Towards a Comprehensive Insight into the Thematic Organization of the Tor Hidden Services

Martijn Spitters, Stefan Verbruggen, Mark van Staalduinen
TNO
Delft, The Netherlands
{martijn.spitters, stefan.verbruggen, mark.vanstaalduinen}@tno.nl

***Abstract*—Tor is a popular 'darknet', a network that aims to conceal its users' identities and online activities. Darknets are composed of host machines that cannot be accessed by conventional means, which is why the content they host is typically not indexed by traditional search engines like Google and Bing. On Tor, web content and other types of services can anonymously be made available as so-called hidden services. Obviously, where anonymity can be a vehicle for whistleblowers and political dissidents to exchange information, the reverse of the medal is that it also attracts malicious actors. In our research, we aim to develop a detailed understanding of what Tor is being used for. We applied classification and topic model-based text mining techniques to the content of over a thousand Tor hidden services in order to model their thematic organization and linguistic diversity. As far as we are aware, this paper presents the most comprehensive content-based analysis of Tor to date.**

I. Introduction

Anonymizing darknets, of which Tor is probably the most popular and well-known example, find increasing interest of users who for some reason wish to stay anonymous when online. Tor is a volunteer-based anonymity network consisting of over 3000 relay servers [1]. The network provides privacy to users accessing internet services, but it also provides a way to anonymously make TCP services available as so-called *hidden services*. Tor is based on the idea of *onion routing*. In a nutshell this means that the data which is sent over the network is first packed in multiple layers of encryption, which are peeled off one by one by each relay on the randomly selected route the package travels.

Even though historically the main goal of such networks was to enable freedom of speech, for instance in countries which employ political censorship, providing location and user anonymity also makes Tor an ideal platform for actors with malicious intentions. In fact, it is not a secret that the Tor network exhibits a wide range of illegal content and activity like drug and weapon trade, money laundering, hacking services, child pornography, and even assassination services. Tor harbors many shady underground marketplaces, for example Silk Road (shut down by the FBI in October 2013, but relaunched only a month later), Black Market Reloaded (shut down by its administrator in December 2013), Agora, and Pandora. Even though some of these services can quite easily be found through Tor signpost pages like the Hidden Wiki, or by using one of the few specialized search engines availabele on the Tor network (e.g. DuckDuckGo, Torsearch, Grams), these resources only cover a limited number of hidden services.

A comprehensive insight into the activities that take place in darknets like Tor may be very useful for many different parties. For instance, law enforcement may use it as a guide to aid forensic investigations or prosecutions, security agencies to detect security flaws and anticipate cyber attacks, and financial institutions to get information about laundering, counterfeit money and stolen account information or credit cards. In this paper, we present our research into mining and analyzing the web content of hidden services found on the Tor network. The main goals of this work are to develop a detailed understanding of what exactly Tor is being used for, to provide a representative and up-to-date model of its thematic organization and linguistic diversity, and to investigate the practical usefulness of such a model.

This paper is outlined as follows. First, we describe the most important related literature in Section II. In Section III we go into detail about the Tor data set we have harvested for the past year, including its linguistic diversity. Section IV is the core of this paper and describes the topic-model based text mining technique we applied to our Tor data, as well as the resulting topic taxonomy. Finally, Section V lists our conclusions.

II. Related work

To date, little research has been published aimed at content analysis of the Tor network, or more broadly speaking, the Dark Web. Biryukov et al. [2] applied language and topic classification to a collection of hidden service data which was mined by exploiting a flaw in the Tor protocol. However, their topic set is relatively broad and it remains unclear how exactly it was composed and on what kind of data the classifier was trained. Other recent studies on how Tor is used are based on monitoring network traffic rather than analyzing the published content [3], [4], [1], or perform an in-depth analysis of a single underground marketplace, but more from a financial-economic perspective [5], [6].

Topic-driven analysis of Dark Web forums has been published in [7], [8]. This work combines topic modeling and social network analysis techniques for identifying key members of Dark Web communities which are specifically talking about certain topics.

Also related to our work, but with a strong focus on extremist/terrorist-generated content, mostly related to Islamic ideology and theology, is the multi-faceted Dark Web research of Chen et al. [9], mainly applied to terrorist groups' websites found on the Dark Web.

978-1-4799-6364-5/14 $31.00 © 2014 IEEE
DOI 10.1109/JISIC.2014.40

TABLE I. STATISTICS OF OUR COLLECTION OF TOR HIDDEN SERVICE CONTENT PAGES IN DECEMBER 2013.

Number of known hidden services	5725
Number of hidden services with downloaded content	1481
Number of known pages (urls)	11.557.357
Number of downloaded pages (urls)	2.196.410
Number of pages with usable content after preprocessing	324.623

III. DESCRIPTION OF THE DATA

We crawled our data set by using the Tor stand-alone client, which functions as a SOCKS-proxy. When software is configured to connect to the internet through this SOCKS-proxy, connections will automatically be routed over the Tor network, providing anonymity and the option to visit hidden services. Details about the inner workings of the Tor protocol when visiting hidden services can be found at [10].

Our crawler only downloads html pages; urls pointing to media content or any other file format than text are ignored. We configured our crawler to harvest data in a breadth-first fashion, considering the number of pages already downloaded from a host, whether or not a host has previously yielded new onion addresses, etc. This way we collect *some* content from as many hidden services as possible, rather than *all* content from only a few hidden services. Table I shows the details about the data set we used for the experiments described in this paper (December 2013). In the past five months, the crawler has discovered almost 2000 new hidden service addresses.

In a preprocessing step we filtered out pages without usable content (e.g. html-embedded error messages, ssh banners, redirection pages), and extracted the text from the html. Very short texts (less than 15 words) were removed as well.

A. Linguistic diversity

In order to get an insight into the linguistic diversity on the Tor network, we applied a language classifier to the downloaded content pages. Our classifier is able to identify 30 languages based on a weighted combination of features (common words, prefixes, suffixes, and character n-grams). Even though a particular page can contain multilingual content, we found this to be rare and mostly occurring on pages with very limited informational content. We therefore assigned a single language to each page.

On 83.27% of the classified content pages, the main written language is English. This number is in line with the one reported in [2]. Figure 1 shows the distribution of all other languages than English, except those that were assigned to less than 0.1% of the pages (Albanian, Slovene, Hungarian, Czech, Greek, Croatian). On 0.17% of the pages, our classifier detected another language than the ones it can identify. On closer inspection, we found that most of these pages were in Chinese and Japanese.

It is interesting to see that the percentage of English content in our data is very similar to that of the surface web during the end of the nineties, at the start of the world-wide popularization of the internet [11]. Besides the fact that the ratio of English to other languages has become more balanced over the last decade (55.7% to 45.3% according to [12]), the distribution in Figure 1 shows a rough correspondence to the overall language use on the surface web (see [12] for a table with

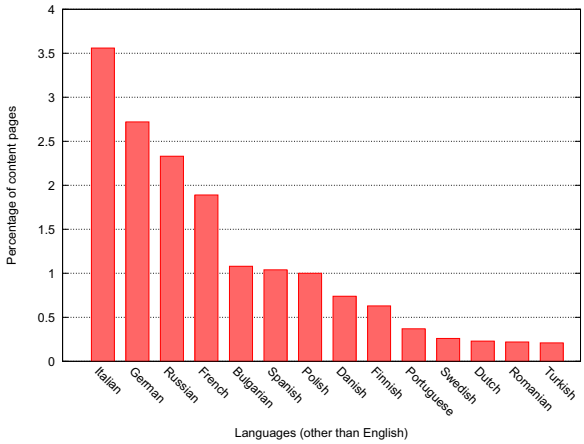

Fig. 1. Percentages of languages other than English, computed on our collection of Tor hidden service content pages. On 83.27% of the analyzed content pages, the main written language was English. Languages with a share less than 0.1% are not shown.

recent estimates). The most striking anomalies are Italian and Bulgarian, which seem to have a much higher relative share on Tor than on the surface web. For Chinese and Japanese the opposite seems to be the case.

B. Uptime analysis

The availability and life span of Tor hidden services can be very unpredictable. Even though our system has identified more than 7000 Tor hidden services (i.e. onion addresses), the lion's share of them is unreachable most of the time. The highest number of concurrently online hidden services that we measured was around 1450. At the start of August 2013 we observed a sudden dramatic fall from 800 to under 400 online hosts, which can be explained by the FBI shutting down Freedom Hosting – Tor's most popular hosting service at that time –, after having arrested its owner and maintainer for facilitating the spread of child pornography. Presumably, during that action the FBI compromised almost half of the sites in the Tor network. Since then, according to measurements on our data the level has recovered more or less linearly to almost 1450 concurrently online services in May 2014.

IV. TOPIC MODELING

We applied topic modeling techniques to discover the topical structure of the hidden service web content in our data set. Topic models represent a family of text mining algorithms that discover topics from document collections without requiring any prior annotations or labeling. The intuition behind topic modeling is that documents typically blend multiple topics. Topic modeling algorithms try to capture this intuition by using term co-occurrence statistics within a document collection to reveal its latent semantic structure. We used a recent and widely used topic modeling technique called Latent Dirichlet Allocation (LDA), created by Blei et al. [13].

A. Latent Dirichlet allocation

LDA is a generative probabilistic model for discrete data, such as our collection of content pages. The intuition behind LDA

TABLE II. THE TOP WORDS FOR SOME SAMPLED TOPICS OF OUR 250-TOPIC LDA MODEL. THE TOPIC LABELS WERE MANUALLY ADDED.

Topic label	Most likely topic words
Trading	product quality packaging order good bitcoin btc buy account money usd payment wallet cash paypal coins currency price cart worldwide
Drugs	mdma cocaine quality lsd pills coke speed ketamine netherlands crystal usa methylone high drugs tested meth gram xtc lab heroin results
Anarchism	anarchist social anarchism power movement class workers revolutionary struggle political revolution anti capitalism left groups society action
Hacking	hack ddos hacker channel irc rat bot anonymous attack anon exploit virus website botnet high remote zeus lulzsec quality hackerspace project
Child Pornography	video boy vids download child sexual kids age women pedo adult young girl abuse nude pedophilia jailbait pics society consent pedoempire
Weapons	9mm rifles guns price automatic ammo glock cart payment big 22lr smoke handguns 45acp shotguns special pistols gauge 380acp redstar ball

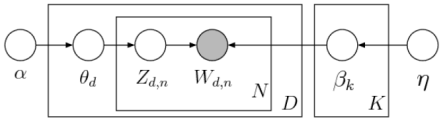

Fig. 2. The probabilistic graphical model for LDA, where K is the number of topics, the D plate denotes the collection of documents in the corpus, and the N plate denotes the collection of words within documents. The shaded node represents the observed variables, i.e. the words of the documents.

TABLE III. EXAMPLES OF TOR HIDDEN SERVICES IN OUR DATABASE AND THEIR TOPICS.

Title	Topics
The Hidden Wiki	Trading, Counterfeit, Child Pornography, Technology, Security, Services, Pornography, Software, Anonymity
Silk Road market place	Trading, Drugs, BTC, Security
Code:Green - hacktivism	Hacking, Security, Software, Anonymity
Pedo Support Community	Child pornography, Communities
HackBB forum	Credit Cards, Trading, Hacking, Paypal, Financial, Bank Accounts, Exploits, Security
Fight back the state	Anarchism, Politics, Law Enforcement
EuroGuns	Weapons, Trading, BTC
Rent-A-Hacker	Hacking, DDoS, Exploits, Services, BTC, Trading

is that documents are mixtures of corpus-wide topics, where each topic is a distribution over words. The topic mixture is drawn from a Dirichlet distribution. Figure 2 shows a graphical model, describing the probabilistic assumptions behind LDA. The generative process of LDA defines a joint probability distribution over observed and hidden random variables. In Figure 2 the nodes are random variables, and edges indicate dependence. A plate means replication. The shaded node represents the observed variables in the model, which are the words of the documents ($w_{d,n}$ is the nth word in document d). The unshaded nodes are hidden: $\beta_{1:K}$ are the topics, θ_d is the topic proportion for document d, and the topic assignments for document d are z_d ($z_{d,n}$ is the topic assignment for the nth word in document d). The goal in LDA is to *infer* the hidden variables – the hidden topic structure of the collection –, i.e. compute their distribution conditioned on the documents: $p(\beta_{1:K}, \theta_{1:D}, z_{1:D}|w_{1:D})$. This posterior can be approximated by using a sampling-based or variational inference algorithm. For an elaborate mathematical description of LDA, including methods for parameter estimation, we refer to [13].

B. Generating a topic taxonomy

The generation process of our Tor topic taxonomy consists of three main steps: (1) model estimation, (2) topic annotation, and (3) topic aggregation to the hidden service level, each of which will be described in the continuation of this section.

1) Model estimation: For modeling our collection of Tor hidden service pages, we used the Mallet implementation of LDA, which uses Gibbs sampling as a posterior approximation technique [14]. In these experiments, we only included English content. After preprocessing and language classification, we had 1021 hidden services with usable English content left. From the raw texts, stop words were removed, and the remaining words were morphologically normalized using a lemmatizer. Because hidden services come in many flavours (wikis, forums, online shops, image libraries, simple web pages, etc.), the number of pages with useful content can vary greatly per hidden service, from just a few pages to many thousands. We estimated various topic models on this set based on different parameter combinations (number of output topics, number of sampling iterations, hyperparameter optimization), and found the 250-topic model to contain the most intuitive and clearly separated topic word lists. Table II shows the most likely words for some of the topics in this model. As we had hoped, these distributions capture some of the underlying topics in the corpus very well. The topic labels in the first column were added manually.

2) Manual topic annotation: We labeled each of the 250 inferred topics in the chosen LDA model by hand. In case the word list clearly represented more than a single topic, we assigned multiple labels (e.g. *DDoS* AND *exploits*; *technology* AND *intelligence*, *trading* AND *weapons*). Topics which we could not relate to any intuitive, coherent theme were labeled 'Unknown'. The result of this manual annotation step is two-fold: first of all it gives us a labeled topic model which can be used for classifying unseen pages, and second, it provides us with a list of topic labels that can serve as the basis for our Tor topic taxonomy.

3) Topic aggregation: In this step we use the topic proportions for each document (i.e. page) in the training collection (which are stored by Mallet during the LDA estimation process) to aggregate the topic annotations to the hidden service level. We apply this step because we consider the hidden service as a whole to be a more logical thematic entity than an individual page. The aggregation step will also correct a bias to the topics of very large hidden services in our data, yielding a more representative global thematic profile of the Tor network. We simply compute the average of each topic's proportions over all pages of a certain hidden service. We keep all topics that exceed a certain threshold, which was empirically measured on a random sample. This threshold is dependent on the type of hidden service, which our system automatically detects – for forums and wiki pages, which usually discuss many different topics, a lower threshold yielded more accurate topic assignments. As an illustration, Table III shows some examples of hidden services from our data set and their aggregated topic assignments.

Given these hidden service topic classifications, we can construct a global topic taxonomy. Figure 3 visualizes the resulting taxonomy as a graph, in which the size of a node expresses the degree of presence over all hidden services in

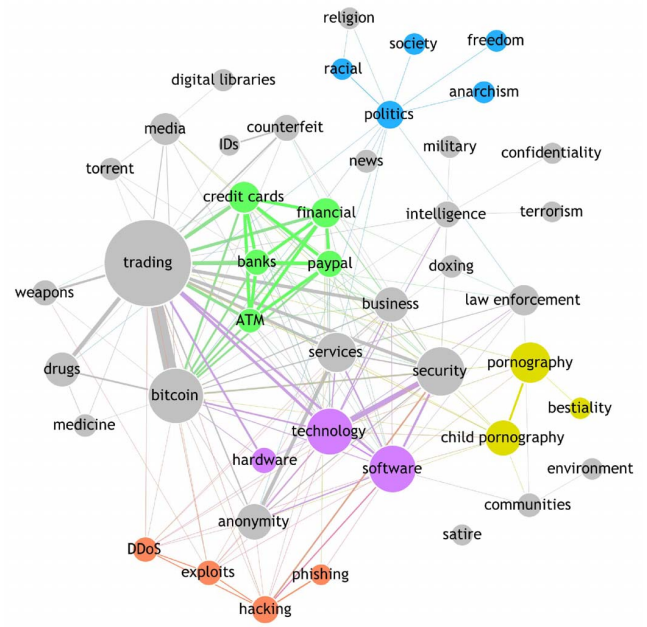

Fig. 3. Topic taxonomy, visualized as a graph, in which the size of a node expresses the degree of presence of the topic on the hidden services, and the weight of an edge expresses the relatedness of the nodes it connects.

our data collection, i.e. the percentage of hidden services for which the score of that topic exceeds the assignment threshold. The weights of the edges is based on the number of times the topics they connect were found together on the same hidden service, and therefore in a way expresses their 'relatedness'. Nodes with the same color form a topical cluster (green is *financial*, purple is *technology*, orange is *hacking*, yellow is *pornography*, blue is *politics*). Note that each of these clusters contains a node with this name, which should be interpreted as the 'other' category within that cluster. The correspondence between node size and actual percentage is as follows: *Trading*: 59%; *Software, Security*: 20-25%; *Child pornography, Drugs*: 10-15%; *Weapons, Anarchy, Doxing*: 2-5%.

The taxonomy in Figure 3 indicates that most Tor hidden services in our data set exhibit illegal or at least controversial content. Trading activity was detected on almost 60% of the hidden services, mostly devoted to drugs, weapons, counterfeit money/documents, stolen credit cards, and hacked accounts. Adult content (or pointers to such) was detected on 17% of the hosts in our collection. About half of it was classified as child pornography. We also found hidden services with more noble intentions, discussing political oppression, freedom, and anonymity. However, because topics cannot easily be mapped to intentions (think of disclosing confidential documents, anarchic communities, doxing), it is very hard to unambiguously divide our Tor data into a 'good' and a 'bad' part.

C. Classification of unseen hidden services

Given our trained topic model, we can automatically assign topics to new, unseen hidden services. The topic model can be used to infer topics on the page level, which we simply aggregate to the hidden service level using the procedure described in the previous section. We are currently in the process of testing and optimizing the topic classifier on new hidden services found since January 2014. A first evaluation on a test sample of 50 new annotated hidden services on the top 3 assigned topics shows an encouraging accuracy of 92%.

V. Conclusions and Future Steps

In this paper we presented our work on content-based mining and analysis of hidden services found on the Tor network. On a collection of over a thousand hidden services, harvested with our Tor crawler, we have successfully applied classification and topic-model based text mining techniques in order to reveal their linguistic diversity and thematic organization. We have inferred a topic taxonomy, which indicates that most hidden services – at least in our data set – exhibit illegal or controversial content.

On the short term, we plan to conduct a thorough evaluation of our model's classification performance on unseen hidden services. Furthermore, we plan on integrating a translation module, so non-English content can be included in the modeling process as well. As a next step, we will also zoom in to a more detailed topical level.

References

[1] A. Biryukov, I. Pustogarov, and R.-P. Weinmann, "Trawling for tor hidden services: Detection, measurement, deanonymization," in *Security and Privacy (SP), 2013 IEEE Symposium on*. IEEE, 2013, pp. 80–94.

[2] ——, "Content and popularity analysis of tor hidden services," *arXiv preprint arXiv:1308.6768*, 2013.

[3] A. Chaabane, P. Manils, and M. A. Kaafar, "Digging into anonymous traffic: A deep analysis of the tor anonymizing network," in *Network and System Security (NSS), 2010 4th International Conference on*. IEEE, 2010, pp. 167–174.

[4] D. McCoy, K. Bauer, D. Grunwald, T. Kohno, and D. Sicker, "Shining light in dark places: Understanding the tor network," in *Privacy Enhancing Technologies*. Springer, 2008, pp. 63–76.

[5] N. Christin, "Traveling the silk road: A measurement analysis of a large anonymous online marketplace," in *Proceedings of the 22nd international conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2013, pp. 213–224.

[6] J. Franklin, A. Perrig, V. Paxson, and S. Savage, "An inquiry into the nature and causes of the wealth of internet miscreants." in *ACM conference on Computer and communications security*, 2007, pp. 375–388.

[7] G. L'Huillier, S. A. Ríos, H. Alvarez, and F. Aguilera, "Topic-based social network analysis for virtual communities of interests in the dark web," in *ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 2010, p. 9.

[8] S. A. Ríos and R. Muñoz, "Dark web portal overlapping community detection based on topic models," in *Proceedings of the ACM SIGKDD Workshop on Intelligence and Security Informatics*. ACM, 2012, p. 2.

[9] J. Xu, H. Chen, Y. Zhou, and J. Qin, "On the topology of the dark web of terrorist groups," in *Intelligence and Security Informatics*. Springer, 2006, pp. 367–376.

[10] The tor hidden service protocol. [Online]. Available: https://www.torproject.org/docs/hidden-services.html.en

[11] P. Gerrand, "Estimating linguistic diversity on the internet: A taxonomy to avoid pitfalls and paradoxes," *Journal of Computer-Mediated Communication*, vol. 12, no. 4, pp. 1298–1321, 2007.

[12] W3Techs. Usage of content languages for websites. [Online]. Available: http://w3techs.com/technologies/overview/content_language/all

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.

[14] A. K. McCallum, "Mallet: A machine learning for language toolkit," 2002.