Cybersecurity researchers build a better 'canary trap'

March 1 2021

Credit: CC0 Public Domain

During World War II, British intelligence agents planted false documents on a corpse to fool Nazi Germany into preparing for an assault on Greece. "Operation Mincemeat" was a success, and covered the actual

Allied invasion of Sicily.

The 'canary trap' technique in espionage spreads multiple versions of false documents to conceal a secret. Canary traps can be used to sniff out information leaks, or as in WWII, to create distractions that hide valuable information.

WE-FORGE, a new data protection system designed at Dartmouth's Department of Computer Science, uses artificial intelligence to build on the canary trap concept. The system automatically creates false documents to protect intellectual property such as drug design and military technology.

"The system produces documents that are sufficiently similar to the original to be plausible, but sufficiently different to be incorrect," said V.S. Subrahmanian, the Distinguished Professor in Cybersecurity, Technology, and Society, and director of the Institute for Security, Technology, and Society.

Cybersecurity experts already use canary traps, "honey files," and foreign language translators to create decoys that deceive would-be attackers. WE-FORGE improves on these techniques by using natural language processing to automatically generate multiple fake files that are both believable and incorrect. The system also inserts an element of randomness to keep adversaries from easily identifying the real document.

WE-FORGE can be used to create numerous fake versions of any technical design document. When adversaries hack a system, they are faced with the daunting task of figuring out which of the many similar documents is real.

"Using this technique, we force an adversary to waste time and effort in

identifying the correct document. Even if they do, they may not have confidence that they got it right," said Subrahmanian.

Creating the false technical documents is no less daunting. According to the research team, a single patent can include over 1,000 concepts with up to 20 possible replacements. WE-FORGE can end up considering millions of possibilities for all of the concepts that might need to be replaced in a single technical document.

"Malicious actors are stealing intellectual property right now and getting away with it for free," said Subrahmanian. "This system raises the cost that thieves incur when stealing government or industry secrets."

The WE-FORGE algorithm works by computing similarities between concepts in a document and then analyzing how relevant each word is to the document. The system then sorts concepts into "bins" and computes the feasible candidate for each group.

"WE-FORGE can also take input from the author of the original document," said Dongkai Chen, a graduate student at Dartmouth who worked on the project. "The combination of human and machine ingenuity can increase costs on intellectual-property thieves even more."

As part of the research, the team falsified a series of computer science and chemistry patents and asked a panel of knowledgeable subjects to decide which of the documents were real.

According to the research, published in *ACM Transactions on Management Information Systems*, the WE-FORGE system was able to "consistently generate highly believable fake documents for each task."

Unlike other tools, WE-FORGE specializes in falsifying technical information rather than just concealing simple information, such as

passwords.

WE-FORGE improves on an earlier version of the system—known as FORGE—by removing the time-consuming need to create guides of concepts associated with specific technologies. WE-FORGE also ensures that there is greater diversity among fakes, and follows an improved technique for selecting concepts to replace and their replacements.

Almas Abdibayev, Deepti Poluru Guarini and Haipeng Chen all contributed to this research while with Dartmouth's Department of Computer Science.

More information: Almas Abdibayev et al, Using Word Embeddings to Deter Intellectual Property Theft through Automated Generation of Fake Documents, *ACM Transactions on Management Information Systems* (2021). DOI: 10.1145/3418289