Characterizing Activity on the Deep and Dark Web

Nazgol Tavabi[1],Nathan Bartley[1],Andrés Abeliuk[1],Sandeep Soni[2],Emilio Ferrara[1],Kristina Lerman[1]
1. University of Southern California, Information Sciences Institute
2. Georgia Institute of Technology

ABSTRACT

The deep and darkweb (d2web) refers to limited access web sites that require registration, authentication, or more complex encryption protocols to access them. These web sites serve as hubs for a variety of illicit activities: to trade drugs, stolen user credentials, hacking tools, and to coordinate attacks and manipulation campaigns. Despite its importance to cyber crime, the d2web has not been systematically investigated. In this paper, we study a large corpus of messages posted to 80 d2web forums over a period of more than a year. We identify topics of discussion using LDA and use a non-parametric HMM to model the evolution of topics across forums. Then, we examine the dynamic patterns of discussion and identify forums with similar patterns. We show that our approach surfaces hidden similarities across different forums and can help identify anomalous events in this rich, heterogeneous data.

CCS CONCEPTS

• **Mathematics of computing → Time series analysis**; **Exploratory data analysis**; *Bayesian nonparametric models*; *Cluster analysis*; • **Information systems → Deep web**; *Clustering*; *Web crawling*; • **Security and privacy** → *Social aspects of security and privacy.*

KEYWORDS

Darkweb; Deepweb; D2web; Cyber Crime; Cyber Security; LDA; Non-Parametric HMM; Beta Process; multivariate Time Series; Cluster Time Series

ACM Reference Format:
Nazgol Tavabi[1],Nathan Bartley[1],Andrés Abeliuk[1],Sandeep Soni[2],Emilio Ferrara[1],Kristina Lerman[1]. 2019. Characterizing Activity on the Deep and Dark Web. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3308560.3316502

1 INTRODUCTION

The web that most people are familiar with—the open and searchable internet of social media sites, online merchants, newspapers and the like—represents just a tiny fraction of the internet. Much larger portions of internet data remain buried within the "deepweb" [18], a term that refers to private corporate intranets and databases, dynamically-generated web pages, and limited access content, such as online academic journals. While only some of the content is hidden behind encrypted protocols in .onion domains, access to the remaining "open" content is generally restricted, requiring registration and authentication [26]. Some portion of the deepweb, also known as the "darkweb," serves as a hub for all kinds of illicit activities. Malicious actors congregate virtually on dark web forums and marketplaces to trade illicit information, goods and services, including ransomware, exploits, hacking tools, stolen media, user credentials, fake ids, prescription medicines and illegal drugs. In addition to serving as a marketplace for these goods, the deep and dark web provides a venue for malicious actors to coordinate cyber attacks [25] and terrorist activity [7]. The growing popularity of the marketplaces within the deepweb can be attributed to the elimination of the risk of violence since there is limited, if any, physical interaction between the buyers and the sellers. Another reason is the use of encrypted protocols to preserve anonymity, encouraging people to express themselves without the risk of getting caught by law enforcement nor being censored by the moderators of a web site [27].

Given the threat posed by these malicious actors, observing their activities on the deep and dark web (d2web) may provide valuable clues both for anticipating and preventing cyber attacks as well as mitigating the fallout from data breaches. However, picking out useful signals in the vast, dynamic and heterogeneous environment of the d2web can be challenging.

In this paper, we use Latent Dirichlet Allocation (LDA) [5] to analyze a large heterogeneous text corpus from the d2web. To understand the dynamics of discussions, we use a non-parametric hidden Markov model [10]—the *Beta Process HMM*. In this approach every forum is represented as a multivariate time series, where variables are the topics found by LDA, and is fed into a Beta Process HMM (BP-HMM). This BP-HMM then finds the shared states among forums, where each state is a distribution over topics. This helps track discussions on different forums and identify anomalous behavior or important events. This approach can also be used to find forums relevant to a specific subject, i.e., forums or time periods within forums where users discuss specific topics, such as hacking techniques and cyber security related issues. This method can also cluster forums into meaningful groups.

We test this framework on data consisting of posts published on 80 D2web forums from 2016 to mid 2017, on topics such as exploits and hacking techniques, selling prescription and non-prescription drugs, and creating fake ids. Overall this paper makes the following contributions:

- Using LDA, we characterize the content of 80 d2web forums where illicit activities are discussed.
- We describe an application of a non-parametric HMM model to learn the forum's shared topic dynamics based on the results obtained from LDA.
- We use the learned shared behaviors as a compact representation of these time series to cluster forums into groups with

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316502

distinct characteristics and analyze learned latent behavioral structures to gain more insight into this data.
- We rank forums based on how their topics of discussions are likely to change.
- Finally, we present case studies revealing how our approach can be used to to identify anomalous activity in d2web data.

2 RELATED WORK

For an overview of the darkweb and deepweb and the challenges these sites pose for researchers and for law enforcement, please see [11, 18]. Researchers have leveraged d2web content in specific applications, but to the best of our knowledge, few have attempted to systematically characterize the topics and dynamics of d2web discussions. For example, Xu et al. [31] analyzed the topology and structure of darkweb networks. Soska et al. [27] analyzed 16 different marketplaces to extract information on goods being sold and money transactions, then trained classifiers on them. Tavabi et al. [28] and Almukaynizi et al. [2] used features from d2web discussions to predict which new vulnerability will be exploited. Goyal et al. [13] and Deb et al. [8] used the frequency of important cyber security-related keywords and the sentiment of d2web posts respectively to predict cyber attacks. Along similar lines, [6] utilized sentiment to analyze communications of extremists on the darkweb and [1] developed a bilingual (English and Arabic) sentiment analysis lexicon for cyber security and radicalism on darkweb forums.

In this paper we propose a statistical approach to analyze discussions of multiple malicious forums by extracting their similarities and their differences through learning hidden Markov models on topic weights obtained from LDA. Latent Dirichlet Allocation (LDA) proposed by Blei et al. [5] is a powerful ubiquitous tool which allows documents to be explained by latent topics. LDA has shown to be effective in the analysis of the darkweb [19, 24]. However, the dynamics of darkweb topics, which can give deeper insight into this data, has not been analyzed previously. Rios et al. [24] used LDA to detect overlapping communities in the darkweb and L'huillier et al. [19] applied LDA to extract key members in darkweb forums. Extensions to LDA, like Dynamic Topic Models [4], have been designed to model evolution of topics across time. However, such models have strong memory requirements making them impractical for a large heterogeneous corpus like the d2web. An alternative approach is performing LDA on the entire corpus and observing fluctuation of weights across time, similar to approaches set forth by [20, 30] and many others. Variants of LDA which combine Hidden Markov Models have also been proposed. The approach proposed by Griffiths et al. [14] models both semantic and syntactic dependencies of documents by defining one state in HMM as the semantic state modeled by LDA and other states as syntactic components. Gruber et al. [15] modeled the relationship between words in a document with an HMM. Their model assumes words in the same sentence have the same topic, and successive sentences are more likely to have the same topics, where in the original LDA paper words in a document are assumed to be independent and documents are modeled as bag of words. These variations were proposed to better capture LDA topics, although Hidden Markov models can also be used to identify shifts and variations in the topics discussed, which is the approach of this work. HMM states can help us recognize important events and anomalies.

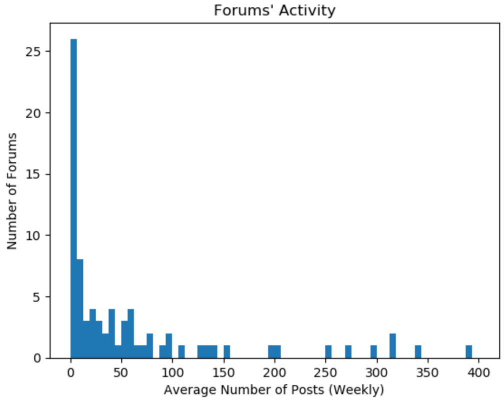


Figure 1: Histogram of activity level in different forums.

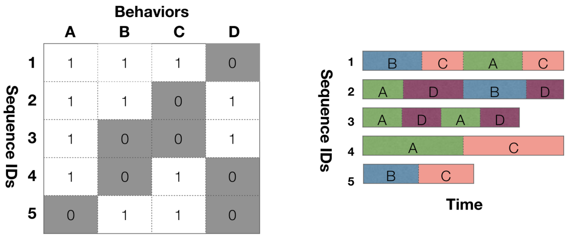
3 METHODS

3.1 D2Web Data Collection

The d2web data we use in the study was collected using the crawling infrastructure described in [22, 25]. This crawling infrastructure uses anonymization protocols, such as Tor and I2P, to access darkweb sites, and handles authentication to access non-indexed deepweb sites on the Internet. The infrastructure includes lightweight crawlers and parsers that are focused on specific sites related to malicious hacking and/or online financial fraud. These sites represent forums and discussion boards where people mostly discuss cyber crime and fraud, although other illicit activities are discussed as well, such as the sale of drugs and other stolen goods. There are also a handful of forums crawled that are on the clearnet but which are mostly white hat (i.e., involved with ethical hacking and/or professional cybersecurity). We include these forums in our analysis primarily to help identify other forums that might discuss similar topics. In all, crawlers scraped data from over 250 d2web forums. The most common languages in which the posts were written were English (accounting for 37.8% of all posts), Russian (22.4% of all posts), and Chinese (15.4% of all posts). Other languages, such as Spanish, Arabic, and Turkish were less frequent, each accounting for less than 7% of the posts. For this analysis we only focus on English posts, though the same structure could be used for multiple languages. Filtering out the non-English posts brings the number of forums down to 155. We pre-processed the posts using NLTK [3], SpaCy [17], and scikit-learn to remove stopwords, tokenize each post, and filter tokens by post frequency to remove frequent words. This gives us a corpus of 1.33 million posts.

3.2 Modeling Topics of Discussion

We applied a popular statistical technique known as Latent Dirichlet Allocation (LDA) [5] to learn the topics of the English-language posts. LDA is used to decompose documents into latent topics, where each topic is a distribution over words, intended to capture the semantic content of documents. In this model each document



	A	B	C	D
1	1	1	1	0
2	1	1	0	1
3	1	0	0	1
4	1	0	1	0
5	0	1	1	0

Figure 2: Illustration of the model. Left image is matrix F. Based on this matrix time series 1 exhibits states A,B and C but doesn't exhibit state D. The right image shows segmentation of time series with hidden states

is treated as a bag-of-words. Once we learn the model we can represent documents as distributions over a fixed number of topics. Doing so gives us low-dimensional representations of documents.

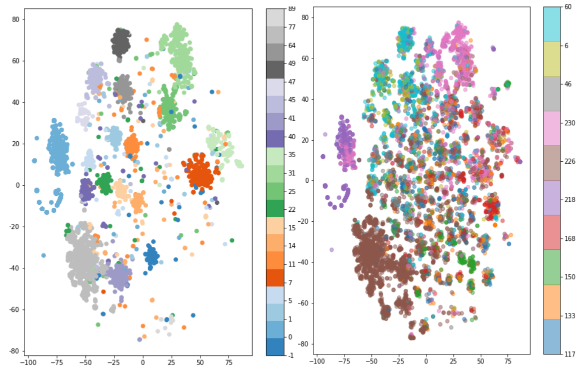
In this paper we use the Gensim implementation of LDA [23] to learn a model with 100 topics. We train the model on all 1.33 million documents to learn the most informative topics. We tested with 50, 100, and 200 topics respectively, and found that 100 topics results in the most coherent and relevant topics. To examine the dynamics, we focused on time period of 2016 until mid September 2017 which has the best coverage in our data. We only looked at forums with at least one month of activity and more than 100 posts overall, which reduced our data set to 80 forums (and approximately 482 thousand posts). Figure 1 shows the level of activity in these 80 forums. The activity is highly heterogeneous, with some forums seeing hundreds of posts per week, and other forums showing little activity.

3.3 Modeling Dynamics of Activity

Hidden Markov Models (HMMs) have been used extensively in modeling dynamic processes and time series in a variety of applications. These generative models segment time series into a predefined number of latent states and learn transition rates between them.

When modeling multiple dynamic processes with HMMs, it is useful to represent them with global states that are shared between these time series, rather than modeling each dynamic process independently and then learning the mapping between states. Joint modeling facilitates comparison of different processes and learns more generalizable models. It is also convenient to work with a non-parametric model that does not fix the number of states a priori.

In this paper, to model activity within the d2web, we describe each forum as a time series of topic vectors representing discussions. We use a generative model proposed by [9, 10], called *Beta Process HMM* (BP-HMM), to identify latent states shared by different time series. Based on the proposed model, different time series are described by a subset of shared latent states. The states are represented by a binary matrix F, where $F_{ij} = 1$ means time series i is associated with state j. An example of F matrix is shown in the left panel of Figure 2. Given matrix F, each time series is modeled as a separate hidden Markov model with the states it exhibits. Each global state is modeled using a multivariate Gaussian distribution, when a time series is in state x, its data is sampled from a Gaussian distribution with mean vector μ_x and covariance matrix Σ_x.



(a) topics

(b) forums

Figure 3: Visualization of the d2web discussions in May 2016 using t-SNE. Each dot represents a post, with its color representing (a) the topic or (b) forum to which the post belongs. The top 20 topics learned by the 100-topic LDA are shown, and the post was assigned to the highest probability topic.

Since the number of such states in the data is not known a priori, the Beta process is used [16, 29] as a prior on matrix F. A Beta process allows for infinite number of states but encourages sparse representations. Consider, as an example, a model with K states. Each state (column of matrix F) is modeled by a Bernoulli random variable whose parameter is obtained from a Beta distribution (Beta Bernoulli process), i.e.,

$$\begin{aligned} \theta_k &\sim \text{Beta}(\alpha/k, 1), k = 1, \cdots, K \\ F_{nk} &\sim \text{Bernoulli}(\theta_k), n = 1, \cdots, N \end{aligned} \tag{1}$$

The underlying distribution when this process is extended to an infinite number of states, as K tends to infinity, is the Beta process. This process is also known as the *Indian Buffet Process* [12, 29] which can be best understood with the following metaphor involving a sequence of customers (time series) selecting dishes (states) from an infinitely large buffet. The first customer enters the buffet and selects servings from $Poisson(\alpha)$ number of dishes. The n-th customer selects dish k with probability m_k/n, where m_k is the popularity of the dish, yielding the so-called "rich-get-richer" effect, and $Poisson(\alpha/n)$ new dishes.

With this approach, the number of states can grow arbitrarily with the size n of the dataset: in other words, the number of states increases if the data cannot be faithfully represented with the already defined states. However, the probability of adding new states decreases according to a $Poisson(\alpha/n)$. Finally, the distribution generated by the Indian Buffet Process is independent of the order of the customers (time series). For posterior computations based on MCMC algorithms, the original work is referenced [9, 10].

3.4 Clustering

To define a similarity measure between two HMMs, one could measure the probability of their state sequences having been generated by the same process. Since each time series is associated with a

Topic (Manually Labeled)	Top 10 Keywords
Vending	
1. Locations	checked, live, united states, unknown, california, carolina, south, ca, nj, new
2. Money	money, people, make, pay, want, free, buy, just, like, sell
3. Pharmaceuticals	buy, online, prescription, cheap, cod, xanax, delivery, overnight, order, day
4. Banking	card, bank, credit, cards, paypal, account, business, debit, accounts, gift
5. Fake IDs	fake, real, id, original, high, english, license, quality, registered, passports
6. Purchase details	order, vendor, days, sent, orders, ordered, package, received, shipped, just
7. LSD	like, just, quote, lsd, really, good, feel, know, tabs, experience
8. Cryptocurrency	bitcoin, btc, wallet, address, send, coins, bitcoins, transaction, sent, account
9. Marijuana	like, good, got, weed, time, bit, high, great, nice, low
10. Markets	market, vendor, vendors, dream, alphabay, markets, scam, ab, hansa, escrow
11. Narcotics	good, cocaine, quality, best, vendor, product, mdma, coke, free, order
Security	
12. Malware	virus, scan, antivirus, file, malware, clean, av, security, download, detected
13. Botnets	bot, attack, malware, used, domain, ddos, botnet, irc, hosting, attacks
14. Windows	windows, build, microsoft, xp, vista, beta, server, ms, longhorn, version
15. Social hacking	email, send, rat, stealer, message, keylogger, mail, facebook, crypter, download
16. Law enforcement	police, law, drug, drugs, enforcement, according, dark, illegal, said, darknet
17. Hacking tutorial	learn, know, want, good, learning, start, like, knowledge, programming, hacking
18. Carding	transfer, dumps, info, sell, cvv, good, track, balance, bank, uk
19. Web Vulnerabilities	web, sql, php, injection, exploit, code, server, site, script, page
20. OS Code	process, code, dll, memory, address, api, function, module, use, hook
21. Network Hacking	network, connect, wifi, wireless, ip, internet, router, connected, pineapple, fon
22. Security	information, data, security, software, used, access, user, users, network, application
23. Proxy	use, tor, using, vpn, internet, proxy, browser, ip, web, access
24. Mobile phones	phone, android, phones, samsung, pixel, battery, note, camera, better, google
25. Update	install, installed, download, just, update, need, use, installing, using, try
Gaming	
26. Gaming Source Code	end, local, return, function, false, mod, script, item, nil, damage
27. Torrents	torrent, quote, download, upload, forget, left, feedback, like, plz, 720p
28. Gameplay	game, complete, level, win, play, kill, mode, team, player, single
29. Games	game, games, new, play, like, xbox, ps4, sony, playstation, console
30. PlayStation Vita	vita, ps, psp, game, firmware, exploit, games, sony, custom, psn
31. Emulators	game, games, vita, version, plugin, homebrew, psvita, emulator, use, play
32. Hacking Consoles	ps3, games, play, tutorial, game, cfw, console, psn, use, need
Other	
33. Contact	contact, pm, need, icq, want, send, add, rue, interested, na
34. Thanks	thanks, thank, man, lot, sharing, bro, thx, share, mate, nice

Table 1: Example topics in the 100-topic LDA model.Topics are labeled (in the first column) manually for convenience.

distinct generative process, we measure two state sequences' similarity as the likelihood that seq_i was generated by the process that gave rise to seq_j, and the likelihood that seq_j was generated by the process giving rise to seq_i. We average the two likelihoods to symmetrize the similarity measure.

$$\forall_{i,j}\text{Sim}(i,j) = \frac{p(\text{seq}_i|T_j) + p(\text{seq}_j|T_i)}{2} \tag{2}$$

The likelihood $p(\text{seq}_i|T_j)$ is computed using the learned transition matrix T_i and Markov process assumption. In transition matrix T_i, which is a square matrix with dimensions equal to the number of states, entry $T_i(m,n)$ gives the probability that time series i transitions from state m to state n. Matrix T_i is stochastic, with the sum of entries in each row equal to 1. Once the similarity between two HMMs is defined, we can perform a number of operations, including clustering similar time series together. For example, we use hierarchical agglomerative clustering method to automatically group forums (represented by their time series) with similar discussions.

4 RESULTS

4.1 Topic Analysis

In this section we explore topics learned by Latent Dirichlet Allocation (LDA) by looking at their most significant words. There are a wide variety of topics covered in our dataset. Table 1 highlights some of the topics learned by the 100-topic LDA model by showing the most significant words associated with each topic. We visualized the topics using t-distributed Stochastic Neighbor Embedding

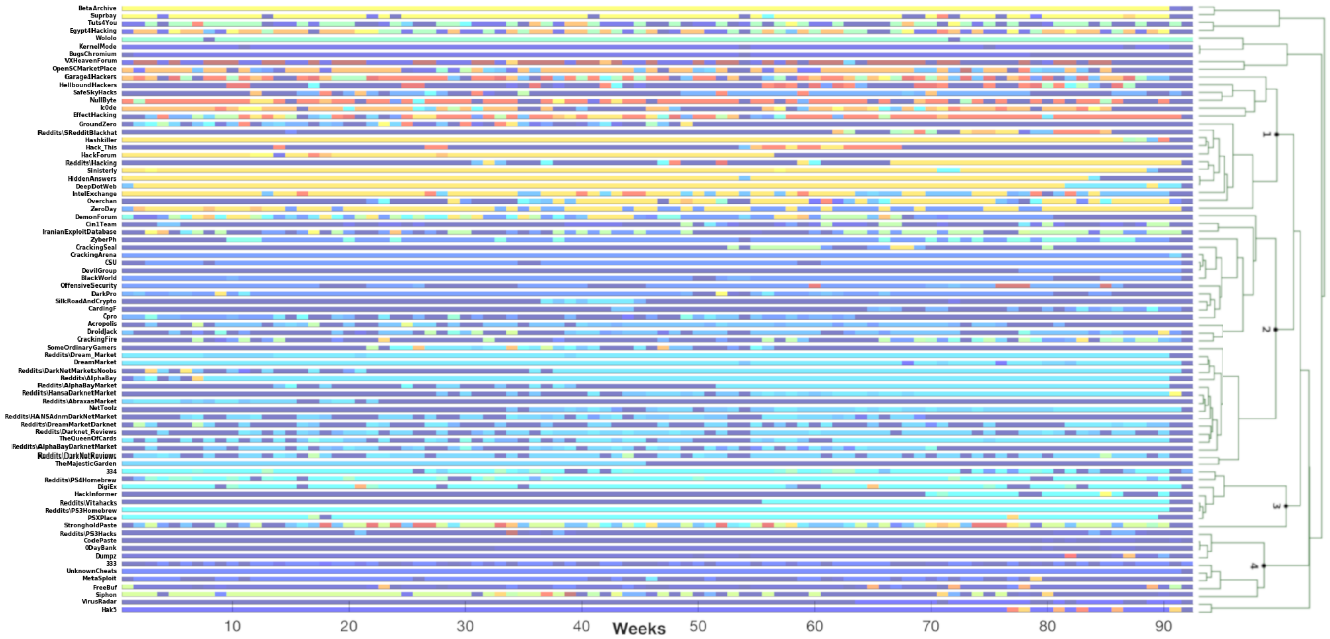


Figure 4: State sequences of forums (each color represents a state) and Dendrogram showing the similarity of forums based on their learned states.

(t-SNE) [21] in Figure 3(a). Each dot on the plot represents a post, colored by its most important topic. Alternatively, we can color the posts in the t-SNE space by the forum on which they were posted (Figure 3(b)). The clustering of posts in the forum space suggests that topics are highly concentrated around specific forums.

4.2 Forum Dynamics

To examine the dynamics of topics on d2web forums, we represent each forum as a time series of topic vectors learned by the 100-topic LDA model. Our unit of time in this analysis is a week; to generate a forum's vector we average the topic vectors of all posts submitted to the forum over the course of a week. The time series of weekly topic vectors were used to both learn HMM states and compute cross-entropy (cross-entropy is discussed further in 4.3).

After training the BP-HMM model on weekly topic distributions of forums the model learned 28 states, i.e., 28 different topic distributions. We clustered the forums according to the similarity of their learned states using the method described in Section 3.4. Figure 4 shows the resulting dendrogram and also the sequences of learned states for each forum. Each line in the figure represents a forum, and different states are represented by different colors. Transitions between states are visible in places the colors alternate.

The clustering results show that the method is able to cluster forums into meaningful groups. Next we examine a few of the main clusters:

Cluster 1 mostly contains forums discussing cyber hacking, including HackForum, GroundZero, ZeroDay, DeepDotWeb, SafeSkyHacks. The two subgroups in this cluster differ mainly in their levels of activity. Forums in the first subgroup are less active (5 posts in a week on average), which is why their average weekly topic vector is more sensitive and their corresponding HMM changes state more frequently. In the active group (more than 50 posts in a week on average) the most common state is the yellow state, which corresponds to activation of the following topics (described in Table 1): 16.Law enforcement, 23.Proxy and 17.Hacking tutorial.

Cluster 2 mostly contains dark web marketplaces such as Abraxas Market, AlphaBay, Dream Market, Hansa Market and BlackWorld, as well as forums dedicated to their reviews. This cluster is also divided into two main subgroups: in the first subgroup the dark blue state dominates representing high activity of discussions regarding topics 1.Locations (which is mostly concerned with the sale of proxy servers), 33.Contact, 34.Thanks and 4.Banking. The second subgroup, depicted with light blue states, corresponds to the topics 6.Purchase details, 10.Markets, 8.Cryptocurrency and 11.Narcotics. Based on the clustering, one can characterize forums in the first subgroup as mostly selling proxy servers and sharing information about other marketplaces, while forums in the second subgroup are more involved in selling drugs. In section 4.4.2 we take a deeper look into forums in the first subgroup.

Cluster 3 is made up of forums related to hacking Playstation video game consoles where the most prominent state in this cluster is the Cyan state. The most active topics in this state are 32.Hacking Consoles and 25.Update.

Cluster 4 (as well as the two forums adjacent) contains forums which are predominantly focused on white hat hacking. Notable forums include Metasploit and Hak5. While 0daybank and FreeBuf are related, they are mostly in Chinese, hence their more active topics have many non-English tokens and are hard to interpret.

With the state sequences obtained from our BP-HMM model, we can track forums' discussions. State transitions indicate a significant change in discussions and could represent an event. However, as shown in Figure 4, some forums change states more frequently, thus their transitions might have less significance. In order to be able to recognize significant transitions, we describe the volatility measure, a forum's likelihood to drastically change its own topic distribution. As each forum is characterized by its learned transition matrix over the global states, we compute a forum's volatility by adding the off diagonal elements of its transition matrix, the probability of changing states. Since we are interested in finding variations in topics discussed rather than the activity of forums, the probabilities of the state corresponding to 0 posts (i.e., no data) were not taken into account. Also to validate the results obtained, a similar volatility measure was computed with cross entropy and is described in section 4.3. Table 2 shows a list of forums with high and low volatility computed via both methods.

Using the described volatility measure with the HMM, we find that the most volatile forums with at least 10 posts per week (on average) are OpenSC Marketplace, Stronghold Paste and Demon Forum and that the least volatile forums are BugsChromium, CSU, and the subreddit PS3Homebrew. An estimate of forum's volatility or lack thereof is also apparent from its state sequence in Figure 4. Stronghold Paste is an onion website similar to Pastebin and covers different topics and hence different states. However, there are two main states it oscillates between: in one of them hacking and cyber security topics (described by topic 19.Web Vulnerabilities) are more prominent, and in the other one, topic 8.Cryptocurrencies has high activation. These results show state transitions in forums like Stronghold Paste have a low probability of being indicative of an event. However transitions in forums like BugsChromium or CSU might be of more interest.

4.3 Topic Dynamics

Since our HMM segments forum dynamics into discrete states, it is less sensitive to noise observed in the data. On the other hand it might not be able to capture small meaningful changes or trends. Hence, we also compute cross entropy of forums as a measure of dispersion over time to validate the results obtained with our non-parametric HMM model. To compute the cross-entropy we use the following formula where Q is the topic distribution for a forum averaged over the entire timespan and P is the topic distribution in a unit of time.

$$H(p, q) = E_P[-\log_2 Q(x)] \tag{3}$$

The majority of the forums have cross-entropy values in the same range. The forum with the lowest average cross-entropy (and therefore lowest volatility) is CSU, a forum primarily concerned with credit card dumps, and as such makes sense that the forum would focus on very few topics of discussion. On the other hand, forums like CodePaste and DroidJack have some of the highest values of cross-entropy, which suggests that the topics of discussion are dispersed and change more often in these forums. Table 2 shows forums with lowest and highest average values of cross-entropy, used as a measure of volatility. This is consistent with the result we got from the HMM model, in the sense that forums with low or high volatility based on the HMM measure also appear in the bottom or top of the ranking based on the cross entropy measure. Results show that large and active forums have wide-ranging discussions on diverse topics however their average topic distribution is usually consistent over time. Forums focused on specific topics like CSU also tend to have low volatility. On the other end of the spectrum there are forums with medium or low activity with discussions spanning a wide range of topics like Stronghold Paste and CodePaste.

HMM-ranked volatility	Cross-entropy-ranked volatility
1. GroundZero	CodePaste
2. OpenSCMarketPlace	DroidJack
3. StrongholdPaste	EffectHacking
4. DemonForum	DemonForum
5. EffectHacking	Overchan
6. CrackingFire	HellboundHackers
7. Overchan	CardingF
8. NullByte	HackForum
9. Siphon	Reddits\ Hacking
...	...
71. UnknownCheats	TheMajesticGarden
72. DevilGroup	Dumpz
73. KernelMode	KernelMode
74. VirusRadar	BlackWorld
75. Reddits\Vitahacks	MetaSploit
76. BetaArchive	0DayBank
77. TheMajesticGarden	Reddits\DarkNetReviews
78. Reddits\PS3Homebrew	Wololo
79. CSU	VirusRadar
80. BugsChromium	CSU

Table 2: Volatility computed via HMM and via Cross-entropy. Ranked highest volatility to lowest.

4.4 Case Studies

4.4.1 Prescription Drugs. In this section we give an example of how this framework could be used to study d2web discussions. In the search for anomalies using the results obtained by our HMM we observed a rare state, exhibited only on 3 first weeks of June 2017 and first week of August 2017 by forum OffensiveSecurity where topic 3.Pharmaceuticals has its highest probability. OffensiveSecurity is considered a low volatility forum based on both measures computed in Table 2 which makes this transition more of interest.

By looking at the posts published on this forum and on the aforementioned dates we retrieved similar posts with variations in the names of the drugs being advertised. An excerpt from one of the posts is as follows: "...buy lynoral cheap buy generic femara buy modafinil online uk can i buy qsymia online buy cytotec online us buy lumigan online canada buy 25 mg lyrica buy vibramycin florida buy diflucan without buy synthroid online next day delivery buy xanax online us where to buy generic qsymia buy adderall no prescription buy cytotec in europe..."

This analysis shows a high and anomalous volume of advertisements for prescription drugs in the specified dates which suggests some precipitating event that merits further investigation.

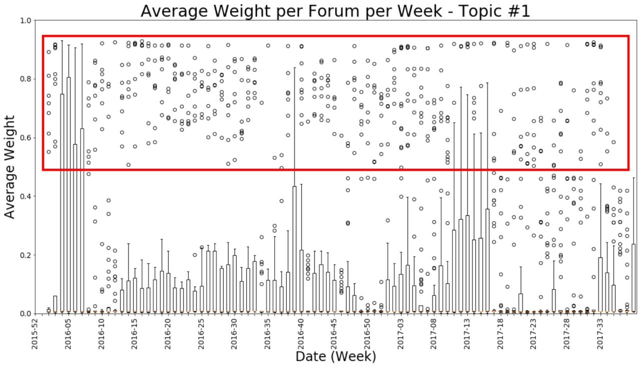


Figure 5: Boxplot Distribution of Proxy-server topic over the timespan of the corpus. Plotted are the 10 top forums in Cluster 2. Each point is the average weight of that topic for a forum for that week. The red box indicates the significant forums in each week.

We also looked at one of the most illicit drug topics, 11.Narcotics. We found that besides 11.Narcotics other prominent topics in the state where this topic is at its highest value are 9. Marijuana and 7.LSD. These are a few forums which exhibit this state: The Majestic Garden, the DarkNetReviews subreddit, and the HANSAdnmDarkNetMarket subreddit.

4.4.2 Proxy Servers. The second case study is around the discussion of proxy servers and the sale of services that can be used to game social media platforms, defraud ad networks, build botnets, and effectively launder illegal activity. We use the clustering from HMM to help identify which kinds of forums have high activity related to proxy servers. Compared to many of the other topics, topic 1.Locations is concerned primarily with the sale of proxy servers. When we examined cluster 2 from the HMM clustering, we noticed a significant activation of this topic. Interestingly, the ten most active forums in cluster 2 seem to capture most of the activation of the topic over time. As seen in Fig. 5, we examine the forum-weeks (points) with significant probability (above 0.50) for topic 1.Locations from forums in cluster 2 and recover approximately 11,000 documents which seem to be automated posts advertising the sale of access to proxies all over the world. These 11,000 documents come mostly from CSU and BlackWorld: both forums have subforums dedicated to the advertising of proxies and are in the top 10 active forums in the cluster. An excerpt from one of these posts is as follows: "...camarillo | ca | unknown | united states | checked at vn5socks.netlive...". When we look for documents pertaining to more specific uses of proxies, we find approximately 20 posts that directly mention "viewbot". Viewbotting is the act of using bots to artificially inflate the number of views on a social media profile (e.g. YouTube and Twitch). As it can be difficult to determine if a viewer is a human or a bot, this can potentially trick the social media platform into thinking a profile is more popular than it actually is and result in more attention than it would obtain organically. Alternate uses of viewbots are to watch video ads on a channel, artificially increasing ad revenue. An excerpt from one of these documents is as follows: "...i viewbotted my vid to 1k views and got 10 slaves..."

Another potential malicious use of proxy servers is for carding. Carding in this case refers to the fradulent use of other people's credit cards, personal and/or financial data to purchase goods, launder money, or generally steal an individual's money. Proxy servers are commonly used to "cash out" stolen credit card information by buying items like pre-paid gift cards through payment processors. In our corpus, there are a number of documents that mention carding. An example that suggests intent to use proxies for carding is as follows: "...proxies are often blacklisted when used for fraud so im looking for a source for fresh proxies for use for carding..."

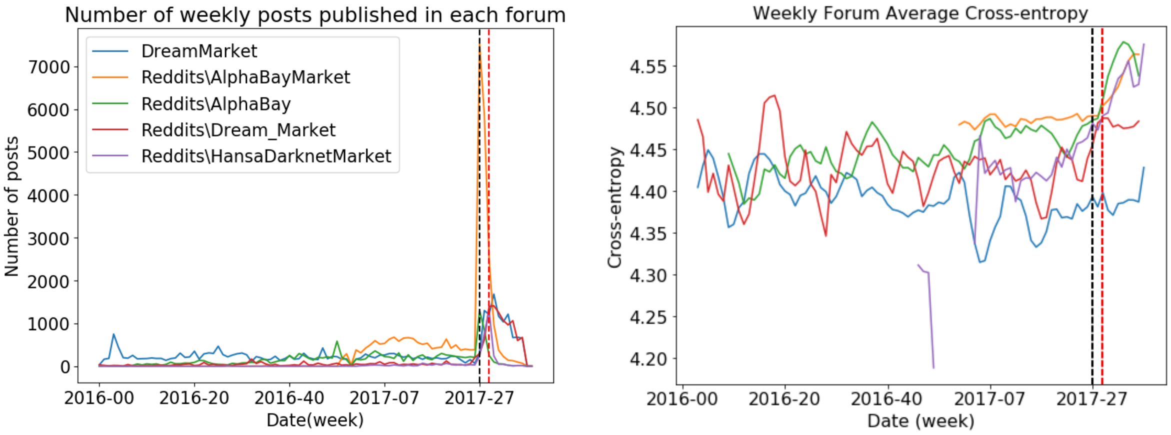
4.4.3 Marketplace Shutdowns. Our last case study regards the seizure of the AlphaBay marketplace by the FBI on July 4th 2017 and the seizure of the Hansa marketplace on July 20th 2017 by the Dutch NHTCU. We show cross-entropy, forum activity and transitions between states to analyze this case study.

Included in our d2web data are forums related to transactions and reviews of these marketplaces, including several private subreddits. We observed that a few of the forums in our dataset have peaks in activity around the same time. Figure 6(a) shows number of posts published in these forums. Forums related to AlphaBay and Hansa have peaks on the date of AlphaBay and Hansa closure respectively. Interestingly, a week after the Hansa closure Dream Market and the subreddits DreamMarket and DreamMarketDarknet had their highest value which suggests that users of these two big market places, AlphaBay and Hansa, have migrated to Dream Market. To check whether the forums respond to these events by changing the topics of discussion, we compute the cross-entropy. As seen in Figure 6(b) we see that two relevant forums, one about AlphaBay and one about DreamMarket, experience a change in their topic vectors after the shutdown (2017 week 27, or 07-04-2017). The cross-entropy increases around that time, suggesting growing difference from the aggregate topic distribution for their respective forums however changes in the volume of posts were more significant. We also observe state transitions in Dream Market forum in all three dates (AlphaBay and Hansa closure and the week after when there is a peak in activity for Dream Market related forums).

5 DISCUSSION & CONCLUSIONS

We used LDA to learn a rich set of latent components from a large corpus of documents spanning different topics and make use of a non-parametric HMM to better understand how those forums relate to one another in terms of the dynamics of their content. We then proceeded to use what we learn from both the states and the topics to identify specific posts discussing malicious activities that are understood to largely come from the dark and deep web.

This work can be extended in a number of different ways. First, one can use our framework to analyze new and unseen forums. Second, we can extend the framework to explicitly consider patterns of activity (e.g., frequency of posts) alongside the semantic information of the topic dynamics. Additionally, using average of the topic distributions of posts in a forum in one week to represent that forum will invariably smooth out events that have a small number of events associated with them. An alternate method that would be more sensitive to smaller fluctuations is to use the maximum value for a topic obtained from any post in that week.



(a)

(b)

Figure 6: (a) Activity of forums relevant to AlphaBay and Hansa closure. (b) Smoothed weekly topic cross-entropy of forums. Cross-entropy is smoothed using a rolling average over 4 weeks. For both figures the black line indicates July 4th, 2017 when AlphaBay was seized. The red line indicates July 20th, 2017 when Hansa was seized.

A promising extension to our system would be to use a dynamic topic model that incorporates a birth-death process for the population of topics, allowing for new topics to emerge and for old topics to die off. Online communities move fast, and a topic model that includes data from even a couple months ago can become obsolete and hinder productive analysis.

Acknowledgments. This work was supported by the Office of the Director of National Intelligence (ODNI) and the Intelligence Advanced Research Projects Activity (IARPA) via the Air Force Research Laboratory (AFRL) contract number FA8750-16-C- 0112, and the Defense Advanced Research Projects Agency (DARPA), contract number W911NF-17-C-0094. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of ODNI, IARPA, AFRL, DARPA, or the U.S. Government.

REFERENCES

[1] Khalid Al-Rowaily, Muhammad Abulaish, Nur Al-Hasan Haldar, and Majed Al-Rubaian. 2015. BiSAL–A bilingual sentiment analysis lexicon to analyze Dark Web forums for cyber security. *Digital Investigation* 14 (2015), 53–62.

[2] Mohammed Almukaynizi, Eric Nunes, Krishna Dharaiya, Manoj Senguttuvan, Jana Shakarian, and Paulo Shakarian. 2017. Proactive identification of exploits in the wild through vulnerability mentions online. In *Cyber Conflict (CyCon US), 2017 International Conference on*. IEEE, 82–88.

[3] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python.* O'Reilly Media.

[4] David M Blei and John D Lafferty. 2006. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*. ACM, 113–120.

[5] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research* 3, Jan (2003), 993–1022.

[6] Hsinchun Chen. 2008. Sentiment and affect analysis of dark web forums: Measuring radicalization on the internet. In *Intelligence and Security Informatics.*

[7] Hsinchun Chen, Wingyan Chung, Jialun Qin, Edna Reid, Marc Sageman, and Gabriel Weimann. 2008. Uncovering the dark Web: A case study of Jihad on the Web. *JASIST* 59, 8 (2008), 1347–1359.

[8] Ashok Deb, Kristina Lerman, and Emilio Ferrara. 2018. Predicting Cyber-Events by Leveraging Hacker Sentiment. *Information* 9, 11 (2018).

[9] Emily Fox, Michael I Jordan, Erik B Sudderth, and Alan S Willsky. 2009. Sharing features among dynamical systems with beta processes. In *NIPS.* 549–557.

[10] Emily B Fox, Michael C Hughes, Erik B Sudderth, Michael I Jordan, et al. 2014. Joint modeling of multiple time series via the beta process with application to motion capture segmentation. *The Annals of Applied Statistics* 8, 3 (2014).

[11] Robert W Gehl. 2016. Power/freedom on the dark web: A digital ethnography of the Dark Web Social Network. *new media & society* 18, 7 (2016), 1219–1235.

[12] Zoubin Ghahramani and Thomas L Griffiths. 2006. Infinite latent feature models and the Indian buffet process. In *NIPS.* 475–482.

[13] Palash Goyal, KSM Hossain, Ashok Deb, Nazgol Tavabi, Nathan Bartley, Andr'es Abeliuk, Emilio Ferrara, and Kristina Lerman. 2018. Discovering Signals from Web Sources to Predict Cyber Attacks. *arXiv preprint arXiv:1806.03342* (2018).

[14] Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. 2005. Integrating topics and syntax. In *NIPS.* 537–544.

[15] Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *Artificial intelligence and statistics.* 163–170.

[16] Nils Lid Hjort et al. 1990. Nonparametric Bayes estimators based on beta processes in models for life history data. *The Annals of Statistics* 18, 3 (1990), 1259–1294.

[17] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017).

[18] George Hurlburt. 2017. Shining Light on the Dark Web. *IEEE Computer* 50, 4 (2017), 100–105.

[19] Gastón L'huillier, Hector Alvarez, Sebastián A Ríos, and Felipe Aguilera. 2011. Topic-based social network analysis for virtual communities of interests in the dark web. *ACM SIGKDD Explorations Newsletter* 12, 2 (2011), 66–73.

[20] Erik Linstead, Cristina Lopes, and Pierre Baldi. 2008. An application of latent Dirichlet allocation to analyzing software evolution. In *Machine Learning and Applications, 2008. ICMLA'08. Seventh International Conference on.* IEEE, 813–818.

[21] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[22] Eric Nunes, Ahmad Diab, Andrew Gunn, Ericsson Marin, Vineet Mishra, Vivin Paliath, John Robertson, Jana Shakarian, Amanda Thart, and Paulo Shakarian. 2016. Darknet and deepnet mining for proactive cybersecurity threat intelligence. *arXiv preprint arXiv:1607.08583* (2016).

[23] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Workshop on New Challenges for NLP Frameworks.* 45–50.

[24] Sebastián A Ríos and Ricardo Muñoz. 2012. Dark Web portal overlapping community detection based on topic models. In *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics.*

[25] John Robertson, Ahmad Diab, Ericsson Marin, Eric Nunes, Vivin Paliath, Jana Shakarian, and Paulo Shakarian. 2017. *Darkweb Cyber Threat Intelligence Mining.*

[26] Jana Shakarian, Andrew T Gunn, and Paulo Shakarian. 2016. Exploring malicious hacker forums. In *Cyber Deception.* Springer, 259–282.

[27] Kyle Soska and Nicolas Christin. 2015. Measuring the Longitudinal Evolution of the Online Anonymous Marketplace Ecosystem. In *USENIX Security Symposium.*

[28] Nazgol Tavabi, Palash Goyal, Mohammed Almukaynizi, Paulo Shakarian, and Kristina Lerman. 2018. DarkEmbed: Exploit Prediction With Neural Language Models. In *AAAI.*

[29] Romain Thibaux and Michael I Jordan. 2007. Hierarchical beta processes and the Indian buffet process. In *Artificial Intelligence and Statistics.* 564–571.

[30] Seshadri Tirunillai and Gerard J Tellis. 2014. Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation. *Journal of Marketing Research* 51, 4 (2014), 463–479.

[31] Jennifer Xu and Hsinchun Chen. 2008. The topology of dark networks. *Commun. ACM* 51, 10 (2008), 58–65.