Association for Information Systems

AIS Electronic Library (AISeL)

WISP 2020 Proceedings

Pre-ICIS Workshop on Information Security and Privacy (SIGSEC)

12-12-2020

# Deception against Deception: Toward A Deception Framework for Detection and Characterization of Covert Micro-targeting Campaigns on Online Social Networks

Jafar Haadi Jafarian

Ersin Dincelli

Keith Guzik

Matthew Michaelis

Farnoush Banaei-Kashani

*See next page for additional authors*

## Authors

Jafar Haadi Jafarian, Ersin Dincelli, Keith Guzik, Matthew Michaelis, Farnoush Banaei-Kashani, and Ashis Biswas

# Deception against Deception: Toward A Deception Framework for Detection and Characterization of Covert Micro-targeting Campaigns on Online Social Networks

**J. Haadi Jafarian[1], Ersin Dincelli, Keith Guzik, Matthew Michaelis,**
**Farnoush Banaei-Kashani, Ashis Biswas**
University of Colorado Denver
Denver, CO, USA

## ABSTRACT

Micro-targeting campaigns on online social networks are an emerging class of social engineering attacks that prime individuals via personalized content for malicious purposes. Detecting micro-targeting campaigns is challenging due to their clandestine nature and the lack of visibility around users' private communications. Our work aims to devise theories, methods, and tools to detect suspected micro-targeting campaigns. To this end, we propose to design and generate a network of decoy personas with characteristics similar to those of targeted groups in order to trap, engage, and identify micro-targeting campaigns. In this paper, we discuss our motivation to conduct this interdisciplinary research effort and introduce our focal research questions and preliminary design for a network of decoy personas.

**Keywords:** Micro-targeting campaigns, online social networks, social engineering, deception.

## INTRODUCTION

While online social networks (OSNs) have become an indispensable medium for communication and information exchange among general population, they have also enabled new types of macro- and micro-targeting social engineering campaigns. Micro-targeting campaigns focus on identifying and priming specific individuals via tailored individualized messages for various goals such as voluntary disclosure of sensitive information or spreading malware (Bossetta 2018). What distinguishes these campaigns from traditional email phishing is

[1] Corresponding author: haadi.jafarian@ucdenver.edu

the wealth of information available about individuals on OSNs, which can be exploited to devise individualized messages with a high likelihood of engagement. Detecting these campaigns is challenging due to the lack of visibility into users' private activities and the stealthy, adaptive nature of micro-targeting campaigns.

This work proposes a security approach against micro-targeting campaigns on OSNs based on deception. To identify potential campaigns against specific vulnerable populations (e.g., senior citizens) on an OSN, we generate and deploy a network of decoy personas (decoy accounts) with attributes that mimic those of the targeted groups. These decoys serve as traps for clandestine micro-targeting campaigns (e.g., contextualized spear-phishing messages) and will constantly adapt to maximize the likelihood of engagement. The data they collect is aggregated and analyzed to identify micro-targeting accounts in real-time. The focal contribution of our approach is twofold: First, we establish the feasibility of using deception strategically to identify clandestine micro-targeting campaigns. Second, we demonstrate the practicality of using deception to identify micro-targeting campaigns by developing a framework able to pick up diluted micro-targeting signals in a data-rich OSN (Twitter).

This paper is organized as follows. First, we define micro-targeting campaigns and present the literature on the detection of micro-targeting campaigns using deception in order to identify our focal research questions. To explain how we address these questions, we next lay out the proposed methodology for designing a network of decoy personas to attract, interact with, and identify malicious actors. Finally, we conclude the paper by discussing our future research.

## RELATED WORK

Micro-targeters use proprietary machine learning or deep learning algorithms for two purposes: (1) classifying users based on their demographic, attitudinal, and other available information about them (their posts on OSN); and (2) generating and delivering personalized messages to each user based on that information. A prime example of micro-targeting occurred in the 2016 U.S. presidential election when Russian operatives sent more than 10,000 spear-phishing tweets to U.S. Department of Defense employees on Twitter (Bossetta 2018). The messages were personalized and generated very high click rates. Bossetta (2018) proposed a kill-chain for how malicious micro-targeting attacks are conducted to conceptualize this new threat model. First, attackers *collect* data on the intended target from their social media posts. These data can be exploited to *construct* fake accounts that appeal to the target's interests. Using these accounts, attackers *contact* targets through communicative modes enabled by the platform, ranging from friend requests to direct messages. Depending on the attacker's intentions, the target may be tricked into revealing information or clicking a malicious link.

Researchers have successfully modeled these attacks. Seymour and Tully (2016), for instance, developed an approach called SNAP_R that uses a long short-term memory (LSTM) neural network to socially engineer specific users into clicking on deceptive URLs. The model is trained using word vector representations of social media posts dynamically seeded with topics extracted from the target's timeline. The approach achieved high success rates, tripling those of historic email attack campaigns. And AI-driven methods have been used to detect fraudulent entities or content on OSNs (by analysis of profile data, posted content, and network patterns) (Varol et al. 2017). But detecting micro-targeting campaigns on OSNs in a timely fashion is a more difficult problem, given the private nature of personal communications. It is difficult to identify suspicious activity until OSN users have been victimized and report what has transpired.

Deception offers a promising avenue for overcoming this obstacle, as it would invite malicious actors to reveal information about themselves that can then be used against them.

Deception has been used successfully in the past to identify bots, for instance. Lee et al. (2011) set up 60 honeypot Twitter accounts that sent random tweets designed to attract bots. Over the course of seven months, they collected over 20,000 bot followers. However, identifying micro-targeting campaigns requires a more dynamic approach to deception, given the intelligence of malicious human actors compared to bots.

These challenges in identifying micro-targeting campaigns motivate the overarching research question informing our research: *how can deception be leveraged to neutralize micro-targeting activities?* Answering this question requires, in turn, a host of technical considerations. *How can effective decoy accounts be designed? Does networking among decoy accounts increase their efficacy in attracting micro-targeters? How can decoy accounts be automated to increase their practicality? How can the data generated by decoy accounts be analyzed to identify malicious accounts? How is success or efficacy defined?* The next section describes our methodology for addressing these questions.

## METHODOLOGY AND RESEARCH OBJECTIVES

**Research Question 1: How to design an effective DP?** The foundation of our approach is decoy personas (DPs), fictitious social media accounts designed to resemble the intended targets of micro-targeting campaigns. Decoy personas should be *plausible* (able to pass as actual human accounts), *playable* (fit the characteristics of profiles targeted by attackers), and *practical* (feasible in ethical, legal, and operational terms). We refer to these criteria as $P^3$.

Prior research examining users' responses to the information on OSN shows that users' reactions depended on the OSN features (e.g., like, share, comment) and the content of the information (DePaula and Dincelli 2018). Therefore, to design DPs according to $P^3$ criteria, we identify a set of behavioral characteristics related to OSN features, information shared (e.g., framing), and users (intrapersonal factors, such as demographics) based on existing literature.

Past research has shown that persuasive messages (e.g., advertisements, political campaigns, phishing messages) are generally more effective when they are tailored to the *interests* and *concerns* of the target audience (Hirsh et al. 2012). Previously, such persuasive messages were targeted towards *demographic* groups (e.g., gender, age) or focused on gain or loss frames (Goel et al. 2017). However, a vast amount of personal data generated by users on OSNs made it possible to predict more detailed insights about users, such as users' *personality traits* and *privacy concerns* (Sumner et al. 2011). Activities on OSNs, such as the history of users' posts and likes, have shown to be predictors of users' personality (Azucar et al. 2018), demographics (age and sex), mental status (e.g., anxious or stressed) (Luerweg 2019), and political ideology (e.g., politically conservative or liberal) (Popov et al. 2018; Tandera et al. 2017). Targeting OSN users based on their personality has been used widely to increase the persuasiveness of messages in various areas, such as advertising and political communication (Stieglitz and Dang-Xuan 2013). For example, Facebook users who received targeted advertisements that were tailored to their personalities were about 50 percent more likely to buy the advertised product than when the advertisement did not match their personalities (Matz et al. 2017).

Thus far, we have identified two categories of characteristics of OSN users that feed into the "persona" we aim to design for the network of decoys: (1) those that serve as *visual cues* and (2) those that affect the *framing of messages* (i.e., posts) produced by the decoy persona. These characteristics play an important role in increasing the plausibility and playability of the decoy accounts.

**Table 1.** Behavioral Characteristics of Online Persona for Decoy Accounts

| Visual Cues | Message Framing |
|---|---|
| Profile pictures | Personality |
| Posts (photographs) | Posts (word choices) |
| OSN features (like, retweet) | Interests (sports, politics) |
| Demographics (age, sex) | Human vulnerabilities (greed, fear) |

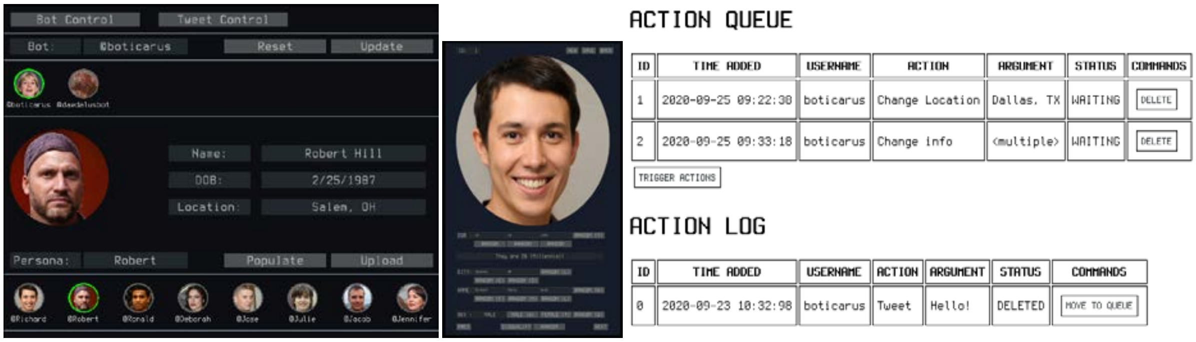
**Research Question 2: How does networking impact the efficacy of DPs?** To answer this question, we build on prior work on cyber deception planning (Jafarian 2017) to develop a planning method that, given an input OSN and attributes of a potential micro-targeting campaign, crafts a network (called DPNet) of DPs. The characteristics of these DPs are determined based on the factors identified in the first research stream. To model the planning problem, we formalize it as a satisfiability modulo theory (SMT) and will use an off-the-shelf SMT solver like Microsoft Z3 to solve it. Given a campaign target definition as a constraint on persona attributes and a budget, the planning module designs a group of DPs based on the following criteria: (1) *maximal coverage* of the campaign target space, (2) *high diversification of personas* based on plausibility, and (3) playability rules to provide a variety of potential targets for the campaign.

**Research Question 3: How to actuate a persona into an account?** We are introducing artificial intelligence (AI) agents that can make conversations based on persona-specific conversational models and can learn to interact with other OSN users while posing persona-specific behaviors as defined by each individual DP. Actuation of these agents requires methods for generating the profile information of a decoy account based on its DP, and also policies and methods for $P^3$-compliant text generation for tweets and replies, issuing retweets, mentions, and (un)follow requests, and managing incoming follow requests.

**Research Question 4: How to adapt DPs to enhance their effectiveness?** We plan to employ a reinforcement learning approach to adapt the design of DPs and their associated OSN accounts. The idea is to approximate a set of policies for the decoy account generator module (i.e., the agent) that adapts decoy accounts over time to enhance their interactions with the micro-targeters. The design is fueled by a specific reward function that translates a set of quantitative and qualitative metrics of the generated decoy accounts into a reward value. Examples of such metrics include the number of messages received with similar information by the agent with respect to a pre-determined group of coordinated agents, the importance of the topics of messages received regarding the context, current trends, etc.

**Research Question 5: How to detect malicious accounts?** We plan to develop supervised and semi-supervised methods that evaluate the data collected by the AI agents (individually and collectively) to identify malicious actors and campaigns, along with their structure, intention, and coordination tactics. Our method uses machine learning techniques to classify an engaged account as malicious or benign by analyzing observed interactions with that account, in addition to its profile, content, and network features.

**Research Question 6: How to evaluate the approach effectively and ethically?** To evaluate our approach, we conduct our evaluation in two stages. First, we plan to build and use a synthetic emulation testbed for a Twitter-like OSN using the publicly available data about a sample target population. We plan to develop social-scientific user studies against social engineering attacks to devise reliable and sound pilot experiments. In the second stage, we plan to deploy our approach on Twitter and develop reliable experiments with qualified volunteers via Amazon Mechanical Turk to evaluate and understand the effectiveness and limitations of our approach in a real-world setup. Figure 1 shows the snapshots of the completed components of the proposed framework, which is currently under development. So far, the tool includes a persona generation tool that allows for generating and auto-populating personas, an interface between planner and Twitter to deploy generated personas, and a unified web interface that allows for remote management of the decoy accounts.



Bot: @boticarus

Name: Robert Hill

DOB: 2/25/1987

Location: Salem, OH

Persona: Robert

ACTION QUEUE

| ID | TIME ADDED | USERNAME | ACTION | ARGUMENT | STATUS | COMMANDS |
|---|---|---|---|---|---|---|
| 1 | 2020-09-25 09:22:30 | boticarus | Change Location | Dallas, TX | WAITING | DELETE |
| 2 | 2020-09-25 09:33:18 | boticarus | Change info | <multiple> | WAITING | DELETE |

ACTION LOG

| ID | TIME ADDED | USERNAME | ACTION | ARGUMENT | STATUS | COMMANDS |
|---|---|---|---|---|---|---|
| 0 | 2020-09-23 10:32:98 | boticarus | Tweet | Hello! | DELETED | MOVE TO QUEUE |

**Figure 1.** Snapshots from the interface of the OSN emulation system (left and middle: persona generation and management module, right: activity management module)

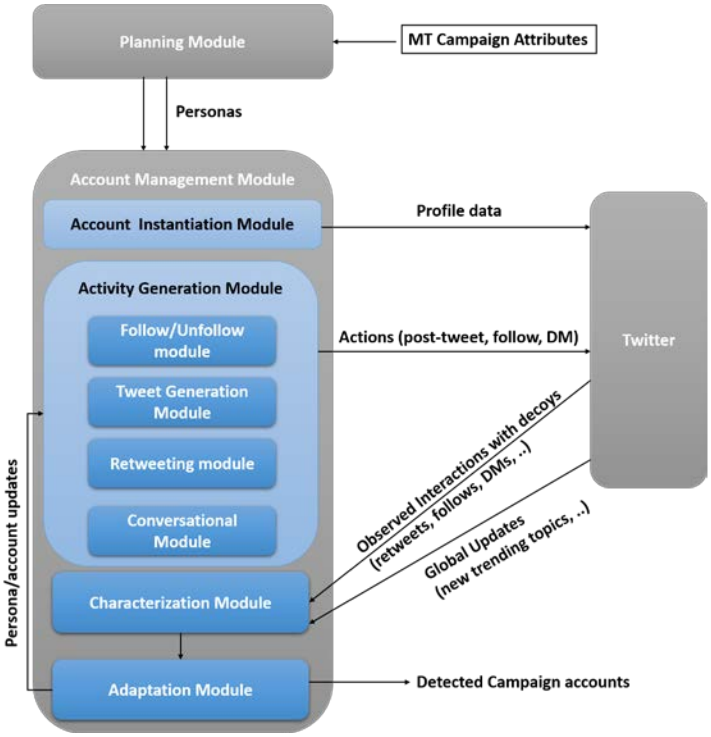

Planning Module
MT Campaign Attributes
Personas
Account Management Module
Account Instantiation Module
Profile data
Activity Generation Module
Follow/Unfollow module
Tweet Generation Module
Retweeting module
Conversational Module
Actions (post-tweet, follow, DM)
Twitter
Observed Interactions with decoys (retweets, follows, DMs, ..)
Global Updates (new trending topics, ..)
Characterization Module
Adaptation Module
Detected Campaign accounts
Persona/account updates

**Figure 2.** The Architecture of the Proposed Framework

Figure 2 shows the architecture of our proposed framework for detecting micro-targeting campaigns on Twitter that is developed based on P$^3$ criteria. The framework includes multiple modules that are responsible for distinct functions. The *planning* module first designs a network of decoy personas that mimic the characteristics of potential targets for a particular campaign. Then, the persona designs are fed into the *account management* module. In the *account management* module, the *profile instantiation* sub-module generates an account from each persona and deploys them on Twitter. Then, the *activity generation* sub-module manages the activities of the account (e.g., follow/unfollow, posting, etc.) based on predefined objectives and policies. The interactions with decoy accounts on Twitter (e.g., retweets, direct messages), as well as new global updates (e.g. new trending topics), are fed back into the *characterization* sub-module of the account management. The *characterization* sub-module performs three distinct tasks: (1) it analyzes the collected data to detect malicious micro-targeting campaigns; (2) it determines the types of adaptations that need to be applied to each persona or their corresponding accounts (e.g., posting about a new topic); and (3) relays those to the *adaptation* sub-module. The *adaptation* sub-module defines the required decoy adaptations and feeds them back into the *account instantiation* and *activity* sub-modules to update the profile or activities of the accounts.

## CONCLUSION AND FUTURE PLANS

In this research-in-progress paper, we proposed our methodology for detecting micro-targeting campaigns through deceptive decoy accounts. Thus far, we have identified the focal research questions and developed an architecture for our framework. In the future, we aim to (1) continue our research on the proposed research questions, (2) complete the development of the emulation system, and (3) conduct experiments with actual Twitter users to evaluate our approach.

## REFERENCES

Azucar, D., Marengo, D., and Settanni, M. 2018. “Predicting the Big 5 Personality Traits from Digital Footprints on Social Media: A Meta-Analysis,” *Personality and Individual Differences* (124), pp. 150-159.

Bodó, B., Helberger, N., and de Vreese, C. H. 2017. “Political Micro-Targeting: A Manchurian candidate or just a dark horse?” *Internet Policy Review* (6:4), pp. 1-13.

Bossetta, M. 2018. “The Weaponization of Social Media: Spear Phishing and Cyberattacks on Democracy,” *Journal of International Affairs* (71:1.5), pp. 97-106.

DePaula, N., and Dincelli, E. 2018. “Information Strategies and Affective Reactions: How Citizens Interact with Government Social Media Content,” *First Monday* (23:4).

Gallagher, R. 2016. “Where Do the Phishers Live? Collecting Phishers’ Geographic Locations from Automated Honeypots,” *ShmooCon XII*, Washington, DC.

Goel, S., Williams, K., and Dincelli, E. 2017. “Got Phished? Internet Security and Human Vulnerability,” *Journal of the Association for Information Systems* (18:1), pp. 22-44.

Hirsh, J. B., Kang, S. K., and Bodenhausen, G. V. 2012. “Personalized Persuasion: Tailoring Persuasive Appeals to Recipients’ Personality Traits,” *Psychological Science* (23:6), pp. 578-581.

Jafarian, J. H. 2017. “Cyber Agility for Attack Deterrence and Deception,” Doctoral dissertation, The University of North Carolina at Charlotte. Publication No. 10686943. ProQuest Dissertations Publishing.

Lee, K., Eoff, B. D., and Caverlee, J. 2011. “Seven Months with the Devils: A Long-Term Study of Content Polluters on Twitter.” *Proceedings of the 5th International AAAI Conference on Web and Social Media (ICWSM)*, Barcelona, Spain, pp. 185-192.

Luerweg, F. 2019. “The Internet Knows You Better Than Your Spouse Does. Scientific American,” www.scientificamerican.com/article/the-internet-knows-you-better-than-your-spouse-does/

Popov, V., Kosinski, M., Stillwell, D., and Kielczewski, B. 2018. “Apply Magic Sauce,” https://applymagicsauce.com/

Seymour, J., and Tully, P. 2016. “Weaponizing Data Science for Social Engineering: Automated E2E Spear Phishing on Twitter,” *Black Hat USA* (37), pp. 1-39.

Stieglitz, S., and Dang-Xuan, L. 2013. “Social Media and Political Communication: A Social Media Analytics Framework,” *Social Network Analysis and Mining* (3:4), pp. 1277-1291.

Sumner, C., Byers, A., and Shearing, M. 2011. “Determining Personality Traits & Privacy Concerns from Facebook Activity,” *Black Hat Briefings*, Abu Dhabi, United Arab Emirates, pp. 197-221.

Varol, O., Ferrara, E., Davis, C. A., Menczer, F., and Flammini, A. 2017. “Online Human-Bot Interactions: Detection, Estimation, and Characterization,” arXiv preprint arXiv:1703.03107.

Walter, D., Ophir, Y., and Jamieson, K. H. 2020. “Russian Twitter Accounts and the Partisan Polarization of Vaccine Discourse, 2015–2017,” *American Journal of Public Health* (110:5), pp. 718-724.

Tandera, T., Suhartono, D., Wongso, R., and Prasetio, Y. L. 2017. “Personality Prediction System from Facebook Users,” *Procedia Computer Science* (116), pp. 604-611.