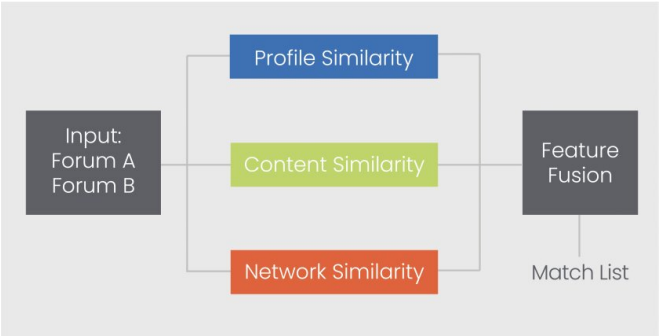
Artificial intelligence shines light on the dark web

14 May 2019, by Kylie Foy



To match users from different forums who are likely the same person, an algorithm calculates similarities in profiles, such as their usernames; in content, such as similar phrasings; and in their network, such as the community with which they interact. Credit: Massachusetts Institute of Technology

Beneath the surface web, the public form of the internet you use daily to check email or read news articles, exists a concealed "dark web." Host to anonymous, password-protected sites, the dark web is where criminal marketplaces thrive in the advertising and selling of weapons, drugs, and trafficked persons. Law enforcement agencies work continuously to stop these activities, but the challenges they face in investigating and prosecuting the real-world people behind the users who post on these sites are tremendous.

"The pop-up nature of dark-web marketplaces makes tracking their participants and their activities extremely difficult," says Charlie Dagli, a researcher in MIT Lincoln Laboratory's Artificial Intelligence Technology and Systems Group. Dagli is referring to the fast rate at which dark-web markets close down (because they are hacked, raided, abandoned, or set up as an "exit scam" in which the site shuts down intentionally after customers pay for unfulfilled orders) and new ones appear. These markets' short lifetimes, from a few months to a couple years, impede efforts to identify their users.

To overcome this challenge, Lincoln Laboratory is developing new software tools to analyze surface- and dark-web data.

These tools are leveraging the one benefit this whack-a-mole-like problem presents—the connections sellers and buyers maintain across multiple layers of the web, from surface to dark, and across dark-web forums. "This constant switching between sites is now an established part of how dark-web marketplaces operate," Dagli says.

Users are making new profiles constantly. Although they may not be employing the same usernames from site to site, they are keeping their connections alive by signaling to each other through their content. These signals can be used to link personas belonging to the same user across dark-web forums and, more revealingly, to link personas on the dark web to the surface web to uncover a user's true identity.

Linking users on the dark web is what law enforcement already tries to do. The problem is that the amount of data that they need to manually shuffle through—500,000 phone numbers and 2 million sex ads posted a month—is too large and unstructured for them to find connections quickly. Thus, only a low percentage of cases can be pursued.

To automate the persona-linking process, Lincoln Laboratory is training machine learning algorithms to compute the similarity between users on different forums. The computations are based on three aspects of users' communications online: "How they identify to others, what they write about, and with whom they write to," Dagli explains.

The algorithm is first fed data from users on a given

Forum A and creates an authorship model for each user. Then, data from users on Forum B are run against all user models from Forum A. To find matches for profile information, the algorithm looks for straightforward clues, such as changes in username spelling like "sergeygork" on Forum A to "sergey gorkin" on Forum B, or more subtle similarities like "joe knight" to "joe nightmare."

The next feature the system looks at is content similarity. The system picks up on unique phrases—for example, "fun in the sun"—that are used in multiple ads. "There's a lot of copy-and-paste going on, so similar phrasings will pop up that are likely from the same user," Dagli says. The system then looks for similarities in a user's network, which is the circle of people that the user interacts with, and the topics that the user's network discusses.

The profile, content, and network features are then fused to provide a single output: a probability score that two personas from two forums represent the same real-life person.

The researchers have been testing these persona-linking algorithms both with open-source Twitter and Instagram data and hand-labeled ground truth data from dark-web forums. All of the data used in this work are obtained through authorized means. The results are promising. "Every time we report a match, we are correct 95 percent of the time. The system is one of the best linking systems that we can find in the literature," Dagli says.

This work is the most recent development in ongoing research. From 2014 to 2017, Lincoln Laboratory contributed to the Defense Advanced Research Projects Agency (DARPA) Memex program. Memex resulted in a suite of surface- and dark-web data analysis software developed collaboratively with dozens of universities, national laboratories, and companies. Ten laboratory technologies spanning text, speech, and visual analytics that were created for Memex were released as open-source software via the DARPA Open Catalog.

Today, more than 30 agencies worldwide are using Memex software to conduct investigations. One of the biggest users, and a stakeholder in Memex's development, is the Human Trafficking Response Unit (HTRU) in the Manhattan District Attorney's Office.

Manhattan District Attorney Cyrus Vance Jr. stated in a written testimony to the U.S. House of Representatives that his office used Memex tools to screen more than 6,000 arrests for signs of human trafficking in 2017 alone. "We also used Memex in 271 human trafficking investigations and in six new sex trafficking indictments that were brought in 2017," he stated. With the introduction of Memex, prostitution arrests screened by HTRU for human trafficking indicators increased from 5 to 62 percent, and investigations of New York Police Department prostitution-related arrests increased from 15 to 300 per year.

Jennifer Dolle, the deputy chief of HTRU, visited the laboratory to present how the unit has benefited from these technologies. "We use these tools every single day. They really have changed how we do business in our office," Dolle says, explaining that prior to Memex, a human trafficking investigation could take a considerably longer time.

Now, Memex tools are enabling HTRU to quickly enhance emerging cases and build sex trafficking investigations from leads that have little information. For example, these tools—including one called TellFinder (built by Memex contributor Uncharted Software) for indexing, summarizing, and searching sex ad data—have been used to identify additional, underage victims from data in a single online prostitution advertisement. "These additional investigative leads allow HTRU to prosecute traffickers on violent felony charges and hold these defendants responsible for the true nature of the crimes they commit against vulnerable victims," says Dolle.

Researchers are continuing to learn how emerging technologies can be tailored to what agencies need and for how the dark web operates. "Data-driven machine learning has become a demonstrably important tool for law enforcement to combat illicit online marketplaces on the dark web," says Lin Li, a principal investigator of this continuous work in the laboratory's Human Dynamic Dark Networks program, which is funded through the laboratory's

Technology Office. "But, some of the ongoing challenges and areas of research include expanding our understanding of the demand economy, disrupting the supply economy, and gaining a better overall situational awareness."

A better understanding of how the supply-and-demand chains of the dark-web economy work will help the team develop technologies to disrupt these chains. Part of the goal is to raise the risks of participating in this illicit economy; linking personas on the dark web to those on the surface web is one potentially powerful tactic.

"This fast-growing illicit economy was shown by DARPA to fund terrorist activities and shown by HTRU as a driver of modern-day slavery. Defeating terrorism and eliminating slavery are national and humanitarian needs," says Joseph Campbell, leader of the Artifical Intelligence Technology and Systems Group. "Our group has extraordinary expertise in AI, machine learning, and the analysis of human networks based on information extracted from multilanguage speech, text, and video combined with network communications and activities. The state-of-the-art technologies that we create, develop, and advance are transferred to our sponsors, who use them daily with tremendous impact for these national and humanitarian needs."

This story is republished courtesy of MIT News (web.mit.edu/newsoffice/), a popular site that covers news about MIT research, innovation and teaching.

Provided by Massachusetts Institute of Technology