IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received November 23, 2020, accepted December 14, 2020, date of publication December 24, 2020,
date of current version January 5, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3047337

Online Social Deception and Its Countermeasures: A Survey

ZHEN GUO[1], JIN-HEE CHO[1], (Senior Member, IEEE), ING-RAY CHEN[1], (Member, IEEE), SRIJAN SENGUPTA[2], MICHIN HONG[3], AND TANUSHREE MITRA[4]

[1]Department of Computer Science, Virginia Tech, Falls Church, VA 22043, USA
[2]Statistics, North Carolina State University, Raleigh, NC 27695, USA
[3]School of Social Work, Indiana University, Indianapolis, IN 46202, USA
[4]Information School, University of Washington, Seattle, WA 98195, USA

Corresponding author: Zhen Guo (zguo@vt.edu)

ABSTRACT We are living in an era when online communication over social network services (SNSs) have become an indispensable part of people's everyday lives. As a consequence, online social deception (OSD) in SNSs has emerged as a serious threat in cyberspace, particularly for users vulnerable to such cyberattacks. Cyber attackers have exploited the sophisticated features of SNSs to carry out harmful OSD activities, such as financial fraud, privacy threat, or sexual/labor exploitation. Therefore, it is critical to understand OSD and develop effective countermeasures against OSD for building trustworthy SNSs. In this paper, we conduct an extensive survey, covering 1) the multidisciplinary concept of social deception; 2) types of OSD attacks and their unique characteristics compared to other social network attacks and cybercrimes; 3) comprehensive defense mechanisms embracing prevention, detection, and response (or mitigation) against OSD attacks along with their pros and cons; 4) datasets/metrics used for validation and verification; and 5) legal and ethical concerns related to OSD research. Based on this survey, we provide insights into the effectiveness of countermeasures and the lessons learned from the existing literature. We conclude our survey with in-depth discussions on the limitations of the state-of-the-art and suggest future research directions in OSD research.

INDEX TERMS Online social deception, cyberattacks, security, defense, prevention, detection, and response, social media, online social networks.

I. INTRODUCTION

A. MOTIVATION

Social media and social network services (SNSs) have become an indispensable part of people's everyday lives. In 2020, approximately 82% of Americans reported using social media [93]. This significant surge is due to various benefits that users enjoy, such as easy communications with others, engagement in civic and political activities, searching jobs, marketing, and/or sharing information or emotional support. Even with these significant benefits, many people have ambivalent feelings about social media due to privacy concerns and/or deceptive activities aiming to harm normal, legitimate users [153]. The proliferation of highly advanced social media technologies has been exploited by perpetrators as convenient tools for deceiving users [7]. The widespread damage due to *online social deception* (OSD) attacks have increased significantly in recent times, with about 25% of people experiencing some types of social deception, such as identity theft, cyberbullying, fraud, or phishing in 2018 [156]. The serious consequences have led to such OSD attacks being defined as *cybercrimes* [139] since early 2000's. The advanced features of SNS technologies further have facilitated the significant increase of serious, sophisticated cybercrimes, beyond simple phishing or spamming, such as human trafficking, online consumer fraud, identity cloning, hacking, child pornography, or online stalking [192]. Therefore, we need to deeply understand OSD and think of how to develop effective countermeasures against OSD for building a trustworthy cyberspace.

Although there have been several papers surveying online social network (OSN) attacks [2], [54], [58], [92], [98], [137], [154], [216], [218], the existing surveys are limited in discussing detection mechanisms using various artificial intelligence (AI) techniques including machine learning, deep learning, or text mining. They did not really embrace a wide

The associate editor coordinating the review of this manuscript and approving it for publication was Hocine Cherifi.

spectrum of defense against OSN attacks such as prevention, detection, and response (or mitigation). Further, there has been lack of discussions on deception which is exploited as the starting point of most OSN attacks.

B. RESEARCH GOAL & QUESTIONS

To fill the gap discussed above, this work aims to deliver a comprehensive, systematic survey for researchers to efficiently and effectively grasp a large volume of the state-of-the-art literature on OSD attacks and its countermeasures in terms of three aspects of defense, such as prevention, detection, and response (or mitigation). To this aim, the scope of our survey focuses on answering the following **research questions**:

RQ1: *How is OSD affected by the fundamental concepts and characteristics of social deception which have been studied in multidisciplinary domains?*
RQ2: *What are new attack types based on the recent trends of OSD attacks observed in real online worlds and how are they related to common social network attacks, cybercrimes, and security breaches based on cybersecurity perspectives?*
RQ3: *How can the cues of social deception and/or susceptability traits to OSD affect the strategies by attackers and defenders in OSNs?*
RQ4: *What kinds of defense mechanisms and/or methodologies need to be explored to develop better defense tools combating OSD attacks?*
RQ5: *What are the key limitations of existing validation and verification methodologies in terms of datasets and metrics?*
RQ6: *What are the key concerns associated with ethical issues in conducting OSD research?*

C. COMPARISON WITH EXISTING SURVEY PAPERS

As social deception leverages OSNs as platforms, there have been several survey papers [2], [54], [58], [92], [98], [137], [154], [216], [218] discussing social network attacks.

Fire *et al.* [54] mainly discussed social network threats targeted at young children in terms of phishing, spamming, fake identity, profile cloning attacks, cyberbullying, and cyber-grooming. Rathore *et al.* [154] surveyed social network attacks with a special emphasis on multi-media security and privacy. Since fake news is an emerging deception attack in OSNs, a recent effort by Kumar and Shah [98] discussed the details of fake news detection methods. Although the existing works stated above [54], [98], [154] proposed mechanisms to mitigate specific social deception threats, they focused on discussing prevention methods and practical security suggestions. An interesting observation is that no work has discussed ethical issues in developing techniques to deal with OSN threats/attacks. Besides, we observed a lack of understanding on the pros and cons of each detection or mitigation technique that combat online social deception attacks.

Rathore et al. [154] conducted a comprehensive survey on social network security. They classified social network security threats in three categories, including multimedia content threats, traditional threats, and social threats with 21 types of threats/attacks. The authors mainly discussed multimedia content threats, along with their definitions, impact, and security response methods, including detection methods for each type of threat. They also compared various security attacks in terms of the nature of attack (attack source), attack difficulty, risk to data privacy/integrity, and attack impact on users. In the end, they proposed a framework to measure and optimize the security of SNSs.

Novak and Li [137] focused on OSN security and data privacy issues. They discussed how to protect user data from attacks by the research in social network inference (e.g., user attributes, location hubs, and link prediction) and in anonymizing social network data. Gao *et al.* [58] discussed the four types of social network attacks, which include privacy breaches, viral marketing attacks, network structural attacks, and malware attacks. The authors compared various attacks, including information leak, de-anonymizing, phishing, Sybil, malware, and spamming, and discussed countermeasure defense mechanisms against them.

Fire *et al.* [54] discussed key OSN threats and solutions against them. The authors outlined OSN threats with an additional focus on attacks against children and teenagers. There are 5 classic threats, 9 modern threats, combination threats and 3 threats targeting children. The defense solutions were techniques provided by OSN operators, commercial companies, and academic researchers and the protection ability of various solutions were discussed. In the end, they provided recommendations for OSN users to protect their security and privacy when using social networks. Kayes and Iamnitchi [92] reviewed the taxonomies of privacy and security attacks and their solutions in OSNs. The authors categorized the attacks based on OSN's stakeholders (users and their OSNs) and entities (i.e., human, computer programs, or organizations) performing the attacks. They discussed attacks on users' information and how to counter leakages and linkages. However, the attacks discussed as social deception are common social network attacks, such as Sybil attacks, compromised accounts and/or spams. The defense techniques to mitigate each attack type were discussed as ways to detect and resist against those attacks.

Kumar and Shah [98] discussed the characteristics and detection of false information on Web and social media, with two knowledge-based types: *opinion-based methods* with ground truth (e.g., fake reviews), and *fact-based methods* without ground truth (e.g., hoaxes and rumors). They described how false information can perform successful deception attacks, and their impacts on the speed of false information propagation and characteristics for each type. Based on the specific characteristics, the authors discussed the detection algorithms for each type utilizing different features and propagation models in terms of the analysis of classification, key actors, impacts, features, and

TABLE 1. Comparison of the key contributions of our survey paper and other existing survey papers.

Criteria	Our Survey	Rathore et al. [154]	Novak and Li [137]	Gao et al. [58]	Fire et al. [54]	Kayes and Iamnitchi [92]	Tsikerdekis and Zeadally [187]
Concepts and Characteristics of Online Social Deception							
Multidisciplinary concepts	✓	✗	✗	✗	✗	✗	✗
Deception cues	✓	✗	✗	✗	✗	✗	Limited
Spectrum of deception with/without intentionality	✓	✗	✗	✗	✗	✗	✓
Properties of social deception	✓	✗	✗	✗	✗	✗	✓
Susceptibility factors to OSD attacks	✓	✗	✗	✗	✗	✗	Limited
Online Social Network Attack Types							
Fake news	✓	✗	✗	✗	✗	✗	✗
Rumors	✓	✗	✗	✗	✗	✗	✗
Information manipulation	✓	✗	✗	✗	✗	✗	✗
Fake reviews	✓	✗	✗	✗	✗	✗	✗
Phishing	✓	✓	✗	✓	✓	✓	✗
Spamming	✓	✓	✗	✓	✓	✓	✗
Fake identity	✓	✓	✓	✓	✓	✓	✓
Compromised account	✓	✗	✗	✗	✗	✓	✓
Profile cloning attack	✓	✓	✗	✓	✓	✗	✓
Crowdturfing	✓	✗	✗	✗	✗	✗	✗
Human trafficking	✓	✗	✗	✗	✗	✗	✗
Cyberbullying	✓	✓	✗	✗	✓	✗	✗
Cyber-grooming	✓	✓	✗	✗	✓	✗	✗
Cyberstalking	✓	✓	✗	✗	✗	✗	✗
Existing OSNs Security Solutions							
Security issues and challenge	✓	✓	✗	✓	✗	Limited	✓
Prevention	✓	Limited	✗	Limited	✓	✗	Limited
Detection	✓	✓	✓	✓	✓	✓	✗
Mitigation	✓	✗	✗	✗	✗	✓	✗
Security suggestions	✓	✓	✗	Limited	✓	Limited	✗
Discussing Limitation, Pros and Cons of Detection							
Ethical Issues	✓	✗	✗	✗	✗	✗	✗
Discussing Key Limitations	✓	✗	✗	✗	✗	✗	✗
Pros and Cons of Techniques	✓	✗	✗	✗	✗	✗	✗

measurements. In addition, they discussed the detection algorithms for opinion-based and fact-based detection mechanisms, respectively.

Wu [216] summarized misinformation in social media, focusing more on the unintentional-spread misinformation, such as meme, spam, rumors, and fake news. It discussed information diffusion models and network structure, misinformation detection and spreader detection, misinformation intervention, and detailed evaluation datasets and metrics. The diffusion models are SIR (Susceptible-Infected-Recovered/Removed), Tipping Point, Independent Cascade, and Linear Threshold model. In the diffusion process, user types can be categorized as *forceful individuals* [2], which refer to users not affected upon belief exchange. Wu and Liu [218] described detecting crowdturfing in social media. The authors summarized the history of astroturfing campaign and crowdturfing. The methods to investigate crowdturfing is mining and profiling social media users as attackers and modeling information diffusion in social media. Finally, crowdturfing detection can be performed in content-based, behavior-based, and diffusion-based approaches in the state-of-the-art research. However, this work [218] limited its scope only to crowdturfing. Hence, we did not include it in TABLE 1 for the comparison of our survey paper with other counterpart survey papers.

Tsikerdekis and Zeadally [187] analyzed the motivations and techniques of online deception in social media platforms. They categorized social media by the extent of media richness and self-disclosure. Due to the user connection and content sharing nature of social media, online deception techniques can involve multiple roles, such as content, sender, and communication channel. They also provided an insightful discussion of challenges in prevention and detection of online deception. However, this work did not discuss any attack behaviors concerned as in our paper.

Based on the existing survey papers [2], [54], [58], [92], [98], [137], [154], [216], [218], we found that there is no comprehensive survey paper on online social deception which sits between OSN threats and cybercrimes. The most related work discussed above focused on security and privacy issues and their solutions in OSNs. Most previous studies analyzed various types of OSN threats and provided detection methods for specific types of security threats. However, they usually discussed traditional types of security issues, which only partially overlap our definitions of social deception threats. We intended to cover more types of OSD threats and provide full ranges of solutions using a wide spectrum of defense strategies, including prevention, detection, and response (or mitigation). To clarify the contributions of our survey paper, we demonstrated the key differences in scope and surveyed

techniques between our survey paper and the existing OSN security and/or attack papers in TABLE 1. We list the key contributions of our survey paper compared to existing survey papers in the following section.

D. KEY CONTRIBUTIONS

We made the following **key contributions** in this paper:

- To understand the fundamental meaning of social deception and its key characteristics, we comprehensively surveyed the multidisciplinary concepts and key properties of social deception. No previous survey paper has addressed all these concepts together to understand the fundamental meanings of social deception.
- We provided a comprehensive set of OSD attacks by following the key properties of social deception (see Section II-D). In particular, we discussed the relationships between social network attacks, OSD attacks, and cybercrimes by describing the relationships between them, major attacks in each category, and the attack goals of OSD in terms of loss of security goals.
- We provided an overview of social deception cues which have been studied in multidisciplinary domains, including individual, cultural, linguistic, physiological, psychological, and technological deception cues. This literature survey on the deception cues is helpful to obtain useful insights for developing better defense tools in terms of prevention, detection, and response against OSD attacks.
- To provide a more comprehensive understanding on a system-level defense framework against OSD attacks, we extensively surveyed the three types of defense mechanisms, including prevention, detection, and response (or mitigation), which are summarized in TABLES 5 – 7.
- We provided pros and cons of major defense approaches to combat OSD attacks and the overall trends of the state-of-the-art OSD defense techniques. This gives a reader to easily identify relevant defense techniques in a given context to conduct research in this area.
- We identified the common datasets and metrics that have been used to validate the performance of defense mechanisms combating the OSD attacks. From this comprehensive survey on datasets and metrics, we also provided useful research directions to enhance the validation and verification methods, which have not been discussed in other existing counterpart survey papers.
- We also comprehensively discussed key findings, insights and lessons learned, limitations, and future research directions based on the extensive survey conducted in this work.

E. PAPER STRUCTURE

The rest of this paper is structured as follows:

- In Section II, we surveyed the multidisciplinary concepts of 'deception' along with goals of deception. In addition, we compared different types of deception in the spectrum of deception in terms of intent and detectability. Further, we discussed the key properties of deception.
- In Section III, we discussed various types of OSD attacks in terms of false information, luring and phishing, fake identity, crowdturfing, and human targeted attacks. Following the major OSD types, the comparisons between social network attacks, social deception attacks, and cybercrimes are discussed. We also discussed the security breach by OSD attacks based on traditional CIA (confidentiality, integrity, and availability) security goals.
- In Section IV, we addressed various cues of social deception, in terms of individual, cultural, linguistic, physiological, psychological, and technological social deception cues. In addition, we discussed the relationships between offline and online social deception cues, mainly identifying their commonalities and differences.
- In Section V, we discussed five different types of key factors that affect susceptibility to online social deception, including demographic, personality, cultural, social and economic, and network structure feature-based factors.
- In Section VI, we surveyed two existing prevention mechanisms against OSD attacks, namely, data-driven analysis, and social honeypots. Although social honeypots are used for both intrusion prevention and intrusion detection, we include them under this intrusion prevention mechanism to preserve its original design purpose as a proactive intrusion prevention mechanism.
- In Section VII, we comprehensively surveyed three existing detection mechanisms against OSD attacks, namely, user profile-based, message content-based, and network structure feature-based. Each class of detection mechanisms are discussed in terms of attack type, key methods, features, and datasets used.
- In Section VIII, we discussed several existing approaches of response mechanisms to detected OSD attacks in terms of mitigation or recovery from OSD attacks.
- In Section IX, we discussed datasets and metrics used for the validation and verification of defense mechanisms against OSD attacks.
- In Section X, since OSD research involves humans and their behaviors, we discussed ethical issues associated with conducting the OSD research.
- In Section XI, based on the comprehensive survey conducted on OSD attacks and their countermeasures, we provided insights and lessons learned along with the limitations of the state-of-the-art OSD research.
- In Section XII, we provided concluding remarks and discussed future research directions in this area.

II. CONCEPTS AND CHARACTERISTICS OF DECEPTION

The concept of deception is highly multidisciplinary and has been studied in various domains. In this section, we discuss

the root definitions of deception and the fundamental properties of deception which have been applied in launching OSD attacks in OSN platforms.

A. MULTIDISCIPLINARY CONCEPT OF DECEPTION

Let us start by looking at the dictionary definition of deception [37]. Deception is defined as: "To cause to believe what is false." However, the definition is too broad and many deception researchers raised doubts on the definition. In the literature, the concepts of deception have been discussed with different perspectives under different disciplines. We briefly discuss how a different discipline has studied deception in the following sections.

1) PHILOSOPHY

In Philosophy, intentional and unintentional (by mistake) deception has been discussed, such as 'inadvertent or mistaken deceiving' [19]. However, the common concept of deception was mostly agreed with 'misleading a belief' by either inadvertently or mistakenly [59], [160]. The core aspects of deception in Philosophy lies in an *intentional* act to *mislead* an entity to believe a *false belief*.

2) BEHAVIORAL SCIENCE

Behavior scientists[1] investigated the concept of deception and its process in the behaviors of animals or humans. Two main concepts of deception are: (1) *Functional deception* for an individual's behavior (i.e., a signal) to mislead the actions of others; and (2) *intentional deception* referring to intentional states, such as beliefs and/or desires, guide an individual's behavior, leading to the misrepresentation of belief states [73], [104], [176].

3) PSYCHOLOGY

Psychologists defined deception as a behavior providing information to mislead subjects to some direction [3] or explicit misrepresentation of a fact aiming to mislead subjects [81], [134]. The major psychological deception study focused on identifying cues as committing a crime [63], psychological symptoms for self-deception [20], [75], individual differences and/or cues to deception [157], verbal or non-verbal communication cues [235].

4) SOCIOLOGY

Sociological deception research has mainly studied the effect of deception in various social context on both positive and negative aspects [123], or deception as a relational, or marketing strategy [150].

5) PUBLIC RELATIONS

In this domain, the concept of self-deception has been studied as a strategic solution to resolve internal or external crisis [168]. The external role of self-deception is described as a way to avoid disastrous impact on an organization [143] by attributing a problem (or guilty) to an individual or victim.

6) COMMUNICATIONS OR LINGUISTICS

In this domain, deception research often aimed to identify either verbal or non-verbal indicators for deceptive communications. Interpersonal deception theory (IDT) views deception as an interactive process between senders and receivers, exchanging non-verbal and verbal behaviors and interpreting their communicative meanings. IDT further explains that deceivers strategically manage their verbal communications to successfully deceive receivers [15], [16]. Experimental studies showed that deceivers produced more words, fewer self-oriented (e.g., I, me, my) and more sense-based words (e.g., seeing, touching) than truth-tellers [72].

7) COMMAND AND CONTROL

In the military domain, deception refers to any planned maneuvers undertaken for revealing false information and hiding the truth to an enemy with the purpose of misleading the enemy and enticing the enemy to undertake the wrong operations [29], [124], [210]. Military deception involves a large number of individuals or organizations as both deceivers and victims and takes place in a long time period [29].

8) COMPUTING AND ENGINEERING

Deceptive behaviors have popularly exhibited by cyber attackers in various forms, such as phishing, social engineering attacks, fraud advertisements, stealthy attack, and so forth [74], [154]. In addition, as the threat of phishing emails increases, an individual online user's susceptibility to phishing attacks is studied in terms of demographics [114], [141], [171] or personality traits [30], [48], [70], [71], [128], [148], [149]. We discuss the details of susceptibility to OSD attacks in Section V. In addition, a lot of detection mechanisms to OSD attacks have been developed in the literature. We discuss them with more details in Section VII as well.

For easy grasping of the key multidisciplinary concepts of deception, we summarized the key deception concepts under different disciplines in FIGURE 1.

B. TYPES OF DECEPTION

Although deception can be intentional or unintentional, we focus on intentional deception in this work, which is more related to an attacker's intent. The intentional deception consists of deception with malicious intent and with non-malicious intent for a deceiver's interest [47].

The goals of malicious deception include:

- *Financial benefit*: Many deceptive behaviors has its purpose to obtain a monetary benefit. Financial benefit is a common reason of an individual's online deceptive behavior. For example, a spammer can be paid from clicking advertisements by attracting online traffic to the specific sites [133]. Malicious users spread phishing links to collect credentials from victims [194].

[1] We consider biologists, ecologists, neuroscientists, and medical scientists as 'behavioral scientists' in this work.

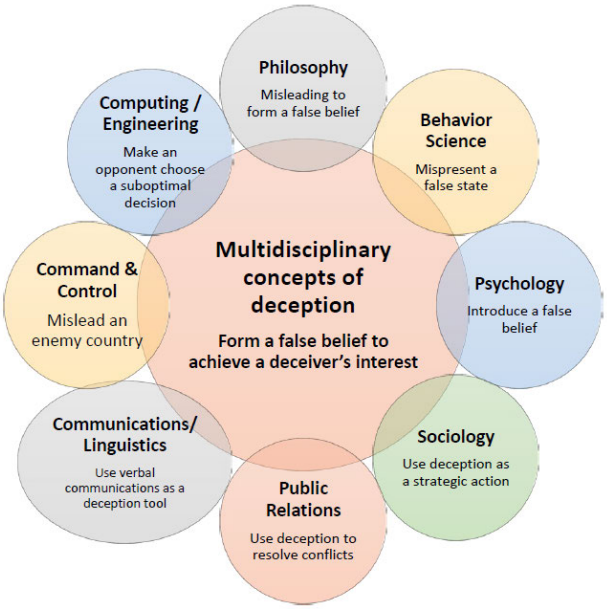


FIGURE 1. The key multidisciplinary concepts of deception.

- *Manipulation of public opinions*: In social media, social and political bots play a role in influencing public opinions [57]. Malicious bots spread spam and phishing links. Politicians and governments worldwide have been using such bots to manipulate public opinions.
- *Cooperative deception*: Cooperation is a strategy of balancing costs and benefits and maintaining stakeholder relationships in the deception or cooperation interactions with opponents [183], often used in public relations.
- *Parasitism* [168]: This refers to 'false framing of responsibility' which can be easily used as a strategy to solve complicated issues without introducing long-term investigations that may cause structural changes.

The goals of non-malicious deception are commonly discussed as follows:

- *Privacy protection*: Deception can be used as a defense for the privacy protection at the organization-level or individual-level. This is also called *defensive deception*. There are a few methods for the individual-level privacy protection in cyberspace. Some privacy techniques add a noise to a user's data for protection against attackers [151] because the data can be modified before being published.
- *Self-presentation*: People use fake presentation to present themselves as certain roles or intents [164]. Self-presentation is an activity to impress others for both liars and truth tellers. Self-presentation is one way of understanding nonverbal communication [35]. Self-presentation can be used as prediction cues to deception [35].
- *Self-deception*: This is to hide true information reflecting conscious mind unconsciously [183], with the two main benefits of not being detected easily and reducing immediate cognitive costs.

TABLE 2. Goal, intent, and security breach according to a different type of social deception.

Goal of social deception	Malicious vs. Non-malicious intent	Breach of security goals
Financial benefits	Malicious	Loss of confidentiality and integrity
Manipulation of public opinions	Malicious	Loss of integrity
Cooperative deception	Malicious	Loss of integrity
Parasitism	Malicious	Loss of integrity
Privacy protection	Non-malicious	Loss of confidentiality
Self-presentation	Non-malicious	Loss of integrity
Self-deception	Non-malicious	Loss of integrity

In TABLE 2, we summarized what social deception is malicious or not and how it is associated with breach of security goal.

C. TAXONOMIES AND SPECTRUM OF DECEPTION

This section discusses the related concepts and spectrum of deception. Deception can be defined and explained by a set of related terminologies in which those concepts should be defined and compared. Deception exists in our daily life in both verbal and nonverbal forms. Deception ranges a wide spectrum with varying intent and detectability (i.e., the extent of deception being detected).

1) KEY TAXONOMIES OF DECEPTION

In this section, we discuss a set of related terminologies related to deception. Most common concepts are defined in the dictionary and discussed in the cybersecurity literature [20], [35], [37], [158], [168].

- *Deceivee* [158]: The victim of a deception.
- *Deceiver* [158]: The perpetrator of a deception.
- *Susceptibility* [37]: Likelihood to be deceived.
- *Exploitation* [37]: The use of resources and benefit from them (e.g., damage to systems) by attackers.
- *Self-deception* [20]: A conscious false belief held with a conflicting unconscious true belief.
- *Trust* [37]: Reliance on the confidentiality and integrity from other sources and with confidence. Earning high trust from a deceivee can be easily exploited by a deceiver.
- *Lying* [35], [158]: Deliberate verbal deceptions. People often lie in pursuit of material gain, personal convenience, or escaping from punishment.
- *White lying* [168]: Normal standards for the lighthearted type of deception.
- *Belief* [37]: A truth in somebody's mind, truth basis.
- *Misbelief* [37]: A misplaced belief (i.e., mistakenly believing in false information)
- *Perception* [37]: The state of being aware of something through the senses.

2) SPECTRUM OF DECEPTION

In daily life and social networks, deception spans a spectrum of verbal and non-verbal behaviors. This section lists a few of the various deceptions based on [45], [158], [173].

- *White lies* [158]: Harmless lies to avoid hurting other's feelings and smooth relationships.
- *Humorous lies* [173]: Jokes that are obvious lies, such as practical jokes.
- *Altruistic lies* [158]: Good lies for protecting others, such as for preventing children from worrying.
- *Defensive lies* [158]: Lies to protect the deceiver, such as lies to get rid of repeated telemarketers.
- *Aggressive lies* [158]: Lies to deceive others for the benefit of the deceivers.
- *Pathological lies* [158]: Lies by a deceiver with psychological disorder.
- *Nonverbal minimization* [45]: Understating an important case in nonverbal camouflage.
- *Nonverbal exaggeration* [45]: Overstating an important case to hide others.
- *Nonverbal neutralization* [45]: Intentionally hiding normal emotions when inquired about emotional things.
- *Nonverbal substitution* [158]: Intentionally changing a sensitive concept with a less sensitive one.
- *Self-deception* [158]: Pushing of a reality into the subconsciousness.

FIGURE 2 represents the spectrum of deception from the lowest detectability to the highest detectability and from lowest bad intent (good intent) to no intent and to highest bad intent. In general, the deception with lower detectability are more with good intent, such as altruistic lies and white lies. Nonverbal deception is usually with bad intent and can be detected by professionals. Those behaviors can also be used as cues to detect lies. The deceptions with neutral intent can also be easily detected. These concepts can be applicable to detect malicious behaviors in online social networks as many offline human behaviors are also easily observed in online user behaviors.

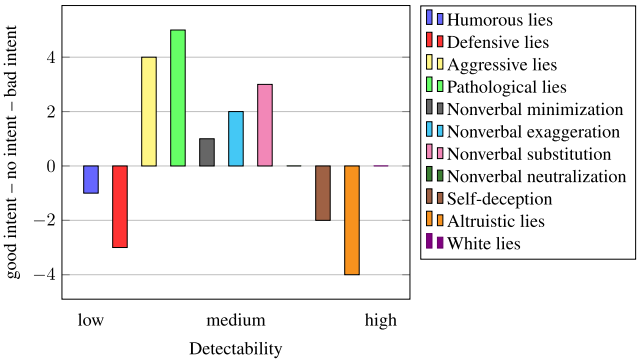


FIGURE 2. The spectrum of deception based on the extent of detectability of deception (x-axis); and the extent of good/bad intent of deception and no intent (y-axis).

D. PROPERTIES OF DECEPTION

Via the in-depth literature review, we observe the following **unique key properties of deception**:

- *Misleading one's belief*: Regardless of intent, deception can mislead one's belief which is actually false. Since deception as an action induces confusion or false information, false beliefs may be formed regardless of its intent or outcome.
- *Impact by deception*: Confusion or misbelief introduced by deception brings an outcome which can be negative or positive based on its original intent or its proper execution. However, when deception with a certain intent is not properly executed as planned or is used mistakenly, the outcome as its impact may not be predictable, resulting in high uncertainty (e.g., uncertain outcome). Hence, if deception is intended, it should be planned with multiple scenarios to lower down the risk introduced by deception in terms of a deceiver's perspective.
- *Success only by a deceivee's cooperation*: For deception to be successful, a deceivee should be deceived by the deception. Even if deception is performed but the deceiver detects the deception, no effect can be introduced.
- *Action as a strategy*: Deception can be used as a strategy to deal with situations with conflicts. The aim of the intentional deception is to mislead a target entity's belief and make the target choose a suboptimal (or poor) action that can be beneficial for the deceiver.
- *Signals as deception cues*: When deception is used, even if it can be very subtle, there exists some signals. Well-known deception strategies are to increase uncertainty (e.g., no signal increases uncertainty) or mislead one's belief (e.g., a false signal leads to false beliefs). Although both deception techniques aim to make a deceiver choose a wrong decision, if deception by misleading with false signal is detected, this provides more information about a deceiver to a deceivee than providing no signal.

Investigating the key properties of deception is critical in developing defense mechanisms to combat OSD attacks as the features of deception-based attacks, distinguished from other common OSN attacks. In this section, we discussed a variety of cues and susceptability traits of social deception behaviors across online and offline platforms. Thanks to the fast advances of social media and OSN technologies, many offline deception characteristics tend to be easily observed even in online deception behaviors. However, due to the limited real-time or interactions for feeling people's presence in online platforms with the current state-of-the-art SNSs and social media technologies, some physiological or psychological cues may not be applicable in detecting online social deception. In addition, upon the detection of the deception, a deceiver can easily get out of the online situation while a deceivee can easily lose a track of the deceiver. Now we look into various types of OSD behaviors currently studied in the literature.

TABLE 3. Classification of online social deception attacks.

OSD Class	Type	Description	Intent & Potential Damage	Source
False Information	Fake news	News contradicts, fabricates or conflates the ground truth and spreads in OSN.	Credibility loss, economical and political misleading, controlling public opinions	[86, 170, 200]
	Rumors	An unverified assertion that starts from one or more sources and spreads over time from node to node in a network.	Misleading people's decision, panic in public, government credibility loss	[199]
	Information manipulation	False information deliberately and often covertly spread in order to influence public opinion or obscure the truth.	Advertising, campaigns	[34]
	Fake reviews	Malicious users write fake reviews, opinions, or comments in social media to mislead other users.	Influencing user's option or decision, advertising, reputation loss	[224]
Luring	Phishing	Attackers trick users into revealing sensitive information related to work, financial credentials, or even personal data to be used in fraudulent activities	Confidential personal data leakage, launch advertising campaigns, pornography	[1, 40, 194]
	Spamming	Attackers send unsolicited messages (spam) in bulk to OSN users	Reputation loss, malicious advertising	[154]
Fake Identity	Fake Profile	Attackers create a huge amount of fake identities for their own benefits.	Personal information leakage, stealing money	[69]
	Compromised account	Attackers hacked legitimate user accounts that are created and used by their fair owners and later used for ill purposes.	Reputation loss, account loss, personal privacy leakage	[44, 92]
	Profile cloning attack	Attacker clones a pre-existing user profile either in the same OSN or a different one.	Reputation loss, sensitive information leakage, account loss	[154]
Crowdturfing	Crowdturfing	Attackers are gathered by crowdsoursing system and speak fake and inaccurate information to mislead people	Speading malicious URLs, forming astroturf campaigns, manipulating opinions	[109, 110, 205, 206, 218]
Human Targeted Attacks	Human trafficking	Traffickers use computers and networks to transport a great number of victims and advertise service across geographic boundaries for labor trade or sex trade	Sexual exploitation, modern slavery, forced labor or services, removal of organs	[51, 66, 105]
	Cyberbullying	Cyberbullying is the deliberate and repetitive online harassing or harming of someone.	Reputation loss, cyber harassment, teen depression	[154]
	Cyber-grooming	Cyber-grooming is when an adult tries to establish an online, emotional connection with a child in order to sexually abuse them.	Reputation loss, cyber harassment	[154]
	Cyberstalking	Attackers exploit their personal information, such as their phone number, home address, location, and schedule, in their SNS user's profile	Reputation loss, personal data leakage, cyber harassment, safety loss	[154]

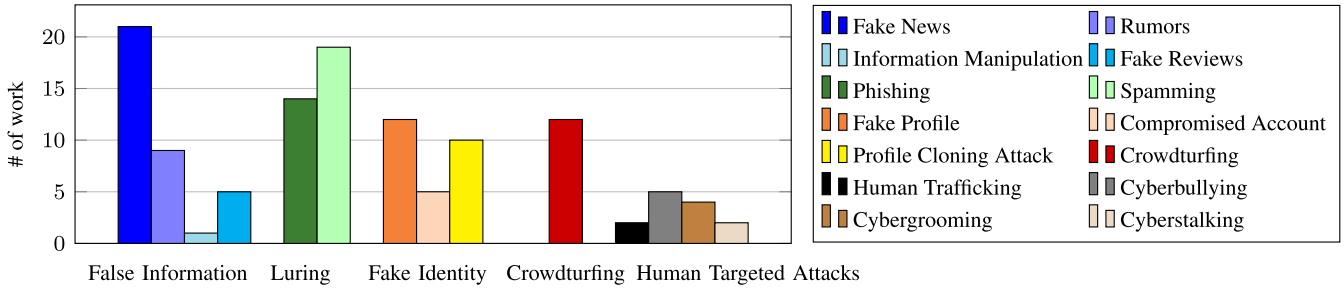


FIGURE 3. The number of works that studied different types of online social deception attacks based on five classes of online social deception. All surveyed works are summarized in TABLES 5-7.

III. TYPES OF ONLINE SOCIAL DECEPTION ATTACKS

Various types of OSD attacks have been discussed in the literature. In this section, we first classify various types of OSD attacks into five classes based on the key intent of each attack class. In addition, since the existing similar studies have used 'online social network attacks' and 'cybercrime' to discuss OSD, we discussed our view on how they are distinguished from and related to each other. All the OSD types are summarized in TABLE 3 and the corresponding work count for each OSD type is illustrated in FIGURE 3. Lastly, we discussed how OSD attacks breach security goals in CIA triad and safety with the aim to give an alert on how serious the OSD can be as a societal problem.

A. FALSE INFORMATION

False information on the web and social media can be classified as *misinformation* and *disinformation*. Misinformation can be considered as 'deception without intent' which mistakenly misleads people's belief due to the false information propagated. Disinformation can be categorized as 'deception with intent,' aiming to mislead people's beliefs. False information can be also categorized as *opinion-based* vs. *fact-based*. Opinion-based false information is propagated without ground truth. On the other hand, fact-based false information can mislead people's beliefs due to the fraud from ground truth, such as hoaxes and fake news in social media [86].

Although no formally accepted terminologies exist to distinguish different kinds of false information, we follow Jiang and Wilson [86]'s two criteria, which are *veracity* and *intentionality* [172], to discuss false information as below:

- *Fake News*: Fake news caused by serious fabrications or large-scale hoaxes [159] has wildly spread via OSNs since the beginning of the 2016 US presidential election cycle. Flintham *et al.* [55] reported that two third of survey respondents accessed news via Facebook. Facebook and Twitter have banned thousands of pages and identified as the major culprit of generating and promoting misinformation [86]. Fact-checking of news articles from different sources becomes a common means to determine the veracity of social media posts. Vosoughi *et al.* [200] found that fake news spread faster than truthful news. The time lag between fake news and fact-checking by fact-checking websites is 10-20 hours [170].
- *Rumors*: Vosoughi *et al.* [199] defined a rumor as an unverified assertion that starts from one or more sources and spreads over time from one user to another in a network. A rumor can be validated as true or false via real-time verification in Twitter or remain unresolved.
- *Information Manipulation*: One of the causes of information manipulation is opportunistic disinformation [34]. This means false information is deliberately and often covertly spread (e.g., planting a rumor) in order to influence public opinions or obscure the truth. Malicious users propagate opportunistic disinformation mainly for financial interest or political purpose.
- *Deceptive Online Comments or Fake Reviews*: Malicious users write fake reviews, opinions, or comments in social media to mislead other users. Usually fake reviews are classified as opinion-based false information [98]. Social bots are often used for automatically generating fake reviews [224].

B. LURING

Luring has been commonly used as one of popular deception strategies. The most common luring techniques in online worlds include:

- *Spamming*: Social media platform users can receive unsolicited messages (spam) that are ranging from advertising to phishing messages [154]. Malicious users usually send spam messages in bulk to influence many legitimate users.
- *Phishing*: Online phishing attacks, such as phishing webpages or phishing emails, are one type of cybercrimes that can lure users to reveal sensitive or credential information and steal private or financial information through social engineering attacks [40] or using other fraudulent, illegal activities [1]. These malicious activities can cause severe economic losses and threaten credibility and financial security of OSN users.

C. FAKE IDENTITY

Attacks using fake identity have their basis on social deception and include:

- *Fake Profile*: In OSNs, attackers create a huge amount of fake identities for their own benefits, which is also called Sybil attack. For example, in Facebook, attackers can leak out other users' personal information, such as e-mail and physical addresses, date of birth, employment data. Identity theft can take financial interests as well as access photographs of the friends of the victims [69].
- *Profile Cloning*: Attackers secretly can create a duplicate of an existing user profile in the same or different social media platforms. Since the cloned profile resembles the current profile, attackers can utilize the friend relationship and deceive and send friend requests to the contacts of the cloned user. By constructing the trust relationship with a potential victim user, the attacker can steal sensitive data from the user's friends. Profile cloning has exposed severe societal threats because attackers can commit more serious cybercrimes, such as cyberbullying, cyberstalking, and blackmail, which can introduce physical threats to potential victims [154].
- *Compromised Accounts*: Legitimate user accounts can be hacked and compromised by attackers [44]. Unlike Sybil accounts, compromised accounts are originally maintained by real users with normal social network usage history and have established social connections with other legitimate users.

D. CROWDTURFING

Malicious, paid human workers can perform malicious behaviors to achieve their employer's goal. This is called *crowdturfing*. For example, participants of an astroturfing campaign are organized by crowdsourcing systems [205]. Crowturfing gathers crowdturfing workers and spreads fake information to mislead people's beliefs and/or public opinions in social media. Crowdturfing activities in social media exploit social networking platforms (e.g., instant message groups, microblogs, blogs, or online forums) as the main information channel of the campaign [218]. Crowdturfing in social media is usually involved with spreading malicious URLs, forming astroturf campaigns, and manipulating public opinions. Usually it is challenging to detect crowdturfing accounts because their social media accounts are mixed with normal posts as a camouflage.

Chinese crowdsourcing sites [205] and Western sites [110] have been studied for the analysis of crowdturfing in campaigns. Three classes of crowdturfers (i.e., professional users, casual users, and middlemen) are identified in Twitter networks. In addition, their profiles, activities, and linguistic characteristics have been analyzed to detect crowdturfing workers [109]. Machine learning (ML)-based crowdturfing detection mechanisms have been considered in Wang *et al.* [206]. Two common types of adversarial attacks are evasion attacks (i.e., attacks changing behavioral features)

and poisoning attacks (i.e., administrators polluting training data) [206].

E. HUMAN TARGETED ATTACKS

Advanced online platforms have provided efficient tools for human targeted criminals to achieve their goals. The cybercriminals start their crime by establishing trust relationships with potential victims. Since this implies that these human targeted attacks are started based on social deception [76], we included the human targeted attacks as one of OSD classes considered in this survey.

The common human targeted OSD attacks include:

- *Human Trafficking*: Offline traditional human trafficking means traffickers kidnap the victims (mostly women and children) for trading with the purpose of labor exploitation or and sex trafficking [51]. Cyber-trafficking means that traffickers leverage cyber platforms for efficiently trafficking a great number of victims by using advertise services across geographic boundaries [66], [105].
- *Cyberbullying*: In this attack, an attacker commits the deliberate and repetitive online harassing of someone, especially adolescents [154]. Cyberbullying causes serious fear and harms for the victims through the online platforms involving deception, public humiliation, malice, and unwanted contact [43].
- *Cybergrooming*: In this attack, adult criminals attempt to establish trust relationships with potential victims, mostly female children, using online social media platforms. Their intent is to have improper sexual relationships with them or produce child pornography products [154], [226].
- *Cyberstalking*: Malicious users can exploit legitimate users' online information and harass them by stalking [154]. Without proper security protection of private information, individual users can expose their private information (e.g., phone number, home address, work location, etc.) in social media platforms without awareness.

F. RELATIONSHIPS BETWEEN ONLINE SOCIAL DECEPTION, SOCIAL NETWORK ATTACKS, AND CYBERCRIMES

Social network attacks, including traditional threats, social threats and multimedia content threats, are the general security threats concerned in the literature [154]. Those security and privacy threats include all the detrimental activities with malicious intent. Social deception is part of social network attacks, as shown in FIGURE 4, because social deception attacks can only be successful when the victims are being deceived from the attacker's perspective.

Four types of social network attacks are considered the OSD attacks: Unsolicited fake information attacks, identity attacks, crowdturfing, and human targeted attacks. The specific types of attacks were described in Section III. Some

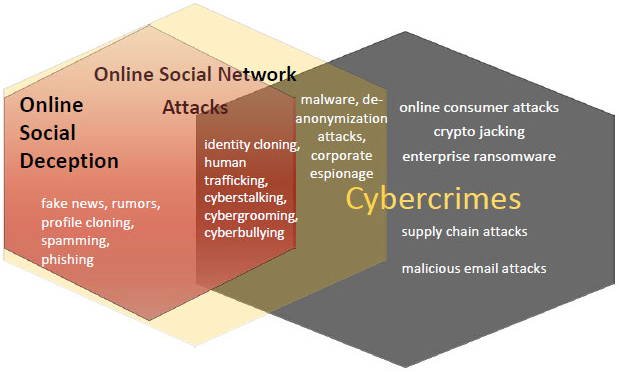


FIGURE 4. The relationships between OSN attacks, social deception, and cybercrime.

OSD attacks, such as personal and confidential information leakout, or identity theft, have been treated as cybercrimes [139] since early 2000's. The advanced features of social network service technologies further facilitated the significant increase of serious, sophisticated cybercrimes, such as human trafficking, online consumer fraud, identity cloning, hacking, child pornography, and/or online stalking [192].

FIGURE 4 illustrates the relationships between OSN attacks, OSD attacks, and cybercrimes. Although cybercrime is considered the most serious as cyberattacks, we can observe there are many attacks that overlap to each other. OSD attacks overlap either OSN attacks or cybercrime or both. Cybercrimes, such as consumer fraud, cryptojacking, enterprise ransomware, supply chain attacks, and malicious email attacks [179], fall in a separate group because these attacks are spread in Internet, which is much broader than OSN platforms. There are no explicit guidelines if certain OSN attacks or threats are illegal or if threats are illegal but their impact may not be direct. For example, when a user's data privacy (or integrity) is breached but no actually loss is found, it is hard to predict if there are future security concerns.

Although cybercriminals caused serious adverse effects to the society and individuals, 44% of the victims reported to the police [62]. Victims' reporting is a beneficial practice to increase the awareness of the communities to defend against potential cybercrimes. Victims may report to not only the police, but also the corporation in an active dialogue environment, or share the victim stories to families and close friends [62]. Cybercriminal profiling is highly challenging, compared to profiles of traditional criminals because cybercriminals can easily leave the platforms. However, it is very beneficial to identify common characteristics of cybercriminals [139] and useful for their early detection. Profiling can follow the procedure in the Behavioral Evidence Analysis [190]. Since most cybercrime victims are corporations and/or their customers, corporations can predict the potential insider criminals more intelligently with the help of cybercriminal profiling [139].

TABLE 4. Impact of online social deception attacks on loss of security goals and safety.

Social Deception Attack	Security Breach
Fake News	Data Integrity
Rumors	Data Integrity
Information Manipulation	Data Integrity
Fake Reviews	Data Integrity
Spamming	Account Confidentiality
Phishing	Account Confidentiality
Fake Profile	Account Integrity
Profile Cloning Attack	Authentication
Compromised Account	Account Integrity, Account Availability
Crowdturfing	Data Integrity, Network Integrity
Human trafficking	Confidentiality, Safety
Cyberbullying	Confidentiality, Safety
Cyber-grooming	Confidentiality, Safety
Cyberstalking	Confidentiality, Safety

G. EFFECT OF ONLINE SOCIAL DECEPTION ATTACKS ON SECURITY GOALS AND SAFETY

The CIA triad security goals play a major role in the information security practice. With the growth of socio-technical security issues, the original CIA triad is expanded with more specialized aspects, such as authentication and non-repudiation [122]. However, they still have limitations in systems and data for the wider organizational and social aspects of security [163]. OSN security has three levels of security goals: network-level, account-level, and message-level. Achieving the CIA security goals can contribute to all social network security levels. In addition to the three security goals, we also added another goal, which is safety. A person and other non-information based assets also needs to be protected in the cyber security practice [197]. For example, cyberbullying can cause direct physical harm to a victim even if there is no loss of information confidentiality, integrity or availability [197]. Therefore, we included human safety as a non-information security goal. For readers' convenience, we summarized how OSD attacks can breach security goals and safety in TABLE 4.

IV. CUES OF SOCIAL DECEPTION

In this section, we discuss various cues of social deception offline and online so that we can investigate how offline deception cues can be applicable in online deception cues. In addition, we aim to deliver insights on how the estimates of those deception cues can provide the key predictors of detecting online social deception.

A. INDIVIDUAL DECEPTION CUES

Riggo and Friedman [157] studied correlations between individual types and behavioral patterns and found individuals vary systematically in displaying certain behavioral cues (e.g., dominance, a social skills measure) are correlated with facial animation behavior. Certain types of individuals can control the display of cues to increase the likelihood of deception. Kraut and Poe [97] found that the occupational status and age were the top predictors of social deception.

B. CULTURAL DECEPTION CUES

Lewis and George [111] showed that individuals from collectivistic cultures were more apt to employ deception in business negotiation than those from individualistic cultures. Heine [75] discussed self-enhancement in Western people where self-enhancement refers to a motivation to make a person feel positive about himself/herself with a high self-esteem [167]. Bond *et al.* [14] showed in the lying settings, Jordanians displayed more behavioral cues than Americans in terms of eye contact and filled pauses.

C. LINGUISTIC DECEPTION CUES

Linguistic or communicative cues exhibiting deception in communications have been studied. Linguistic profiles are studied in deceptive communication, choice and use of languages, and linguistic patterns in deceptive messages [15], [16]. The example linguistic deception cues include use of more word quantity [72], [132], third-person pronounce use [182], use of emotion words, and markers of cognitive complexity (i.e., lying requires less complex cognitive process) [152].

D. PHYSIOLOGICAL DECEPTION CUES

Physiological or behavioral cues are the emotions in deceiving that liars are expressing because they are indicators of guilt [35]. In the studies of behavioral cues to deception [35] and physiological cues to identifying deception [201], liars may have at least one of emotions, content complexity, and attempted control phenomena. The examples of behavioral cues include less blinks or decreased hand and finger movement due to increased cognitive load [201], [202], [204], higher-pitched voices and faster speech [35], or displacement activities (e.g., high anxiety or conscious deception) [184].

E. PSYCHOLOGICAL DECEPTION CUES

Psychological or cognitive cues include nonverbal anxiety responses that are consciously revealed in the intentional deception [94]. Mitchell [124] described the mental process of deceptions from a social cognitive perspective based on children verbal deception and nonverbal deception in sports. Knapp et al. [94] used controlled lab settings to determine the characteristics of intentional deception with verbal and nonverbal cues. The example psychological cues include increased cognitive load [183], [201], [202], nervousness [35], [183], [201], or controlled behavior [183], [201].

Trivers [183] emphasized nervousness, control and cognitive load as three key deception cues. In addition, other anxiety responses are discussed [94]. Deceivers tend to exhibit cognitive cues, such as more uncertainty, vagueness, nervousness, reticence, dependence, and/or unpleasantness as a negative effect.

F. TECHNOLOGICAL DECEPTION CUES

Ferrara *et al.* [52] discussed the impact and detection of social bots which are the outcome of abusing new technologies.

Social bots with malicious intents caused several levels of damage to society. Early bots automatically posted content and can be spotted by the cues of a high volume of content generation. Several social honeypot approaches attracted social bots followers by carefully designed bots and analyzed the technology cues of social bots. However, sophisticated social bots are becoming more intelligent and tend to mimic human-like behaviors, making it hard to detect the social bots. The advanced detection strategy leveraged the technological cues from social graph structure, such as densely connected communities, and behavioral patterns. The proposed behavioral signature contains classes of features including network, user, friends, timing, content, and sentiment [52].

G. RELATIONSHIPS BETWEEN DECEPTION CUES OF OFFLINE AND ONLINE PLATFORMS

Via the in-depth survey of deception cues, we identified the commonalities and differences between online and offline deceptive behaviors as below.

1) COMMONALITIES BETWEEN ONLINE AND OFFLINE DECEPTIVE BEHAVIORS

Deception usually spreads via communication between deceivers and deceivees. The online media platforms support chat-based or synchronous communications similar to the traditional face-to-face chatting or interviews [187]. Interpersonal deception theory [16] discusses several verbal and non-verbal deception cues for traditional offline communications. Most of the verbal deception cues (e.g., linguistic cues) are relevant to both offline and online deception [36]. Messages and posts are the main source of online information so that the linguistic cues are most useful cues for online deception [230]. These days online platforms also provide face-to-face chatting. Although it is limited to some extent, some physiological cues and/or body movement can be captured.

2) DIFFERENCES BETWEEN ONLINE AND OFFLINE DECEPTIVE BEHAVIORS

Although face-to-face social media platforms make people feel much closer to each other by delivering body movement and facial expressions, feeling some physiological cues or subtle behavioral changes may not be captured like face-to-face interactions [187]. In addition, typing behavior (e.g., response time and the number of edits) for online chatting were studied as cues of online deception [36], which is not often observed in offline interactions. In addition, online behaviors are known different from offline behaviors in their motivations and attitudes [33].

V. SUSCEPTIBILITIES TO ONLINE SOCIAL DECEPTION

Attackers aim to achieve their attack goals as efficient as possible with minimum cost. To this end, the attackers may target highly susceptible people to the OSD attacks. In this section, we discuss various types of susceptability traits to the OSD attacks in order to help researchers develop protection tools for susceptible users in OSNs.

A. INDIVIDUAL OR SOCIETY-BASED SUSCEPTIBLE FACTORS

Demographic factors were studied to investigate the susceptibility to OSD attacks. Young age groups between 18 and 25 are known more susceptible to phishing than other age groups [171]. Young children were also identified as key potential victims to cybergrooming [12], [214]. Women are found more susceptible to phishing than men [171]. In particular, old women were found the most vulnerable populations to phishing [114], [141]. People's risk perception capabilities and knowledge about risk are shown as the key factor to prevent online deception [64], [118], [215], [220].

Personality traits are studied to investigate their impact on susceptibility to scams or phishing attacks [30], [48], [70], [71], [128], [148], [149] using the Big Five personality traits model [189]. However, due to the sample bias and lack of subjects covering a wide range of personality traits, the findings are not generalizable. In order to overcome the issues of limited sampling, Cho et al. [25] developed a mathematical model based on Stochastic Petri Nets to investigate the effect of user personality traits on phishing susceptibility. Ding *et al.* [39] classified phishing emails in terms of their corresponding target victims based on personality traits. Weir [6] also studied a user's susceptibility to social engineering attack by proposing a user-centric framework considering socio-psychological, habitual, socio-emotional, and perceptual user attributes.

Cultural factors have been studied as factors to influence susceptibility to OSD attacks. A well-known classification of cultural values is Hofstede's two cultural dimensions [77]: individualism vs. collectivism. In the individualistic culture, individuals are loosely tied to one another and a sense of 'I' and an individual's 'privacy' are valued. On the other hand, in the collectivistic culture, individuals are tightly connected emphasizing 'we-ness' and 'belongings' to each other. Since culture has been studied as a key factor impacting trust in a society where trust affects deceptive behavior, existing studies also have looked at how culture influences deception.

Social and economic factors are also studied as factors affecting the susceptibility to OSD attacks. Vulnerable status in a socio-economic ladder in the off-line world seems to be transferable to the online world. For example, low education and/or income may influence the level of knowledge and awareness about online social deception (or phishing) or related threat [90], [181]. However, there is a lack of empirical evidence to insist the relationships between individual characteristics related to social and economic status [90].

B. ONLINE ACTIVITY-BASED SUSCEPTIBLE FACTORS

Wagner *et al.* [203] found that a user's out-degree is identified as a key network feature social bots can target as their victim since higher out-degree in OSNs means more friends a user has. Susceptible users tend to be more active (e.g.,

retweet, mention, follow or reply) in the Twitter network and interact with more users, but their communication is mainly for conversational purpose rather than informational purpose. Susceptible users tend to use more social words and show more affection. Similarly, in Facebook, susceptible users tend to more engage in posting activities with less restrictive privacy settings, naturally resulting in higher vulnerability to privacy threats [70]. Social isolation (loneliness) and risk-taking online behaviors are the indirect factors of vulnerable people, such as victims of cybergrooming [211], [213]. Albladi and Weir [6] analyzed various user characteristics, such as a level of involvement, for vulnerability of social engineering attacks.

Engagement in social media is one of the most prominent attributes contributing to high susceptibility to social deception. Habitual use of social media measured by the size of social network and time spent in social media increases the likelihood of being victims for social attacks in OSNs [196]. Highly active social network users can be more favorable targets for attackers as they have more exposures to social media and accomplish their attacks through the active users' networks [6]. More use of social media is significantly associated with a higher level of risks for sexual exploitation [12], [214] and cyberbullying [41].

It is critical to look into what individual, cultural, network, or interaction traits introduce high susceptibilities to OSD attacks because protecting highly susceptible users first can be the key to prevent the OSD attacks. However, there has been little work that developed protection tools for susceptible users with high priority in the literature.

VI. PREVENTION MECHANISMS OF ONLINE SOCIAL DECEPTION

In this section, as proactive defense mechanisms, we discuss two types of OSD prevention mechanisms: *Data-driven prevention mechanisms* and *social honeypots*. The surveyed OSD prevention research works are listed in TABLE 5.

A. DATA-DRIVEN PREVENTION MECHANISMS

Prevention mechanisms against OSD attacks have been little explored. We discuss several types of data-driven prevention mechanisms that have been commonly used to deal with OSD attacks as follows:

- *Fake News Prevention*: Saad *et al.* [161] proposed a blockchain-based system to fight against fake news by recording a transaction in blockchain when posting a news article and applying authentication consensus of the record. The result was measured by an authentication indicator along with the post. In this design, when a user saw a post, the authentication indicator associated with the post was shown as the status of verification: successful, failed or pending. This mechanism addressed the following services for preventing fake news spread in the OSN: (i) Determine the authenticity of the news by users' consensus to ensure the trustworthiness of posts; (ii) identify a malicious user from the transaction record; and (iii) delete false information posts with a penalty applied to the fake news attackers. In general, the malicious attackers are the normal users but normal users do not have write access to the blockchain. Only the information source from a group of publishers or a group of a social network is allowed to commit transactions to the blockchain.
- *Phishing Prevention*: Florêncio and Herley [56] proposed a low-delay phishing prevention method where a client reports the reuse activities of user password in unknown websites and a server makes decisions and updates the blocked list. Gupta and Pieprzyk [68] proposed a defense model to classify web-pages on a collaborative platform PhishTank. This defense model uses a plug-in method into a browser to check blacklisting and blocking lists.
- *Identity Theft Prevention*: Tsikerdekis [186] discussed a proactive approach of identity deception prevention using social network data. Data in common contribution networks are used to establish a community's behavioral profile. Malicious accounts can be barred before joining a community based on the deviation of user behaviors from the community's profile.
- *Cyberbullying Prevention*: Dinakar *et al.* [38] proposed a dashboard reflective user interface in social network platforms for both cyberbullying attackers and victims. The reflective user interface integrated notifications, action delay, and interactive education. Their user study revealed that the in-context dynamic help in the user interface is effective for the end-users.

Pros and Cons: Preventing OSD attacks needs assessment of users or information in order to determine whether to allow the user or information can stay or be propagated in a given OSN. However, the so-called trust assessment is not clear. The key merit of the prevention mechanisms should be how quickly false information or malicious users are detected. Otherwise, it is not distinguishable from detection mechanisms. In addition, the effectiveness of the prevention mechanisms is still measured by detection accuracy. There should be more useful metrics that can capture the nature of proactiveness of the prevention mechanisms. In addition, no real-world implementation using the prevention mechanisms is considered, which limits applicability of the prevention mechanisms as well.

B. SOCIAL HONEYPOTS

Recently, the concept of good bots has appeared by creating social network avatars to identify malicious activities by highly intelligent, sophisticated attacks, such as advanced persistent attacks (APTs) [195]. *Honeypots technology* is not new and has been popularly used in communication networks as a *defensive deception* to proactively deal with attackers by luring them to honeypots for preventing them from accessing

TABLE 5. Online social deception prevention mechanisms.

Type	Method	Features	Datasets	Ref.
Fake news	Prototype of conventional Blockchain system	State tuple of unique transaction identifier, news payload, timestamp of news generation, hash of news payload, and user identifier	Synthetic data	[161]
Phishing	A server-client system to report password reuse and update whitelist	Password, user ID, domain in the protected list	Phishing attacks data from third party vendor in three weeks	[56]
	Blacklisting, heuristics and moderation based phishing prevention platform	Number of moderators, availability of moderators, unanimity of decision, network resources	Simulated data	[68]
Fake profile	Proactive sub-community behavioral profiles: support vector machine (SVM), random forest (RF), adaptive boosting	Closeness, betweenness, eigenvector centrality	Dataset *giftcardexchange* from Reddit banned users	[186]
	Weka toolkit and Decorate algorithm	Tweet similarity (TS) and 9 other content-based features; honeypot features: ratio of malicious accounts interacted (MAR), average daily new follower number of a social honeypot (DFN), social honeypots an account interacts with (AIN), and social honeypot follow back an account (AFB)	Seven types of Twitter accounts on the blue bird network: the Social Star, the Butterfly, the Distant Star, the Private Eye, the Cycler, the Listener and the Egghead	[135]
	LogitBoost, RF, XGBoost; evaluate robustness by individual attack model and coordinated attack model	4 sets: 3 profile features, 8 posting activity features, 2 page liking features, 3 social attention features. Temporal features: change rate of # of liked pages and of category entropies in 30 days	Fake Liker in Fiverr and Microworkers and legitimate liker in Facebook conference group	[9]
Spammer	Decorate, logistic regression (LR) and LibSVM	User profile features (longevity of account, average tweets per day, ratio of following/followers, percentage of bidirectional friends) and tweets features (number of URLs and @username in 20 most recent tweets)	MySpace and Twitter social honeypot deployment	[107]
	RF, standard boosting and bagging, feature grouping	Link payloads, user behavior over time, and followers/following network dynamics; 2 User demographics (longevity of account), 5 user friendship networks, 8 user content (average content similarity), user history (change rate of number of following)	60 Twitter honeypot accounts, Twitter dataset of content polluters and legitimate users from http://infolab.tamu.edu/data	[108]
	Random forest	Following/followers ratio, URL ratio, message similarity, friend choice, number of messages sent, friend number	900 MySpace, Facebook and Twitter honey-profiles	[177]
	Random forest	Tweet behavior (tweet frequency, tweet keyword, tweet topics), follow behavior, and application usage	Tweeter accounts	[222]
Socialbot	Account monitoring simulation, analysis of variance (ANOVA, $p = 0.05$)	Attack strategy (no knowledge attacker, partial knowledge attacker, full knowledge attacker); defense strategy: random, most connected, eigenvector, PageRank, and cost eigenvector 1 and 2	Stanford Large Network Data Collection, 50 communities in each of Friendster, LiveJournal, and Orkut	[145, 146]
	Statistical method	Profile: acceptance rate of friend requests sent, incoming friend request, insider's incoming friend requests, Discounted cumulative gain. Email: total received, spam and suspicious, from Xing, from LinkedIn, DCG score	7 Social Honeypot profiles in a European organization; messages in Xing, LindkedIn	[147]
	Equilibrium simulation	Honeybot deployment (HD); Honeybot exploitation (HE) and Protection and Alert System (PAS)	Generated network	[233, 234]
Cyberbullying	Dashboard reflective user interface: notifications, action delay, and interactive education	TF-IDF, Ortony lexicon for negative affect, list of profane words, part-of-speech tags, label-specific unigrams and bigrams	Two datasets from YouTube comments and Formspring with expert annotation	[38]

a target [27]. The existing approaches using *social honeypots* have mainly focused on detecting social spammers, socialbots [234], or malware [107], [108], [145]–[147], [177], [208] as a passive monitoring tool. These works use some profiles of attackers to detect them based on the features collected from the social honeypots placed as fake SNS accounts (e.g., Facebook or Twitter).

Although the original purpose of social honeypots was to proactively prevent attackers from accessing system/network resources, they have been used as a complement to detect various OSN attacks. However, the original purpose of social honeypots lies in a proactive intrusion prevention mechanism. In addition, although the social honeypots can be used as a detection tool for OSN or OSD attacks, their goal is an early detection or mitigation based on the proactive defense in nature. Hence, we include social honeypots as prevention mechanisms of OSD attacks.

For the social honeypots to be used as detection mechanisms, they are defined as information resources that monitor a spammer's behaviors and log their information (e.g., their profiles and contents in social networking communities) [107]. This early study detected deceptive spam profiles in MySpace and Twitter by social honeypot deployment. Based on the spammer they attracted, a SVM spam classifier was trained to identify spammers and legitimate users. An ML-based classifier was also developed to identify unknown spammers with high precision in two social network communities. Lee *et al.* [108] detected content polluters in Twitter by designing Twitter-based social honeypots. The 60 social honeypot accounts followed other social honeypot accounts and posted four types of tweets to each other. They investigated the harvested users to nine clusters via the Expectation-Maximization (EM) algorithm. They used content polluters classification by Random Forest and improved

the results by standard boosting and bagging and by different feature group combinations.

Haddadi and Hui [69] focused on privacy and fake profiles by characterizing fake profiles and reducing the threats of identity theft. They set social honeypots using the fake identities of celebrities and ordinary people and analyzed the different behaviors (e.g., a number of friends, friends requests, and public/private messages) between those fake accounts. Stringhini et al. [177] studied 900 honey-profiles to detect spammers in three social network communities (e.g., MySpace, Facebook, and Twitter) where their honey-profiles have geographic networks. They collected activity data for a long time (i.e., one year). In addition, this work identified both spam profiles and spam campaigns based on the shared URL.

Virvilis *et al.* [195] described the common characteristics of APT attackers and malicious insiders and discussed multiple deception techniques for early detection of sophisticated attackers. They created social network avatars in attack preparation phase (information gathering), along with fake DNS records and HTML comments. Zhu *et al.* [234] showed the analysis and simulation of infiltrating social honeybots defense into botnets of social networks. The framework SODEXO (SOcial network Deception and EXploitation) had three components: HD, HE, and PAS. The HD deployed a moderate number of honeybots in the social network. The HE modeled the dynamics and utility optimization of honeybots and botmaster by a Stackelberg game model. The results showed that a small number of honeybots could significantly decrease the infected population (i.e., a botnet) in a large social network.

Paradise *et al.* [145], [146] simulated defense account monitoring attack strategies in OSNs. The attackers sent friend requests to some community members chosen by different attacker strategies. In addition, the attackers may have full knowledge of the defense strategies. The defender chose a set of accounts to monitor based on various criteria. They analyzed the acceptance rate, hit rate, a number of friends before hit, and monitored cost between combinations of attackers and defenders. The result showed that under the sophisticated attackers with the full knowledge of defense strategies, defense using PageRank and most connected profiles had the best detection with minimum cost. Paradise *et al.* [147] targeted at detecting the attackers in the reconnaissance stage of APT. The social honeypot artificial profiles were assimilated into an organizational social network (Xing and LinkedIn) and received the friend requests to organization employees. The authors analyzed the attacker profiles collected in the social honeypot.

Badri Satya *et al.* [9] collected the so called 'fake likers' on Facebook, who are paid workers to propagate fake likes using linkage and honeypot pages. The authors extracted the four types of profiles and behavior features and trained classifiers to detect the fake likers. The temporal features were cost-efficient compared to the previous research. They also evaluated the robustness of their work by modifying features using individual attack model and coordinated attack model. De Cristofaro *et al.* [31] studied paying for 'likes fraud' in Facebook and linking the campaigns to honeypot pages to collect data. They analyzed the page advertising and promotion activities. Nisrine *et al.* [135] discovered malicious profiles by social honeypots and used both feature-based strategy and honeypot feature-based strategy to collect data. Combining honeypot features can increase the ML accuracy and recall, compared to the scheme with traditional features only.

Zhu [232] defined ''active honeypots'' as active Twitter accounts, which capture more than 10 new spammers everyday, similar to the spammer network hubs. They extracted 1,814 those accounts from the Twitter space and studied the properties and identification of active honeypots. Yang *et al.* [222] deployed passive social honeypots to capture spammers' preferences by designing social honeypots with various behaviors. The design considered tweet behavior (i.e., tweet frequency, tweet keywords, and tweet topics), followed behaviors of famous people's accounts and application installation. They analyzed which type of social honeypots has the highest capture rate and designed advanced social honeypots based on their results. They demonstrated that the advanced honeypot can capture spammers 26 times faster than the normal social honeypots.

Pros and Cons: Social honeypots would be highly effective particularly when it is well deployed to attract targeted attackers. However, so far, the existing studies discussed above did not consider key, unique characteristics of vulnerable victim profiles to develop social honeypots. The effectiveness of existing social honeypots is evaluated based on intrusion detection accuracy rather than the coverage of attack types or the main attack types attracted to the social honeypots. Since an individual honeypot did not target a particular attack, it is not clear what types of attackers are more attractive to certain characteristics of the social honeypots from the existing approaches. In addition, developing social honeypots with fake accounts may introduce ethical issues because the use of the social honeypots itself is based on deceiving all other users as well.

VII. DETECTION MECHANISMS OF ONLINE SOCIAL DECEPTION

Most existing defense mechanisms against OSD attacks focus on detecting those attacks. We discuss those detection mechanisms based on three types: *user profile-based*, *message content-based*, and *network feature-based*.

A. USER PROFILE-BASED DECEPTION DETECTION MECHANISMS

Most profile cloning studies utilized the user profiles [91], [95], [169]. To identify cloned profiles, they calculated profile similarities using various methods based on user profile attributes. Kontaxis *et al.* [95] proposed three components to detect profile cloning: an information distiller, a profile hunter, and a profile verifier. The profile verifier component

TABLE 6. Data-driven deception detection mechanisms.

Type	Method	Features	Datasets	Ref.
Spam URL	Random forest	Behavioral factor of URL posting (posting count, posting standard deviation, posting intensity, posting user network) and click (rises+falls, spikes+troughs, peak difference, total clicks, average clicks, clicking days, max clicks, effective average clicks, click standard deviation, mean median ratio)	List labeled dataset, URL-category website URLBlacklist (http://urlblacklist.com) and manually labeled dataset	[17]
Spambot	Longest common substring (LCS), supervised and unsupervised classifier	Twitter account behavior (type (tweet, reply, retweet) and content (entities in tweets)) as DNA string of characters, and LCS curve, LCS value	Two Twitter dataset	[28]
Spam	SVM, adaboost, and random forests	Topic distribution (LDA) for each user, two new topic based features: Local Outlier Standard Score (LOSS) and Global Outlier Standard Score (GOSS)	Soical honeypot Twitter dataset from [107], synthesized Weibo dataset	[115]
	Labeled latent Dirichlet allocation (L-LDA) model and SVM	Word-based (*TF-IDF* scheme), topic-based (group of words that have a high probability of co-occurrence, normalized topic frequencies), and user-based features (average time interval of posting (ATI), and the average similarity (AS) of two adjacent comments)	YouTube social spam dataset	[175]
	Random forest, C4.5 decision tree, Bayes network, naive Bayes, k-nearest neighbor, and support vector machine	12 lightweight features: user-based (account_age, no_of followers, no_of followings, no_userfavourites, no_lists, and no_tweets) and tweet-based (no_retweets, no_hashtags, no_usermentions, no_urls, no_chars, and no_digits) and feature discretization	Four datasets from Twitter's streaming API with spam to non-spam ratios and continuous sampling method, ground truth from commercial tool	[21]
	Naïve Bayes, logistic regression, RF and semi-supervised spam detection	Hashtag, content, user and domain features	HSpam14 dataset of 15 days of tweets	[166]
	Semi-supervised clue fusion algorithm, boosting-based fusion	Content, behavior (variance of posting times during a period, night activity, regularity of posting), relationship (ratio of follower to followee, average number of neighbors' messages/ followers), and interaction features (average number of comments per message, average number of repost, mentions fraction)	Data crawler from Weibo API	[22]
Phishing	CNN-LSTM algorithm, XGBoost	URL deep features, URL statistical features, webpage code features, webpage text features	Two historical data were crawled from PhishTank website	[223]
	Search-engine based, heuristic-based and logistic regression	Phishing vocabulary similarities, 37 ULR lexical features (information entropy, confused string, length)	Data source from PhishiTank, Yahoo, URLB and DMOZ	[40]
Fake comments	Markov chain model on topic-crowd opinion pairs	Second-order Markov chain probabilities; five topics (Information, Science, Entertainment, Humour, and Adult) and three Crowd opinions (Positive, Negative, and Neutral)	User comments from Reddit; test dataset with 300 automated texts	[49]
	ComLex: word embedding and unsupervised spectral clustering. linear regression, SVM and nearest neighbors (NN)	ComLex linguistic signals from user comments, keep emojis, special tokens *snopesref* or *politifactref*, 100-dimention vector; EmoLex; Linguistic Inquiry and Word Count (LIWC)	5303 social media posts from Politifact and Snopes with 2614374 user comments from Facebook, Twitter, and YouTube	[86]
Rumor	Logistic regression, SVM, naive Bayes, and decision tree	Five new features from rumor publisher's behavior (verified user or not, number of followers, average number of followees per day, average number of posts per day, and number of possible microblog sources) with existing behavior-based features as follower's comments and reposting	Published rumor data from Sina Weibo	[113]
	Temporal model: dynamic time warping (DTW), and hiddern Markove models (HMM); non-temporal model: SVM and logistic regression	Time-series features: 4 linguistics (ratio of tweets containing negation, average formality & sophistication of tweets, opinion & insight (LIWC), inferring & tentative tweets), 6 user involved (controversiality, originality, credibility, influence, role, and engagement), and 7 temporal propagation dynamics features, feature contribution is ranked by Chi-squre test	209 manually annotated rumors from Snopes.com and FactCheck.com and Twitter historical API	[199]
	One-class support vector machine, support vector data description, k-nearest neighbors, principle component analysis (PCA), k-means, autoencoder	Content features (13 syntactic and 13 semantic), 18 user profile features, and 8 meta-message features	Two published Twitter datasets	[50]

calculated the profile similarity score between testing social profiles and the user's original profile. Both the information field and profile pictures contributed to estimating the profile similarity. Kamhoua *et al.* [91] detected user profiles across multiple OSNs in a supervised learning classifier. The method consists of three steps: the profile information collection from a friend request, the friend list identity verification, and the report of possible colluders. The binary classifier was based on both the profile attributes similarity and friend list similarity. Shan *et al.* [169] simulated profile cloning attacks by snowball sampling and iteration attack and then detected the attackers by a detector called 'ChoneSpotter.' The context-free detection algorithm includes the profile information and friendship connections. The input features include recently used IPs, a friend list, and the profile and its similarity. A cloned profile was determined by using the same IP prefix and the similarity over a certain threshold.

User profile features and user behavior/activity features were extracted to detect malicious accounts [9], [17], [28], [113], [147], [175], [207] in Sybil attacks, fake reviews, or spamming attacks. Badri Satya *et al.* [9] studied the feature engineering from the account of 'fake likers.' They

TABLE 6. *(Continued.)* **Data-driven deception detection mechanisms.**

Type	Method	Features	Datasets	Ref.
Fake news	Deep Recurrent Neural Networks (RNN) model	Neural embedding, n-gram TF vector, and external features including polarity, lexiton based sentiment difference	FNC-1 challenge from Emergent dataset [53]	[13]
	Naïve Bayes Multinomial algorithm	Basic word usage pattern (word vector); seven themes (code book) between fake news and satire	Created: https://github.com/jgolbeck/fakenews	[61]
Profile cloning	A binary classifier calculating the attributes and friend list similarities from different OSNs	Profile attributes similarities, friend list similarity, friend request information, friends lists	Synthetic dataset of 2000 people's profiles	[91]
	Three components: Information Distiller, Profile Hunter, and Profile Verifier	User-identifying terms, profile-records, profile similarity	LinkedIn automated profile creation	[95]
	CloneSpotter: A real-time context-free detection algorithm	Recently used IPs, friend list, profile, profile similarity	Synthesized accounts in Renren network	[169]
Fake account	Decorate emsemble classifier	Blacklist: 50 top LDA topic words, 500 fake word from TF-IDF. 14 content-based features: fake word ratio, mean time between tweets, extreme idle duration between tweets	1KS-10KN dataset and social honeypot dataset on Twitter	[178]
Spam account	Gradient boosting, RF, extremly randomized trees (ExtraTrees), maximum-entropy (MaxEnt), multi-layer perceptron (MLP), SVM	Lightweight features: user profile features (user name, screen name, location and description), account features (account age and verification flag), pairwise engage-with features (user activities) and engaged-by features (indirect from users)	Honeypot dataset from [108] and manually annotated dataset	[82]
Malicious / compromised account	Weka	21 account features from contents and account activity behaviors and Petri net based features	Twitter data	[162]
	SVM	Semantic representation of clickstreams	Synthetic data banksim and paysim	[209]
	k-NN	n-gram built baseline to identify writing style and identify users; keep updating the baseline with new posts	Twitter dataset of 1000 users: https://wiki.illinois.edu//wiki/display/forward/Dataset-UDI-TwitterCrawl-Aug2012#Dataset-UDI-TwitterCrawl-Aug2012-4.Creation	[10]
	CADET: nonlinear autoencoders	Feature embeddings from tweets content, source, location, and timing	Twitter data of posted geotagged tweets	[193]
Crowdturfing	Random forest, decision tree, SVM, Bayesian probability models	9 user profile fields (FFRatio, reciprocity, user tweets per day, account age, ratio of tweets with URLs and mentions), 8 user interactions (comments and retweets), 5 tweeting clients (devices), and temporal behavior (12 tweet burstiness and 1 entropy regularity)	Two baseline datasets: authenticated dataset and active users from Sina Weibo accounts from three-year crowdturfing campaigns;	[206]
	CrowdTarget to detect crowdturfing targets: Ada Boost, Gaussian naive Bayes, k-nearest neighbors	Retweet-based features from crowdturfing targets: retweet time distribution (mean, standard deviation, skewness, and kurtosis); ratio of the most dominant application; number of unreachable retweeters; ratio of number of received clicks to the number of retweets for tweets containing URLs	Twitter, crowdturfing sites, and five black-market sites, e.g. `retweets.pro` and `socialshop.co`	[174]
	Character-level RNN, SVM linguistic classifier	Temperature, a parameter used in softmax function; character-level probability distribution $P(X_{t+1} = x_{t+1}\|x_{1,\ldots,t})$; 1 similarity feature, 4 structural features, 6 syntactic features, 4 semantic features, 62 LIWC features	2017 Yelp Challenge Dataset https://www.yelp.com/dataset/challenge and Attack dataset, replacing fake reviews with RNN generated reviews	[224]
	Random forest, Naïve Bayes, logistic regression, SVM	Four groups 92 features: user demographics 5 features, user friendship networks 4 features, user activity 12 features (behavioral), user content 3 features including 68 LIWC dictionary	Random sample 10,000 twitter users	[109]
Cybercrime account	Content-sensitive Gibbs sampling (CSLDA)	Semantic labels (transactional or collaborative), Laplacian semantic ranking score	2 cybercrime related corpora from Twitter and online forums	[106]
Cyberbullying	JRip, J48, SVM from Weka tool	TF-IDF, Ortony lexicon for negative affect, list of profane words, part-of-speech tags, label-specific unigrams and bigrams	Two datasets from YouTube (comments) and Formspring (young people) with expert annotation	[38]

considered profile features, such as the length of user introduction, the longevity of an account, and the number of friends. Social activities represent a unique attribute observed in OSNs and consist of the behavior features of an account, such as sending friend request, posting, retweeting, liking/disliking and social attention [9]. More specific features under each activity category can be further extracted, such as the acceptance of a friend request sent from [147] and the average time interval of posting from [175]. Wang *et al.* [207] investigated several behavioral signatures for the output of crowdturfing campaigns and tasks. Cao and Caverlee [17] studied the behavioral features to detect spam URLs in OSNs. They used fifteen click and posting-based features in Random Forest classifiers and evaluated the top six features.

Cresci *et al.* [28] proposed a novel DNA-inspired social fingerprinting approach of behavioral modeling to detect spambot accounts. Twitter account behaviors were encoded as a string of behavioral units (e.g., tweet, reply and retweet). This new model can deal with the new type of spambots which can be easily missed by most traditional tools. Social

fingerprinting sequences are characterized by the LCS curve. Spambots are related to high LCS values by sharing suspicious long behavioral patterns. The LCS curve from behavioral model is used to detect more sophisticated types of crowdsourcing spammers.

User profiles and activities are the key features to detect OSD attacks (e.g, advanced spammers or crowdturfing), along with other content-based and graph-based features [82], [107]–[109], [199], [206]. We will discuss those hybrid detection examples in Section VII-D.

Pros and Cons: User profile information provides specific activity features and behaviors about each user. However, some profile information is private; thus, collecting private information itself is the violation of a user's privacy right. In addition, even if the information itself is open to the public, how to use the information should be agreed with the owner of the information. Since each user enters his/her profile information, if the user is malicious, it is easy to enter fake information for making self-presentation look attractive, which is one of self-deception. Besides, collecting profile and behavioral data incurs high cost and/or time under privacy protection of the social media platforms.

B. MESSAGE CONTENT-BASED DECEPTION DETECTION MECHANISMS

In TABLE 6, we showed that the majority of social deception detection approaches have used content-based features because the text of user posts and reviews can be easily collected and analyzed using existing linguistic models. The proliferation of social media and/or network applications allowed numerous types of raw and advanced content features available. Topic modeling and sentiment-based features have been popularly utilized for the linguistic analysis of deceptive messages.

1) TOPIC MODELING-BASED DETECTION

Most of the work developed topic distributions by using Latent Dirichlet Allocation (LDA) [106], [115], [175], [178], [217]. If each user's posts are collected as a document, LDA generates the topic probability distribution of the user's document. Liu *et al.* [115] extended the topic features to two new features. A GOSS indicates a user's interests in specific topics, compared to other users while a LOSS indicates a user's interests in various topics. By adding those two topic-based features to classifiers, the averaged F1-score shows better performance. Swe and Myo [178] built a keyword ''blacklist'' to detect fake accounts by extracting topics from LDA and keywords from TF-IDF (term frequency-inverse document frequency) algorithms. The blacklist contributed to 500 fake words. The number and ratio of fake words and a few other content-based features were extracted for their classifier. The result using a ''blacklist'' showed better accuracy than the traditional spam word list by reducing false positive rate. Wu *et al.* [217] extracted the topic distribution of 18 topics for one message following the official Weibo topic categories. The probability of 18 topics was used as one feature vector for the SVM classifier.

The LDA algorithm has been enhanced to detect cybercriminal accounts and spams. Lau *et al.* [106] developed a weakly supervised cybercriminal network mining method supported by a probability generative model and a novel context-sensitive Gibbs sampling algorithm (CSLDA). The algorithm can extract the semantically rich representations of latent concepts to predict transactional and collaborative relationships (e.g., cybercriminal indicator) in publicly accessible messages posted on social media. Song *et al.* [175] used Labeled LDA (L-LDA) to indicate the probability of co-occurrence. The latent topics were normalized to topic-based features, which have distinct properties with TF-IDF generated word-based features.

Golbeck *et al.* [61] detected two types of false article stories, which are fake news and satires by themes and word vectors. Then they defined a theme by a new codebook with 7 theme types, such as conspiracy theory and hyperbolic criticism. Multiple themes can be labelled to an article as a theme coding. The proposed classifier worked better for articles under a certain type of theme.

Pros and Cons: The topic features can be easily obtained. However, there would be unique network features distinguishing attackers from normal users. That is, the content-only features may not be able to capture other features of dynamic interactions with other users, such as likes, friend acceptance, or frequency of leaving comments or sharing. In addition, topic models are highly sensitive to datasets and topic models may perform differently in detection accuracy depending on datasets.

2) FEATURE-BASED DECEPTION DETECTION

TABLE 6 lists the feature set used by the papers surveyed in this work. The commonly used features include raw features, such as word vector, word embedding, hashtags, links and URLs [119]. Advanced features include deep content features, statistics, LIWC and other metadata, such as location, source, or time [193]. Most ML-based models use supervised learning. Among the supervised models, random forest, SVM, Naïve Bayes, logistic regression, and k-nearest neighbors are the most favorable classifiers for detection. Neural networks models, such as Recurrent Neural Networks [224] and Convolutional Neural Networks with Long Short-Term Memory (CNN-LSTM) [223], are used for textural features. Temporal models, such as DTW and HMM [49], [199], are discussed in rumor detection. The boosting-based ensemble models are implemented for spammer detection [82], [223]. A few studies used semi-supervised models [82], [166] when the labeled dataset was not available.

Everett *et al.* [49] studied the veracity of the automated online reviews provided by regular users. They used the text generated by second-order Markov chain model. The key findings include: (i) The negative crowd's opinion reviews are more believable to humans; (ii) light-hearted topics are easier to deceive than factual topics; and (iii) automated

text on adult content is the most deceptive. Yao *et al.* [224] investigated attacks of fake Yelp restaurant reviews generated by an RNN model and LSTM model. The model considers the reviews themselves only, not including metadata as reviewers. Similarity feature, structural features, syntactic features, semantic features, and LIWC features were used in SVM to compare the character-level distribution. They found that information loss was incurred in the process of generating fake reviews from RNN models and the generated reviews can be detected against real reviews. Song *et al.* [174] detected crowdturfing targets and retweets from crowdturfing websites and black-market sites.

Pros and Cons: Feature-based models generate high accuracy and low false positive rates. The raw content features are easily obtainable although the extraction of sophisticated features incurs high cost. However, the temporal pattern of messages influences the detection performance. The semantic analysis methods may ignore hidden messages and background knowledge and require tuning many input parameters, which leads to high complexity and labor-intensive.

3) SENTIMENT-BASED DECEPTION DETECTION

Sentiment of social media messages serves as extra features of message contents. Sentiment provides emotional involvement, such as like, agree, or negation, calculated by lexicon analysis [13], [38], [79], [86], [198]. Jiang and Wilson [86] introduced a novel emotional and topical lexicon, the so called *ComLex*. The authors analyzed the linguistic signals in user comments, regarding misinformation and fact-checking. Specifically, they discussed the signals from user comments to misinformation posts, veracity of social media posts, or fact-checking effects. There are signals for positive fact-checking effect as well as signals (e.g., increased swear word usage) indicating potential ''backfire'' effects [138], where attempts to intervene against misinformation may only entrench the original false belief.

Sentiment features are often used along with TF-IDF word vectors. Supervised classifiers in current research utilize sentiment analysis to improve prediction. Bhatt *et al.* [13] detected fake news stances from neural embedding, *n*-gram TF vector and sentiment difference between news headline-body TF vector pair. Dinakar *et al.* [38] proposed a sentiment analysis to predict bullying, aiming at discovering goals and emotions behind the contents. Note that Ortony lexicon [144] maintains a list of positive and negative words describing the affect. The lexicon of negative words was only added in the feature list to detect bully-related rude comments.

Pros and Cons: Sentiment analysis includes more emotional and background information, in addition to the explicit content, which can increase the prediction accuracy, when compared to semantic-only methods. However, the use of sentiment analysis cannot fully leverage the linguistic information in the contents where the lexicon is domain-specific. In addition, more elaborated dimensions of emotions or sentiments should be considered in order to capture fake information and its intent.

C. NETWORK STRUCTURE FEATURE-BASED DETECTION

Several general network features were extracted in supervised learning methods, such as topology, node in-degree and out-degree, edge weight, and clustering coefficient [100], [155], [199]. Wu *et al.* [216] summarized false information spreader detection based on network structures. Ratkiewicz *et al.* [155] built a *Truthy* system to enable the detection of astroturfing on Twitter. The proposed Truthy system extracted a whole set of basic network features for each meme and sent those features with a meme mood by sentiment analysis to the supervised learning toolkit. Kumar *et al.* [100] developed four feature sets, including network features to identify hoaxes in Wikipedia. The network features measure the relation between the references of the article in the Wikipedia hyperlink network. The performance of features sets was evaluated in a random forest classifier.

In the following sections, we discuss algorithms and supervised learning methods specifically designed for the network structure, such as propagation-based models, graph optimization algorithms, and graph anomaly detection algorithms. TABLE 7 lists all the surveyed works under Section VII-C.

1) EPIDEMIC MODELS

Epidemic model is a direct way to model and simulate the diffusion of disease [131]. Since the spread of disease in a certain population is similar to the propagation of false information in the social media communities, epidemic models have been often modified to quantify the extent of false information propagation [87]. The epidemic models are agent-based, where an individual node is modeled as an agent. Different types of agents are characterized by distinct states and behaviors, such as the agents Susceptible (S), Infectious (I), and Recovered (R) in the traditional SIR (Susceptible, Infectious, and Recovered) model [129] in false information propagation. In OSNs, agents in the SIR model represent a group of users in each state as follows: (i) *Susceptible* (*S*): Users who have not received information (e.g., rumor posts or fake news) yet but are susceptible to receive and believe it; (ii) *Infectious* (*I*): Users who received the information and can actively spread it; and (iii) *Recovered* (*R*): Users who received the information and refuse to spread it [227].

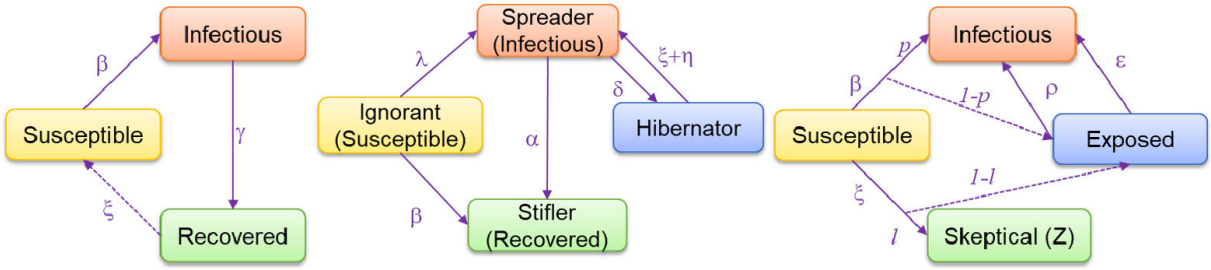
The state transitions are S to I by infection rate β, and I to R by recovery rate γ depicted in FIGURE 5a. The current false information propagation research has two tracks employing the epidemic models: (i) Adding more links and parameters to the traditional SIR model; or (ii) Building SEIZ model (Susceptible, Exposed, Infected, and Skeptic–Z; discussed below) to fit to the OSN data.

a: SIR MODEL WITH VARIATIONS

Many variants of the basic SIR models have been proposed in the current false information propagation research. Zhao *et al.* [227] added forgetting mechanisms to

TABLE 7. Network structure-driven deception detection mechanisms.

Type	Method	Features	Datasets	Ref.
Fake news	Opinion model with subjective logic, epidemic model	Opinion consisting of belief, disbelief and uncertainty, agent features (prior belief and centrality degree), agent types (disinformers, misinformers and true informers)	Facebook dataset with 1033 nodes and 26747 edges	[26]
	Hierarchical propagation model	Three-layer credibility network of event, sub-events and messages	Microblog datasets SW-2013 and SW-MH370	[88]
	Credibility propagation network	News verification by mining conflicting viewpoints	Built upon Sina Weibo	[89]
	Independent cascading model and Bayesian inference	User's flagging activity, user's observed activity, utility of blocking a news	Social circles Facebook graph of 4039 users and 88234 edges	[185]
	Cascades representation, LSTM-RNN sequence classifier	Embedding of users	Twitter API used for certain topics and the dataset consisting of 68892 news with 288591 posts and 121211 users	[219]
Fake reviews	FraudEagle algorithm: scoring by loopy belief propagation (LBP) and grouping by cross-association clustering	User nodes and product nodes with signed links and prior belief	New SoftWare Marketplace (SWM) dataset from an app store database	[5]
	Random forest	Linguistic traits, activities, reply network structure	Disqus communities datasets	[101]
	REV2, unsupervised and supervised RF prediction	Quality/trust scores from fairness, goodness and reliability for users, ratings, and products	Flipkart, Bitcoin OTC, Bitcoin Alpha, Epinions, and Amazon dataset	[102]
Rumor	SEIZ model	4 topics and 5 geographic regions, state Exposed (E) users taking some time to post, state Skeptic (Z) users who heard a news but decided not to retweet, retweet cascades	8 Twitter dataset for real news and rumors	[87]
	Correct information diffusion in SIR-extended diffusion model	SIR agents, condition to become R, diffusion of corrected information and new situations	Tweets collected from two weeks before and after Great East Japan Earthquake	[140]
	Propagation structure via tree kernel	Cascades similarity, SVM-time series, decision tree based ranking, random forest	Twitter15 and Twitter16 data	[116]
	Cascades representation, LSTM-RNN	Bottom-up tree, top-down tree	Twitter15 and Twitter16 data	[117]
	Temporal model: DTW and HMM; non-temporal model: SVM and logistic regression	8 temporal propagation dynamics features (time-inferred diffusion, fraction of low-to-high diffusion, fraction of nodes in largest connected component, average depth to breadth ratio, ratio of new users, ratio of original tweets, fraction of tweets containing outside links, fraction of isolated nodes)	209 manually annotated rumors from Snopes.com and FactCheck.com and Twitter historical API	[199]
	Cascades similarity, hybrid SVM with graph kernel and RBF	23 features from propagation tree and 8 new features (topic type LDA, search engine, user type, ave sentiment, ave doubt, ave surprise, ave emotion, and re-post time)	Sina Weibo data of 2601 false rumors from the official management center	[217]
Misinformation	Profit minimization of misinformation (PMM), influence maximization	Total activity profit of edges	Three real-world datasets from existing work: soc-wiki-Vote, p2p-Gnutella08 and ca-HepTh	[23]
Astroturfing	Klatsch data model, Gephi toolkit	Topology of the largest connected component, number of nodes and edges, mean degree and strength of nodes, mean edge weight, clustering coefficient of largest connected coefficient, standard deviation and skew of in-degree/out-degree	Dataset from Tweets of political keywords, Yahoo Meme, Google Buzz	[155]
Spam	Malicious relevance score propagation algorithm, criminal account inference algorithm	Social relationship graph: criminal supporters community (social butterflies, social promoters, dummies), social relationships and semantic coordination	Twitter half million account (2060 spammer, 5924 criminal supporters)	[221]
	User ranking algorithm Collusionrank to penalize users who connected to spammers	Node rank, followers, indegree, outdegree, indegree/outdegree ratio, social capitalists	Twitter dataset	[60]
	SSDM: directed Laplacian formulation to model social networks, SVM and elastic net (EN)	Network information by adjacency matrix	Crawled Twitter dataset from Twitter search API	[78]
	Lockstep propagation algorithm	Adjacency matrix, "block," "ray" and "pearl" subspace, lockstep score	Weibo network with 100 million nodes, synthetic data	[85]
	CatchSync algorithm	Topology	Twitter and Weibo dataset, synthetic data	[84]
	Troll identification algorithm (TIA), five de-cluttering operations	Centrality measures: Freaks, Fans Minus Freaks (FMF), PageRank (PR), Signed Spectral Ranking (SSR), Negative Ranking (NR), Singed Eigenvector Centrality (SEC), Modified HITS (M-HITS), Bias and Deserve (BAD)	Slashdot Zoo dataset	[99]
	Incremental tensor analysis	Tensor decomposition and deflation	Phone-call network data, computer-traffic network	[8]
Sybil	MailRank algorithm: basic and personalized scores	Email interactions link analysis, sender rating by social reputation score based on PageRank score	Synthetic dataset	[24]
	A near-optimal defense algorithm SybilLimit with node trust ranking	Undirected social network with nodes and trust relations, each node consisting of a suspect and a verifier	Three crawled datasets from Friendster, LiveJournal, DBLP and Kleinberg	[225]
Abnormal nodes	OddBall algorithm, scalable and unsupervised	Features for egonets: density, weights, ranks and eigenvalues	Bipartite network: Auth2Conf from DBLP, Don2Com and Com2Cand from donations of political candidates	[4]



(a) SIR Model (b) SIHR Model (c) SEIZ Model

FIGURE 5. Three types of agent-based epidemic models. The solid line arrows are transitions from one state to another states with probabilities. The dotted line arrows are the transaction that may not exist at all times. (a) SIR model: β is infection rate, γ is recovery rate, and ξ is the rate of Recovered to Susceptible. (b) SIHR model: α is stifling rate, β is refusing rate, γ is spreading rate, δ is forgetting rate, η is wakened remembering rate, and ξ is spontaneous remembering rate. (c) SEIZ model: β is infection rate, ϵ is self-adoption rate, ϕ is contact rate, and ξ is skeptic rate. The details of p and l and the whole model were explained in [87].

the SIR model for rumor spreading, so that the spreader (I) can be converted to stiflers (R). Stiflers are defined similar to Recovered state. They used the population size of R to measure the impact of rumor. They found that a forgetting mechanism can help reduce rumor influence and the rumor saturation threshold can be influenced by the average degree of nodes in the network. Another Hibernator state (i.e., users who refuse to spread rumor just because they forgot) was added to the SIHR (Susceptible, Infectious, Hibernator, and Recovered) model [228] to measure forgetting rate α and remembering mechanism η. The new remembering mechanism was proved to delay the rumor termination time and reduce rumor maximum influence. The direct link from S to R was added by [228] and were extended by [229]. The update was that all users in state S were finally converted to either I or R state if they had the chance to be exposed to spreaders (I). FIGURE 5a and FIGURE 5b describe the SIR and SIHR models, respectively.

Cho *et al.* [26] extended the basic SIR model by replacing the transition between states to a decision based on the agent's belief on the extent of uncertainty in the agent's opinion. The Subjective Logic opinion model is used to model an agent's opinion composition and update based on the extent of uncertainty. The three states in the SIR are defined based on the degree of each dimension of an opinion which is defined by belief, disbelief, and uncertainty. The opinion update involved interaction similarity between two agents, a conflict measure between belief and disbelief, and opinion decay upon no interactions between agents for opinion updates. Based on the degree of uncertainty in a given opinion, an agent's opinion can move from any state to any other state. This work investigated the effect of misinformation and disinformation in terms of how well false information can be effectively mitigated by propagating countering (true) information by selecting a good set of true informers.

The evolutionary SIR model simulation has been used to model decision strategies in fake news attacks [96]. The state transitions in the SIR model was replaced by the decision model Iterated Prisoner's Dilemma (IPD). The deception strategies can modify the prior knowledge of the agents by either adding uncertainty or changing false perceptions. In their expensive simulation experiments, only a small population of fake news attackers can initiate the spread but the fitness of attackers was sensitive to the cost of deception.

b: SEIZ MODEL WITH VARIATIONS

Jin *et al.* [87] captured diffusion of false and true news by the SEIZ epidemic model. Instead of considering the Recovered state, they modeled a state of users being heard of the rumor but not spreading it (Skeptic, Z) and influenced users (E) posting the rumor with an exposure delay. The SEIZ model was accurately capturing the diffusion patterns in real news and rumors events and was evaluated to be better than the simple SIS (Susceptible, Infectious, and Susceptible) model. They also proposed a ratio R_{SI}, the transition rates entering E from S to the transition rates exiting E to I, to differentiate rumor and real news events data. Isea and Lonngren [83] extended the SEIZ model by considering a forgetting rate of rumor posts. The forgetting rate is defined as a probability a user forgets the rumors across all the states. FIGURE 5c shows the key components of the SEIZ model and its process with the states and rates given from one state to another state.

c: PROS AND CONS

Epidemic models provide a direct and straightforward mathematical model for the diffusion dynamics of the false information. The agent density plot with time is a good way of observing the differences between the simulation and real values. However, simulation tests face a common issue as the population size is unknown and stable, and initial variable values are unknown. If the population size is as large as the real social media network, the computational cost cannot be ignored. In addition, in the SIR model, the state change is controlled by probability; but this autonomous behavior ignores a user's intention and belief. To complement this, there have been some efforts [26], [96] focusing on modeling and evaluating the effect of subjective, uncertain opinion and trust of agents and the role of more agents in terms of false information diffusion.

2) CREDIBILITY-BASED MODELS

In OSNs, one of the detection mechanisms for false information attackers, Sybil accounts, or spammers is modeling the credibility score in the network [88], [89], [225]. Existing works used various ways to represent credibility scores, such as reputation scores, trust scores, and belief scores. Credibility in OSNs can be modeled by two methods: *classification-based* and *credibility propagation*. A classification-based approach uses supervised learning algorithms [130]. On the other hand, *the credibility propagation approach* constructs a network to propagate credibility scores among users, tweet contents, events and activities [88]. Based on the credibility scores, ranking algorithms of users and posts can be conducted, such as PageRank [5], [24], [60], [225].

Negm *et al.* [130] used 5Ws (i.e., who, what, when, where, and why) credibility to distinguish credible news and RSS (Rich Site Summary) files from news agencies to extract publication dates, headlines, contents, and locations to feed into different algorithms to calculate the credibility of a news agency. The compared algorithms include TF-IDF, TF-IDF with location, Latent Semantic Index (LSI), and TF with LSI and log entropy. They concluded that TF-IDF and TF-IDF with location performed the best in calculating credibility. More recently, Norambuena *et al.* [136] leveraged the 5W1H extraction and news summarization techniques to propose the Inverted Pyramid Score (IPS) to distinguish structural differences between breaking and non-breaking news, with the long-term goal of contrasting reporting styles of mainstream and non-mainstream fake outlets.

Jin *et al.* [88] have introduced a credibility propagation network for news content composed of three layers: message, sub-event, and event. The event layer talks about the main event the news covers, the sub-event layer relates events to the main event, and the message layer holds the content of the news article. A graph optimization problem is formulated to calculate the credibility in this hierarchical network. All the layers are content-based, and have direct relations with the credibility of the news. Jin *et al.* [89] further proposed a verification method on credibility in a propagation model by using a topic modeling technique. Mitra and Gilbert [125] constructed the CREDBANK corpus by tracking tweets, topics, events, and associated in-situ human credibility judgements to systematically study credibility of social media events tracked over real-time. They later leveraged this corpus to construct language and temporal models for credibility assessment [126], [127]. By identifying theoretically grounded linguistic dimensions, the authors presented a parsimonious model that maps language cues to perceived levels of credibility. For example, hedge words and positive emotion words were associated with lower credibility. Additionally, by examining the temporal dynamics of the event reportages, they found that the amount of continued collective attention given to an event contained useful information about its associated levels of credibility [126].

Akoglu *et al.* [4] proposed the so-called *OddBall* algorithm to detect anomaly behavior like malicious posts and fake donations. They studied a sub-graph (egonets) of a target node with its neighbors. They analyzed various scoring and ranking methods by using feature patterns in density, weights, principle eigenvalues, and ranks and compared their performance in different network topologies.

Kumar *et al.* [102] detected fake reviewers in user-to-item rating networks. They developed a new trust system to rank users, products and ratings by fairness, goodness, and reliability, respectively. The intrinsic scores are calculated by combining network and behavior properties. Users rated with low reliability are more likely to be fake reviewers [102]. Akoglu *et al.* [5] developed the so called FraudEagle algorithm to spot fraudsters as well as fake reviews in online review platforms. There are two steps in the FraudEagle algorithm in terms of scoring users and reviews and grouping the analyzed results. For each review, the sentiment from true and false is only analyzed to assign the belief score. The grouping step reviews top-ranked users in a subgraph by clustering and merging more evidence to reveal fraudsters.

Ghosh *et al.* [60] developed the *CollusionRank* algorithm for detecting link farming type spammer attacks. The influence scores were given to the users and web pages. By decreasing the influence scores of the users connected to spammers, the follow-back behavior of social capitalists was discouraged. Yu *et al.* [225] developed the *SybilLimit* ranking algorithm for detecting Sybil attacks. A Sybil node was identified by calculating the node's trust score. Chirita *et al.* [24] developed the *MailRank* algorithm for detecting Sybil attacks in the email network. A sender is assessed by a global and personalized reputation score.

Pros and Cons: Credibility models can be applied in different stages and levels based on contents, user behaviors, and posts/comments in highly heterogeneous networks. In addition, a credibility model based on network features is agnostic to platforms and languages because the model only needs network features. However, how to accurately evaluate initial credibility values is not a trivial problem. Considering credibility at multiple levels makes the computation more complex and expensive so it may not be preferred. Further, credibility may be subjective and cannot be ported across platforms and/or networks. Lastly, a credibility model may not be able to detect sudden changes caused by instances which are not easily observable, thus impacting the accuracy of the credibility score assessment.

3) CASCADES FEATURES-BASED MODELS

Information network propagation patterns can be represented by a cascading structure depicting the flow of OSD information flow that users time-travelled through, posted, tweeted, and retweeted. The cascading structure has two forms: *hop-based cascades* and *time-based cascades* [231]. The cascades features can be grouped into two approaches: (i) Calculating the similarity of cascades between true and false information; and (ii) representing cascades using

informative representation and features in a supervised learning model.

a: CASCADES SIMILARITY

Cascades similarity is computed between fake news and true news. A graph kernel [231] was used as a common strategy for computing the cascades similarity. Wu *et al.* [217] proposed a fake news detection method using a hybrid kernel function. This graph kernel function calculates the similarity between different propagation trees. It also discussed about Radial Basis Function (RBF) kernel which calculates the distance between two vectors of traditional and semantic features. The sentiment and doubt scores for user posts need to be verified for fakes news. Ma *et al.* [116] proposed a top-down tree structure using RNNs for false information detection. The RNN learns the representation from tweets content, such as embedding various indicative signals hidden in the structure to improve rumors identification.

b: CASCADES REPRESENTATION

Cascades representation pursues informative representation as features to distinguish fake news from true news. For example, the number of nodes is a feature in a non-automated way. Alternatively cascades representation can fit deep learning models [219]. Wu and Liu [219] used LSTM-RNN to model propagation cascades of a message. This work combines the propagation pathways with user embedding, which forms a heterogeneous network. A message is represented by a sequence of its spreaders. A modularity maximization algorithm is used to cluster nodes with embedding vectors. Ma *et al.* [117] proposed propagation trees using Propagation Tree Kernel (PTK) for rumor detection. It can explore the suggested feature space when calculating the similarity between two objects.

c: PROS AND CONS

Similarity-based approaches consider the roles of users in propagating false information. Computing similarity between two cascades may require high computational complexity [231]. Representation-based methods automatically represent news to be verified; however, the depth of cascades may challenge such methods as it is equal to the depth of the neural network. All the approaches only provided experimental data to show their effectiveness. However, it may not properly reflect real world settings. Training data is a time-consuming process and is often computationally expensive.

4) GAME THEORETIC MODELS

This explores the deception and defense by reward and penalty model in OSD attacks. In game theory, the actions and decisions of the players are mainly based on the reward and penalty of their previous activities and the other players' actions [180].

Kopp *et al.* [96] discussed a game theoretic false information propagation model as a deception model that simulates the propagation of fake news in the OSNs. They used three types of game theories: Greenberg's deception model [65], Li and Cruz's deception model [112], and hypergame theory [11]. The Greenberg's deception model investigated the effect of deception on players' payoffs [65]. Kopp *et al.* [96] mapped false information to Greenberg's false signal model. Li and Cruz [112] used passive and active deception strategies by introducing noise and randomization, respectively, to increase uncertainty. Kopp et al. [96] used the deception game in [112] for consistently monitoring constraints and conditions, which affects game strategies. Bennett and Dando [11] used hypergame theory to model a deception game where players had subjective perception and understandings of a complicated game. Kopp *et al.* [96] also used [11] to consider players' subjective belief which may introduce uncertainty as well. Kopp *et al.* [96] proposed the information theoretic model that attackers' deceptive behavior can be significantly mitigated when the cost of deception is fairly expensive.

Pros and Cons: Game theoretic approaches to model OSD attacks add extra features over and other conventional network structure-based approaches above by considering the cost and benefit of performing a deceptive behavior by users in OSNs. Game theoretic deception detection is a promising approach that reflects human behaviors aiming to take an optimal action based on the expected outcome. However, game theoretic approaches have been rarely adopted in modeling and analyzing online social deceptive behaviors, compared to data-driven deception detection approaches. Due to this reason, the effectiveness of game theoretic deception detection approaches has not been fully investigated in the literature. In addition, aligned with a conventional drawback in using game theory, a large number of deceptive actions may introduce a high solution complexity. Uncertain, subjective beliefs of users should be carefully considered in terms of modeling incomplete information and/or imperfect information in game theory.

5) BLOCKCHAIN-BASED MODELS

Huckle and White [80] developed a tool called *Proventor* to prove the origin of the media. The Proventor is based on Blockchain storing provenance metadata for users to trust the authenticity of the metadata. Provenator can be used to validate news for news outlets like CNN and BBC where information and news is sometimes gathered from independent sources. However, since Provenator uses Blockchain and cryptography, a small difference, such as one pixel difference between two images, can make the result vastly different, leading to generating numerous false alarms and human interventions for validation, which is labor-intensive. McEvily *et al.* [121] proposed a social media platform called *Steem* (i.e., a database) based on Blockchain technology for building a community reward system. The reward system relies on users for consensus voting, reading content, and commenting.

Pros and Cons: The original design of Blockchain has security benefits in terms of provenance, integrity and

immutability. The Blockchain system is a heterogeneous network that incorporates other stakeholders to detect and control OSD activities. In addition, it is resilient against OSD attacks. Managing the large ledger size in Blockchain is an issue as shared information in social media and news outlets grows exponentially. Since both flagging accuracy and consensus verification rely on the contribution of crowd signals, it may break when too many users are malicious. For example, if a large volume of attackers contribute to the crowd activities and even control the system, a user cannot access to write transactions. In addition, the authorized party may be compromised by advanced attackers.

6) OTHER NETWORK OPTIMIZATION MODELS

Several graph optimization algorithms were proposed in graph anomaly detection and community detection problems. Hu et al. [78] developed a matrix factorization-based algorithm to detect social spammers on Twitter. Their framework utilized both content information and network information of an adjacency matrix and solved a non-smooth convex optimization problem. Several approaches have been taken to detect link farming attacks via network structure-based algorithms. Araujo et al. [8] detected temporal communities in cell networks and computer-traffic networks based on Tensor analysis. Jiang *et al.* [85] detected behavior patterns in OSNs where the spectral subspaces had different patterns and different lockstep behaviors. In addition, Jiang *et al.* [84] identified synchronized behaviors from spammers. Kumar *et al.* [99] considered trolling as a social deception activity. They proposed a *decluttering algorithm* to break a network into smaller networks on which the detection algorithm could be run. Kumar *et al.* [101] considered *sockpuppets* as an OSD attack where users created multiple identities to manipulate a discussion. They found that sockpuppets could be distinguished from normal users by having more clustered egonets.

Pros and Cons: Graph-based features are more available compared to the user profiles and/or user interaction features without violating privacy issues. In addition, graph-based algorithms can be agnostic to any datasets with high applicabilities in diverse platforms. However, collecting graph-based features, such as centrality measures, and solving graph optimization often incurs high computational overhead. This hinders its applicability to platforms that require real-time or lightweight detection for streaming data.

D. HYBRID DETECTION

Since ML/DL-based models can take an abundant amount of features, one can train a hybrid feature set combining the user profile, message content, and network features to detect OSD attacks. Unlike several existing survey papers which discussed only individual feature categories [98], [218], our discussion will focus on dealing with OSD attacks using hybrid features [82], [107]–[109], [199], [206].

Lee *et al.* [109] detected crowdturfers from Twitter users. A total of 92 features were divided into 4 groups: User demographics, user friendship networks, user activity (behavior-based features), and user content similarity including linguistic feature from LIWC dictionary. Vosoughi *et al.* [199] developed a tool called *Rumor Gauge* for automatically verifying rumors and predicting their veracity before they are verified by trusted channels. Since rumors are temporal, time-series features are extracted as the rumor spreads. A total of 17 features (e.g., linguistics, user involved, and propagation dynamics) were studied. They found that the fraction of low-to-high diffusion in the diffusion graph is the most predictive feature to represent the veracity of rumors. The time-series features are processed in DTW and HMM models but DTW assumes all the time-series are independent and assigns equal weight to all 17 features. The experiment evaluated the performance of the Rumor Gauge in terms of the accuracy of veracity prediction, contribution of each individual feature, and contribution of three groups of features and accuracy as a function of latency.

Pros and Cons: Hybrid detection takes advantages of hybrid feature sets and can improve the accuracy in detecting rumors, spammers, and crowdturfings. A drawback of the hybrid detection approach is expensive feature engineering and acquisition. Furthermore, the training process is time-consuming with the increase of complexity as the feature size increases.

VIII. RESPONSE MECHANISMS TO ONLINE SOCIAL DECEPTION

In this section, we survey existing mitigation or recovery mechanisms after OSD attacks are detected along with early detection mechanisms of OSD attacks [38], [56], [216]. Florêncio and Herley [56] developed a mitigation strategy to deal with compromised accounts by detecting password reuse events and timely reporting it to financial institutions. The aftermath actions were to take down identified phishing sites, restore the compromised accounts, and rescue users from bad decisions.

Dinakar et al. [38] took a mitigation action to counter cyberbullying with two steps: (i) early detection; and (ii) reflective user interfaces that popped up notices and suggestions on user behaviors. Most efforts made to mitigate OSD attacks in OSNs mainly focused on reducing the effect of false information propagation. Wu et al. [216] summarized two misinformation intervention methods: (i) detecting and preventing misinformation from spreading in an early stage; and (ii) developing a competing campaign to fight against misinformation. To limit the spread of fake news, a sample of fake news with maximal utility was identified in [185]. Within a certain constraint, this sample of fake news kept the largest number of users away from fake news posts. Their algorithm was robust against a high amount of spammers. Huckle and White [80] also made an effort to mitigate fake news spread based on the validity proof of digital media data, such as a picture in the fake news. The blockchain technology was used to prove the origins of digital media data; however, this method cannot prove the authenticity of the whole news article. Kumar and Shah [98] summarized misinformation

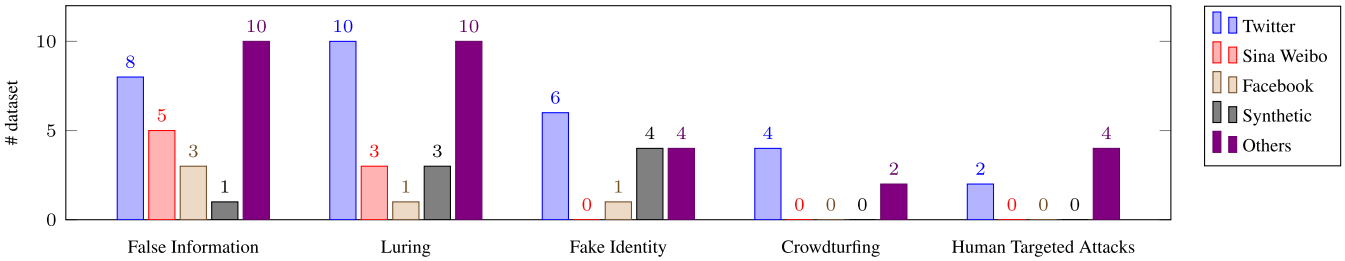


FIGURE 6. The types of datasets and the frequency of their use under the five online social deception studies. The datasets are collected from all the approaches for the prevention, detection, and mitigation of OSD attacks in TABLES 5-7.

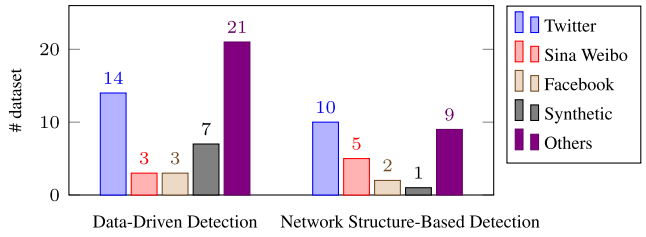


FIGURE 7. The types of datasets and the frequency of their use based on two types of approaches, data-driven OSD detection techniques shown in TABLE 6 and network structure-based OSD detection techniques shown in TABLE 7.

mitigation by modeling true and false information. From the existing four different approaches, the authors concluded that these algorithms are effective in detecting the spread of rumor and their simulations could suggest rumor mitigation strategies. Okada *et al.* [140] studied rumor diffusion by an SIR-extended information diffusion model and developed a mitigation mechanism to ask high influential users to spread correction diffusion. The authors examined how false rumor diffuses and converges when help and/or correct information is given and how fast the convergence appears.

Pros and Cons: Mitigation and recovery mechanisms relied heavily on early detection. The simulation model of spreading true information can mitigate the negative influence. However, most studies are based on simulation models, limited in using real world datasets, or has not been validated based on the implementation in real-world platforms. Although it is highly challenging for the developed model to be deployed in real platforms, there should be more efforts of using empirical, real datasets for the validation of the developed recovery models. Recovery in OSNs is more difficult than offline social networks because the relationships can be easily dropped. Only one research [56] designed a system for account restoration. More research efforts should be made to effectively mitigate the aftermath actions upon early detection.

TABLE 8 summarizes the classification of OSD defense mechanisms including prevention, detection, and mitigation/response discussed in Sections VI–VIII. Existing works mostly focused on detection of OSD attacks we classified in Section III. Less attention has been paid to prevention and mitigation where the main focuses include false information, luring, and identity theft. There are still open questions to build trustworthy cyberspace against human targeted attacks, especially for protecting children.

TABLE 8. Classification used for the defense mechanisms to deal with online social deception attacks in this survey.

Technique	OSD Attacks Considered
Attack Prevention	
Data-Driven	Fake news, phishing, fake profile, cyberbullying
Social honeypots	Spamming, fake profile, socialbot
Attack Detection	
User profile	Rumor, fake review, spam, fake profile
Message content	Fake news, rumor, fake review, phishing, spam, fake account, compromised account, crowdturfing, cyberbullying
Network structure	Fake news, rumor, fake review, false information, spam sybil attacks, crowdturfing
Attack Response	
Early detection	Compromised accounts, cyberbullying, false information propagation
Information propagation mitigation	False information (or fake news), spamming
Blockchain-based authenticity	Fake news, rumor

IX. VALIDATION & VERIFICATION

A. DATASETS

We summarized all the datasets used in existing OSD prevention and detection approaches in TABLES 5–7. Most datasets are from various social media platforms, including Twitter, Sina Weibo, Facebook, YouTube, and Reddit. FIGURE 6 demonstrates the frequency distribution of each data source for the five types of OSD attacks considered in this work. Twitter, Weibo and Facebook platforms are used with synthetic datasets and datasets from all other sources. Twitter is the most frequently used data source probably because of the user friendly API for public users to download tweets in a certain time period. Datasets for false information attacks (e.g., rumors, fake news and fake reviews) and luring attacks (e.g., spamming and phishing) draw the most attention from researchers. It demonstrates the diversity of the sources of datasets used in the literature.

FIGURE 7 illustrates the dataset platforms distribution for two types of OSD attack detection approaches, namely, data-driven detection and network structure-based

detection. FIGURE 7 shows the datasets distribution in data-driven approaches (see the left part of the figure) summarized in TABLE 6. Twitter datasets are broadly used in all types of OSD attack detection mechanisms, such as spambot, malicious account, fake account, compromised account, rumors, and crowdturfing. Other data sources include LinkedIn, YouTube, online forums Reddit, blacklisting websites, fact-checking websites, crowdturfing worker sites, and PhishTank websites, depending on the type of OSD attacks. Several benchmark datasets are frequently used, such as a social honeypot dataset [108] in which the authors collected a lot of spammer accounts by using social honeypots deployed in Twitter networks for seven months.

FIGURE 7 also shows the dataset distribution used in network structure-based detection (see the right side of the figure) in TABLE 7. Twitter, Weibo, and Facebook are the top three individual data sources. The others include fact-checking websites, app store database, online forums, and rating platforms. The datasets for network structure-based approaches can be divided into simulation research and detection research. Synthetic datasets are more frequently used in simulation models, such as epidemic models and/or credibility/ranking-based models.

Based on our survey of the datasets used in the OSD research, as shown in FIGURE 7, most existing approaches rely on the analysis of static datasets. Although it is not easy to deploy a defense mechanism in a dynamic, real platform, agent-based models where the agent's behavior is modeled based on real datasets can provide better insights on how the defense mechanisms work under dynamic environments.

B. METRICS

Most data-driven approaches have used metrics to estimate the detection accuracy of OSD attacks. The following metrics have been considered in the literature:

- *Confusion Matrix* [10], [17], [21], [28], [40], [49], [50], [61], [78], [88], [89], [91], [95], [102], [107], [108], [113], [115], [117], [135], [162], [166], [174], [175], [177], [178], [186], [199], [206], [222]: The confusion matrix is made of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN). They are the basic components for other accuracy metrics, such as precision and recall.
- *Precision* [10], [17], [21], [28], [40], [50], [61], [78], [82], [88], [89], [91], [102], [107], [113], [115], [135], [162], [166], [175], [186], [193], [217], [219], [224]: This metric simply estimates the true positives over positives detected including true positives and false positives by:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{1}$$

- *Recall* [17], [21], [28], [40], [50], [61], [78], [82], [88], [89], [91], [107], [113], [115], [135], [162], [175], [186], [217], [219], [224]: This metric captures the true positives over the actual positives include true positives and false negatives. This metric is estimated by:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2}$$

- $\mathbf{F}_1$ *Score or Measure* [17], [21], [28], [38], [40], [50], [61], [78], [82], [88], [89], [107]–[109], [113], [162], [186], [217], [219]: This metric is an indicator of the accuracy of detection based on both precision and recall. It is measured by:

$$\text{F}_1 = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{3}$$

- *Accuracy* [9], [10], [13], [28], [38], [40], [49], [60], [82], [84], [86], [88], [89], [107]–[109], [116], [117], [135], [166], [175], [186], [199], [217], [219], [223]: This metric measures correct detection for true positives and true negatives. However, when the datasets are not balanced such as too large true positives with too small true negatives or vice-versa, this metric may mislead. It is given by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

  There is also a weighted accuracy score [13] with different weights on labels. Accuracy can also be used to evaluate the contribution of each features or feature sets [82], [166], [199], [217].
- *False Positive Rate (FPR)* [9], [21], [49], [109], [162], [174], [178], [186], [206], [223]: This metric is to measure misdetection in terms of false alarms among the ones detected as positives and computed by:

$$\text{FPR} = \frac{FP}{FP + TN} \tag{5}$$

- *False Negative Rate (FNR)* [9], [10], [109], [206], [223]: This metric captures how many positives are missed and is estimated by:

$$\text{FNR} = \frac{FN}{TP + FN} \tag{6}$$

- *Specificity* [10], [21], [28], [162], [174]: This metric measures the extent of correctly detecting negatives over the actual number of negatives and is obtained by:

$$\text{Specificity} = \frac{TN}{TN + FP} = 1 - FPR \tag{7}$$

- *Weighted Cost* (W_{cost}) [223]: In phishing detection, since the ratio of legitimate websites to phishing website is high, a legitimate website misclassified to a phishing one (FPR) has severe effects than the reverse (FNR). The weighted cost is used to balance the performance of FPR and FNR. W_{cost} is estimated by:

$$\text{W}_{cost} = \text{FNR} + \lambda \times \text{FPR}, \quad \lambda > 1. \tag{8}$$

  where λ is the weight of FPR. Higher values of λ means larger influence of FPR value.
- *Receiver Operating Characteristic (ROC) Curve* [10], [82], [106], [174], [175], [206]: ROC curve draws a

plot of classifier's true positive rate (TPR) against FPR at various detection threshold scenarios. This curve is used to measure and compare stability between several classifier models.

- *Area Under the Curve (AUC)* [10], [17], [22], [61], [82], [102], [106], [108], [109], [162], [174]: AUC is calculated by the the area under the ROC curve. It measures the probability of a classifier to correctly identify a true-positive data. Since AUC is insensitive to imbalance between classes, it can be better than *Accuracy* in evaluating imbalanced dataset. AUC is another metric of classifier stability and classification quality for different settings.
- *Discounted Cumulative Gain (DCG)* [147]: DCG measures the effectiveness of an algorithm, an alternative measure to AUC. A higher DCG is indictive of an early identification of suspicious cases and estimated by:

$$\text{DCG} = r[1] + \sum_{i=2}^{n} \frac{r[i]}{\log_2 i}, \tag{9}$$

where $r[i]$ is 1 if the i^{th} friend request was defined as suspicious or 0 if the i^{th} friend request was defined as legitimate, and n is the number of total incoming requests that require further investigation [147].

- *Matthews Correlation Coefficient (MCC)* [28], [61], [162], [186]: MCC measures the correlation between predicted class and real class of users. This metric is considered as the unbiased version of F_1-*measure* and given by:

$$\text{MCC} = \frac{TP \times (TN - FP) \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}, \tag{10}$$

where MCC $\approx$ 1 means high prediction accuracy. MCC ≈ 0 means the prediction is no better than random guessing. $MCC \approx -1$ means that the prediction is in disagreement with the real class.

- *Cohen's Kappa Value (κ)* [38]: This metric is a measure of reliability for two classifiers or raters, which considers true positive agreement by chance. Cohen's Kappa Value is used when *Accuracy* alone is insufficient to evaluate model reliability [38]. Cohen's Kappa is calculated as:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{11}$$

where P_o is the observed agreement in classification, the same as *Accuracy*, and P_e is the hypothetical probability of agreement by chance. High Cohen's Kappa Value ($0.8 \leq \kappa \leq 1$) indicates good reliability [18].

- *Mean Absolute Error (MAE)* [87], [216]: Many detection algorithms for OSD attacks use MAE to estimate their detection accuracy. In addition, this metric is used to measure the simulation fitting error of an epidemic model by calculating the absolute values of errors at each time points.

$$\text{MAE} = \frac{1}{|U|} \sum_{i \in U} |p_i - l_i|, \tag{12}$$

where U is a user set, p_i is a prediction result, l_i is a true label, and i is a data index.

- *2-norm Error* [87]: This measures the simulation fitting error of an epidemic model as one of the performance measures of model fitting and optimization. A good model would reduce this error through iterations. This metric is estimated by:

$$\text{2-norm Error} = \frac{\| \text{I}(t) - \text{Tweets}(t) \|^2}{\| \text{Tweets}(t) \|^2}, \tag{13}$$

where $\text{I}(t)$ is the number of users (agent I) that spread the rumor tweet at time t. $\text{Tweets}(t)$ is the number of tweets at time t from the real data.

- *Mean Fraction of Recovered Agents Per Time Unit (*R*)* [26]: This is a specific case of the statistics and plot metric. Instead of plotting the count of each agent at each time point, the average fraction of recovered agents during the total session time T is calculated.

$$\text{R} = \frac{\sum_{t=1}^{T} R(t)}{T}, \tag{14}$$

where $R(t)$ is the number of agents recovered from false information (i.e., not believing in false information) and T is the total simulation time.

- *Spearman's Rank Correlation Coefficient (ρ)* [49], [86]: This metric measures the rank correlation between the predicted labels and the ground truth and is obtained by:

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}, \tag{15}$$

where n ranks are distinct integers and d_i is the difference of two ranks between an element. ρ ranges in $[-1, 1]$ as a real number where 0 refers to random guess while 1 indicates positive correlation [212].

- *Label Ranking Average Precision (LRAP)* [86]: This measures the ability to give more accurate prediction for each post message, with a prefect prediction of 1. LRAP is measured by:

$$\text{LRAP} = \frac{1}{n} \sum_{i=0}^{n-1} \frac{1}{\| y_i \|_0} \sum_{j, y_{ij}=1} \frac{|L_{ij}|}{\text{rank}_{ij}}, \tag{16}$$

where n is number of data points, y_i is the vector of ground truth labels of the ith data point, $\| \cdot \|_0$ is number of non-zero elements in a vector, y_{ij} is the binary label of jth label from ground truth vector y_i, $|L_{ij}|$ is number of positive labels for a given data point i, and rank_{ij} is the rank of predicted label (p_{ij}) in predicted label vector (p_i) for a given i [165].

- *Label Ranking Loss (LRL)* [86]: This metric estimates the number of times that irrelevant labels are ranked higher than relevant labels. Due to its large volume of

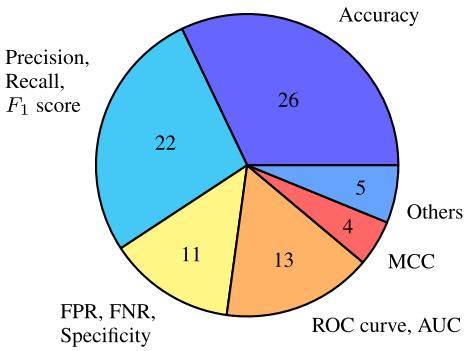


FIGURE 8. Counts of research works in TABLES 5-7 by metrics.

complex description, the interested readers can refer to [188] for more details.

FIGURE 8 illustrates the frequency of each metric used in the existing approaches surveyed in this work. Since most of the current studies are to develop OSD attack detection mechanisms, the majority of the metrics is related to measuring detection accuracy. Among all the detection metrics, Precision, Recall, F_1 score, Accuracy are the most popular metrics used in the existing works. FPR, FNR, Specificity, ROC, and AUC are also obtained based on the Confusion Matrix. They are used to compare the performance of multiple classifiers. However, algorithmic complexity of defense algorithms is rarely considered.

X. ETHICAL ISSUES OF SOCIAL DECEPTION

Ethical issues in social deception research have been discussed as follows:

- Privacy issues may be raised when conducting social deception research in terms of setting up social honeypots and fake profiles, collecting data from those accounts, and capturing users behaviors (e.g., making friends and posting texts). Elovici *et al.* [46] strongly recommended sharing datasets in the public. This allows other researchers to avoid taking unnecessary procedures associated with any ethical issues which are often encountered in the process of data collection. If many public datasets for research are available, new researchers can reduce the need to crawl their own dataset. In addition, if the OSN provider has an advanced way of anonymization, the researcher can follow those standards to protect the identity when handling the collected data. The authors also discussed a coordinated emergency response team (CERT) to handle vulnerability disclosures from the new research results [46] in terms of strictly anonymizing users' identities and handling findings with great care.
- Since social honeypots research involves human subjects-based experiments, it should be regulated by the institutional review board (IRB) approval [42] particularly in terms of privacy issues that may be raised in personal data analysis, stakeholder analysis, and human deception analysis. However, Many ethical issues still remain even not discussed [67], [107], [233].
- Several online social deception studies have discussed legal and ethical issues [31], [147], [222]. However, their discussions are limited in that if no malicious activities do not directly involve normal, legitimate users, their design is safe to normal users. There may be indirect influences of social honeypots that can introduce to normal users, such as normal users approaching to social honeypots.
- Although one community seriously concerns ethical issues related to privacy in conducting social deception research, the other community takes a position of advocating online social deception research in terms of safeguarding society and vulnerable people. Hence, their perspective is that there are neither unethical nor illegal issues associated with conducting online social deception research [120].
- Some researchers claim that creating fake accounts as social honeypots is only for detecting spammers, not to take benefits from normal users or buy compromised accounts [191], [232]. However, it seems not clear whether social honeypots using fake accounts do not introduce any harms to normal users.
- To prevent risks from using crowdsourcing methods, some guidelines of controls and protections toward unethical behaviors are discussed, such as privacy violation [142]. The system design and research procedures should include how to prevent sensitive data sharing and to enforce users' security education and training.
- For misinformation propagation experiments, some researchers claim that since misinformation itself (e.g., fake news) is from public information, it does not require any informed consent [32]. However, spreading the public misinformation itself can even amplify its influence in OSNs, which can still manipulate public opinions.

The ethical issues associated with conducting online social deception research have been hotly debated because this issue touches conflicting aspects of the fundamental values, which is privacy vs. safety. In the current state of the OSD research, there have been a lot of obscure aspects in conducting human subject involved research in online platforms. Since human users are the key part of OSNs and the key entities to be protected in OSNs, there should be very specific guidelines and regulations which can facilitate researchers to safely solve OSD problems within the legal boundary. Otherwise, although solving the OSD problems is highly critical to ensure the public good and safety in our society, extra hassle derived from ethical issues may significantly hinder researchers from tackling the OSD research.

XI. DISCUSSIONS: INSIGHTS & LIMITATIONS

Based on the extensive survey conducted, we identify the following insights:

- **Deception domains and intent**: Deception is defined across multidisciplinary domains with varying intent and detectability in type and extent. Although social

deception is frequently considered as a negative connotation with low integrity and maliciousness, not necessarily all socially deceptive behaviors have bad intent. Rather, social deception can play a defensive social role for self-protection or self-presentation.

- **OSD type category**: Like OSN attacks and cybercrimes, OSD can be defined by deceptive intent. However, unlike OSN attacks or cybercrimes, a unique aspect of the OSD is that OSD is only possible when a deceivee cooperates with a deceiver. Hence, training and education of deceivees is highly critical for preventing OSD attacks.
- **Importance of social deception cues**: Traditional offline deception cues and vulnerabilities are from several domains: individual, cultural, linguistic, physiological and psychological. The cues and vulnerabilities of OSD have variations compared to face-to-face communication. For serious OSD attacks which mainly belong to cybercrimes, such as human targeted attacks (e.g., human trafficking, cyberbullying, cyberstalking, or cybergrooming), if OSD cues are effectively captured, there is a much higher chance to prevent and detect OSD attacks than offline social deception due to much less real-time interactions which trigger much less risky situations from the safety perspective.
- **Ethical design considerations of social honeypots**: A social honeypot is one of broadly studied OSD prevention/detection mechanism. They are deployed to passively collect attackers account profiles. However, since social honeypots deal with human users, there should be careful legal or ethical considerations in their design features. To this aim, there should be more specific, clear guidelines and regulations available for the researchers.
- **OSD detection mechanisms**: Three dominant OSD detection approaches surveyed in this work are user-profile-based, message content-based, and network structure-based. They each have pros and cons in different scenarios. In particular, if a detection mechanism uses only network structure features to detect OSD attacks, it would better preserve user privacy but need to develop lightweight algorithms to efficiently calculate expensive network features, such as centrality values requiring knowledge of the entire network topology and high computation cost to estimate centrality values. To maximize the synergy of all three approaches, hybrid approaches incorporating all are promising.
- **Metrics for performance evaluation**: As the majority of OSD defense mechanisms are explored to effectively detect OSD attacks, most works have used accuracy metrics to measure the performance of their proposed work. A few of the metrics are based on correlations and ranks, which are mainly used to identify key signals to detect OSD attacks.

We also found the following **limitations** of the existing OSD detection approaches:

- **Lack of systematic, comprehensive defense strategies to combat OSD attacks**: Fighting against OSD attacks requires systematic, comprehensive, and active defense strategies covering prevention, detection, and mitigation/response. However, existing approaches have been heavily explored in detection strategies, rather than prevention or mitigation strategies. In addition, some approaches are embracing multiple roles with a single mechanism. For example, most current OSD mitigation approaches are based on the results from early detection. Further, since a social honeypot collects attacker profiles, the analysis of social honeypots is used to design classifiers for both prevention and detection.
- **Lack of experiments with real-time, dynamic datasets**: Current prevention and detection methods are based on simulation and/or real datasets, but only a few studies discussed effective training and detection using streaming data, such as Twitter API. In addition, the high computational and time complexity for real-time detection remains an open issue.
- **Insufficient proactive defense**: The inherent role of a social honeypot is proactively finding targeted attackers (i.e., a particular type of attackers). This way allows a system to identify targeted OSD attackers and proactively take actions to prevent vulnerable users from being victimized by the targeted OSD attackers. Although honeypots are used in communication networks as a proactive intrusion prevention mechanism, social honeypots are passively used in OSNs due to potential legal and ethical issues. Without clarifying the legal/ethical design guidelines and regulations, the function and exploitation of social honeypots cannot be fully benefited and even can be improved further to deal with highly intelligent attackers. In particular, to deal with real human-based OSD attacks, such as crowdturfing by paid workers to conduct social deception activities, more active social honeypot designs should be allowed while preserving normal user privacy and ethical rights.
- **High complexity of features and models**: We substantially surveyed the features for data-driven detection methods in Sections VII-A and VII-B and network/epidemic models for network structure feature-based methods in Section VII-C. The complexity of extracting and evaluating features and the model optimization grows fast with the size of datasets. How to reduce the solution complexity and improve solution efficiency for OSD detection is still an open issue.
- **Lack of qualitative analysis for cues of OSD attacks**: Most OSD defense mechanisms have focused on dealing with attacks by machines (or bots). However, for more serious OSD attacks (i.e., human targeted attacks), appropriate cues should be first carefully identified through qualitative analysis based on multidisciplinary research efforts with behavioral scientists.

XII. CONCLUSION & FUTURE WORK

In this section, we discuss the **key findings** from this survey to answer the **research questions** raised in Section I-B as follows:

RQ1: *How is OSD affected by the fundamental concepts and characteristics of social deception which have been studied in multidisciplinary domains?*

Answer: The fundamental meanings and intent of social deception are commonly present in both offline and online social deception as we find surprisingly common trends and characteristics observed in socially deceptive behaviors. The common goal is 'misleading a potential deceivee for the benefit of a deceiver' by increasing the deceivee's misbelief or confusion. In both online and offline platforms, social deception is successful only when the deceivee cooperates with actions taken by the deceiver. Due to the unique characteristics of an online environment such as less real-time/face-to-face interactions without physical presence to each other, both the deceivee and deceivers can take advantages of them in terms of defense (i.e., prevention, detection, and response/mitigation) and attack (e.g., anonymous attacks or easily running away if something goes wrong).

RQ2: *What are new attack types based on the recent trends of OSD attacks observed in real online worlds and how are they related to common social network attacks, cybercrimes, and security breaches based on cybersecurity perspectives?*

Answer: More serious human targeted attacks (e.g., human trafficking, cyberstalking, cybergrooming, or cyberbullying) have emerged as new OSD attack types. The seriousness has grown as online deception often leads to offline crimes, which become indeed the major concern of cybercrimes. While human targeted attacks become a more serious social issue, there is a lack of cyber laws to respond to this serious social deception attack, easily leading to cybercrimes. Human targeted attacks also bring the discussion of security breach of a person and non-information assets. In this sense, human safety needs to be protected against the new types of OSD attacks.

RQ3: *How can the cues of social deception and/or susceptability traits to OSD affect the strategies by attackers and defenders in OSNs?*

Answer: Many cues and susceptability traits of offline social deception behaviors are present in online social deception behaviors. The examples include intentionality of social deception, its cues from linguistic, cultural, and/or technological contexts, and various susceptibility factors including demographics, cultural, and/or network structure feature-based traits. Moreover, due to the limited real-time and/or interactions feeling people's presence in online platforms, some cues such as physiological and/or psychological cues may be missed while they can be highly useful cues for detecting social deception. However, as more advanced features of online platform-based interactions emerge, more physiological/psychological cues can be captured to improve deception detection (e.g., heart beats can be fed back to a detection mechanism).

RQ4: *What kinds of defense mechanisms and/or methodologies need to be explored to develop better defense tools combating OSD attacks?*

Answer: Most defense mechanisms to combat OSD attacks only focused on detection, particularly in terms of data-driven approaches using machine/deep learning techniques. Prevention mechanisms are substantially limited and have often been considered along with detection mechanisms (e.g., social honeypots or data-driven approaches). Response mechanisms after the detection of the OSD are even much less explored than prevention mechanisms.

RQ5: *What are the key limitations of existing validation and verification methodologies in terms of datasets and metrics?*

Answer: Popular datasets used in existing OSD research are from Twitter, Sina Weibo, and Facebook along with other synthetic datasets collected from simulation, as shown in FIGURES 6 and 7. In particular, to study human targeted attacks, there is a lack of datasets available because online human targeted deception data are based on individual chats or dyadic interactions. In addition, most metrics are to measure detection accuracy of OSD attacks, which is natural to observe as most defense mechanisms mainly focus on detection. Hence, there is a lack of efficiency metrics that can capture cost or complexity of the proposed defense techniques against OSD attacks.

RQ6: *What are the key concerns associated with ethical issues in conducting OSD research?*

Answer: The OSD research is inherently involved with human users and may introduce ethical issues. However, to conduct meaningful experiments, some real testbed-based validation/verification should be conducted to obtain high confidence in the developed technologies under realistic settings. However, when deploying defense techniques in a real testbed (e.g., Facebook, Twitter, etc), the defense process may encounter inevitable deception towards normal, legitimate users. In addition, privacy is a big concern in cybersecurity and there is an inherent trade-off between preserving users privacy and improving the quality of defense tools against OSD attacks (i.e., privacy vs. safety). To investigate serious OSD attacks, such as human targeted attacks, most interactions are peer-to-peer, such as dyadic conversations/chats, which is mostly unavailable. As a result, there is a lack of real datasets in studying highly serious human targeted attacks, such as human trafficking, cyberstalking, or cybergrooming attacks. In addition, there is a lack of systematic legal and/or ethical guidelines and regulations on how to proceed the OSD research with involvement of human users in real testbed settings.

We suggest the following **future research directions** in the online social deception and its countermeasure research:

- **Multidimensional research approaches to solve online social deception**: Although various concepts,

properties, and cues of social deception have been studied in diverse disciplines, the multidisciplinary nature of social deception has not been appropriately considered in developing defense mechanisms against OSD attacks. In particular, deceivers and deceivees are both humans via online platforms. Without understanding the way deceivers and deceivees communicate and/or interact to each other, it is hard to detect deception easily. Deception can be easily deployed on top of firm, trust relationships. In order to distinguish deception from truthfulness, in-depth understanding of deception based on multidisciplinary research effort is a must for developing effective defense mechanisms against OSD attacks.

- **Distinction of benign deception from malicious deception**: In the cybersecurity domain, deception refers to a deceptive action with malicious intent. However, in a social network, many users may use OSD to promote self-presentation/protection for privacy protection. Therefore, if OSD is treated as a form of attacks, it can possibly result in a high false positive rate (i.e., detecting benign users as malicious users). In order to prevent this, we need to develop deception-specific online defense tools that can differentiate benign deception from malicious deception.
- **Culture-aware defense against OSD attacks**: Based on our survey, different cultural deception cues have been observed [14], [75], [111], [167]. Since deception cues are sensitive to cultural characteristics, culture-aware defense mechanisms should be developed to effectively deal with OSD attacks that consider unique cultural characteristics of a social network.
- **Detectability-aware and intent-aware defense against OSD attacks**: As discussed in FIGURE 2, the spectrum of deception can span into a wide range of detectability and intent. Intelligent OSD attackers may establish trust relationships with potential victims and exploit the established trust to deceive the victims. This is especially observed in human targeted attacks, such as human trafficking or cybergrooming, which is categorized as serious cybercrimes [226]. Hence, we need to develop detectability-aware and intent-aware cues against highly subtle hard-to-detect OSD attacks.
- **Security protection of adolescent online users in multiple roles**: Adolescents have high vulnerability to OSD attacks, as discussed in Section V. Deceptions, such as cyberbullying, have exposed severe social, behavioral and security issues introduced by adolescents. Educational and habitual guidelines, parental control, and/or security guard tools cannot protect potential deceivees or victims. Social media platforms need to enhance their effective OSD prevention mechanisms especially for young users by identifying their vulnerability factors for more proactive protection.
- **Dynamic, updated defense mechanisms to obfuscate highly advanced attackers**: Recent studies showed that OSD attackers can build advanced social bots by analyzing the current detection models and fooling the existing models by leveraging adversarial machine learning (AML) techniques [103]. One countermeasure is to collect new datasets and retrain the classifiers. However, it is challenging to support updating the models with additional datasets. In addition, the cost of repeatedly training the classifiers with the whole dataset is particularly high. Hence, we need to develop lightweight ML algorithms. Another countermeasure can be identifying unknown deception features based on linguistic, behavioral, and technological cues.
- **Defense against human attackers vs. social bots**: A human attacker is another type of advanced attackers where a real human is behind the social network platforms performing OSD attacks. They can bypass detection because the conversation is from real humans or the accounts are mimicking normal users. There also exist crowdturfing workers who spread deceptive information in social media and get paid. More research work is needed to investigate how to detect and differentiate social bots from human attackers.
- **Measurement of physiological and/or psychological cues to develop better prevention techniques against OSD attacks**: Due to the unique characteristics of online platforms, some critical deception cues are missing and must be identified first, such as physiological and/or psychological cues. Measuring those cues can be critical in terms of improving prevention and early detection against OSD attacks.
- **Extra effort for developing prevention and response mechanisms to defend against OSD attacks**: In terms of the techniques used across all defense mechanisms, while machine/deep learning approaches are popularly used, game theoretic and/or network structure feature based approaches are still to be further explored to produce more mature approaches. They have extra merits over data-driven approaches in that the game theoretic approach can predict an attacker's next move. For prevention, although early detection as an OSD prevention strategy is receiving a high attention with growing amounts of recent works to fight against OSD attacks, there should be more prevention mechanisms that can provide more proactive defense, such as identifying potential attacks even before the attacks occur. Response/mitigation after OSD detection, such as mitigation after false information spread or recovery after OSD attacks are launched, is little explored in the literature and calls for more efforts to further investigate effective mechanisms to minimize risk and aftermath effect after OSD detection.
- **Effective deception cues-based approach to combat OSD attacks without violating user privacy**: Due to a lack of effective deception cues/datasets, it is difficult to conduct OSD research to defend against serious human targeted OSD attacks for validation and verification.

A future direction is to develop techniques to capture clear deception cues without violating user privacy.

- **Integrated defense needed to prevent, detect, and mitigate false information propagation**: As discussed in Section III-A, false information embraces fake news, unverified rumors, manipulated information, deceptive online comments or fake reviews. False (or unverified or forged) information is mostly propagated with undesirable intent to influence public opinions. Although there have been a rich volume of defense mechanisms developed to detect fake news, fake reviews, or fake comments, the adverse impact of propagated fake news has not been significantly mitigated. A more holistic approach is in a critical need by integrating the defense mechanisms for prevention, early detection, and fast mitigation of false information.
- **More efficiency metrics to expedite the defense process**: Efficiency metrics for measuring algorithmic complexity of defense techniques have not been sufficiently used in existing approaches. More meaningful complexity/efficiency metrics should be considered in order to expedite the speed of prevention, detection, and recovery as a defense against OSD.
- **Systematic legal and/or ethical guidelines for conducting meaningful OSD research**: Since humans are the key factors in solving the problems associated with the OSD attacks, the research community and government need to provide clear guidelines on conducting OSD research without violating user privacy. In communication networks, the research community appears to have reached some accord about using defensive deception techniques to defend against cyberattacks by emphasizing its benefits. However, for cybersecurity research on OSN platforms likely involving human subjects, there is little research, let along a consensus, on what methodologies are allowed and what level of user privacy must be preserved before achieving the goal of defense effectiveness.

REFERENCES

[1] H. Abutair, A. Belghith, and S. Alahmadi, "CBR-PDS: A case-based reasoning phishing detection system," *J. Ambient Intell. Hum. Comput.*, vol. 10, no. 7, pp. 2593–2606, Jul. 2019.

[2] D. Acemoglu, A. Ozdaglar, and A. ParandehGheibi, "Spread of (MIS) information in social networks," *Games Econ. Behav.*, vol. 70, no. 2, pp. 194–227, 2010.

[3] J. Adair, T. Dushenko, and R. Lindsay, "Ethical regulation and their impact on research practice," *Ethical Regulation Impact Res. Pract.*, vol. 40, no. 1, pp. 59–72, 1985.

[4] L. Akoglu, M. McGlohon, and C. Faloutsos, "Oddball: Spotting anomalies in weighted graphs," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2010, pp. 410–421.

[5] L. Akoglu, R. Chandy, and C. Faloutsos, "Opinion fraud detection in online reviews by network effects," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, 2013, pp. 2–11.

[6] S. Albladi and G. Weir, "User characteristics that influence judgment of social engineering attacks in social networks," *Hum.-Centric Comput. Inf. Sci.*, vol. 8, no. 1, p. 5, 2018.

[7] J. Anderson and J. Cho, "Software defined network based virtual machine placement in cloud systems," in *Proc. IEEE Mil. Commun. Conf. (MILCOM)*, Oct. 2017, pp. 876–881.

[8] M. Araujo, S. Papadimitriou, S. Günnemann, C. Faloutsos, P. Basu, A. Swami, E. E. Papalexakis, and D. Koutra, "COM2: Fast automatic discovery of temporal ('comet') communities," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Springer, 2014, pp. 271–283.

[9] P. R. Badri Satya, K. Lee, D. Lee, T. Tran, and J. J. Zhang, "Uncovering fake likers in online social networks," in *Proc. 25th ACM Int. Conf. Inf. Knowl. Manage.*, 2016, pp. 2365–2370.

[10] S. Barbon, R. A. Igawa, and B. B. Zarpelão, "Authorship verification applied to detection of compromised accounts on online social networks," *Multimedia Tools Appl.*, vol. 76, no. 3, pp. 3213–3233, 2017.

[11] P. G. Bennett and M. R. Dando, "Complex strategic analysis: A hypergame study of the fall of France," *J. Oper. Res. Soc.*, vol. 30, no. 1, pp. 23–32, Jan. 1979.

[12] I. R. Berson, M. J. Berson, and J. M. Ferron, "Emerging risks of violence in the digital age," *J. School Violence*, vol. 1, no. 2, pp. 51–71, Mar. 2002.

[13] G. Bhatt, A. Sharma, S. Sharma, A. Nagpal, B. Raman, and A. Mittal, "Combining neural, statistical and external features for fake news stance identification," in *Proc. Web Conf. Companion, Int. World Wide Web Conf. Steering Committee*, 2018, pp. 1353–1357.

[14] C. F. Bond, A. Omar, A. Mahmoud, and R. N. Bonser, "Lie detection across cultures," *J. Nonverbal Behav.*, vol. 14, no. 3, pp. 189–204, Sep. 1990.

[15] D. B. Buller, J. K. Burgoon, A. Buslig, and J. Roiger, "Testing interpersonal deception theory: The language of interpersonal deception," *Commun. Theory*, vol. 6, no. 3, pp. 268–289, Aug. 1996.

[16] D. Buller and J. Burgoon, "Interpersonal deception theory," *Commun. Theory*, vol. 6, no. 3, pp. 203–242, Aug. 1996.

[17] C. Cao and J. Caverlee, "Detecting spam urls in social media via behavioral analysis," in *Proc. Eur. Conf. Inf. Retr.* Springer, 2015, pp. 703–714.

[18] J. Carletta, "Assessing agreement on classification tasks: The kappa statistic," *Comput. Linguistics*, vol. 22, no. 2, pp. 249–254, 1996.

[19] T. L. Carson, *Lying and Deception: Theory and Practice*. London, U.K.: Oxford Univ. Press, 2010.

[20] Z. Chance and M. I. Norton, "The what and why of self-deception," *Current Opinion Psychol.*, vol. 6, pp. 104–107, Dec. 2015.

[21] C. Chen, J. Zhang, Y. Xie, Y. Xiang, W. Zhou, M. M. Hassan, A. AlElaiwi, and M. Alrubaian, "A performance evaluation of machine learning-based streaming spam tweets detection," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 65–76, Sep. 2015.

[22] H. Chen, J. Liu, Y. Lv, M. H. Li, M. Liu, and Q. Zheng, "Semi-supervised clue fusion for spammer detection in Sina Weibo," *Inf. Fusion*, vol. 44, pp. 22–32, Nov. 2018.

[23] T. Chen, W. Liu, Q. Fang, J. Guo, and D.-Z. Du, "Minimizing misinformation profit in social networks," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 6, pp. 1206–1218, Dec. 2019.

[24] P.-A. Chirita, J. Diederich, and W. Nejdl, "Mailrank: Using ranking for spam detection," in *Proc. 14th ACM Int. Conf. Inf. Knowl. Manage.*, 2005, pp. 373–380.

[25] J.-H. Cho, H. Cam, and A. Oltramari, "Effect of personality traits on trust and risk to phishing vulnerability: Modeling and analysis," in *Proc. IEEE Int. Multi-Disciplinary Conf. Cognit. Methods Situation Awareness Decis. Support (CogSIMA)*, Mar. 2016, pp. 7–13.

[26] J.-H. Cho, S. Rager, J. O'Donovan, S. Adali, and B. D. Horne, "Uncertainty-based false information propagation in social networks," *ACM Trans. Social Comput.*, vol. 2, no. 2, pp. 1–34, Oct. 2019.

[27] F. Cohen, "The use of deception techniques: Honeypots and decoys," *Handbook Inf. Secur.*, vol. 3, no. 1, pp. 646–655, 2006.

[28] S. Cresci, R. Di Pietro, M. Petrocchi, A. Spognardi, and M. Tesconi, "Social fingerprinting: Detection of spambot groups through DNA-inspired behavioral modeling," *IEEE Trans. Dependable Secure Comput.*, vol. 15, no. 4, pp. 561–576, Aug. 2018.

[29] D. C. Daniel and K. L. Herbig, *Strategic Military Deception: Pergamon Policy Studies on Security Affairs*. Amsterdam, The Netherlands: Elsevier, 2013.

[30] A. Darwish, A. E. Zarka, and F. Aloul, "Towards understanding phishing victims' profile," in *Proc. Int. Conf. Comput. Syst. Ind. Inform.*, 2012, pp. 1–5.

[31] E. De Cristofaro, A. Friedman, G. Jourjon, M. A. Kaafar, and M. Z. Shafiq, "Paying for likes: Understanding Facebook like fraud using honeypots," in *Proc. Conf. Internet Meas. Conf.*, 2014, pp. 129–136.

[32] M. Del Vicario, A. Bessi, F. Zollo, F. Petroni, A. Scala, G. Caldarelli, H. E. Stanley, and W. Quattrociocchi, "The spreading of misinformation online," *Proc. Nat. Acad. Sci. USA*, vol. 113, no. 3, pp. 554–559, 2016.

[33] K. J. Denker, J. Manning, K. B. Heuett, and M. E. Summers, "Twitter in the classroom: Modeling online communication attitudes and student motivations to connect," *Comput. Hum. Behav.*, vol. 79, pp. 1–8, Feb. 2018.
[34] Department of Homeland Security. (2018). *Countering False Information on Social Media in Disasters and Emergencies*. [Online]. Available: https://www.dhs.gov/sites/default/files/publications/SMWG_Countering-False-Info-Social-Media-Disasters-Emergencies_Mar2018-508.pdf
[35] B. M. DePaulo, J. J. Lindsay, B. E. Malone, L. Muhlenbruck, K. Charlton, and H. Cooper, "Cues to deception," *Psychol. Bull.*, vol. 129, no. 1, pp. 74–118, 2003.
[36] D. C. Derrick, T. O. Meservy, J. L. Jenkins, J. K. Burgoon, and J. F. Nunamaker, "Detecting deceptive chat-based communication using typing behavior and message cues," *ACM Trans. Manage. Inf. Syst.*, vol. 4, no. 2, pp. 1–21, Aug. 2013.
[37] *Definition of 'Deception'*, Oxford English Dictionary, London, U.K., 1989.
[38] K. Dinakar, B. Jones, C. Havasi, H. Lieberman, and R. Picard, "Common sense reasoning for detection, prevention, and mitigation of cyberbullying," *ACM Trans. Interact. Intell. Syst.*, vol. 2, no. 3, pp. 1–30, Sep. 2012.
[39] K. Ding, N. Pantic, Y. Lu, S. Manna, and M. I. Husain, "Towards building a word similarity dictionary for personality bias classification of phishing email contents," in *Proc. IEEE 9th Int. Conf. Semantic Comput.*, Feb. 2015, pp. 252–259.
[40] Y. Ding, N. Luktarhan, K. Li, and W. Slamu, "A keyword-based combination approach for detecting phishing Webpages," *Comput. Secur.*, vol. 84, pp. 256–275, Jul. 2019.
[41] M. Diomidous, K. Chardalias, A. Magita, P. Koutonias, P. Panagiotopoulou, and J. Mantas, "Social and psychological effects of the Internet use," *Acta Inf. Medica*, vol. 24, no. 1, pp. 66–68, 2016.
[42] D. Dittrich, "The ethics of social honeypots," *Res. Ethics*, vol. 11, no. 4, pp. 192–210, Dec. 2015.
[43] A. N. Doane, S. Ehlke, and M. L. Kelley, "Bystanders against cyberbullying: A video program for college students," *Int. J. Bullying Prevention*, vol. 2, no. 1, pp. 41–52, Mar. 2020.
[44] M. Egele, G. Stringhini, C. Kruegel, and G. Vigna, "COMPA: Detecting compromised accounts on social networks," in *Proc. NDSS*, 2013, pp. 1–17.
[45] P. Ekman, *Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage*. New York, NY, USA: Norton, 2009.
[46] Y. Elovici, M. Fire, A. Herzberg, and H. Shulman, "Ethical considerations when employing fake identities in online social networks for research," *Sci. Eng. Ethics*, vol. 20, no. 4, pp. 1027–1043, Dec. 2014.
[47] E. E. Englehardt and D. Evans, "Lies, deception, and public relations," *Public Relations Rev.*, vol. 20, no. 3, pp. 249–266, 1994.
[48] F. Enos, S. Benus, R. L. Cautin, M. Graciarena, J. Hirschberg, and E. Shriberg, "Personality factors in human deception detection: Comparing human to machine performance," in *Proc. 9th Int. Conf. Spoken Lang. Process.*, 2006, pp. 813–816.
[49] R. M. Everett, J. R. C. Nurse, and A. Erola, "The anatomy of online deception: What makes automated text convincing?" in *Proc. 31st Annu. ACM Symp. Appl. Comput. (SAC)*, 2016, pp. 1115–1120.
[50] A. Ebrahimi Fard, M. Mohammadi, Y. Chen, and B. Van de Walle, "Computational rumor detection without non-rumor: A one-class classification approach," *IEEE Trans. Comput. Social Syst.*, vol. 6, no. 5, pp. 830–846, Oct. 2019.
[51] D. A. Feingold, "Human trafficking," *Foreign Policy*, vol. 32, no. 150, pp. 26–30, Sep. 2005.
[52] E. Ferrara, O. Varol, C. Davis, F. Menczer, and A. Flammini, "The rise of social bots," *Commun. ACM*, vol. 59, no. 7, pp. 96–104, Jun. 2016.
[53] W. Ferreira and A. Vlachos, "Emergent: A novel data-set for stance classification," in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics, Hum. Lang. Technol.*, 2016, pp. 1163–1168.
[54] M. Fire, R. Goldschmidt, and Y. Elovici, "Online social networks: Threats and solutions," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 2019–2036, 4th Quart., 2014.
[55] M. Flintham, C. Karner, K. Bachour, H. Creswick, N. Gupta, and S. Moran, "Falling for fake news: Investigating the consumption of news via social media," in *Proc. CHI Conf. Hum. Factors Comput. Syst.*, 2018, pp. 1–10.
[56] D. Florêncio and C. Herley, "Evaluating a trial deployment of password re-use for phishing prevention," in *Proc. Anti-Phishing Work. Groups 2nd Annu. eCrime Researchers Summit*, 2007, pp. 26–36.
[57] M. Forelle, P. Howard, A. Monroy-Hernández, and S. Savage, "Political bots and the manipulation of public opinion in Venezuela," 2015, *arXiv:1507.07109*. [Online]. Available: http://arxiv.org/abs/1507.07109
[58] H. Gao, J. Hu, T. Huang, J. Wang, and Y. Chen, "Security issues in online social networks," *IEEE Internet Comput.*, vol. 15, no. 4, pp. 56–63, Jul./Aug. 2011.
[59] B. Gert, *Morality: Its Nature and Justification*, 6th ed. London, U.K.: Oxford Univ. Press, 2005.
[60] S. Ghosh, B. Viswanath, F. Kooti, N. K. Sharma, G. Korlam, F. Benevenuto, N. Ganguly, and K. P. Gummadi, "Understanding and combating link farming in the Twitter social network," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 61–70.
[61] J. Golbeck *et al.*, "Fake news vs satire: A dataset and analysis," in *Proc. 10th ACM Conf. Web Sci.*, 2018, pp. 17–21.
[62] W. Goucher, "Being a cybercrime victim," *Comput. Fraud Secur.*, vol. 2010, no. 10, pp. 16–18, Oct. 2010.
[63] P. A. Granhag and M. Hartwig, "A new theoretical perspective on deception detection: On the psychology of instrumental mind-reading," *Psychol., Crime Law*, vol. 14, no. 3, pp. 189–200, Jun. 2008.
[64] S. Grazioli and S. L. Jarvenpaa, "Perils of Internet fraud: An empirical investigation of deception and trust with experienced Internet consumers," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 30, no. 4, pp. 395–410, Jul. 2000.
[65] I. Greenberg, "The role of deception in decision theory," *J. Conflict Resolution*, vol. 26, no. 1, pp. 139–156, 1982.
[66] V. Greiman and C. Bain, "The emergence of cyber activity as a gateway to human trafficking," *J. Inf. Warfare*, vol. 12, no. 2, pp. 41–49, 2013.
[67] C. Grier, K. Thomas, V. Paxson, and M. Zhang, "@Spam: The underground on 140 characters or less," in *Proc. 17th ACM Conf. Comput. Commun. Secur.*, 2010, pp. 27–37.
[68] G. Gupta and J. Pieprzyk, "Socio-technological phishing prevention," *Inf. Secur. Tech. Rep.*, vol. 16, no. 2, pp. 67–73, May 2011.
[69] H. Haddadi and P. Hui, "To add or not to add: Privacy and social honeypots," in *Proc. IEEE Int. Conf. Commun. Workshops*, May 2010, pp. 1–5.
[70] T. Halevi, J. Lewis, and N. Memon, "A pilot study of cyber security and privacy related behavior and personality traits," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 737–744.
[71] T. Halevi, N. Memon, and O. Nov, "Spear-phishing in the wild: A real-world study of personality, phishing self-efficacy and vulnerability to spear phishing attacks," Dept. Comput. Sci. Eng., NYU Polytech. School Eng., New York, NY, USA, Tech. Rep., 2015, doi: 10.2139/ssrn.2544742.
[72] J. Hancock, L. E. Curry, S. Goorha, and M. Woodworth, "On lying and being lied to: A linguistic analysis of deception in computer-mediated communication," *Discourse Process.*, vol. 45, pp. 1–23, Jan. 2008.
[73] M. D. Hauser, "Minding the behavior of deception," in *Machiavellian Intelligence II: Extensions and Evaluations*. Cambridge, U.K.: Cambridge Univ. Press, 1997.
[74] R. Heartfield and G. Loukas, "A taxonomy of attacks and a survey of defence mechanisms for semantic social engineering attacks," *ACM Comput. Surv.*, vol. 48, no. 3, pp. 1–39, Feb. 2016.
[75] S. J. Heine, "Evolutionary explanations need to account for cultural variation," *Behav. Brain Sci.*, vol. 34, no. 1, pp. 26–27, Feb. 2011.
[76] M. Hernández-Álvarez, "Detection of possible human trafficking in Twitter," in *Proc. Int. Conf. Inf. Syst. Softw. Technol. (ICI2ST)*, 2019, pp. 187–191.
[77] G. Hofstede, "Dimensionalizing cultures: The Hofstede model in context," *Online Readings Psychol. Culture*, vol. 2, no. 1, p. 8, 2011.
[78] X. Hu, J. Tang, Y. Zhang, and H. Liu, "Social spammer detection in microblogging," in *Proc. 23rd Int. Joint Conf. Artif. Intell.*, 2013, pp. 2633–2639.
[79] X. Hu, J. Tang, H. Gao, and H. Liu, "Social spammer detection with sentiment information," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 180–189.
[80] S. Huckle and M. White, "Fake news: A technological approach to proving the origins of content, using blockchains," *Big Data*, vol. 5, no. 4, pp. 356–371, Dec. 2017.
[81] R. Hyman, "The psychology of deception," *Annu. Rev. Psychol.*, vol. 40, no. 1, pp. 133–154, 1989.
[82] I. Inuwa-Dutse, M. Liptrott, and I. Korkontzelos, "Detection of spam-posting accounts on Twitter," *Neurocomputing*, vol. 315, pp. 496–511, Nov. 2018.

[83] R. Isea and K. E. Lonngren, "A new variant of the seiz model to describe the spreading of a rumor," *Int. J. Data Sci. Anal.*, vol. 3, no. 4, pp. 28–33, 2017.

[84] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "CatchSync: Catching synchronized behavior in large directed graphs," in *Proc. 20th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2014, pp. 941–950.

[85] M. Jiang, P. Cui, A. Beutel, C. Faloutsos, and S. Yang, "Inferring strange behavior from connectivity pattern in social networks," in *Proc. Pacific–Asia Conf. Knowl. Discovery Data Mining*. Springer, 2014, pp. 126–138.

[86] S. Jiang and C. Wilson, "Linguistic signals under misinformation and fact-checking: Evidence from user comments on social media," in *Proc. ACM Hum.-Comput. Interact. (CSCW)*, vol. 2, 2018, pp. 1–23.

[87] F. Jin, E. Dougherty, P. Saraf, Y. Cao, and N. Ramakrishnan, "Epidemiological modeling of news and rumors on Twitter," in *Proc. 7th Workshop Social Netw. Mining Anal.*, 2013, pp. 1–9.

[88] Z. Jin, J. Cao, Y.-G. Jiang, and Y. Zhang, "News credibility evaluation on microblog with a hierarchical propagation model," in *Proc. IEEE Int. Conf. Data Mining*, Dec. 2014, pp. 230–239.

[89] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proc. 13th AAAI Conf. Artif. Intell.*, 2016, pp. 2972–2978.

[90] V. Kalmus, A. Realo, and A. Siibak, "Motives for Internet use and their relationships with personality traits and socio-demographic factors," *Trames, J. Hum. Social Sci.*, vol. 15, no. 4, pp. 385–403, 2011.

[91] G. A. Kamhoua, N. Pissinou, S. S. Iyengar, J. Beltran, C. Kamhoua, B. L. Hernandez, L. Njilla, and A. P. Makki, "Preventing colluding identity clone attacks in online social networks," in *Proc. IEEE 37th Int. Conf. Distrib. Comput. Syst. Workshops (ICDCSW)*, Jun. 2017, pp. 187–192.

[92] I. Kayes and A. Iamnitchi, "Privacy and security in online social networks: A survey," *Online Social Netw. Media*, vol. 3, pp. 1–21, Oct. 2017.

[93] S. Kemp. (2020). *More Than Half of the People on Earth Now Use Social Media*. [Online]. Available: https://blog.hootsuite.com/simon-kemp-social-media/

[94] M. L. Knapp, R. P. Hart, and H. S. Dennis, "An exploration of deception as a communication construct," *Hum. Commun. Res.*, vol. 1, no. 1, pp. 15–29, Sep. 1974.

[95] G. Kontaxis, I. Polakis, S. Ioannidis, and E. P. Markatos, "Detecting social network profile cloning," in *Proc. IEEE Int. Conf. Pervas. Comput. Commun. Workshops (PERCOM Workshops)*, Mar. 2011, pp. 295–300.

[96] C. Kopp, K. B. Korb, and B. I. Mills, "Information-theoretic models of deception: Modelling cooperation and diffusion in populations exposed to 'fake news,'" *PLoS ONE*, vol. 13, no. 11, 2018, Art. no. e0207383.

[97] R. E. Kraut and D. B. Poe, "Behavioral roots of person perception: The deception judgments of customs inspectors and laymen," *J. Pers. Social Psychol.*, vol. 39, no. 5, p. 784, 1980.

[98] S. Kumar and N. Shah, "False information on Web and social media: A survey," 2018, *arXiv:1804.08559*. [Online]. Available: http://arxiv.org/abs/1804.08559

[99] S. Kumar, F. Spezzano, and V. S. Subrahmanian, "Accurately detecting trolls in slashdot zoo via decluttering," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2014, pp. 188–195.

[100] S. Kumar, R. West, and J. Leskovec, "Disinformation on the Web: Impact, characteristics, and detection of Wikipedia hoaxes," in *Proc. 25th Int. Conf. World Wide Web, Int. World Wide Web Conf. Steering Committee*, 2016, pp. 591–602.

[101] S. Kumar, J. Cheng, J. Leskovec, and V. Subrahmanian, "An army of me: Sockpuppets in online discussion communities," in *Proc. 26th Int. Conf. World Wide Web*, 2017, pp. 857–866.

[102] S. Kumar, B. Hooi, D. Makhija, M. Kumar, C. Faloutsos, and V. Subrahmanian, "Rev2: Fraudulent user prediction in rating platforms," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 333–341.

[103] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*. [Online]. Available: http://arxiv.org/abs/1611.01236

[104] D. D. Langleben, L. Schroeder, J. A. Maldjian, R. C. Gur, S. McDonald, J. D. Ragland, C. P. O'Brien, and A. R. Childress, "Brain activity during simulated deception: An event-related functional magnetic resonance study," *NeuroImage*, vol. 15, no. 3, pp. 727–732, Mar. 2002.

[105] M. Latonero, "Human trafficking online: The role of social networking sites and online classifieds," SSRN, Tech. Rep., 2011, doi: 10.2139/ssrn.2045851.

[106] R. Y. K. Lau, Y. Xia, and Y. Ye, "A probabilistic generative model for mining cybercriminal networks from online social media," *IEEE Comput. Intell. Mag.*, vol. 9, no. 1, pp. 31–43, Feb. 2014.

[107] K. Lee, J. Caverlee, and S. Webb, "Uncovering social spammers: Social honeypots + machine learning," in *Proc. 33rd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, 2010, pp. 435–442.

[108] K. Lee, B. D. Eoff, and J. Caverlee, "Seven months with the devils: A long-term study of content polluters on Twitter," in *Proc. 5th Int. AAAI Conf. Weblogs Social Media*, 2011, pp. 185–192.

[109] K. Lee, P. Tamilarasan, and J. Caverlee, "Crowdturfers, campaigns, and social media: Tracking and revealing crowdsourced manipulation of social media," in *Proc. 7th Int. AAAI Conf. Weblogs Social Media*, 2013, pp. 331–340.

[110] K. Lee, J. Caverlee, and C. Pu, "Social spam, campaigns, misinformation and crowdturfing," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 199–200.

[111] C. C. Lewis and J. F. George, "Cross-cultural deception in social networking sites and face-to-face communication," *Comput. Hum. Behav.*, vol. 24, no. 6, pp. 2945–2964, Sep. 2008.

[112] D. Li and J. B. Cruz, "Information, decision-making and deception in games," *Decis. Support Syst.*, vol. 47, no. 4, pp. 518–527, Nov. 2009.

[113] G. Liang, W. He, C. Xu, L. Chen, and J. Zeng, "Rumor identification in microblogging systems based on users' behavior," *IEEE Trans. Comput. Social Syst.*, vol. 2, no. 3, pp. 99–108, Sep. 2015.

[114] T. Lin, D. E. Capecci, D. M. Ellis, H. A. Rocha, S. Dommaraju, D. S. Oliveira, and N. C. Ebner, "Susceptibility to spear-phishing emails: Effects of Internet user demographics and email content," *ACM Trans. Comput. Hum. Interact.*, vol. 26, no. 5, pp. 32:1–32:28, Jul. 2019.

[115] L. Liu, Y. Lu, Y. Luo, R. Zhang, L. Itti, and J. Lu, "Detecting 'smart' spammers on social network: A topic model approach," 2016, *arXiv:1604.08504*. [Online]. Available: http://arxiv.org/abs/1604.08504

[116] J. Ma, W. Gao, and K.-F. Wong, "Detect rumors in microblog posts using propagation structure via kernel learning," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, vol. 1, 2017, pp. 708–717.

[117] J. Ma, W. Gao, and K.-F. Wong, "Rumor detection on Twitter with tree-structured recursive neural networks," in *Proc. 56th Annu. Meeting Assoc. Comput. Linguistics, Long Papers*, vol. 1, 2018, pp. 1980–1989.

[118] N. K. Malhotra, S. S. Kim, and J. Agarwal, "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model," *Inf. Syst. Res.*, vol. 15, no. 4, pp. 336–355, Dec. 2004.

[119] B. Markines, C. Cattuto, and F. Menczer, "Social spam detection," in *Proc. 5th Int. Workshop Adversarial Inf. Retr. Web*, 2009, pp. 41–48.

[120] A. M. Matwyshyn, A. Cui, A. D. Keromytis, and S. J. Stolfo, "Ethics in security vulnerability research," *IEEE Secur. Privacy Mag.*, vol. 8, no. 2, pp. 67–72, Mar. 2010.

[121] N. McEvily, D. Novaes, K. Panesar, J. Moyer, A. Karr, B. Ng, and W. Ryan. (2018). *An Incentivized Blockchain Enabled Multimedia Ecosystem*. [Online]. Available: https://crushcrypto.com/wp-content/uploads/2018/02/CRNC-Whitepaper.pdf

[122] P. Mell and T. Grance, "The NIST definition of cloud computing," Comput. Secur. Division, Inf. Technol. Lab., Nat. Inst. Standards Technol., Tech. Rep., 2011.

[123] B. M. Meltzer, "Lying: Deception in human affairs," *Int. J. Sociol. Social Policy*, vol. 23, nos. 6–7, pp. 61–79, Jun. 2003.

[124] R. W. Mitchell, "A framework for discussing deception," in *Deception Perspectives on Human and Non-Human Deceit*. Albany, NY, USA: State Univ. of New York Press, 1986, pp. 3–40.

[125] T. Mitra and E. Gilbert, "Credbank: A large-scale social media corpus with associated credibility annotations," in *Proc. 9th Int. AAAI Conf. Web Social Media*, 2015, pp. 258–267.

[126] T. Mitra, G. Wright, and E. Gilbert, "Credibility and the dynamics of collective attention," in *Proc. ACM Hum.-Comput. Interact.*, vol. 1, Dec. 2017, pp. 1–17.

[127] T. Mitra, G. P. Wright, and E. Gilbert, "A parsimonious language model of social media credibility across disparate events," in *Proc. ACM Conf. Comput. Supported Cooperat. Work Social Comput.*, 2017, pp. 126–145.

[128] D. Modic and S. E. Lea, "How neurotic are scam victims, really? The big five and Internet scams," in *Proc. Conf. Int. Confederation Advancement Behav. Econ. Econ. Psychol.*, Exeter, U.K., 2011, pp. 1–24.

[129] E. Mussumeci and F. C. Coelho, "Modeling news spread as an SIR process over temporal networks," 2016, *arXiv:1701.07853*. [Online]. Available: http://arxiv.org/abs/1701.07853

[130] T. M. Negm, M. A. Rezqa, and A. F. Hegazi, "News credibility measure utilizing ontologies & semantic weighing schemes (NCMOSWS)," in *Proc. 2nd World Conf. Smart Trends Syst., Secur. Sustainability (WorldS4)*, Oct. 2018, pp. 57–64.
[131] M. E. Newman, "Spread of epidemic disease on networks," *Phys. Rev. E, Stat. Phys. Plasmas Fluids Relat. Interdiscip. Top.*, vol. 66, no. 1, pp. 016128:1–016128:11, 2002.
[132] M. L. Newman, J. W. Pennebaker, D. S. Berry, and J. M. Richards, "Lying words: Predicting deception from linguistic styles," *Pers. Social Psychol. Bull.*, vol. 29, no. 5, pp. 665–675, May 2003.
[133] Nextgate. (2019). *Research Report 2013 State of Social Media Spam*. [Online]. Available: https://www.slideshare.net/prayukth1/2013-state-of-social-media-spam-research-report
[134] S. D. Nicks, J. H. Korn, and T. Mainieri, "The rise and fall of deception in social psychology and personality research, 1921 to 1994," *Ethics Behav.*, vol. 7, no. 1, pp. 69–77, Mar. 1997.
[135] M. Nisrine, "A security approach for social networks based on honeypots," in *Proc. 4th IEEE Int. Colloq. Inf. Sci. Technol. (CiSt)*, Oct. 2016, pp. 638–643.
[136] B. K. Norambuena, M. Horning, and T. Mitra, "Evaluating the inverted pyramid structure through automatic 5w1h extraction and summarization," in *Proc. Comput. Journalism Conf.*, 2020, pp. 1–7.
[137] E. Novak and Q. Li, "A survey of security and privacy in online social networks," Dept. Comput. Sci., College William Mary, Williamsburg, VA, USA, Tech. Rep., 2012, pp. 1–32.
[138] B. Nyhan and J. Reifler, "When corrections fail: The persistence of political misperceptions," *Political Behav.*, vol. 32, no. 2, pp. 303–330, Jun. 2010.
[139] N. Nykodym, R. Taylor, and J. Vilela, "Criminal profiling and insider cyber crime," *Comput. Law Secur. Rev.*, vol. 21, no. 5, pp. 408–414, Jan. 2005.
[140] Y. Okada, K. Ikeda, K. Shinoda, F. Toriumi, T. Sakaki, K. Kazama, M. Numao, I. Noda, and S. Kurihara, "SIR-extended information diffusion model of false rumor and its prevention strategy for Twitter," *J. Adv. Comput. Intell. Intell. Informat.*, vol. 18, no. 4, pp. 598–607, 2014.
[141] D. Oliveira, H. Rocha, H. Yang, D. Ellis, S. Dommaraju, M. Muradoglu, D. Weir, A. Soliman, T. Lin, and N. Ebner, "Dissecting spear phishing emails for older vs young adults: On the interplay of weapons of influence and life domains in predicting susceptibility to phishing," in *Proc. Conf. Hum. Factors Comput. Syst. (CHI)*, 2017, pp. 6412–6424.
[142] A. Onuchowska and G.-J. de Vreede, "Disruption and deception in crowdsourcing: Towards a crowdsourcing risk framework," AIS eLibrary, Tech. Rep., 2018.
[143] G. Ortmann, "On drifting rules and standards?" *Scandin. J. Manage.*, vol. 26, no. 2, pp. 204–214, 2010.
[144] A. Ortony, G. L. Clore, and M. A. Foss, "The referential structure of the affective lexicon," *Cognit. Sci.*, vol. 11, no. 3, pp. 341–364, Jul. 1987.
[145] A. Paradise, R. Puzis, and A. Shabtai, "Anti-reconnaissance tools: Detecting targeted socialbots," *IEEE Internet Comput.*, vol. 18, no. 5, pp. 11–19, Sep. 2014.
[146] A. Paradise, A. Shabtai, and R. Puzis, "Hunting organization-targeted socialbots," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining*, Aug. 2015, pp. 537–540.
[147] A. Paradise, A. Shabtai, R. Puzis, A. Elyashar, Y. Elovici, M. Roshandel, and C. Peylo, "Creation and management of social network honeypots for detecting targeted cyber attacks," *IEEE Trans. Comput. Social Syst.*, vol. 4, no. 3, pp. 65–79, Sep. 2017.
[148] J. Parrish, J. L. Bailey, and J. F. Courtney, "A personality based model for determining susceptibility to phishing attacks," Univ. Arkansas Little Rock, Little Rock, AR, USA, Tech. Rep., 2009.
[149] M. Pattinson, C. Jerram, K. Parsons, A. McCormac, and M. Butavicius, "Why do some people manage phishing e-mails better than others?" *Inf. Manage. Comput. Secur.*, vol. 20, no. 1, pp. 18–28, 2012.
[150] A. Patwardhan, S. M. Noble, and C. M. Nishihara, "The use of strategic deception in relationships," *J. Services Marketing*, vol. 23, no. 5, pp. 318–325, Jul. 2009.
[151] J. Pawlick, E. Colbert, and Q. Zhu, "A game-theoretic taxonomy and survey of defensive deception for cybersecurity and privacy," *ACM Comput. Surv.*, vol. 52, no. 4, pp. 1–28, Sep. 2019.
[152] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annu. Rev. Psychol.*, vol. 54, no. 1, pp. 547–577, Feb. 2003.
[153] L. Rainie. (2018). *Americans' Complicated Feelings About Social Media in an Ear of Privacy Concerns*. [Online]. Available: https://pewrsr.ch/2pJczTZ
[154] S. Rathore, P. K. Sharma, V. Loia, Y.-S. Jeong, and J. H. Park, "Social network security: Issues, challenges, threats, and solutions," *Inf. Sci.*, vol. 421, pp. 43–69, 2017.
[155] J. Ratkiewicz, M. Conover, M. Meiss, B. Gonçalves, S. Patil, A. Flammini, and F. Menczer, "Truthy: Mapping the spread of astroturf in microblog streams," in *Proc. 20th Int. Conf. Companion World Wide Web*, 2011, pp. 249–252.
[156] R. J. Reinhart. (2018). *One in Four Americans Have Experienced Cybercrime*. [Online]. Available: https://news.gallup.com/poll/245336/one-four-americans-experienced-cybercrime.aspx
[157] R. E. Riggo and H. S. Friedman, "Individual differences and cues to deception," *J. Pers. Social Psychol.*, vol. 45, no. 4, pp. 899–915, 1983.
[158] N. C. Rowe and J. Rrushi, *Introduction to Cyberdeception*. Cham, Switzerland: Springer, 2016.
[159] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: Three types of fakes," in *Proc. 78th ASIS&T Annu. Meeting, Inf. Sci. Impact, Res. Community (ASIST)*, Silver Springs, MD, USA: American Society for Information Science, 2015, pp. 83:1–83:4.
[160] L.-M. Russow, "Deception: A philosophical perspective," in *Deception Perspectives on Human and Non-Human Deceit*. Albany, NY, USA: State University of New York Press, 1986, pp. 3–40.
[161] M. Saad, A. Ahmad, and A. Mohaisen, "Fighting fake news propagation with blockchains," in *Proc. IEEE Conf. Commun. Netw. Secur. (CNS)*, Jun. 2019, pp. 1–4.
[162] S. R. Sahoo and B. B. Gupta, "Hybrid approach for detection of malicious profiles in Twitter," *Comput. Electr. Eng.*, vol. 76, pp. 65–81, Jun. 2019.
[163] S. Samonas and D. Coss, "The CIA strikes back: Redefining confidentiality, integrity and availability in security," *J. Inf. Syst. Secur.*, vol. 10, no. 3, pp. 21–45, 2014.
[164] B. R. Schlenker and M. R. Leary, "Social anxiety and self-presentation: A conceptualization model," *Psychol. Bull.*, vol. 92, no. 3, pp. 641–669, 1982.
[165] Scikit-Learn. (2019). *Metrics and Scoring: Quantifying the Quality of Predictions*. [Online]. Available: https://scikit-learn.org/stable/modules/model_evaluation.html
[166] S. Sedhai and A. Sun, "Semi-supervised spam detection in Twitter stream," *IEEE Trans. Comput. Social Syst.*, vol. 5, no. 1, pp. 169–175, Mar. 2018.
[167] C. Sedikides and M. J. Strube, "The multiply motivated self," *Pers. Social Psychol. Bull.*, vol. 21, no. 12, pp. 1330–1335, Dec. 1995.
[168] J. Seiffert-Brockmann and K. Thummes, "Self-deception in public relations. A psychological and sociological approach to the challenge of conflicting expectations," *Public Relations Rev.*, vol. 43, no. 1, pp. 133–144, Mar. 2017.
[169] Z. Shan, H. Cao, J. Lv, C. Yan, and A. Liu, "Enhancing and identifying cloning attacks in online social networks," in *Proc. 7th Int. Conf. Ubiquitous Inf. Manage. Commun.*, 2013, pp. 1–6.
[170] C. Shao, G. L. Ciampaglia, A. Flammini, and F. Menczer, "Hoaxy: A platform for tracking online misinformation," in *Proc. 25th Int. Conf. Companion World Wide Web, Int. World Wide Web Conf. Steering Committee*, 2016, pp. 745–750.
[171] S. Sheng, M. Holbrook, P. Kumaraguru, L. Cranor, and J. Downs, "Who falls for phish? A demographic analysis of phishing susceptibility and effectiveness of interventions," in *Proc. Conf. Hum.-Comput. Interact. (CHI)*, Atlanta, GA, USA, 2010, pp. 373–382.
[172] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD Explor. Newslett.*, vol. 19, no. 1, pp. 22–36, 2017.
[173] H. A. Smith, *The Compleat Practical Joker*. New York, NY, USA: Morrow, 1953.
[174] J. Song, S. Lee, and J. Kim, "Crowdtarget: Target-based detection of crowdturfing in online social networks," in *Proc. 22nd ACM SIGSAC Conf. Comput. Commun. Secur.*, 2015, pp. 793–804.
[175] L. Song, R. Y. K. Lau, and C. Yin, "Discriminative topic mining for social spam detection," in *Proc. Pacific Asia Conf. Inf. Syst. (PACIS)*, 2014, pp. 378–394.
[176] S. A. Spence, T. F. D. Farrow, A. E. Herford, I. D. Wilkinson, Y. Zheng, and P. W. R. Woodruff, "Behavioural and functional anatomical correlates of deception in humans," *Neuroreport*, vol. 12, no. 13, pp. 2849–2853, Sep. 2001.

[177] G. Stringhini, C. Kruegel, and G. Vigna, "Detecting spammers on social networks," in *Proc. 26th Annu. Comput. Secur. Appl. Conf.*, 2010, pp. 1–9.

[178] M. M. Swe and N. Nyein Myo, "Fake accounts detection on Twitter using blacklist," in *Proc. IEEE/ACIS 17th Int. Conf. Comput. Inf. Sci. (ICIS)*, Jun. 2018, pp. 562–566.

[179] Symantec. (2019). *Internet Security Threat Report*. [Online]. Available: https://www.symantec.com/security-center/threat-report

[180] S. Tadelis, *Game Theory*. Princeton, NJ, USA: Princeton Univ. Press, 2013.

[181] R. Tembe, O. Zielinska, Y. Liu, K. W. Hong, E. Murphy-Hill, C. Mayhorn, and X. Ge, "Phishing in international waters: Exploring cross-national differences in phishing conceptualizations between Chinese, Indian and American samples," in *Proc. Symp. Bootcamp Sci. Secur. (HotSoS)*. New York, NY, USA: Association for Computing Machinery, 2014, pp. 1–7.

[182] L. ten Brinke and S. Porter, "Cry me a river: Identifying the behavioral consequences of extremely high-stakes interpersonal deception," *Law Hum. Behav.*, vol. 36, no. 6, pp. 469–477, 2012.

[183] R. Trivers, *Deceit and Self-Deception*. Springer, 2010, pp. 373–393.

[184] A. Troisi, "Displacement activities as a behavioral measure of stress in nonhuman primates and human subjects," *Stress*, vol. 5, no. 1, pp. 47–54, 2002.

[185] S. Tschiatschek, A. Singla, M. G. Rodriguez, A. Merchant, and A. Krause, "Fake news detection in social networks via crowd signals," in *Proc. Web Conf. Companion, Int. World Wide Web Conf. Steering Committee*, 2018, pp. 517–524.

[186] M. Tsikerdekis, "Identity deception prevention using common contribution network data," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 1, pp. 188–199, Jan. 2017.

[187] M. Tsikerdekis and S. Zeadally, "Online deception in social media," *Commun. ACM*, vol. 57, no. 9, pp. 72–80, Sep. 2014.

[188] G. Tsoumakas, I. Katakis, and I. Vlahavas, "Mining multi-label data," in *Data Mining and Knowledge Discovery Handbook*. Springer, 2009, pp. 667–685.

[189] E. C. Tupes and R. E. Christal, "Recurrent personality factors based on trait ratings," *J. Pers.*, vol. 60, no. 2, pp. 225–251, Jun. 1992.

[190] B. E. Turvey, *Criminal Profiling: An Introduction to Behavioral Evidence Analysis*. New York, NY, USA: Academic, 2011.

[191] Twitter Help. (2019). *The Twitter Rules*. [Online]. Available: https://help.twitter.com/en/rules-and-policies/twitter-rules

[192] S. G. A. van de Weijer, R. Leukfeldt, and W. Bernasco, "Determinants of reporting cybercrime: A comparison between identity theft, consumer fraud, and hacking," *Eur. J. Criminol.*, vol. 16, no. 4, pp. 486–508, Jul. 2019.

[193] C. VanDam, P.-N. Tan, J. Tang, and H. Karimi, "CADET: A multi-view learning framework for compromised account detection on Twitter," in *Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM)*, Aug. 2018, pp. 471–478.

[194] M. Vergelis, T. Shcherbakova, and T. Sidorina. (2019). *Spam and Phishing in Q1 2019*. [Online]. Available: https://securelist.com/spam-and-phishing-in-q1-2019/90795/

[195] N. Virvilis, B. Vanautgaerden, and O. S. Serrano, "Changing the game: The art of deceiving sophisticated attackers," in *Proc. 6th Int. Conf. Cyber Conflict (CyCon)*, Jun. 2014, pp. 87–97.

[196] A. Vishwanath, "Habitual facebook use and its impact on getting deceived on social media," *J. Comput.-Mediated Commun.*, vol. 20, no. 1, pp. 83–98, Jan. 2015.

[197] R. von Solms and J. van Niekerk, "From information security to cyber security," *Comput. Secur.*, vol. 38, pp. 97–102, Oct. 2013.

[198] S. Vosoughi and D. Roy, "A human-machine collaborative system for identifying rumors on Twitter," in *Proc. IEEE Int. Conf. Data Mining Workshop (ICDMW)*, Nov. 2015, pp. 47–50.

[199] S. Vosoughi, M. Mohsenvand, and D. Roy, "Rumor gauge: Predicting the veracity of rumors on Twitter," *ACM Trans. Knowl. Discovery Data*, vol. 11, no. 4, pp. 1–36, 2017.

[200] S. Vosoughi, D. Roy, and S. Aral, "The spread of true and false news online," *Science*, vol. 359, no. 6380, pp. 1146–1151, Mar. 2018.

[201] A. Vrij, "Why professionals fail to catch liars and how they can improve," *Legal Criminol. Psychol.*, vol. 9, no. 2, pp. 159–181, Sep. 2004.

[202] A. Vrij, R. Fisher, S. Mann, and S. Leal, "Detecting deception by manipulation cognitive load," *Trends Cognit. Sci.*, vol. 10, no. 4, pp. 141–142, 2006.

[203] C. Wagner, S. Mitter, C. Körner, and M. Strohmaier, "When social bots attack: Modeling susceptibility of users in online social networks," in *Proc. MSM*, 2012, pp. 41–48.

[204] H. G. Wallbott and K. R. Scherer, "Stress specificities: Differential effects of coping style, gender, and type of stressor on autonomic arousal, facial expression, and subjective feeling," *J. Pers. Social Psychol.*, vol. 61, no. 1, pp. 147–156, 1991.

[205] G. Wang, C. Wilson, X. Zhao, Y. Zhu, M. Mohanlal, H. Zheng, and B. Y. Zhao, "Serf and turf: Crowdturfing for fun and profit," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 679–688.

[206] G. Wang, T. Wang, H. Zheng, and B. Y. Zhao, "Man vs. machine: Practical adversarial detection of malicious crowdsourcing workers," in *Proc. 23rd USENIX Secur. Symp. (USENIX Security)*, 2014, pp. 239–254.

[207] T. Wang, G. Wang, X. Li, H. Zheng, and B. Y. Zhao, "Characterizing and detecting malicious crowdsourcing," *ACM SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 537–538, 2013.

[208] S. Webb, J. Caverlee, and C. Pu, "Social honeypots: Making friends with a spammer near you," in *Proc. CEAS*, 2008, pp. 1–10.

[209] T. Weller, "Compromised account detection based on clickstream data," in *Proc. Web Conf. Companion, Int. World Wide Web Conf. Steering Committee*, 2018, pp. 819–823.

[210] B. Whaley, "Toward a general theory of deception," *J. Strategic Stud.*, vol. 5, no. 1, pp. 178–192, Mar. 1982.

[211] H. Whittle, C. Hamilton-Giachritsis, A. Beech, and G. Collings, "A review of young people's vulnerabilities to online grooming," *Aggression Violent Behav.*, vol. 18, no. 1, pp. 135–146, Jan. 2013.

[212] Wikipedia. (2019). *Spearman's Rank Correlation Coefficient*. [Online]. Available: https://en.wikipedia.org/wiki/Spearman%27s_rank_correlation_coefficient

[213] E. J. Williams, A. Beardmore, and A. N. Joinson, "Individual differences in susceptibility to online influence: A theoretical review," *Comput. Hum. Behav.*, vol. 72, pp. 412–421, Jul. 2017.

[214] J. Wolak, D. Finkelhor, K. Mitchell, and M. Ybarra, "Online 'predators' and their victims: Myths, realities, and implications for prevention and treatment," *Amer. Psychol.*, vol. 63, no. 2, pp. 111–128, 2010.

[215] R. Wright, S. Chakraborty, A. Basoglu, and K. Marett, "Where did they go right? Understanding the deception in phishing communications," *Group Decis. Negotiation*, vol. 19, no. 4, pp. 391–416, Jul. 2010.

[216] B. Wu, F. Morstatter, X. Hu, and H. Liu, *Mining Misinformation in Social Media*. Boca Raton, FL, USA: CRC Press, 2016, pp. 135–162.

[217] K. Wu, S. Yang, and K. Q. Zhu, "False rumors detection on Sina Weibo by propagation structures," in *Proc. IEEE 31st Int. Conf. Data Eng.*, Apr. 2015, pp. 651–662.

[218] L. Wu and H. Liu, *Detecting Crowdturfing in Social Media*. New York, NY, USA: Springer, 2017, pp. 1–9.

[219] L. Wu and H. Liu, "Tracing fake-news footprints: Characterizing social media messages by how they propagate," in *Proc. 11th ACM Int. Conf. Web Search Data Mining*, 2018, pp. 637–645.

[220] X. Dong, J. A. Clark, and J. Jacob, "Modelling user-phishing interaction," in *Proc. Conf. Hum. Syst. Interact.*, May 2008, pp. 627–632.

[221] C. Yang, R. Harkreader, J. Zhang, S. Shin, and G. Gu, "Analyzing spammers' social networks for fun and profit: A case study of cyber criminal ecosystem on Twitter," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 71–80.

[222] C. Yang, J. Zhang, and G. Gu, "A taste of tweets: Reverse engineering Twitter spammers," in *Proc. 30th Annu. Comput. Secur. Appl. Conf.*, 2014, pp. 86–95.

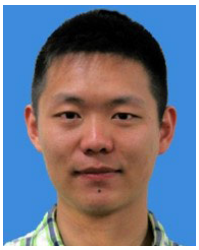
[223] P. Yang, G. Zhao, and P. Zeng, "Phishing Website detection based on multidimensional features driven by deep learning," *IEEE Access*, vol. 7, pp. 15196–15209, 2019.

[224] Y. Yao, B. Viswanath, J. Cryan, H. Zheng, and B. Y. Zhao, "Automated crowdturfing attacks and defenses in online review systems," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, Oct. 2017, pp. 1143–1158.

[225] H. Yu, P. B. Gibbons, M. Kaminsky, and F. Xiao, "Sybillimit: A near-optimal social network defense against sybil attacks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2008, pp. 3–17.

[226] P. Zambrano, J. Torres, L. Tello-Oquendo, R. Jácome, M. E. Benalcázar, R. Andrade, and W. Fuertes, "Technical mapping of the grooming anatomy using machine learning paradigms: An information security approach," *IEEE Access*, vol. 7, pp. 142129–142146, 2019.

[227] L. Zhao, Q. Wang, J. Cheng, Y. Chen, J. Wang, and W. Huang, "Rumor spreading model with consideration of forgetting mechanism: A case of online blogging LiveJournal," *Phys. A, Stat. Mech. Appl.*, vol. 390, no. 13, pp. 2619–2625, Jul. 2011.
[228] L. Zhao, J. Wang, Y. Chen, Q. Wang, J. Cheng, and H. Cui, "SIHR rumor spreading model in social networks," *Phys. A, Stat. Mech. Appl.*, vol. 391, no. 7, pp. 2444–2453, Apr. 2012.
[229] L. Zhao, H. Cui, X. Qiu, X. Wang, and J. Wang, "SIR rumor spreading model in the new media age," *Phys. A, Stat. Mech. Appl.*, vol. 392, no. 4, pp. 995–1003, Feb. 2013.
[230] L. Zhou and D. Zhang, "Following linguistic footprints: Automatic deception detection in online communication," *Commun. ACM*, vol. 51, no. 9, pp. 119–122, 2008.
[231] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," 2018, *arXiv:1812.00315*. [Online]. Available: http://arxiv.org/abs/1812.00315
[232] H. Zhu, "Fighting against social spammers on Twitter by using active honeypots," Ph.D. dissertation, Dept. Elect. Comput. Eng., McGill Univ. Libraries, Montreal, QC, Canada, 2015.
[233] Q. Zhu, A. Clark, R. Poovendran, and T. Basar, "SODEXO: A system framework for deployment and exploitation of deceptive honeybots in social networks," 2012, *arXiv:1207.5844*. [Online]. Available: http://arxiv.org/abs/1207.5844
[234] Q. Zhu, A. Clark, R. Poovendran, and T. Başar, "Deployment and exploitation of deceptive honeybots in social networks," in *Proc. 52nd IEEE Conf. Decis. Control*, Dec. 2013, pp. 212–219.
[235] M. Zuckerman, B. M. DePaulo, and R. Rosenthal, "Verbal and nonverbal communication of deception," in *Advances in Experimental Social Psychology*, vol. 14, L. Berkowitz, Ed. New York, NY, USA: Academic, 1981, pp. 1–59.



ZHEN GUO received the M.S. degree in biological sciences and the M.S. degree in computer science from Fordham University, New York City, in 2013 and 2016, respectively. He is currently pursuing the Ph.D. degree in computer sciences with the Virginia Polytechnic Institute and State University, Falls Church, VA, USA. His research interests include online social deception and social capital-based friending decision networks.

JIN-HEE CHO (Senior Member, IEEE) received the M.S. and Ph.D. degrees in computer science from Virginia Tech, in 2004 and 2008, respectively. She has been an Associate Professor with the Department of Computer Science, Virginia Tech, since 2018. Prior to joining the Virginia Tech, she has also been working as a Computer Scientist with the U.S. Army Research Laboratory (USARL), Adelphi, MD, USA, since 2009. She has published over 100 peer-reviewed technical articles in leading journals and conferences 140 the areas of trust management, cybersecurity, metrics and measurements, network performance analysis, resource allocation, agent-based modeling, uncertainty reasoning and analysis, information fusion / credibility, and social network analysis. She is also a member of ACM. She received the best paper awards in IEEE TrustCom'2009, BRIMS'2013, IEEE GLOBECOM'2017, 2017 ARL's publication award, and IEEE CogSima 2018. She is a winner of the 2015 IEEE Communications Society William R. Bennett Prize in the Field of Communications Networking. In 2016, she was selected for the 2013 Presidential Early Career Award for Scientists and Engineers (PECASE).

ING-RAY CHEN (Member, IEEE) received the B.S. degree from National Taiwan University, and the M.S. and Ph.D. degrees in computer science from the University of Houston. He is currently a Professor with the Department of Computer Science, Virginia Tech. His research interests include trust and security, network and service management, and reliability and performance analysis of mobile wireless networks and cyber physical systems. He was a recipient of the IEEE Communications Society William R. Bennett Prize in Communications Networking and the U.S. Army Research Laboratory (ARL) Publication Award. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON SERVICES COMPUTING, the IEEE TRANSACTIONS ON NETWORK AND SERVICE MANAGEMENT, and *The Computer Journal*.

SRIJAN SENGUPTA received the B.Stat. and M.Stat. degrees from the Indian Statistical Institute, in 2007 and 2009, respectively, and the Ph.D. degree in statistics from the University of Illinois at Urbana-Champaign, in 2016. He is currently an Assistant Professor with the Department of Statistics, North Carolina State University. Before joining North Carolina State University, he worked as an Assistant Professor of statistics with Virginia Tech from 2016 to 2020, and a Risk Management Actuary from 2009 to 2011. His research interests include statistical network analysis, bootstrap and related resampling/subsampling methods, and machine learning. He is also a member of ASA, ICSA, and IISA. His awards include the Norton prize for outstanding Ph.D. thesis from the University of Illinois at Urbana-Champaign, the Birla Sun Life Academic Excellence Award from the Institute of Actuaries of India, and the IMS New Researcher Travel Award.

MICHIN HONG received the B.S. and M.S.W. degrees from Ewha Womans University, South Korea, and the Ph.D. degree from the University of Maryland. She is currently an Associate Professor with the Indiana University School of Social Work. Her research interests include social determinants affecting ethnic/racial disparities in health and access to health care. Recently, she has expanded her research to explore individuals' vulnerability in the online world.

TANUSHREE MITRA received the M.S. degree in computer science from Texas A&M University, in 2011, and the Ph.D. degree in computer science from the Georgia Institute of Technology, in 2017. She is currently an Assistant Professor with the Information School, University of Washington. From 2017 to 2020, she was an Assistant Professor with the Computer Science Department, Virginia Tech. She studies and builds large-scale social computing systems to understand and counter problematic information online. Her work employs a range of interdisciplinary methods from the fields of human–computer interaction, data mining, machine learning, and natural language processing. She received best paper honorable mention awards from ACM CHI 2015 and ACM CSCW 2020, the Virginia Tech College of Engineering's Outstanding New Assistant Professor Award in 2020, and the Georgia Tech's GVU Center's Foley Scholarship for research innovation and potential impact in 2015.