SYSML: StYlometry with Structure and Multitask Learning: Implications for Darknet Forum Migrant Analysis

Pranav Maneriker Yuntian He Srinivasan Parthasarathy
The Ohio State University
{maneriker.1@,he.1773,parthasarathy.2}@osu.edu

Abstract

Darknet market forums are frequently used to exchange illegal goods and services between parties who use encryption to conceal their identities. The Tor network is used to host these markets, which guarantees additional anonymization from IP and location tracking, making it challenging to link across malicious users using multiple accounts (sybils). Additionally, users migrate to new forums when one is closed further increasing the difficulty of linking users across multiple forums. We develop a novel stylometry-based multitask learning approach for natural language and model interactions using graph embeddings to construct low-dimensional representations of short episodes of user activity for authorship attribution. We provide a comprehensive evaluation of our methods across four different darknet forums demonstrating its efficacy over the state-of-the-art, with a lift of up to 2.5X on Mean Retrieval Rank and 2X on Recall@10.

1 Introduction

Crypto markets are *"online forums where goods and services are exchanged between parties who use digital encryption to conceal their identities"* (Martin, 2014). They are typically hosted on the Tor network, which guarantees anonymization in terms of IP and location tracking. The identity of individuals on a crypto-market is associated only with a username; therefore, building trust on these networks does not follow conventional models prevalent in eCommerce. Interactions on these forums are facilitated by means of text posted by their users. This makes the analysis of textual style on these forums a compelling problem.

Stylometry is the branch of linguistics concerned with the analysis of authors' style. Text stylometry was initially popularized in the area of forensic linguistics, specifically to the problems of author profiling and author attribution (Juola, 2006; Rangel et al., 2013). Traditional techniques for authorship analysis on such data rely upon the existence of long text corpora from which features such as the frequency of words, capitalization, punctuation style, word and character n-grams, function word usage can be extracted and subsequently fed into any statistical or machine learning classification framework, acting as an author's 'signature'. However, such techniques find limited use in short text corpora in a heavily anonymized environment.

Advancements in using neural networks for character and word-level modeling for authorship attribution aim to deal with the scarcity of easily identifiable 'signature' features and have shown promising results on shorter text (Shrestha et al., 2017). Andrews and Witteveen (2019) drew upon these advances in stylometry to propose a model for building representations of social media users on Reddit and Twitter. Motivated by the success of such approaches, we develop a novel methodology for building authorship representations for posters on various darknet markets. Specifically, our key contributions include:

First, a *representation learning* approach that couples temporal content stylometry with access identity (by levering forum interactions via *meta-path graph context information*) to model and enhance user (author) representation;

Second, a novel framework for training the proposed models in a *multitask setting* across multiple darknet markets, using a small dataset of labeled migrations, to refine the representations of users within each individual market, while also providing a method to correlate users across markets;

Third, a detailed drill-down *ablation study* discussing the impact of various optimizations and highlighting the benefits of both graph context and multitask learning on forums associated with four darknet markets - *Black Market Reloaded*, *Agora Marketplace*, *Silk Road*, and *Silk Road 2.0* - when compared to the state-of-the-art alternatives.

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 6844–6857
November 7–11, 2021.

2 Related Work

Darknet Market Analysis: Content on the dark web includes resources devoted to illicit drug trade, adult content, counterfeit goods and information, leaked data, fraud, and other illicit services (Lapata et al., 2017; Biryukov et al., 2014) . Also included are forums discussing politics, anonymization, and cryptocurrency. Biryukov et al. (2014) found that while a vast majority of these services were in English (about 84%), a total of about 17 different languages were detected. Analysis of the volume of transactions and number of users on darknet markets indicates that they are resilient to closures; rapid migrations to newer markets occur when one market shuts down (ElBahrawy et al., 2019).

Recent work (Fan et al., 2018; Hou et al., 2017; Fu et al., 2017; Dong et al., 2017) has levered the notion of a heterogeneous information network (HIN) embedding to improve graph modeling, where different types of nodes, relationships (edges) and paths can be represented through typed entities. Zhang et al. (2019) used a HIN to model marketplace vendor sybil[1] accounts on the darknet, where each node representing an object is associated with various features (e.g. content, photography style, user profile and drug information). Similarly, Kumar et al. (2020) proposed a multi-view unsupervised approach which incorporated features of text content, drug substances, and locations to generate vendor embeddings. We note that while such efforts (Zhang et al., 2019; Kumar et al., 2020) are related to our work, there are key distinctions. First, such efforts focus only on vendor sybil accounts. Second, in both cases, they rely on a host of multi-modal information sources (photographs, substance descriptions, listings, and location information) that are not readily available in our setting - limited to forum posts. Third, neither effort exploits multitask learning.

Authorship Attribution of Short Text: Kim (2014) introduced convolutional neural networks (CNNs) for text classification. Follow-up work on authorship attribution (Ruder et al., 2016; Shrestha et al., 2017) leveraged these ideas to demonstrate that CNNs outperformed other models, particularly for shorter texts. The models proposed in these works aimed at balancing the trade-off between vocabulary size and sequence length budgets based on tokenization at either the character or word level. Further work on subword tokenization (Sennrich et al., 2016), especially byte-level tokenization, have made it feasible to share vocabularies across data in multiple languages. Models built using subword tokenizers have achieved good performance on authorship attribution tasks for specific languages (e.g., Polish (Grzybowski et al., 2019)) and also across multilingual social media data (Andrews and Bishop, 2019). Non-English as well as multilingual darknet markets have been increasing in number since 2013 (Ebrahimi et al., 2018). Our work builds upon all these ideas by using CNN models and experimenting with both character and subword level tokens.

Multitask learning (MTL): MTL (Caruana, 1997), aims to improve machine learning models' performance on the original task by jointly training related tasks. MTL enables deep neural network-based models to better generalize by sharing some of the hidden layers among the related tasks. Different approaches to MTL can be contrasted based on the sharing of parameters across tasks - strictly equal across tasks (hard sharing) or constrained to be close (soft-sharing) (Ruder, 2017). Such approaches have been applied to language modeling (Howard and Ruder, 2018), machine translation (Dong et al., 2015), and dialog understanding (Rastogi et al., 2018).

3 SYSML Framework

Motivated by the success of social media user modeling using combinations of multiple posts by each user (Andrews and Bishop, 2019; Noorshams et al., 2020), we model posts on darknet forums using *episodes*. Each *episode* consists of the textual content, time, and contextual information from multiple posts. A neural network architecture f_θ maps each episode to combined representation $e \in \mathbb{R}^E$. The model used to generate this representation is trained on various metric learning tasks characterized by a second set of parameters $g_\phi : \mathbb{R}^E \rightarrow \mathbb{R}$. We design the metric learning task to ensure that episodes having the same author have *similar* embeddings. Figure 1 describes the architecture of this workflow and the following sections describe the individual components and corresponding tasks. Note that our base modeling framework is inspired by the social media user representations built by Andrews and Bishop (2019) for a single task. We add meta-path embeddings and multitask objectives to

[1] a single author can have multiple users accounts which are considered as *sybils*

enhance the capabilities of SYSML. Our implementation is available at: https://github.com/pranavmaneriker/SYSML.

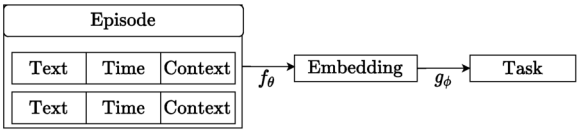

Figure 1: Overall SYSML Workflow.

3.1 Component Embeddings

Each episode e of length L consists of multiple tuples of texts, times, and contexts $e = \{(t_i, \tau_i, c_i) | 1 \leq i \leq L\}$. Component embeddings map individual components to vector spaces. All embeddings are generated from the forum data only; no pretrained embeddings are used.

Text Embedding First, we tokenize every input text post using either a character-level or byte-level tokenizer. A one-hot encoding layer followed by an embedding matrix E_t of dimensions $|V| \times d_t$ where V is the token vocabulary and d_t is the token embedding dimension embeds an input sequence of tokens $T_0, T_1, \ldots, T_{n-1}$. We get a sequence embedding of dimension $n \times d_t$. Following this, we use f sliding window filters, with filters sized $F = \{2, 3, 4, 5\}$ to generate feature-maps which are then fed to a max-over-time pooling layer, leading to a $|F| \times f$ dimensional output (one per filter). Finally, a fully connected layer generates the embedding for the text sequence, with output dimension d_t. A dropout layer prior to the final fully connected layer prevents overfitting, as shown in Figure 2.

Time Embedding The time information for each post corresponds to when the post was created and is available at different granularities across darknet market forums. To have a consistent time embedding across different granularities, we only consider the least granular available date information (date) available on all markets. We use the day of the week for each post to compute the time embedding by selecting the corresponding embedding vector of dimension d_τ from the matrix E_w.

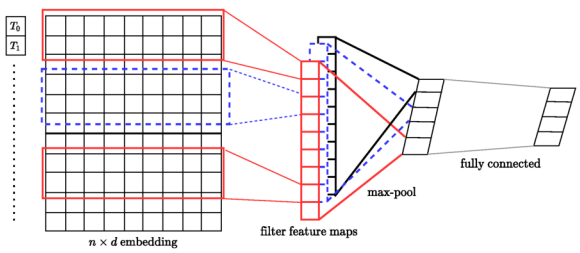

Figure 2: Text Embedding CNN (Kim, 2014).

Structural Context Embedding The context of a post refers to the threads that it may be associated with. Past work (Andrews and Bishop, 2019) used the subreddit as the context for a Reddit post. In a similar fashion, we encode the subforum of a post as a one-hot vector and use it to generate a d_c dimensional context embedding. In the previously mentioned work, this embedding is initialized randomly. We deviate from this setup and use an alternative approach based on a *heterogeneous graph* constructed from forum posts to initialize this embedding.

Definition 3.1 (Heterogeneous Graph)**.** A heterogeneous graph $G = (V, E, T)$ is one where each node v and edge e are associated with a 'type' $T_i \in T$, where the association is given by mapping functions $\phi(v) : V \to T_V$, $\psi(e) : E \to T_E$, where $|T_V| + |T_E| > 2$

The constraint on $T_{V,E}$ ensures that at least one of T_V and T_E have more than one element (making the graph heterogeneous). Specifically, we build a graph in which there are four types of nodes: user (U), subforum (S), thread (T), and post (P), and each edge indicates either a post of new thread (U-T), reply to existing post (U-P) or an inclusion (T-P, S-T) relationship. To learn the node embeddings in such heterogeneous graphs, we leverage the metapath2vec (Dong et al., 2017) framework with specific meta-path schemes designed for darknet forums. Each meta-path scheme can incorporate specific semantic relationships into node embeddings. For example, Figure 3 shows an instance of a meta-path 'UTSTU', which connects two users posting on threads in the same subforum and goes through the relevant threads and subforum. Our analysis is user focused; to capture user behavior, we consider *all* metapaths starting from and ending at a user node. Thus, to fully capture the semantic relationships in the heterogeneous graph, we use seven meta-path schemes: UPTSTPU, UTSTPU, UPTSTU, UTSTU, UPTPU, UPTU, and UTPU. As a result, the learned embeddings will preserve the semantic relationships between each subforum, included posts as well as relevant users (authors). Metapath2vec generates embeddings by maximizing the probability of heterogeneous neighbourhoods, normalizing it across typed contexts. The

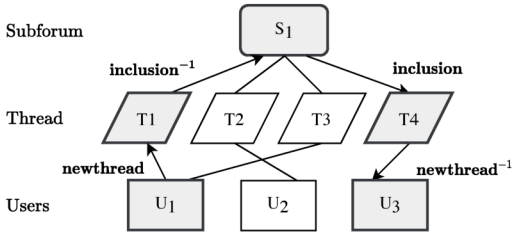

Figure 3: An instance of meta-path 'UTSTU' in a subgraph of the forum graph.

optimization objective is:

$$\arg\max_{\theta} \prod_{v \in V} \prod_{t \in T_v} \prod_{c_t \in N_t(v)} p(c_t|v;\theta)$$

Where θ is the learned embedding, $N_t(v)$ denotes v's neighborhood with the t^{th} type of node. In practice, this is equivalent to running a word2vec (Mikolov et al., 2013) style skip gram model over the random walks generated from the meta-path schemes when $p(c_t|v;\theta) =$ is defined as a softmax function. Further details of metapath2vec can be found in the paper by Dong et al. (2017).

3.2 Episode Embedding

The embeddings of each component of a post are concatenated into a $d_e = d_t + d_\tau + d_c$ dimensional embedding. An episode with L posts, therefore, has a $L \times d_e$ embeddings. We generate a final embedding for each episode, given the post embeddings using two different models. In **Mean Pooling**, the episode embedding is the mean of L post embeddings, resulting in a d_e dimensional episode embedding. For the **Transformer**, the episode embeddings are fed as the inputs to a transformer model (Devlin et al., 2019; Vaswani et al., 2017), with each post embedding acting as one element in a sequence for a total sequence length L. We follow the architecture proposed by Andrews and Bishop (2019) and omit a detailed description of the transformer architecture for brevity (Figure 4 shows an overview). Note that we do not use positional embeddings within this pooling architecture. The parameters of the component-wise models and episode embedding models comprise the episode embedding $f_\theta : \{(t, \tau, c)\}^L \rightarrow \mathbb{R}^E$.

3.3 Metric Learning

An important element of our methodology is the ability to learn a distance function over user representations. We use the username as a label for the episode e within the market M and denote each username as a unique label $u \in U_M$. Let $W = |U_M| \times d_E$ represent a matrix denoting the weights corresponding to a specific metric learning method and let $x^* = \frac{x}{||x||}$. An example of a metric learning loss would be Softmax Margin, i.e., cross-entropy based softmax loss.

$$P(u|e) = \frac{e^{W_u d_e}}{\sum_{j=1}^{|U_M|} e^{W_j d_e}}$$

We also explore alternative metric learning approaches such as Cosface (CF) (Wang et al., 2018), ArcFace (AF) (Deng et al., 2019), and MultiSimilarity (MS) (Wang et al., 2019).

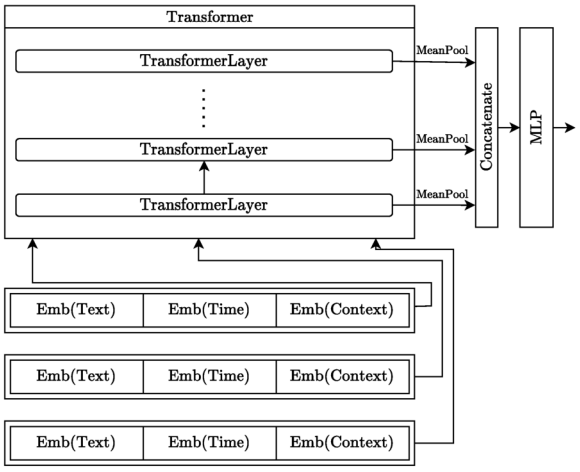

Figure 4: Architecture for Transformer Pooling.

3.4 Single-Task Learning

The components discussed in the previous sections are combined together to generate an embedding and the aforementioned tasks are used to train these models. Given an episode $e = \{(t_i, \tau_i, c_i)|1 \leq i \leq L\}$, the componentwise embedding modules generate embedding for the text, time, and context, respectively. The pooling module combines these embeddings into a single embedding $e \in \mathbb{R}^E$. We define f_θ as the combination of the transformations that generate an embedding from an *episode*. Using a final metric learning loss corresponding to the task-specific g_ϕ, we can train the parameters θ and ϕ. The framework, as defined in Figure 1, results in a model trainable for a single market M_i. Note that the first half of the framework (i.e., f_θ) is sufficient to generate embeddings for episodes, making the module invariant to the choice of g_ϕ. However, the embedding modules learned from

these embeddings may not be compatible for comparisons across different markets, which motivates our multi-task setup.

3.5 Multi-Task Learning

We use authorship attribution as the metric learning task for each market. Further, a majority of the embedding modules are shared across the different markets. Thus, in a multi-task setup, the model can share episode embedding weights (except context, which is market dependent) across markets. A shared BPE vocabulary allows weight sharing for text embedding on the different markets. However, the task-specific layers are not shared (different authors per dataset), and sharing f_θ does not guarantee alignment of embeddings across datasets (to reflect migrant authors). To remedy this, we construct a small, manually annotated set of labeled samples of authors known to have migrated from one market to another. Additionally, we add pairs of authors known to be distinct across datasets. The *cross-dataset* consists of all episodes of authors that were manually annotated in this fashion. The first step in the multi-task approach is to choose a market ($\mathcal{T}_M$) or cross-market ($\mathcal{T}_{cr}$) metric learning task $\mathcal{T}_i \sim \mathcal{T} = \{\mathcal{T}_M, \mathcal{T}_{cr}\}$. Following this, a batch of N episodes $\mathcal{E} \sim \mathcal{T}_i$ is sampled from the corresponding task. The embedding module generates the embedding for each episode $f_\theta^N : \mathcal{E} \rightarrow \mathbb{R}^{N \times E}$. Finally, the task-specific metric learning layer $g_\phi^{\mathcal{T}_i}$ is selected and a task-specific loss is backpropagated through the network. Note that in the *cross-dataset*, new labels are defined based on whether different usernames correspond to the same author and episodes are sampled from the corresponding markets. Figure 5 demonstrates the shared layers and the use of *cross-dataset* samples. The overall loss function is the sum of the losses across the markets: $\mathcal{L} = \mathop{\mathbb{E}}_{\mathcal{T}_i \sim \mathcal{T},\, \mathcal{E} \sim \mathcal{T}_i} [\mathcal{L}_i(\mathcal{E})]$.

4 Datasets

Munksgaard and Demant (2016) studied the politics of darknet markets using structured topic models on the forum posts across six large markets. We start with this dataset and perform basic preprocessing to clean up the text for our purposes. We focus on four of the six markets - *Silk Road* (**SR**), *Silk Road 2.0* (**SR2**), *Agora Marketplace* (**Agora**), and *Black Market Reloaded* (**BMR**). We exclude 'The Hub' as it is not a standard forum but an 'omni-forum' (Munksgaard and Demant, 2016) for discussion of other marketplaces and has a significantly different structure, which is beyond the scope of this work. We also exclude 'Evolution Marketplace' since none of the posts had PGP information present in them and thus were unsuitable for migration analysis.

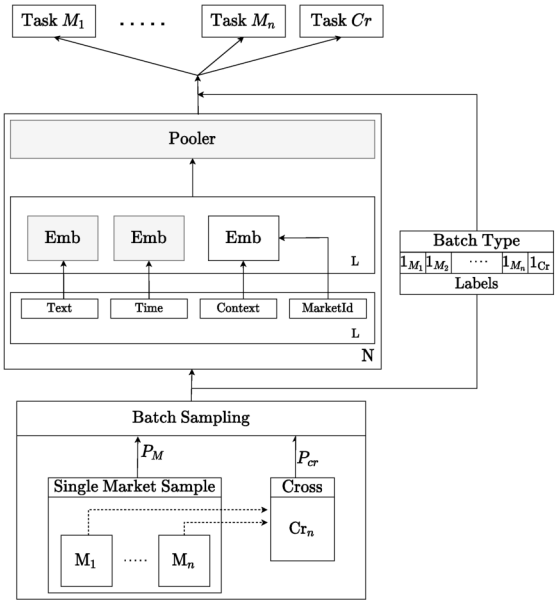

Figure 5: Multi-task setup. Shaded nodes are shared

Pre-processing We add simple regex and rule based filters to replace quoted posts (i.e., posts that are begin replied to), PGP keys, PGP signatures, hashed messages, links, and images each with different special tokens (`[QUOTE]`, `[PGP PUBKEY]`, `[PGP SIGNATURE]`, `[PGP ENCMSG]`, `[LINK]`, `[IMAGE]`). We retain the subset of users with sufficient posts to create at least two episodes worth of posts. In our analysis, we focus on episodes of up to 5 posts. To avoid leaking information across time, we split the dataset into approximately equal-sized train and test sets with a chronologically midway splitting point such that half the posts on the forum are before that time point. Statistics for data after pre-processing is provided in Table 1. Note that the test data can contain authors not seen during training.

Market	Train Posts	Test Posts	#Users train	#Users test
SR	379382	381959	6585	8865
SR2	373905	380779	5346	6580
BMR	30083	30474	855	931
Agora	175978	179482	3115	4209

Table 1: Dataset Statistics for Darkweb Markets.

Cross-dataset Samples Past work has established PGP keys as strong indicators of shared authorship

on darkweb markets (Tai et al., 2019). To identify different user accounts across markets that correspond to the same author, we follow a two-step process. First, we select the posts containing a PGP key, and then pair together users who have posts containing the same PGP key. Following this, we still have a large number of potentially incorrect matches (including scenarios such as information sharing posts by users sharing the PGP key of known vendors from a previous market). We manually check each pair to identify matches that clearly indicate whether the same author or different authors posted them, leading to approximately 100 reliable labels, with 33 pairs matched as migrants across markets.

5 Evaluation

While ground truth labels for a single author having multiple accounts are unavailable, individual models can still be compared by measuring their performance on authorship attribution as a proxy. We evaluated our method using retrieval-based metrics over the embeddings generated by each approach. Denote the set of all episode embeddings as $E = \{e_1, \dots e_n\}$ and let $Q = \{q_1, q_2, \dots q_\kappa\} \subset E$ be the sampled subset. We computed the cosine similarity of the query episode embeddings with all episodes. Let $R_i = \langle r_{i1}, r_{i2}, \dots r_{in} \rangle$ denote the list of episodes in E ordered by their cosine similarity with episode q_i (excluding itself) and let $A(.)$ map an episode to its author. The following measures are computed.

Mean Reciprocal Rank: (MRR) The RR for an episode is the reciprocal rank of the first element (by similarity) with the same author. MRR is the mean of reciprocal ranks for a sample of episodes.

$$MRR(Q) = \frac{1}{\kappa} \sum_{i=1}^{\kappa} \frac{1}{\min_j \left(A(r_{ij}) = A(e_i)\right)}$$

Recall@k: (R@k) Following Andrews and Bishop (2019), we define the R@k for an episode e_i to be an indicator denoting whether an episode by the same author occurs within the subset $\langle r_{i1}, \dots, r_{ik} \rangle$. R@k denotes the mean of these recall values over all the query samples.

Baselines We compare our best model against two baselines. First, we consider a popular short text authorship attribution model (Shrestha et al., 2017) based on embedding each post using character CNNs. While the method had no support for additional attributes (time, context) and only considers a single post at a time, we compare variants that incorporate these features as well. The second method for comparison is invariant representation of users (Andrews and Bishop, 2019). This method considers only one dataset at a time and does not account for graph-based context information. Results for episodes of length 5 are shown in Table 2

6 Analysis

6.1 Model and Task Variations

To compare the variants using statistical tests, we compute the MRR of the data grouped by market, episode length, tokenizer, and a graph embedding indicator. This leaves a small number of samples for paired comparison between groups, which precludes making normality assumptions for a t-test. Instead, we applied the paired two-samples Wilcoxon-Mann-Whitney (WMW) test (Mann and Whitney, 1947). The first key contribution of our model is the use of meta-graph embeddings for context. The WMW test demonstrates that using pre-trained graph embeddings was significantly better than using random embeddings ($p < 0.01$). Table 2 shows a summary of these results using ablations. For completeness of the analysis, we also compare the character and BPE tokenizers. WMW failed to find any significant differences between the BPE and character models for embedding (table omitted for brevity). Many darkweb markets tend to have more than one language (e.g., BMR had a large German community), and BPE allows a shared vocabulary to be used across multiple datasets with very few out-of-vocab tokens. Thus, we use BPE tokens for the forthcoming multitask models.

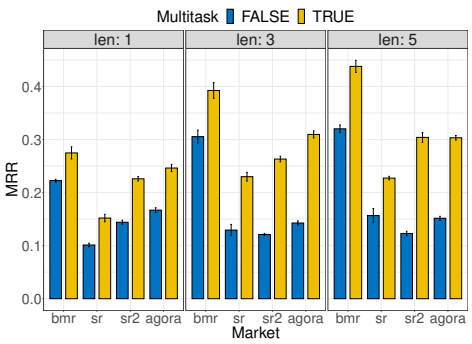

Figure 6: Drill-down: one-at-a-time vs. multitask.

Multitask Our second key contribution is the multitask setup. Table 2 demonstrates that SYSML (multitask) outperforms all baselines on episodes of length 5. We further compare runs of the best

Method	BMR		Agora		SR2		SR	
	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10
Shrestha et al. (2017) (CNN)	0.07	0.165	0.126	0.214	0.082	0.131	0.036	0.073
+ time + context	0.235	0.413	0.152	0.263	0.118	0.21	0.094	0.178
+ time + context + transformer pooling	0.219	0.409	0.146	0.266	0.117	0.207	0.113	0.205
Andrews and Bishop (2019) (IUR)								
mean pooling	0.223	0.408	0.114	0.218	0.126	0.223	0.109	0.19
transformer pooling	0.283	0.477	0.127	0.234	*0.13*	*0.229*	0.118	0.204
SYSML (single)	*0.32*	*0.533*	*0.152*	*0.279*	0.123	0.21	*0.157*	*0.266*
- graph context	0.265	0.454	0.144	0.251	0.089	0.15	0.049	0.094
-graph context - time	0.277	0.477	0.123	0.198	0.079	0.131	0.04	0.08
SYSML (multitask)	**0.438**	**0.642**	**0.303**	**0.466**	**0.304**	**0.464**	**0.227**	**0.363**
- graph context	0.396	0.602	**0.308**	**0.469**	0.293	0.442	0.214	0.347
- graph context - time	0.366	0.575	0.251	0.364	0.236	0.358	0.167	0.28

Table 2: Best performing results in **bold**. Best performing single-task results in *italics*. All $\sigma_{MRR} < 0.02$, $\sigma_{R@10} < 0.03$, For all metrics, higher is better. Results suggest single-task performance largely outperforms the state-of-the-art (Shrestha et al., 2017; Andrews and Bishop, 2019), while our novel multi-task cross-market setup offers a substantive lift (**up to 2.5X on MRR and 2X on R@10**) over single-task performance.

single task model for each market against a multitask model. Figure 6 demonstrates that multitask learning consistently and significantly (WMW: $p < 0.01$) improves performance across all markets and all episode lengths.

Metric Learning Recent benchmark evaluations have demonstrated that different metric learning methods provide only marginal improvements over classification (Musgrave et al., 2020; Zhai and Wu, 2019). We experimented with various state-of-the-art metric learning methods (§3.3) in the multi task setup and found that softmax-based classification (SM) was the best performing method in 3 of 4 cases for episodes of length 5 (Figure 7). Across all lengths, SM is significantly better (WMW: $p < 1e-8$) and therefore we use SM in SYSML.

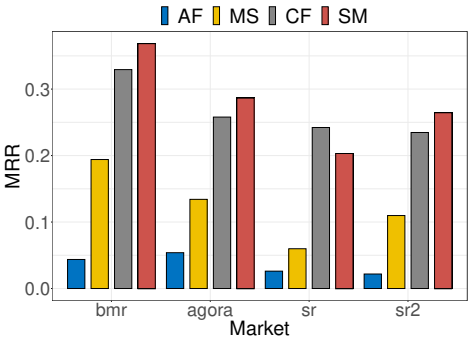

Figure 7: Task comparison: SM and CF are better performing two methods, with SM better in 3 of 4 cases.

6.2 Novel Users

The dataset statistics (Table 1) indicate that there are users in each dataset who have no posts in the time period corresponding to the training data. To understand the distribution of performance across these two configurations, we compute the test metrics over two samples. For one sample, we constrain the sampled episodes to those by users who have at least one episode in the training period (Seen Users). For the second sample, we sample episodes from the complement of the episodes that satisfy the previous constraint (Novel Users). Figure 8 shows the comparison of MRR on these two samples against the best single task model for episodes of length 5. Unsurprisingly, the first sample (Seen Users) have better query metrics than the second (Novel Users). However, importantly both of these groups outperformed the best single task model results on the first group (Seen Users), which demonstrates that the *lift offered by the multitask setup is spread across all users.*

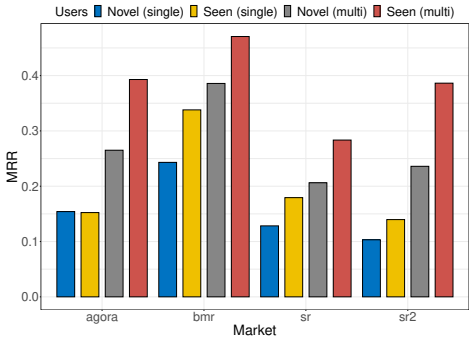

Figure 8: Lift on the multitask setup across users.

Episode Length Figure 9 shows a comparison of the mean performance of each model across various episode lengths. We see that compared to the baselines, SYSML can combine contextual and stylistic information across multiple posts more effectively.

Additional results (see appendix), indicate that this trend continues for larger episode sizes.

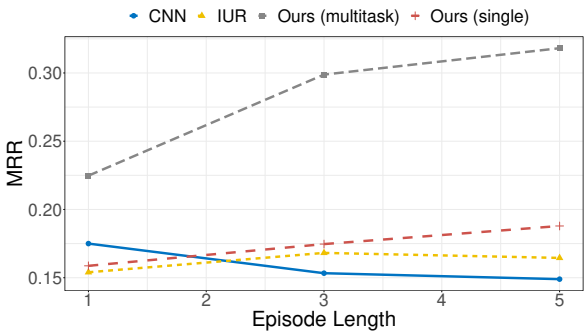

Figure 9: SYSML is more effective at utilizing multi post stylometric information

7 Case Study

7.1 Qualitative Analysis of Attribution:

In this section, we consider the average (euclidean) distance between each pair of episodes by the same author as a heuristic for stylometric identifiability (SI), where lower average distance corresponds to higher SI and vice versa. Somewhat surprisingly, authors with a small number of total episodes (< 10) were found at both extremes of identifiability, while the authors with the highest number of episodes were in the intermediate regions, suggesting that SI is not strongly correlated with episode length. Next, we further investigate these groups.

High SI authors: Among the 20 users with the lowest average distance between episodes, a single pattern is prominent. This first group of high SI users are "newbie" users. On a majority of analyzed forums, a minimum number of posts by a user is required before posting restrictions are removed from the user's account. Thus, users create threads on 'Newbie Discussion' subforums. Typical posts on these threads include repeated posting of the same message or numbered posts counting up to the minimum required. As users tend to make all these posts within a fixed time frame, the combination of repeated, similar stylistic text and time makes the posts easy to identify. Exemplar episodes from this "newbie" group are shown in Table 3.

After filtering these users out, we identified a few more notable high SI users. These include an author on BMR with frequent '£' symbol and ellipses ('...') and an author on Agora who only posted referral links (with an eponymous username 'ReferralLink'). Finally, restricting posts to those made by 200 most frequently posting users (henceforth, T200), we found a user (labeled HSI-Sec[2]) who frequently provided information on security, where character n-grams corresponding to 'PGP', 'Key', 'security' are frequent (Table 4). Thus, SYSML is able to leverage vocabulary and punctuation-based cues for SI.

Thread	Posts
Spam to 50 & Get out of Noobville Post 30 Times . . . To Post Anywhere Spam to 50 . . . . . . use my link . . .	26, 27, 28, 29, 30 7, 8, 9, . . . 46, 47, . . . , Yeah 50 Spam! [LINK], Here is my ref link [LINK], Try this link [LINK], . . .

Table 3: Examples of highly identifiable posts.

Low SI authors: Here, we attempt to characterize the post episode styles that are challenging for SYSML to attribute to the correct author. Seminal work by Brennan and Greenstadt (2009); Brennan et al. (2012) has demonstrated that obfuscation and imitation based strategies are effective against text stylometry. We analyze the T200 authors who had high inter-episode distances to ascertain whether this holds true for SYSML. For the least (and third least) identifiable author among T200, we find that frequent word n-grams are significantly less frequent than those for the most identifiable author from this subset (most frequent token occurs ~ 600 times vs. ~ 4800 times for identifiable) despite having more episodes overall. Further, one of the most frequent tokens is the `[QUOTE]` token, implying that this author frequently incorporates other authors' quotes into their posts. This strategy is analogous to the imitation based attack strategy proposed by Brennan et al. (2012). For the second least identifiable T200 author, we find that the frequent tokens have even fewer occurrences, and the special token `[IMAGE]` and its alternatives are among the frequent tokens - suggesting that an obfuscation strategy based on diversifying the vocabulary is effective. Some samples are presented in Table 4 under LSI-1 and LSI-2.

Gradient-based attribution: To cement our preceding hypotheses, we investigate whether the generated embedding can be attributed to phrases in the input which were mentioned in the previous section. We use Integrated Gradients (Sundararajan et al., 2017), an axiomatic approach to input attribution. Integrated Gradients assign an importance score to each feature which corresponds to an approximation of the integral of the gradient of a model's output with respect to the input features

[2]pseudonym

Author	Word Importance
HSI-Sec	. . . 2 cents, anyway . . . PGP Key Fingerprint = . . .
	. . . PGP Key Fingerprint . . . . security is NOT retroactive .
	. . . Is it possible for a gpg key to request that)
LSI-1	Check out the link in my sig . . . [IMAGE alt=8)]
	Hey dude, just run a search . . . I can not help much . . . Im sure if you ask . . . German , he may be willing to lend a hand. Good luck
	freind [IMAGE alt=8)]
LSI-2	[QUOTE] From: . . . Just my opinion, I 've done just about every-
	thing, . . . IMAGE alt=8)] couldnt agree more
	[QUOTE] From : . . . strangely enough, when im in . . . I too
	jabber meaningless jibberish . . .
	Negative Neutral Positive

Table 4: Integrated Gradient based attribution of posts

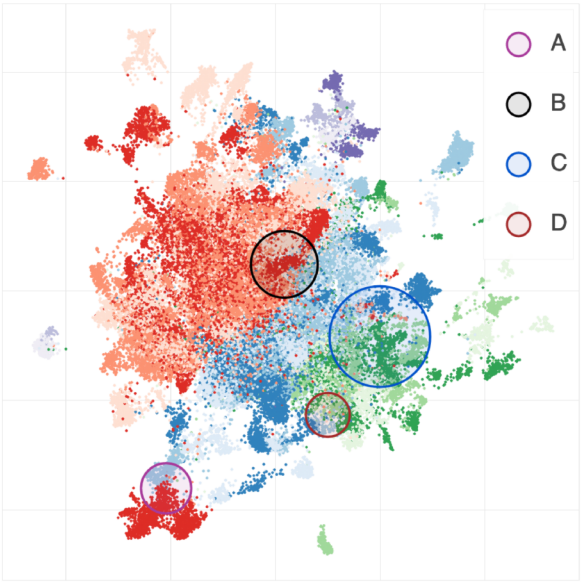

Figure 10: UMAP visualization of cross dataset embeddings for the top 200 authors, one hue per market. Circles denote the same user in two different markets.

along a path from some reference baseline value (in our case, all [PAD] tokens) to the input feature. In Table 4, the highlight color corresponds to the attribution importance score for the presented posts. We observed that the attribution scores correspond to our intuitions: HSI-Sec had high importance for security words, LSI-1 had obfuscated posts due to the presence of common image tokens, and LSI-2 had quotes mixed in, lead to misattribution (imitation-like strategy).

7.2 Migrant Analysis

To understand the quality of alignment from the episode embeddings generated by our method, we use a simple top-k heuristic: for each episode of a user, find the top-k nearest neighboring episodes from other markets, and count the most frequently occurring user among these (candidate sybil account). Figure 10 shows a UMAP projection for T200. Users of each market are colored by sequential values of a single hue (i.e., reds - SR2, blues - SR, etc.). The circles in the figure highlight the top four pairs of users (top candidate sybils) with a frequent near neighbor from a different market. We find that each of these pairs can be verified as sybil accounts, either by a shared username (A, C, D) or by manual inspection of posted information (B). Note that none of these pairs were pre-matched using PGP - none were present in the high-precision matches. Thus, SYSML is able to identify high ranking sybil matches reflecting users that migrate from one market to another.

8 Conclusion

We develop a novel stylometry-based multitask learning approach that leverages graph context to construct low-dimensional representations of short episodes of user activity for authorship and identity attribution. Our results on four different darknet forums suggest that both graph context and multitask learning provides a significant lift over the state-of-the-art. In the future, we hope to evaluate how such methods can be levered to analyze how users maintain trust while retaining anonymous identities as they migrate across markets. Further, we hope to quantitatively evaluate the migration detection to assess the evolution of textual style and language use on darknet markets. Characterizing users from a small number of episodes has important applications in limiting the propagation of bot and troll accounts, which will be another direction of future work.

Acknowledgements

This work has been supported by NSF grants SES-1949037, CCF-2028944, and OAC-2018627. All content represents the opinion of the authors, and is not necessarily endorsed by their sponsors. The authors thank the OSU CLIPPERS group, Micha Elsner, Sean Current, Saket Gurukar, and anonymous reviewers for helpful discussions and feedback on this work. Additionally, the authors thank the Ohio Supercomputing Center (Center, 1987) and the OSU RI2 cluster for providing the computational resources used for conducting the experiments in this paper.

References

Martin Andrews and Sam Witteveen. 2019. Unsupervised natural question answering with a small model. In *Proceedings of the Second Workshop on Fact Extraction and VERification (FEVER)*, pages 34–38, Hong Kong, China. Association for Computational Linguistics.

Nicholas Andrews and Marcus Bishop. 2019. Learning invariant representations of social media users. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1684–1695, Hong Kong, China. Association for Computational Linguistics.

Alex Biryukov, Ivan Pustogarov, Fabrice Thill, and Ralf-Philipp Weinmann. 2014. Content and popularity analysis of tor hidden services. In *2014 IEEE 34th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, pages 188–193. IEEE.

Gwern Branwen, Nicolas Christin, David Décary-Hétu, Rasmus Munksgaard Andersen, StExo, El Presidente, Anonymous, Daryl Lau, Delyan Kratunov Sohhlz, Vince Cakic, Van Buskirk, Whom, Michael McKenna, and Sigi Goode. 2015. Dark net market archives, 2011-2015. `https://www.gwern.net/DNM-archives`.

Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Transactions on Information and System Security (TISSEC)*, 15(3):1–22.

Michael Robert Brennan and Rachel Greenstadt. 2009. Practical attacks against authorship recognition techniques. In *IAAI*.

Roderic Broadhurst, Matthew Ball, Chuxian Jiang, Joy Wang, and Harshit Trivedi. 2021. Impact of darknet market seizures on opioid availability. *Broadhurst R, Ball, M, Jiang, CX, et al*.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

Ohio Supercomputer Center. 1987. Ohio supercomputer center.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 4690–4699. Computer Vision Foundation / IEEE.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China. Association for Computational Linguistics.

Yuxiao Dong, Nitesh V. Chawla, and Ananthram Swami. 2017. metapath2vec: Scalable representation learning for heterogeneous networks. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*, pages 135–144. ACM.

Mohammadreza Ebrahimi, Mihai Surdeanu, Sagar Samtani, and Hsinchun Chen. 2018. Detecting cyber threats in non-english dark net markets: A cross-lingual transfer learning approach. In *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 85–90. IEEE.

Abeer ElBahrawy, Laura Alessandretti, Leonid Rusnac, Daniel Goldsmith, Alexander Teytelboym, and Andrea Baronchelli. 2019. Collective dynamics of dark web marketplaces. *ArXiv preprint*, abs/1911.09536.

Yujie Fan, Yiming Zhang, Yanfang Ye, and Xin Li. 2018. Automatic opioid user detection from twitter: Transductive ensemble built on different meta-graph based similarities over heterogeneous information network. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 3357–3363. ijcai.org.

Tao-Yang Fu, Wang-Chien Lee, and Zhen Lei. 2017. Hin2vec: Explore meta-paths in heterogeneous information networks for representation learning. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 1797–1806. ACM.

Piotr Grzybowski, Ewa Juralewicz, and Maciej Piasecki. 2019. Sparse coding in authorship attribution for Polish tweets. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 409–417, Varna, Bulgaria. INCOMA Ltd.

Shifu Hou, Yanfang Ye, Yangqiu Song, and Melih Abdulhayoglu. 2017. Hindroid: An intelligent android malware detection system based on structured heterogeneous information network. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax,*

NS, Canada, August 13 - 17, 2017, pages 1507–1515. ACM.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Patrick Juola. 2006. Authorship attribution. *Found. Trends Inf. Retr.*, 1(3):233–334.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, and Orion Reblitz-Richardson. 2020. Captum: A unified and generic model interpretability library for pytorch.

Ramnath Kumar, Shweta Yadav, Raminta Daniulaityte, Francois R. Lamy, Krishnaprasad Thirunarayan, Usha Lokala, and Amit P. Sheth. 2020. edarkfind: Unsupervised multi-view learning for sybil account detection. In *WWW '20: The Web Conference 2020, Taipei, Taiwan, April 20-24, 2020*, pages 1955–1965. ACM / IW3C2.

Mirella Lapata, Phil Blunsom, and Alexander Koller, editors. 2017. *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. Association for Computational Linguistics, Valencia, Spain.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

James Martin. 2014. *Drugs on the dark net: How cryptomarkets are transforming the global trade in illicit drugs*. Springer.

Tomas Mikolov, Kai Chen, G. Corrado, and J. Dean. 2013. Efficient estimation of word representations in vector space. In *ICLR*.

Rasmus Munksgaard and Jakob Demant. 2016. Mixing politics and crime–the prevalence and decline of political discourse on the cryptomarket. *International Journal of Drug Policy*, 35:77–83.

Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. 2020. A metric learning reality check.

Nima Noorshams, Saurabh Verma, and Aude Hofleitner. 2020. TIES: temporal interaction embeddings for enhancing social media integrity at facebook. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, CA, USA, August 23-27, 2020*, pages 3128–3135. ACM.

Francisco Rangel, Paolo Rosso, Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at pan 2013. In *CLEF Conference on Multilingual and Multimodal Information Access Evaluation*, pages 352–365. CELCT.

Abhinav Rastogi, Raghav Gupta, and Dilek Hakkani-Tur. 2018. Multi-task learning for joint language understanding and dialogue state tracking. In *Proceedings of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, pages 376–384, Melbourne, Australia. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *ArXiv preprint*, abs/1706.05098.

Sebastian Ruder, Parsa Ghaffari, and John G Breslin. 2016. Character-level and multi-channel convolutional neural networks for large-scale authorship attribution. *ArXiv preprint*, abs/1609.06686.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Prasha Shrestha, Sebastian Sierra, Fabio González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 2017. Convolutional neural networks for authorship attribution of short texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 669–674, Valencia, Spain. Association for Computational Linguistics.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Xiao Hui Tai, Kyle Soska, and Nicolas Christin. 2019. Adversarial matching of dark net market vendor accounts. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019*, pages 1871–1880. ACM.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural*

Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA, pages 5998–6008.

Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. 2018. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 5265–5274. IEEE Computer Society.

Xun Wang, Xintong Han, Weilin Huang, Dengke Dong, and Matthew R. Scott. 2019. Multi-similarity loss with general pair weighting for deep metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 5022–5030. Computer Vision Foundation / IEEE.

Andrew Zhai and Hao-Yu Wu. 2019. Classification is a strong baseline for deep metric learning. In *30th British Machine Vision Conference 2019, BMVC 2019, Cardiff, UK, September 9-12, 2019*, page 91. BMVA Press.

Yiming Zhang, Yujie Fan, Wei Song, Shifu Hou, Yanfang Ye, Xin Li, Liang Zhao, Chuan Shi, Jiabin Wang, and Qi Xiong. 2019. 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, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, pages 3448–3454. ACM.

A Ethics Statement

The research conducted in this study was deemed to be *exempt research* by the Ohio State University's Office of Responsible Research Practices, since the forum data is classified as 'publicly available'. Darknet forum data is readily available publicly across multiple markets (Branwen et al., 2015; Munksgaard and Demant, 2016) and we follow standard practices for the darkweb (Kumar et al., 2020) limiting our analysis to publicly available information only. The data was originally collected to study the prevalence of illicit drug trade and the politics surrounding such trades.

Limiting Harm To the best of our knowledge, the collected data does not contain leaked private information (Munksgaard and Demant, 2016). Beyond relying on the exempt nature of the study, we also strive to take further steps for minimizing harms from our research. In accordance with the ACM Code of Ethics and to limit potential harm, we carry out substantial pre-processing (§4) to remove links, images, and keys that may contain sensitive information. Towards respecting the privacy of subjects, we do not connect the identity of users to any private information; our method serves only to link users across markets. Further, in this study, we restrict our analysis to darknet markets that have been inactive for several years. The darknet market community has itself taken steps over the past few years to link identities of trustworthy members across market closure via development of information hubs such as Grams, Kilos, and Recon (Broadhurst et al., 2021). Our efforts aim to understand the formative years that lead towards this centralization.

Inclusiveness Our methods do not attempt to characterize any traits of the users making the posts. Based on our analysis, the datasets contain posts in English, German, and Italian. Thus, our methods may be limited in applicability and biased in performance for languages belonging to these and related Indo-European languages.

Potential for Dual Use Our goal is to understand how textual style evolves on darknet markets and how users on such markets may misuse them for scams and illicit activities. This digital forensic analysis can be put to good use for understanding trust signalling on these markets. We understand the potential harm from dual use; stylometric methods could be used for the identification of users who may not want their identity to be made public, especially when they are subject of hostile governments. We believe that making the information about the existence of such stylometric advances public and providing prescriptions for avoidance techniques (§7.1) would aid users who may not know of strategies that they can use to preserve their anonymity. Existing work (Noorshams et al., 2020; Andrews and Bishop, 2019) has already expanded the use of stylometry to the open web. Thus, we have made the analysis of patterns that lower stylometric identifiability one focus of our case study.

B Reproducibility

We describe the various hyperparameter settings used for the models trained by us. All deep learning models are implemented in python using Pytorch[3], and the original C++ implementation of metapath2vec is used for generating metapath embeddings[4]. We used an implementation from the Captum python library (Kokhlikyan et al., 2020) that uses the Gauss-Legendre quadrature rule for

[3]https://pytorch.org/

[4]https://ericdongyx.github.io/metapath2vec/m2v.html

approximating the gradient.

C Training Hyperparameters

We use batches of size 256. The Adam optimizer is used for training each network. The initial learning rate is set to $1e-3$, with a multiplicative decay factor of 0.5 if the validation metrics do not improve after 5 epochs. Each model is trained for 30 epochs, and each configuration is run 5 times. We used a V100 GPU to train each run, with the average running time of 27:17 per run (mm:ss). For each run, 10% of the dataset is used for validation. The best model is selected on the basis of minimum validation loss.

C.1 Model Hyperparameters

C.1.1 Text Embedding Model

Character vocabularies of size 1k and BPE vocabularies of size 30k are trained using only training portion of the datasets. The HuggingFace Tokenizers[5] library is used to build the byte-level BPE vocab. We use a Text CNN for embedding text across all settings. Each token has 32 dimensional embeddings, and the final embedding dimension for a text sequence is set to 128. Filters of sizes $\{2, 3, 4, 5\}$ are used, and the dropout probability is set to 0.1 for the final layer.

C.1.2 Time Embedding

The time embedding dimension is set to 64.

C.1.3 Context Embedding

The context embedding dimension is set to 128. For metapath2vec, we generate 1000 walks for each user (author) node, the number of negative samples for each user is 5, and the window size in the skip-gram model is set to 7. These hyperparameters are also used in the metapath2vec work. For the length of each sampled walk, we set it to 80, which is widely used in many representative skip-gram based embedding methods such as node2vec.

C.1.4 Pooling Transformer

The pooling transformer model has a feed forward layer dimension and final dimension of 128. There are 4 layers, each with 4 heads. The droupout probability for the final feed forward leyer layer is set to 0.1, and the output dimension is set to 32.

[5]https://github.com/huggingface/tokenizers

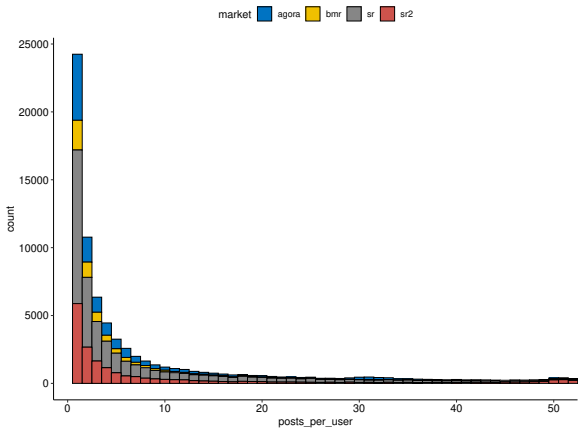

Figure 11: Frequency of number of posts per user

C.1.5 Metric Learning Techniques

We use the pytorch metric learning[6] package for implementing the different metric learning approaches, with the default parameters for each approach from their corresponding papers. i.e.,

- **CosFace**: $m = 0.35$, $s = 64$
- **ArcFace**: $m = 28.6°$, $s = 64$
- **MultiSimilarity**: $\alpha = 2$, $\beta = 50$, $\lambda = 0.5$

D Parameter Search

Most hyperparameter comparisons are reported in the paper. For the multitask dataset sampling, we ran the multitask model with $P_{cr} \in \{0.01, 0.02, 0.04, 0.1\}$, with $P_M = 1 - P_{cr}$, $P_{M_I} \propto |M_i|$ with similar performances up to $P_{cr} = 0.04$ and a drop at $P_{cr} = 0.1$. All results reported in the paper have $P_{cr} = 0.01$

E Metrics

All metrics are computed using a sample of the episode embeddings. The sample size used for computing the metrics is $\kappa = 1000$

F Additional Results

From Figure 11, we see that the number of users reduces rapidly as the posts per user decrease. Thus, we limited our analysis to up to 5 posts per episode. For completeness, we also provide additional results for 7 and 9 posts per episode in Table 5 and 6 respectively. Note that the histogram has some non-smooth bumps at around 10, 50, 100 posts as they act as the minimum number of posts for

[6]https://ericdongyx.github.io/metapath2vec/m2v.html

Method	BMR		Agora		SR2		SR	
	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10
SYSML (singletask)	0.305	0.508	0.186	0.32	0.159	0.273	0.14	0.246
SYSML (multitask)	0.484	0.689	0.349	0.519	0.401	0.556	0.292	0.429

Table 5: Additional results for 7 posts per episode

Method	BMR		Agora		SR2		SR	
	MRR	R@10	MRR	R@10	MRR	R@10	MRR	R@10
SYSML (singletask)	0.264	0.48	0.146	0.249	0.165	0.272	0.194	0.319
SYSML (multitask)	0.4667	0.648	0.357	0.498	0.377	0.522	0.299	0.449

Table 6: Additional results for 9 posts per episode

different levels of forum users. As explained in a previous section, users post on 'newbie' forums until they reach a specific number of posts, leading to these unusual bumps in the histogram. We note that the performance of our methods continues to improve as the posts per episode are increased (at a cost to coverage - number of users studied), though the improvement is higher in the bigger markets as these tend to have a sufficiently large number of individuals with a higher number of total posts.