Who are in the Darknet? Measurement and Analysis of Darknet Person Attributes

Meiqi Wang*†‡, Xuebin Wang*†‡, Jinqiao Shi*†, Qingfeng Tan§, Yue Gao*†‡, Muqian Chen*†‡ and Xiaoming Jiang*
* Institute of Information Engineering Chinese Academy of Sciences, Beijing,China
† National Engineering Laboratory for Information Security Technologies, Beijing, China
‡ School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
§ Cyberspace Institute of Advanced Technology, GuangZhou University, GuangZhou, China

***Abstract*—The high anonymity of Darknet makes it attractive to users who want to avoid Internet censorship and surveillance. As a result, in recent years, Darknet is abused for various illegal purposes. Undoubtedly, measurement and analysis towards the attributes of people in the Darknet can obtain a comprehensive characterization of dangerous users and help trace malicious users, reducing cybercrimes. However, it is still challenging to extract person attributes in Darknet scenario due to its anonymity and content sparsity. Therefore, in this paper, we propose a new person attribute extraction method consisting of three steps: block filtration, attribute candidate generation and attribute candidate verification. Experiments show that our extraction method performs better than traditional extraction methods. Using the extracted information as input, we measure and analyze the number of attributes, Top-K name entities, email domain name, etc. of people in Darknet, revealing the characteristics of the person attributes in the dark web pages.**

I. Introduction

Darknet provides users with anonymity, so it carries a large amount of illegal content such as contraband trade, terrorism, child pornography etc., posing a serious threat to cyberspace and national security. For example, towards recent reports [1], the Silk Road marketplace is known to the general public for trade of drugs, arms and other kinds of goods, raising increasing concerns of cyber crimes. Although it is known that Darknet contains person information, the distribution, categories and content of the attribute information of the darkweb pages are still unknown, and further detection and analysis are needed. The measurement and analysis towards the attributes of people in the Darknet can quickly and accurately obtain a comprehensive characterization of dangerous users, support for the traceability of the users, and help reduce cyber crimes and maintain national security.

On the one hand, because of the anonymity of the Tor protocol, it is very difficult to crack the anonymity of Tor from the protocol level; On the other hand, some users engage in illegal activities on the Darknet, meanwhile, they will leave their own contact information such as e-mail, qq number, etc., which provides the possibility to extract and further analyze the person attribute information on the Darknet. However, person information extraction is still challenging in Darknet scenario. Current available person information extraction methods can be divided into two categories according to the application scenario: 1) situations where the content of the person information is substantial [2]–[4] and 2) scenes where the information content of the characters is indefinite and nonuniform [5]–[11]. The attributes of some pages are sparse, while for some pages the attributes are rich. However, in the Darknet, due to privacy, security, and other issues, the person properties of most pages are seriously missing, which means the methods in the above two scenarios are not applicable. At present, there is a lack of relevant research work in the scenario where the information content of people is sparse. In order to extract the Darknet person information, there is a need for a new person attribute information extraction technique for scenes where attributes are sparse on the page. In the domestic and foreign researches, the measurement and analysis of Darknet content are more and more comprehensive, which reveals the current development, status, and threats of Darknet. However, the measurement and analysis of the person attributes of Darknet are still not comprehensive and systematic enough to reveal the distribution, density, behavior habits, and network traces of users in Darknet. As a result, it is necessary to systematically and comprehensively measure and analyze person attributes in Darknet.

We, therefore, propose a method that adapt to the Darknet scenario consisting of three steps: block filtration, attribute candidate generation and attribute candidate verification. For a raw web page, the text block is cleaned firstly so that the irrelevant content can be filtered out. Then all attribute candidate sets are extracted for every attribute class. Finally, the most likely attribute of the target person in every candidate set is selected through classification. Experiments show that our extraction method performs better than the basic methods proposed in WePS [12], [13], a character retrieval competition organized by Javier Artiles et al. After the person

Corresponding author: Xuebin Wang, Email:wangxuebin@iie.ac.cn

978-1-5386-4210-8/18/$31.00 ©2018 IEEE
DOI 10.1109/DSC.2018.00151

information extracted, we make a deep analysis in order to obtain a comprehensive characterization of Darknet users. The experiments have been conducted on WEPS2 data sets and real data collected from Tor [14], one of the most famous Darknet. The major contributiions of our work are:

1) A method of person attribute information extraction for Darknet is proposed. For the sparseness and lack of the attribute of the dark web page, the existing character information extraction technology is not adaptable. Our method performs higher precision, recall, and F1 score in scenes with missing character attributes comparing to basic methods.
2) Measurement and analysis of the Darknet person attributes. We measure and analyze from the dark web page the number of attributes, Top-K name entities, email domain name, etc., revealing the characteristics of the person attributes in the dark web page, facilitating better understanding of Darknet person information, and helping trace the source of dark network users.

In the rest of paper, we introduce related work in Section II. Our extraction method is proposed in Section III with experiments and discussions. Then we analyse the extracted person information in Section IV. Finally, Section V includes our conclusions and future works.

II. Related Work

There have been extensive and ongoing works both on person profiles extracting and darknets measuring. In this section, we will summarise the previous researches.

A. Extraction of Person Information

At present, the existing person attributes extraction methods can be divided into two categories according to their application scenario: one is applied to scenes where the personal information is enriched, and the other is applied to scenes that contain indefinite or nonuniform personal information.

The homepage of a scholar is one of the typical examples of scenes with enriched personal information. In 2008, Tang et al. put forward a method of extracting researcher profiles automatically from the Internet [3]. Firstly, they use the researchers' names as search keywords to get a batch of web pages by Google API. To identify whether the page is a scholar's personal homepage or not, a binary SVM classifier is trained. When preprocessing, the text is divided into tokens, and each token is given a potential label which is associated with a attribute. Finally, each label is validated through the CRF model and a unique sequence of labels is output. Experimental results show that the average F1 score of this method is as high as 83.37%.

In more scenes, due to privacy, security and other issues, the personal information in the web pages is nonuniform. In the WePS [12], [13] character retrieval competition organized by Javier Artiles et al, multiple teams [5]–[11] proposed their methods. Among them, Chen et al. [5] proposed a rule-based AE system that attempts to capture typical patterns in web pages. This method breaks the limitation that the pattern can only be learned from a single sentence. Lan et al. [7] used preprocessing methods such as web content cleaning and structure cleaning. They first thoroughly cleaned the web pages and then tried to use named entity recognition, regular expression matching and other methods for information extraction. Watanabe et al. [10] proposed a two-step method. The first step is to use named entity recognition, regular expression matching, dictionaries and many other ways to extract candidate sets of attributes. And the second step determines which attribute values of candidates belong to the retrieved person. In this competition, the personal information in web pages is far less than that in the scholars' homepages, so the difficulty is much greater than the former. The results show that the average F1 score of multiple attributes is 12.2%.

B. Measurement Towards Darknet

With its increasing popularity, there are more and more researches on the measurement or analysis of Darknet.

In 2014, Biryukov et al. analyzed the content and popularity of Tor hidden services in [15]. They collected 39,824 .onion addresses and 6,579 of whom used either HTTP or HTTPS. Measured by the number of client requests, their results showed that the most popular hidden services are the command and control centers of botnets. In 2013, Christin N et al. [16] performed an in-depth measurement and analysis of Silk Road, which runs as a Tor hidden service and uses Bitcoin as its foreign currency. They collected and analyzed 8-month data from the end of 2011 to learn the categories of merchandise sold on the Silk Road, and the benefits of sellers and operators. In 2016, by acquiring the GPS location information in the pictures of merchandise, two students at Harvard University, Paul Lisker and Michael Rose, obtained a physical map of some of the global darknet businesses[1].

While prior person attributes extraction technologies may be applied to specific scenes, more work should be done towards extraction technology for scenes where sparse information are contained in web pages; As for Darknet measurement, there is a lack of measurement towards people on Darknet.

III. Person Attribute Extraction

Person Attribute Extraction (AE) gives us access to the people information in the dark web for further analysis, extracting the attributes from unstructured text in the dark web pages. Generally speaking, the AE task makes use of traditional Information Extraction (IE) and Named Entities Recognition (NER) approaches, and goes beyond them. That is, it not only extracts people potential attributes on raw web pages, but also determines whether the extracted information belongs to the target person. In this section we begin with a basic AE method, and then introduce our proposed method.

A. Basic Method

Basic AE method is illustrated in Fig.1 and is mainly composed of two steps: the attribute candidate generation

[1] http://toutiao.secjia.com/harvard-students-show-darkweb-merchant-physical-map

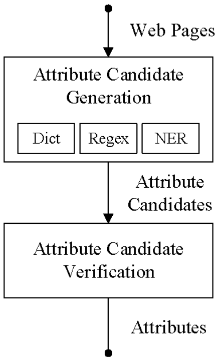

Web Pages
Attribute Candidate Generation
Dict
Regex
NER
Attribute Candidates
Attribute Candidate Verification
Attributes

Fig. 1: An Abstract Diagram Presenting the Basic Attribute Extraction Method

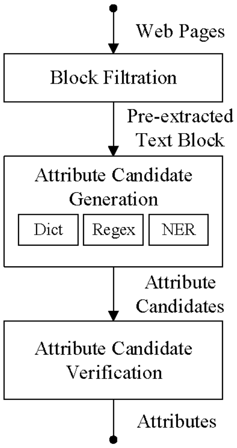


Fig. 2: An Abstract Diagram Presenting Our Proposed Attribute Extraction Method

and the attribute candidate verification. The former refers to extracting a plurality of possible attribute values from the web text for the target attribute, while the latter refers to selecting, from the candidate sets, attribute values that are considered most likely to belong to the target person, based on a model or a text distance or the like. For example, in the attribute candidate generation step, for the nationality attribute, China, US, and Japan may be extracted from the page as candidate sets. And in the attribute candidate verification step, for a candidate set as above, the classifier may select US as the most likely nationality attribute value for the target person.

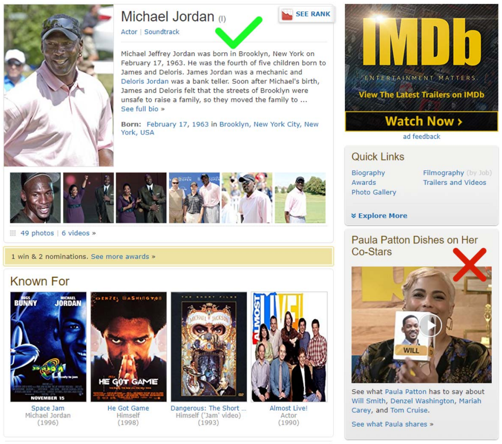
The common person AE method described above can be used directly in Darknet scenarios. However, due to the inherent privacy and concealment of the Darknet, the attribute information of the dark web pages is very sparse and the interference factors are pretty large. As a result, in Darknet scenarios, using the basic AE method and taking the entire page text as a pre-extracted text block will undoubtedly cause great noise. Therefore, we propose our AE method adapted to Darknet scenarios as below.

B. Proposed Method

A graphical diagram presenting our method is shown in Fig.2. Our proposed method consists of three steps: block filtration, attribute candidate generation and attribute candidate verification. For a raw web page, the text block is cleaned firstly so that the irrelevant content can be filtered out. Then all attribute candidate sets are extracted for every attribute class. Finally, the most likely attribute of the target person in every candidate set is selected through classification. A more detailed discussion for each step will be presented in the following subsections.

1) Block Filtration: According to [7], unlike conventional data or plain text, web pages have a lot of information that is not relevant to the main contents of the pages, including: (1) rich resource of formats and functional codes, e.g., HTML tags, script codes, CSS, etc, (2) irrelevant and noisy information, e.g., contextual Ads, navigation banner, Rich Media Ads, copyright notices, and even fraud anchor words with links, etc, (3) confusing information, e.g., in most cases, even though the web pages contain many mentions of people attributes, they are not relevant to the target person, for example, the email addresses of the web masters, friends, colleagues, or even other person who make comments in this web page, etc. This situation is particularly serious in Darknet scenarios. Undoubtedly, all such irrelevant information in web pages, i.e. web page noise, can severely harm the AE performance. Therefore, above all, we should use block filtration to decide which content of the page is meaningful and relevant to the target person and filter out the noisy part of the web page.

It is not difficult to find through observation that most of the information belonging to the target person is near the person's keyword (the person's name/username in most case). As shown in Fig.3, Michael Jordan's birthplace, date of birth, and other information are all located below the person's name. The distracter is located on the right side of the person's name. After stripping away all HTML tags, the distracter is farther away from the person's name. Therefore, the core idea of the text block filtration strategy is to select the sentence near the keyword as pre-extracted text.



Michael Jeffrey Jordan was born in Brooklyn, New York on February 17, 1963. He was the fourth of five children born to James and Deloris. James Jordan was a mechanic and Deloris Jordan was a bank teller. Soon after Michael's birth, James and Deloris felt that the streets of Brooklyn were unsafe to raise a family, so they moved the family to ...
See full bio »

Born: February 17, 1963 in Brooklyn, New York City, New York, USA

1 win & 2 nominations. See more awards »

See what Paula Patton has to say about Will Smith, Denzel Washington, Mariah Carey, and Tom Cruise.
See what Paula shares »

Fig. 3: An Example of Block Filtration

As the information content of the person attributes of different pages is various, for different pages, the number of nearby sentences selected should vary with the content of the page's attribute information. The richer the attribute information of the page is, the more sentences should be selected nearby. Therefore, here we need to quantify the richness of the page's attribute information. We use a value between [0,1] to represent the richness of the page's attribute information. The larger the value, the richer the information content. Then we try 5, 7, 10 as the maximum sentences extracted and find little difference through experiments, so the middle value 7 is selected as the upper limit. That is, we extract

TABLE I: The Weight of Different Metrics

Metrics	Maximum attribute value	Weights
name	5	0.05
location	5	0.05
organization	5	0.05
time	3	0.05
nation	3	0.05
word number	3800	0.05
keywords	0	0.55
email	10	0.1
qq	10	0.1
phone	10	0.1
bitcoin	10	0.1
skype	10	0.1
wechat	10	0.1

TABLE II: Definition of Extracted Person Attributes

Attribute class	Examples of attribute value
realname	Alice
birthday	1st May 1900
birthplace	London
occupation	Teacher
location	California
organization	XXX Foundation
nationality	UK
email	xxx@yyy.com
qq	1111111111
phone	11111111111
school	University of New York
wechat	1111111

at least the sentence where the keyword is located, up to seven sentences near the keyword. The number of sentences to be extracted is the rounded up result of the product of 7 and the quantized value.

The quantization method is described in detail as below. We give different weights to different metrics. The metrics are divided into the following three categories:

1. Common attributes: such as names, geographic locations, etc.

2. Strong attributes: such as mail address, phone number, etc.

3. Other metrics: such as the number of words on the page, the number of keywords, etc.

The weights of different metrics are shown in Table I. The maximum attribute value in the table indicates that when the attribute value reaches the upper limit, corresponding weight values can be obtained. For example, if the location attribute value is greater than or equal to 5, then the weight is increased by 0.05; if the location attribute value is 1, the weight is added by 0.05*1/5. The metric ‘keywords’ indicates the number of words corresponding to the attributes, e.g. for the word birthplace, the keywords are birthday, birth, born, etc., which are relevant to the word birthplace.

Through the above method, after obtaining the plain text for each web page by removing HTML tags, the quantized value can be calculated. In order to verify the rationality of the quantification method, we randomly selected 2,397 pages, and counted the number of attribute types for each page and calculated the quantified values. In general, the larger the number of attribute types the page contains, the larger the quantized value is. That is, there should be a linear relationship between the two factors. As shown in Fig. 4, the experimental results show that the quantification value increases with the number of attribute types, and they have a linear relationship. Therefore, the quantization method proposed in this paper has certain rationality.

2) Attribute Candidate Generation: Attribute candidate generation refers to extracting possible attribute values of all attributes from the retrieved various pages utilising information extraction technology. For a given person attribute, its attribute value is usually a specific form of a noun. For example, for a mail address, the attribute value must match the format of the mail address; for the birthplace, the attribute value must conform to the geographic format. According to the different categories of attributes to be extracted, Han X et al. [6] divided the categories of person attributes into three categories, as shown in Table III.

In Table III, the first attribute category is a traditional named entity (person name, etc.). The existing named entity recognition tools can obtain acceptable results. Therefore, we do not do independent research on the NER method, and directly adopt the Stanford University named entity recognition tool[2] which is recognized by the industry. We use the trained model given by the Stanford team directly, which formed as: Time, Location, Organization, Person, Money, Percent, Date.

The second attribute category is special types of entities such as mail addresses, telephone numbers, website addresses, etc. There are usually certain rules for such entities. In our experiments, different regular expressions are constructed for mail addresses, phone numbers, and so on. The results show that regular expressions can accurately extract candidate person attribute values for this type of entity.

The third category is special noun phrases such as positions and degrees that can be exhaustive noun phrases. For the entities of this class, because of their exhaustiveness, a dictionary of each attribute is constructed here via Wikipedia[3].

In general, multiple candidate attribute values of pre-extracted attributes can be obtained accurately by the above three methods. The attributes extracted by us are shown in Table II. However, usually only one of the multiple candidate attribute values belongs to the target person. If multiple candidate attribute values are simultaneously used as the target person’s attribute value, or are only randomly selected or proximately selected, it will be hard to measure and analyse the person attributes accurately. Therefore, on the basis of the candidate set generation, it is necessary to further verify the candidate set and select the attribute value that is most likely to be the search person. Therefore, in the following sections, the method for verifying candidate sets will be described in detail.

[2] http://nlp.stanford.edu/software/CRF-NER.shtml

[3] https://zh.wikipedia.org

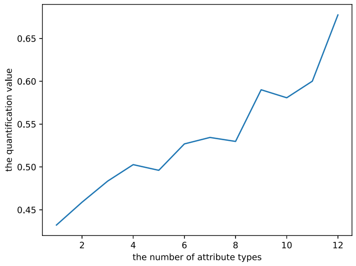
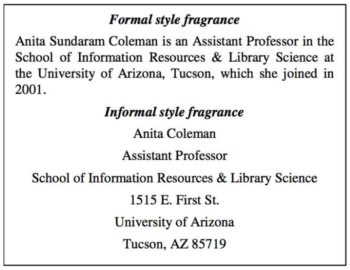


Fig. 4: The Linear Relationship Between the Number of Attribute Types and the Quantification Value



Formal style fragrance

Anita Sundaram Coleman is an Assistant Professor in the School of Information Resources & Library Science at the University of Arizona, Tucson, which she joined in 2001.

Informal style fragrance

Anita Coleman
Assistant Professor
School of Information Resources & Library Science
1515 E. First St.
University of Arizona
Tucson, AZ 85719

Fig. 5: Formal and Informal Text Style

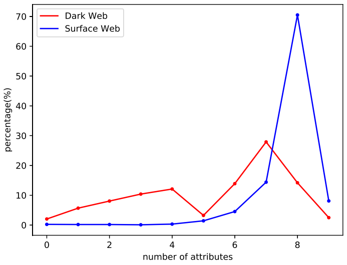


Fig. 6: The Distribution of the Number of Attributes in the Surface Web and Dark Web

TABLE III: Extraction Methods for Different Categories of Person Attributes

Categories	Person attributes	Extraction method
traditional named entity	birthday, birthplace, organization, school, name	NER tools
special types of entities	email address, phone number, website address, fax	rule-based extraction methods
special noun phrases	position, degree, major, profession	dictionary

TABLE IV: The Statistics of the Data Set

Number of documents in the test data	3468
Number of documents ignored	585
Number of documents used for the test	3468
Number of documents with at least one attribute value	2421
Number of documents with no attribute value	462

3) Attribute Candidate Verification: Candidate sets are verified using a binary classifier. The training set is used to obtain a trained model, and then a classifier for each attribute is used to select the most probable attribute values from a plurality of candidate sets. The data used in this paper is from WePS2[4]. The test data included a corresponding data set of 30 person names coming from three different sources: Wikipedia, ACL'08, and the U.S. census. For each name, WePS2 obtained the top 150 search results from the Internet search engine (Yahoo! API). At the same time, they provided training data consisting of 17 names from WePS1. Among them, the person attributes data is marked by 4 independent workers. The statistics of the data set are shown in Table IV.

The candidate set is validated using the SVM model, and a total of 30 features are extracted from the candidate based on the candidate set's context, extracted sentences attribute, parts of speech and so on. The features are as shown in Table V. The features of different dimensions can describe different aspects of candidate words. For instance, the word's position, whether there are candidate words in the context, etc. belong to the context features; the ratio of capital letters, the number of special characters, etc. belong to the sentences attribute features; And the proportion of nouns is part of speech features.

Chen et al. [5] divided the web text into formal text and informal text according to its style of writing. As shown in Fig.5, the formal text is a paragraph that describes the position or other information of the person; the informal text lists the personal information by lines without descriptive vocabulary. Through observation, it is easy to find that the proportion of capital letters in the formal text is significantly smaller than that of the informal text. Therefore, formal and informal text can be distinguished by the proportion of uppercase letters (we set 0.5 as the threshold). People tend to be more inclined to leave personal information in the informal text, which makes the personal attributes in the informal text are relatively dense. Therefore, the style of the text can be regarded as an important feature.

Due to the large interference factors, it is difficult to extract candidate words accurately, which results in the candidate set often containing much noise and further affects the verification. For example, the expected pre-extracted attribute value is Columbia University, but since (B.S.) is a degree information and is located near the attribute, it is easy to obtain Columbia University (B.S.) as a candidate set. Similar phenomena occur in the extraction of nationality and degree. If the extraction result is simply compared with the tagged value, many negative examples will be involved, which will seriously affect the overall effect of the trainer. Therefore, for the above text, this paper uses the word vector to calculate the similarity of two words for the matching. When the similarity is greater than 0.9, the candidate is considered positive. SpaCy[5] is a natural language processing toolkit written by Python and in this paper we adopt its built-in word2vec API directly.

[4] http://nlp.uned.es/weps/weps-2

[5] https://spacy.io/

TABLE V: Selected Features

Context Feature	word position	Sentences Feature	ratio of capital letters
	times of word shows up		number of capitalised words
	candidate word nearby		ratio of capitalised words
	keyword nearby		number of digits
Speech Feature	number of nouns		ratio of digits
	ratio of nouns		number of special characters
	number of prepositions		ratio of special characters
Sentences Feature	number of words		paragraph style
	number of capital letters		contain url

TABLE VI: Results of Person Attribute Extraction Using Proposed Method

Attribute	P	TP	Total	Precision(%)	Recall(%)	F1(%)
organization	321	121	417	37.7	29	37.8
birthplace	5	3	10	60.0	50.0	40.0
nationality	34	18	49	52.9	36.7	40.0
location	79	21	156	26.6	13.5	17.9
school	22	11	24	50.0	45.8	47.8
position	246	96	511	39.0	18.8	25.4
email	19	12	24	63.2	50.0	55.8
birthday	11	5	41	45.5	12.2	19.2
name	449	401	1302	89.3	30.8	45.8
average	1186	688	2534	58.0	27.2	37.0

TABLE VII: Results of Person Attribute Extraction on WePS

System	Precision(%)	Recall(%)	F1(%)
PolyUHK	30.4	7.6	12.2
ECNU_1	6.8	18.8	10.0
ECNU_2	8.0	17.6	11.0
MIVTU	5.7	15.5	8.3
CASIANED	8.5	19.0	11.7
UC3M_1	2.5	2.2	2.3
UC3M_2	2.4	2.2	2.3
UC3M_3	2.2	2.0	2.1
UC3M_4*	2.2	2.0	2.1
UC3M_5*	8.0	3.6	5.0
UvA_1	2.7	27.3	5.0
UvA_2	4.4	27.4	7.6
UvA_3	0.7	0.2	0.2
UvA_5	3.3	2.8	3.1

C. Evaluation Experiments

In this part of the experiment, the WePS dataset was used. The experimental results were measured in terms of accuracy, recall rate, and F1 score. As shown in Table VI, the accuracy rate of the person's name is the highest, 89.3%, the average accuracy rate is 58%; the recall rate of the school is the highest, 45.8%, the average recall rate is 27.2%; the highest F1 score of the position is 55.8%, the average F1 score is 37%.

By comparison, as shown in Table VII, it is not difficult to find that in the absence of some attributes, the results of this paper are better than those of WePS in terms of accuracy, recall, and F1 score. There are two main reasons for the result: 1) The attributes information in all of the web pages used in this paper is sparse, while the personal information in the web pages of WePS is nonuniform. Obviously, the heterogeneity increases the difficulties. 2) For the sparsity, this paper adds pre-extracted block filtration and verification of noise processing based on WePS.

IV. Analysis

After extracting person attributes information from Darknet, we deeply analyze and measure the attributes and characteristics of the entities. In order to show a comprehensive landscape, we describe our analysis results in detail in this section.

A. The Measurement and Analysis of the Information Content of the People in the Darknet

In order to obtain attribute information of people in the Darknet, we randomly select more than 50,000 pages from thousands of dark web pages that we have crawled. These pages contains person attribute information such as email, bitcoin, etc. Then we use the attribute extraction method mentioned in Section III. More than 50,000 attribute records are extracted from these pages, each attribute record corresponds to a dark web page. Next, in order to compare to the information content of people on the surface web page, we randomly select 2,092 web pages from the surface network (in WEPS2 data sets) and extract person attributes, and count the information content of each record.

Fig.6 shows the distribution of the number of attributes, both in the surface web and dark web. We can find that most of the surface web pages are rich in people information; However, most of the dark web pages have sparse information. The most likely reason is that the Darknet users pay more attention to the privacy of their personal information than the surface web users, and do not arbitrarily disclose personal attributes. On the other side, Dark Web is relatively lacking in pages which is rich in personal information such as personal homepages.

B. The Analysis of Top-K Active Entities in the Darknet

In this section, we analyze distribution of common people, organizations, and countries in the Darknet and compare them with the surface network to discover the characteristics of these entities in the Darknet. Therefore, we measure and analyze the Top-10 people, Top-10 organizations, and Top-20 countries in the dark web pages.

In order to extract entities from the 50,000 pages mentioned in the last subsection, we use the NER (Stanford Named Entity Recognizer) to recognize the people names, organizations and countries entities in the web pages. As shown in the Table

TABLE VIII: Top-10 Organization Entities

Rank	Organization	Num
1	Google	112,068
2	Facebook	103,492
3	First Cams	78,277
4	Twitter Ban	73,980
5	UTC	69,755
6	Mircosoft	27,082
7	Paypal	25,293
8	Magyar	23,193
9	BTX	22,920
10	LTC	15,614

TABLE IX: Top-20 Country Entities

Rank	Country	Num	Rank	Country	Num
1	US	121,911	11	France	13,432
2	UK	83,878	12	Germany	13,187
3	Europe	83,708	13	Russia	12,887
4	Korea	48,267	14	Mexico	11,752
5	South Korea	47,653	15	Canada	11,418
6	Indonesia	19,657	16	China	10,921
7	Portugal	15,419	17	Romania	10,591
8	Australia	14,720	18	Nassau	10,481
9	Japan	14,546	19	Bulgaria	10,097
10	Pandaria	13,847	20	India	10,080

XII, the most common people in the Darknet is Bli, with 85598 times occuring in the web pages. More importantly, common Darknet people names are generally not common in the surface network. The Table VIII shows that most of the Top-10 organizations in the dark pages are well-known companies. The most common organization is Google, with 112,068 occurrences. Well-known Internet companies such as Facebook, Twitter, and Microsoft are also on the top 10 list. The emergence of First Cams, BTC, and LTC on the list shows the privacy and anonymity of Darknet. As for the Top-20 countries, Table IX shows that the United States had the highest number of 121,911. China is ranked at 16th. The distribution of Top-10 organizations and Top-20 countries basically meet our expectations. However, the distribution of Top-10 people is quite different from the surface web. We guess this is because the person names discussed on the Darknet are very different from the surface network due to the anonymity of the Darknet.

TABLE X: the Geographical Distribution of Pictures in Darkweb

Country	Num
US	2,965
Germany	23
Austria	22
Poland	5
Brazil	2
England	1
Russia	1

TABLE XI: Top-9 Email Domain in Surface Web

Rank	Email Domain
1	Gmail
2	Outlook
3	Yahoo
4	GMX
5	AOL
6	Zoho
7	Lycos
8	inbox
9	Hushmail

C. *The Analysis of User's Geography Distribution*

Because of the anonymity of dark networks and the high security awareness of users, it is difficult to directly obtain the user's geographical location. Therefore, most researches indirectly statistically analyze the Darknet user's geographical location information through side channels. In this section, we use the phone number and Exif header of pictures to locating the geographical location of users in the Darknet.

To locate users through the Exif header, we use the method mentioned by Paul Lisker[6]. After analyzing 621,904 pictures collected from Darknet, we find that only 3020 pictures has GPS information. Table X shows the geographical distribution of the location extracted from these pictures. Most of the pictures are token in the United States.

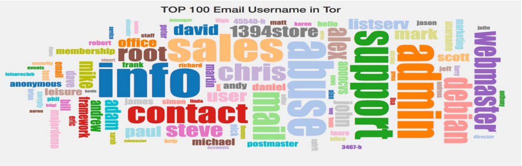


Fig. 7: Top-100 Email Username in Tor

To locate users through the phone number, we get 1030 Chinese phone number extracted from Darknet web pages. Fig.8 shows that most of phone numbers concentrate in Guangxi Province and Guangdong Province.

D. *The Analysis of Email Address in the Darknet*

In order to understand how many users will leave their own email addresses, the distribution of email domain names in the Darknet, the proportion of anonymous email address, etc., we make a statistical analysis of the domain names and user names of the email address.

In this section, we have extracted a total of 187,769 email addresses from the Darknet through regular expressions, and statistically analyzed the top-10 domain name of the email addresses and compared the result with that of the surface network. At the same time, in order to analyze the semantic information content of the email user name, we make a visual analysis of the top-100 emails users.

As shown in Fig.9, the Gmail is most popular email domain in the Darknet. Further more, there are more non-anonymous mailboxes than anonymous mailboxes. Comparing with distribution of email domain in the surface web (Table XI), the most popular mailbox systems in the Darknet are not common in the surface network. For the Top-100 user name of email, we use the word cloud for visual analysis (Fig.7). The results show that most of the users are named info, sales, etc. The semantic information content is quite rare with barely no value.

V. Conclusion

The Darknet carries a large amount of sensitive information. Given the scarcity of personal attributes on the Darknet,

[6] http://toutiao.secjia.com/harvard-students-show-darkweb-merchant-physical-map

TABLE XII: Top-10 Name Entities

Rank	Dark Web		Surface Web Boys		Surface Web Girls	
	Name	Percentage	Name	Percentage	Name	Percentage
1	Bli	2.51	Noah	0.963	Emma	1.05
2	Lauri Love	2.13	Liam	0.902	Olivia	1.01
3	Ett	1.28	Mason	0.816	Sophia	0.895
4	Deutsch	0.93	Jacob	0.78	Ava	0.842
5	kom	0.83	William	0.78	Isabella	0.801
6	Lax SameSite	0.82	Ethan	0.74	Mia	0.766
7	Polski	0.75	James	0.726	Abigail	0.636
8	Dudu	0.46	Alexander	0.713	Emily	0.606
9	Matti	0.42	Michael	0.707	Charlotte	0.586
10	Eauide	0.42	Benjamin	0.671	Harper	0.529

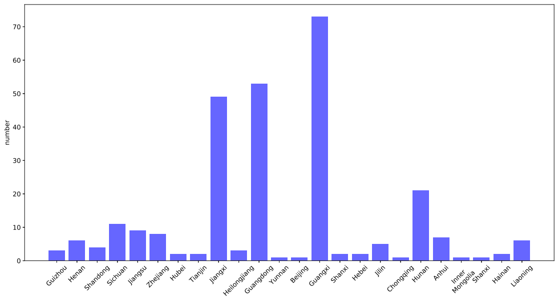


Fig. 8: the Geographical Distribution of Phone Number in Darkweb

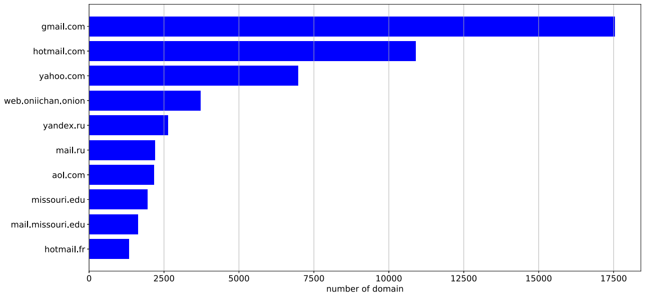


Fig. 9: Top-10 Email Domain in Tor

our paper modifies the existing person information extraction methods and gets higher scores of accuracy, recall rate, and F1. Based on our method, we measured and analyzed the attributes of people on the Darknet from the perspectives of top-k names, the number of attributes, email domain, geographic distribution and etc.

For the future work, the person information extraction towards Darknet might be examined further to get better effect of extraction, and we hope to implement a Darknet People Retrieval System based on our extraction technology.

Acknowledgment

This work was supported in part by the National Key Research and Development Program of China under Grant No.2016YFB0801003.

References

[1] James Ball, "Silk Road: the online drug marketplace that officials seem powerless to stop," http://www.theguardian.com/world/2013/mar/22/silk-road-online-drug-marketplace.

[2] J. Tang, J. Zhang, D. Zhang, L. Yao, C. Zhu, and J. Li, "Arnetminer: An expertise oriented search system for web community," in *Proceedings of the 2007 International Conference on Semantic Web Challenge-Volume 295*. CEUR-WS. org, 2007, pp. 1–8.

[3] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 990–998.

[4] J. Tang, J. Zhang, L. Yao, and J. Li, "Extraction and mining of an academic social network," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 1193–1194.

[5] Y. Chen, S. Y. M. Lee, and C.-R. Huang, "Polyuhk: A robust information extraction system for web personal names," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[6] X. Han and J. Zhao, "Casianed: People attribute extraction based on information extraction," *City*, pp. 20–24, 2009.

[7] M. Lan, Y. Z. Zhang, Y. Lu, J. Su, and C. L. Tan, "Which who are they? people attribute extraction and disambiguation in web search results," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[8] C. De Pablo-Sanchez and P. Martínez Fernández, "Uc3m at weps2-ae: Acquiring patterns for people attribute extraction from webpages," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*. Citeseer, 2009.

[9] D. Pinto, M. Tovar, D. Vilario, H. Dıaz, and H. Jiménez-Salazar, "An unsupervised approach based on fingerprinting to the web people search task," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*. Citeseer, 2009.

[10] K. Watanabe, D. Bollegala, Y. Matsuo, and M. Ishizuka, "A two-step approach to extracting attributes for people on the web," in *2nd Web People Search Evaluation Workshop (WePS 2009), 18th WWW Conference*, 2009.

[11] I. Nagy and R. Farkas, "Person attribute extraction from the textual parts of web pages." *Acta Cybern.*, vol. 20, no. 3, pp. 419–440, 2012.

[12] J. Artiles, J. Gonzalo, and S. Sekine, "Weps 2 evaluation campaign: overview of the web people search clustering task," in *2nd web people search evaluation workshop (WePS 2009), 18th www conference*, vol. 9, 2009.

[13] E. Amigó, J. Artiles, J. Gonzalo, D. Spina, B. Liu, and A. Corujo, "Weps-3 evaluation campaign: Overview of the online reputation management task," in *CLEF 2010 (Notebook Papers/LABs/Workshops)*, 2010.

[14] R. Dingledine, N. Mathewson, and P. Syverson, "Tor: The second-generation onion router," Naval Research Lab Washington DC, Tech. Rep., 2004.

[15] A. Biryukov, I. Pustogarov, F. Thill, and R.-P. Weinmann, "Content and popularity analysis of tor hidden services," in *Distributed Computing Systems Workshops (ICDCSW), 2014 IEEE 34th International Conference on*. IEEE, 2014, pp. 188–193.

[16] 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*. ACM, 2013, pp. 213–224.