DarkJargon.net: A Platform for Understanding Underground Conversation with Latent Meaning

Dominic Seyler
dseyler2@illinois.edu
University of Illinois at Urbana-Champaign, USA
Department of Computer Science

Wei Liu
weil8@illinois.edu
University of Illinois at Urbana-Champaign, USA
Department of Computer Science

Yunan Zhang
yunanz2@illinois.edu
University of Illinois at Urbana-Champaign, USA
Department of Computer Science

XiaoFeng Wang
xw7@indiana.edu
Indiana University Bloomington, USA
Center for Security Informatics (CSI)

ChengXiang Zhai
czhai@illinois.edu
University of Illinois at Urbana-Champaign, USA
Department of Computer Science

ABSTRACT

An essential part of underground conversation are dark jargon terms. They are benign-looking, but have hidden, sometimes sinister meanings and are used by participants of underground forums for illicit behavior. For example, the dark term "rat" is often used in lieu of "Remote Access Trojan". We present a novel online platform that caters to the understating of underground conversation with latent meaning. Our system enables researchers, law enforcement agents and "white-hat" hackers to gain invaluable insights into underground communication by providing them with a tool to (1) look-up dark jargon terms in a dictionary; (2) explore the usage of dark jargon over time and interpret their meaning; (3) collaborate and contribute their own research findings. Furthermore, we introduce a novel dark jargon interpretation method that leverages masked language modeling of a transformer-based architecture.

CCS CONCEPTS

• **Information systems** → **Specialized information retrieval**; **Language models**; **Deep web**; *Crowdsourcing*; *Web interfaces*.

KEYWORDS

dark jargon interpretation, underground forum, dark net

ACM Reference Format:
Dominic Seyler, Wei Liu, Yunan Zhang, XiaoFeng Wang, and ChengXiang Zhai. 2021. DarkJargon.net: A Platform for Understanding Underground Conversation with Latent Meaning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3404835.3462801

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3462801

User post: "you can easly [sic] get track2 off a card just swiping it in a msr"
track2: Track2 is an American Banking Association (ABA) format for storing information on the magnetic stripe on a credit card. [*debitcard*, *decrypter*]
msr: An MSR is a device that converts information on the magnetic stripe of a credit card into data that can be understood by retail software. [*card*, *slot*]

Figure 1: Example user post in the Dark0de [1] underground forum. The descriptions (sources: [11, 12]) of each dark term and the most-likely clean terms, retrieved by our method (i.e., words in brackets) are taken from DarkJargon.net.

1 INTRODUCTION

When participants in underground forums (e.g., Dark0de [1]) converse, they often obfuscate their true intentions by using dark jargon. These are terms that may look innocent to an outsider that is not a regular participant in underground activity, but the true meaning of these terms is hidden and sometimes sinister. For example, the dark term "rat" is often used as an acronym of "Remote Access Trojan" or "popcorn" is used to refer to a type of marijuana (i.e., popcorn nugs). For researchers and law enforcement alike, it is essential to have a good understanding of these latent meanings in order to gain valuable insights into underground activity. To this end, several systems for explaining domain-specific jargon have been developed, however, we find that their focus is often narrow and many dark jargon terms that we discover are not contained in these resources. Furthermore, existing computational approaches for jargon interpretation suffer from a lack of expressiveness, which we overcome by leveraging the more expressive dark term interpretation framework recently proposed in Seyler et al. [14] and a novel interpretation method based on masked language modeling.

To cater to the understanding of underground conversation we present our system *DarkJargon.net*[1] (**http://darkjargon.net**). We envision this to be a system that caters to the needs of researchers,

[1]We provide the code and data here: https://github.com/dom-s/dark-jargon.

law enforcement agents and "white-hat" hackers to enable them to make sense of underground conversation. In an IR setting, the system can help query intent understanding by improving the matching of documents with queries. Furthermore, the task of collecting dark jargon is inherently collaborative and in order to make it scalable, the system is powered by a novel method for the interpretation of dark jargon words and provides a platform to share and evaluate findings by the dark jargon research community. To achieve these goals, our system has three major use-cases:
(1) **Dark Jargon Lookup:** The system provides a human-curated list of dark jargon terms that were found in underground forums by cybersecurity experts.
(2) **Dark Jargon Exploration:** *Usage:* The system shows detailed usage statistics for a given dark term by presenting its usage over time. If needed, these statistics can also be further analyzed by "drilling-down" into specific underground forums. *Meaning:* To gain further insights into the latent meaning of dark jargon, the system displays the most likely clean[2] terms according to a novel dark term interpretation method based on masked language modeling and a method proposed in Seyler et al. [14]. As meanings of dark terms can vary depending on the underground community, the system shows the probability distributions over separate forums.
(3) **Dark Jargon Contribution/Collaboration:** To facilitate a community that explores the meaning of dark terms and shares its findings, the system provides a collaboration interface. *Explore/Validate:* Users can explore existing dark jargon terms that have been found by other users and give feedback on the correctness/incorrectness of terms. *Contribute Novel Dark Jargon:* The system allows users to contribute novel dark terms that they found in their research.

To demonstrate a use-case of our system, consider, the user post in Figure 1. Looking at the post there are two major challenges: (1) The dark term has a hidden meaning (e.g., "track2" could be the second song of an album) and/or (2) the dark term is ambiguous (e.g., "msr" could also mean Microsoft Research). Using our system the researcher is able to get a definition for both dark terms (Figure 2) and will be able to resolve any ambiguity by looking at the most common clean terms on our exploration interface (Figure 4). Furthermore, looking at the usage over time (Figure 3) gives the user an impression of when and how often the term was used overall and in different underground communities. If said user happens to be a researcher, she can evaluate existing dark jargon findings and contribute her own using the collaboration interface in Figure 5.

2 RELATED SYSTEMS

The systems most related to our work are human-curated jargon dictionaries, as well as computational approaches for dark jargon interpretation. There are multiple slang dictionaries, which usually cover a narrow domain: For example, The Jargon File [13] is a collection of terms that are used in "hacker culture". Even though this is a good effort to collect hacker jargon, its application is narrow as it only focuses on computing-related terminology and is heavily outdated (latest version: December 2003). A large collection of general slang terms is Urban Dictonary [6]. Here, the trade-off is generality for specificity, as the dictionary contains many "pop-culture" slang terms, however, we found that many dark jargon terms we discover are not contained in Urban Dictionary or are listed with different meanings. There is also some effort by government entities to create lists that help understand the communication of criminals. Some examples here are the list of slang terms and code words [7] created by the Drug Enforcement Administration (DEA) and the Cybersecurity Glossary [10] created by the National Initiative for Cybersecurity Careers and Studies (NICCS). While these lists are respectable efforts to understand domain-specific jargon they cannot be directly applied to underground forums, since dark terms are often community-specific and their usage varies greatly over time [15, 17]. Furthermore, Yuan et al. [17] proposed CantReader [16], which is a computational approach for identifying and interpreting dark jargon using word vectors, derived from a dark term's context. The method categorizes a dark term into one of five general classes for interpretation. For example, "popcorn" is categorized as "drug" and not as "marijuana", which would be more beneficial for interpretation. In contrast, our system leverages the more expressive dark term interpretation framework proposed in Seyler et al. [14], which enables us to get a probability distribution over clean words rather than a limited set of categories.

3 METHODOLOGY

3.1 Corpora

For dark jargon interpretation and exploration, the system makes use of five corpora, which are scrapes of four underground forums and one clean forum published by Yuan et al. [17]. *Dark0de* [1] contains 7,417 threads from a hacking technique forum about malware and illicit services. *Hackforums* [2] includes 52,670 threads from a blackhat hacking technique forum. *Nulled* [3] contains 121,499 threads for trading leaked or hacked information. *Silk Road* [5] has 195,403 threads of posts for selling illegal goods, mostly drugs. *Reddit* [4] contains a web scrape of 1.2 million threads from 1,697 top subreddits in terms of the number of subscribers.

3.2 Human-curated Definitions

To create our human-curated dark term definitions, we have a cybersecurity expert gather evidence for the hidden meaning of a candidate dark term. To arrive at our candidate terms, we take the 10,000 most frequent terms of each dark forum, after stopword removal, and randomly sample 10%. We then create the clean term mappings for each candidate term using our interpretation methods (Sections 3.3 and 3.4). For each candidate dark term, the expert searches on the web for the term and its most likely clean term according to both methods and each of the forums. The expert investigates up to 5 search results to find a definition of the dark term. If the candidate dark term is ambiguous or cannot be explained using the search, the expert can also look at individual communication traces that mention the candidate term, as captured in our underground corpora.

3.3 Word Context Distribution Interpretation

For dark jargon interpretation, we leverage the word context distribution method proposed in Seyler et al. [14]. Here, dark terms are

[2]Clean terms are defined as words with no hidden meaning, which can be found in "clear" web corpora or dictionaries, such as Reddit or Wikipedia. For a more detailed definition see Seyler et al. [14].

represented by the global context they appear in, using a context-based word distribution estimated on occurrences in a dark and clean (i.e., Reddit) corpus. The intuition is that terms with hidden meanings will appear in the same context independent of the forum. Therefore, if a dark term appears in very similar contexts than a clean term, there is a strong indication that the meaning of both is the same. Once we have estimated the probability of a dark term w_d given an underground forum corpus C_{dark} as $P(w_d|C_{dark})$ and similarly for a clean term w_c given a clean forum corpus C_{clean} as $P(w_c|C_{clean})$, we arrive at a mapping by minimizing the KL-Divergence as in Equation 1, for a given dark term w_d.

$$\arg\min_{w_c \in C_{clean}} KL(P(w_d|C_{dark})||P(w_c|C_{clean})) \quad (1)$$

3.4 Masked Language Modeling Interpretation

In addition to the word context distribution based method, we present a novel method for the dark jargon interpretation based on Masked language modeling (MLM). As leveraged in Devlin et al. [9], MLM is an unsupervised training objective, where the model needs to predict a hidden (i.e., masked) input token from the input sequence. Thus, the model needs to leverage the context of the masked token within the sequence for this inference. The MLM objective is directly applicable to our problem setup of mapping hidden dark terms to clean terms, since: (1) The meaning of dark terms is context dependent and (2) a pre-trained model, such as BERT [9], is trained on "clean" text (i.e., Wikipedia) and therefore will predict the most likely term with no hidden meaning given its context. Applied to an underground forum setting with a target dark term, we can input the communication trace containing the masked target dark term and the model will predict the most likely token with no hidden meaning.

For a sequence S made up of words w over a vocabulary V, the objective of masked language modeling is to estimate a probability distribution over V for a masked word $m \in S$ as $P(m|S^{\setminus m};\theta)$, where $S^{\setminus m}$ is the word sequence without m (i.e., the context) and model θ. Since BERT is trained on general-purpose corpora, we can assume that if a dark term is fed as a masked token into the model, the model would predict the most-likely clean term, given its context. For a given dark word w_d, we infer $P(w_d|S_i^{\setminus w_d};\theta)$, for every sequence S_i that contains w_d in our underground corpora C. Further, let $rank(w|P(x))$ be a function that returns the rank of a word $w \in V$, ranked according to the probability distribution $P(x)$. Then, we arrive at an aggregate score over all S_i for a target word w as in Equation 2. Finally, we filter out tokens that are not alpha-numeric or are contained in the NLTK [8] stopword list.

$$score^{MLM}(w, w_d) = \sum_{S_i \in C} 1/rank(w|P(w_d|S_i^{\setminus w_d};\theta)) \quad (2)$$

4 SYSTEM FUNCTIONS

4.1 Dark Term Lookup

A common scenario is when a person observes some conversation in an underground forum and does not know the meaning of certain dark terms (e.g., Figure 1). For this scenario, the system provides a dictionary of dark terms and their definitions (Figure 2), which were evaluated by a cybersecurity expert (Section 3.2). The system also lists the URL of where the definition was found, which can be inspected by the user in case more info than the definition is required. In order to explore more details about the meaning and usage of a certain dark term, the user can click on the dark term itself, which guides her to the exploration interface.

4.2 Dark Term Exploration

The exploration interface has two major components: (1) The user can retrieve detailed information about the usage of the selected dark term over time and (2) browse different clean term mapping methods, in order to explore the meaning of a dark term.

Usage Over Time. The user can leverage this feature to get further insights about the usage of a selected dark term in the underground community. The system plots the usage of a term over time (Figure 3), where the x-axis is the temporal dimension and the y-axis is the number of occurrences for a certain time point. By default, the system shows the aggregate usage over all underground forums. If desired, the user can further "drill-down" into specific underground forums, by selecting a forum using the drop-down menu and clicking "update". Then, the plot will interactively update with the forum-specific data.

Meaning Exploration. With this feature the user can explore computational methods for dark term interpretation and therefore get insights into the latent meaning of dark jargon words. Interpretation is done in the form of mapping a dark term to a list of clean terms, according to different mapping methodologies. The interface (Figure 4) shows the most probable words for a selected dark term and mapping method for the combination of all underground forums ("total") and separate forums. Furthermore, the user can select different mapping methods using the drop-down menu and clicking "update". Then, all plots will interactively update with the method-specific data. Available methods are "KL-Divergence" (Section 3.3) and "Masked Language Modeling (BERT)" (Section 3.4).

4.3 Collaboration

DarkJargon.net allows for the collective exploration and collection of dark terms. The system provides a collaboration interface, where users can browse dark terms identified by the community, give feedback and contribute novel dark terms. During the extended existence of the system, community-sourced dark terms will be periodically checked by cybersecurity experts and, once confirmed, added to the main dictionary.

Explore/Validate Existing Contributions. The main part of the collaboration interface is a listing of dark jargon terms that have been discovered by other users of the website (upper part of Figure 5). The listing contains the dark term, its definition and source and the name of the user that submitted the entry. On the right-hand side there are two buttons that can be used to evaluate existing dark terms by giving relevance feedback. More specifically, when browsing the list of collected dark terms, a user can press the "thumbs-up" button to indicate that a dark term is valid or the "thumbs-down" button to indicate that it is not valid. Below the icon, each button shows the count of positive and negative votes for a specific dark term. When experts evaluate the validity of dark terms, they can choose to consider this feedback mechanism in order to make their decisions. Clicking on a dark term will guide the user to the previously introduced exploration interface. As stated before, this

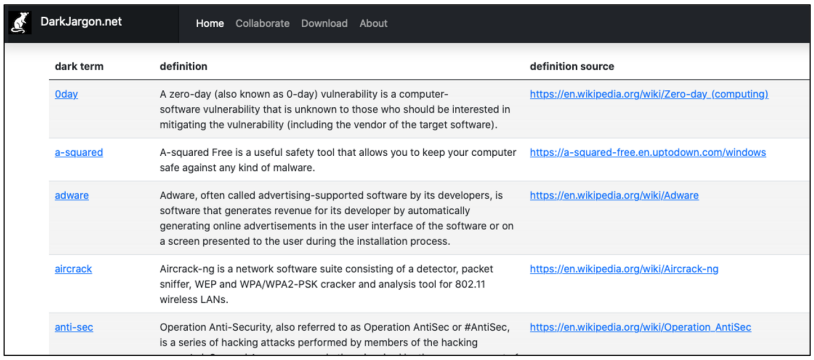

Figure 2: Dictionary listing of expert-evaluated dark terms, their definition and URL to the source of the definition. Clicking on a dark term brings the user to the exploration interface.

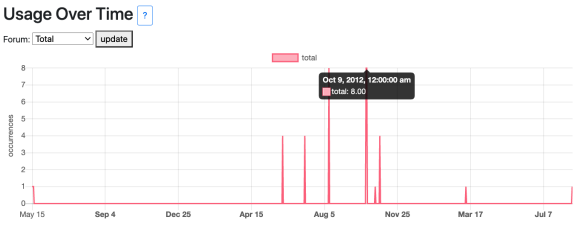

Figure 3: Usage statistics for dark term "msr". Hovering over the time series shows details about specific data points.

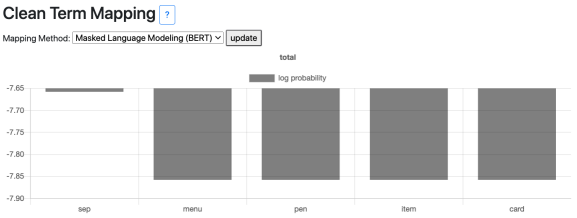

Figure 4: Clean term mapping for dark term "msr", according to our MLM method. Terms are ranked according to the smallest log probability (i.e., the most likely).

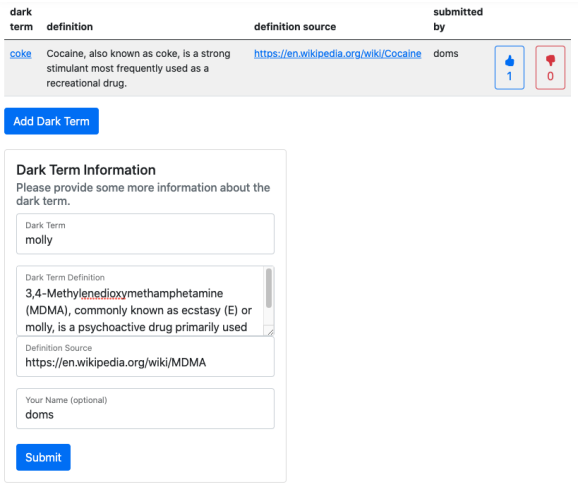

Figure 5: The collaboration interface lists dark terms identified by the community. Users can give feedback by up/down-voting dark terms and contributing novel dark terms.

interface can be used to gain further insights into the usage and meaning of a dark term. For missing data, the system indicates that a dark term has not been observed in our datasets.

Contribute Novel Dark Terms. To contribute to the collection of dark terms, a user can submit a novel dictionary entry. To do this, she can press the "Add Dark Term" button, which opens a form to be filled out (lower part of Figure 5). This form asks to provide the dark term, a short definition, the source of the definition and a user name. After clicking "Submit", the entry will appear in the listing.

5 DEMONSTRATION SCENARIOS

We plan to demonstrate the following scenarios of DarkJargon.net:

1. Dark Jargon Understanding: Using an underground communication trace containing dark terms, we will show how to understand their hidden meaning. We demonstrate this using a lookup of the definition in the dictionary and browsing the definition source website. We gain further understanding of the meaning by examining the clean term mappings in the exploration interface.

2. Dark Jargon Usage: We will show how to retrieve usage statistics over different underground forums for a given dark term using the exploration interface.

3. Dark Jargon Validation: We will browse the list of community contributed dark terms and give feedback on correctness or incorrectness using the collaboration interface.

4. Dark Jargon Contribution: We will demonstrate how to contribute novel dark terms using the collaboration interface.

5. Dataset Download: We will show how researchers can download the DarkJargon.net datasets for their own research.

ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant No. 1801652.

REFERENCES

[1] [n.d.]. Dark0de (forum). https://en.wikipedia.org/wiki/Dark0de
[2] [n.d.]. Hackforums. https://hackforums.net
[3] [n.d.]. Nulled (forum). https://www.nulled.to
[4] [n.d.]. reddit (forum). https://www.reddit.com
[5] [n.d.]. Silk Road (marketplace). https://en.wikipedia.org/wiki/Silk_Road_(marketplace)
[6] [n.d.]. Urban Dictionary. https://urbandictionary.com
[7] Drug Enforcement Administration. 2018. Slang Terms and Code Words. https://www.dea.gov/documents/2018/07/01/2018-slang-terms-and-code-words
[8] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit.* " O'Reilly Media, Inc.".
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*.
[10] National Initiative for Cybersecurity and Studies. [n.d.]. Cybersecurity Glossary. https://niccs.cisa.gov/about-niccs/cybersecurity-glossary
[11] Ryan Graham. 2020. Track2 Character Encoding. https://medium.com/@ryandeangraham/track2-character-encoding-61d6150ecb95
[12] HarperCollins Publishers. [n.d.]. Definition of 'MSR'. https://www.collinsdictionary.com/us/dictionary/english/msr
[13] Eric S. Raymond. [n.d.]. The Jargon File. http://catb.org/jargon/
[14] Dominic Seyler, Wei Liu, XiaoFeng Wang, and ChengXiang Zhai. 2021. Towards Dark Jargon Interpretation in Underground Forums. In *European Conference on Information Retrieval*.
[15] Hao Yang, Xiulin Ma, Kun Du, Zhou Li, Haixin Duan, Xiaodong Su, Guang Liu, Zhifeng Geng, and Jianping Wu. 2017. How to learn klingon without a dictionary: Detection and measurement of black keywords used by the underground economy. In *Symposium on Security and Privacy (SP)*. 751–769.
[16] Kan Yuan, Haoran Lu, Xiaojing Liao, and Xiaofeng Wang. 2018. CantReader Code Release. https://sites.google.com/view/cantreader
[17] Kan Yuan, Haoran Lu, Xiaojing Liao, and Xiaofeng Wang. 2018. Reading Thieves' Cant: Automatically Identifying and Understanding Dark Jargons from Cybercrime Marketplaces. In *USENIX Security Symposium*.