View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Lancaster E-Prints

Detecting Insider Threats Through Language Change

This article may not exactly replicate the final version published in the APA journal. It is not the copy of record. Visit: http://psycnet.apa.org/journals/lhb/37/4/267/

Abstract

The act of conducting an insider attack carries with it cognitive and social challenges that may affect an offender's day-to-day work behavior. We test this hypothesis by examining the language used in emails that were sent as part of a 6-hour workplace simulation. The simulation involved participants (N=54) examining databases and exchanging information as part of a four-stage organized crime investigation. After the first stage, 25% of the participants were covertly incentivized to act as an 'insider' by providing information to a provocateur. Analysis of the language used in participants' emails found that insiders became more self-focused, showed greater negative affect, and showed more cognitive processing compared to their co-workers. At the interpersonal level, insiders showed significantly more deterioration in the degree to which their language mimicked other team members over time. Our findings demonstrate how language may provide an indirect way of identifying employees who are undertaking an insider attack.

Detecting Insider Threats Through Language Change

In 2001, Robert Hanssen, a FBI Supervisory Special Agent with 25 years' service, pleaded guilty to spying for the Soviet and Russian intelligence service for twenty-two years. During this period, in exchange for money, Hanssen provided his handlers with significant information about US counter-intelligence operations, including the identity of three Russian agents who were secretly giving information to US authorities (Wise, 2003). Often described as the worst intelligence disaster in US history, Hanssen's case demonstrates the significant cost that can be incurred when an employee misuses his or her legitimate access and knowledge of an organization (Schultz, 2002). It makes clear the importance of developing methods for identifying insider behavior so that it may be investigated effectively (Maybury et al., 2006).

The majority of studies of 'insiders' have sought to identify ways of detecting suspicious behaviors that relate to the attack (Christoph et al., 1995; Spitzner, 2003). For example, in their large-scale company simulation, Maloof and Stephens (2007) showed that intellectual property theft may be detected by comparing an insider's document, printing and Internet use with the pattern of use expected for their organizational role. Of the employees whose behavior differed from the norm, 84% were carrying out insider attacks (Stephens & Maloof, 2009). Another, more person-centered approach seeks to encourage workers to 'whistle-blow' on suspicious co-worker behavior by promoting a collective ownership of security within the organization (Colwill, 2009; Williams, 2008). The success of this approach comes not from identifying attack behaviors directly, but from the possibility that co-workers are in a position to notice the changes in work behavior that accompany an insider's efforts to hide his or her duplicity.

In this paper we build on the idea that insiders may show suspicious behavior in their work behavior by examining the extent to which insiders change their language use when conducting an attack. A large body of research shows that language use may reflect a person's affective and cognitive state (Pennebaker, 2012), and that joint language use can be predictive of the degree of cooperation in dyadic and group interactions (Gonzales, Hancock, & Pennebaker, 2010; Taylor & Thomas, 2008). This suggests that language may provide a mechanism through which to detect insiders. Specifically, it suggests that differences in language use may reflect the changes in affective, cognitive and social processes that insiders would be predicted to show when planning and committing an attack. We explore this idea by examining the language used in emails sent as part of an immersive workplace simulation.

Language of an Insider

In reviewing the nature of insider attacks, a number of authors have highlighted being distant from co-workers and disinterested in work as markers of malicious intent (Colwell, 2009; Shaw, Fischer, & Rose, 2009). As insiders become immersed in their activity, they can become more focused on personal success over the collective goals of the workgroup and less concerned with maintaining positive relationships with their co-workers. Consistent with this idea are several findings relating to the motivations of insiders. For example, in their analysis of espionage cases in the 1990s, Herbig and Wiskoff (2002) found that divided loyalty was the most common reason for insider activity. Similarly, Stanton, Stam, Guzman, and Caldera (2003) demonstrated a significant negative relationship between self-reported organizational commitment and low-level information security breaches.

A number of language variables have been shown to relate to this kind of self-focus and social distancing (Cohen, Mehl, & Pennebaker, 2004). Of these, the most studied is personal pronoun use (Pennebaker, 2012). Studies report elevated first-person pronoun use (e.g., 'I', 'me') among individuals with high self-focus, such as suicidal poets (Stirman & Pennebaker, 2001), depressed students (Rude, Gortner, & Pennebaker, 2004), and people going through personal crises (Pennebaker & Lay, 2002). Similarly, use of second person pronouns may reflect a speaker who wishes to distinguish self from the immediate recipient of his or her message (Pennebaker & Lay, 2002). By contrast, the use of first-person plural pronouns, or 'we' words, is negatively related to distancing and positively related to having a strong sense of community. For example, in the first two weeks after September 11, 2001, people showed heightened use of plural pronouns and other socially related words as communities came together to deal with the crisis (Cohn, Mehl, & Pennebaker, 2004). Collectively this evidence suggests that pronouns may provide a method for detecting self-focus and social distancing in the work behavior of insiders. Thus, we predict that, in comparison to co-workers, insiders will show more use of first-person singular and second person pronouns, and less use of first-person plural pronouns (Hypothesis 1).

A second behavior that is highlighted as a cue to insider activity stems from evidence that the motivation for insider attacks is often frustration with the organization and its failure to acknowledge a person's status and accomplishments (Keeney et al., 2005; Kowalski, Cappelli, & Moore, 2008). According to a survey of professionals, this 'disgruntlement' can be exhibited through a number of work behaviors, including anger outbursts, confrontational behavior, and general negative affect (Greitzer, Noonan,

Kangas, & Dalton, 2010). This is consistent with Randazzo, Cappelli, Keeney, Moore, and Kowalski's (2004) finding that increased outbursts at co-workers and refusals to work with supervisors are common behaviors in the work history of past insiders. Moreover, it links with organizational research showing that strong negative affect is a precursor to counterproductive work behaviors (e.g., absenteeism, theft) that have parallels with insider activity (Hollinger & Clark, 1983; Workman & Gathegi, 2007). Consistent with this evidence, we predict that insiders will show heightened negative affect and expression of feelings compared to co-workers (Hypothesis 2).

A final difference in behavior that may be expected of an insider stems from the cognitive load that is likely to be associated with undertaking an insider attack (cf. Spence et al., 2004; Vrij et al., 2008). Attacks may well be more cognitively demanding than regular work behavior, for much the same reasons that lies can be more demanding that truths (Vrij et al., 2008). First, insiders are less likely than co-workers to take their credibility for granted and may, as a result, make an additional effort to appear open and trustworthy around co-workers. Second, insiders may feel the need to be less categorical about the tasks at hand so as to avoid saying something that raises the suspicions of co-workers, and to adapt their behavior if they are getting close to being caught. Third, insiders must manage the 'two worlds' of their regular work tasks and their insider attack, and they may become pre-occupied over ensuring that there is no slippage between the two. All of these tasks require a mental effort and so would be expected to be associated with more complex language and less cognitive clarity (Walczyk et al., 2005). Such changes in language have been shown in other domains, for example, when an individual

is responding to a series of personal and professional upheavals within a short space of time (Pennebaker & Lay, 2002)

A number of studies suggest that there are linguistic correlates of cognitive processing. For example, in the Reality Monitoring approach, coders consider cognitive operations such as expression of thoughts or reasoning, and suppositions of a series of experiences, as indicators of the construction and encoding of imagined rather than experienced events (Masip et al., 2005). Similarly, in their study of the linguistic markers of deception, ten Brinke and Porter (2011) found that liars used more tentative words than truth-tellers probably because they sought to avoid committing to a concrete version of their story. This link between language use and cognitive processing is also evident outside of deception research. For example, Junghaenel, Smyth, and Santner (2008) found that psychiatric patients asked to tell a story made less use of words indicating discrepancies and tentativeness compared to a community group. This difference, they argue, is consistent with the fact that many psychiatric patients are less able (or less willing) to think through alternative scenarios and viewpoints. In line with this evidence, we predicted that insiders would use more words relating to cognitive processes than co-workers (Hypothesis 3).

Interacting with Others

As well as providing clues to individuals' thoughts and feelings, language also plays a role at the interpersonal level. The mimicking of language across individuals is fundamental to social behavior and positively associated with cooperation (Garrod & Pickering, 2004). For example, Communication Accommodation Theory (Coupland & Giles, 1998) explores how in many settings speakers adjust the content and timing of

their speech to emphasize or minimize the social distance between themselves and the other person (Cappella & Panalp, 1981). According to this theory, dialogue between speakers whose relationship is positive and cohesive is characterized by high levels of matching in turn-taking, paralinguistic and linguistic cues. A wealth of research supports the idea that social synchrony can be used as a metric of positively functioning social dynamics (Tickle-Degan & Rosenthal, 1990). What is remarkable about this dynamic is that it appears to occur, or is at least be reliably measured, not at the level of absolute word matching but at the level of matching in function words (e.g., pronouns, articles). This kind of language mimicry has been shown to relate to better group problem-solving (Gonzales et al., 2010), stability in romantic relationships (Ireland, Slatcher, Eastwick, Scissors, Finkel, & Pennebaker, 2011), and even success in police hostage negotiations (Taylor & Thomas, 2008).

Given the link between language mimicry and positive social dynamics, it seems reasonable to predict that insiders will show less mimicry with co-workers than that typically found amongst co-workers. That is, because insiders are focused on their own goals and will not wish to become too involved with others, so they will show less natural coordination and accommodation in their language use as they communicate with their 'out-group' co-workers (Hypothesis 4).

Method

Participants and Procedure

Fifty-four final year undergraduate psychology students were paid £50 to take part in a 6-hour workplace simulation known as Confidential Operations Simulation (iCOS). iCOS is designed to simulate the investigative tasks and organizational environment of a

police investigation into organized crime. The simulation involves three teams—the Fraud, Human Trafficking, and Narcotics teams—of four investigators working together to complete four stages of an increasingly complex investigation. For each of the four stages, participants are required to assimilate information from various databases that are available at their computer terminal, and make inferences from this information about which suspects to arrest (i.e., the arrest order), and where such arrests should be carried out (i.e., the arrest locations). The tasks require collaboration because no single participant has access to all of the databases from his or her computer (and, thus, no one person has access to all of the available information). The solution in each round requires participants to exchange information systematically, recognize the 'connections' across databases, and engage in collaborative problem solving. Each round lasts approximately one hour.

The iCOS simulation was carried out as though it were a 'working day,' with the four stages punctuated with regular 15 minute breaks and a lunch hour (both of which are valuable to an insider's efforts to complete his or her tasking; see below). At the beginning of the day, participants were randomly assigned to a role within one of the three teams (either Administrator, Field agent, Intelligence analyst, or Tactical investigator). They were then given instructions about the kind of investigative tasks they were to complete, provided training on how to use the investigative databases, and given time to familiarize themselves with the databases and the office-like surroundings. In order to simulate a secure working environment and efficiently capture language use, participants were not provided with paper and pens and worked in 'silent' offices. They were encouraged to make notes using common desktop publishing and spreadsheet

software, and to exchange information using email. They also had access to a printer in a separate room, should they wish to produce a hardcopy of their notes (the printer again providing insiders with an opportunity to access others' material by reading their printouts prior to them being collected).

Once participants were familiar with the iCOS environment, one member of each team received instructions about the first crime to be investigated. After these initial instructions, all further interaction with participants was conducted via email with 'Gold Command' (a confederate). As well as issuing instructions for subsequent tasks, Gold Command ensured the simulation ran smoothly in a number of other ways. These included responding to participants' requests for permission to access other databases (both co-workers and insiders made such requests, with insiders using it to complete their task), sending periodic emails to ensure teams were on task, and providing help to participants who were having technical difficulties. Gold Command was also the person to whom teams reported their decision regarding the arrest order and arrest locations for each period. By embedding task instructions into the simulation, we hoped to enhance participants' immersion in the simulation (Druckman, 2005).

The first of the four periods was used as a familiarization phase that allowed participants to acquaint themselves with the gameplay and develop working relationships with one another. At the beginning of the second period, after the teams had been assigned their investigative task, up to two players ($M = 1.44$, $n = 14$) were approached covertly and asked to provide information to a provocateur for an additional reward of £20. Specifically, in this period, the participants were asked to obtain a piece of information relating to one of the individuals under investigation, and to email this

information to a specified email address. The approach occurred face-to-face and out of sight of the other participants (e.g., in the corridor) in order to replicate the 'real-world' necessity to not leave a digital trace of the approach. These 'insiders' were selected randomly. They were instructed to complete the insider task using whatever approach they wished, but to avoid raising suspicion from teammates, since they were able to report their concerns to Gold Command. All of the participants approached to be an insider agreed to the task.

The multiple teams and sequence of periods in iCOS were designed to provide the insiders multiple opportunities to complete their tasking. For example, they may utilize friendships with members of other teams to hide malicious information gathering, or distribute their activity across multiple periods to make it more difficult to spot patterns of malicious activity. Similarly, the work-breaks taken by co-workers (e.g., for the bathroom or between game periods) afforded opportunities for players to exhibit behavior that compromised security. For example, we observed an average of four unattended laptops per simulation, despite instructions to not do so, and we observed participants leaving their password cards on their desk on five occasions across the simulations. Both of these provided an insider the opportunity to note down details before joining the others on their break, and we observed one participants taking advantage of this opportunity at the beginning of a lunch break.

The third and fourth period of iCOS matched the second period in terms of structure, but in order to ensure the simulation remained interesting and challenging, the investigative and insider tasks became progressively more difficult. For example, in each period, the same 'insider' participants were covertly invited (via email because it was

possible to obscure the true intent of the message) to complete an insider act for a further £20. In the second period this act required discovering and reporting some non-task relevant information from a database to which they had direct access. In the third period the act required reporting information from their team's databases to which they had no direct access. In the fourth period they were required to report two pieces of related information from another team's database. In no instances did an insider refuse to complete the task offered to them. In two instances during the fourth period the insider managed to only retrieve one of the two pieces of required information (they were retained as insiders within the analysis, since they completed most of the tasks successfully).

Once the four periods of the simulation were completed, players moved to a meeting room where they were informed that the security of the information to which they had access may have been compromised, and that their behavior during the simulation would be the subject of further investigation. They were then asked to complete a post-simulation interview and questionnaire (collected to test different kinds of interview protocol rather than insider detection through language use, and so not examined here), after which they were debriefed.

Analysis of Behavior

To derive measures of language use, participants' emails were analyzed using the text analysis program Linguistic Inquiry and Word Count (LIWC; Pennebaker, Booth, & Francis, 2007). LIWC calculates the proportion of words in a text that match a set of categories that relate to affective, cognitive and social dimensions. These categories have been shown to be both reliable (Tausczik & Pennebaker, 2010) and valuable in their

contribution to the analysis of written texts (Hancock, Woodworth, & Porter, in press; ten Brinke & Porter, 2011; Vrij et al., 2007). In the current study we focused on nine LIWC categories and sub-categories that are relevant to the hypotheses. Specifically, to examine change in self-focus we measured participants' use of personal pronouns, including first-person singular pronouns (e.g. 'I', 'me'), first-person plural pronouns (e.g., 'we', 'our') and second-person singular pronouns (e.g., 'you', 'you're'). To examine changes in emotional state we measured participants' use of negative emotion words (e.g., 'awkward', 'dislike') and words relating to the expression of feelings (e.g., 'fragile', 'pressed'). To examine changes in cognitive processes we measured participants' use of words relating to overall cognitive processes (e.g., 'evaluate', 'infer'), and those relating specifically to discrepancies (e.g., 'ought', 'should') and tentative presentation (e.g., 'kinda', 'maybe').

Linguistic mimicry. To measure the degree of language mimicry in participants' emails, we used an index known as Language Style Matching (LSM). LSM measures the extent to which two speakers match in their use of nine categories of function words: adverbs (e.g., 'very', 'well'), articles (e.g., 'a', 'the'), auxiliary verbs (e.g., 'am', 'have'), conjunctions (e.g., 'but', 'therefore'), indefinite pronouns (e.g., 'this', 'it'), negations (e.g., 'no', 'not'), personal pronouns (e.g., 'I', 'we'), prepositions (e.g., 'in', 'around'), and quantifiers (e.g., 'many', 'few') (for more details see Gonzales et al., 2010; Ireland et al., 2011) . These nine categories cover more than half of the vocabulary of daily speech (Rochon, Saffran, Berndt, & Schwartz, 2000) and correlate with social behaviors such as dominance, deception, and social bonding (Tausczik & Pennebaker, 2010).

For each of the nine categories, LSM is calculated using the following formula

(the articles category is used here as an example):

$$LSM_{articles} = 1 - [(|articles_P - articles_S|) / (articles_P + articles_S + .0001)],$$

where $articles_P$ is the percentage of articles used by the participant, $articles_S$ is the percentage of articles used by the other team player. The addition of .0001 in the denominator is used to prevent division by zero. The resulting nine category-specific LSM scores were then averaged to produce a single LSM score for matching. This score varies between .00 and 1.00, where a higher score indicates greater linguistic similarity between the two speakers. For the simulation data, we obtained an LSM score for the participants' interaction with each other member of his or her team.

We assessed the internal consistency of the LSM measure by examining the extent to which the nine function word categories produced similar scores. Cronbach's alpha computed across the nine categories suggested that the LSM index had good internal consistency for the first ($\alpha = .77$), second ($\alpha = .82$), third ($\alpha = .74$), and fourth ($\alpha = .85$) time period (cf. Ireland & Pennebaker, 2010).

Results

Game Communication

Participants wrote an average of 325.4 words (SD = 197.0) in each round of the game, and there was no difference in the average number of words typed across the rounds, $F(3,147) = 1.93$, $p = .12$. There was also no difference in the number of words written by insiders and co-workers across the rounds, $F < 1$, *ns*. Table 1 shows the mean percentage of words for the nine LIWC categories, as a function of game period and whether or not the participant acted as an insider. These data were submitted to a series of 4 (Game period: 1 through 4) x 2 (Intention: Insider vs. Co-worker) mixed ANOVAs that

controlled for differences in assigned organizational role and team membership. These ANOVAs were followed up with Helmert interaction contrasts that tested for differences in insiders and co-workers' behavior before (i.e., the first game period) and after (i.e., the second through fourth game period) participants were tasked to be an insider.

Consistent with Hypothesis 1, insiders showed greater self-focus than co-workers after becoming insiders. Insiders' use of personal pronouns was associated with a Game period x Intention interaction, $F(3,147) = 4.92$, $p < .01$, $\eta^2 = .10$, with insiders using significantly more personal pronouns than their co-workers in the second game period, $t(52) = 3.49$, $p < .01$, 95% CI [1.25, 4.64], and third game period, $t(52) = 2.56$, $p < .01$, 95% CI [.46, 3.85]. The Helmert contrast was consistent with this pattern, showing divergence between insiders' and co-workers' behavior when comparing the second game period to later periods, $F(1,49) = 7.05$, $p < .05$, $\eta^2 = .14$, 95% CI [1.20, 8.69]. A series of equivalent analyses for the various types of personal pronouns added support to the prediction that insiders seek to distance themselves from their co-workers. There were significant interaction effects for the singular pronoun categories that distinguish self from other, namely first-person pronouns, $F(3,147) = 4.26$, $p < .01$, $\eta^2 = .09$, and second-person pronouns, $F(3,147) = 2.86$, $p < .05$, $\eta^2 = .06$. By contrast, there was no significant interaction for the category that combines self with other, namely, first-person plural pronouns, $F < 1$.

There was less evidence to support the prediction that insiders would show more negative affect compared to their co-workers (Hypothesis 2). The Game period x Intention interactions were non-significant for use of negative emotion words, $F(3,147) = 1.31$, $p = .27$, and use of words expressing feelings, $F(3,147) = 1.45$, $p = .23$. However,

the planned Helmert contrasts showed a trend toward insiders using more negative emotion words than their co-workers once assigned their task, $F(1,49) = 2.29$, $p = .14$, $\eta^2 = .05$, 95% CI [-.34, 2.40], and a similar, marginally significant trend for insiders using more feeling related words once assigned their task, $F(1,49) = 3.69$, $p = .06$, $\eta^2 = .08$, 95% CI [-.06, 1.53]. A further examination revealed that the marginal nature of these findings stemmed from the fact that the difference across insiders and co-workers diminished after the initial round of the tasking. For example, insiders used significantly more negative emotion words than co-workers during the second game period, $t(52) = -2.22$, $p < .05$, 95% CI [-.96, -.05], marginally more during the third game period, $t(52) = 1.34$, $p = .09$, but no more during the fourth time period, $t < 1$.

Consistent with Hypothesis 3, the Game period x Intention interaction suggested that insiders and co-workers showed a significant difference in their use of language associated with cognitive processing, $F(3,147) = 3.45$, $p < .05$, $\eta^2 = .07$. The Helmert contrast revealed that participants tasked to be an insider showed more cognitive processing than their co-workers, $F(1,49) = 4.64$, $p < .05$, $\eta^2 = .09$, 95% CI [.64, 18.60]. As might be expected, this difference was largest during the second period when insiders had just been assigned their tasks, $t(52) = 2.22$, $p < .05$, 95% CI [.29, 5.53]. As with the previous hypotheses, we conducted equivalent analyses on the sub-categories of language that comprise the cognitive processing category. These analyses revealed that the overall effect was driven primarily by changes in insiders' use of discrepancy words, $F(1,49) = 10.42$, $p < .01$, $\eta^2 = .21$, 95% CI [1.59, 6.82], and tentative words, $F(1,49) = 4.72$, $p < .01$, $\eta^2 = .10$, 95% CI [.30, 7.60].

Language Style Matching

The interpersonal data were submitted to the equivalent 4 (Period) x 2 (Intention) mixed ANOVA as the individual data, but with Linguistic Style Matching (LSM) scores as the Dependent Variable. Figure 1 shows the mean LSM scores for the three periods in which insiders were active relative to the mean LSM scores in the first period (i.e., the baseline). As can be seen from Figure 1, there was decreasing language synchrony for both insiders and co-workers over the four game periods, $F(3,801) = 26.08$, $p < .001$, $\eta^2 = .10$, which is consistent with the fact that team members disagreed more over the difficult tasks of later periods. However, consistent with Hypothesis 4, this main effect was subsumed by a significant period x intention interaction, $F(3,801) = 4.66$, $p < .01$, $\eta^2 = .02$. The linear decrease in LSM over the periods was significantly sharper for insiders compared to co-workers, $F(1,267) = 12.61$, $p < .001$, $\eta^2 = .05$, 95% CI [-.26, -.08]. Once participants accepted the insider task, their language matching with other players deteriorated to a level significantly below co-workers matching with one another.

To examine this change in more detail, we examined LSM scores for each of the nine function word categories separately. This analysis revealed significant period x intent interaction trends, equivalent to that found for the overall LSM scores, for auxiliary verbs, $F(1,267) = 10.10$, $p < .01$, $\eta^2 = .04$, 95% CI [-.33, -.08], quantity modifiers, $F(1,267) = 13.50$, $p < .001$, $\eta^2 = .05$, 95% CI [-.60, -.18], and, for the final two periods, prepositions, $F(1,267) = 10.16$, $p < .01$, $\eta^2 = .04$, 95% CI [-.21, -.05], and conjunctives, $F(1,267) = 14.50$, $p < .001$, $\eta^2 = .05$, 95% CI [-.38, -.12]. In all of these instances, insiders showed increasingly less synchrony in the content of their emails with fellow co-workers, compared to co-workers' synchrony with one another. Collectively this trend across the

four word categories suggests that insiders fail to match the degree of concreteness shown by co-workers: they did not match co-workers' specificity in describing issues, as achieved through auxiliary verbs (e.g., 'ought', 'should') and quantitative modifiers (e.g., 'few,' 'some'), and they did not match co-workers' descriptions of how issues transpired or how they relate to one another, as provided by prepositions (e.g., 'by,' 'with') and conjunctions (e.g., 'and,' 'but').

Identification of Insiders

To explore the extent to which the observed differences in linguistic behavior allow insiders to be distinguished from their co-workers, we conducted a binary logistic regression in which the four individual categories of personal pronouns, negative emotions, feelings, and cognitive processes, together with LSM scores, were regressed on participants' role (i.e., insider or co-worker). In this regression, the linguistic variables from the second, third and fourth periods were entered as predictors in blocks (i.e., one period per block), after controlling as before for assigned organizational role and team membership. This approach enabled us to model the extent to which the accumulating knowledge of participants' behavior across the simulation periods enabled the discrimination of insiders from their co-workers.

Collectively linguistic behavior predicted group membership for the second period, $\chi^2(8) = 28.76$, $p < .01$, AUC = .804, 95% CI [.65, .96], second and third period, $\chi^2(13) = 32.90$, $p < .01$, AUC = .863, 95% CI [.75, .98], and second through fourth period, $\chi^2(18) = 52.68$, $p < .01$, AUC = .959, 95% CI [.91, 1.00]. As information from each period was incorporated into the model, classification of insiders and co-workers improved from the chance level of 74.1% (since 14 of the 54 participants were insiders)

to 83.3%, 85.2%, and 92.6%. Table 2 summarizes the regression models derived as each simulation period was entered into the analysis. As expected, the coefficients in Table 2 mirror the ANOVA contracts, with insiders on the whole distinguishing themselves from co-workers by using more personal pronouns, showing subtle differences in affect and cognitive processing, and by reducing their degree of language style matching.

One limitation of this analysis is that the predictor to case ratio was low, particularly when all of the blocks were entered into the regression. This may have inflated the accuracy of classification and likely accounts for the large and variable beta values associated with predictors across the periods. In these circumstances, a better assessment of the significance of individual predictors may come from examining each period separately, since this reduces the number of predictors in each model. When stepwise binary logistic regressions were run for each period separately, significant predictors were found for Period 2: personal pronouns, $b = .445$, SE = .150, Wald = 8.84, $p < .05$, and LSM, $b = -6.143$, SE = 2.58, Wald = 5.66, $p < .05$; for Period 3: $b = .365$, SE = .154, Wald = 5.61, $p < .05$, and LSM, $b = -8.382$, SE = 3.30, Wald = 6.46, $p < .05$; and for Period 4: LSM, $b = -4.71$, SE = 2.53, Wald = 3.48, $p = .062$. These results suggest that personal pronouns and LSM play an important role in discriminating insiders from their co-workers.

Discussion

Outside of interventions that catch perpetrators ‘red handed,’ the challenge for investigators of insider threat is to identify methods that allow legitimate but suspicious behavior to be identified and investigated. A common approach is to encourage employee whistle-blowing, but this relies on an employee’s motivation, vigilance, and belief that

such action will be received positively by co-workers and supervisors (Conchie, Taylor, & Donald, 2012). In this article, we explored the value of an alternative approach that works unobtrusively by examining the language behavior of mock employees carrying out day-to-day tasks. Overall, we found that those tasked to undertake an insider attack showed change in their language use that were consistent with predicted change in work motivation and affect towards co-workers and the organization.

The language of insiders showed a shift in personal presentation and cognitive focus when compared to both their own behavior before the insider tasking, and to the behavior of their co-workers during the attack. Specifically, compared to these two 'baselines,' insiders used significantly more personal pronouns, particularly first-person singular pronouns, suggesting a shift in focus toward self and away from issues relating to the collective efforts of the group (Pennebaker & Lay, 2002). These increases were not matched by a change in first-person plural pronoun use. Insiders showed no increase in their use of first-person plural pronouns, arguably reinforcing the fact that their perception is one of being separate from the immediate work group. This shift in language is consistent with evidence suggesting that an increase in first-person pronoun use reflects increasing social isolation and an absence of social connection (Pennebaker, 2012).

In contrast to our hypothesis, we found only a trend for insiders using more negative emotion and feeling words compared to their co-workers. The difference across insiders and co-workers was most evident immediately after the insiders were tasked, with any difference diminishing in later stages. In one sense this is not surprising since our participants, having only spent one day in the simulation, were not likely to have the

kind of personal investment that might be true for employees turned insiders (cf. Donohue & Taylor, 2007). The difference in affective language may reflect a short-lived response to the challenge of undertaking the insider task, rather than a pervasive shift in feelings for co-workers. In a more intensive simulation, or in real-world scenarios, the greater personal investment may lend itself to greater shifts in affect when undertaking the insider attack. However, it may also reflect the fact that emotion words may not provide the clearest insight into a person's affective state, which can be expressed in many different ways (Pennebaker, Mehl, & Niederhoffer, 2003). One of these ways is through non-verbal and para-linguistic behaviors, such as intonation and facial expressions, as expressed during work-related interactions with co-workers. It is perhaps these kinds of behaviors that are picked up by co-workers when they whistle-blow on suspicious behaviors. This suggests that there may be mileage in having future 'whistle-blowing' initiatives highlight suspicious emotional presentation as something to report.

Consistent with changes in social and identity language use, insiders also showed change in the way they thought and expressed themselves. Compared to their co-workers, insiders presented with a greater number of words associated with cognitive complexity. Central to this difference was greater use of words that acknowledge discrepancies within an account and words that make assertions tentative. Consistent with previous research, these language changes suggest an increase in cognitive processing that may reflect the greater task challenge required for an insider who must complete both their legitimate and illegitimate work. However, a second explanation for this language use recognizes that it is the kind of language that is used when justifying a particular action (e.g., "I know it may look strange but…") or mitigating the unusual nature of a request (e.g., "I

know what I ought to have done was…"). Thus, in the insider threat scenario, the increase in cognitive complexity likely centers on both increased task load and a tendency to present the 'exceptions' and 'shades of gray' to others, in order to avoid suspicion and lower the extent of direct deceit.

Interestingly, for cognitive processes, the difference between insiders and co-workers was most prevalent in the first insider period (the second of the simulation's four periods) and reduced in prevalence across the final two periods. This pattern of behavior fits nicely with the notion that individuals recover and adapt to the dual nature of the insider task over time (Wise, 2003). After the initial difficulty of managing both tasks, and the consequences that this has on both work behavior and social interactions, the insider may have begun to develop methods of coping with their role (cf. Griffin, Neal, & Parker, 2007). These methods may have been strategies relating to how to ensure that there was no slippage across the roles when communicating with co-workers. Or, it may have related to strategies for conducting the insider task in an effective way. This speculation has implication for the wider question of whether it is easier to identify deceivers at initial interview when they are unprepared for their lie, or in later interviews, when they must work hard to maintain the consistency of their lie (Leins, Fisher, Vrij, Leal, & Mann, 2011; Porter & ten Brinke, 2010). Within the insider scenario, our data suggest that suspicious behaviors may sometimes be most evident in the early stages before insiders learn to manage the challenges of remaining undiscovered.

Limitations

Although our simulation embodied many features of organizational practice, its scale meant that there were differences between what our participants experienced during

the simulation and what might occur in practice. For example, one key aspect of the insider threat problem is its low base rate of occurring. In part recognition of this, we broke away from the traditional paradigm of balancing the number of genuine and deceptive participants in a study (e.g., Vrij et al., 2008) by having only 25% of participants act as insiders. However, even this proportion is likely higher than the prevalence of insiders within an organization, which may be one amongst hundreds of co-workers. In recognizing this difference it is easy to make the inference that an insider within a large organization may be harder to identify, since his or her suspicious behavior is situated in a larger set of genuine behaviors. This would mean that our results are an over-optimistic account of how language can help identify insiders. However, this conclusion ignores the fact that insiders within larger co-worker communities have a degree of anonymity that may lead to overconfidence in their actions (Robinson & O'Leary-Kelly, 1998). Thus, while a low base-rate presents a challenge for insider threat detection, it may equally be the case that our results reflect insiders who were being particularly cautious about the way in which they behaved.

A second, related limitation of this study concerns the absence of a 'world' outside of the simulation. There are two aspects that are likely to be important. The first is that employees often communicate with individuals outside of their own organization, which increases the heterogeneity of the email traffic beyond that which occurred in our simulation. The second is that some insiders receive collaborative help from outsiders to conduct the attack. These two possibilities highlight the range of other variables that may come into play when techniques like the one explored in this paper are transferred to the work domain. There are a variety of ways in which insider attacks can be committed, and

a variety of organizational structures within which such an attack can be carried out. An interesting development of this research, therefore, is both to identify how these factors influence the availability of suspicious behaviors, and to examine whether certain forms of organizational structure promote the availability of such behaviors. In our simulation, the notable absence of a world outside the simulation took away some of the complexities that might make the identification of insiders more difficult.

In relation to this, it is notable that almost all of our insiders completed their task successfully, suggesting that variations in the complexity of the insider's task and the security of the organization's systems may impact not only the visibility of insider behavior but, arguably, also the type of insider behavior observed. For example, in an organizational environment where it is not possible to leave a terminal unlocked or a password unattended, the insider will likely need to utilize alternative approaches (e.g., social engineering, Stajano & Wilson, 2011), and this may result in a different kind of interpersonal behavior.

A third limitation concerns the nature of the participants who took part in our simulation. Our insiders were chosen at random with no consideration given to their personality type, motivation, or personal circumstances. Although this allocation of participants to conditions minimizes any unintended biases, it may not tally with how insiders emerge, or are chosen by provocateurs, in reality. Insiders may be motivated by precipitating events (e.g., perceived organizational injustice, Greitzer et al., 2010) or adverse life circumstances (e.g., excessive debt, Maybury et al., 2006), or even chosen by provocateurs based on certain personality traits (Shaw, Ruby, & Post, 1998). It should be possible to examine such factors using a more elaborate simulation that focuses on

‘vulnerable’ personalities and manipulate at least the organizational pre-cursors prior to monitoring behavior.

Our findings demonstrate how language can provide an indirect way of identifying people who are undertaking an insider attack. The fact that changes in insider behavior were identifiable even without extensive ‘profiling’ of their typical behavior use is impressive, and suggests that this approach may add value to more technical efforts to identify insiders directly.

References

Cappella, J. N., & Panalp, S. (1981). Talk and silence in sequences in informal conversations: Interspeaker influence. *Human Communication Research, 7,* 117-132. doi:10.1111/j.1468-2958.1981.tb00564.x

Christoph, G. G., Jackson, K. A., Neuman, M. C., Sicilliano, C. L. B., Simmonds, D. D., Stallings, C. A., & Thompson, J. L. (1995). UNICORN: Misuse detection for UNICOS. *Proceedings of the IEEE/ACM SC95 conference* (p. 56). doi:10.1109/SUPERC.1995.241777.

Colwill, C. (2009). Human factors in information security: The insider threat—who can you trust these days? *Information Security Technical Report, 14,* 186-196. doi:10.1016/j.istr.2010.04.004

Cohn, M. A., Mehl, M. R., & Pennebaker, J. W. (2004). Linguistic markers of psychological change surrounding September 11, 2001. *Psychological Science, 15,* 687-693. doi:10.1111/j.0956-7976.2004.00741.x

Conchie, S. M., Taylor, P. J., & Donald, I. J. (2012). Promoting safety voice with safety-specific transformational leadership: The mediating role of two dimensions of trust. *Journal of Occupational Health Psychology, 17,* 105-115. doi: 10.1037/a0025101

Coupland, N., & Giles, N. (1998). Introduction: The communicative contexts of accommodation. *Language and Communication, 8,* 175- 182. doi:10.1016/0271-5309(88)90015-8

Donohue, W. A., & Taylor, P. J. (2007). Role effects in negotiation: The one-down phenomenon. *Negotiation Journal, 23,* 307-331. doi:10.1111/j.1571-9979.2007.00145

Druckman, D. (2005). *Doing research: Methods of inquiry for conflict analysis*. Thousand Oaks: California.

Garrod, S., & Pickering, M. J. (2004). Why is conversation so easy? *Trends in Cognitive Science, 8,* 8-11. doi:10.1016/j.tics.2003.10.01

Gonzales, A. L., Hancock, J. T., & Pennebaker, J. W. (2010). Language style matching as a predictor of social dynamics in small groups. *Communication Research, 31,* 3-19. doi: 10.1177/0093650209351468

Greitzer, F. L., Noonan, C. F., Kangas, L. J., & Dalton, A. C. (2010). *Identifying at-risk employees: A behavioral model for predicting potential insider threats*. Pacific Norhtwest National Laboratory report (PNNL-19665). Richland, WA.

Griffin, M. A., Neal, A., & Parker, S. K. (2007). A new model of work role performance: Positive behavior in uncertain and interdependent contexts. *Academy of Management Journal, 50,* 327-347. doi:view/UQ:136277

Hancock, J. T., Woodworth, M. T., & Porter, S. (in press). Hungry like a wolf: A word-pattern analysis of the language of psychopaths. *Legal and Criminological Psychology*. doi:10.1111/j.2044-8333.2011.02025.x

Herbig, K. L., & Wiskoff, M. F. (2002). *Espionage against the United States by American citizens 1947-2001*. Defense Personnel Security Research Center (technical report: PERSEREC-TR 02-5).

Hollinger, R. C., & Clark, J. P. (1983). *Theft by employees*. Lexington, MA: Heath.

Ireland, M. E. & Pennebaker, J. W. (2010). Language style matching in writing: Synchrony in essays, correspondence, and poetry. *Journal of Personality and Social Psychology, 99*, 549-571. doi:10.1037/a0020386

Ireland, M. E., Slatcher, R. B., Eastwick, P. W., Scissors, L. E., Finkel, E. J., & Pennebaker, J. W. (2011). Language style matching predicts relationship initiation and stability. *Psychological Science, 22*, 39-44. doi:10.1177/095679761039292

Junghaenel, D. U., Smyth, J. M., Santner, L. (2008). Linguistic dimensions of psychopathology: A quantitative analysis. *Journal of Social and Clinical Psychology, 27,* 36-55. doi:10.1521/jscp.2008.27.1.36

Keeney, M., Kowalski, E., Cappelli, D., Moore, A., Shimeall, T., & Rogers, S. (2005). *Insider threat study: Computer system sabotage in critical infrastructure sector*. Technical report. US Secret Service and CERT Program, Software Engineering Institute, Carnegie Mellon University, Pittsburgh, PA.

Kowalski, E., Cappelli, D., & Moore, A. (2008). *U.S. secret service and CERT/SIE insider threat study: Illicit cyber activity in the information technology and telecommunications sector*. Carnegie Mellon University.

Leins, D., Fisher, R. P., Vrij, A., Leal, S., & Mann, S. (2011). Using sketch drawing to induce inconsistency in liars. *Legal and Criminological Psychology, 16,* 253-265. doi:10.1348/135532510X501775

Maloof, M. A., & Stephens, G. D. (2007). ELICIT: A system for detecting insiders who violate need-to-know. *Lecture Notes in Computer Science, 4637,* 146-166. doi:10.1007/978-3-540-74320-0_8

Masip, J., Sporer, S., Garrido, E., & Herrero, C. (2005). The detection of deception with the reality monitoring approach: A review of the empirical evidence. *Psychology, Crime, and Law, 11,* 99-122. doi:10.1080/10683160410001726356

Maybury, M., Chase, P., Chiekes, B., Pietravalle, R., Sebring, J., Costa, M., Brackney, D., Lehtola, P., Matzner, S., Hetherington, T., Wood, B., Sibley, C., Marin, J., Longstaff, T., Spitzner, L., Haile, J., Copeland, J., & Lewnadowski, S. (2006). Detection of malicious insider activity using models of insider behavior. *Journal of Intelligence Community Research and Development*.

Pennebaker, J. W. (2012). *The secret life of pronouns: What our words say about us*. New York: Bloomsbury Press.

Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). Linguistic Inquiry and Word Count (LIWC2007): *A computer-based text analysis program [computer software]*. Austin, TX: LIWC.net.

Pennebaker, J. W., & Lay, T. C. (2002). Language use and personality during crises: Analyses of Mayor Rudolph Giuliani's press conferences. *Journal of Research in Personality, 36,* 271-282. doi:10.1006/jrpe.2002.2349

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology, 54,* 547-577. doi:10.1146/annurev.psych.54.101601.145041

Porter, S., & ten Brinke, L. (2010). The truth about lies: What works in detecting high-stakes deception? *Legal and Criminological Psychology, 15,* 57-75. doi: 10.1348/135532509X433151

Randazzo, M. R., Cappelli, D., Keeney, M., Moore, A., & Kowalski, E. (2004). *U.S. secret service and CERT coordination center/SEI insider threat study: Illicit cyber activity in the banking and finance sector*. Carnegie Mellon University. Available at: www.cert.org/archive/pdf/bankfin040820.pdf

Robinson, S. L., & O'Leary-Kelly, A. M. (1998). Monkey see, monkey do: The influence of work groups on the antisocial behavior of employees. *Academy of Management Journal, 41,* 658-672. doi:10.2307/256963

Rochon, E., Saffran, E. M., Berndt, R. S., Schwartz, M. F. (2000). Quantitative analysis of aphasic sentence production: Further development and new data. *Brain and Language, 72,* 193-218. doi:10.1006/brln.1999.2285

Rude, S. S., Gortner, E. M., & Pennebaker, J. W. (2004). Language use of depressed and depression-vulnerable college students. *Cognition and Emotion, 18,* 1121-1133. doi:10.1080/02699930441000030

Schultz, E. E. (2002). A framework for understanding and predicting insider attacks. *Computers & Security, 21,* 526-531. doi:10.1016/S0167-4048(02)01009-X

Shaw, E., Fischer, L., & Rose, A. (2009). *Insider risk evaluation and audit (technical report 09-02)*. Defence Personnel Security Research Center. Monterey, CA. Available from: http://www.dhra.mil/perserec/reports/tr09-02.pdf

Shaw, E., Ruby, J. G., & Post, J. M. (1998). The insider threat to information systems: The psychology of the dangerous insider. *Security Awareness Bulletin, 2-98,* 27-46.

Spence, S. A., Farrow, T. F. D., Herford, A. E., Wilkinson, I. D., Zheng, Y., & Woodruff, P. W. R. (2001). Behavioural and functional anatomical correlates of deception in humans. *Neuroreport, 12,* 2849-2853. doi: 10.1002/cbm.785

Spitzner, L. (2003). *Honeypots: Catching the insider threat*. Proceedings of the 19th annual Computer Security Applications conference (pp. 170-179). doi:10.1109/CSAC.2003.1254322

Stajano, F., & Wilson, P. (2011). Understanding scam victims: Principles for systems security. Communications of the ACM, 54, 70-75. doi:10.1145/1897852.1897872

Stanton, J. M., Stam, K. R., Guzman, I., & Caldera, C. (2003). Examining the linkage between organizational commitment and information security. *IEEE: Systems, Man and Cybernetics, 3,* 2501-2506. doi:10.1109/ICSMC.2003.1244259.

Stirman, S. W., & Pennebaker, J. W. (2001). Word use in the poetry of suicidal and nonsuicidal poets. *Psychosomatic Medicine, 63,* 517-522. doi:0033-3174/01/6304-0417

Stephens, G. D., & Maloof, M. A. (2009). Detecting insider theft of trade secrets. *IEEE Security & Privacy, 7,* 14-21. doi:10.1109/MSP.2009.110

Taylor, P. J., & Thomas, S. (2008). Linguistic style matching and negotiation outcome. *Negotiation and Conflict Management Research, 1,* 263-281. doi:10.1111/j.1750-4716.2008.00016.x

Tausczik, Y. R., & Pennebaker J. W. (2010). The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology, 29,* 24-54. doi:10.1177/0261927X09351676

ten Brinke, L., & Porter, S. (2011). Cry me a river: Identifying behavioral consequences of extremely high-stakes interpersonal deception. *Law and Human Behavior.* doi:10.1037/h0093929.

Tickle-Degnen, L., & Rosenthal, R. (1990). The nature of rapport and its nonverbal correlates. *Psychological Inquiry, 1,* 285-293. doi:10.1207/s15327965pli0104_1

Vrij, A., Mann, S., Kristen, S., & Fisher, R. P. (2007). Cues to deception and ability to detect lies as a function of police interview style. *Law and Human Behavior, 31,* 499-518. doi:10.1007/s10979-006-9066-4

Vrij, A., Mann, S. A., Fisher, R. P., Leal, S., Milne, R., & Bull, R. (2008). Increasing cognitive load to facilitate lie detection: The benefit of recalling an event in reverse order. *Law and Human Behavior, 32,* 253-265. doi:10.1007/s10979-007-9103-y

Walczyk, J. J., Schwartz, J. P., Clifton, R., Adams, B., Wei, M., & Zha, P. (2005). Lying person-to-person about live events: A cognitive framework for lie detection. *Personnel Psychology, 58,* 141-170. doi:10.1111/j.1744-6570.2005.00484.x

Williams, P. A. H. (2008). In a 'trusting' environment, everyone is responsible for information security. *Information Security Technical Report, 13,* 207-215.

Wise, D. (2003). *Spy: The insider story of how the FBI's Robert Hanssen betrayed America*. New York: Random House.

Workman, M., & Gathegi, J. (2007). Punishment and ethics deterrents: A comparative study of insider security contravention. *Journal of the American Society of Information Science and Technology, 58,* 318-342. doi:10.1002/asi.20474

Table 1.

LIWC Word Categories and Sub-Categories Scores as a Function of Time Period and Insider (I) versus Co-Worker (C)

Language Category	Period							
	1		2		3		4	
	I	C	I	C	I	C	I	C
Personal Pronoun	7.88	7.15	8.67	5.72	7.61	5.45	5.33	5.83
First-person singular	3.05	2.31	3.54	1.94	2.93	2.22	1.60	2.24
First-person plural	2.05	2.30	2.04	1.50	1.16	.95	1.32	1.31
Second-person	.89	.84	1.02	.78	1.61	.71	.71	.72
Negative Emotion	.40	.45	1.02	.51	1.04	.69	.71	.63
Feelings	.13	.23	.32	.20	.56	.31	.93	.58
Cognitive Processes	13.45	16.01	16.38	13.47	12.17	12.48	12.52	13.68
Discrepancies	1.31	1.95	2.40	1.25	1.50	.96	1.03	.88
Tentative	2.75	3.28	3.66	3.18	2.38	2.89	4.29	2.83

Table 2.

Language Predictors of Insiders as a Function of Cumulative Simulation Period

Language Predictors	Model (Blocks)					
	Period 2	Period 2 + 3		Period 2 + 3 + 4		
	2	2	3	2	3	4
Personal Pronouns	.624*	.414	.402	2.05	3.39	-2.00
Negative Emotion	-.285	.259	-.660	-1.11	-2.26	-.566
Feelings	.370	.369	-.070	2.45	6.01	-2.63
Cognitive Processes	-.086	-.072	-.144	.098	-.615	.118
LSM	-6.33*	-.441	-6.41	34.19	-21.51	-57.97

* $p < .05$

Figure 1. Linguistic Style Matching scores relative to baseline (Period 1) as a function of game period and intention (Error bars = 95% CI)