Technical Report: A inventory of open and dark web marketplace for identity misrepresentation *

Arunkumar Bagavathi, Sai Eshwar Prasad Muppalla, Siddharth Krishnan, and Bojan Cukic

Department of Computer Science
University of North Carolina at Charlotte
bcukic@uncc.edu

Abstract. Biometric identification is a critically important technology in traveler, immigration and refugee management. The technology itself and the processes related to human identification and identity management are a prime target for identity theft, tampering, spoofing, and impersonation. In the past year, our team developed a systematic methodology for identification of biometric technology vulnerabilities and identity management process limitations. We defined several attack vectors and tried to establish objective measures of risk exposure. But to establish risk exposure, one needs to understand the social process that may lead to biometric attacks. Therefore, we developed a methodology to monitor publicly available information sources that may reveal the extent of threats, availability and sophistication of attack tools and how-to recipes for biometric attacks at US Ports of Entry. This report presents the search techniques we developed and overviews current results.

Keywords: Border Security · Legitimate trade and travel - technology · Immigration - identity fraud.

1 Introduction

Research literature in the field of biometric security, including liveness detection and anti-spoofing is evolving. A good understanding of the risks stemming from zero-effort attacks obtained in recent years (biometric misidentification rates due to the probabilistic nature of the technology) is just the tip of the iceberg of identity misrepresentation risks in homeland security. Subverting biometric recognition with artificial materials (gummy fingers, intense face make-up or patterned

* This material is based upon work supported by the U.S. Department of Homeland Security under Grant Award Number 2015-ST-061-BSH001. This grant is awarded to the Borders, Trade, and Immigration (BTI) Institute: A DHS Center of Excellence led by the University of Houston, and includes support for the project A Systematic Process for Vulnerability Assessment of Biometric Systems at Borders awarded to the University of North Carolina at Charlotte. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the U.S. Department of Homeland Security.

lenses in eyes) is becoming known once the reports of successful attacks surface in public. In addition to technical limitations of biometric collection and matching, some of the broader identity verification processes (limitations of trust placed on foreign ID documents or exception handling due to unavailable biometric, for example) may inadvertently allow inaccurate identification or intentional acts of identity tampering. Yet, the accuracy of human identification for travelers, immigrants and refugees is the cornerstone of trustworthy immigration system and US borders, the subject of intense national scrutiny.

As the government clamps on illegal immigration and strengthens border control, the illegal and malicious activities to bypass or get through the identity checks and security vetting are likely to increase. The new generations of individuals that want to visit US as travelers, become immigrants or receive refugee status are avid technology users. For example, Syrian refugees have used social networks and smartphone apps to plan different legs of their journey to Europe, and to locate resources, aid and in some cases smugglers [8]. In addition, some refugees may have been using online social media sites to communicate and share their experiences and offer services [2]. Individuals trying to hide their identity will very likely use technology to their advantage. In addition to the open source web space, these activities will inevitably be extended in the dark web space, the anonymous TOR based web, where the identity of users is secret and where the sale and distribution of illicit activities and resources can be acquired anonymously too. TOR based web forums and market places are likely to provide an Amazon-like experience that allows users to surf and purchase tools and how-to kits about avoiding identification.

Continuation of our research on reducing the identity misrepresentation risk and misuse opportunities for travelers, immigrants and refugees requires us to understand resources and techniques for identity concealment that could be acquired from open and black markets, and open and dark websites. This information will allow us to tune-up the risk model with collected metrics, costs and possible experiences related to attack vectors. The technical report provides an overview of the search techniques we developed and the results we acquired related to the presence of "know-how" information that may assist in identity spoofing at US ports of entry. We investigated open and dark websites and report here the availability of tools, hardware and software required for faking one's identity as they may relate to attacks to US immigration services. The rest of this report is organized as follows. Section 2 discusses related research work in automated and semi-automated large-scale searche of Web domains. Section 3 reports the methodology developed and used in search for the results. Section 4 lists some of the results. Section 5 concludes the report.

2 Related work

Regardless of the knowledge doamain, every search is initiated by a manually designed finite set of terms can capture only limited information from massive resources like, Web pages, blogs and social media. Query expansion is an infor-

mation retrieval technique that models a set of query terms from a given set of query items to optimize the effectiveness of gathering useful information and insights. Rocchio's Smart Retrieval System is the first query expansion model [14]. Query expansion can be categorized into three varieties according to [9]: manual, automatic and interactive. Manual query expansion requires a user to find insightful queries from the retrieved information and the updated query list is used for the next iteration. This process continues until the user is satisfied with the set of results. *Automatic query expansion*, on the other hand, involves no human interaction on query expansion process [3] by expanding the query terms based on some measures. In the literature, automatic query expansion techniques are mostly based on term frequency [6,12], term proximity [18], semantics between terms [15], domain specific terms [19] and query expansion based on a supervised model [11]. *Interactive query expansion* provides more control to the user by providing them suggestions [17,15]. It is up with users to select the query terms. Such a technique is useful for semi-automated 'human-in-the-loop' scenarios and it is the most appropriate approach for our work. Other information retrieval research on query expansion from a set of images [4] are considered out of the scope of this report.

A step in our query expansion process includes *Wordclouds*. They have been used in multiple applications to determine the most frequently occurring terms from the given set of documents [10]. Wordcloud representations are engaging for users¿ They make it easy to identify tectual terms and patterns due to the use od multiple colors, fonts and shapes. Even though wordclouds mostly capture words that have the highest frequency compared to other words in the vocabulary, they can also be used to identify words that are rarely occurring or to cluster words that are similar based on similarity or semantic measures. In the *interative query expansion* process we are about to describe in the next Section, wordclouds are used by search - domain experts to direct queries and identify more appropriate query terms.

Open web content like Wikipedia are easily accessible with crawlers. However, this is not the case for dark web. Since dark webs have a distinct domain (mostly *.onion*), there are only a few data crawlers for collecting dark web data. Some of the big challenges in crawling dark web are accessibility and safety. This was an important problem for our project as we did not expose our research team, and students in particular, to potentially risky behaviors and safety risks. Over the past decade, many researchers have proposed techniques for searching *topic-specific* dark web [13] and focused crawling on dark web [7]. Apart from the research, there is now a commercial application programming interface (API), *Webhose.io* [1] available to collect and summarize dark web data. This is a paid service. It hides the identity of students and faculty performing dark web search and it is very effective. For these reasons, we decided to use it for all dark web interactions in this project.

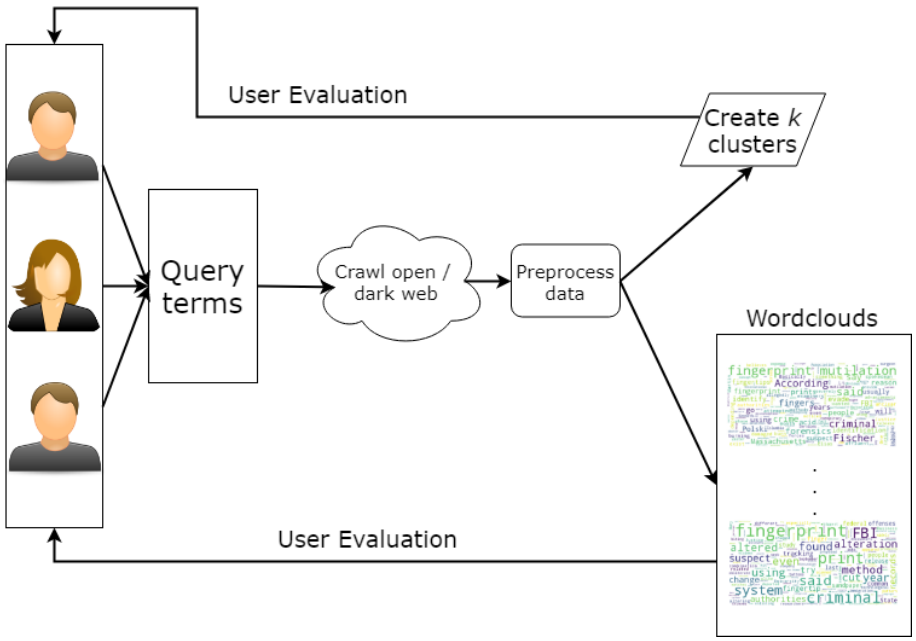

Fig. 1. Interactive query expansion method overview. The process starts with experts giving initial set of key words. We scrape / collect data from open / dark web from the given query words.The resulting webpages are pre-processed and clustered into k groups. k wordclouds are built to represent each of the k clusters. Wordclouds are evaluated by experts to get next set of query terms. At that point the process repeats until the result are deemed satisfactory.

3 Methodology

In this section, we describe the methods that we followed to perform interactive query expansion. We focus on illicit activities, such as falsifying biometrics, obtaining or forging identity documents to enter the country. We are looking for services or advice offered with or without compensation that could inspire a valid attack vector. Towards that end, we collect data from the open web and the dark web. An enormous amount of text may potentially need to be processed and visualized,. Therefore, rather than reading the text, the user can evaluate visualized information feed as a way to create a new set of queries as input to the system. Figure 1 gives an overview of our methodology to expand the list of query words. This process consiste of the following steps:

1. Build a vocabulary of queries from human knowledge.
2. Query open and dark web from the given input queries to collect a set of webpages.
3. Preprocess the webpage text and extract features of cleaned words.
4. Create k clusters from the results.
5. Visualize the data with k wordclouds.
6. Let the user pick more relevant key words from wordclouds and k clusters.
7. repeat until the results are satisfactory.

3.1 Building vocabulary of key words

The retrieval process starts with a set of seed words to initiate the process. We formulated basic query terms using subject matter expertise on biometrics and security. Our basic key words like *biometric attacks*, *fingerprint falsification*, *fingerprint alteration immigration*, which are more related to words that poses an indication for people talking about illegal immigration using falsifying or spoofing biometric devices. We considered on various biometric devices for scanning fingerprints, faces and eyes for deriving the seed key words.

3.2 Query webpages

With the seed words, we search on open web and dark web to collect multiple posts that talk about the given keywords. For the open web data, we used Google search API and for the dark web data, we used Webhose.io API [1]. The complete data collection process is further outlined in the following sections.

Open web querying Collecting data using the Google search API is a *2-step process*. Given a set of key words to the API, it returns a number of webpage links (according to API rate limits). Upon collecting these webpage URL's, we can scrape the data available in these webpages. Ideally, the production version of the system would not be limited by the number of URL's offered by the free version of Google's search API, but this was acceptable for the proof od concept.

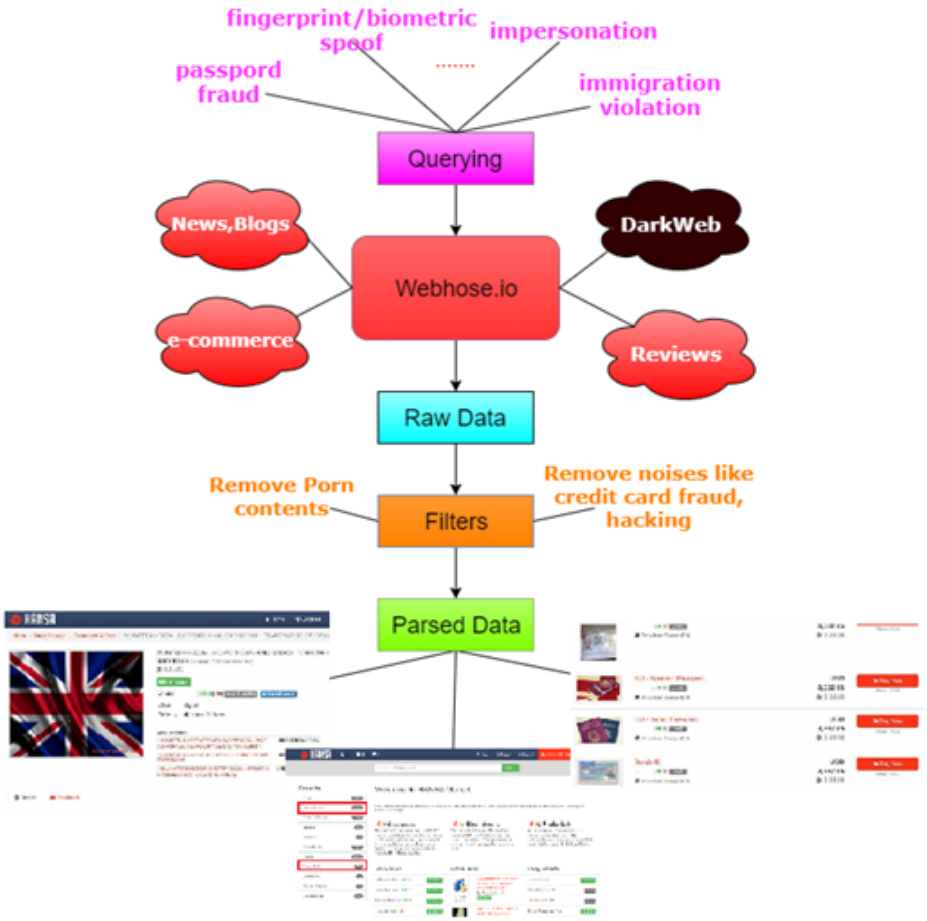

Fig. 2. Dark web data collection process.

Dark web querying Collecting dark web data using the webhose.io API is a simple process. Webhose.io is a commercial API to collect data of multiple streams like news and blogs, movie or hotel reviews, e-commerce, television and dark web. Given a set of key words, the API returns a dictionary of data with all URL's, webpage text and images. But this simple process comes with a cost of getting some irrelevant outputs from the API too. We noticed that many returned dark web results were irrelevant for our searches, a much larger proportion than in our open web search. Therefore, we incorporated several filters to ensure the higher quality of the output. Figure 2 gives an overview on our process of filtering keywords and posts collected from the API. These filters reduced the exposure of our student researchers to unwanted dark web content, including credit card fraud, hacking, pornography and similar illicit activities. While filtering is never 100 percent effective, it helped a lot.

3.3 Data preprocessing

After collecting online posts, from both open and dark web, we execute text preprocessing methods [16], including *tokenization, stop words removal, lemmatization* and *term frequency* on the collected documents or web pages.

Tokenization and Stop words removal: *Tokenization* is the process of converting sentences or paragraphs into words. Technically, consider a text $\mathbb{T} = \bigcup_{i=1}^{s} \mathbb{S}_i$, where $\mathbb{S}_i$ is the i^{th} sentence in the given text $\mathbb{T}$ and each sentence can comprise of words separated by a delimiter α. With *Tokenization*, we split all sentences $\bigcup_{i=1}^{s} \mathbb{S}_i$ using the delimiter α to extract all words, without removing duplicate words, from $\mathbb{T}$. All extracted words are not useful. English text is made of most commonly occurring words like *articles*, *conjunctions*, *pronouns* and *prepositions*. Not all of these are equally important for extracting useful knowledge from the text. Thus, in all text processing, it is mandatory to remove some of the less important words from the given text prior to the analysis.

Lemmatization: Even though we extracted useful words and eliminated all stop words, there remain inflections in the available words. For example, plural form of words like *biometrics* and *fingerprints*. The traditional approach to handle this is *stemming*, which cuts the stem of a word. For example, *happiness* and *happy* gets converted to *happi*. Thus, *stemming* can alter the underlying meaning of a word and, occasionally, it may not create understandable content. Unlike stemming, *lemmatization* follows a careful approach by relying on lexical knowledge to obtain a root of the word. Since we require human readable words from the given text, we use *lemmatization* in our experiments and avoid stemming.

Term Frequency and Inverse Document Frequency: At this point in our research, we consider a simple and most commonly used feature for the representation of a body of text called *Term Frequency-Inverse Document Frequency*. From this form, we can conduct clustering. *Term Frequency* differs from the general meaning of *Word count*. Word count simply gets the number of times the given word $\mathfrak{t}$ occurs in total of all documents. *Term Frequency* on the other hand measures frequency of the word $\mathfrak{t}$ in a given document(webpage) $\mathbb{D}_i$. Term Frequency of a word $\mathfrak{t}$ in the document $\mathbb{D}_i$ can be given in the formula as

$$TF(\mathfrak{t}, \mathbb{D}_i) = \frac{f(\mathfrak{t}, \mathbb{D}_i)}{\sum\limits_{\mathfrak{t}' \in \mathbb{D}_i} f(\mathfrak{t}', \mathbb{D}_i)}$$

where $f(\mathfrak{t}, \mathbb{D}_i)$ is the word count of a word $\mathfrak{t}$ in $\mathbb{D}_i$ and $\sum\limits_{\mathfrak{t}' \in \mathbb{D}_i} f(\mathfrak{t}', \mathbb{D}_i)$ is the sum of word counts of all words $\mathfrak{t}'$ in $\mathbb{D}_i$. Given a set of 'N' documents $\mathbb{D} = \{\mathbb{D}_1, \mathbb{D}_2, \ldots \mathbb{D}_N\}$, the *Inverse Document Frequency* can be given as

$$IDF(\mathfrak{t}, \mathbb{D}) = log \frac{N}{\left|\mathbb{D}_i \in \mathbb{D} \mid \mathfrak{t} \in \mathbb{D}_i\right|}$$

where $\left|\mathbb{D}_i \in \mathbb{D} \mid \mathfrak{t} \in \mathbb{D}_i\right|$ is the number all documents that contain the word $\mathfrak{t}$. With the above given formula, we can calculate $TF - IDF$ score of a term $\mathfrak{t}$ by

$$TF - IDF(\mathfrak{t}) = TF(\mathfrak{t}, \mathbb{D}_i) * IDF(\mathfrak{t}, \mathbb{D})$$

3.4 k-means clustering

We use an iterative unsupervised machine learning algorithm called *k-means clustering* [5] to group the words from text into k clusters, where k is defined by the user. The algorithm starts with random k clusters based on random k centroids chosen from the data. Each data point is allocated to exactly one nearest cluster. The distance of a data point to the cluster is calculated using a distance / similarity measure. The most common distance / similarity measures are: *cosine similarity, Euclidean distance* and *Manhattan distance*. Once all data points are allocated to clusters, cluster centroids are recalculated by taking the mean of all data points in the cluster. In the next iteration, distance is calculated from data points to the new clusters and the algorithm continues to optimize allocation similarly for n iterations.

3.5 Interactive query expansion

With the pre-processed data, we collect a group of wordclouds based on word frequencies of each word from each webpage. These wordclouds are presented to users. They can select a new set of query terms apart from terms available in the current search string. Let $\mathbb{V}_0$ be the initial set of m query words given by experts

and $\{\mathbb{W}_i\}_{i=0}^{n}$ be the set of n wordclouds from webpages crawled from open web and dark web, where $|\mathbb{W}_i| = q$. From these k wordclouds, say the experts are finding a unique set of words $\mathbb{V}'_0$, where $\mathbb{V}_0 \cap \mathbb{V}'_0 = \emptyset$. For the next iteration of data collection process, we use the latest set of query words $\mathbb{V}'_0$ to crawl the data from open and dark web and process continues until we receive empty results from APIs or the iteration reaches the maximum count p.

4 Results

We represent the results of our algorithm through a simple example.

4.1 Human engineered query terms

The first phase of our method is to design a set of query terms based on experts knowledge with respect to the domain problem (biometric impersonation risks for points of entry). For our test purpose, we considered all potential biometric systems, but give consideration preference to the results which refer to *fingerprints, face recognition* and *passport fraud*. Figure 3 gives a wordcloud of a set of initial query words selected to fit the problem being considered.

Fig. 3. Wordcloud of query terms selected by experts.

4.2 Results based on *k-means clustering*

Using the crawled webpages from open and dark web, we remove all stop words, that are most common English words. We then extract Term Frequency-Inverse

Document Frequency(*TF-IDF*) scores of words from the pre-processed documents. With the available *TF-IDF* scores as word features, we experiment using *k-Means* clustering with k value set as '*2*'. The value of k is so low to make it easier to follow the process. We experimented with and analyzed the values of k in the range from dozens to hundreds. Figure 4 gives sample clusters of words from the *2-means* clustering algorithm.

Fig. 4. Wordclouds from the results of *2-means* clustering algorithm based on TF-IDF features of the data collected from Open web and Darkweb

4.3 Representing web content as wordclouds

We generate wordclouds from collected web posts on open and dark web. Figure 5 depicts a sample of wordclouds from the posts on Darkweb using keywords *fingerprint alteration, immigration, fake biometric, attack* and *artificially manufactured biometric*.

Fig. 5. Wordclouds from the data extracted from Darkweb using keywords ””

4.4 Interesting finds

We apply human knowledge on the wordclouds based on *k-Means clustering* and *word counts* to extract more query terms related to the problem domain. Apart from retrieving new query words for our system, we also get more useful webpage links for to analyze web content and future analysis of such web pages. Some of such retrieved web sites (open web) are:

- `http://www.banderasnews.com/howto/fingerprints.htm`
- `https://careertrend.com/how-2122642-fake-fingerprints.html`
- `https://betanews.com/2017/01/20/from-spoofing-to-iris-scanning-\ the-future-of-biometrics/`
- `https://www.ncbi.nlm.nih.gov/pubmed/21216360`
- `https://www.theguardian.com/technology/2014/dec/30/hacker-fakes-\ german-ministers-fingerprints-using-photos-of-her-hands`

URLs of the results we collected from the Dark Web would not be as useful to include here. The content in the Dark Web is often revoked or removed making its retrieval possible only from the version-control datasets. Luckily, *Webhose.io* offers the search over the version controlled content. The following is the sample of our findings on the popular black markets:

1. HANSA, AlphaBay: top black markets that acts as e-commerce websites to sell fake identity passports, SSNs and driving licenses of various countries. Both were taken down by FBI in July 2017
2. Counterfeit Guru, House of Lions Forums: currently inactive websites similar to HANSA and AlphaBay support discussions of illegal identity manipulation - related activities.
3. WSM Forum : Currently active dark website to sell fake identity documents from different countries;
4. Hidden Answers : Discussion forums on illegal topics

Example questions posted in Hidden Answers that match our interest include the following:

- So I heard that there were ways to get a U.S. Citizenship simply by "buying" one. I know that there is a way by marrying and then divorcing (sorry if I spelled that wrong) but that's not what I'm looking for. Is there an actual way to actually buy one? I'm willing to pay a very high price if I must. These guys over here "claim" they can give you a citizenship.
  (http://xfnwyig7olypdq5r.onion/,
  http://sla2tcypjz774dno.onion/uscitizen.html)
- Please tell me if either one are scammers or actual people for actual services. [Posted on Nov 23, 2016].

Example responses to the above question include these:

1. Hey if you are very serious getting to U.S and want to be there, Hit me up. i got a way. I am not selling you anything but will give you this Idea. you can try your luck. Its still working for many and hopefully wont be burnt soon. [Posted On Nov 24,2016]
2. I can help you get without paying. [Posted On Dec 2,2016]

The data we collected contains many dark web posts related to the sales of fake SSNs, passports and driving licenses of various countries. For example:

STORY1 - US,UK Passports:

UK Passports - Buy real UK passports, become a UK citizen now. Our passports are no fake passports, they are real passports.
Products Login Register FAQs Your UK Passport - Name of your choice! We are selling original UK Passports made with your info/picture.
Also, your info will get entered into the official passport database.
So its possible to travel with our passports.
How we do it? Trade secret!
Information on how to send us your info and pictures will be given after purchase!
You can even enter the UK/EU with our passports, we can just add a stamp for the country you are in!
Ideal for people who want to work in the EU/UK. Product Price Quantity Your original UK passport with your info/pictures 2000 GBP

STORY2 - US Fake passports - hidden answers:

Do fake passports actually work in getting in and out of usa? - Hidden Answers
COMMENT1 : Here is the link of high qualitative documents service. I've found on hidden wiki: * http://fakeidskhfik46ux.onion - Fake documents service online. 3-5 days FREE express delivery worldwide. commented Jan 18, 2016 by Danlos N00b 101 (60 points)
COMMENT2 : the ones you could find on the DW, are usually only good for NON government type dealings where you may need to present some form of ID, any kind of "Official" will flag it pretty quickly, though there may be one on here somewhere, all the ones I've seen are either too cheap to be the real deal, or clearly say they cannot be used in government situations, or are obvious scams, good luck bro answered Jan 8, 2016
COMMENT3 : You won't find anything like that here.
But if you have a good quality fake and you are going into a low risk country like Canada, yeah it will work. However if you are flying to or from some high risk area like North Korean or Syria, where they are on high alert and look hard at everyone and every thing, you are fucked. answered Jan 8, 2016 by MrBlack Champ (46,630 points)

STORY - 3:

Where I can buy fake USA passport? - Hidden Answers COMMENT : First tell ,you want it from deep web or from your nearby.
if from DW, then you will find on Hidden Wiki. answered Feb 15, 2016 by Nova Grand Master (22,400 points)
COMMENT : You can buy it from http://fakeidskhfik46ux.onion answered Feb 15, 2016 by Danlos N00b 101 (60 points)
COMMENT : Just done a quick check, AlphaBay http://pwoah7foa6au2pul.onion Dr. D's Multilingual Market http://drddrddig5z3524v.onion and Python http://25cs4ammearqrw4e.onion have them, I can't personally vouch for any of the vendors though. answered Feb 15, 2016 by CaptainCapsaicin Master First-Class (10,575 points)

STORY - 4:

Order FINGERPRINTING TUTORIAL - MAKE FAKE FINGERPRINT 3 TUTORIAL ABOUT FINGERPRINTING: - HOW TO MAKE FAKE FINGERPRINT -Impact of Artificial "Gummy" Fingers on Fingerprint Systems -Biometrical Fingerp... 5.00 USD Order UK black Visa prepaid debit/CC with NO yearly limit How to apply for a UK black Visa prepaid debit/CC with NO yearly limit:... 5.00 USD Order

5 Summary

Over the past five months, we developed a novel and original human-in-the-loop approach for searching open and Dark Web for topics of interest. The purpose of this research step was to better understand the risks of occurrence of biometric impersonation attacks at US Points of Entry.

We believe that the proposed search algorithm is very promising. We have been able to find information about the interest in illegal identity manipulation and a market of services that cater to such requests. The majority of our findings so far are about passport fraud. We have detected interest in spoofing biometric identity too. However, so far we have detected only a few cryptic descriptions of services that might be available on the Dark Web. While our results indicate that the demand for identity spoofing exists, so far this is not creating substantial input into the risk modeling which we hoped to collect. In spite of the results so far being more modest than expected, this report indicates that open and dark web need to be continually monitored and exploited for open source intelligence efforts to prevent eventual outbursts of infoirmation thst might entice travelers, immigrants or refugees to try illegal entry through the biometric ports of entry.

References

1. (2018), `https://webhose.io/data-feeds/dark-web/`, [Online; Last accessed: 05/20/2018]

2. Brunwasser, M.: A 21st-century migrants essentials: Food, shelter, smartphone (2015), https://www.nytimes.com/2015/08/26/world/europe/a-21st-century-migrants-checklist-water-shelter-smartphone.html, [Online; Last accessed: 06/03/2018]
3. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Surveys (CSUR) **44**(1), 1 (2012)
4. Chum, O., Philbin, J., Sivic, J., Isard, M., Zisserman, A.: Total recall: Automatic query expansion with a generative feature model for object retrieval. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8. IEEE (2007)
5. Cohen, M.B., Elder, S., Musco, C., Musco, C., Persu, M.: Dimensionality reduction for k-means clustering and low rank approximation. In: Proceedings of the forty-seventh annual ACM symposium on Theory of computing. pp. 163–172. ACM (2015)
6. Collins-Thompson, K., Callan, J.: Estimation and use of uncertainty in pseudo-relevance feedback. In: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 303–310. ACM (2007)
7. Fu, T., Abbasi, A., Chen, H.: A focused crawler for dark web forums. Journal of the Association for Information Science and Technology **61**(6), 1213–1231 (2010)
8. Gillespie, M.: Phones crucial to survival for refugees on the perilous route to europe (2017), http://theconversation.com/phones-crucial-to-survival-for-refugees-on-the-perilous-route-to-europe-59428, [Online; Last accessed: 06/03/2018]
9. He, B.: Query expansion models. In: Encyclopedia of database systems, pp. 2257–2260. Springer (2009)
10. Heimerl, F., Lohmann, S., Lange, S., Ertl, T.: Word cloud explorer: Text analytics based on word clouds. In: System Sciences (HICSS), 2014 47th Hawaii International Conference on. pp. 1833–1842. IEEE (2014)
11. Kotov, A., Zhai, C.: Tapping into knowledge base for concept feedback: leveraging conceptnet to improve search results for difficult queries. In: Proceedings of the fifth ACM international conference on Web search and data mining. pp. 403–412. ACM (2012)
12. Lee, K.S., Croft, W.B., Allan, J.: A cluster-based resampling method for pseudo-relevance feedback. In: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. pp. 235–242. ACM (2008)
13. L'huillier, G., Alvarez, H., Ríos, S.A., Aguilera, F.: Topic-based social network analysis for virtual communities of interests in the dark web. ACM SIGKDD Explorations Newsletter **12**(2), 66–73 (2011)
14. Rocchio, J.J.: Relevance feedback in information retrieval. The SMART retrieval system: experiments in automatic document processing pp. 313–323 (1971)
15. Singh, J., Sharan, A.: Co-occurrence and semantic similarity based hybrid approach for improving automatic query expansion in information retrieval. In: International Conference on Distributed Computing and Internet Technology. pp. 415–418. Springer (2015)
16. Srividhya, V., Anitha, R.: Evaluating preprocessing techniques in text categorization. International journal of computer science and application **47**(11), 49–51 (2010)

17. Zha, Z.J., Yang, L., Mei, T., Wang, M., Wang, Z.: Visual query suggestion. In: Proceedings of the 17th ACM international conference on Multimedia. pp. 15–24. ACM (2009)
18. Zhao, J., Huang, J.X., Ye, Z.: Modeling term associations for probabilistic information retrieval. ACM Trans. Inf. Syst. **32**(2), 7:1–7:47 (Apr 2014)
19. Zhao, L., Chen, F., Dai, J., Hua, T., Lu, C.T., Ramakrishnan, N.: Unsupervised spatial event detection in targeted domains with applications to civil unrest modeling. PloS one **9**(10), e110206 (2014)