, G. Stringhini, What's in a name?: Understanding profile name reuse on twitter, in: Proceedings of the 26th International Conference on World Wide Web, WWW ’17, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, 2017, pp. 1161–1170. `doi:10.1145/3038912.3052589`.
URL `https://doi.org/10.1145/3038912.3052589`

[39] M. Edwards, A. Rashid, P. Rayson, A systematic survey of online data mining technology intended for law enforcement, ACM Computing Surveys (CSUR) 48 (1) (2015) 15.

[40] R. Overdorf, C. Troncoso, R. Greenstadt, D. McCoy, Under the underground: Predicting private interactions in underground forums, arXiv preprint arXiv:1805.04494 (2018).

[41] D. R. Thomas, S. Pastrana, A. Hutchings, R. Clayton, A. R. Beresford, Ethical issues in research using datasets of illicit origin, in: Proceedings of the 2017 Internet Measurement Conference, ACM, 2017, pp. 445–462.

[42] Skype username tips, `https://itstillworks.com/skype-username-tips-22188.html`, last Accessed: 17-12-2020 (2017).

[43] Ipv4 vs ipv6: What's the difference?, `https://www.guru99.com/difference-ipv4-vs-ipv6.html`, last accessed: 17-12-2020 (2020).

[44] A. Bancroft, P. Scott Reid, Challenging the techno-politics of anonymity: the case of cryptomarket users, Information, Communication & Society 20 (4) (2017) 497–512.

[45] M. Paquet-Clouston, B. Haslhofer, B. Dupont, Ransomware payments in the bitcoin ecosystem, Journal of Cybersecurity 5 (1) (2019) tyz003.

[46] I. Ladegaard, We know where you are, what you are doing and we will catch you: Testing deterrence theory in digital drug markets, The British Journal of Criminology 58 (2) (2018) 414–433.

[47] S. Kethineni, Y. Cao, C. Dodge, Use of bitcoin in darknet markets: Examining facilitative factors on bitcoin-related crimes, American Journal of Criminal Justice 43 (2) (2018) 141–157.

[48] S. Meiklejohn, M. Pomarole, G. Jordan, K. Levchenko, D. McCoy, G. M. Voelker, S. Savage, A fistful of bitcoins: characterizing payments among men with no names, in: Proceedings of the 2013 conference on Internet measurement conference, 2013, pp. 127–140.

[49] S. Pastrana, G. Suarez-Tangil, A first look at the crypto-mining malware ecosystem: A decade of unrestricted wealth, in: Proceedings of the Internet Measurement Conference, 2019, pp. 73–86.

[50] Regular expression to "validate" a bitcoin address, `http://mokagio.github.io/tech-journal/2014/11/21/regex-bitcoin.html`, last accessed: 17-12-2020 (2014).

[51] A. Caines, S. Pastrana, A. Hutchings, P. J. Buttery, Automatically identifying the function and intent of posts in underground forums, Crime Science 7 (1) (2018) 19.

[52] J. Ramos, Using tf-idf to determine word relevance in document queries, in: Proceedings of the first instructional conference on machine learning, Vol. 242, New Jersey, USA, 2003, pp. 133–142.

[53] J. Hughes, B. Collier, A. Hutchings, From playing games to committing crimes: A multi-technique approach to predicting key actors on an online gaming forum, in: 2019 APWG Symposium on Electronic Crime Research (eCrime), IEEE, 2019, pp. 1–12.

[54] S. Pastrana, A. Hutchings, D. Thomas, J. Tapiador, Measuring ewhoring, in: Proceedings of the Internet Measurement Conference, 2019, pp. 463–477.

[55] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, D. McCoy, Doppelgänger finder: Taking stylometry to the underground, in: 2014 IEEE Symposium on Security and Privacy, IEEE, 2014, pp. 212–226.

[56] W. E. Winkler, String comparator metrics and enhanced decision rules in the fellegi-sunter model of record linkage., Tech. rep., ERIC (1990).

[57] W. W. Cohen, P. Ravikumar, S. E. Fienberg, et al., A comparison of string distance metrics for name-matching tasks., in: IIWeb, Vol. 3, 2003, pp. 73–78.

[58] S. kulkarni, Jaro winkler vs levenshtein distance, `https://srinivas-kulkarni.medium.com/jaro-winkler-vs-levenshtein-distance-2eab21832fd6`.
Last visited 19th July 2021 (March 2021).

[59] A. W. McDonald, S. Afroz, A. Caliskan, A. Stolerman, R. Greenstadt, Use fewer instances of the letter “i”: Toward writing style anonymization, in: Proceedings of the Privacy Enhancing Technologies Symposium, Springer, 2012, pp. 299–318.

[60] D. Dittrich, E. Kenneally, et al., The menlo report: Ethical principles guiding information and communication technology research, Tech. rep., US Department of Homeland Security (2012).