Towards Dark Jargon Interpretation in Underground Forums

Dominic Seyler
dseyler2@illinois.edu

Wei Liu
weil8@illinois.edu

XiaoFeng Wang
xw7@indiana.edu

ChengXiang Zhai
czhai@illinois.edu

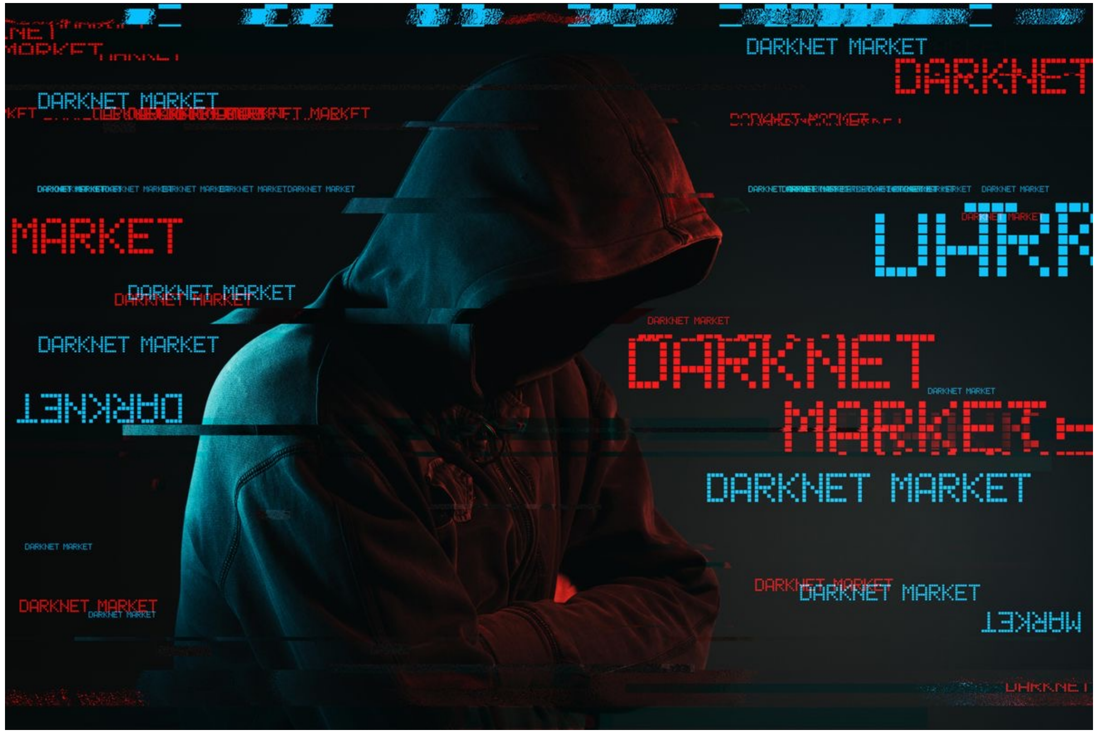
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

What is *dark jargon*?

- **Dark jargon:** benign-looking words that have hidden, sinister meanings.
- Used by participants in underground forums to hide their illicit activities.
- “rat”→ Remote Access Trojan (acronym)
- “popcorn” → type of marijuana (short for ”popcorn nugs”)
- “ecstasy” → MDMA (“Methylenedioxymethamphetamine”)

Motivation

- Identifying real meaning is essential for understanding cybercrime facilitated through social media platforms.
- Previous work has been successful at detection, but interpretation is still an open issue.



General Framework

- Idea: use words with no hidden meaning (i.e., ”clean” words) as direct explanation of dark jargon words (i.e., “dark” words).
- Create mapping: $hidden\_meaning(V_{dark}) \ \rightarrow \ V_{clean}$
  - Find mapping for each dark word w_d ∈ V_{dark} to clean word vocabulary V_{clean}
- Mapping can be probabilistic where for each w_d we obtain a probability distribution over clean words w_c ∈ V_{clean}

Problem Setup

- Concrete instantiation of the general framework:
  - Given dark corpus C_{dark} and clean corpus C_{clean}
  - Build joint vocabulary V, which is the most frequent words in $C_{dark} \cup C_{clean}$
  - For each dark word $w_d \in C_{dark}$, find a clean word $w_c \in C_{clean}$ that expresses the hidden meaning of w_d.
- We propose two methodologies to achieve this mapping:
  1. Word distribution modeling and Kullback-Leibler-divergence (KL-divergence).
  2. Cross-context Lexical Analysis (CCLA) [1].

[1] Massung 2017. "Beyond topic-based representations for text mining". Ph.D. dissertation.

Word Distribution Modeling and KL-Divergence

- Intuition:
  - Dark word (“rat”) will appear in different context than the clean word “rat”.
  - It’s context will be more similar to the clean word “malware”, as to “mouse”.
- Build word distribution for each word $w \in V$ using “sliding-window”.
  - This is done separately for C_{dark} and C_{clean} to estimate two probability distributions $P(w_d |\, C_{dark})$ and $P(w_c |\, C_{clean})$.

$P(\text{“rat”} \mid C_{dark})$	$P(\text{“rat”} \mid C_{clean})$	$P(\text{“malware”} \mid C_{clean})$
…	…	…
computer 0.0083	scared 0.0033	anti-virus 0.0073
windows 0.0071	running 0.0029	computer 0.0069
…	…	…

Word Distribution Modeling and KL-Divergence

- Get dissimilarity of two words using KL-Divergence:

$$dissim(w_d, w_c) = KL(P(w_d|C_{dark})||P(w_c|C_{clean}))$$

- For each w_d, define its hidden meaning:

$$\arg\min_{w_c \in C_{clean}} dissim(w_d, w_c)$$

Cross-context Lexical Analysis

- **CCLA**: analyze differences and similarities of words across different contexts. Contexts are defined over document collections.
- Use word embeddings trained on separate corpora (clean and dark in our setting).
- Compute the similarity of the top-k closest neighbors of a word in separate corpora.
- We modify CCLA slightly: measure similarity of a pair of words.
- For each w_d, we maximize the similarity according to CCLA over each clean word $w_c \in C_{clean}$

Experimental Setup

- Datasets:
  - Dark Corpus: 376,989 posts scraped from four major underground forums [1].
  - Clean Corpus: 1.2 million reddit threads from the most popular subreddits.
- *clean-clean* Environment (simulated):
  - Split clean corpus and replace random words (simulated dark terms) with individual placeholders.
  - Map placeholder in split 1 to original word in split 2.
  - Measure mean reciprocal rank (MRR) of placeholder and original word.
- *dark-clean* Environment (real-world):
  - Use dark corpus instead of simulated clean corpus.
  - Map each word in the dark corpus to a word in the clean corpus.
  - Perform manual evaluation of word meanings.

[1] Yuan et al. 2018. "Reading thieves' cant: Automatically identifying and understanding dark jargons form cybercrime marketplaces." In USENIX

Experimental Results

- Setup: clean-clean
- Measure MRR for all words in corpus and simulated dark terms.

Method	MRR all words	MRR dark words
KL	0.909	**0.892**
CCLA	**0.974**	0.479

- KL method performs well, outperforming CCLA for dark words.
- CCLA performs better on all words, but much worse for dark words.
- Insight: KL method is preferable for detecting/interpreting dark words.

Manual Evaluation

	Dark Word	Clean Word	Meaning
drugs	gdp	kush	Grand Daddy Purps (type of marijuana)
	blueberry	kush	type of marijuana
	coke	cocaine	nickname for cocaine
	klonopin	xanax	sedative medication
	shrooms	lsd	hallucinogenic drug similar to LSD
	bubba	kush	type of marijuana
	ecstasy	mdma	nickname for mdma
	dilaudid	oxy, morphine	strong painkiller (aka: hospital heroin)
	pineapple	kush	type of marijuana
malware	zeus	botnet	botnet malware
	rat	malware	Remote Access Trojan (malware)

Summary

Dark Jargon	• Benign-looking words that have hidden, sinister meanings.
General Framework	• Use words with no hidden meaning (i.e., ”clean” words) as direct explanation of dark jargon words (i.e., “dark” words).
Word Distribution Modeling	• Intuition: Word meaning is context dependence. • Use differences in context distributions for mapping.
Results	• 0.89 MRR on simulated dataset. • Manual evaluation shows efficacy of method.

Thank You!

more info at: https://dominicseyler.com

Dominic Seyler

dseyler2@Illinois.edu

Wei Liu

weil8@Illinois.edu

XiaoFeng Wang

xw7@Indiana.edu

ChengXiang Zhai

czhai@Illinois.edu

UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN