Towards an Automated Recognition System for Chat-based Social Engineering Attacks in Enterprise Environments

Nikolaos Tsinganos
University of Macedonia
Thessaloniki, Greece
tsinik@uom.edu.gr

Georgios Sakellariou
University of Macedonia
Thessaloniki, Greece
geosakel@uom.edu.gr

Panagiotis Fouliras
University of Macedonia
Thessaloniki, Greece
pfoul@uom.edu.gr

Ioannis Mavridis
University of Macedonia
Thessaloniki, Greece
mavridis@uom.edu.gr

ABSTRACT

Increase in usage of electronic communication tools (email, IM, Skype, etc.) in enterprise environments has created new attack vectors for social engineers. Billions of people are now using electronic equipment in their everyday workflow which means billions of potential victims of Social Engineering (SE) attacks. Human is considered the weakest link in cybersecurity chain and breaking this defense is nowadays the most accessible route for malicious internal and external users. While several methods of protection have already been proposed and applied, none of these focuses on chat-based SE attacks while at the same time automation in the field is still missing. Social engineering is a complex phenomenon that requires interdisciplinary research combining technology, psychology, and linguistics. Attackers treat human personality traits as vulnerabilities and use the language as their weapon to deceive, persuade and finally manipulate the victims as they wish. Hence, a holistic approach is required to build a reliable SE attack recognition system. In this paper we present the current state-of-the-art on SE attack recognition systems, we dissect a SE attack to recognize the different stages, forms, and attributes and isolate the critical enablers that can influence a SE attack to work. Finally, we present our approach for an automated recognition system for chat-based SE attacks that is based on Personality Recognition, Influence Recognition, Deception Recognition, Speech Act and Chat History.

CCS CONCEPTS

• **Security and privacy** → **Phishing**; • **Computing methodologies** → *Supervised learning*;

KEYWORDS

Social Engineering, Personality, Persuasion, Deception, Speech Act

ARES 2018, August 27–30, 2018, Hamburg, Germany

ACM ISBN 978-1-4503-6448-5/18/08…$15.00
https://doi.org/10.1145/3230833.3233277

ACM Reference Format:
Nikolaos Tsinganos, Georgios Sakellariou, Panagiotis Fouliras, and Ioannis Mavridis. 2018. Towards an Automated Recognition System for Chat-based Social Engineering Attacks in Enterprise Environments. In *ARES 2018: International Conference on Availability, Reliability and Security, August 27–30, 2018, Hamburg, Germany*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3230833.3233277

1 INTRODUCTION

In an assessment made in 2006 about users' awareness in Social Engineering (SE) methods in the form of email phishing attacks, Karakasiliotis et al. [17] reported that out of 179 participants 36% were successful in identifying legitimate emails, versus 45% that were successful in spotting illegitimate ones. Almost ten years later, in a similar assessment, Verizon in the 2015 Data Breach Investigation Report [37] presented the results of a test conducted by sending 150,000 emails; they reported that within the first hour, 50% of the recipients had opened the email and clicked on phishing links. The first user clicked the phishing link only after 82 seconds. Social Engineering is also recognized as the second reason for security breaches at 35%, right behind traditional hacking methods. Furthermore, today, it is a very common practice in workspaces to enable employees to use their own computers or other electronic mobile devices under 'bring your own device' (BYOD) policies. This increase in working at home magnifies the SE problem due to insufficiently protected personal computers. A social engineer's successful attack on an employee could results also in compromise of entire employer' s information system.

Until now, various methods have been used in order to protect the weakest link in the cyber security chain, the human. Such methods are penetration tests using social engineering techniques, security awareness training programs for the employees, creation and enforcement of corporate cyber security policies and development of security-aware organizational culture. Trying to uncover the social engineer's behavior, cyber security researchers noticed that this category of attacks needed an interdisciplinary approach that would help understand the inner workings of the attack, and the methods of social engineers in combination with the psychological characteristics of the human being manipulated. SE attacks are here to stay and threaten all users in enterprises, government agencies and every single individual.

Although, much research have been done regarding several forms of SE attacks, the rise of cyber communication tools usage is a strong motivation to design stronger defenses for chat-based SE attacks. While many of social engineering attack vectors and different communication channels exist, direct human-to-human communication offers attackers critical advantage and instantaneous results.

In Section 2, a comprehensive literature review for the current state-of-the-art in SE attack recognition systems, focusing on attacks that involve text-based conversation between the attacker and the victim, is presented. Section 3 summarizes our findings regarding the SE attack cycle, the various forms of SE attacks and the related attack attributes. Section 4 presents the SE attack enablers, namely human Personality, influence, deception, speech act and chat history. Our proposed approach towards an automated recognition system for chat-based SE attacks in enterprise environments is presented in Section 5 and we conclude in Section 6.

2 RELATED WORK

Hoeschele and Rogers [14] presented the Social Engineering Defense Architecture (SEDA) - an architecture for detecting social engineering attacks over phone conversation in real-time. The architecture uses a storing facility to save caller details (voice signature, etc.) and can provide authentication services, too. Subsequently, Hoeschele [15], presented a SEDA proof-of-concept model where some simple SE attack detection processes were implemented along with a database to store all gathered information. The model managed to process and then detect all attacks resulting in 100% accuracy. Nevertheless, that system lacks the use of previous activity history and personality recognition characteristics for both attacker and victim. In [3] the authors propose an architecture called Social Engineering Attack Detection Model (SEADM). Their system helps users decide by using a simple binary decision tree model. The authors make many unrealistic assumptions in order to justify the logic behind their proposed system. SEADM had a second chance in [23] and also an android implementation as a proof of concept. The authors revised SEADM to cater for SE attacks utilization of unidirectional, bidirectional and indirect communication between the attacker and the victim. The proposed and revised SEADMv2 extends the previous model. Bhakta et al. [4] argue that the most effective SE attacks involve a dialog between the attacker and the victim. Their approach uses a predefined Topic Blacklist (TBL) against which dialog sentences are checked. The TBL is manually populated with pairs of verbs (actions) and nouns (objects) using security policy documents or other expert knowledge. The dialogues are then processed using natural language processing (NLP) techniques. The authors claim 100% precision and 88.9% recall using their approach. Unfortunately, they do not present a classification accuracy value.

Following the above work, Sawa et al. [29] used more advanced language processing techniques striking a balance between syntactic and semantic analysis. A handful of various tools are used (Stanford parser, Penn tagset symbols, Tregex tool and others) in order to generate a parse tree and then search for questions or commands. This approach is still using the TBL and the results shown are 100% precision and 60% recall for the first corpus (a fabricated dataset composed of three phone conversations between professional social engineers and unaware victims). The results on the second corpus (Supreme Court Dialog Corpus) are showing zero false positive. The researchers did not present an accuracy value, while at the same time the dataset is very small for measuring precision and recall. Due to the same reason (small dataset with only three conversations) the results of precision and recall are weak as a success measure. Furthermore, the researchers did not take into account any context information during the classification process. Therefore, the algorithm is unaware of the intricacies of the specific environment it is operated upon. Another disadvantage is that the process is not fully automated, since the creation of the TBL requires human involvement. Furthermore, the authors did not consider the target as a factor of influence in the process and they did not use cognitive models or any other personality traits.

Finally, Uebelacker et al. [36] propose a SE attacks taxonomy based on Cialdini's Influence principles. More specifically, they study the relationship between the Big-5 Theory (personality traits) and Cialdini's influence (Persuasion) principles and finally propose a theory-based SE Personality Framework (SEPF). Moreover, they propose a complete research road map for their future work on SEPF . They define three domains related to cyber security, namely: physical, digital, and social. They focus on the social domain related to the victims (employees) excluding the attackers. After a thorough study of the related literature they summarize their findings as follows: *"Conscientiousness, Extraversion, and Openness... show both increased and decreased susceptibility to SE depending on context and sub-traits"*. Furthermore, *"Agreeableness increases and Neuroticism decreases susceptibility to SE"*.

3 BACKGROUND

In a typical social engineering attack, the attacker acts in a predetermined manner, where she initially *gathers information* using every possible technique or tool, then approaches the potential victim and *develops a trust relationship*. Next, she *exploits* this trust relationship to *manipulate* the victim to perform an action that would enable her to violate the respective information system. At the final stage, the attacker reaches her original target violating a CIA triad member (confidentiality, integrity, availability) of informational resources.

In order for the attacker to develop a trust relationship, she relies on specific human (victim) *personality traits* treating them as vulnerabilities and adapting her tactics accordingly. Her aim is to *influence* the victim's way of thinking, and to *persuade* him to behave in a mistaken way. The act of *deception* is underlying throughout the attacker's effort. A communication scenario between the SE attacker and her victims involves message exchange through an electronic chat system. This is the point where our efforts on recognizing SE attacks are focusing.

A SE attack is mainly related to deception and concerns every human activity, making it difficult to precisely predefine and recognize it by only syntactic or semantic analysis of the chat messages. Furthermore, human language ambiguity makes discriminating a sentence as malicious or not, even harder. To cope with this challenge, a researcher has to employ a toolkit (e.g., machine learning tools) to process all available data and to infer in a probabilistic

way. Moreover, for an automated SE attack recognition system to be efficient it has to embrace several scientific disciplines.

3.1 SE Attack Cycle

SocialEngineering is defined in [31] as "a deceptive process in which crackers 'engineer' or design a social situation to trick others into allowing them access to an otherwise closed network, or into believing a reality that does not exist." According to Mitnick et al. [22] a SE attack, also known as the SE attack cycle, is composed of four stages:

- Information Gathering (IG)
- Development of Relationship (RD)
- Exploitation of Relationship (RE)
- Execution to achieve objective (EX)

The attacker gathers information from various public sources at *"Information Gathering"*, develops a trusting relationship with the victim at *"Relationship Development"*, exploits this relationship in order to steal valuable information at stage *"Relationship Exploitation"* and finally, having all necessary knowledge, attacks the real target in stage *"Execution"*. These four stages correspond to the attacker's steps during a SE attack. For an attacker to be successful and move from one stage to the other some conditions should be met. We focus on these conditions, which we call SE Attack Enablers. ISACA [1] defines enablers as *"Factors that, individually and collectively, influence whether something will work"*. SE Attack Enablers are further discussed in section 4.

3.2 SE Attack Attributes

A SE attack can be either human- or computer-based. In human-based attacks we have a human-to-human interaction (e.g., phone conversation), while computer-based attacks require the use of a digital medium [26]. SE attacks can also be categorized as direct, if the attacker is interacting with the victim (phone conversation, social media chat, etc.) or as indirect if some electronic medium mediates (phishing email, rogue website, etc.).

In [24], the author proposes a new model in order to describe the SE attack cycle. This model is called "The cycle of Deception" and is more of a conceptual model that combines models for the defense cycle, the victim behavior cycle and the attack cycle. Janczewski et al. [16] conducted an interview experiment of IT practitioners and proposed the following concepts as relevant to every SE attack: people, security awareness, psychological weakness, technology, defenses, attack methods, security strategy, technical controls, security-enhanced product, and education.

In [34], Tetri et al. tried to analyze functions of different techniques by extrapolating three dimensions: *persuasion*, *fabrication*, and *data gathering* in which they dissect all SE attacks to be easier to understand. Heartfield et al. [12] claim that SE attacks aiming at deceiving the user by means of phishing emails, scareware, or spoofed websites are semantic attacks. The authors present a taxonomy for semantic attacks and defense mechanisms. Another interesting taxonomy of SE attacks is presented by Krombholz et al. in [18], where the authors define three main categories: *Channel* which is the medium that the attacker uses (e.g., email, telephone, physical to contact the target), *Operator* which is a way to differentiate between human-based and automated attacks, and *Type* which is one of socio-technical, technical, physical or social attack category. This taxonomy seems more agile and easy to classify existing or new attack vectors as shown in the same work.

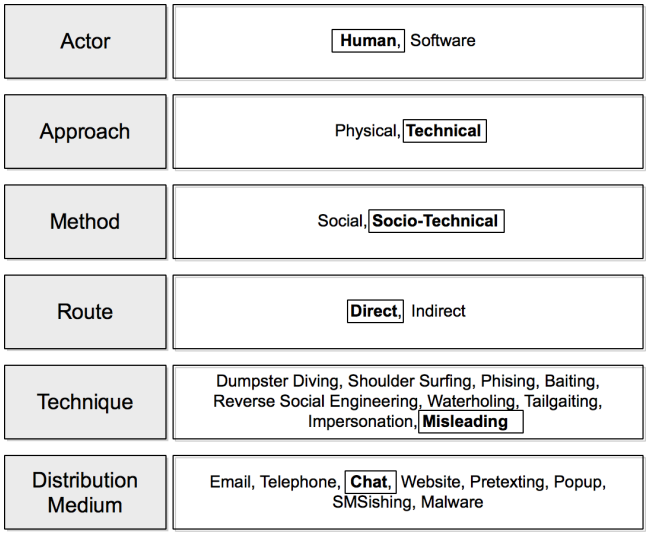

Figure 1: SE attack attributes.

Fig. 1 summarizes the previous works and presents a unified view of the most common attack attributes: actor, approach, method, route, technique and medium used to manipulate victims. Our work focus is shown in bordered, bold fonts for every different attribute.

4 SE ATTACK ENABLERS

Social engineering is a term that characterizes the general phenomenon of deception involving the field of information systems. Its success depends on specific traits of human personality. These personality traits define the way of human behavior. Our interested lies in traits that:

- Enhance the attacker's ability to influence and deceive.
- Make the victim vulnerable to manipulation.

An employee's previous conversations can also help us draw a more complete picture of his vulnerability level and trigger an alarm with more confidence if a threshold is exceeded. In the following sub-sections, the main SE attack enablers are presented that, in our belief, are decisive for the success or failure of a SE attack.

4.1 Personality

In psychology, human personality *"refers to individual differences in characteristic patterns of thinking, feeling and behaving"* and, although there is no universal acceptance, the Big-5 Theory analyzes a five-factor model (FFM) of the personality traits, or otherwise called factors to classify personalities. These factors are believed to capture most of the individual differences in terms of personality. The five factors, usually measured between 0 and 1, are [33]:

- conscientiousness: *"The degree to which individuals are hard-working, organized, dependable, reliable, and persevering versus lazy, unorganized, and unreliable."*

Table 1: Mapping of Influence Principles and Factors.

Cialdini (2001) [27]	Harl (1997) [11]	Gragg (2003) [8]	Granger (2001) [9]	Peltier (2006) [26]
Authority		Authority	Impersonation	
Scarcity		Strong Affect, Overloading		
Liking & Similarity	Diffusion of Responsibility, Personal Persuasion, Ingratiation	Deceptive Relationship, Diffusion of Responsibility	Ingratiation, Impersonation, Diffusion of Responsibility, Friendliness	Diffusion of Responsibility, Ingratiation, Trust Relationship,
Reciprocation	Co-operation	Reciprocation		
Social Proof	Involvement, Moral Duty	Moral Duty		Guilt
Commitment & Consistency	Conformity	Integrity/Consistency	Conformity	

- extraversion: *"The extent to which individuals are gregarious, assertive, and sociable versus reserved, timid, and quiet."*
- agreeableness: *"The degree to which individuals are cooperative, warm, and agreeable versus cold, disagreeable, rude, and antagonistic."*
- openness: *"the extend to which an individual has richness in fantasy life, aesthetic sensitivity, awareness of inner feelings, need for variety in actions, intellectual curiosity, and liberal values."*
- neuroticism: *"the degree to which one has negative effect, and also disturbed thoughts and behaviors that accompany emotional distress"*

Research in [5], [21], reports that high values on conscientiousness, extraversion and openness sometimes increase and sometimes decrease susceptibility to SE attacks. High values on agreeableness increase susceptibility and high values on neuroticism decrease susceptibility to SE attacks. The results are contradictory in many situations and they do not lead to a direct conclusion. Up till now, researchers have examined the relation between personality traits and social engineering by combing knowledge of human behavior in other fields (marketing, etc.). It would be of great benefit to analyze and measure the exact relation of personality traits with specific SE techniques.

Nevertheless, after the work of [19], several attempts have been made to exploit the results and apply the findings to different research fields.

4.2 Influence

As Schneier points out [28], human risk perception has evolved over thousands of years. Nevertheless, progress in technology has changed our lives very fast without allowing enough time for our risk perception to adjust to new threats. This vulnerability in human design is exploited by social engineers and then transferred to information systems to compromise them. Schneier discusses also heuristics (called shortcuts) in human behavior and biases. Both are causal factors for wrong appraisals and decisions. Robert Cialdini [35] agrees with Schneier and discusses the principles of influence and how heuristics and biases are exploited by a human to manipulate another human. Cialdini also argues that there are two types of influence: *compliance* and *persuasion*. Using persuasion the attacker sends a message and then the victim changes his behavior, attitude or knowledge as a result of the received message. Compliance forces the change of a behavior as a result of a direct request. The request can be explicit (hard) or implicit (soft). Cialdini conducted experiments and field studies on sales and marketing department employees, and defined six influence principles:

- Reciprocation: a social norm that make us repay others for what we have received. It builds trust between humans and we are all trained to adhere or suffer severe social disapproval. Humans feel obliged after receiving a gift.
- Commitment and Consistency: humans commit by stating who they are, based on what they do or think. They also like to be consistent because that builds character. Attackers exploit that kind of belief by initially asking for a small favor, then a bigger one and finally the big bad favor. Humans that have already served an attacker feel they have to show commitment and be consistent with their prior behavior.
- Social Proof: humans tend to believe what others do or think as right.
- Liking: if someone likes us and makes it obvious, it is hard to resist not to like him back. After that it is easier for him to ask us a favor and difficult for us to deny him one. On the opposite direction we all want to be liked
- Authority: humans tend to trust and obey experts or someone in a high hierarchical position. It is difficult for an employee to deny a request from an IT manager, for example.
- Scarcity: limited information leads to wrong decisions and limited resources are more desirable. If an attacker knows that an employee wants a specific application then she can offer it (after injecting an exploit), or claim a reason to request a favor based on evidence that only the user possesses.

Apart from Cialdini, many researchers tried to capture the psychological aspects of human behavior related to influence. Gragg [8] presents a list of such principles and calls them triggers: Strong affect, Overloading, Reciprocation, Deceptive Relationships, Diffusion of Responsibility, Authority, Integrity and Consistency. Scheeres [30] makes obvious the relationship between Gragg's and Cialdini's treatment by correlating all these principles and triggers. Granger

[9] and Peltier [26] present similar factors of influence based on their point of view.

Table 1 summarizes the mapping of the above factors along with Cialdini's principles. In our approach Cialdini's influence principles are chosen because there is a major overlap with all of the factors proposed by the other researchers.

4.3 Deception

An [2] describes Deception as *"an act or statement intended to make people believe something that the speaker does not believe to be true, or not the whole truth"*. A more precise definition for Deception is given in [10] where *"Deception is a successful or unsuccessful attempt, without forewarning, to create in another a belief which the communicator considers to be untrue"*. Over the years the research community became very interested in the detection of deception. Due to the interdisciplinary nature of the phenomenon, researchers from various scientific fields (psychology, computer science, linguistics, philosophy, etc.) have already presented their results by studying and analyzing several different deceptive cues (e.g., biometric indicators, facial indicators or gestural indicators). There are two categories of deception [2]:

- face saving: when humans lie to protect themselves, to avoid tension and conflict in a social interaction, or to minimize hurt feelings and ill will,
- malicious: when humans lie with harmful intent.

Our primary interest is in detecting a malicious deception attempt in a text-based conversation and use this finding as an extra indicator for recognizing a social engineering attempt. So far, several research attempts have been made studying verbal or nonverbal cues in order to detect deceptive behavior [25], [7]. Current work in deception detection is mainly based on verbal cues and has shown that it is possible to reliably predict a deception attempt [38]. In most of the works researchers have collected data and manually annotated them for deceptive status. After that, the labeled data were fed to a classification algorithm for supervised learning. The features extracted for text-based deception detection are critical and directly connected to prediction accuracy [25], [7].

The common scientific approach is to use three types of features, namely lexical, acoustic, and speech features. The most frequently used techniques for lexical analysis are: Linguistic Inquiry and Word Count (LIWC), N-gram, Part-of-speech (POS), and Dictionary of Affect in Language (DAL).

LIWC is primarily used for detecting psychological characteristics by calculating several metrics for usage of different word categories, usage of casual words, existence of positive or negative emotions in text, etc. In [13], [25] researchers used LICW to examine text-based communication and managed to extract valuable knowledge regarding people's personality, and cognitive and emotional characteristics. The above research works differ in accuracy results due to the use of different datasets that lead to accurate or less accurate machine learning algorithms. DAL is mostly used to analyze emotive content and its main difference from LICW is that it has a narrower focus. N-gram is usually combined with other more advanced techniques, like LICW to train binary classifiers (e.g., Naive Bayes, SVM, etc.) during a lexical analysis.

4.4 Speech Act

Theoretical linguistics inquire into the nature of human language and seek to answer fundamental questions as to what a language is, or the inner workings of it. Several different levels of analysis are defined, such as syntactic (studies the structure of the visible/audible form of the language), semantic (studies the relations and dependencies between different language structures and their potential meanings), and pragmatic (studies the issues related to language use due to context and uncovers the intention of the speaker in an utterance).

Our study on chat-based conversations can benefit by finding the ordering and patterns of interaction between two interlocutors. Our interest is in uncovering the actions that are hidden between the words and pragmatic analysis seems to be the appropriate approach from such a language/action perspective [39]. The starting point to study the pragmatics of language action is Speech Act Theory (SAT). According to SAT [32], the uttering of a sentence is an action, and in short form says that "saying is doing" or similarly "words are deeds". Austin claimed *"all utterances, in addition to meaning, perform specific acts via the specific communicative force of an utterance"* and introduced a three-fold distinction among the acts one simultaneously performs when saying something:

- Locutionary act: the production of a meaningful linguistic expression.
- Illocutionary act: the action intended to be performed by a speaker in uttering a linguistic expression, either explicitly or implicitly. Examples include: accusing, apologizing, refusing, ordering, etc.
- Perlocutionary act: the effect of the illocutionary act on the hearer such as persuading, deterring, surprising, misleading or convincing.

For example, the phrase of an IT technician: *"The operating system will reboot in five minutes."* results in saying that the OS will reboot in 5 minutes (locutionary act) and informs the users of the imminent rebooting of the OS (illocutionary act). By producing his utterance the IT technician intents to make users believe that the OS will reboot in 5 minutes and urges them to do housekeeping activities (perlocutionary act). The IT technician performs all these speech acts, at all three levels, just by uttering the above sentence.

Searle proposed speech acts to be classified into five categories along four dimensions (illocutionary point, direction of fit between words and world, psychological state, and propositional content):

- **Representatives** express the speaker's beliefs. Examples include claiming, reporting, asserting, stating and concluding. Using representatives the speaker makes words fit the world by representing the world as he believes it is.
- **Directives** express the speaker's desire to get the hearer act in a specific way. Examples include commands, advice, orders and requests. Using directives, the speaker intends to make the world match the words via the hearer. E.g., "Double-click this file."
- **Commissives** are used to express the speaker's intention and commitment to do something in the future. Examples include offers, pledges, promises, refusals, and threats. Using commissives, the speaker adapts the world to the words; e.g., "I'll never give you access to your account."

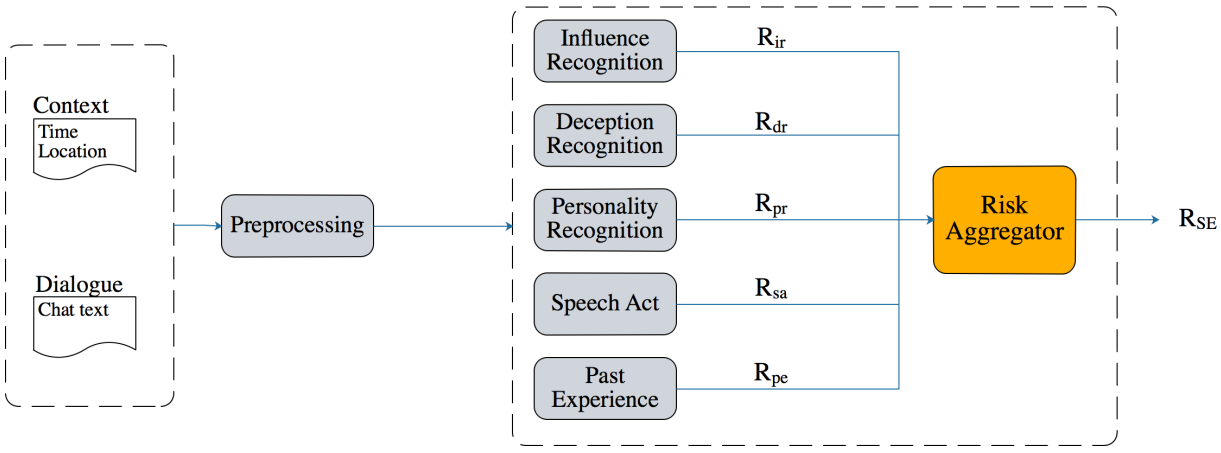

Figure 2: System Architecture.

- **Expressives** express the psychological state of the speaker such as joy and sorrow. Examples include praising, blaming, apologizing, and congratulating. There is no direction of fit for expressives; e.g. "Well done, John!"
- **Declaratives** are used to express immediate changes in the current state of some affair. Examples are firing (from employment), declaring war, etc. Both directions of fit, suit these type of speech act (words-to-world and world-to-words). E.g., "I object, Your Honor."

4.5 Chat History

This enabler refers to the technical challenge of assessing the risk of a potential SE attack through the history of a user's chat dialogues. In many cases, SE attacks take place in multiple repeated phases, where the offender is properly prepared before the attack commences. In particular, she creates a 'trust' relationship, which requires time to explore her victim until she finds the right spot for the attack to take place. Therefore, the purpose of this process is to utilize all previous chat dialogues between the same interlocutors, transform them to a measurable value and use it as an extra indicator for detecting a SE attack.

5 SYSTEM ARCHITECTURE

In order to protect users from SE attacks through person-to-person text communication, a technical solution is needed, beyond the training programs and psychological preparation. Such a technical solution could make use of all important factors to develop and implement an automated process for risk assessment during a chat conversation. However, it seems challenging as human personality traits can lead someone to be influential, persuasive, and deceptive while at the same time another human can be more or less vulnerable to deceptive acts.

Automated SE attack recognition means that there must be a clear decision making (even though probabilistically) on whether a person aims to intentionally deceive another person. Working in this direction, we designed an automated recognition system which functions in a linear manner based on Natural Language Processing (NLP) techniques along with psychological characteristics detection for both interlocutors. The system includes five recognition subsystems, namely: Influence Recognition (IR), Deception Recognition (DR), Personality Recognition (PR), Speech Act (SA) and Past Experience (PE). Each subsystem calculates a separate risk value (Rir, Rdr, Rpr, Rsa, Rpe), which is then fed to the Risk Aggregator that calculates the overall probability distribution of SE attack risk R_{SE} . Figure 2 presents a conceptual diagram of the automated SE attack recognition system. The tools and techniques used in every stage of a SE attack (Information Gathering - IG, Relationship Development - RD, Relationship Exploitation, and Execution to achieve objective - EX) in correspondence with the associated SE Enablers are depicted in Table 2.

Table 2: Enablers, Stage and Techniques

Enablers	Stages	Techniques
Personality Traits	IG, RD	Classification
Deception	IG, RD	Classification, Conversation for Action (Speech Act)
Influence/Persuasion	RD	Classification, Semantic Analysis
Speech Act	RE	Conversation for Action (Speech Act), Typed Dependency Trees, Named-Entity Recognition
Past Experience	IG, RD, RE, EX	Value Threshold

5.1 Dialogs, Context and Preprocessing

The dialogue text between the two interlocutors is the initial input for all system processes, along with contextual information. Contextual information may include time and location details, which can be used by the Past Experience subsystem, as described below.

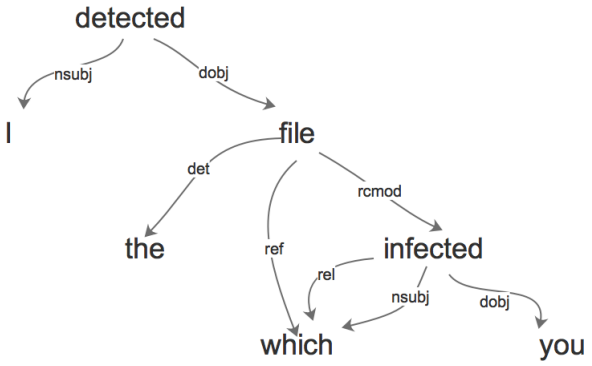

Figure 3: Typed Dependency Tree.

Location details (e.g., in the form of an IP address) is useful for separating insiders from outsiders and for controlling the use of different nicknames from the same location.

As a first step, the captured dialogue text is pre-processed with Natural Language Processing techniques. Depending on the original raw text, pre-processing (also collectively known as tokenization), comprises cleaning of unwanted tags and labels (e.g., HTML tags) and unnecessary capitalization, stemming (lemmatization), stop-words removal, syntactic analysis, vocabulary creation (for blacklist creation, topic modeling etc.), and annotation (e.g., POS tagging). Subsequently, the vectorized representation of the pre-processed dialogue text and contextual information is inputted to the recognition subsystems.

To efficiently process a sentence and extract valuable information, typed dependency trees are utilized to represent the structure of a sentence and all the dependencies between the individual words. All the dependencies are also labeled with grammatical relations (e.g., subject, object, indirect object, etc.). After parsing a sentence and representing it as a typed dependency tree, information about predicate-argument structures, which are not readily available from other structure parses, can be extracted easily.

A hierarchy of grammatical relations rooted with the most generic relation is created where the relations between heads and their dependents can be easily identified. The creation of such a tree is based on special rules/patterns that are applied on the corresponding phrase structure tree [6]. First, a dependency extraction is performed where a sentence is parsed using a phrase structure grammar parser, followed by a dependency typing where the head of each word of the sentence is identified using modified rules to retrieve the semantic head of each word rather than the syntactic head. In Figure 3, an example of a typed dependency tree is depicted, where *nsubj* means nominal subject, *dobj* means "direct object", *det* means determiner, *ref* means referent, *rel* means relative (word introducing a rcmod), and *rcmod* means relative clause modifier.

5.2 Influence Recognition

The system calculates the degree of influence of the attacker by analyzing the text as described in section 4.2. We are interested in modeling persuasion arguments using neural networks and perform semantic analysis of the dialogue to predict persuasiveness. Based on Cialdini's model (authority, scarcity, liking & similarity, reciprocation, social proof and commitment & consistency) well-known binary classifiers (Naive Bayes, Support Vector Machines) are used which are effective in feature vector models. Feature vectors are populated with metric values for topic initiation, topic control, sentence structure and dialogue goal. Furthermore, two commonly used features in NLP, word unigrams and bigrams are used along with the implied Bag-of-Word model.

5.3 Deception Recognition

The system is able to calculate the degree of deception that is hidden in the attacker's writings according to the section 4.3. In our approach, deception detection is treated as a classification problem where lexical features are used to apply machine learning algorithms. There are many algorithms, like SVM, capable of handling large number of features. To extract the lexical features Linguistic Inquiry and Word Count (LIWC), Part-of-Speech (POS) and N-gram techniques are utilized. Discovering positive emotion words is a main objective of the Deception Recognition subsystem because a great proportion of these appears more frequently in deceptive speech than in truthful speech [13]. Similar measurements are performed using DAL, while N-gram is used in conjunction with LIWC to train the classifiers.

Zuckerman [40] argues that deception can be categorized in three categories, namely: emotional stress, cognitive effort, and attempted behavioral control. Emotion recognition is simultaneously performed in DR subsystem to detect emotional stress that is generally caused (fear, guilt, delight, etc.) while an attacker tries to deceive. A deceiver might feel fear that she will be caught, or she might feel guilty doing something wrong, or even she could feel delighted by fooling someone else.

5.4 Personality Recognition

Personality recognition is performed using classification tools that are utilizing the results of Mairesse [19] . The personality traits of the victim are used to calculate the related risk of being vulnerable to a SE attack. The main objective of the Personality Recognition subsystem is to identify the personality category (as defined in the *Big-5 theory*) of both interlocutors (attacker and victim) based on the captured dialog. To this extend, a document-modeling technique [20] is utilized, based on a convolutional neural network features extractor. Chat dialog sentences are fed to convolution filters to create a sentence model in the form of n-gram feature vectors. Each text-based dialog is then represented as aggregated vectors of its sentences. The vectors are created at the preprocessing stage based on Mairesse's features and then they are concatenated. All emotionally neutral sentences are discarded from the text-based dialog to further improve the results.

5.5 Speech Act

Identifying and tracking proposed actions and corresponding responses over communication channels (like text-based chat) is crucial for protecting users from SE attacks. These are difficult tasks due to syntactical, grammatical and structural idiosyncrasies of chat-based conversations. In our approach, every action is decomposed in different lexical units with accompanying parameters while every response can be in different states (acceptance, denial). Actions

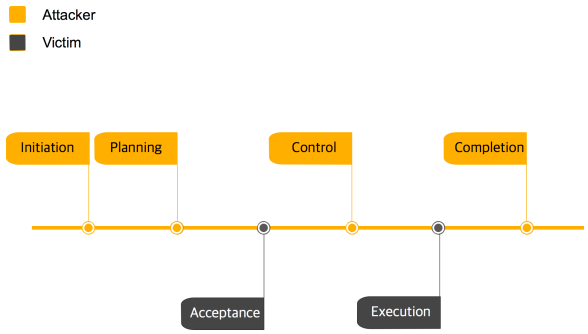

Figure 4: Timeline of chat offering a bait-project

and responses are then identified based on features extracted from the captured dialogues. We assume that every chat has a short lifecycle depending on the particular interlocutors (attacker, victim). A time-line depicting the aforementioned chat steps is shown in Figure 4. An attacker can initiate a chat (Initiation), and plan/offer a bait-project (Planning). If the victim's response is acceptance (Acceptance) then the attacker has taken control (Control) of the situation. After that, the victim executes the action (Execution), and the attack reaches completion (Completion).

Our focus is on human chat-based conversations from a perspective based on language as action. Therefore, the Speech Act subsystem defines actions and responses based on extracted features from text conversation (typed dependency) in the context of SE attacks, identifies the response state, determines achieved steps in the chat conversation timeline, and monitors the corresponding SE attack progression to raise an alert. Here, our main interest is in identifying a "conversation for action" in which the attacker (A) makes a request to the victim (B) either to do something or say something (e.g., reveal information). The state transition diagram in Fig. 5 is an adapted version of the one that Winograd [39] developed to represent a Conversation for Action (CfA) as a pattern of a Speech Act.

More specifically, the state transition diagram represents a CfA initiated by a request from an Attacker (A) to a Victim (V). The circles represent conversation states and the labeled lines represent speech acts. After the initial request of A, V can accept, decline, or counter-offer. A makes a request to V and V can promise to fulfill the request, reject it or counter-offer. V can accept the counter offer, counter again or withdraw. In case V promises to fulfill the request, he can later assert that the request is done. A can declare the request done, not done, or withdraw. To identify requests for action by the attacker and monitor the flow from state to state in a CfA, we utilize NLP techniques, Typed Dependencies Trees, and Named-Entity Recognition techniques (NER).

5.6 Past Experience

The Past Experience process analyses features from dialogues captured in a long period of time, along with accompanying previously stored risk values. History is expressed in number of dialogues rather than some time metric. Since many SE attacks last long and take place in several phases, it is beneficial to use this past history. The PE subsystem handles the following values: the risk values of all previous chats between the same interlocutors, together with the proportion of the same user's conversations.

The exclusivity of an attacker's conversations with a particular victim is calculated as a ratio. The importance of calculating this ratio results from the fact that most attackers form a deceptive relationship with their victims before the attack begins. Therefore, the elevated rate can signal a possible attack preparation. Specifically, whenever a chat conversation is detected, the nickname used by an attacker is also recorded together with his network connection details. Thus, considering the number of conversations recorded in the past, (where the attacker had the same nickname in which the victim participates), the ratio can be calculated.

Fig 6 shows the utilization of the SE attack recognition subsystems during the evolution of the SE attack stages. Past Experience is able to utilize content from historic data from every SE attack stage. For Personality Recognition and Deception Recognition data are gathered during the IG and RD stages. The Influence Recognition subsystem monitors the dialogues during the RD stage. Finally, during the EX stage data are gathered by the Speech Act subsystem.

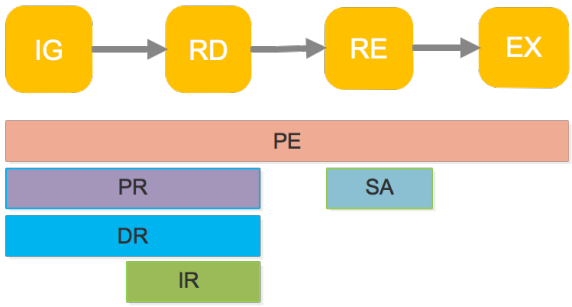

Figure 6: Utilization of SE attack Enablers during SE attack stages

6 CONCLUSIONS

In this paper, we demonstrated that SE attacks is a persistent cyber threat in enterprise environment and that detection is needed in early stages. A thorough review of related works is conducted which revealed the shortage of automated recognition systems for chat-based SE attacks. A dissection of the separate SE attack stages was presented along with the related SE attack attributes and the various forms of the attacks. The major enablers were identified for every stage, namely: personality traits, influence (persuasion), deception, speech act and past experience. Finally a system capable of recognizing chat-based SE attacks in early stages is proposed by combining the related corresponding indicators to the aforementioned SE attack enablers. The proposed system is required to comply with the European General Data Protection Regulation (GDPR) and other related international data protection regulations.

ACKNOWLEDGMENTS

This work has been partially supported by the European Commission through project FORTIKA funded by the European Union Horizon 2020 programme under Grant Agreement No. 740690 . The

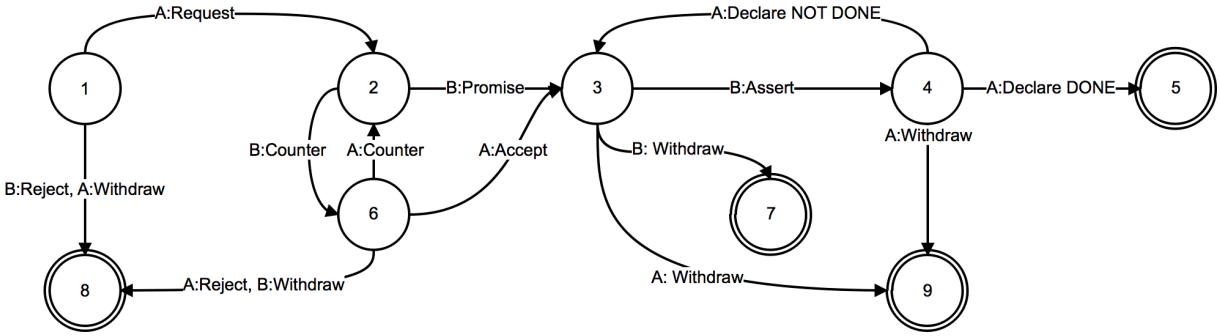

Figure 5: Conversation for Action state diagram adapted for SE Attack recognition.

opinions expressed in this paper are those of the authors and do not necessarily reflect the views of the European Commission.

REFERENCES

[1] 2018. Information Technology - Information Security - Information Assurance | ISACA. (2018). https://www.isaca.org/pages/default.aspx

[2] Guozhen An. 2015. Literature review for Deception detection. *Dr. Diss. City Univ. New York* (2015).

[3] Monique Bezuidenhout, Francois Mouton, and Hein S Venter. 2010. Social engineering attack detection model: Seadm. In *Information Security for South Africa (ISSA), 2010*. IEEE, 1–8.

[4] Ram Bhakta and Ian G Harris. 2015. Semantic analysis of dialogs to detect social engineering attacks. In *Proc. 2015 IEEE 9th Int. Conf. Semant. Comput. IEEE ICSC 2015*. IEEE, 424–427. https://doi.org/10.1109/ICOSC.2015.7050843

[5] Ali Darwish, Ahmed El Zarka, and Fadi Aloul. 2012. Towards Understanding Phishing Victims' Profile. In *2012 Int. Conf. Comput. Syst. Ind. Informatics*. IEEE, 13–17.

[6] Marie-Catherine De Marneffe, Bill MacCartney, Christopher D Manning, and Others. 2006. Generating typed dependency parses from phrase structure parses. In *Proceedings of LREC*, Vol. 6. Genoa Italy, 449–454.

[7] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers-Volume 2*. Association for Computational Linguistics, 171–175.

[8] David Gragg. 2003. A Multi-Level Defense Against Social Engineering Social. *SANS Inst.* (2003), 21. https://doi.org/10.9780/22307850

[9] Sarah Granger. 2001. Social Engineering Fundamentals, Part I: Hacker Tactics | Symantec Connect. *Secur. Focus. December* 1527 (2001). http://www.symantec.com/connect/articles/social-engineering-fundamentals-part-i-hacker-tactics

[10] Pär Anders Granhag and Leif A Strömwall. 2004. *The Detection of Deception in Forensic Contexts*. Vol. 9780521833. Cambridge University Press, Cambridge. 1–348 pages. https://doi.org/10.1017/CBO9780511490071 arXiv:arXiv:gr-qc/9809069v1

[11] Harl. 1997. Psychology of Social Engineering. (1997). http://barzha.cyberpunk.us/lib/cin/se10.html

[12] Ryan Heartfield and George Loukas. 2015. A Taxonomy of Attacks and a Survey of Defence Mechanisms for Semantic Social Engineering Attacks. *ACM Comput. Surv.* 48, 3 (2015), 1–39. https://doi.org/10.1145/2835375

[13] Julia Hirschberg, Stefan Benus, Jason M Brenier, Frank Enos, Sarah Friedman, Sarah Gilman, Cynthia Girand, Martin Graciarena, Andreas Kathol, Laura Michaelis, Bryan Pellom, Elizabeth Shriberg, and Andreas Stolcke. 2005. Distinguishing Deceptive from Non-Deceptive Speech. *Proc. Interspeech 2005* (2005), 1833–1836. https://doi.org/10.1.1.59.8634

[14] Michael D Hoeschele and Marcus K Rogers. 2004. CERIAS Tech Report 2005-19 DETECTING SOCIAL ENGINEERING. (2004).

[15] Michael D Hoeschele and Marcus K Rogers. 2006. CERIAS Tech Report 2006-15 DETECTING SOCIAL ENGINEERING by Michael Hoeschele Center for Education and Research in Information Assurance and Security , Purdue University , West Lafayette , IN 47907-2086. (2006).

[16] Lech J Janczewski and Lingyan Fu. 2010. *Social Engineering-Based Attacks: Model and New Sealand Perspective*. IEEE, Piscataway, NJ.

[17] A Karakasiliotis, S M Furnell, and M Papadaki. 2006. Assessing end-user awareness of social engineering and phishing. (2006), 4–5. https://doi.org/10.4225/75/57a80e47aa0cb

[18] Katharina Krombholz, Heidelinde Hobel, Markus Huber, and Edgar Weippl. 2014. Advanced Social Engineering Attacks. *J. Inf. Secur. Appl.* 22 (2014), 11.

[19] Francois Mairesse and Marilyn Walker. 2000. Words Mark the Nerds: Computational Models of Personality Recognition through Language. In *28th Annu. Conf. Cogn. Sci. Soc.*, Vol. 28. 543–548.

[20] Navonil Majumder, Soujanya Poria, Alexander Gelbukh, and Erik Cambria. 2017. Deep learning-based document modeling for personality detection from text. *IEEE Intelligent Systems* 32, 2 (2017), 74–79.

[21] Maranda McBride, Lemuria Carter, and Merrill Warkentin. 2012. Exploring the role of individual employee characteristics and personality on employee compliance with cybersecurity policies. *2011 Dewald Roode Work. Inf. Syst. Secur. Res.* (2012), 1–13.

[22] Kevin D Mitnick and William L Simon. 2011. *The art of deception: Controlling the human element of security*. John Wiley & Sons.

[23] Francois Mouton, Louise Leenen, and H S Venter. 2015. Social Engineering Attack Detection Model: SEADMv2. In *2015 Int. Conf. Cyberworlds*. IEEE, 216–223. https://doi.org/10.1109/CW.2015.52

[24] Marcus Nohlberg. 2008. *Securing Information Assets: Understanding, Measuring and Protecting against Social Engineering Attacks*. Ph.D. Dissertation. Department of Computer and Systems Sciences (together with KTH), Stockholm University, Kista.

[25] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T Hancock. 2011. Finding Deceptive Opinion Spam by Any Stretch of the Imagination. In *Proc. 49th Annu. Meet. Assoc. Comput. Linguist. Hum. Lang. Technol. 1*. Association for Computational Linguistics, 309–319. https://doi.org/10.1145/2567948.2577293 arXiv:1107.4557

[26] Thomas R Peltier. 2006. Social engineering: Concepts and solutions. *Inf. Syst. Secur.* 15, 5 (nov 2006), 13–21. https://doi.org/10.1201/1086.1065898X/46353.15.4.20060901/95427.3

[27] Alan J Resnik and Robert B Cialdini. 1986. Influence: Science & Practice. *J. Mark. Res.* 23, 3 (1986), 305. https://doi.org/10.2307/3151490 arXiv:arXiv:1011.1669v3

[28] Serge Vaudenay (eds.) Samuel Galice Marine Minier (auth.). 2008. *Progress in Cryptology - AFRICACRYPT 2008, First International Conference on Cryptology in Africa, Casablanca, Morocco, June 11-14, 2008. Proceedings* (1 ed.). Lecture Notes in Computer Science 5023 Security and Cryptology, Vol. 5023. Springer-Verlag Berlin Heidelberg.

[29] Yuki Sawa, Ram Bhakta, Ian G Harris, and Christopher Hadnagy. 2016. Detection of Social Engineering Attacks Through Natural Language Processing of Conversations. In *2016 IEEE Tenth Int. Conf. Semant. Comput.* IEEE, 262–265. https://doi.org/10.1109/ICSC.2016.95

[30] Jamison W Scheeres. 2008. *Establishing the Human Firewall: Reducing an Individual'S Vulnerability To Social Engineering Attacks*. Technical Report. DTIC Document. 49 pages.

[31] Bernadette H Schell and Clemens Martin. 2006. *Webster's New World Hacker Dictionary*. Wiley Pub, Indianapolis, IN. 387 pages.

[32] John R Searle, Ferenc Kiefer, Manfred Bierwisch, and Others. 1980. *Speech act theory and pragmatics*. Vol. 10. Springer.

[33] Charles Donald Spielberger. 2004. *Encyclopedia of applied psychology*. Elsevier Academic Press.

[34] Pekka Tetri and Jukka Vuorinen. 2013. Dissecting social engineering. *Behav. Inf. Technol.* 32, 10 (oct 2013), 1014–1023. https://doi.org/10.1080/0144929X.2013.763860

[35] David R Tobergte and Shirley Curtis. 2013. *INFLUENCE The Psychology of Persuasion*. Vol. 53. Harper Collins. 1689–1699 pages. https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[36] Sven Uebelacker and Susanne Quiel. 2014. The social engineering personality framework. In *Proc. - 4th Work. Socio-Technical Asp. Secur. Trust. STAST 2014 - Co-located with 27th IEEE Comput. Secur. Found. Symp. CSF 2014 Vienna Summer Log. 2014*. 24–30. https://doi.org/10.1109/STAST.2014.12

[37] Verizon. 2015. 2015 Data Breach Investigations Report. *Verizon Bus. J.* 1 (may 2015), 1–70. https://doi.org/10.1017/CBO9781107415324.004 arXiv:arXiv:1011.1669v3

[38] Aldert Vrij. 2014. Detecting lies and deceit: Pitfalls and opportunities in nonverbal and verbal lie detection. In *Interpers. Commun.* 321–346. https://doi.org/10.1515/9783110276794.321

[39] Terry Winograd. 1986. A language/action perspective on the design of cooperative work. In *Proceedings of the 1986 ACM conference on Computer-supported cooperative work*. ACM, 203–220.

[40] Miron Zuckerman, Bella M Depaulo, and Robert Rosenthal. 1981. Verbal and nonverbal communication of deception. *Adv. Exp. Soc. Psychol.* 14, C (1981), 1–59. https://doi.org/10.1016/S0065-2601(08)60369-X arXiv:2066187