Use of Interpersonal Deception Theory in Counter Social Engineering

Grace Hui Yang
The InfoSense Group
Department of Computer Science
Georgetown University
huiyang@cs.georgetown.edu

Yue Yu
The InfoSense Group
Department of Computer Science
Georgetown University
yy476@georgetown.edu

ABSTRACT

Social engineering attacks exploit human vulnerabilities rather than computer vulnerabilities. Ranging from straightforward spam emails to sophisticated context-aware social engineering, social engineering has demonstrated rich varieties. Surprisingly, even the simplest type of attacks are able to fool numerous innocent people. The more sophisticated ones are even more "successful" in achieving their malicious purposes. In order to mitigate and combat these attacks, we need better automated counter social engineering algorithms and tools. In this position paper, we propose a reinforcement learning framework that incorporates interpersonal deception theory to fight against social engineering attacks on social media sites.

KEYWORDS

Interpersonal Deception Theory, Reinforcement Learning, Social Engineering

1 INTRODUCTION

Social engineering attacks have become increasing threats in online life. These attacks are made possible by the rapid increase in usage of social networks, mobile devices, and web platforms. Social engineering attacks are maliciously planned by online criminals. Social engineering attacks, sometimes also called phishing attacks, aim to collect sensitive and personal information about users, including identification numbers, passwords, and bank account information [1, 8] via emails, chatrooms, dating sites and other forms of social media.

Early social engineering attacks include only cosmetic deceptions where the appearance of a graphical user interface (GUI) are imitated. For instance, part of a social media platform user interface may contain a component that could be used for social engineering. The victims trust that these components on the GUI are used as they intended. However, the malicious attacks exploit this trust by showing the known components but embedding malicious codes in them to obtain users' sensitive information.

Recent attacks have become more sophisticated. They focus more on mimicking a legitimate system's behavior or a legitimate user's behavior rather than imitating GUI components. Online social networks, such as Facebook, Twitter, LinkedIn, as well as dating sites, are used to chat with the victims first, gain their trusts, and then obtain their sensitive information via these interpersonal communications. There are also hybrid attacks that combine both aspects of "looking like" and "behaving like" known values to make the deception even more convincing.

Social engineering is within the scope of interpersonal deception [3]. It is a type of interpersonal communication where the messages knowingly transmitted by a sender to foster a false belief or conclusion by the receiver and to obtain sensitive or personal information of the receiver. It belongs to the type of strategic behaviors with a clear goal-oriented nature. The deception happens when communicators control the information contained in their messages to convey a fake meaning. In this research, we propose to employ general theories and understandings developed in **Interpersonal Deception Theory** (IDT) [3] for modeling and combating social engineering from a broader scope and a deeper level.

Detecting social engineering attacks can be complex. Many counter social engineering methods share a few common stages. The first stage is planning or orchestration. The following questions are asked – "how is the victim or the target chosen?" "How does the attack reach the target?" and "Is the attack automated?". The second stage is about "Is it behavior that deceives the target?" and "Does the deception occur in the system or external?". The third stage cares about "Does the deception take one step or multiple steps?" "Is it persistent?" [6]. Counter social engineering is a type of task that requires good understanding of the cognitive and behavioral patterns of the criminals and that of the victims.

Interpersonal communications are governed by a set of cognitive and behavioral theories. As a subtype of interpersonal deception, social engineering makes no exception. In this position paper, we propose to make use of known theories on interpersonal deception and model them into a multi-agent reinforcement learning framework. We aim to bridge the understandings in psychology with modern machine learning algorithms and tools. We make use of the influence during interactions, pre-interactions and post-interactions among the victims, the social engineering attackers and our counter social engineering agents.

The proposed reinforcement learning framework is flexible. We can design the states and actions to reflect the factors that have studied and proved to be useful in IDT. It would be important to model completeness, directness, knowledge, clarity of the messages as well as to model the personality, vulnerability, arousal, negative affect, cognitive effort, suspicion, and attempted control of the message senders. Inspired by IDT, we also explore the impact of context (e.g. personal and contextual data about individuals) and relationships (e.g. friend network) to social engineering. We not

CIKM'18, The Second International Workshop on Rumours and Deception in Social Media Workshop, October 22–26, 2018, Turin, Italy

only model them but also implement tools to collect context and make use of social network information to both detect social engineering attacks and generate counter social engineering messages and activities to investigate the attackers.

In this position paper, we discuss the possibilities for modeling theories about interpersonal deception to create new counter social engineering strategies for underlying artificial intelligence (AI) and machine learning (ML) algorithms. The research is proposed to be built on top of a multi-agent reinforcement learning framework with the capabilities to model and use interpersonal deception theories for combating the attacks.

2 RELATED WORK

Previous research has investigated the psychological side of the problem – why phishing works [4]. The largest factors have to do with people's misconceptions about computer security and attack templates. There is a general lack of knowledge about computer systems, computer security, and security indicators (or the absence of security indicators). Many social media users have only basic or even incorrect assumptions and heuristics when deciding how to respond to emails or chat messages asking for sensitive information. For instance, some assume that once a business already has their personal information, it is safe to give it again.

As a result, educating social media users about making the right decisions when receiving social engineering attacks is a very important component in preventing such attacks. Unfortunately, most existing approaches only focus on awareness training and hope users make better decisions the next time they are faced with a questionable email or chat message. The effectiveness of this strategy is limited. Instead, our research focuses on performing automated active detection and intervention for potential attacks, relieving users of the pressure of self-protection.

The automated approaches include sandboxing, authorisation-authentication-accounting (AAA), monitoring via Honeypots, integrity checking, machine learning [6]. Those countermeasures aim to prevent and detect attacks before and after the victim data are collected and used [1].

Existing machine learning techniques used for counter social engieering have focused on classification algorithms. Both linear and non-linear supervised machine learning models are used to make the decision on whether an email or a website is social engineering. Algorithms such as Support Vector Machines (SVM), Naïve Bayes, and k-Nearest Neighbor are widely used. The challenging here is to identify the good features that are able to distinguish normal emails and text messages from the engineering ones.

We summarize a list of popular features and cues used in the machine learning appraoches:

- URL features such as IR address characteristics, geographic properties, domain names. [1]
- Content-based features which examine how suspicious the content is, e.g. asking for money, asking for a bank account, asking for a password. [1]
- Document structure features, including a web page's main page, component files, DOM structures etc.
- Linguistic cues for deceptions, such as the length of unique words, the length of sentences, word diversity, type-token ratio, six-letter words, the number of verbs being used, tentative words, modal verbs. [5]
- Complexity of language use, such as exclusive words, causation relations, certainty, negations, negative emotions including anger, sadness words and pleasant and unpleasantness words. [5]

Unsupervised approaches such as clustering, mining and statistical language models are also used in social engineering detection. For instance, simple techniques such as term frequency and inverse document frequency (TF-IDF), regular expressions representing social engineering text patterns, as well as latent semantic analysis (LSA) [1], are still used in social engineering attack detection. Zhou and Shi have shown in [12] that with two n-gram statistical language models (SLM), one using deception data and the other using legitimate data, together with the Kneser-Ney smoothing technique, the language modeling approach can outperform SVM. In this proposed research, we fully explore the features and cues, including n-grams, and take advantage of already proposed features in the literature.

There are also patents invented for detecting and fighting against social engineering attacks [7, 11]. Some patents propose developing decoy systems to trap the attackers. The decoy systems contain hardware components as well as decoy documents and other digital information. They have a more realist understanding of how a deception system works. For instance, a deception system could generate receipts, tax documents, and other form-based documents with credentials, names, emails, addresses or login information collected online or within an organization [11]. These patents are quite complex in terms of their designs. Even though no effectiveness metrics are reported, these approaches sound quite practical. However, within each component of these patented system, the detailed features do not seem as effective as what has been studied in the research community. For example, the linguistic features mentioned in those patents are quite naive. Keywords such as "top secret" and "privileged" are hard-coded into the decoy system and no advanced machine learning techniques nor more flexible methods are used to make the patented system scalable to large scales.

Spear phishing is the form of social engineering that deceives the victims by creating emails, text messages or chats with context relevant to the victim. The relevant content is collected from the Internet. An adversary can digitally "stalk" a victim (a Web user) and discover as much information as possible about the victim, either through direct observation of posted information or by inferring knowledge using simple inference logic. Such knowledge includes a person's race, relationship status, estimated income level, and religion [9, 10]. Current technologies for counter spear phishing are still in its infancy. Most existing techniques overlap largely with the privacy community and data mining community in terms of understanding how the contextual information is crawled and collected for an individual from publicly available online data.

In the Artificial Intelligence community, research in dialogue-based systems and adversarial search are relevant to counter social engineering in terms of their common goal of being interactive and adaptive. However, there is no prior study on counter social enineering in this context. The work by Banerjee and Peng [2] is perhaps the most similar to what we propose here. They proposed

a multi-agent reinforcement learning framework for countering deception. However, their work is in the domain of gaming and adversarial search. Moreover, they do not show how to incorporate existing strategies into a reinforcement learning algorithm, which is our focus.

3 MULTI-AGENT REINFORCEMENT LEARNING

This research develops a general reinforcement learning framework for modeling two teams of agents, the social engineering attackers, and counter social engineering agents, into one interactive, dynamic environment. We propose elements, framework, placeholders for interpersonal deception theories, and ways to model human-programmable policies for modeling and combating social engineering attacks.

Multi-agent learning (MAL) lies at the intersection of distributed artificial intelligence and reinforcement learning. A multi-agent system (MAS) contains multiple agents as its name suggests. Agents in a MAS typically operate in large, complex, dynamic and unpredictable environments which is a key difference between MAL and typical supervised machine learning. MAL is also an area where game theory meets with reinforcement learning. Game theory has been extensively studied for adversarial search in artificial intelligence as well as for concurrent reinforcement learning. Most algorithms for multi-agent reinforcement learning have been proposed mostly in the space of stationary environment. That is, one agent is explicitly formulated based on stationary policies for self-play and the other agents are following stationary policies and assuming explicit knowledge of the agent and the domain. Nonetheless, there are quite few MAL algorithms available.

In the problem of counter social engineering, there are multiple agents. There are one or more attackers. There are also one or more victims. They are the agents in the game. Here we also assume a clear division of attackers and victims as two sides of the agents. Different from the dynamic search problem that we have just mentioned, counter-deception is uncooperative, which makes it closer to the traditional AI problem of adversarial search where both teams of players would like to win the other team. When we model the counter social engineering problem, its uncooperative nature needs to be taken into account and to be properly represented in the model.

In the multi-agent stochastic game (SG), we propose it to be a tuple $< S, A_{ph}, A_c, f, R_{ph}, R_c >$, where S is the discrete set of states, A_{ph} is the set of actions that the phishers take, A_c is the set of actions that the counter social engineering agents take. Both actions A_{ph} and A_c yield a joint action set $A = A_{ph} \times A_c$. f is the state transition probabilistic function and is defined over $S \times A \times S \rightarrow [0, 1]$. The phisher reward function R_{ph} is defined over $S \times A_{ph} \times S \rightarrow \mathbb{R}$ and counter social engineering agent reward function R_c is defined over $S \times A_c \times S \rightarrow \mathbb{R}$.

States S is a discrete set of states.

Actions A is a discrete set of actions that an agent can take. For instance, the criminal's actions include searching for a name and collecting context for an individual.

Observations Ω is a discrete set of observations that an agent makes about the states. O is the observation function which represents a probabilistic distribution for making observation o given action a and landing in the next state s'.

Transitions T is the state transition function $T(s_i, a, s_j) = Pr(s_i, a, s_j)$ ranging from 0 to 1. It is the probability of starting in state s_i, taking action a, and ending in state s_j. The sum over all actions give the total state transition probability $T(s_i, s_j) = Pr(s_i, s_j)$.

Reward $r = R(s, a)$ is the immediate reward, also known as *reinforcement*. It gives the expected immediate reward of taking action a at state s. An agent in an MDP usually maximizes its own long-term reward.

Long term reward is the sum of all past and future rewards in the entire process: $\sum_{t=1}^{\infty} r$. It can be optionally discounted for the future states: $\sum_{t=1}^{\infty} \gamma^t r_t$, where γ is the discount factor.

A **policy** π describes the behaviors of an agent. A non-stationary policy is a sequence of mapping from states to actions. It is also the solution that we usually seek in a Markov Decision Process. A policy π makes a decision that which action should be taken for a state. We optimize π to decide how to move around in the state space in order to optimize the long-term reward $\sum_{t=1}^{\infty} r$ in the entire process. The policy studies $\pi : S \rightarrow A$, such that π optimizes the long-term reward that is represented in a value function V.

The goal for counter social engineering can be defined as the following: select or suggest the most suitable strategy π_c for counter social engineering agent c to best fulfill the long-term expected rewards. Given that different counter social engineering strategies demonstrate significantly different algorithms and result representations, this research concentrates on how to defining the elements of the multiple-agent reinforcement learning framework: its states, actions, rewards, etc.

4 MATHEMATICAL MODELING

The research proposed is a new attempt for studying interactions in the interpersonal deception process. Here we assume a game between the two groups of agents, the social engineering attackers, and the counter social engineering agents. When the game turns into the case that the agents have different goals, it is very interesting for us to see how to detect the social engineering attacks and perform effective counter social engineering activities. In this research, we focus on how to represent theories in interpersonal deception into policies that a machine can understand and execute.

A multi-agent reinforcement learning algorithm is modeled as a stochastic game with a set of joint actions $A = A_1 \times A_2 \times A_3 \times ... A_n$, where each set of actions A_i is the possible actions of agent i. The goal of the i^{th} agent is assumed to find a strategy or a policy π_i which maximizes the agent's expected sum of discounted long term rewards for state s, i.e. the value function,

$$v_{\pi_i}(s) = \sum_{i=0}^{\infty} \gamma^i E(r_t^i | \pi_i, \pi_{-i}, s_0 = s)$$

where r_t^i is the reward for the i^{th} agent at time t, s_0 is the initial joint state, π_{-i} is the strategy of the i^{th} agent's opponent, and γ is the discount factor.

Here we propose a general framework for policy presentation, where both the social engineering strategy and the counter social

engineering strategy can be present in the same framework. For instance, we could model the two sides of strategies as a bimatrix game, in which a part of matrices, M_1 and M_2. The entries in the matrices $M_k(a_1, a_2)$ are used to represent the payoff of the k^{th} agent for the joint actions (a_1, a_2). Here the two matrices are of size $|A_1| \times |A_2|$ each. Here 1 means social engineering agents, and 2 means counter social engineering agents. If our game is a zero-sum game, then the matrices can be written as

$$M_1 = \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}, M_2 = -M_1.$$

If we consider a simple two agent game here at a single stage, we further assume that both learners are naive Q-learners that maintain a Q-table for the values of their possible actions with the updating function

$$Q_{t+1}(a) = (1 - \alpha)Q_t(a) + \alpha r_{t+1}$$

where α is the learning rate ranging from 0 to 1 and r_t is the reward at time t. The policy that the agent takes would output the action a_t as the maximizing action $a_t = \arg\max_b Q_t(b)$, where b is also an action.

Note that the above is only for two agents. For agents more than two, there would be other possibilities and other solutions. We here study and investigate new presentations for the even complex settings, especially how agents form two teams to combat with each other.

The transitions T and the rewards R are the main interests of a designed policy. For instance, one strategy for detecting social engineering attacks is to see if an attacker sends similar social engineering messages to multiple victims, probably in the same organization. This strategy begins with putting any sender of emails into the picture. Then it is expected for the email sender to send multiple email messages (with the "multiple" action enabled), possibly with form letter writing skills; and those messages are received by multiple receivers. The received messages are validated by their content similarity. If the similarity passes a threshold, the messages are further validated to see if majority of those message are "asking" for sensitive information such as money. If this action is further confirmed, then we decide it is probably a social engineering attack.

5 CONCLUSION

Social engineering is a common type of malicious attack on social media. It is a complex process. Its complexity comes from the involvement of many factors ranging from cognitive and behavioral theories to the latest web technologies, and to the new digital lifestyle of everyone. Social engineering attacks have escalated in complexity. Recent types of attacks, such as the context-aware attacks, are more difficult to detect than social engineering attacks from a few years ago.

In this position paper, we discuss a new counter social engineering method that oversees many factors during social engineering attacks, coordinates strategy optimization to pro-act appropriately at various stages in detecting and combating the attacks. In particular, we propose a new framework that incorporate interpersonal deception theories into multi-agent reinforcement learning to combat social engineering attacks. Our approach helps bridge the gap between the human understanding of interpersonal deceptions and machine discovered knowledge, improving reaction and response to novel attack types and enabling the use of known interpersonal deception strategies and wisdom.

6 ACKNOWLEDGEMENTS

This research is supported by National Science Foundation grant IIS-145374. Any opinions, findings, conclusions, or recommendations expressed in this paper are of the authors, and do not necessarily reflect those of the sponsor.

REFERENCES

[1] Ahmed Aleroud and Lina Zhou. 2017. Phishing Environments, Techniques, and Countermeasures: A Survey. 68 (04 2017).

[2] Bikramjit Banerjee and Jing Peng. 2003. Countering deception in multiagent reinforcement learning. In *Proceedings of the Workshop on Trust, Privacy, Deception and Fraud in Agent Societies at AAMAS-03, Melbourne, Australia.* 1–5.

[3] David B Buller and Judee K Burgoon. 1996. Interpersonal deception theory. *Communication theory* 6, 3 (1996), 203–242.

[4] Rachna Dhamija, J Doug Tygar, and Marti Hearst. 2006. Why phishing works. In *Proceedings of the SIGCHI conference on Human Factors in computing systems.* ACM, 581–590.

[5] Valerie Hauch, Iris Blandón-Gitlin, Jaume Masip, and Siegfried L Sporer. 2015. Are computers effective lie detectors? A meta-analysis of linguistic cues to deception. *Personality and Social Psychology Review* 19, 4 (2015), 307–342.

[6] Ryan Heartfield and George Loukas. 2015. A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks. *ACM Comput. Surv.* 48, 3, Article 37 (Dec. 2015), 39 pages. https://doi.org/10.1145/2835375

[7] A.D. Keromytis and S.J. Stolfo. 2016. Systems, methods, and media for generating bait information for trap-based defenses. (Sept. 22 2016). http://www.google.sr/patents/US20160277444 US Patent App. 15/155,790.

[8] A. N. Shaikh, A. M. Shabut, and M. A. Hossain. 2016. A literature review on phishing crime, prevention review and investigation of gaps. In *2016 10th International Conference on Software, Knowledge, Information Management Applications (SKIMA).* 9–15. https://doi.org/10.1109/SKIMA.2016.7916190

[9] Lisa Singh, Grace Hui Yang, Micah Sherr, Andrew Hian-Cheong, Kevin Tian, Janet Zhu, and Sicong Zhang. 2015. Public information exposure detection: Helping users understand their web footprints. *2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)* 00 (2015), 153–161. https://doi.org/doi.ieeecomputersociety.org/10.1145/2808797.2809280

[10] Lisa Singh, Hui Yang, Micah Sherr, Yifang Wei, Andrew Hian-Cheong, Kevin Tian, Janet Zhu, Sicong Zhang, Tavish Vaidya, and Elchin Asgarli. 2015. Helping Users Understand Their Web Footprints. In *Proceedings of the 24th International Conference on World Wide Web (WWW '15 Companion).* ACM, New York, NY, USA, 117–118. https://doi.org/10.1145/2740908.2742763

[11] S.J. Stolfo, A.D. Keromytis, B.M. Bowen, S. Hershkop, V.P. Kemerlis, P.V. Prabhu, and M.B. Salem. 2016. Methods, systems, and media for baiting inside attackers. (Nov. 22 2016). https://www.google.com/patents/US9501639 US Patent 9,501,639.

[12] Lina Zhou, Yongmei Shi, and Dongsong Zhang. 2008. A statistical language modeling approach to online deception detection. *IEEE Transactions on Knowledge and Data Engineering* 20, 8 (2008), 1077–1081.