DFRWS
DIGITAL FORENSIC RESEARCH CONFERENCE

Analyzing the DarkNetMarkets Subreddit for Evolutions of Tools and Trends Using LDA Topic Modeling

By

Kyle Porter

From the proceedings of

The Digital Forensic Research Conference

DFRWS 2018 USA

Providence, RI (July 15th - 18th)

Digital Investigation 26 (2018) S87–S97

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

DFRWS 2018 USA — Proceedings of the Eighteenth Annual DFRWS USA

Analyzing the DarkNetMarkets subreddit for evolutions of tools and trends using LDA topic modeling

Kyle Porter

Department of Information Security and Communication Technology, NTNU, Gjøvik, Norway

Keywords:
Topic modeling
Latent dirichlet allocation
Web crawling
Datamining
Semantic analysis
Digital forensics
Surface-web monitoring

A B S T R A C T

Darknet markets, which can be considered as online black markets, in general sell illegal items such as drugs, firearms, and malware. In July 2017, significant law enforcement operations compromised or completely took down multiple international darknet markets. To quickly understand how this affected the markets and the choice of tools utilized by users of darknet markets, we use unsupervised topic modeling techniques on the DarkNetMarkets subreddit in order to determine prominent topics and terms, and how they have changed over a year's time. After extracting, filtering out irrelevant posts, and preprocessing the text crawled from the subreddit, we perform Latent Dirichlet Allocation (LDA) topic modeling on a corpus of posts for each month from November 5th, 2016 to November 5th, 2017. Our results indicate that even analyzing public forums such as the DarkNetMarkets subreddit can reveal trends and keywords related to criminal activity and their methods, and that users of the dark web appear to be becoming increasingly more security-minded due to the recent law enforcement events.

1. Introduction

The *dark web*, websites which can only be reached through anonymity networks such as Tor (Dingledine et al., 2004) and are not indexed by any search engine, is well known for hosting criminal marketplaces. These *darknet markets* sell illicit items such as drugs, weapons, and hacking tools. Recent research shows that the markets and forums on the dark web have been valuable sources of cyber threat intelligence (Deliu, 2017; Nunes et al., 2016; Samtani et al., 2015), as well as general sources of intelligence for law enforcement agencies (Van Buskirk et al., 2016) to monitor the state of darknet markets to identify emerging trends.

In July 2017, two of the most popular darknet markets, AlphaBay and Hansa, were shut down by various law enforcement agencies (Gibbs and Beckett, 2017). From this, one may conjecture that the state of darknet markets and their customers may be going through a more tumultuous period than usual. To analyze the effect of these real world events, and to identify changes in behavior by darknet customers, we gather data from public sources on Reddit.[1]

Reddit is a social media platform with specific interest oriented forums called subreddits, and in this work we extract intelligence from the subreddit called DarkNetMarkets. We crawl this subreddit for a year's worth of posts and data to obtain a corpus for each month between November 5th, 2016 to November 5th, 2017. Ultimately, 15,400 posts were gathered from the subreddit, and since manually analyzing the corpora is exceedingly time consuming, we utilize Latent Dirichlet Allocation (LDA) (Blei et al., 2003)) unsupervised topic modeling to extract month to month information pertaining to the state of the darknet markets, the security and anonymity tools used by visitors to the darknet markets, and the cryptocurrency and related services used when purchasing items over the darknet.

The primary contribution of this work are the topics and terms produced by performing LDA topic modeling on the data from the DarkNetMarkets subreddit, wherein we can relatively quickly observe how the tools and the trends in markets, security, and cryptocurrency have changed from November 5th, 2016 to November 5th, 2017. From analyzing this data, following the July 2017 busts we can see an increase of uncertainty on the part of the users of the darknet markets, as well as an increase in security-mindedness. We note that law enforcement agencies are already monitoring this subreddit, so this information may already be known, but we none-the-less empirically show the effectiveness of LDA on criminally associated subreddits to quickly come to understand important content of a year's worth of subreddit posts.

The following is an outline of the paper. First we describe background information regarding details of Reddit, how darknet markets generally operate, and a brief overview of Latent Dirichlet

E-mail address: kyle.porter@ntnu.no.
[1] https://about.reddit.com/.

https://doi.org/10.1016/j.diin.2018.04.023

Allocation. The next section describes our experimental methodology, including our methods for extracting, filtering, and cleaning our dataset before applying topic modeling. Afterwards, we describe our topic modeling results for each month of posts on the DarkNetMarkets subreddit and analysis. Finally, we describe the related work, and conclude with a summary and discussion.

2. Background

In this section, we discuss attributes of Reddit, the darknet market community, and aspects of Latent Dirichlet Allocation so the topics produced by our experiments are more understandable.

2.1. The corpus: Reddit and subreddits

Reddit is a news aggregation and discussion website, where posts are organized into subreddits of specific interests. Subreddits are much like a standard "board" on a forum, and for our purposes the posts on these subreddits serve as the typical "threads". Each post has a title, potentially self-posted information by the author, and comments in response to the title or what was said by the author. Oftentimes, posts have a "flair", which is put in place by the author of the post or moderator of the subreddit that classifies the type of post being made. Furthermore, every post has a Unix timestamp associated with it, and therefore the corpus can be analyzed with respect to any given timeframe. The subreddit used for our experiments is "www.reddit.com/r/DarkNetMarkets".

2.2. DarkNetMarkets subreddit

The subreddit "DarkNetMarkets" is a public subreddit, where users discuss the goods and services of the black markets that can only be reached via an anonymity network such as Tor. Topics of conversation appear to mostly revolve around drugs, and the purchasing of drugs with cryptocurrency. More interesting topics of conversation include advice on increasing anonymity, operational security, tools used to improve stealthy financial transactions, and the state of markets and vendors. Vendors are essentially drug dealers who use the darknet markets as their platform to do business. Users purchase from vendors who they believe they can trust over darknet markets they believe are uncompromised using cryptocurrency.

The DarkNetMarkets subreddit has been a source of controversy in the past. In 2015, the FBI requested Reddit to reveal personal information regarding some of the DarkNetMarkets contributors (Knibbs, 2015). Surprisingly, the subreddit was banned on March 21, 2018 due to a new rule from Reddit administrators that forbids communities to use Reddit as a medium to exchange or perform transactions of prohibited goods or services (Franceschi-Bicchierai, 2018).

2.3. Latent Dirichlet Allocation

To perform unsupervised topic modeling on the data extracted from the Darknet Markets subreddit we use Latent Dirichlet allocation (LDA) (Blei et al., 2003)). This methodology was chosen as it is simple and is often used in a variety of sciences for topic modeling text corpora, which can be used as a type of text summarization of a large set of documents. LDA produces a model of a corpus of documents, where the model assumes that each of the documents in the corpus are derived from a generative process where each document consists as a distribution of a finite set of topics, each topic is a multinomial distribution of the vocabulary of words in the corpus, and each word of the document is drawn from one topic in the generative process.

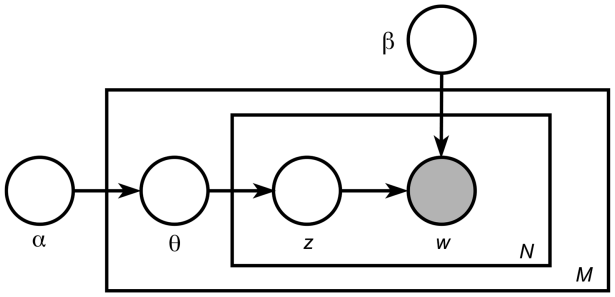

Fig. 1. Latent dirichlet allocation graphical model (Blei et al., 2003)).

Fig. 1 shows a graphical model of LDA, where the value w represents a vector of N words in document i of a total of M documents. A topic z is assigned to each word w_j of a document i, and therefore makes each document a composition of topics represented by some topic distribution θ over the document i. High α values represent that each document has a relatively even distribution of the topics, whereas low values of α indicate that the documents have a sparse distribution of all topics. Similarly, a high β value represents if topics are a relatively even distribution of the vocabulary of words, versus a low value of β which represents a sparse distribution of words per topic. Both α and β are set to default values of 1 divided by the number of topics for our experimentation (0.1). The latent elements we learn from running the LDA algorithm are the distributions of topics per document, and the distribution of words per topic.

A common issue regarding topic modeling via LDA is that the topics generated are not always interpretable or coherent by humans (Chang et al., 2009). To increase the certainty of being capable of classifying our generated topics, we use a relevancy metric introduced by Sievert and Shirley (2014). Typically, topics output a ranked list of the most probable terms in a topic, but this is often problematic as common terms in the corpus generally rank highly in the lists of multiple topics. This can make distinguishing topics difficult. The equation for relevance is given below.

$$\text{rel}(\text{term w} \mid \text{topic t}) = \lambda * p(w|t) + (1-\lambda) * p(w|t)/p(w) \quad (1)$$

After generating the topic model of a corpus, we can adjust the weight λ to influence the word ranking per topic according to relevance. When $\lambda = 1$, the standard ranking is returned as it is simply the conditional probability of the word w given the topic t. As λ approaches 0, the weight of the ratio of the word-topic probability $p(w|t)$ to overall word probability $p(w)$ increases. In this fashion, words with high probability $p(w)$ are ranked lower as λ approaches 0.

3. Experimental methodology

In our experiment we wish to extract tools and trends as well as changes in tools and trends in the DarkNetMarkets subreddit from the topic models generated by the LDA algorithm with the subreddit posts as input. Of specific interest is to observe how these items have changed after the July 2017 busts. To accomplish this, we create a corpus of subreddit posts for each 12 months of the year, where we began to extract subreddit posts from 00:00 November 5th, 2016, and limit our data extraction to 00:00 November 5th, 2017 (UTC). From this corpus, we compose smaller corpora consisting of posts from the 5th of each month to the 5th of the next month, where then we preprocess each monthly corpus and prepare it as input into

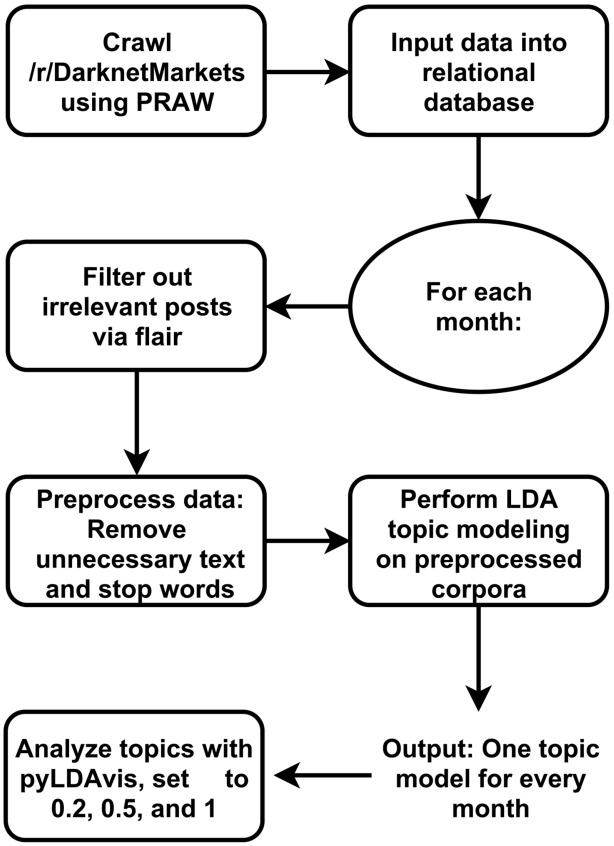

Fig. 2. Basic methodological process of topic modeling experiments.

the LDA algorithm from the gensim Python library.[2] The experimental methodology is summarized in Fig. 2.

3.1. Data extraction

We web crawled for text data from every post on the DarkNetMarkets subreddit from November 5th, 2016 to November 5th 2017 using the Python Reddit API Wrapper (PRAW)[3] to directly extract the post title, author text, comment text, post flair, and the time of creation for the post in the Unix timestamp format. This data is input into a relational database, where each row of the database contains the previously listed information per post. We consider each individual post to be a document, where the content of the document consists of the title of the post as well as the text written by the author and the replying comments. 15,400 different posts and their corresponding information were added to our database.

Whenever we wish to perform topic modeling on some set of documents over a specified period, we pull the raw text data from the set of documents corresponding to the UTC dates of the period from the database, where certain documents are filtered out due to their post flair signaling an irrelevant post. While the majority of posts share a small set of flairs, there were over 1000 different flairs we could potentially filter out for our experiments. Posts that had flairs related to humor or regarding delivery logistics were not included in our corpora. A brief summary of the corpora for months between November 2016 and November 2017, where posts have been filtered by flair, is described in Table 1 below. For each monthly corpus, we give the number of documents that the corpus consists of and the average number of words per document. The smallest document per month had 12 words or less, the largest documents per month typically 10,000 words or more, and the standard deviation of the words per document per month was between 652.93 and 1239.52.

Table 1
Corpora of posts filtered by flair.

Month	# Docs	Average # Words Per Doc
Nov 2016	1167	619
Dec 2016	1156	612
Jan 2017	1120	716
Feb 2017	1175	654
Mar 2017	1219	638
Apr 2017	1011	613
May 2017	1047	577
Jun 2017	919	695
Jul 2017	1795	689
Aug 2017	1328	693
Sep 2017	940	667
Oct 2017	1148	727

3.2. Preprocessing the data

Each document was preprocessed separately, as we would often adjust the preprocessing methodology to identify changes in the results from running the LDA algorithm on the data. The following was our standard method of preprocessing.

- Replace characters not in {A-Za-z0-9} with a single space
- Lowercase text
- Remove popular English contractions
- Replace trailing whitespaces with a single space
- Replace line breaks in the document for a single space
- Tokenize the data by spaces, and remove common *stop-words* (using the NLTK Python library[4])

As the text that we are analyzing is rather informal language, the standard set of stop-words, words which occur frequently but add little semantic value such as “a” or “the”, needed to be extended. Since each document is essentially a conversation amongst multiple people using colloquial language, our initial results from inputting the preprocessed data without filtering for the colloquial terms led to topics which focused on these terms. The topics generated revolved around swearing or common words such as “know”, “like”, or “even”. The process of identifying such useless terms was simply performed by generating the corpora for each month, and observing which of the twenty most highly relevant words per topic ($\lambda = 1$) added no insight to the potential topic. Such words were removed and topics were again generated to see if the topics now included other useless terms. This process was repeated until the twenty most relevant words per topic did not contain any obviously useless words, and as a result the coherence of topics became much clearer. The additional list of stop-words can be found in Appendix A.

We note that we did not perform word stemming, as we feared this could obscure the names of vendors, markets, or other services that use atypical language and could ultimately reduce the ability to understand topics (de Waal et al., 2008).

[2] https://radimrehurek.com/gensim/.
[3] https://praw.readthedocs.io/en/latest/.
[4] http://www.nltk.org/.

3.3. LDA parameters and methods for analysis

Other than the corpus and its associated vocabulary of words, the only parameters for the LDA algorithm that are required to be set by the user are the values for α, β (called ``eta'' by gensim), and the number of topics. As mentioned in the section discussing Latent Dirichlet Allocation, we set the values of α and β to their default values of 1 divided by the number of topics (0.1). While we did not attempt to optimize these values, our initial readings of the DarkNetMarkets posts led to assumptions about the topic-document distributions and word-topic distributions. Posts are typically focused on the title or description of the post, with some expected deviation, therefore indicating low values of α. The topics being discussed themselves are generally distinct, indicating a low value of β, but the word distribution is not extremely sparse. For instance, the mention of drugs will no doubt appear in many different topics. There is potential to improve our results with more careful choices of α and β.

The number of topics were decided on through repeated experimentation (as was done by Fang et al., (2016) and Deliu (2017)). Using too few topics creates vague topics, but choosing too many topics causes overfitting. After some time we found that generating models using 10 topics oftentimes provided sufficiently coherent topics to analyze our data effectively.

It is worth noting that we chose to perform LDA on a month by month basis as opposed to Dynamic Topic modeling (DTM) (Blei and Lafferty, 2006) because we were uncertain that topics would consistently be generated from month to month. DTM is valuable for observing the change in the word distribution per topic, however in our scenario new topics appear over time (such as the July 2017 busts). Furthermore, LDA algorithms and analytical tools are well documented.

Analyzing the topics generated by each corpus was assisted by using the Python library for interactive topic model visualization (pyLDAvis)[5] which is based on the work by Sievert and Shirley (2014). Using this tool, we can modify the rankings of the word distributions per topic by relevance (see Section 2.3). To obtain a clearer understanding of the topics generated, we examine the word distribution results when setting λ to 1, 0.5, and 0.2. We recorded the different rankings if the new terms and rankings introduced valuable information or helped clarify the topic (as seen in Appendix B).

4. Results and analysis

In this section we give a summary of our full results which may be found in Appendix B. The full results show the topics generated for each month from November 5, 2016 to November 5, 2017 from the documents in the DarkNetMarkets subreddit. For each month, we list the twenty most relevant terms for each topic under different settings for λ. We provide labels for each topic generated by the LDA algorithm, where the topics per month are listed from most *prevalent* (where prevalence defines the percentage of the corpus that the topics are comprised of) to the least prevalent. The bolded words represent terms of interest, and we omit some of the terms in repetitive topics such as “bot posts”, “moderator posts”, or “vendor reviews” to show more valuable terms in other topics.

Table 2 shows the topics relevant for intelligence gathering for each month, where each topic is paired with its prevalence (a ranking from 1 to 10). Question marks next to topic names indicate we are uncertain that this is truly the topic, and question marks by themselves indicate we are uncertain of what the topic is in general. All the topics included in Table 2 contain valuable or insightful terms. Before analyzing the relevant terms, we first examine how the topics have changed over these twelve months.

For the most part, the topics are consistent throughout the year with the exception of the topics that emerge due to ongoing real world events. For instance, stories such as the market busts in July 2017 emerge as a topic, and in October 2017 a topic emerged regarding an undercover IRS agent obtaining identifiable information about a hacker on the DarkNetMarkets subreddit (Franceschi-Bicchierai, 2017). Furthermore, even after July 2017, topics regarding cryptocurrency or security did not tend to increase in prevalence rank. What did change however, are the most relevant words per topic.

4.1. The state of the darknet markets

We can observe how the general state of the darknet markets has changed by examining the topics discussing different marketplaces and law enforcement. Prior to July 2017, Alphabay and Hansa (two darknet markets) are often listed as very relevant terms in a variety of these topics, and law enforcement terms (such as “le”, “law”, “police’, “europol”, etc.) are relatively scarce. For instance, in the months prior to July we did not find any law enforcement terms in the top twenty most relevant terms for any topic where $\lambda = 1$. Allowing for the other values of λ, law enforcement terms appeared in the top twenty most relevant terms in nine topics over these months. From July 2017 onwards, law enforcement related terms became far more relevant for topics regarding markets, vendors, cryptocurrency, and security. When $\lambda = 1$, law enforcement related terms appeared in the top twenty most relevant words in 17 topics. After July, the only market that is consistently listed as a relevant term is Dream. Even without reading the original data from the DarkNetMarkets subreddit, one could guess that a law enforcement event happened that affected the dominant market Alphabay, and that Dream seems to have replaced it. Other markets such as Aero, Agora, Traderoute, Sourcery, and Trishula appear have come into relevancy by October 2017, as well as concepts such as decentralized darknet markets like OpenBazaar. It seems that these markets may be filling the void left by Alphabay and Hansa, though some of their relevancy may be due to denial of service attacks on various markets in October.

The current state of darknet markets and their users that we may gather from these topics is one of concern and uncertainty. This can be seen as by October the number of markets relevant to the topics listed increased greatly, as well as the general change of terms in some topics. The topic which consistently appeared to be the most popular was general discussion about markets and vendors, and before July 2017 the most relevant terms within the word distribution per topic often contained words such as “address”, “btc”, “drugs”, “days”, “pack”, and “time”. This seems to indicate that these discussions often focused on the purchase and delivery of drugs from certain markets. However, from July 2017 onward the tone of these conversations appears to change as words such as “le”, “pgp” (Pretty Good Privacy), “multisig”, “2fa” (Two-Factor Authentication)’, “gg”, “phishing”, and “escrow” begin to appear as more relevant terms for the vendor and market topics. Security terms seem to be becoming more relevant in the largest topic of conversation. Additionally, the inclusion of the word “gg” (which stands for good game, indicating that a game is over) in the list of these relevant terms humorously seems to indicate that some users of darknet markets may believe that these markets are no longer viable.

Interestingly, it is not only law enforcement that users of darknet markets are worrying about, but they have become targets for hackers and untrustworthy markets as well. In addition to being

[5] https://github.com/bmabey/pyLDAvis.

Table 2
Topics with their prevalence ranking for each month from November 5th, 2016 to November 5th, 2017

Month	Topics
Nov 2016	General conversations about markets (1), general conversations about drugs (4), buying drugs (5), security (7), ??? (9), news articles ??? (10)
Dec 2016	Markets and vendors (3), cryptocurrency (4)
Jan 2017	Discussions about vendors and markets (1), cryptocurrency (2), operating systems and hacking (9)
Feb 2017	General conversation or bust??? (2), cryptocurrency and security (4), vendors and markets (7)
Mar 2017	General conversations about vendors and orders (1), anonymity, drugs, and general conversation (5), disputes??? (9), crytpocurrency (10)
Apr 2017	General market conversation (1), security or anonymity (4), vendors??? (8)
May 2017	Cryptocurrency and security (3), markets (7), transactions (8)
Jun 2017	Markets and vendors (2), ??? (6), ??? (8)
Jul 2017	Markets and vendors (1), Hansa and Alphabay busts (4), cryptocurrency (7), ??? (8)
Aug 2017	Markets and vendors (1), drugs and law enforcement (3), cryptocurrency (5), discussion about Silk Road (7), anonymity (8), DHL market??? (9)
Sep 2017	Vendor discussion or review??? (1), cryptocurrency and anonymity (2), drugs (9), DDoS attack article??? (10)
Oct 2017	Markets and vendors (1), anonymity and story about undercover agent (5), security and drugs (6), secure transactions (7), Silk Road??? (9)

afflicted by DDoS attacks, markets have been hacked and credentials, passwords, and private conversations have been leaked. For this reason, security features such as two-factor authentication may be expected as standard in the future of these markets. Terms such as “escrow” and “multisig” appear to be relevant to prevent markets which may scam users into purchasing items that they will never receive (“exit scamming”), which appears to have regained significant interest after the July 2017 busts. This may be currently relevant as users may be more willing to use new markets since many of the largest and most trusted markets have been shut down or under attack.

4.2. Tools used by darknet market users

Topics related to cryptocurrency, security, and anonymity appear to be consistently relevant, and the word distributions for these topics often contain the different tools darknet market users utilize.

The topics regarding cryptocurrency do not always appeared to be related specifically to security, but we go over the tools and services that they use as it may be useful for intelligence gathering for law enforcement. Bitcoin and Monero appear to be the cryptocurrencies most commonly used, but more interestingly are some of the services that have been used in conjunction with these currencies. Mixing or tumbling is a service which attempts to increase the anonymity of cryptocurrency transactions, where a group of users exchange cryptocoins with each other to increase the difficulty in tracing transactions (Ruffing et al., 2014). Mixing services such as Dash, Helix, and Bitmixer (which has since July 2017 shut down) appear in some of the relevant terms in the cryptocurrency topics. Other services that appear are cryptocurrency exchanges or brokers such as Coinbase, Seraphim, Localbitcoins, Bitbay, Shapeshift, and Viabtc.

Tools used by the darknet market to increase anonymity go beyond that of simply using Tor. The most common operating system suggested by the topics is Tails,[6] where all software is configured to connect to the internet through Tor. Less prevalent operating systems found in these topics are Whonix and Qubes. The use of VPNs is a commonly found in the security related topics for the purpose of increasing anonymity, where the only VPN listed in the highly relevant terms is PureVPN. Lastly, for the purpose of confidential communication, PGP appears to be the most commonly used tool.

Besides seeing that PGP was a more relevant term in the final seven months, we could not distinguish how the use of these tools changed over time. The only obvious obversation we make is that occurrence of real world events would spark discussions about specific tools. For instance, compromises of anonymity of markets or users (such as in July and August), or a spike in the price of bitcoin (as in August) led to topics where the previously listed tools were highly relevant.

4.3. Limitations of topic modeling results for forensic purposes

Topic modeling and the word distributions per topic are only made practical when paired with the original data source, as it is easy to misinterpret the results. From the time spent on this work, we find that these models are useful for developing hypotheses of the content within the subreddit, that we then later would confirm by searching the original data source for some of the terms listed in the topics. Perhaps the most useful aspect of topic modeling results are the word distributions per topic, as the topic puts the terms into a context. For instance, there are many words for markets, tools, and services we would not have recognized if they were not contextualized by the topics in which they were discovered in. These terms can then be used as keywords for further investigation.

A relevant complaint about LDA algorithms is the time required to build the models, where building topic models for extremely large corpora can take hours (Noel and Peterson, 2014) (where for us, building a model from a monthly corpus took a manner of minutes due to the smaller corpus sizes). Additionally, it is worth noting while analyzing topic models is a relatively fast method for understanding some content of a large corpus, the analysis still takes a generous amount of time.

5. Related work

As far as we know, our work is the first which applies LDA topic modeling for exploring subreddits for extracting forensic intelligence such as the identification of tools and services used in criminal behavior, as well as assessing the state of a criminal community. However, these topics have all been researched separately and in different contexts.

Topic modeling has been used extensively to gather intelligence directly from darknet markets and forums. Grisham et al. used LDA topic modeling to determine the types of items being sold on Alphabay, and the top vendors selling them (Grisham et al., 2016). Semani et al. gathered data from darknet hacker forums and performed topic modeling on source code, attachments, and hacking tutorials to better understand hacker assets (Samtani et al., 2015). To organize the source code posts, they used a support vector machine (SVM) to classify the code by programming language. Similarly, Deliu analyzed hacker forums, and performed binary

[6] https://tails.boum.org/about/index.en.html.

classification on the comments of the forums using an SVM to automate the process of filtering security irrelevant comments (as well as mutli-class classification with respect to specific security concepts), and then used LDA topic modeling to identify malware related topics discussed on the forum (Deliu, 2017). Unlike Deliu's work, our posts were slightly more organized due to the fact that the Reddit posts were often paired with flairs to indicate the overall relevancy of the post. However, we could likely improve the accuracy of our results if we performed binary classification on a comment by comment basis. Lastly, Fang et al. used variations of LDA (including Dynamic Topic Modeling) to explore popular topics on Chinese hacker forums (Fang et al., 2016). Their use of Dynamic Topic Modeling (over a twelve-year period) showed the emergence of new communication methods, specific security mechanisms, and even inferences that hackers on these forums became more cautious of faulty transactions.

Topic modeling has also been used for semantic analysis of Reddit posts as well. Shen and Rudzics used several semantic analysis techniques, including LDA, to detect anxiety related posts on Reddit, where the Reddit API was used to extract the text from multiple subreddits associated with anxiety (Shen and Rudzicz, 2017).

Several researchers have applied the use of topic modeling to specifically digital forensic contexts. Okolica et al. used an extension of LDA called Author-Topic modeling (Rosen-Zvi et al., 2004) where each author is associated with a multinomial distribution over topics, and this was used to create social network graphs to identify insider threats (Okolica et al., 2007). Since Reddit posts include usernames alongside each comment or post, it might be useful to apply Author-Topic modeling on Reddit data in which there is in interest to identify users with behavior of interest. Beebe and Liu compared different clustering algorithms (including LDA) for clustering search hit results, where LDA was applied prior to k-means or Kohonen Self-Organizing Maps so that documents were clustered by topic as opposed to terms (Beebe and Liu, 2014). They found that clustering by topics performed better than clustering by terms for keyword searching. Noel and Peterson used LDA topic modeling as a method of relaxing keyword selection for search (Noel and Peterson, 2014). While they suggest that LDA search should not replace regular expression search, they note benefits of LDA search such as the ablity to find documents which do not contain the searched term and an improvement to data browsing through topics.

6. Conclusion

In this work we performed Latent Dirichlet Allocation topic modeling on all posts from the DarkNetMarkets subreddit for each month between November 5th, 2016 and November 5th, 2017 to obtain forensic intelligence regarding the darknet market community and determine the effects of the July 2017 darknet market busts. Items of interest included the state of the darknet market and the tools utilized by visitors to the darknet markets. To increase the coherence of our topic modeling results and to find additional keywords, we used the relevancy metric by Sievert and Shirley (2014). Adjusting this metric allows words that have a relatively high conditional probability of appearing in a topic but a low overall probability of appearing in the corpus to be ranked even higher in a topic.

Our results indicate that the general state of the darknet markets (with respect to Reddit users) has gone from casual and relaxed to its current state of concern, uncertainty, and security-mindedness. After the July 2017 busts, words associated to law enforcement (such as “le” and “police”) are often highly relevant in many topics, and the void left by the previously most popular markets seems to have been filled by a multitude of newer or smaller markets. However, these markets do not appear to be inherently trustworthy, as users in recent months often discuss methods of secure transaction between potentially untrustworthy markets (such as two-factor authentication, “escrow”, and “multisig”). Additionally, the relevance of the term “pgp” is more apparent in the last seven months, indicating an increased use or desire of authenticated and confidential communication.

Cryptocurrency and security tools appear to have been consistent topics of conversation throughout the year. Popular cryptocurrencies are Monero and Bitcoin, where darknet market users often use additional services such as “mixing” or “tumbling” to enhance their anonymity. We also extracted a number of popular bitcoin exchanges or brokers used by darknet market users. To enhance anonymity and operational security, the preferred choice of operating system appears to be Tails, as it automatically configures software to connect to the internet via Tor, and there additionally is interest in VPN services. For confidential communication, PGP appears to be the most commonly used method. We did not observe much of an evolution of tools, only that they become more relevant when real world events involving the price of cryptocurrency or law enforcement occur.

This is the first work that we are aware of that uses LDA topic modeling on subreddits to extract tools and trends related to criminal behavior on the darknet market. However, since our data source is public, much of the information provided here can be found in various locations across the web. The advantage to our approach is that it was useful for relatively quickly understanding this online community and inferring possible trends within this community. Such information may be useful to law enforcement to understand how real world events have effects on criminal online communities.

Future work may improve on our experiments by including an SVM classifier to filter out irrelevant comments (Deliu, 2017), and word stemming should be applied to the preprocessing procedure to determine if it improves the topic results. Furthermore, it would worthwhile to see how effective these topic modeling approaches are on other criminal related subreddit communities, especially since the DarkNetMarkets subreddit was banned in March, 2018. Perhaps the most interesting future work would be utilizing methods to automate or hasten the analysis of these topic models.

Acknowledgments

We would like to thank the reviewers for their comments and help for preparing this paper for publication.

This work was supported by the Research Council of Norway program IKTPLUSS, under the R&D project “Ars Forensica - Computational Forensics for Large-scale Fraud Detection, Crime Investigation & Prevention”, grant agreement 248094/O70.

Appendix A. List of additional stop-words

The following words often appeared as the most probable words per topic, and therefore were removed during preprocessing:

know, knowing, like, liking, would, one, get, people, think, thinking, even, could, go, going, fucking, fuck, shit, also, use, using, much, got, good, make, making, well, way, feel, really, see, want, need, sure, say, right, still, take, taking.

Appendix B. Full topics for every month from November 5th, 2016 to November 5, 2017

Table B.3
Topics for November 2016.

#	Topic	Relevancy λ and terms
1	General conversation about markets	$\lambda = 1$: vendor, com, https, deleted, reddit, r, order, time, post, vendors, www, account, day, never, drugs, something, address, back, **ab**, market $\lambda = 0.5$: vendor, com, dispute, reddit, deleted, https, www, address, r, account, post, vendors, order, day, comments, drugs, **le**, bond, money, buyer $\lambda = 0.2$: dispute, bond, chatsecure, confirmations, dna, trappy, klonopins, taper, turkey, fee, fubinaca, fibf, **botnet**, disputes, uncomfirmed, montfort, router, **seized**, buyer, remindmebot
2	Bot posts	$\lambda = 1$: please, onion, http, vendor, r, message, darknetmarkets, contact, product, mdma, always, questions, subreddit, moderators, automatically, compose, concerns, bot, vendors, action
3	Formated vendor review	$\lambda = 1$: 10, vendor, time, product, shipping, stealth, review, order, price, quality, days, 5, shipped, ordered, 3, great, first, coke, 4, experience
4	General conversations about drugs	$\lambda = 1$: time, reddit, deleted, u, meth, post, https, us, never, bbmc, com, drugs, lol, r, back **tor**, something, always, around, said
5	Buying drugs	$\lambda = 1$: account, deleted, drugs, never, time, someone, fent, drug, pills, anyone, money, real, getting, **bitcoin**, coins, probably, anything, buy, look, something
6	Vendor review	$\lambda = 1$: 1, x, 5, value, 2, cost, vendor, 0, 4, 3, package, shipped, price, shipping, marked, cut, processed, low, resolution, product
7	Security	$\lambda = 1$: https, **tails**, **bitcoin**, **monero**, com, 2, wallet, **btc**, r, reddit, **blockchain**, time, new 0, usb, deleted, www, 1, post, **multisig**
8	Drugs	$\lambda = 1$: acetone, cocaine, com, blast, new, wash, thanks, r, drug, time, www, drugs, u, mdma, try, always, https, key, put, deleted
9	???	$\lambda = 1$: time, wallet, **btc**, address, package, vendor, 2, **hansa**, money, first, tracking, friends, something, **multisig**, 2, back, buy, order, never, send
10	News articles???	$\lambda = 1$: trump, service, mail, **tor**, time, attacks, marijuana, drugs, **ddos**, onion, said better, day, https, 1, com, please, first, president, hours

Table B.4
Topics for December 2016.

#	Topic	Relevancy λ and terms
1	General conversation about drugs and life	$\lambda = 1$: deleted, time, r, reddit, com, https, lol, life, drugs, please, vendor, yeah, never, lsd, sub, u, www, someone, something, man $\lambda = 0.5$: life, reddit, sub, deleted, drugs, r, com, lol, man, lot, yeah, time, someone, meth, lsd, https, u, hard, thing, ca $\lambda = 0.2$: pasitheas, drama, life, duterte, trump, sober, peitho, president, hbb, ak, documentary, bailey, bbmv, vote, ukb, sub, jay, narcan, alexa, philippines
2	Formated vendor reviews	$\lambda = 1$: 10, vendor, product, shipping, review, price, order, stealth, 5, quality, shipped, time, ordered, days, great, mdma, 1, tabs, 2, us
3	Markets and vendors	$\lambda = 1$: vendor, market, order, time, **ab**, vendors, back, **alphabay**, account, **scam**, deleted, **exit**, never, **dream**, money, markets, **bitcoin**, **btc**, days, new
4	Cryptocurr.	$\lambda = 1$: **btc**, buy, cash, money, 5, **bitcoin**, circle, deleted, time, **lbc**, bank, **coinbase**, buying, pay, wallet, back, atm, card, find, though
5	Adverts for .onion sites	$\lambda = 1$: onion, http, please, php, listing, r, grams, contact, message, questions, darknetmarkets, id, subreddit, automatically, moderators, concerns, bot, action, compose, performed
6	Mailing discussions	$\lambda = 1$: vendor, time, https, post, **tails**, com, www, vendors, order, deleted, drugs, used, years, package, law, packages, drug, something, pills, 2
7	Formated vendor reviews	$\lambda = 1$: 1, x, cost, value, 3, 4, 5, 2, 0, package, vendor, low, shipped, marked, cut, processed, resolution, product, price, bad
8	Review about nborne?	$\lambda = 1$: vendor, time, reddit, lsd, never, 2, message, please, r, u, post, lol, days, com, https, nborne, www, deleted, order, drug
9	Moderator posts	$\lambda = 1$: r, darknetmarkets, please, reddit, rules, www, com, post, https, subreddit, message, new, automatically, account, hours, submission, modpost, wiki, 6, vendor
10	???	$\lambda = 1$: new, year, press, anything, captcha, buds, pretty, deleted, around, ago, used, **tails**, time, blue, nw, buy, someone, try, find, half

Table B.5
Topics for January 2017.

#	Topic	Relevancy λ and terms
1	Discussion about vendors and markets	$\lambda = 1$: market, time, post, reddit, never, **ab**, vendor, **alphabay**, vendors, com, back, r, deleted, https, please, address, www, something, used, money $\lambda = 0.5$: market, post, address, **ab**, **alphabay**, users, **le**, markets, never, reddit, time, mods, superlist, **tumbling**, used, vendors, coins, sub, something $\lambda = 0.2$: superlist, address, **le**, **tumbling**, bug, mods, **tails**, market, jerome, users, private, staff, helix, coins, address, wallet, xavier, proof, **ab**
2	Cryptocurr.	$\lambda = 1$: **btc**, **bitcoin**, time, **monero**, buy, post, com, cash, r, u, lol, https, someone, deleted, never, drugs, price, anything, back $\lambda = 0.5$: **btc**, **monero**, **bitcoin**, cash, fee, bank, **coinbase**, money, buy, card, **seraphim**, coin, gold, **xmr**, currency, **lbc**, mod, hair, fees, pay
3	Vendor review	$\lambda = 1$: vendor, order, time, days, ordered, feedback, product, reddit, mdma, https, r, never, shippping, review, com, anyone, price, **ab**, thanks, 2
4	Drugs/review	$\lambda = 1$: meth, time, high, 5, 3, 2, day, trip, 1, review, though, test, coke, little, first, lsd, 4, product, dose, mdma
5	Adverts for .onion sites	$\lambda = 1$: onion, http, please, r, darknetmarkets, contact, message, subreddit, questions, automatically, compose, moderators, action, bot, performed, concerns, https, grams, vendor, reddit
6	Formated vendor review	$\lambda = 1$: 10, vendor, stealth, shipping, product, shipped, time, price, review, order, days, 2, 5, fe, quality, market, 3, communication, yes, ordered
7	jesusofrave lsd bust story	$\lambda = 1$: john, jor, lsd, tabs, **tmg**, time, tt, deleted, new, post, https, account, back, never, reddit, com, lol, made, anything, yeah
8	Vendor review???	$\lambda = 1$: 1, x, cost, value, 5, package, 2, 4, 3, vendor, 0, package, https, shipped, cut, marked, shipping, low, processed, resolution, market
9	Operating systems and hacking	$\lambda = 1$: **windows**, shirt, **linux**, com, drugs, gloves, drug, **tails**, account, https, wear, someone, **whonix**, dna, lawyer, www, deleted, first, legal, buy
10	Bot comments	$\lambda = 1$: com, http, remindmebot, message, reddit, imgur, np, subject, https, comments, grams, compose, 0a, vendor, time, r, delete, 2017, www, 1

Table B.6
Topics for February 2017.

#	Topic	Relevancy λ and terms
1	Formated Vendor Review	See November 2016 for similar terms.
2	General conversation or bust???	$\lambda = 0.5$: **btc**, **monero**, **bitcoin**, cash, fee, bank, **coinbase**, money, buy, card, **seraphim**, coin, gold, **xmr**, currency, lbc, mod, hair, fees, pay
3	General conversation about drugs	$\lambda = 1$: time, life, never, back, mdma, heroin, deleted, fent, something, high, drugs, meth, day, drug, yeah, though, lol, dose, anything, probably
4	Cryptocurr. and security	$\lambda = 0.5$: **tails**, wallet, fee, **tor**, transaction, **electrum**, **blockchain**, **usb**, **bitcoin**, **vpn**, **btc**, r, transactions, coins, wiki, https, backup, www, **monero**, reddit $\lambda = 0.2$: **tails**, wallet, **electrum**, fee, **usb**, **blockchain**, **vpn**, transaction, backup, **tor**, confirmations, unconfirmed, **monero**, modpost, darknetmarketsnoobs, **password**, transactions, submission, persistence, **viabtc**
5	Adverts for.onion sites	$\lambda = 1$: onion, http, please, r, vendor, darknetmarkets, market, message, contact, **ab**, automatically, moderators, **alphabay**, subreddit, questions, reddit, action, php, compose, performed, concerns
6	Bot/Mod Post???	See November 2016 for similar terms.
7	Vendors and markets	$\lambda = 1$: vendor, order, **exit**, money, **scam**, fe, **btc**, **ab**, time, cash, market, growmore, **hansa**, back, vendors, **bitcoin**, deleted, anyone, **dream**, buy $\lambda = 0.5$: **exit**, growmore, cash, **scam**, fe, **btc**, **lbc**, money, **hansa**, **coinbase**, **scamming**, **bitcoin**, order, coins, deposit, **scammed**, **ab**, **escrow**, refund, employee
8	Drugs	$\lambda = 1$: fentanyl, drug, fent, test, results, drugs, testing, vendors, heroin, us, work, china, used, many, getting, service, https, said, deleted, cut
9	???	$\lambda = 1$: https, jpg, never, com, mind, meth, put, many, trying, actually, remember, something, name, reviews, believe, together, ok, vendor, deleted, time
10	???	$\lambda = 1$: damn, banned, fa, bro, p, strain, u, grande, ariana, taste, high, fappy, dob, page, 10, com, sweet, im, perfect, 4fa

Table B.7
Topics for March 2017.

#	Topic	Relevancy λ and terms
1	General conversations about vendors and orders	$\lambda = 1$: time, order, vendor, deleted, back, money, vendors, never, lol, drugs, post, first, something, **btc**, someone, days, buy, anything, pack, getting $\lambda = 0.5$: money, order, time, back, deleted, **le**, vendors, btc, mail, drugs, never, lol, trust, pack, house, someone, post, scam, guy, orders
2	Drugs	$\lambda = 1$: lsd, deleted, vendor, time, try, tabs, lol, something, though, never, anyone, meth, maybe, 2, back, said, mdma, day, acid, pretty
3	Moderator posts	$\lambda = 1$: r, reddit, please, darknetmarkets, com, message, https, rules, www, subreddit, automatically, compose, contact, questions, bot, moderators, action, concerns, performed, post
4	Formated vendor review	$\lambda = 1$: 10, vendor, time, review, product, order, shipping, price, stealth, quality, https, reviews, ordered, days, 2, 3, vendors, great, cocaine, never
5	Anonymity, drugs, and general conversation	$\lambda = 1$: **tor**, **tails**, time, deleted, 2, cocaine, **vpn**, first, 1, new, used, every, 3, 5, 0, around, water, cia, work, man $\lambda = 0.2$: **tails**, **vpn**, **isp**, **usb**, **browser**, **miners**, **tor**, **wpa**, **windows**, **fingerprinting**, **qubes**, **vpns**, cia, **reaver**, **xiopan**, **linux**, persistence, **veracrypt**, filter, democrats
6	Adverts for .onion sites	$\lambda = 1$: onion, http, php, please, vendor, listing, r, id, message, darknetmarkets, questions, contact, subreddit, moderators, automatically, shipping, bot, compose, concerns
7	Vendor review	$\lambda = 1$: 1, vendor, 5, x, value, 2, shipped, 3, product, cost, shipping, 4, review, price, 0, package, marked, stealth, communication, market
8	Drugs	$\lambda = 1$: fentanyl, fent, drugs, heroin, drug, hcl, markets, buy, ban, water, someone, crack, something, base, glass, market, harm, cocaine, vendors, point
9	Disputes???	$\lambda = 1$: vendor, dispute, 2, order, used, time, product, days, label, took, long, back, last, anyone, 1, **ab**, never, market, box, shipping
10	Cryptocurr.	$\lambda = 1$: **monero**, **btc**, **bitcoin**, market, **eth**, price, time, **ethereum**, buy, **dash**, value, crack, **alphabay**, com, high, back, looking, crypto, meth, 5

Table B.8
Topics for April 2017.

#	Topic	Relevancy λ and terms
1	General market conversation	$\lambda = 1$: vendor, order, vendors, time, deleted, days, fent, ordered, pack, anyone, never, feedback, **ab**, 2, market, something, first, back, us, said
2	Drugs??? or Reviews???	$\lambda = 1$: 10, time, high, lsd, vendor, deleted, price, product, drug, trip, mdma, 2, acid, lol, 5, 4, around, took, drugs, tabs
3	Moderator posts	$\lambda = 1$: reddit, http, onion, com, r, please, vendor, deleted, time, https, message, www, darknetmarkets, wallet, sub, rules, compose, questions, always, contact
4	Security or anonymity	$\lambda = 1$: never, deleted, time, account, **tails**, back, vendor, someone, time, **btc**, 2, lol, darknetmarkets, **tor**, **pgp**, used, something, questions, always, contact $\lambda = 0.2$: gmail, email, 2cbking, password, **vpn**, phone, **tor**, provider, signal, **tails**, cash, fingerprint, bank, snapchat, secure, os, **burner**, card, upgrade, sms
5	Vendor Rreview	$\lambda = 1$: 1, x, vendor, 5, value, review, cost, product, 2, shipping, 3, price, 4, 0, package, shipped, stealth, communication, market, marked
6	Moderator or bot posts	$\lambda = 1$: market, **pgp**, r, message, please, u, encrypt, time, darknetmarkets, reddit, never, automatically, vendor, contact, questions, concerns, subreddit, compose, bot, users
7	Moderator posts	$\lambda = 1$: r, please, darknetmarkets, reddit, https, com, www, message, automatically, rules, subreddit, questions, compose, bot, action, contact, moderators, performed, concerns, post
8	Vendors???	$\lambda = 1$: com, http, **ab**, vendor, deleted, time, **dhl**, vendors, www, imgur, market, https, **bitcoin**, back, lot, anyone, **alphabay**, order, read, pretty
9	Drugs???	$\lambda = 1$: 4, heroin, u, 3, mdma, day, https, 5, 2, lsd, time, vendor, little, tabs, pretty, ordered, lol, first, year, meth
10	???	$\lambda = 1$: time, back, heroin, ketamine, k, tried, morphine, ca, great, u, work, tramadol, first, around, man, im, best lot, **mxe**, hard

Table B.9
Topics for May 2017.

#	Topic	Relevancy λ and terms
1	General vendor discussion	$\lambda = 1$: vendor, order, time, ordered, pack, vendors, never, days, lol, said, name, deleted, dispute, package, drugs, day, anything, back, orders, feedback $\lambda = 0.5$: order, vendor, dispute, ordered, pack, packs, feedback, vendors, bars, orders, name, box, mail, lol, ordering, never, said, xanax, package, house
2	General vendor/market discussion	$\lambda = 1$: vendor, order, time, market, us, 5, days, vendors, deleted, back, 2, post, never, product, 3, day, great, message, something, last
3	Cryptocurr. and security	$\lambda = 1$: **bitcoin**, **btc**, money, drugs, buy com, deleted, time, dnm, https, back, **monero**, cash, someone, **tails**, drug, ca, coins, reddit market, $\lambda = 0.2$: **monero**, **bitcoin**, **lbc**, **tails**, **xmr**, **ethereum**, **litecoin**, exchange, **zcash**, cash, anonymous, **localbitcoins**, banks, **eth**, currency, bible, privacy, guide, **shapeshift**, anonymity
4	Moderator posts	$\lambda = 1$: r, darknetmarkets, please, reddit, message, com, https, www, automatically, subreddit, rules, vendor, contact, questions, post, compose, action, bot, moderators, concerns,
5	Formated review	$\lambda = 1$: 10, 5, 1, vendor, x, review, price, value, product, shipping, 3, 2, cost, 4, shipped, stealth, quality, 0, package, market
6	General conversation	$\lambda = 1$: drugs, time, package, **opsec**, never, 1, said, u, life, lol, house, pretty, real, deleted, 2, day, though, 4, something, money
7	Markets	$\lambda = 1$: **ab**, money, account, back, time, **btc**, deleted, https, com, wallet, u, never, man, drugs, **dream**, days, market, meth, 2, lol
8	Transactions	$\lambda = 1$: fee, transaction, wallet, fees, **btc**, transactions, time, **electrum**, buy, low, drug, send, unconfirmed, days, high, set, drugs, higher, pay, **bitcoin**
9	Drugs	$\lambda = 1$: cocaine, test, lsd, dog, https, drug, time, drugs, used, high, acid, water, www, org, never, dogs, meth, 2, coke, lol
10	???	$\lambda = 1$: time, 5, high, day, around, never, lsd, drug, meth, work, kratom, drugs, always, back, com, though, better, pretty, bad, tolerance

Table B.10
Topics for June 2017.

#	Topic	Relevancy λ and terms
1	General vendor conversation???	$\lambda = 1$: time, deleted, lol, vendor, drugs, never, order, said, us, 2, money, back, drug, days, meth, something, yeah, u, day, first $\lambda = 0.5$: deleted, lol, drugs, time, meth, said, sell never, yeah, gay, tell, us, money, drug, pills, life, order, xanax, day, something
2	Markets and vendors	$\lambda = 1$: vendor, market, 1, 2, **ab**, x, **btc**, 3, vendors, time, order, 0, wallet, **bitcoin**, **alphabay**, account, never, new, **hansa**, cost $\lambda = 0.5$: market, x, wallet, **ab**, **btc**, 0, 1, **monero**, vendor, **escrow**, **bitcoin**, 2, **hansa**, cost, **alphabay**, fee, coin, **pgp**, 3, vendors
3	Moderator posts	$\lambda = 1$: r, darknetmarkets, please, post, reddit, time, message, order, com, vendor, new, https, subreddit, automatically, package, rules, name, www, mail, account
4	Formatted review	$\lambda = 1$: 10, vendor, 5, review, product, price, shipping, stealth, quality, market, time, shipped, us, communication, value, 1, mdma, https, days, high, weight
5	Moderator posts	$\lambda = 1$: com, reddit, https, r, please, message, darknetmarkets, www, compose, rules, automatically, contact, questions, bot, subreddit, action, moderators, concerns, performed, removed
6	???	$\lambda = 1$: deleted, drug, time, post, said, vendor, drugs, **vpn**, never, man, someone, **tails**, used, getting, something, account, maybe, **tor**, pretty, year
7	Drugs	$\lambda = 1$: time, drug, said, dmt, 10, lsd, doctors, fentanyl, 2, hard, 3, day, long, xanx, 4, gg, back, tabs, acid, drugs
8	???	$\lambda = 1$: fentanyl, u, vendor, man, com, feedback, mining, court, back, **tails**, us, message, fent, last, day, deleted, since, used, said, anyone
9	???	$\lambda = 1$: 10, vendor, mdma, ec, lsd, 100, hcl, shipping, deleted, powder, yes, stealth, 77, gel, u, excellent, tabs, quality, pure, price
10	jwh article???	$\lambda = 1$: jwh, state, vendor, legal, deleted, 018, moonshine, **btc**, dnm, weed, buy, company, indictment, stuff, fedex, purplelab, ordering, find, thing, youtube

Table B.11
Topics for July 2017.

#	Topic	Relevancy λ and terms
1	Markets and vendors	$\lambda = 1$: market, vendor, vendors, **hansa**, **dream**, time, **le**, **pgp**, order, key, **ab**, markets, deleted, never, new, back, address, used, 2, message $\lambda = 0.2$: vendors, **dream**, key, **pgp**, **compromised**, **le**, vendor, **hansa**, market, encrypt, order, buyers, auto, **2fa**, address, markets, **escrow**, **ab**, orders, decrypt
2	General conversation???	$\lambda = 1$: fent, drugs, drug, market, time, fentanyl, life, us, markets, money, dnm, many, never, **le**, deleted, things, years, something, better
3	Bot posts	See November 2016 for similar terms.
4	Hansa and Alphabay busts	$\lambda = 1$: https, **alphabay**, **hansa**, market, **ab**, www, com, **le**, reddit, us, **fbi**, email, site, time, web, post, server, servers, maybe, dark $\lambda = 0.5$: **alphabay**, **alpha02**, **fbi**, web, cazes, email, **europol**, servers, criminal, investigation, trappy, admin, dark, desnake, justice, server, **ab**, thailand, gov, article
5	Formatted vendor review	$\lambda = 1$: 10, vendor, 1, 2, 5, review, market, product, https, 3, shipping, price, x, shipped, value, 4, stealth, time, order, **hansa**
6	Article???	$\lambda = 1$: time, com, mail, http, said, lol, us, drugs, drug, **police**, https, anything, ca, www, never, lawyer, name, years, man, fud
7	Cryptocurr.	$\lambda = 0.2$: **bitcoin**, aktif, **bcc**, blockchain, **fork**, **coinbase**, bch, **btc**, **monero**, exchanges, **electrum**, **bitmixer**, **localbitcoins**, wallet, **shapeshift**, **tumbling**, **helix**, aug, **tumble**, uahf
8	???	$\lambda = 1$: **dhl**, server, market, www, coins, mods, 1, com, user, buy, **tor**, post, https, cash, com, **bitcoins**, **vpn**, mod, u, money
9	Ads for .onion sites	$\lambda = 1$: onion, http, register, darkheroesq46awl, **monero**, **xmr**, 0, 2, deleted, meth, thanks, market, **dhl**, **bitcoin**, wallet, back, time, new, free, ca
10	???	$\lambda = 1$: email, tcp, glass, open, vendor, starbucks, weed, **i2p**, heard, accounts, market, **tor**, getting, jar, **alphabay**, tried, anyone, filtered, thanks, dont

Table B.12
Topics for August 2017.

#	Topic	Relevancy λ and terms
1	Markets and vendors	$\lambda = 1$: market, vendor, **dream**, vendors, time, order, markets, **le**, never, account, 2, **pgp**, site, back, post, **hansa**, deleted, address, money, anyone $\lambda = 0.2$: market, **dream**, **multisig**, markets, password, site, **hansa**, vendors, **sig**, support, **scam**, **compromised**, **phishing**, **2fa**, **multi**, **sourcery**, **le**, passwords, admin, **escrow**
2	Vendor review	See November 2016 for similar terms.
3	Drugs and law enforcement	$\lambda = 1$: drugs, drug, com, https, fent, time, reddit, **le**, fentanyl, darknet, **police**, www, said, used, us, **law**, please, deleted, r, lol $\lambda = 0.5$: drugs, drug, fentanyl, fent, darknet, **police**, web, **law**, cops, **enforcement**, arrested, dark, 2017, authorities, federal, arrest, illegal, article, investigation, **le**
4	Moderator posts	See December 2016 for similar terms.
5	Cryptocurr.	$\lambda = 0.2$: **monero**, **xmr**, **btc**, **bitcoin**, exchange, **shapeshift**, cash, **coinbase**, bcc, **tumbling**, **lbc**, exchanges, currency, wallet, 50k, fees, 08, mymonero, bch, kraken
6	Formatted vendor review	See November 2016 for similar terms.
7	Discussion about Silk Road	$\lambda = 1$: 01, 2013, money, 27, **monero**, ross,
8	Anonymity	$\lambda = 1$: **tor**, https, **anonymity**, **vpn**, transactions, com, set, time, **monero**, network, **tails**, work, **security**, 1, address, **i2p**, node, transaction, never, used $\lambda = 0.5$: **tor**, **vpn**, **anonymity**, **i2p**, utxo, node, **zerocash**, network, nodes, perpetrator, **vpns**, farber, ricochet, iota, js, ripple, apple, **whonix**, algorithm, **ipv6**
9	DHL market???	$\lambda = 1$: **dhl**, u, deleted, https, sub, r, sub, r, reddit, comments, banned, com, time, please, post, back, www, image, lol 3, vendor, market
10	Vendor reviews???	See November 2016 for similar terms.

Table B.13
Topics for September 2017.

#	Topic	Relevancy λ and terms
1	Vendor discussion or review???	$\lambda = 1$: vendor, order, never, time, 2, days, market, back, vendors, 1, 3, **dream**, 5, **tr**, post, day, gg, lol, said, shipping
2	Cryptocurr. and anonymity	$\lambda = 1$: **btc**, **le**, market, **bitcoin**, wallet, **tor**, time, **monero**, **dream**, https, com, dnm, money, drugs, used, never, actually, something, coins, many $\lambda = 0.5$: **btc**, wallet, **bitcoin**, **monero**, **tor**, **le**, coins, **tails**, **xmr**, dnm, **blockchain**, drugs, money, **vpn**, deposit, phone, **opsec**, **dream**, fee, caught $\lambda = 0.2$: **btc**, wallet, **blockchain**, **monero**, **xmr**, **vpn**, **bitcoin**, fee, deposit, coins, **tor**, **tails**, beard, privacy, vallerius, **openbazaar**, phone, **shapeshift**, decentralized, **le**
3	Moderator posts	$\lambda = 1$: r, please, darknetmarkets, message, reddit, https, com, subreddit, www, automatically, contact, compose, questions, moderators, bot, action, concerns, performed, rules, wiki
4	Vendor review	$\lambda = 1$: vendor, 10, product, review, coke, order, vendors, shipping, test, time, quality, heroin, price, best, better, lol, back, never, 3, days
5	Vendor review???	$\lambda = 1$: 10, lsd, vendor, price, review, time, product, meth, shipping, 5, stealth, 3, tabs, trip, alice, quality, usa, 2, bob, 100
6	Vendor review???	$\lambda = 1$: 10, https, vendor, review, com, price, product, stealth, shipping, time, 5, image, message, net, reddit, anonimage, comments, 3, bars, high
7	General conversation?	$\lambda = 1$: lol, u, used, north, drug, time, said, bad, korea, butane, dmt, someone, drugs, meth, cup, punch, come, com, anyone, new
8	Drugs???	$\lambda = 1$: https, mdma, com, www, v, time, watch, drug, drugs, youtube, 1, lol, mda, password, marijuana, first, 2, years, delted, money
9	Drugs	$\lambda = 1$: fent, **tails**, fentanyl, **tor**, 3, 1, heroin, drugs, com, **electrum**, said, probably, 2, drug, **monero**, **alpha02**, every, https, usb, new
10	DDoS attack article???	$\lambda = 1$: **ddos**, **pgp**, magento, attack, key, **tr**, app, **le**, admin, site, signature, attacks, signed, time, mirrors, onion, **alphabay**, **traderoute**, code, password

Table B.14
Topics for October 2017.

#	Topic	Relevancy λ and terms
1	Markets and vendors	$\lambda = 1$: market, vendor, **dream**, vendors, time, markets, never, money, **tr**, back, order, **scam**, **le**, **exit**, new, **aero**, site, reddit, 2, something $\lambda = 0.5$: market, **dream**, vendors, markets, money, time, **scam**, **exit**, **tr**, never, back, **le**, site, **aero**, order, **escrow**, everyone, coins, **btc**
2	Moderator posts	See December 2016 for similar terms.
3	Formatted vendor review	See November 2016 for similar terms.
4	Drugs	$\lambda = 1$: drug, drugs, fentanyl, fent, u, heroin, said, pills, time, lol, years, life, selling, china, bars, xanax, **police**, money, lot, us
5	Anonymity and story about undercover agent	$\lambda = 1$: https, **vpn**, com, **tor**, reddit, web, **le**, **logs**, **police**, us, time, **ip**, www, phishkingz, information, message, r, used, 1 $\lambda = 0.2$: **vpn**, **logs**, phishkingz, insanitydrm, pia, **purevpn**, **trishula**, **vpns**, **bitbay**, **cve**, imguralbumbot, eg, lin, motherboard, ignoreme, autplayed, hacking, delet, web, administrators
6	Security and drugs	$\lambda = 1$: **tails**, nitty, time, life, never, **usb**, post, drugs, drug, heroin, **tor**, toh, dnm, https, pharma, us, 2cb, suboxone, dnms, man $\lambda = 0.5$: nitty, **tails**, **usb**, toh, suboxone, 2cb, laptop, gloss, pharma, gbl, emboss, 2cbking, microsoft, **burner**, cocaethylene, hbr, **whonix**, coked, stored, ghb
7	Secure transactions	$\lambda = 1$: key, **pgp**, tx, https, vendor, message, vendor, public, address, **keys**, signed, **multisig**, sign, com, private, used, u, transaction, reddit, r
8	???	$\lambda = 1$: vendor, **mxe**, time, us, 2017, order, dont, u, post, every, vendors, mdma, **pgp**, may, never, nl, r, though, address, trying
9	Silk Road???	$\lambda = 1$: package, 2013, **police**, drugs, house, https, com, name, information, mail, silk, road, warrant, u, message, open, pack, marijuana, vendor, packages
10	Drugs	$\lambda = 1$: meth, tt, time, flair, test, never, someone, point, trust, lsd, lol, since, actually, stuff, tmg, year, every, maybe, vendor

References

Beebe, N.L., Liu, L., 2014. Clustering digital forensic string search output. Digit. Invest. 11 (4), 314–322.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. In: Proceedings of the 23rd International Conference on Machine Learning. ACM, pp. 113–120.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M., 2009. Reading tea leaves: how humans interpret topic models. In: Advances in Neural Information Processing Systems, pp. 288–296.

de Waal, A., Venter, J., Barnard, E., 2008. Applying topic modeling to forensic data. In: IFIP International Conference on Digital Forensics. Springer, pp. 115–126.

Deliu, I., 2017. Extracting cyber threat intelligence from hacker forums (Master's Thesis). NTNU.

Dingledine, R., Mathewson, N., Syverson, P., 2004. Tor: the Second-generation Onion Router. Tech. rep. Naval Research Lab, Washington DC.

Fang, Z., Zhao, X., Wei, Q., Chen, G., Zhang, Y., Xing, C., Li, W., Chen, H., 2016. Exploring key hackers and cybersecurity threats in Chinese hacker communities. In: Intelligence and Security Informatics (ISI), 2016 IEEE Conference on. IEEE, pp. 13–18.

Franceschi-Bicchierai, L., 2017. IRS Cybercrime Agent Lurks Dark Web Subreddit Looking for Hackers. https://motherboard.vice.com/en_us/article/wjxwvm/irs-cybercrime-agent-lurks-dark-web-subreddit-looking-for-hackers. (Accessed 17 November 2017).

Franceschi-Bicchierai, L., 2018. Reddit Bans Subreddits Dedicated to Dark Web Drug Markets and Selling Guns. https://motherboard.vice.com/en_us/article/ne9v5k/reddit-bans-subreddits-dark-web-drug-markets-and-guns. (Accessed 23 March 2018).

Gibbs, S., Beckett, L., 2017. Dark Web Marketplaces AlphaBay and Hansa Shut Down. https://www.theguardian.com/technology/2017/jul/20/dark-web-marketplaces-alphabay-hansa-shut-down. (Accessed 3 November 2017).

Grisham, J., Barreras, C., Afarin, C., Patton, M., Chen, H., 2016. Identifying top listers in Alphabay using latent dirichlet allocation. In: Intelligence and Security Informatics (ISI), 2016 IEEE Conference on. IEEE, 219–219.

Knibbs, K., 2015. Feds Want Reddit to Give up Personal Info of Darknet Market Redditors. https://gizmodo.com/feds-want-reddit-to-give-up-personal-info-of-darknet-ma-1694608548. (Accessed 15 November 2017).

Noel, G.E., Peterson, G.L., 2014. Applicability of latent dirichlet allocation to multi-disk search. Digit. Invest. 11 (1), 43–56.

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

Okolica, J.S., Peterson, G.L., Mills, R.F., 2007. Using Author Topic to detect insider threats from email traffic. Digit. Invest. 4 (3–4), 158–164.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., Smyth, P., 2004. The author-topic model for authors and documents. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence. AUAI Press, pp. 487–494.

Ruffing, T., Moreno-Sanchez, P., Kate, A., 2014. CoinShuffle: practical decentralized coin mixing for Bitcoin. In: European Symposium on Research in Computer Security. Springer, pp. 345–364.

Samtani, S., Chinn, R., Chen, H., 2015. Exploring hacker assets in underground forums. In: Intelligence and Security Informatics (ISI), 2015 IEEE International Conference on. IEEE, pp. 31–36.

Shen, J.H., Rudzicz, F., 2017. Detecting anxiety through reddit. In: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—from Linguistic Signal to Clinical Reality, pp. 58–65.

Sievert, C., Shirley, K.E., 2014. Ldavis: a method for visualizing and interpreting topics. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces, pp. 63–70.

Van Buskirk, J., Naicker, S., Bruno, R., Breen, C., Roxburgh, A., 2016. Drugs and the internet. Issue 7.