Doppelgänger Finder: Taking Stylometry To The Underground

Sadia Afroz*, Aylin Caliskan-Islam†, Ariel Stolerman†, Rachel Greenstadt† and Damon McCoy‡
*University of California, Berkeley †Drexel University ‡George Mason University

***Abstract*—Stylometry is a method for identifying anonymous authors of anonymous texts by analyzing their writing style. While stylometric methods have produced impressive results in previous experiments, we wanted to explore their performance on a challenging dataset of particular interest to the security research community. Analysis of underground forums can provide key information about who controls a given bot network or sells a service, and the size and scope of the cybercrime underworld. Previous analyses have been accomplished primarily through analysis of limited structured metadata and painstaking manual analysis. However, the key challenge is to automate this process, since this labor intensive manual approach clearly does not scale.**

We consider two scenarios. The first involves text written by an unknown cybercriminal and a set of potential suspects. This is standard, supervised stylometry problem made more difficult by multilingual forums that mix l33t-speak conversations with data dumps. In the second scenario, you want to feed a forum into an analysis engine and have it output possible doppelgängers, or users with multiple accounts. While other researchers have explored this problem, we propose a method that produces good results on actual separate accounts, as opposed to data sets created by artificially splitting authors into multiple identities.

For scenario 1, we achieve 77% to 84% accuracy on private messages. For scenario 2, we achieve 94% recall with 90% precision on blogs and 85.18% precision with 82.14% recall for underground forum users. We demonstrate the utility of our approach with a case study that includes applying our technique to the Carders forum and manual analysis to validate the results, enabling the discovery of previously undetected doppelgänger accounts.

I. Introduction

Underground forums are used as a rendezvous location for cybercriminals and play a crucial role in increasing efficiency and promoting innovation in the cybercrime ecosystem. These forums are frequently used by cybercriminals around the world to establish trade relationships and facilitate the exchange of illicit goods and services such as the sale of stolen credit card numbers, compromised hosts, and online credential theft. Linking different aliases to the same individual across sources of data to increase knowledge of a cybercriminal's activities is a powerful ability. An anecdotal example of this analysis performed manually is the case of the Rustock botnet operator where his accounts were manually linked together from multiple leaked data sources including underground forum posts [1]. All this information provides valuable insights, about how much he was earning, who else he was dealing with, which paints a fairly rich picture of a botnet operator's role in the underground cyber ecosystem.

Other information gleaned from underground forums is providing security researchers, law enforcement, and policy makers valuable information on how the market is segmented and specialized, the social dynamics of the community, and potential bottlenecks that are vulnerable to interventions. These advances have been accomplished primarily through analysis of limited structured metadata and painstaking manual analysis. Because of the size of the datasets and the labor intensity of the task, there are limitations to what can be accomplished by these techniques.

In fiction and folklore, a doppelgänger is an apparition or double of a living person. Many underground forums use the word *doppelgänger* to refer to a duplicate account of a user in the forum. The use of doppelgängers is forbidden in these forums because it undermines the fragile trust between pseudonymous users engaged in risky, illegal behavior and enables them to take advantage of each other. Users suspected of using multiple accounts are commonly banned. Understanding how and why users persist in maintaining multiple identities can help identify the dynamics of trust relationships in these forums. In this work we use stylometry, or linguistic analysis, to detect doppelgängers and study their use in these forums.

Linguistic analysis has recently been applied successfully to many security problems from using stylometry to identify anonymous bloggers [2], to using topic modeling to find job postings for web service abuse [3]. However, the underground forums present a particular challenge for text analytic techniques. The messages are short and tend to mix conversations with "products" such as credit card and bank account numbers, URLs, IP addresses, etc. Furthermore, the forums are written in a multilingual 133t-speak slang that renders most natural language processing tools such as part-of-speech taggers inaccurate—this language is often intentionally difficult to parse and speak even for native human speakers and serves to weed out outsiders. As such they represent a stress test of sorts for these approaches.

Our key contributions include:

1. Adapting authorship attribution to underground forums. Authorship attribution is useful in the scenario where an analyst has an unknown piece of text and wishes to attribute it to one out of a set of suspects. This scenario may be useful in underground analysis on its own, but we also use it as a subroutine in our multiple account detection algorithm.

Although some language-agnostic authorship attribution methods are available [4], [5] for this task, most of the highly accurate attribution methods [2], [6] are language specific for

standard English. We show that by using language-specific function words and parts-of-speech taggers, our authorship attribution method provides high accuracy even with over 1000 authors in difficult, foreign language texts. We create a feature set that incorporates the informal language, such as l33tsp34k, used in underground forums and data preprocessing methods that can remove non-conversational products from messages. These as a whole improve our accuracy by 10-15% beyond current state of the art methods directly applied to underground forums.

2. A general multiple author detection algorithm. Unlike standard authorship attribution, identifying doppelgängers is an unsupervised learning problem and requires novel methods where all pairs of accounts are compared against each other. Existing methods for this problem [7], [8] based on distance have been evaluated by artificially splitting authors into multiple identities. We find that these methods have reduced accuracy when applied to actual separate accounts—such as multiple blogs by the same author—and that improved methods are needed. Non-textual methods used to identify fraud or spam accounts are insufficient because they do not catch the high-value alternate identities used in these forums. Our approach *Doppelgänger Finder* evaluates all pairs of a set of authors for duplicate identities and returns a list of potential pairs, ordered by probability. This list can be used by a forum analyst to quickly identify interesting multiple identities. We validated our algorithm on real-world blogs using multiple separate blogs per author and using multiple accounts of members in different underground forums. Code for the algorithm is available at https://github.com/sheetal57/doppelganger-finder.

3. A practical manual analysis of an underground forum to identify previously unknown multiple identities. Using *Doppelgänger Finder* on a German carding forum Carders, we show how to discover and group unknown identities in cases when ground truth data is unavailable.

We discovered at least 10 new author pairs (and an additional 3 probable pairs) automatically which would have been hard to discover without time consuming manual analysis. These pairs are typically high value identities—in one case we found a user who created such identities for sale to other users on the forum. We end with an analysis of how and why these identities are created by these users and the purposes they serve in the forums.

II. Related Work

A. Underground Markets

Most of the past research on the underground market has focused on either analyzing structured metadata (i.e. social graphs, and trade ratings) in underground forums or performing a manual analysis of products and prices. One of the first studies by Franklin et al. performed an analysis of underground chat messages in public IRC channels to gain insight into prices and types of products traded [9]. Another study performed an analysis of an underground carders forum to understand how they propagate credentials in large scale data breaches [10]. A separate study explored how trust models were formed in underground forums [11], Yip et al. preformed an analysis of structural metadata in underground forums to examine the dynamics of social graphs in these communities [12]. Finally, another study did an analysis of activities taking place on Chinese underground markets [13]. McCoy et al. [14] analyzed the underground forums of three pharmaceutical affiliate programs and provided a detailed cost accounting of the overall business model. Recent research has investigated using underground market data to disrupt fraudulent activities. Thomas et al. identified patterns in fraudulent account usernames/emails by purchasing twitter accounts from an underground market [15].

When one forum is disrupted, these cybercriminals often create or join another forum using the same or different identities. Previous research tried to understand why these cybercriminals choose forums for doing their business [16] and what properties make underground forums sustainable [17]. We focus on a solution to identify when multiple accounts are controlled by the same person based on automated analysis of the unstructured message contents. Our research can help identify known cybercriminals by analyzing their conversation, even when they change online identities.

B. Authorship Attribution

Users are unique in many ways and an extensive amount of research exploits different aspects of behavior to deanonymize users in anonymized datasets. For example, a user can be identified based on how and what he types [18], his browser setup [19], which movie he prefers [20], who he connected with in a social network [21], when and what he writes in his blog or social network or on product reviews [22], [2], [23] and even how he fills bubbles in a paper form [24]. In the leaked underground forum, we only have the users' posts and their social network information. But deanonymizing these users using their social links from other social networks [21] is challenging as these relationships are ephemeral business relationships. Also, often these posts are from different time frames, so linking users using timing analysis, as previous work did to deanonymize flickr and twitter users is not possible [22], [20].

While stylometry has been applied to chat data in the past [6], large numbers of authors [2], as well as foreign language and translated texts [25], the combination of these properties in our data set is unique. The Writeprints [6] work evaluated their techniques on instant messaging chat logs from CyberWatch (www.cyberwatch.com). This data set is probably the closest to the forum data sets that we worked with. However, they had fewer words per author (an average of 1,422 words), but were in English. We were able to achieve better accuracy with more authors.

A few previous works explored the question of identifying multiple identities of an author. The Writeprints method can be used to detect similarity between two authors by measuring distance between their "writeprints." Qian et al.'s method, called "Learning by similarity," learns in the similarity space by creating a training set of similar and dissimilar docu-

ments [8] and comparing the distances between them. This method was evaluated using users who participated in Amazon book reviews. Almishari et al. [7] also used a similar distance-based approach using reviews from yelp.com to find duplicate authors. Koppel et al. [26] used a feature subsampling approach to detect whether two documents are written by the same author. But all of these methods were evaluated by creating *artificial* multiple identities per author by splitting a single author into two parts. In our experiments we noticed that identifying users writing about similar topics is easier than when they write about different topics. We evaluated our method on a real world blog dataset where users themselves created different identities in different blogs and in many cases different blogs by the same user were not about the same topic.

C. Detecting Fraudulent Accounts

Perito et al. [27] showed that most users use similar usernames for their accounts in different sites, e.g., daniele.perito and d.perito. Thus different accounts of a user can be tracked by just using usernames. This does not hold when the users are deliberately trying to hide their identities, which is often the case in the underground forums (example of usernames in multiple accounts are in Table XI). Usernames and other account information and behavior in the social network have often used to identify Sybil/spam accounts [28], [29], [30]. Our goal is different from these works as we are trying to identify duplicate accounts of highly active users, who would be considered as *honest* users in previous fraud detection papers. For example, these users are highly connected with other users in the forum, unlike spam/sybil accounts. Their account information (usernames, email addresses) are similar to spam accounts with mixed language, special characters and disposable email accounts, however, these properties hold for most users in these forums, even the ones who are not creating multiple identities.

III. Underground Forums

We analyzed four underground forums: AntiChat (AC), BlackhatWorld (BW), Carders (CC), L33tCrew (LC) (summarized in Table I). For each of these four forums we have a complete SQL dump of their database that includes user registration information, along with public and private messages. Each of these SQL forum dumps has been publicly "leaked" and uploaded to public file downloading sites by unknown parties.

A. Forums

This section gives an overview of the forums, in particular, it shows the relationship between a member's rank and his activities in the forum. In all forums, high-ranked members had more posts than low-ranked members. Access to special sections of these forums depends on a member's rank. Having the full SQL dump gives us the advantage of seeing the whole forum, which would have been unavailable if we had crawled the forums as an outsider or as a newly joined member. In general, the high-ranked users have more reputation, a longer post history, and consequently more words for our algorithms to analyze.

1) Antichat: Antichat started in May 2002 and was leaked in June 2010. It is a predominantly Russian language forum with 25871 active users (users with at least one post in the forum). Antichat covers a broad array of underground cybercrime topics from password cracking, stolen online credentials, email spam, search engine optimization (SEO), and underground affiliate programs.

Anybody with a valid email address can join the forum, though access to certain sections of the forum is restricted based on a member's rank. At the time of the leak, there were 8 advanced groups and 8 user ranks in our dataset[1]. A member of level N can access groups at level $\leq$ N. Admins and moderators have access to the whole forum and grant access to levels 3 to 6 by invitation. At the time of the leak, there were 4 admins and 89 moderators in Antichat.

Members earn ranks based on their reputation which is given by other members of the forum for any post or activity[2]. Initially each member is a *Beginner (*Новичок*)* [3], a member with at least 50 reputation is *Knowledgeable (*Знающий*)* and 888 reputation is a *Guru (*Гуру*)* (all user reputation levels are shown in Table II). A member can also get negative reputation points and can get banned. In our dataset there were 3033 banned members. The top reasons for banning a member are having multiple accounts and violating trade rules.

Rank	Rep.	Members	Members with $\geq$4500 words
Ламер (Lamer)	-50	646	22
Чайник (Newbie)	-3	340	4
Новичок (Beginner)	0	38279	553
Знающий (Knowledgeable)	50	595	256
Специалист (Specialist)	100	658	413
Эксперт (Expert)	350	271	177
Гуру (Guru)	888	206	153
Античатовец (Antichatian)	5555	1	1

Table II
ANTICHAT MEMBERS RANK

Antichat has a designated "Buy, Sell, Exchange" forum for trading. Most of the transactions are in WebMoney[4]. To minimize cheating, Antichat has paid "Guarantors" to guarantee product and service quality[5]. Sellers pay a percentage of the value of one unit of goods/services to the guarantor to verify their product quality. Members are advised not to buy non-guaranteed products. In case of a cheating, a buyer is paid off from the guarantor's collateral value.

2) BlackhatWorld: BlackhatWorld is primarily an English speaking forum that focuses on blackhat SEO techniques, started in October 2005 and is still active. At the time of the leak (May 2008) Blackhat had 4489 active members.

Like Antichat, anybody can join the forum and read most public posts. At the time of the leak, a member needed to pay

[1] http://forum.antichat.ru/thread17259.html
[2] Member rules are described https://forum.antichat.ru/thread72984.html
[3] Translated by Google translator
[4] http://www.wmtransfer.com/
[5] https://forum.antichat.ru/thread63165.html

Forum	Primary Language	Date covered	Posts	Private msgs	Users	Lurkers
Antichat (AC)	Russian	May 2002-Jun 2010	2160815	194498	41036	15165 (36.96%)
BlackhatWorld (BW)	English	Oct 2005-Mar 2008	65572	20849	8718	4229 (48.5%)
Carders(CC)	German	Feb 2009-Dec 2010	373143	197067	8425	3097(36.76%)
L33tCrew (LC)	German	May 2007-Nov 2009	861459	501915	18834	9306 (46.41%)

Table I
SUMMARY OF FORUMS

$25 to post in a public thread.[6] A member can have 8 ranks depending on his posting activities and different rights in the forum based on his rank. This rank can be achieved either by being active in the forum for a long period or by paying fees. A new member with less than 40 posts is a *Blacknoob* and 40-100 posts is a *Peasant*, both of these ranks do not have access to the "Junior VIP" section of the forum which requires at least 100 posts[7]. The "Junior VIP" section is not indexed by any search engines or visible to any non Jr. VIP members. At the time of the leak, a member could pay $15 to the admin to access this section. A member is considered active after at least 40 posts and 21 days after joining the forum. Member ranks are shown in Table III. The forum also maintains an "Executive VIP" section where membership is by invitation and a "Shitlist" for members with bad reputations. There were 43 banned members in our dataset. Most of the members in our BlackhatWorld dataset were Blacknoobs.

Rank	Members	Members with $\geq$4500 words
Banned Users	43	4
21 days 40 posts	7416	4
Registered Member	248	74
Exclusive V.I.Ps	7	7
Premium Members (PAID/Donated)	191	19
Admins and Moderators	8	8

Table III
BLACKHATWORLD MEMBERS RANK

In our dataset any member with over 40 posts was allowed to trade. This rule has currently been changed, now a member has to be at least a Junior VIP to trade in the BlackhatWorld marketplace, the "Buy, Sell, Trade" section[8]. Each post in the marketplace must be approved by an admin or moderator. In our dataset, there were 3 admins and 5 moderators. The major currency of this forum is USD. Paypal and exchange of products are also accepted.

3) Carders: Carders was a German language forum that specialized in stolen credit cards and other accounts. This forum was started in February 2009 and was leaked and closed in December 2010 [9].

At the time of the leak, Carders had 3 admins and 11 moderators. A regular member can have 9 ranks, but unlike other forums the rank was not dependent only on the number of posts (Table IV). Access to different sections of the forum was restricted based on rank. Any member with a verified email can be a *Newbie*. To be a *Full Member* a member needed at least 50 posts. A member had to be at least a *Full Member* to sell tutorials. *VIP Members* were invited by other high-ranked members. To sell products continuously a member needed a *Verified vendor* license which required at least 50 posts in the forum and €150+ per month. For certain products, for example, drugs and weapons, the license costs at least €200. Carders maintained a "Ripper" thread where any member can report a dishonest trader. A suspected ripper was assigned *Ripper-Verdacht!* title. Misbehaving members, for example, spammers, rippers or members with multiple accounts, were either banned temporarily or permanently depending on the severity of their action. In our dataset, there were 1849 banned members. The majority of the members in our Carders dataset are Newbie.

Rank	Members	Members with $\geq$4500 words
Nicht registriert (Not registered)	1	0
Email verification	323	1
Newbie	4899	23
Full Member	1296	431
VIP Member	7	6
Verified Vendor	16	6
Admins	14	13
Ripper-Verdacht! (Ripper suspected)	14	7
Time Banned	6	2
Perm Banned	1849	193

Table IV
CARDERS MEMBERS RANK

Other products traded in this forum were cardable shops (shops to monetize stolen cards), proxy servers, anonymous phone numbers, fake shipping/ delivery services and drugs. The major currencies of the forum were Ukash[10], PaySafeCard (PSC)[11], and WebMoney.

4) L33tCrew: Like Carders, L33tCrew was a predominantly carding forum. The forum was started in May 2007 and leaked and closed in Nov 2009. We noticed many users joined Carders after L33tCrew was closed. At the time of the leak, L33tCrew had 9528 active users.

L33tCrew member rank also depended on a member's activity and number of posts. With 15 posts a member was allowed in the base account area. The forum shoutbox, which was used to report minor problems or off topic issues, was visible to members with at least 40 posts. A member's ranking was based on his activity in the forum (Table V). On top of that, a member could have 2nd and 3rd level rankings. 100–150 posts were needed to be a 2nd level member. Members could rise to 3rd level after "proving" themselves in 2nd level and proving that they had non-public tools, tricks, etc. The

[6]The posting cost is now $30

[7]http://www.blackhatworld.com/blackhat-seo/misc.php?do=vsarules

[8]http://www.blackhatworld.com/blackhat-seo/bhw-marketplace-rules-how-post/387929-marketplace-rules-how-post-updated-no-sales-thread-bumping.html

[9]Details of carders leak at http://www.exploit-db.com/papers/15823/

[10]https://www.ukash.com/

[11]https://www.paysafecard.com/

proof included sending at least three non-public tools to the admin or moderators.

Rank	Min. posts	Members	Members with ≥4500 words
Newbie	0-30	715	93
Half-Operator	60	158	67
Operator	100	177	121
Higher Levels	150	412	398
Unranked Members	–	16410	679
Banned	–	913	197
Admins	–	11	11
Invited	–	33	8
Vorzeitig in der Handelszone	–	5	2

Table V
L33TCREW MEMBERS RANK

B. Member overlap

We identified common active users in the forums by matching their email addresses. Here "active" means users with at least one private or public message in a forum. Among the four forums, Carders and L33tCrew had 563 common users based on email addresses, among which 443 were active in Carders and 439 were active in L33tCrew. Common users in other forums are negligible.

C. Hiding identity

In all of the forums, multiple identities were strictly prohibited. On Carders and Antichat one of main reasons for banning a member is creating multiple identities. We wanted to check whether the users were taking any measures to hide their identities. We found several users were using disposable email addresses (562 in Carders, 364 in L33tCrew) from top well-known disposable email services, e.g. trashmail.com, owlpic.com, 20minutemail.com.

Carders used an alternative-ego detection tool (AE detector)[12] which saves a cookie of history of ids that log into Carders. Whenever someone logs into more than one accounts, it sends an automated warning message to forum moderators saying that the forum has been accessed from multiple accounts. The AE detector also warns the corresponding members. We grouped these multiple account holders based on whether or not they received these warning messages from the AE detector. We found 400 multiple identity groups with total 1692 members, where group size varies from 2 to 466 accounts (shown in Figure 1).

We suspect that the AE detector does not reflect multiple account holders perfectly. There are possible scenarios that would trigger the AE detector, e.g. when two members use a shared device to log into Carders or use a NAT/proxy. The corresponding users in these situations were considered as doppelgängers by the AE detector, which does not reflect the ground truth. Likewise, the AE detector may not catch all the alter egos, as some users may take alternate measures to log in from different sources. These suspicions were supported by our stylometric and manual analyses of Carders posts.

[12]http://www.vbulletin.org/forum/showthread.php?t=107566

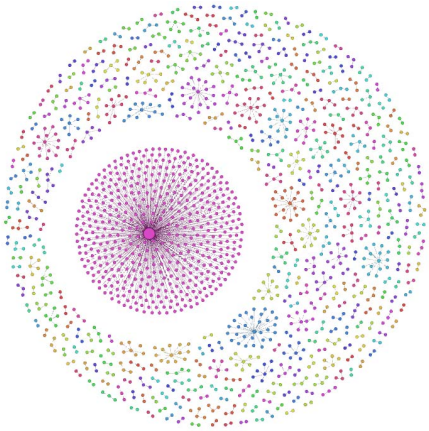

Figure 1. Duplicate account groups within Carders as identified by the AE detector. Each dot is one user. There is an edge between two users if AE detector considered them as duplicate user.

D. Public and private messages

In a forum a member can send public messages to public threads and private messages to other members. In our dataset we had both the public and private messages of all the members. Public messages are used to advertise/request products or services. In general, public messages are short and often have specific formats. For example, Carders specifies a specific format for public thread titles.

Private messages are used for discussing details of the products and negotiating prices. Sometimes members use their other email, ICQ or Jabber address for finalizing trades.

IV. AUTHORSHIP ATTRIBUTION

Our goal in this section is to see how well stylometry works in the challenging setting of underground forums and adapt stylometric methods to improve performance.

A. Approach

We consider a **supervised authorship attribution problem** that given a document D and a set of authors $\mathcal{A} = \{A_1, ..., A_n\}$ determines who among the authors in $\mathcal{A}$ wrote D. The authorship attribution algorithm has two steps: training and testing. During training, the algorithm trains a classifier using F features extracted from the sample documents of the authors in $\mathcal{A}$. In the testing step, it extracts features predefined in F from D and determines the probability of each author in $\mathcal{A}$ of being the author of D. It considers an author A_{max} to be the author of D if the probability of A_{max} being the author of D, $Pr(A_{max}\ wrote\ D)$, is the highest among all $Pr(A_i\ wrote\ D), i = 1, 2, ...n$.

k-attribution is the relaxed version of authorship attribution that outputs k top authors, ranked by their corresponding probabilities, $Pr(A_i\ \ wrote\ \ D)$, where $i = 1, 2, ...k$ and $k \leq n$.

B. Feature extraction

Our feature set contains lexical, syntactic and domain specific features. The lexical features include frequency of n-grams, punctuation and special characters. The syntactic features include frequency of language-specific parts-of-speech and function words. In our dataset we used English, German, and Russian parts-of-speech taggers and corresponding function words. For English and German parts-of-speech tagging we used the Stanford log-linear parts-of-speech tagger [31] and for Russian parts-of-speech tagging we used TreeTagger [32] with Russian parameters[13]. Function words or stop words are words with little lexical meaning that serve to express grammatical relationships with other words within the sentence, for example, in English function words are prepositions (to, from, for), and conjunctions (and, but, or). We used German and Russian stop words from Ranks.nl (http://www.ranks.nl/resources/stopwords.html) as function words. Similar feature sets have been used before in authorship analysis on English texts [6], [2], [33]. We modified the feature set for the multilingual case by adding language specific features. As the majority of the members use leetspeak in these forums, we used the percentage of leetspeak per document as a feature. Leetspeak (also known as Internet slang) uses combinations of ASCII characters to replace Latin letters, for example, leet is spelled as l33t or 1337. We defined leetspeak as a word with symbols and numbers and used regular expressions to identify such words.

Feature	Count
Freq. of punctuation (e.g. ',' '.')	Dynamic
Freq of special characters (e.g., '@', '%'	Dynamic
Freq. of character ngrams, n =1-3	150
Length of words	Dynamic
Freq. of numbers ngrams, n=1-3	110
Freq. of parts-of-speech ngrams, n=1-3	150
Freq. of word ngrams, n=1-3	150
Freq. of function words, e.g. for, to, the.	Dynamic
Percentage of leetspeak, e.g, l33t, pwn3d	-

Table VI
FEATURE SET

We used the JStylo [33] API for feature extraction, augmenting it with leetspeak percentage and the multilingual features for German and Russian.

C. Classification

We used a linear kernel Support Vector Machine (SVM) with Sequential Minimal Optimization (SMO) [34]. We performed 10-fold cross-validation, that is, our classifier was trained on 90% of the documents (at least 4500 words per author) and tested on the remaining 10% of the documents (at least 500 words per author). This experiment is repeated 10 times, each time randomly taking one 500-word document per author for testing and the rest for training. To evaluate our method's performance we use precision and recall. Here *true positive* for author A means number of times a document written by author A was correctly attributed to author A and *false positive* for author A means number of times a document written by any author other than A was misclassified to author A. We calculate per author precision/recall and take the average to show overall performance.

[13]http://corpus.leeds.ac.uk/mocky/

D. Removing product data

One of the primary challenges with this dataset is the mixing of conversational discussion with product discussions, e.g., stolen credentials, account information with passwords, and exploit code. This is particularly pronounced in the most active users who represent the majority of the trading activities. As the classifier relies on writing style to determine authorship, it misclassifies when two or more members share similar kinds of product information in their messages. Removing product information from conversation improved our classifier's performance by 10-15%. Identifying product information is also useful for understanding what kind of products are being traded in the forums.

Our product detector is based on two observations: 1) product information usually has repeated patterns, 2) conversation usually has verbs, but product information does not have verbs. To detect products, we first tag all the words in a document with their corresponding parts-of-speech and find sentence structures that are repeated more than a threshold of times. We consider the repeated patterns with no verbs as products and remove these from the documents.

To find repeated patterns, we measured Jaccard distance between each pair of tagged sentences. Due to errors in parts-of-speech tagging, sometimes two similar sentences are tagged with different parts-of-speech. To account for this, we considered two tagged sentences as similar if their distance is less than a threshold. We consider a post as a product post if any pattern is repeated more than three times. Note that our product detector is unsupervised and not specific to any particular kind of product, rather it depends on the structure of product information.

To evaluate our product detector we randomly chose 10,000 public posts from Carders and manually labeled them as product or conversation. 3.12% of the posts contained products. Using a matching threshold of 0.5 and repetition threshold of 3, we can detect 81.73% of the product posts (255 out of 312) with 2.5% false positive rate.[14]

E. Results

1) Minimum text requirement for authorship attribution: We trained our classifier with different numbers of training documents per author to see how much text is required to identify an author with sufficient accuracy. We performed this experiment for all the forums studied. In our experiments, accuracy increased as we trained the classifier with more words-per-author. On average, the accuracy did not improve when more than 4500 words-per-author were used in training (Figure 2).

[14]Note that false positives are not that damaging, since they only result in additional text being removed.

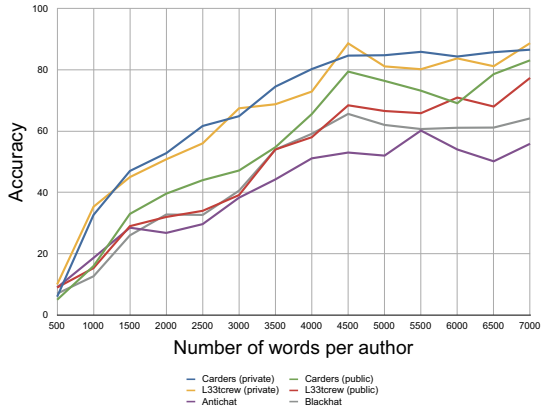

Figure 2. **Effect of number of words per user on accuracy**

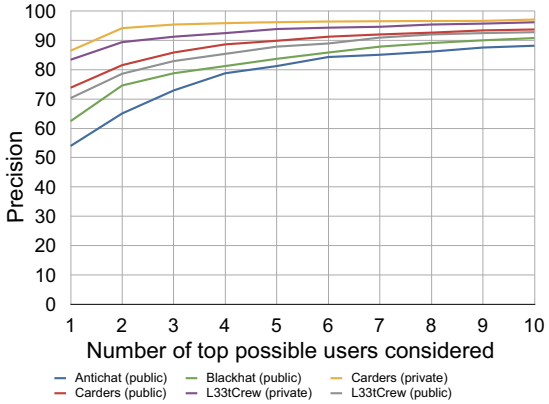

Figure 3. User attribution on 50 randomly chosen authors.

2) Attribution within forums: Many users were removed from the data set due to insufficient text, especially after products and data dumps were removed. Table VII shows the number of authors remaining in each forum and our results for authorship attribution in each forum which are mostly the high ranked members (section III-A). Results are for the public and private messages respectively. Aside from this, performance on private messages ranged from 77.2% to 84% precision. Recall results were similar, as this is a multi-class rather than a binary decision problem and precision for all authors was averaged (a false positive for one author is a false negative for another author). This is comparable to results on less challenging stylometry problems, such as English language emails and essays [6]. Performance on public messages, which were shorter and less conversational— more like advertising— was worse, ranging from 60.3% to 72%. The product detection and changes to the features set we made increased the overall accuracy by 10-15% depending on the setting.

However, it is difficult to compare the performance across different forums due to the differing number of authors in each forum. To compare performance in different forums we randomly chose 50 authors from each forum and performed k-attribution. Figure 3 shows the results of k-attribution for $k = 1$ to $k = 10$ where the $k = 1$ case is strict authorship attribution. This result shows that the differences between private and public messages persist even in this case and that the accuracy is not greatly affected when the number of authors scale from 50 to the numbers in Table VII. Furthermore, we found that the results are best for the Carders forum. The higher accuracy for Carders and L33tCrew may be due to the more focused set of topics on these forums or possibly the German language. Via manual analysis, we noted that the part-of-speech tagger we used for Russian was particularly inaccurate on the Antichat data set. A more accurate parts-of-speech tagger might lead to better results on Russian language forums.

Relaxed or k-attribution is helpful in the case where stylometry is used to narrow the set of authors in manual analysis. As we allow the algorithm to return up to 10 authors, we can increase the precision of results returned to 96% in the case of private messages and 90% in the case of public messages.

Forum	Public		Private	
	Members	**Precision**	**Members**	**Precision**
AntiChat	1459	44.4%	25	84%
Blackhat	81	72%	35	80.7%
Carders	346	60.3%	210	82.8%
L33tCrew	1215	68.8%	479	77.2%

Table VII
AUTHOR ATTRIBUTION WITHIN IN A FORUM.

F. Importance of features

To understand which features were the most important to distinguish authors, we calculated the Information Gain Ratio (IGR) [35] of each feature F_i over the entire dataset:

$$IGR(F_i) = (H(A) - H(A|F_i))/H(F_i) \quad (1)$$

where A is a random variable corresponding to an author and H is Shannon entropy.

In all the German, English and Russian language forums punctuation marks (comma, period, consecutive periods) were some of the most important features (shown in Table VIII). In German and English forums leetspeak percentage was highly ranked. Interestingly, similar features are important across different forums, even though the predominant languages of the forums are different.

V. DETECTING MULTIPLE IDENTITIES

In a practical scenario, an analyst may want to find any probable set of duplicate identities within a large pool of authors. Having multiple identities per author is not uncommon, e.g., many people on the Internet have multiple email addresses, accounts on different sites (e.g. Facebook, Twitter, G+) and blogs. Grouping multiple identities of an author is a powerful ability as the easiest way to change identity on the Internet is to create a new account.

[15]mfg is an abbreviation of a German greeting "Mit Freundlichen Gruessen" (English: sincerely yours).

[16]German subordinating conjunctions (e.g. weil (because), daß (that), damit (so that))

German forums	English forums	Russian forums
Char. trigram: mfg [15]	Punctuation: (')	Char. 1-gram: (ё)
Punctuation: Comma	Punctuation: Comma	Function word: ещё (English: more)
Leetspeak	Foreign words	Punctuation: Dot
Punctuation: Dot	Leetspeak	Char. 3-grams: ени
Char 3-gram:(...)	Function word: i'm	Char. bigrams: (,)
Nouns	Punctuation: Dot	Word-bigrams:что бы (English: that would)
Uppercase letters	POS-bigram (Noun,)	
Function word: dass (that)	Char. bigram: (,)	
Conjunctions [16]		
Char. 1-gram: ∧		

Table VIII
FEATURES WITH HIGHEST INFORMATION GAIN RATIO IN DIFFERENT FORUMS.

Grouping all the identities of an author is not possible by using only the traditional supervised authorship attribution. A supervised authorship attribution algorithm, trained on a set of unique authors, can answer who, among the training set, is the author of an unknown document. If the training set contains multiple identities of an author, supervised AA will identify only one of the identities as the most probable author, without saying anything about the connection among the authors in the training set.

A. Approach

The goal of our work is to identify multiple identities of an author. We leverage supervised authorship attribution to group author identities. For each pair of authors A and B we calculate the probability of A's document being attributed to B ($Pr(A \rightarrow B)$) and B's document being attributed to A ($Pr(B \rightarrow A)$). We consider A and B are the same if the combined probability is greater than a threshold. To calculate the pairwise probabilities, for each author $A_i \in \mathcal{A}$ we train a model using all other authors in $\mathcal{A}$ except A_i and test using A_i. The algorithm is described in Procedure 1. We call this method *Doppelgänger Finder*.

This method can be extended to larger groups. For example, for three authors A, B and C we compute P(A==B), P(B==C) and P(C==A). If A=B and C=B, we consider A, B and C as the three identities of one author.

Procedure 1 *Doppelgänger Finder*

Input: Set of authors $\mathcal{A} = A_1, ..A_n$ and associated documents, D, and threshold t
Output: Set of multiple identities per authors, M

```
F ⇐ Add weight k with every feature frequency (default k=10)
F' ⇐ Features selected using PCA on F
▷ Calculate pairwise probabilities
for A_i ∈ 𝒜 do
    n = Number of documents written by A_i
    C ⇐ Train on all authors except A_i using F'
    R ⇐ Test C on A_i (R contains the probability scores per author.)
    for A_j ∈ R do
        Pr(A_i → A_j) = (Σ_{x=1}^{n} Pr(A_jx)) / n
    end for
end for
▷ Combine pairwise probabilities
for (A_i, A_j) ∈ 𝒜 do
    P = Combine(Pr(A_i → A_j), Pr(A_j → A_i))
    if P > t then
        M.add(A_i, A_j, P)
    end if
end for
return M
```

B. Feature extraction

To identify similarity between two authors we use the same features used for regular authorship attribution (Table VI), with two exceptions: 1) exclude the word n-grams because this makes the feature extraction process much slower without any improvement in the performance; and 2) instead of limiting the number of other n-grams, we use all possible n-grams to increase the difference between authors, e.g., if author A uses a bi-gram "ng" but author B never uses it, then "ng" is an important feature to distinguish A and B. If we include all possible n-grams instead of only the top 50, we can catch many such cases, especially the rare author-specific n-grams.

After extracting all the features, we add weight to the feature frequencies to increase distance among authors. This serves to increase the distance between present and not present features and gives better results. As our features contain all possible n-grams, the total number of features per dataset is huge (over 100k for 100 authors). All the features are not important and they just make the classification task slower without improving the accuracy. To reduce the number of features without hurting performance, we use Principal Component Analysis (PCA) to weight and select only the features with high variance.

Principal component analysis (PCA) is a widely used mathematical tool for high dimension data analysis. It uses the dependencies between the variables to represent the data in a more tractable, lower-dimensional form. PCA finds the variances and coefficients of a feature matrix by finding the eigenvalues and eigenvectors. To perform PCA, the following steps are performed:

1) Calculate the covariance matrix of the feature matrix F. The covariance matrix measures how much the features vary from the mean with respect to each other. The covariance of two random variables X and Y is:

$$cov(X,Y) = \sum_{i=1}^{N} \frac{(x_i - \bar{x})(y_i - \bar{y})}{N} \quad (2)$$

where $\bar{x} = mean(X)$, $\bar{y} = mean(Y)$ and N is the total number of documents.

2) Calculate eigenvectors and eigenvalues of the covariance matrix. The eigenvector with the highest eigenvalue is the most dominant principle component of the dataset (PC1). It expresses the most significant relationship between the data dimensions. Principal components are calculated by multiplying each row of the eigenvectors with the sorted eigenvalues.
3) One of the reasons for using PCA is to reduce the number of features by finding the principal components of input data. The best low-dimensional space is defined as having the minimal error between the input dataset and the PCA (eq. 3).

$$\frac{\sum_{i=1}^{K} \lambda_i}{\sum_{i=1}^{N} \lambda_i} > \theta \quad (3)$$

where K is the selected dimension, N is the original dimension and λ is an eigenvalue. We chose $\theta = 0.999$ so that the error between the original dataset and the projected dataset is less than 0.1%.

C. Probability score calculation

We use Logistic regression with 'L1' regularization and regularization factor $C = 1$ as a classifier in Procedure 1 to calculate pairwise probabilities. We experimented with linear kernel SVM, which was slower than Logistic regression without any performance improvement. Any machine learning method that gives probability score can be used for this. After that we need to calculate $P(A == B)$ by combining the two probabilities: $P(A \rightarrow B)$ and $P(B \rightarrow A)$. We experimented with three ways of combining the probabilities:

1) Average: Given two probabilities $Pr(A \rightarrow B)$ and $Pr(B \rightarrow A)$, combined score is $\frac{Pr(A \rightarrow B) + Pr(B \rightarrow A)}{2}$.
2) Multiplication: Given two probabilities, combined score is $Pr(A \rightarrow B) * Pr(B \rightarrow A)$. We can consider the two probabilities as independent because when $Pr(A \rightarrow B)$ was calculated A was not present in the training set. Similarly B was not present when $Pr(B \rightarrow A)$ was calculated. Also in this case if any of the one-way probabilities is zero, the combined probability would be zero.
3) Squared average: The combined score is $\frac{Pr(A \rightarrow B)^2 + Pr(B \rightarrow A)^2}{2}$.

All the three approaches give similar precision/recall. We finally used the multiplication approach as its performance is slightly higher in the high recall region.

D. Baseline

We implement two distance based methods, as suggested by previous work, to compare our performance.

1) Unsupervised: Calculate the euclidean distance between any two authors. Choose a threshold. Two authors are the same if the distance between them is less than the threshold.
2) Supervised: Train a classifier using the euclidean distance between any two authors in the training set. Test it using the euclidean distance between the authors in the test set.

We use the same features and classifiers for both our method and the baseline method. Note that, we did not try different feature sets and weighting schemes to improve accuracy. The distance method might provide different results with different feature sets and classifiers.

E. Evaluation

1) Data: To evaluate *Doppelgänger Finder* we used a real world blog dataset used in the Internet scale authorship experiment by Narayanan et al.[2]. These blogs were collected by scanning a dataset of 3.5 million Google profile pages for users who specify multiple blogs. From this list of blog URLs, RSS feeds and individual blog posts were collected, filtered to remove HTML and any other markups and only the blogs with at least 7500 characters of text across all the posts were retained. This resulted in total 3,628 Google profiles where 1,663 listed a pair of blogs and 202 listed three to five blogs.

Out of the 1,663 pairs of blogs, many were group blogs with more than one author. We removed the group blogs from the dataset and then manually verified 200 blogs written by 100 authors. Each author in the dataset has at least 4500 words. Among the 200 blogs, we used 100 blogs as our development dataset, we call it **Blog-dev** and the other 100 as a test dataset **Blog-test**. We use the Blog-dev dataset to measure the effect of different feature sets and probability scores. The Blog-test dataset is used to verify that our method provides similar performance on different datasets. The two sets are mutually exclusive.

2) Methodology: To evaluate our method's performance we use precision and recall. Note that, this is a binary task, not multiclass classification discussed in section IV. The precision-recall curve (PR curve) shows the precision and recall values at different probability scores. We chose the PR curve instead of ROC curve as we have more false cases (no match between two authors) than true cases, which makes the false positive rate very low even when the number of false positive is very high[17] Area under a curve (AUC) value shows area under the PR curve. Higher value of AUC denotes better performance.

3) Result: Figure 4 shows the precision-recall curve for Blog-dev using different feature sets. The algorithm performs best when all the features are used, although only one feature class, e.g., char n-grams or function words, also gives high performance. All features give higher combined probability scores than one feature set (Figure 5). The combined probability scores are high when two authors are the same and

[17]For example, in the case of 100 authors with 50 true pairs, number of true cases is 50 but number of false cases is 10000-50=9950. So, the false positive rate would be 1% even when number of false positives is 100.

low (almost zero) when they are not (Figure 7). Our method has similar performance on the Blog-test set of 100 authors with 50 pairs (Figure 6). On average, distances between two blogs written by the same author is 0.0001, which is lower than when the blogs are from different authors (0.0003). The distance based method performs much worse than our method on Blog-test set, specially the supervised method performs similar to a random classifier.

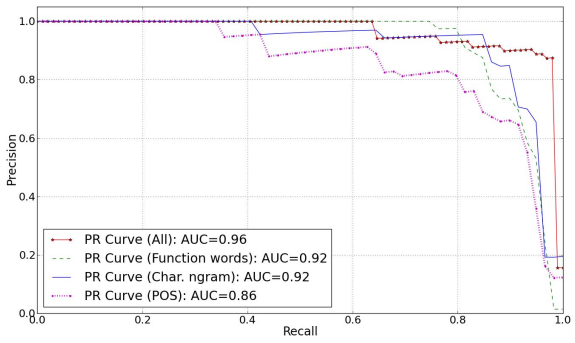

Figure 4. *Doppelgänger Finder*: Precision/Recall curve on Blog-dev dataset.

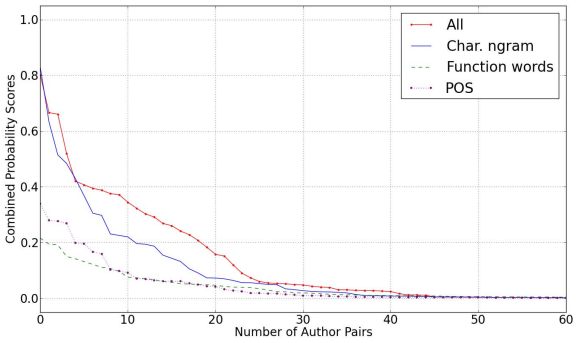

Figure 5. Probability scores on Blog-dev dataset.

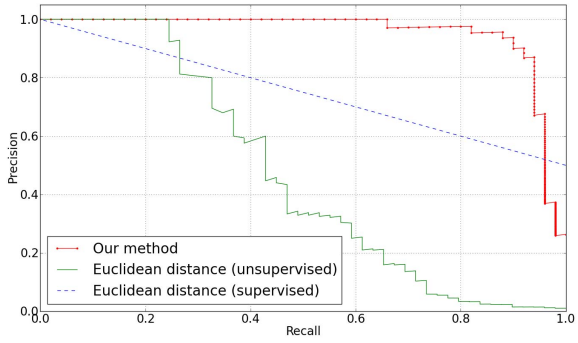

Figure 6. Comparing *Doppelgänger Finder* on Blog-test dataset.

F. Discussion

The goal of our method is to identify possible multiple identities from a dataset by ranking the author pairs in case where any training set is unavailable. However, the actual score may vary depending on the properties of the dataset, such as size of the dataset and language of the text. For example, in the Blog-dev dataset the threshold of *0.004* gave the best performance (Table 5), but in the Blog-test set 0.003 provided the best recall. The recommended approach of using it for manual analysis is to plot the probability curve (as in Figure 5) and verify author pairs in decreasing order. We provide a detailed manual analysis of an underground forum in the following section.

We also experimented with unsupervised clustering algorithms like k-Nearest Neighbour with k=2, but it could cluster 6 out of 50 pairs of blogs.

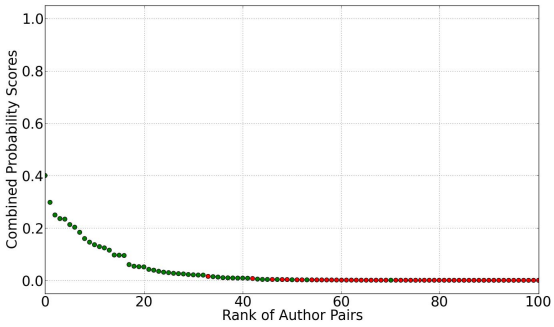

Figure 7. Combined probability scores on Blog-test dataset. Each dot represents a pair of blogs, green dot means both blogs belong to the same person and red dot means the blogs belong to different people. This graph shows when both of the blogs belong to the same person the probability scores are higher than when the blogs belong to different authors.

Dataset	Threshold	Precision	Recall
Blog-dev	**0.004**	**0.90**	**0.94**
	0.01	0.91	0.82
	0.04	1.0	0.64
Blog-test	**0.003**	**0.90**	**0.92**
	0.004	0.95	0.88
	0.01	0.95	0.78
	0.04	1.0	0.46
L33tCrew-Carders	**0.004**	**0.85**	**0.82**
	0.01	0.87	0.71
	0.04	0.92	0.39

Table IX
PRECISION-RECALL AT DIFFERENT THRESHOLDS. THRESHOLD IN **BOLD** GIVES THE BEST PERFORMANCE.

VI. Multiple Identities in Underground Forums

In this section we show how our method can be used to identify duplicate accounts by performing a case study on the underground forums. In the forums, many users create multiple identities to hide their original identity (reasons for doing so are discussed later) and they do so by changing the obvious identity indicators, e.g. usernames and email addresses. So we did not have any strong ground truth information for the multiple identities in a forum. We do, however, have some common users across two forums. We treat the common identities in different forums as one dataset and use that to

evaluate *Doppelgänger Finder* in underground forum. After that we run it on a forum and manually verify our results.

A. Multiple identities across forums

We collected users with same email address from L33tCrew and Carders. We found 563 valid common email addresses between these two forums. Among them, 443 users were active (had at least one post) in Carders and 439 were active in L33tCrew. Out of these 882 users, 179 had over 4500 words of text. We performed *Doppelgänger Finder* on these 179 authors which included 28 pairs of users (the rest of the 123 accounts did not have enough text in the other forum so merely served as distractor authors for the algorithm). Our method provides 0.85 precision and 0.82 recall when the threshold is 0.004 with exactly 4 false positive cases (Table IX and Figure 8).

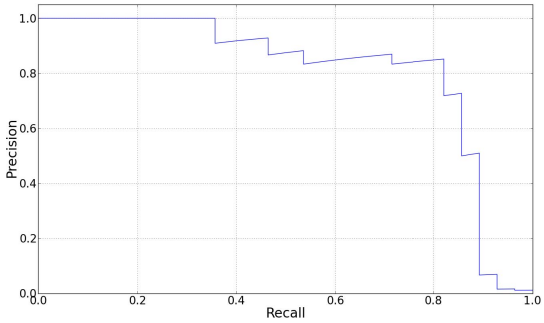

Figure 8. *Doppelgänger Finder*: With common users in Carders and L33tCrew: 179 users with 28 pairs. AUC is 0.82.

B. Multiple identities within forum

We used *Doppelgänger Finder* on Carders and manually analyzed the member-pairs with high scores to show that they are highly likely to be the same user. We selected all the Carders users with at least 4500 words in their private messages, which resulted total 221 users. We chose only private messages as our basic authorship attribution method was more accurate in private messages than in public messages. After that we ranked the member pairs based on the scores generated by our method. The highest combined probability score of the possible pairs is 0.806 and then it goes down to almost zero after the first 50 pairs (Figure 9).

1) Methodology: Table X shows the criteria we use to validate the possible doppelgängers. We manually read their private and public messages in the forum and information used in the user accounts to extract these features. The first criterion is to see if two users have the same ICQ numbers a.k.a UINs which is used by most traders to discuss details of their transactions. ICQ's are generally exchanged in private messages. Our second criterion is to match signatures. In all the forums users can enable or disable a default signature on their forum profiles. Signatures could be generic abbreviations of common phrases such as 'mfg,' or 'Grüße' or pseudonyms in the forum. We also investigate the products traded, payment methods used, topics of messages, and user information in the

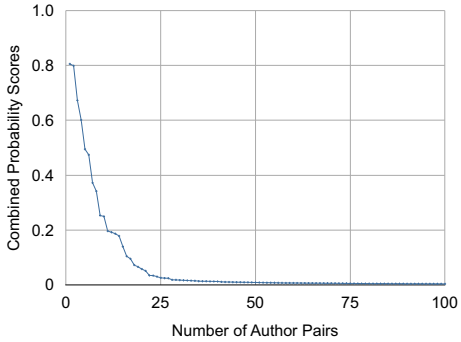

Figure 9. Combined probability scores of the top 100 pairs from Carders.

Criteria	Description
Username	Whether their usernames are same
ICQ	If two users have the same ICQ numbers
Signature (Sig.)	Whether they use the same signatures
Contact Information	Phone number and other contact information shared
Acc. Info	Information in the user table, e.g, their group membership, join and ban date, activity time
Topics	Their topic of discussion
OR AE	At least one of the users trigger the AE detector.
Interaction (Intr.)	Do they talk with each other?
Other	Other identity indicators, e.g., users mention their other accounts or the pair is banned for having the same IP address.

Table X
CRITERIA FOR VERIFYING MULTIPLE ACCOUNTS

user table, e.g., join date, banned date if banned, rank in the forum and groups the user joined. We check whether or not they set off the Alter-Ego detector on Carders. Lastly we check whether or not members in a pair sent private messages to each other because that would indicate that they are likely not the same person. We understand that there are many ways to verify identity but in most cases these serve as good indicators.

The *Doppelgänger Finder* algorithm considered $\binom{221}{2}$ possible pairs. We chose all the pairs with score greater than 0.05 for our manual analysis (21 pairs). We limit our analysis to limit the number of pairs to analyze as it could be quite time consuming. We also chose three pairs with low score (rank 22-24 in Table XI) to illustrate that higher score pairs are more likely to be true pairs (belong to the same person) than the lower score pairs. Note that, all of the top possible doppelgängers use completely different usernames. To protect the members' identity we only show the first three letters of their usernames in Table XI.

There are five possible outcomes of our manual analysis: True, Probably True, Unclear, Probably False and False. *True* indicates that we have conclusive evidence that the pair is doppelgängers, e.g., sometimes the pair themselves admit in their private/public messages about their other accounts or the pair shares same IM/payment accounts. *Probably True* indicates that the members share similar uncommon attributes but there are no conclusive evidences of them being the same.

Unclear indicates that some criteria are similar in both and some are very different and there are no conclusive attributes either way. *Probably False* means there are very few to no similarity between the members but no evidence that they are not the same. *False* indicates that we found conclusive evidence that the members in a pair are not the same, e.g., the members trade with each other [18].

2) Result and Discussion: We found that in Carders, as in the blog and cross-forum experiments, the accounts produced at the high end of the probability range were doppelgängers. The 12 pairs with the highest probabilities were assessed as **True** or **Probably True**. After that, there is a range where both the manual and linguistic evidence is thinner but nonetheless contains some true pairs (pairs 13-17). The manual analysis suggested that pairs below this probability threshold were likely not doppelgängers. Thus, our manual analysis overall agreed with the linguistic analysis performed by *Doppelgänger Finder*. In the following sections we give detailed examples of the five cases.

*a) **True**:* True cases are particularly seen when users explicitly state their identities and/or use the same ICQ numbers in two separate accounts. For example, each pair of users in **Pair 1-3, 5, 6, 8, 9, 10, and 16** provides an ICQ number in their private messages that is unique to that pair. The users in **Pair-11** use the same jabber nickname. One of the users in **Pair 1** (user name **per****) was asking the admins to give his other account back and telling other members that he is *Smi***.

Other cases had just as convincing, but more subtle evidence. The accounts in **Pair-8** both use *trashmail* which provides disposable email addresses, which shows that these users are careful about hiding their identities. However, the most convincing evidence of their connection was a third doppelgänger account, which we will call user-8c, who did not have enough text to be in our initial user set, but was brought to our attention by the linguistic similarity between the accounts in Pair-8. Both users in Pair-8 share the same ICQ number with user-8c. User-8b explicitly writes two messages from User-8c's account, one in Turkish and one in English revealing his user-8b username. These users do not send private messages to each other. These findings imply that the three user accounts belong to the same person.

*b) **Probably True**:* These accounts do not have a "smoking gun" like a shared ICQ number or Jabber account, however, we are able to observe that the accounts shared have similar interests or other properties. We consider how common these similar properties are in the entire forum and assess as probably true accounts that share uncommon properties.

In the case of **Pair-4**, user-4a does not have an ICQ number, but user-4b frequently gives out an ICQ number. User-4a wants to buy new ICQ numbers. This suggests that he uses ICQ and hides his own ICQ number. They both use a similar signature: 'mfg', but this is common. They trade similar products and talk about similar topics such as Kokain and D2 numbers. Since these are not common, this suggests they might be the same user. User-4a is a newbie while user-4b is a full member. The accounts were active during the same period.

The accounts in **Pair-7** have different ICQ numbers. However, both user-7a and user-7b deal with online banking products, PS3, Apple products, Amazon accounts and cards. They both use Ukash. They both use the same signature such as 'grüße' or 'greezz'. User-7a is a full member and user-7b is permanently banned. They have both been active account holders at the same period. User-7a has a 13th level reputation and user-7b has a 11th level reputation.

Similarly, the accounts in **Pair-12** use the same, rare signature 'peace' and both are interested in weed.

*c) **Unclear**:* The accounts in **Pair-13** do not have common ICQ numbers, even though they have the same ICQ numbers with other users (suggesting they do use doppelgänger accounts with lower text, lower reputation accounts). User-13a is a full member with a reputation level of 8. User-13b is a full member with a reputation level of 15. User-13a's products are carding, ps, packstation, netbook, camcorder, and user-13b's products are carding, botnets, cc dumps, xbox, viagra, iPod.

*d) **Probably False**:* The **Pair-14** accounts have different ICQs. User-14a products are tutorials, accounts, Nike, ebay and ps. User-14b's products are cameras and cards. User-14a is a full member with reputation level of 5. User-14b is permanently banned with a reputation level of 15.

One of the users in **Pair-17**, User-17b shares two ICQ numbers with another user but not with User-17a. User-17a's products are iPhone, iPad, macbook, drops, and paypal and User-17b's products are paypal, iPhone, D2 pins, and weed.

*e) **False**:* These users have specific and different signatures and also they use different ICQ numbers. These accounts sometimes interact, suggesting separate identities.

Pairs such as **20** send each other private messages to trade and complete a transaction, suggesting they are business partners not doppelgängers.

The accounts in **Pair-24** do not have any common UINs. They have different signatures, User-24a uses the signature 'LG Carlos' and 'Julix' interchangeably. User-24b never uses 'Carlos' or 'Julix' but he sometimes uses 'mfg' or 'DingDong' at the end of his messages. User-24a's products are iPhone, ebay, debit, iTunes cards, drop service, pack station, fake money while User-24b's products are camera, ps3, paypal, cards, keys, eplus, games, perfumes. They do not talk to each other.

Pair-21 is a special case of false labels. User-21a and user-21b are group accounts shared by both of the users. In one private message User-21a told user-21b: "*You think it is good that they think we are the same.*", because they got a warning from the admins for using the same computer. They also stated that they were meeting at each other's houses in person for business, which implies that they might be using the same accounts. They sent many messages to other people mentioning each other's names to customers.

[18]It is possible for a member to generate fake trades between his two accounts to prove uniqueness of the accounts. For the purpose the analysis we assume that is unlikely as we do not have any evidence of this happening.

Rank	Score	Usernames	ICQ	Sig.	Contact	Acc.	Topics	OR AE	Other	Intr.	Decision
1	0.806	per**, Smi**	X		icq		weed	X	X	0	True
2	0.799	Pri**, Lou**	X					X	X	0	True
3	0.673	Kan**, deb**	X					X		0	True
4	0.601	Sch**, bob**	–	mfg	–		Kokain	–		0	Probably True
5	0.495	Duk**, Mer**	X	–				–		0	True
6	0.474	Dra**, Pum**	X					X	X	0	True
7	0.372	p01**, tol**	–	greezz			X	–		0	Probably True
8	0.342	Qui**, gam**	X			X		X		0	True
9	0.253	aim**, sty**	X					X		0	True
10	0.250	Un1**, Raz**	X					X	X	0	True
11	0.196	PUN**, soc**	–		Jabber		X	–	X	0	True
12	0.192	Koo**, Wic**	–	peace		X	weed	X		0	Probably True
13	0.187	Ped**, roc**	–				X	–		0	Unclear
14	0.178	Tzo**, Haw**	–				X	X		0	Probably False
15	0.140	Xer**, kdk**	–			X	X	X		0	Unclear
16	0.105	sys**, pat**	X					X		0	True
17	0.095	Xer**, pat**	–			–	X	X		0	Probably False
18	0.072	Qui**, Sco**	–					X		0	False
19	0.066	Fru**, DaV**	–			–	–	–		0	Probably False
20	0.058	Ber**, neo**	–							5	False
21	0.051	Mr.**, Fle**	–					X	X	26	False*
22	0.01	puT**, pol**	–	–	–	–	–	–		0	False
23	0.001	BuE**, Fru**	–	–	–	–	–	–		0	False
24	0.0001	Car**, Din**	–	–	–	–	–	–		0	False

Table XI
MANUAL ANALYSIS OF USERS: X INDICATES SAME, – INDICATES DIFFERENT, EMPTY MEANS THE RESULT IS INCONCLUSIVE OR COMPLICATED WITH MANY VALUES.

VII. DISCUSSION

A. Lessons learned about underground markets

Doppelgänger Finder helped us detect difficult to detect dopplegänger accounts. We performed a preliminary analysis on L33tCrew and Blackhat and found similar results as Carders. Our manual analysis of these accounts improves our understanding of why people create multiple identities in underground forums, either within or across forums.

Banning. Getting banned in a forum is one of the main reasons for creating another account within a forum. Rippers, spammers or multiple account holders get penalized or banned once the admins become aware of their actions. Users with penalties get banned once their infraction points go over a certain threshold. There are hundreds of users within forums that have been banned and they open new accounts to keep actively participating in the forums. Some of the new accounts get banned again because the moderators realize that they have multiple accounts, which is a violation of forum rules.

Sockpuppet. Some forum members create multiple accounts in order to raise demand and start a competition to increase product prices.

Accounts for sale. Some users maintain multiple accounts and try to raise their reputation levels and associate certain accounts with particular products and customers. Once a certain reputation level is reached, they offer to sell these extra accounts.

Branding. Some users appear to setup multiple accounts to sell different types of goods. One reason to do this is if one class of goods is more risky, such as selling drugs, the person can be more careful about protecting his actual identity when using this account. Another reason to do this might be to have each account establish a "brand" that builds a good reputation selling a single class of goods.

Cross-forum accounts. Many users have accounts in more than one forums potentially as a method to grow in their sales by reaching more people not present on the same forums and to purchase goods not offered in a single forum.

Group accounts. In some cases groups of people work together as an organization and each member is responsible for a specific operation among a variety of products that are traded across different accounts. How to adapt stylometry algorithms to deal with multi-authored documents is an open problem that is left as future work.

B. Lessons learned about Stylometry

We found that any stylometric method can be used in non-English languages by using a high quality parts-of-speech tagger and function words of that language. We have access to one more forum called *BadhackerZ* whose primary language is transliterated Hindi using English letters. We did not have a POS tagger that could handle the mixture of these two languages. We were not able to get meaningful results by applying stylometry to *BadhackerZ*, therefore we excluded this forum from stylometric analysis. Similarly, the Russian POS tagger we used produced poor results on our dataset. POS tags generally have high information gain in stylometric analysis and as a result play a crucial role in stylometry. Future work might involve experimenting with other POS taggers or improving their efficacy by producing manually annotated samples of forum text.

C. Doppelgänger detection by forum administators

One of the primary reasons for banning accounts on these underground forums is because of users creating multiple accounts. This shows that forum administators are actively looking for these types of accounts and removing them since they can be used to undermine underground forums. They use a number of methods ranging from automated tools, such as AE detector, and more manual methods, such as reports from other members. As we have seen from analysis all of these methods are error prone and result in many false positives and false negatives. Many of the false positives were probably generated by users using proxies to hide their IP and location. In addition, when static tools with defined heuristics (IPs, browser cookies, etc.) are used to detect doppelgänger accounts' users can take simple precautions to avoid detection. Many of the accounts detected by doppelgänger finder were not detected by these methods potentially because that user was actively evading known detection methods.

D. Performance

Our method needs to run N classifiers for N authors. Each classifier is independent, thus can be run in parallel. Performance also depends on number of documents per author. Using only 4 threads on a quad core Macbook pro laptop the blog experiments with 100 authors and at least 9 documents per author took around 10 minutes and the underground forum experiment took around 35 minutes, which can be made faster with more threads.

E. Hybrid doppelgänger finder methods

Based on what we have learned from our manual analysis of our doppelgänger finder results on Carders, we could potentially build a hybrid method that integrates both stylometry and more underground specific features. For instance, some of the doppelgänger accounts could be identified with simple regular expressions that find and match contact information, such as ICQ numbers. In other cases manual analysis revealed more subtle features, such as two accounts selling the same uncommon product or talking about a similar set of topics, which can be a good indicator of being doppelgängers.

Custom parsers and pattern matchers could be created and combined with our doppelgänger finder tool to improve its results. However, it is difficult to know a priori what patterns to look for in different domains. Thus, using doppelgänger finder and performing manual analysis would make this task of designing and adding additional custom tools easier.

F. Methods to evade doppelgänger finder

There are several limitations to using stylometry to detect doppelgängers. The most obvious limitation is that it requires a large number of words from a single account. A forum member could stop using his account and create a new one before reaching this amount of text, but as pointed out in Section III parts of the forum are closed off to new/ less active members, thus less activity is not beneficial to them. They are often not allowed to engage in commerce until they have payed a fee and built up a good reputation by posting.

Another way to evade our method is for the author to intentionally change their writing style to deceive stylometry algorithms. As shown in previous research this is a difficult, but possible task [36], and tools such as Anonymouth can give hints as to how to alter writing style to evade stylometry [33]. We do not currently see any evidence of this technique being used by members of underground forums, but Anonymouth could be integrated into forums.

VIII. Conclusion

Doppelgänger Finder enables easy analysis of a forum for high-value multiple identities. Our analysis of Carders has already produced insights into the use of multiple identities within these forums. We have confidence that it can be applied to other forums, given the promising results on blogs and cross-forum accounts. This technique can also be used to detect multiple identities on non-malicious platforms.

This work also motivates the need for improved privacy enhancing technologies such as Anonymouth [33] for authors who wish to not have their pseudonymous writings linked.

Acknowledgment

We want to thank Vern Paxson, Ling Huang, Vaibhav Garg and the anonymous reviewers for their useful feedback. We also thank the authors of [2], especially Neil Zhenqiang Gong and Emil Stefanav, for providing us access to the blog dataset. This work is supported by Intel through the ISTC for Secure Computing, DARPA N11AP20024, National Science Foundation grant 1237076, 1253418 and CNS-1347151 and a gift from Google. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

References

[1] S. Savage, G. M. Voelker, J. Fowler *et al.*, "Beyond technical security: Developing an empirical basis for socio-economic prespectives," http://www.sysnet.ucsd.edu/frontier/proposal.pdf, 2012.

[2] A. Narayanan, H. Paskov, N. Gong, J. Bethencourt, E. Stefanov, R. Shin, and D. Song, "On the feasibility of internet-scale author identification," in *Proceedings of the 33rd conference on IEEE Symposium on Security and Privacy*. IEEE, 2012.

[3] D. Kim, M. Motoyama, G. Voelker, and L. Saul, "Topic modeling of freelance job postings to monitor web service abuse," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*. ACM, 2011, pp. 11–20.

[4] P. Juola, J. Sofko, and P. Brennan, "A prototype for authorship attribution studies," *Literary and Linguistic Computing*, vol. 21, no. 2, pp. 169–178, 2006.

[5] M. Koppel, J. Schler, and S. Argamon, "Computational methods in authorship attribution," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 1, pp. 9–26, 2009.

[6] A. Abbasi and H. Chen, "Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace," *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 1–29, 2008.

[7] M. Almishari, P. Gasti, G. Tsudik, and E. Oguz, "Privacy-preserving matching of community-contributed content," in *Computer Security–ESORICS 2013*. Springer, 2013, pp. 443–462.

[8] T. Qian and B. Liu, "Identifying multiple userids of the same author," in *EMNLP 2013*, 2013.

[9] J. Franklin, V. Paxson, A. Perrig, and S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants," in *ACM Conference on Computer and Communications Security (CCS)*, 2007.

[10] K. Peretti, "Data breaches: what the underground world of carding reveals," *Santa Clara Computer & High Tech. LJ*, vol. 25, p. 375, 2008.

[11] M. Motoyama, D. McCoy, K. Levchenko, S. Savage, and G. Voelker, "An analysis of underground forums," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 2011, pp. 71–80.

[12] M. Yip, N. Shadbolt, and C. Webber, "Structural analysis of online criminal social networks," in *Intelligence and Security Informatics (ISI), 2012 IEEE International Conference on*. IEEE, 2012, pp. 60–65.

[13] J. Zhuge, T. Holz, C. Song, J. Guo, X. Han, and W. Zou, "Studying malicious websites and the underground economy on the chinese web," *Managing Information Risk and the Economics of Security*, pp. 225–244, 2009.

[14] D. McCoy, A. Pitsillidis, G. Jordan, N. Weaver, C. Kreibich, B. Krebs, G. M. Voelker, S. Savage, and K. Levchenko, "Pharmaleaks: Understanding the business of online pharmaceutical affiliate programs," in *Proceedings of the 21st USENIX conference on Security symposium*. USENIX Association, 2012, pp. 1–1.

[15] K. Thomas, D. McCoy, C. Grier, A. Kolcz, and V. Paxson, "Trafficking fraudulent accounts: the role of the underground market in twitter spam and abuse," in *USENIX Security Symposium*, 2013.

[16] M. Yip, N. Shadbolt, and C. Webber, "Why forums? an empirical analysis into the facilitating factors of carding forums," in *ACM Web Science 2013*, 2013.

[17] S. Afroz, V. Garg, D. McCoy, and R. Greenstadt, "Honor among thieves: A commons analysis of underground forums," in *eCrime Researcher's Summit*, 2013.

[18] P. Chairunnanda, N. Pham, and U. Hengartner, "Privacy: Gone with the typing! identifying web users by their typing patterns," in *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*. IEEE, 2011, pp. 974–980.

[19] P. Eckersley, "How unique is your web browser?" in *Privacy Enhancing Technologies*. Springer, 2010, pp. 1–18.

[20] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *IEEE Symposium on Security and Privacy*. IEEE, 2008, pp. 111–125.

[21] ——, "De-anonymizing social networks," in *30th IEEE Symposium on Security and Privacy*. IEEE, 2009, pp. 173–187.

[22] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, T. Renata *et al.*, "Exploiting innocuous activity for correlating users across sites," in *WWW'13 Proceedings of the 22nd international conference on World Wide Web*, 2013.

[23] M. Almishari and G. Tsudik, "Exploring linkability of user reviews," in *Computer Security–ESORICS 2012*. Springer, 2012, pp. 307–324.

[24] J. A. Calandrino, W. Clarkson, and E. W. Felten, "Bubble trouble: Off-line de-anonymization of bubble forms." in *USENIX Security Symposium*, 2011.

[25] A. Caliskan and R. Greenstadt, "Translate once, translate twice, translate thrice and attribute: Identifying authors and machine translation tools in translated text," in *Semantic Computing (ICSC), 2012 IEEE Sixth International Conference*. IEEE, 2012, pp. 121–125.

[26] M. Koppel and Y. Winter, "Determining if two documents are by the same author," *Journal of the American Society for Information Science and Technology*, 2013.

[27] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils, "How unique and traceable are usernames?" in *Privacy Enhancing Technologies*. Springer, 2011, pp. 1–17.

[28] D. M. Freeman, "Using naive bayes to detect spammy names in social networks," in *Proceedings of the 2013 ACM workshop on Artificial intelligence and security*. ACM, 2013, pp. 3–12.

[29] G. Danezis and P. Mittal, "Sybilinfer: Detecting sybil nodes using social networks." in *NDSS*, 2009.

[30] L. Alvisi, A. Clement, A. Epasto, U. Sapienza, S. Lattanzi, and A. Panconesi, "Sok: The evolution of sybil defense via social networks," in *IEEE security and privacy*, 2013.

[31] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180.

[32] H. Schmid, "Improvements in part-of-speech tagging with an application to german," in *In Proceedings of the ACL SIGDAT-Workshop*. Citeseer, 1995.

[33] A. W. McDonald, S. Afroz, A. Caliskan, A. Stolerman, and R. Greenstadt, "Use fewer instances of the letter i: Toward writing style anonymization," in *Privacy Enhancing Technologies*, 2012, pp. 299–318.

[34] J. C. Platt, "Sequential minimal optimization: A fast algorithm for training support vector machines," *Advances in Kernel Methods Support Vector Learning*, vol. 208, no. MSR-TR-98-14, pp. 1–21, 1998.

[35] J. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[36] M. Brennan, S. Afroz, and R. Greenstadt, "Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity," *ACM Transactions on Information and System Security (TISSEC)*, vol. 15, no. 3, 2012.