Winter 12-13-2018

# Over-claiming as a Predictor of Insider Threat Activities in Individuals

Stan Zavoyskiy
*University at Albany*

Nicholas Rizzo
*University at Albany*

Sanjay Goel
*University at Albany*

Kevin Williams
*University at Albany*

# Over-claiming as a Predictor of Insider Threat Activities in Individuals

**Stan Zavoyskiy**
University at Albany
Albany, New York, USA

**Nicholas Rizzo**
University at Albany
Albany, New York, USA

**Sanjay Goel**[1]
University at Albany
Albany, New York, USA

**Kevin Williams**
University at Albany
Albany, New York, USA

## ABSTRACT

Insiders can engage in malicious activities against organizations such as data theft and sabotage. Prior research on insider threat behavior indicates that once motivated to commit malicious activity, insiders seek opportunity where they can act without being detected. In this research we set up an experiment where we leverage this opportunistic behavior and present participants with messages signaling opportunity for data theft. In the experiment, students were engaged in routine tasks with a bonus based on their performance. While working on their assigned tasks, they were presented with opportunities (probes) to steal data that would increase their payout. Their pre and post probe behavior was observed to test if they engaged in behavior that was deemed suspicious when they received the probe. The goal of the project is to test whether the overclaiming personality trait is a predictor of malicious insider behavior and this was measured through the Over Claiming questionnaire developed by Paulhaus (Paulhaus et al. 2003) The results indicated that over claiming proved to be a strong predictor of malicious insider behavior.

**Keywords:** Over-Claiming, Behavioral Security, Cybersecurity, Insider Threat Detection

[1] Corresponding author. goel@albany.edu +1 518-956-8323

## INTRODUCTION

Insider threats continue to comprise a large proportion of cyber-security breaches in organizations and constitute approximately half of all such incidents (Richardson 2008). Healthcare, education, and government agencies all have proven particularly susceptible to the problem posed by insider threats (Verizon 2018). Many experts believe that cybersecurity incidents caused by insiders are more damaging than external breaches (RSA 2016). Excessive user privilege, the litany of devices that have access to sensitive information, and the increasing complexity of networks are examples of the challenges that leave 90% of businesses feeling vulnerable to insider threats. (C.A Technologies 2018).

There is a strong impetus to identify insider threats prior to their manifestation in the organization. It is imperative that we can understand the motivations, proclivities, and dispositional characteristics predictive of human behavior that triggers malicious behavior of insiders. Resting upon the axiom that humans tend not to act randomly, we can come to understand the psychological underpinnings of the malicious insider's actions, and furthermore look to assess the factors that trigger the need for stealing data. Insider threat can be examined through what is referred to as the insider threat kill chain, this begins with radicalization of the worker (a trigger point) wherein he becomes motivated to commit a malicious act. This is followed by searching for opportunities to engage in malicious insider behavior, and finally leads to the malicious act (e.g, exfiltrating data) when the opportunity arises (Colwill 2009).

We can mitigate the damage from insider attacks by intervening in the process at any point in the insider threat kill chain prior to data exfiltration. Recently, research was conducted that manipulated whether users received bogus messages indicating that they were performing below average, followed by a probing message that informed participants as to the status of the

network, as relevant to cyber-security. It was revealed that when looking to data obtained prior to the reconnaissance stage where subjects were seeking opportunities to exfiltrate data, researchers were able to predict, with reasonable accuracy, the malicious from the benevolent insider (Goel et al. 2016; Goel et al. 2017). In this experiment, honeytokens (fake, baited opportunities) were presented to the subjects with the contention that malicious subjects, seeking opportunity, would respond differently to the threats compared to benign subjects. This work leverages the experimental setup of this previous work to look at "overclaiming", which simply put is a measure of exaggeration in identifying one's knowledge, as a predictor of malicious insider behavior.

A confluence of factors motivates an individual to engage in conscious insider theft. The factors involved are both dispositional (personality) and situational (circumstances), and it is important to note that the *interaction* between the two is generally of greatest interest as it is often the case that circumstance serves as the catalyst for the realization of particular, malicious actions. The dispositional factors that we have previously investigated included the Big Five personality characteristics (McCrae and Costa 1999), the Dark Triad (Paulhus and Williams 2002; measures subclinical: Narcissism, Psychopathy, and Machiavellianism), self-esteem (Rosenberg 1965), general self-efficacy (Gardner and Pierce, 1998, technological proficiency. The current research attempts to understand the interactions in individuals who over-claim their knowledge on a subject and their proclivity to steal data.

In the present research, we have designed a task where participants were assigned with researching and finding notable information on hacking groups, with a monetary incentive provided for participation as well as the prospect of additional monetary compensation for excellent performance. Unbeknownst to the subject, we provided opportunities to steal data that

allowed one to improve productivity during the experiment and measured their behavior via sensors installed on the computer. We then examined the role of personality, including data from the personality inventories collected from subjects prior to the task and included these as variables in a series of statistical analyses that attempted to elucidate predictive factors of the malicious insider threat. The rest of the paper is described as follows. Section 2 provides a brief review of the literature followed by a detailed description of personality as well as the impact of over-claiming in the context of insider threats in Section 3. Section 4 provides the details of the experiments and section 5 the results of the experiments. Section 6 provides a brief conclusion and plans for prospective future research.

## LITERATURE REVIEW

Since individuals have begun to attack computer systems, researchers have attempted to build profiles to understand the underlying psychological characteristics of hackers. Landreth and Rheingold (1985) first proposed the idea of classifying hackers by skill level and building a psychological profile around a hacker's abilities. This idea was recently notably revisited by Kandias et al. (2010), who proposed using user's technical sophistication and their role on a computer system or network as elements in a prediction model for insider threats. Additionally, Kandias et al. (2010) combined these variables with self-reported responses to the Computer Crime and Social Learning Questionnaire (Rogers 2001) to determine the likelihood a stressful event could result in an individual acting maliciously. The major limitation of this approach is that the responses are likely to be affected by self-report bias, as insiders are unlikely to comfortably admit in an anonymous survey that they are acting against the organization that has placed trust in them.

The situational factors involved in scenarios where individuals make the transition from trusted insider to insider threat are well reviewed in the current literature (e.g., Shaw 2006). Shaw (2006) produced a major review of the behavioral characteristics exhibited by insider threats. The major findings presented in this review indicated that 81% of insiders planned their attacks in advance of carrying them out, and 85% of insiders told a third party of their plans to commit a malicious action. Shaw (2006) also reported that individuals who become insiders are likely to be experiencing behavioral and emotional issues. It was also reported that malicious insiders were especially likely to have a strong negative affect towards their workplace. Shaw (2006) also analyzed the email correspondence of individuals who exhibited insider threat behavior and discovered that insiders displayed superior intelligence, high rigidity, arrogance, and generally speaking, the characteristics of a social "loner".

Beyond the work of Shaw (2006), Greitzer and Hohimer (2011) proposed using behavioral modeling to predict if individuals were likely to become insiders, albeit in a different fashion. They first proposed using linguistic cues to detect insiders, through the terms they use to respond to cues presented (e.g., a potential insider would respond to _ight with fight, non-insiders would respond with tight). Additionally, the authors propose applying the well-known Stroop Task for the detection of insiders. In this task, words are presented in a fashion that inhibits reaction time by the respondent (Stroop 1935). The Emotional Stroop Task (McKenna and Sharma 1995) provides fear-eliciting stimuli to subjects, which much like the original Stroop Task, inhibit response times for those items which induce fear, as would be the case when an individual who is worried about revealing their insider tendencies is presented a word relevant to their malicious activities (e.g., "steal") would be inhibited from responding (Richards et al. 1992).

Furthermore, there has been an examination of the role of dispositional and situational factors in promoting compliant behavior. Johnston et al. (2016) used a scenario-based survey approach to test how the dispositional and situational factors impact an employee's security policy violation intentions when aggregated (Johnston et al. 2016). The authors identify two meta-traits (i.e., Stability and Plasticity) that serve as moderators of the relationships between perceptions derived from situational factors and intentions to violate information security policy (DeYoung 2006). The researchers concluded that the differences between individuals who exhibit the stability meta-trait versus the plasticity meta-trait precludes the use of standardized interventions to prevent insider activity (Johnston et al. 2016).

Despite the modicum of ingenuity in attempting to profile and predict malicious insider threats, there is a clear paucity in well-validated methods that combine digital threat detection with psychological profiling to predict malicious insider behavior. Despite a handful of experiments that have looked at traditional personality traits (i.e., The Big Five), nuance and sophistication have been limited in such approaches, and as such, there is a distinct lack of efficacy in proactively detecting insider threat. The present research attempts to more effectively use psychological variables in hopes of ultimately furthering the sophistication of methods in the early detection of the malicious insider. To the best of our knowledge this is the first examination of the interaction between over-claiming as a personality characteristic investigating the malicious insider threat.

## OVERCLAIMING

Researchers have often struggled with self-report measures, particularly because "self-enhancement" is considered by the psychological community as one of the three main needs for the self (Baumeister 1982). What this means is that individuals naturally look to self-present in a

fashion that depicts them more favorably to others. The search for an effective means to counter this problem has largely been disappointing (Paulhus 1991). The most notable and prevalent attempts to counter this problem have been the use of scales to determine to what extent individuals respond in a way that is deemed favorable based on the group norm, intrapsychic measures where the extent to which individuals rate themselves as better than average is assessed, and credible criterion discrepancy measures (Paulhus et al. 2003), which essentially attempt to identify forms of exaggeration, or self-enhancement, that is wholly unwarranted given a specific context. Largely, these and similar scales have not reached the level of empirical rigor that is deemed necessary to mollify the problem at hand.

In response to the weaknesses of such prior attempts to mollify this issue, The Over-Claiming Questionnaire (OCQ; Paulhus et al. 2003) was implemented as a measure of "faking" behavior, in that it presents participants with a set of items from different domains (social sciences, history, arts, etc.), 20% of which are bogus. It should also be noted that in our iteration of the over-claiming scale, we adopted items (from a repository made available online by Paulhus and colleagues) relevant to modern technology and computers and created our own category of cyber-security items, including a set of original foils that represent no existing cyber-security construct. In completing the OCQ, participants rated the familiarity of items presented to them on a 7-point scale anchored at 0 (*never heard of it)* and 6 *(know it very well).* The extent to which participants rate familiarity with items that simply do not exist indicates their bias for over-claiming. Additionally, bias on the OCQ can be used as a covariate in statistical analyses by accounting for exaggeration in socially desirable responding, and thus bolstering the accuracy of standard self-reported measures by statistically controlling for faking behavior, as we have done in this research.

However, the OCQ also is inherently a measure of faking, or rather exaggeration. Such behavior has been referred to in various ways including but not limited to positive self-presentation, and self-deceptive enhancement (Paulhus 1984; Paulhus and Reid 1991): two different motivated actions that at face value seem rather similar. It is to be noted for the purposes of the present research, that in low demand conditions, where there is little prospect of being unfavorably judged (e.g., completing the OCQ knowing that you will not, and cannot be identified), high levels of bias in over-claiming has been referred to as *narcissistic self-enhancement* (Paulhus 1998), as individuals exaggerate their self-perception without much reward. In this sense, the OCQ is seen in the present experiment as a low-demand scenario, and as such serves as a robust methodological analog to deviance relevant to various forms of deception, and specific to the present experiment, the exfiltration of other's data.

## EXPERIMENTAL DESIGN

In this experiment undergraduate students studying computer science and cyber-security were recruited as participants and were given a monetary incentive to participate in a hacker-research task which had participants identify particular notable hacking groups, their members, notable characteristics, known attacks, etc. Participants were given forty minutes to research as many hacker groups as they could. Subjects were told to save their results in a publicly accessible shared directory so that their results could be evaluated by the experimenters. Participants were told that they would be compensated based on their performance on the task, with additional rewards for outstanding performance. The nature of the monetary compensation made it so there was an incentive to cheat via the copying of another subject's results from the shared drive. In addition to the pressure that resulted from the design, a message was sent to half of the subjects at the fifteen-minute mark that fostered a sense of urgency, by asking them how many groups

they had completed their research on, and then informing half of the subjects them that they were performing below the average subject in terms of the number of groups adequately researched. Subsequently, all of the subjects were sent a probe message that indicated an opportunity for data theft. By sending messages to participants, we manipulated the perceived affordances and level of risk associated with data exfiltration. A computer laboratory was set up for the experiments where the computers were equipped with Keylogging software (Veriato360 and NetVisor). Data theft behavior was measured through the use of this keylogging software such that keystrokes were recorded and then analyzed to detect file and directory access (indicating data theft). Three primary forms of action were operationalized and analyzed based on data collected via user activity monitoring software:

[1] Curiosity: The opening of other participant's folders on the shared drive
[2] File Opening: Opening of other participant's files in the shared drive
[3] Theft: Copying another participant's work from files that were improperly accessed on the shared drive

The specific task for participants was to conduct research on hacker groups and catalog their organizational structure and activities in a file located on a shared drive on the network; each subject was assigned a separate file to post their results. Additionally, a project directory was created with a pre-filled catalog of hacker groups giving subjects an ability to copy from both other users and the project directory. Five different opportunities were presented to the users to exfiltrate data (we call these probes) as shown below:

**Table 1.** List of Probes

| Number | Probe |
|---|---|
| 1 | Alert to System Administrator: Microsoft Group Policy Editor has been corrupted, all security controls have been disabled |
| 2 | Due to group policy misalignment some participants are unable to access their folders. So, we have changed the configuration so that all participants can access the entire shared drive |

| | |
|---|---|
| 3 | Please be informed that for the last 5 minutes of the task we will activate software to monitor your progress and inform you if you adequately cover all of the key elements in your report. |
| 4 | USER NOTICE: File permission error detected (0x0011532A). Patch applied (v 1.2.9_R1). Security update in: 10:00 mins. Save all files prior to install to prevent data loss. |
| 5 | Please note that at the end of the experiment we will run all reports through SafeAssign, a software program that checks for plagiarism. |

The behavior was correlated with the personality of the users and the results are discussed below.

## RESULTS

Given that the dependent variables in the present research were categorical and binary, in that they either represented a "yes" or a "no" in regard to an action of interest (e.g., "Did they open other individual's folders?"), we conducted a series of Binary Logistic Regressions. Results demonstrated that for pre-probe curiosity, operationalized as opening a "taboo" folder and keeping it open (prior to the presentation of the probe), there was no significant effect of any other personality variable other than over-claiming. Greater over-claiming bias was associated with increased curiosity behavior in line with predictions that those who inflate their true knowledge would be more likely to engage in at least, "pre-malicious" behavior.

Whether or not users opened files belonging to others was the primary dependent variable of interest, particularly given that operationalizing whether an individual stole data and used it to bolster their own performance proved to be difficult. Exploratory analyses revealed a high rate of error and a greatly wanting level of reliability for this dependent measure. This was a result of the fact that many subjects did not copy the data from their peers' work as originally expected, but rather summarized the work product of many of their counterparts after reviewing a great number of documents.

There was an interaction found between probe and self-evaluation, $p < .05$, as individuals were more likely to access other's file post-probe for the messages that increased activity when they

were also informed that their performance was below average. In analyzing the role of moderating or mediating personality variables, the sample size that we acquired ($N$=76) left us with a very low observed power; particularly so for three-way interactions. This limitation was a function of the number of participants we were able to recruit within the allotted time-frame available to us. It should also be noted that given the small sample, results should be treated with caution as further research is necessary to firmly establish the potential role of a moderating or mediating personality variables.

Additionally, the results demonstrated that greater over-claiming led to increased odds of opening another's files, $p < .05$. This is in line with our predictions in that greater "faking" behavior as measured by the OCQ was believed to be positively related to a higher incidence of deviant behavior. Further, technological proficiency significantly predicted the greater likelihood of opening another's files, $p < .05$. While it may be tempting to claim that this provides evidence that those with higher technological proficiency differ significantly in regard to their drive to cheat, a more probable explanation is that these individuals are the most likely to have the technical knowledge to more easily exfiltrate data, as compared to less technologically literate individuals. Also notable was that higher General Self-Esteem predicted a lesser likelihood of opening another's files. In line with the prevalent view of self-esteem, which states that individuals with high self-esteem tend to be significantly less likely to engage in deviant or self-deprecatory actions (e.g., Rosenberg et al. 1989) whether it be physical or social, given that their needs for esteem are met. (Baumeister et al. 1993). General self-efficacy, or the belief in one's ability to realize a particular goal, conversely was significant in predicting a greater likelihood of opening another's files, $p < .05$. This finding while counter to our original hypotheses, actually seems compelling in that individuals with high self-efficacy, who feel they have a greater

capacity to achieve personal goals may be more likely to perceive less threat and act more boldly in his context, conferring in a higher incidence in malicious insider behavior.

The ecological validity in this particular case may be questionable, as, in other contexts (other than within a college classroom) that are typically less punitive in terms for stringency and consequences of cheating, one would expect more cheating behavior than average. Thus, this effect could flip as a function of the demands of the unique social environment. The present research would be considered far closer to the former, as cheating behavior within a university context is notoriously associated with highly undesirable consequences, such as expulsion. This limitation may conversely also introduce an area of future investigation regarding how differences in perceived risk, norms, and other contextual variables associated with how deviant behavior might be differentially perceived may affect insider actions.

## CONCLUSIONS

In this research, we demonstrated that over-claiming is robust both in helping to mitigate error stemming from biases inherent to self-report data, particularly in contexts where the social desirability of actions may be particularly relevant. It is our hope that future research will look to adopt more complex and nuanced methodologies in psychologically profiling malicious insiders. Whereas machine-learning algorithms, Bayesian threat detection systems, and other digital tools that attempt to thwart cyber-attacks do demonstrate a level of efficacy in deterring insider threats, it is largely *post hoc*. We should note that malicious cyber threats from within an organization, is typically related to an insider with certain predispositions, motivations, varied perceptions of their environment, and that such actors rarely act without some level of deliberation, logic, and the assessment of the likely ramifications of their actions. Carefully,

crafted experiments that provide opportunities can reveal malicious intentions of employees prior to their malicious actions.

## References

Baumeister, R. F. 1982. "A Self-Presentational View of Social Phenomena," *Psychological Bulletin* (91:1), pp. 3–26. (https://doi.org/10.1037/0033-2909.91.1.3).

Baumeister, R. F., Heatherton, T. F., and Tice, D. M. 1993. "Baumeister, Heatherton, Tice - 1993 - When Ego Threats Lead to Self-Regulation Failure Negative Consequences of High Self-Esteem.Pdf," *Journal of Personality and Social Psychology* (64:1), pp. 141–156. (https://doi.org/10.1177/0146167206289408).

C.A Technologies. 2018. "Insider Threat: 2018 Report." (https://www.ca.com/content/dam/ca/us/files/ebook/insider-threat-report.pdf).

Colwill, C. 2009. "Human Factors in Information Security: The Insider Threat - Who Can You Trust These Days?," *Information Security Technical Report* (14:4), pp. 186–196. (https://doi.org/10.1016/j.istr.2010.04.004).

DeYoung, C. G. 2006. "Higher-Order Factors of the Big Five in a Multi-Informant Sample.," *Journal of Personality and Social Psychology* (91:6), pp. 1138–1151. (https://doi.org/10.1037/0022-3514.91.6.1138).

Goel, S., Williams, K., Zavoyskiy, S., and Rizzo, N. 2017. "Detecting Insider Threats Using Active Indicators Using Active Probes to Detect Insiders Before They Steal Data," in *AMCIS 2017 Proceedings*. (http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1045&context=amcis2017).

Goel, S., Williams, K., Zavoyskiy, S., and Williams, K. 2016. "Stopping Insiders before They Attack : Understanding Motivations and Drivers," *WISP 2016 Proceedings*, pp. 1–15. (http://aisel.aisnet.org/wisp2016/2).

Greitzer, F. L., and Hohimer, R. E. 2011. "Modeling Human Behavior to Anticipate Insider Attacks," *Journal of Strategic Security* (4:2), pp. 25–48. (https://doi.org/10.5038/1944-0472.4.2.2).

Johnston, A. C., Warkentin, M., McBride, M., and Carter, L. 2016. "Dispositional and Situational Factors: Influences on Information Security Policy Violations," *European Journal of Information Systems* (25:3), pp. 231–251. (https://doi.org/10.1057/ejis.2015.15).

Kandias, M., Mylonas, A., Virvilis, N., Theoharidou, M., and Gritzalis, D. 2010. "An Insider Threat Prediction Model," in *TrustBus - International Conferenceon Trust, Privacy and Security in Digital Business*, pp. 26–37. (https://doi.org/10.1007/978-3-642-15152-1_3).

Landreth, B., and Rheingold, H. 1985. *Out of the Inner Circle*, (First.), Microsoft Press.

McKenna, F. P., and Sharma, D. 1995. "Intrusive Cognitions: An Investigation of the Emotional Stroop Task," *Journal of Experimental Psychology: Learning, Memory, and Cognition*. (https://doi.org/10.1037/0278-7393.21.6.1595).

Paulhus, D. L. 1984. "Two-Component Models of Socially Desirable Responding," *Journal of Personality and Social Psychology* (46:3), pp. 598–609. (https://doi.org/10.1037/0022-3514.46.3.598).

Paulhus, D. L. 1991. "Measurement and Control of Response Bias," in *Measures of Personality and Social Psychological Attitudes*, pp. 17–59. (https://doi.org/10.1016/B978-0-12-590241-0.50006-X).

Paulhus, D. L. 1998. "Interpersonal and Intrapsychic Adaptiveness of Trait Self-Enhancement: A Mixed Blessing?," *Journal of Personality and Social Psychology* (74:5), pp. 1197–1208. (https://doi.org/10.1037/0022-3514.74.5.1197).

Paulhus, D. L., Harms, P. D., Bruce, M. N., and Lysy, D. C. 2003. "The Over-Claiming Technique: Measuring Self-Enhancement Independent of Ability," *Journal of Personality and Social Psychology* (84:4), pp. 890–904. (https://doi.org/10.1037/0022-3514.84.4.890).

Paulhus, D. L., and Reid, D. B. 1991. "Enhancement and Denial in Socially Desirable Responding.," *Journal of Personality and Social Psychology* (60:2), pp. 307–317. (https://doi.org/10.1037/0022-3514.60.2.307).

Richards, A., French, C. C., Johnson, W., Naparstek, J., and Williams, J. 1992. "Effects of Mood Manipulation and Anxiety on Performance of an Emotional Stroop Task," *British Journal of Psychology* (83:4), pp. 479–491. (https://doi.org/10.1111/j.2044-8295.1992.tb02454.x).

Richardson, R. 2008. "CSI Computer Crime and Security Survey," *Computer Security Institute* (Vol. 1). (https://doi.org/10.1007/978-3-319-04307-4).

Rogers, M. K. 2001. "A Social Learning Theory and Moral Disengagement Analysis of Criminal Computer Behavior: An Exploratory Study."

Rosenberg, M., Schooler, C., and Schoenbach, C. 1989. "Self-Esteem and Adolescent Problems: Modeling Reciprocal Effects," *American Sociological Review*. (https://doi.org/10.2307/2095720).

RSA. 2016. "2016: Current State of Cybercrime." (https://www.rsa.com/content/dam/premium/en/white-paper/2016-current-state-of-cybercrime.pdf).

Shaw, E. D. 2006. "The Role of Behavioral Research and Profiling in Malicious Cyber Insider Investigations," *Digital Investigation* (3:1), pp. 20–31. (https://doi.org/10.1016/j.diin.2006.01.006).

Stroop, J. R. 1935. "Studies of Interference in Serial Verbal Reactions," *Journal of Experimental Psychology* (18:6), pp. 643–662. (https://doi.org/10.1037/h0054651).

Verizon. 2018. "2018 Data Breach Investigations Report (DBIR)," *Verizon Business Journal*. (http://bfy.tw/HJvH).