Cyber Frauds and Fakes / Summer 2018

Dongwon Lee, Ph.D. (dongwon@psu.edu)

Team Assignment (60 points)

DUE: June 28 Thursday 6:00PM

Goal: The aim of this assignment is for students to demonstrate their ability to (1) analyze a challenging cyber fraud problem and (2) develop an effective data-driven solution to address the fraud problem.

Deliverable: For one of the problems below, develop your data-driven solution on-the-paper, describing an idea without any experimentation or implementation. While you do not need to demonstrate the effectiveness of your solution by experimenting or implementing it, still, your design should include full details to the extent that another person is able to implement the solution based on your design report. Your design may address some of the following aspects:

- How to formulate and model the problem precisely using which computational or machine learning framework?
- Does the problem occur already? Else, when is the problem likely to occur? In what applications or scenarios?
- Are there related solutions? Has the problem been solved already? What are their pros/cons?
- How to identify or collect right dataset? Which of existing datasets to use, or how to create new one? How to design training, validation, and/or test dataset?
- How to pre-process the collected dataset?
- How to evaluate the model?

At the end, submit: (1) a PDF report in English (limit is 3 pages) of your team’s solution, and (2) a PPT presentation in English (limit is 10 pages including all) to present your team’s solution in class. Use a reasonable format in your report and presentation.

Grading: A team of 4 students work together in this assignment. At the end, all students in the same team would receive the same points. Grading will be done based on how innovative and plausible your solution is. Due to time constraint, only a selected number of teams are invited to present on Day 5. If so, the teams will use the PPT that they have submitted to present their solutions.

Copyright: Do NOT use the ideas of these problems and their descriptions outside of this class.
Problem #1: Reverse Turing Test for Chatbot

Chatbot or smart assistants such as Apple's Siri or Amazon's Alexa are quickly gaining popularity, and many people view them as one of top consumer applications for AI. The gist of such chatbot related technologies is an AI engine to simulate human-like conversation. If human cannot tell if she is conversing with another human or chatbot, then such chatbot has succeeded in simulating human-like conversation, and can rightfully claim that it has passed the Turing Test. However, imagine a scenario where an adversary is attacking people using such a human-like chatbot, pretending to be real human, in conversation. Then, if one can accurately tell if the other party in conversation is human or chatbot, such an information can be useful to avoid potential online frauds.

Research Question (RQ): Given a transcript of conversation between two unknown parties A and B as an input, design a data-driven solution to determine if A or B is human or chatbot (i.e., binary classification). Or, one of 3 types: human-human, human-machine, or machine-machine (i.e., multi-class classification).

Problem #2: Challenge-Response for Phishing Attacks

Dodge et al\textsuperscript{1}. reported an interesting experiment done in the West Point in 2004. In the experiment, researchers sent a phishing email to 512 cadets, pretending it to be coming from a fictitious Colonel, asking them to click a malicious link regarding a grade change problem. Surprisingly, 80\% of cadets clicked the link, without checking the authenticity of the email or the sender therein. As many cadets simply trusted the sender for being a Colonel, the click rate was unusually high. To make the matters worse, if the sender of the email in the experiment was recipient’s immediate supervisor (i.e., spear phishing from an inside attacker), then even the remaining 20\% of cadets might have fallen for such an attack. In general, such phishing attacks are

extremely difficult to verify if the sender is really who she claims to be, or an attacker with malicious purposes (pretending to be someone else). If the content of an email does not contain a useful clue (e.g., a bogus link) but only asks for sensitive information urgently, determining if such an email is an attack or not becomes even harder.

One potential solution to address such spear phishing is to "ask-back" to the sender, requesting her to prove her identity. If the sender can answer a question that only a close friend or supervisor would know, then one may consider the email as "safe" and the sender as a friend, not an attacker. We envision that such a technique be useful for defending against social engineering attacks such as spear phishing, where it becomes critical of: (1) knowing if the sender of an email is a friend or attacker, and/or (2) delaying an attack by forcing attackers to solve challenge questions has critical values.

**Research Question (RQ):** Design a challenge-response authentication technique that exploits latent knowledge hidden in a user's past emails as a corpus to auto-generate challenge questions that are: (1) hard for attackers to guess, (2) easy for honest senders to respond, and (3) cheap for machines to verify.

---

**Problem #3: Measuring Real Impacts of Fake News**

Many existing socio-technical methods toward fake news, by and large, currently focus on the issue of "correlation" and "discriminative models"—e.g., how to build a machine learning model that learns latent patterns among false articles, or how to detect bot accounts from the information propagation patterns. We argue that an important missing piece in this picture is the "causation"—what triggers and why? That is, reading and sharing a tweet with false news is one thing, but attending a protest or buying a gun after being exposed to fake news is entirely different and more serious matter. What prompts users who read a false article to go out and do a real action? Can one identify characteristics of such susceptible users to fake news, and build a machine learning model to predict them?

**Research Question (RQ):** Design a solution that can identify susceptible users to fake news, monitor them in social media, and alert when they are about to engage in actions related to the fake news.