2005

Detecting Online Deception and Responding to It

Rowe, Neil C.

Monterey, California. Naval Postgraduate School

Encyclopedia of Virtual Communities and Technologies, Hershey, PA: Idea Group, 2005.

Detecting Online Deception and Responding to It

Neil C. Rowe

U.S. Naval Postgraduate School
Code CS/Rp, 833 Dyer Road, Monterey, California 93943 USA
email ncrowe@nps.edu

ABSTRACT

Since many forms of online deception are harmful, it is helpful to enumerate possible detection methods. We discuss low-levels clues such as pauses and overgenerality as well as ?cognitive? clues such as noticing of factual discrepancies. While people are generally poor at detecting deception using their intuitions, the online environment provides the ability to automate the analysis of clues and improve the likelihood of detection by doing ?data fusion?. Appropriate responses to deception must differ with the type, as some deceptions like deliberate provocation are best handled by ignoring them while other deceptions like fraud are best handled by exposure.

This article appeared in the *Encyclopedia of Virtual Communities and Technologies*, Hershey, PA: Idea Group, 2005.

INTRODUCTION

An important problem in online communities is detection of deception by their members. Deception is a form of manipulation, and can have many varied negative consequences in a virtual community, especially once discovered (Joinson & Dietz-Uhler, 2002) but even if undiscovered. Virtual communities need to be aware of the problems and need to agree on policies for detecting deception and responding to it.

BACKGROUND

Online deception is encouraged by the special circumstances of online communities (George and Carlson, 1999). Studies have shown that deception occurrence is inversely related to communications bandwidth, or the rate at which data can be transmitted between people (Burgoon et al, 2003). In other words, people feel more inclined to deceive the more remote and less familiar they are to the deceivees, and both factors usually apply online. Unfortunately, people are less effective at detecting deception than they think they are (Eckman 2001). Online deception is especially difficult to detect; in many cases it is never discovered or is discovered much later, due to the lack of authority in cyberspace and the temporary nature of much cyberspace data.

DECEPTION DETECTION METHODS

There is a large literature on the detection of deception in conventional face-to-face social interaction. Although people are often poor at detecting deception, they can improve some with training (Ford, 1996).

People doing detection can use both low-level and high-level clues. Low-level clues can be both nonverbal and verbal (see Table 1). Nonverbal clues ("cues") are generally more telling since they are often harder to suppress by the deceiver (Miller & Stiff, 1993). One must be cautious because not all popularly ascribed clues are effective: Polygraphs or electronic "lie detectors" have not been shown to do better than chance. Note some nonverbal clues appear even without audio and video connections; for example, (Zhou & Zhang, 2004) showed four nonverbal factors they called "participation" were correlated in experiments with deception in text messaging, such as the pause between messages.

Table 1: Low-level clues to interpersonal deception.

Visual clues	Vocal clues	Verbal clues
increased blinking (video)	hesitation (text, audio, video)	overgenerality (text, audio)
increased self-grooming actions (video)	shorter responses and shorter pauses (text, audio, video)	increased use of negatives (text, audio)
increased pupil dilation (video)	increased speech errors (audio)	increased irrelevance (text, audio)
	higher voice pitch (audio)	increased hyperbole (text, audio, video)

High-level clues (or "cognitive" ones) involve discrepancies in information presented (Bell and Whaley, 1991; Heuer, 1982), and they can occur in all forms of online interaction. For instance, if a person A says they talked to person B but B denies it, either A or B is deceiving. Logical fallacies often reveal deception, as in advertising (Hausman, 1999); for instance, a diet supplement may claim you can lose ten pounds a week without changing your diet. In deception about matters of fact like news reports, checks of authoritative references can reveal the deception. Inconsistency in tone is also a clue to deception, as when someone treats certain people online very differently than others.

Suspiciousness of clues is enhanced by secondary factors: the less clever the deceiver, the more emotional the deceiver, the less time they have to plan the deception, the less chance they will be caught, the higher the stakes, the less familiarity of the deceiver and deceivee, and the more pleasure the deceiver attains from a successful deception (Eckman & Frank, 1993). The perceived likelihood of deception can be estimated as the opposite of the likelihood that a sequence of events could have occurred normally.

Specialized statistical methods can also be developed for recognizing common online deceptions like fraud in online commercial transactions (MacVittie, 2002), criminal aliases (Wang, Chen, & Akabashch, 2004) and the doctoring of Web pages to get better placement in search engines (Kaza, Murthy, & Hu, 2003). For instance, clues that online transactions involve stolen credit-card numbers are an email address at a free email service, a difference between the shipping and billing addresses, and an IP address (computer identity code) for the originating computer that is geographically inconsistent with the billing address (MacVittie, 2002).

DATA FUSION FOR BETTER DETECTION OF DECEPTION

It is important for detection to consider all available clues for deception, since clues can be created inadvertently by nondeceptive people. Thus we have a problem of "data fusion" or of combining evidence. Besides observed clues from the suspected deceiver themselves, we can include the reputation of a person within a virtual community as in EBay-style reputation-management systems (Barber & Kim, 2001; Yu & Singh, 2003).

Several researchers have proposed mathematical formulations of the fusion problem. If clues are independent, then the probability of deception is the inverse of the product of the inverses of the probabilities of deception given each clue, where the inverse is one minus the probability. A generalization of this is the Bayesian network where related non-independent probabilities are grouped together (Rowe, 2004). Other approaches also appear successful (Carofiglio, de Rosis, & Castelfranchi, 2001). Distrust is psychologically different from trust, and tends to increase more easily than decrease (Josang, 2001), so the mathematics must reflect that.

Fusion can be automated although that is difficult for many of the clues. Automation has been achieved in some specialized applications, notably programs that detect possible credit-card fraud, and "intrusion-detection systems" for protecting computers and networks by noticing when suspicious behavior is present (Proctor, 2002).

RESPONDING TO DECEPTION

Serious online crimes like fraud can be prosecuted in courts. For less serious matters, virtual communities are societies, and societies can establish their own rules and laws for behavior of their members. Members who engage in disruptive or damaging forms of deception can have privileges revoked, including automatically as by "killfiles" for ignoring messages of certain people. Less serious forms of deception can often be effectively punished by ignoring it or ostracizing the perpetrator just as with real-world communities; this is effective against "trolls", people deceiving to be provocative (Ravia, 2004). In moderated newsgroups, the moderator can delete postings they consider to be deceptive and/or disruptive. On the other hand, deception involving unfair exploitation is often best handled by exposure and publicity, like that of "shills" or people deceptively advancing their personal financial interests.

In all these cases, some investigation is required to justify punishment. Computer forensics techniques (Prosise & Mandia, 2000) may help determine the employment of a newsgroup shill, who started a libelous rumor, or how and by whom a file was damaged. Private-investigator techniques help to determine the identity of a disruptive or masquerading member in a newsgroup like comparing aliases against directories, Web sites, and other newsgroups; and false identities can be detected by linguistic quirks of the masquerader (Ravia, 2004).

FUTURE TRENDS

Technology is making deception easier in virtual communities, and cyberspace is becoming more representative of traditional societies in its degree of deception. While detection methods are not systematically used today, the increasing problems will force more extensive use of them. To counteract identity deception and other forms of fakery, we will see more use of online "signatures" or "certificates" for identifying people, either formal (as with cryptography), or informal (as by code phrases (Donath, 1998)). We will also see more methods from computer forensics investigations like those that collect records of the

same person from different communities or network resources to see patterns of misuse or criminal activity.

CONCLUSION

Many clues are available to detect online deception. So although it is more difficult than detecting deception in face-to-face interactions, tools are available, some of which are automated. If honesty is important in an online setting, they are many ways to improve its likelihood.

REFERENCES

Barber, R. S., and Kim. J. (2001). Belief revision process based on trust: agents evaluating reputation of information sources. In Falcone, R., Singh, M., and Tan, Y.-H. (eds.), *Trust in Cyber-Societies*, LNAI 2246 (Berlin: Springer-Verlag), 73-82.
Bell, J. B., & Whaley, B. (1991). *Cheating*. New York: Transaction Publishing.
Burgoon, J., Stoner, G., Bonito, J., & Dunbar, N. (2003). Trust and deception in mediated communication. *Proc. 36th Hawaii Intl. Conf. on System Sciences*, Honolulu, HI, 44-54.
Carofiglio, V., de Rosis, F., & Castelfranchi, C. (2001). Ascribing and weighting beliefs in deceptive information exchanges. *Proc. Conf. on User Modeling*, 222-224.
Eckman, P. (2001). *Telling lies: clues to deceit in the marketplace, politics, and marriage*. New York: Norton.
Eckman, P., & Frank, M. (1993). Lies that fail. In Lewis, M, & Saarni, C. (Eds.), *Lying and deception in everyday life* (pp. 184-200). New York: Guilford Press.
Ford, C. (1996). *Lies! Lies!! Lies!!! The psychology of deceit*. Washington, DC: American Psychiatric Press.
George, J., & Carlson, J. (1999). Group support systems and deceptive communications. *Proc. 32nd Hawaii Intl. Conf. on System Sciences*, Maui, HI.
Hausman, C. (2000). *Lies we live by: defeating double-talk and deception in advertising, politics, and the media.* New York: Routledge.
Heuer, R. (1982). Cognitive factors in deception and counterdeception. In Daniel, D., & Herbig, K. (Eds.), *Strategic Military Deception* (pp. 31-69). New York: Pergamon.
Joinson, A., & Dietz-Uhler, B. (2002). Explanations for perpetrations of and reactions to deception in a virtual community. *Social Science Computer Review, 20* (3), 275-289.
Josang, A. (2001, June). A logic for uncertain probabilities. *Intl. Journal. of Uncertainty, Fuzziness, and Knowledge-Based Systems*, *9* (3), 279-311.
Kaza, S., Murthy, S., & Hu, G. (2003, October). Identification of deliberately doctored text documents using frequent keyword chain (FKC) model. *IEEE Intl. Conf. on Information Reuse and Integration*, 398-405.
MacVittie, L. (2002). Online fraud detection takes diligence. *Network Computing, 13* (4), 80-83.
Miller, G. R., & Stiff, J. B. (1993) *Deceptive Communications*, Newbury Park, UK: Sage Publications.
Proctor, P. (2001). *Practical intrusion detection handbook*. Upper Saddle River, NJ: Prentice-Hall PTR.
Prosise, C., & Mandia, K. (2001). *Incident response*. New York: McGraw-Hill Osborne Media.
Ravia, F. (2004). Trolling lore. Retrieved November 23, 2004 from www.searchlores.org/trolls.htm.
Rowe, N. (2004, December). Designing good deceptions in defense of information systems. *Computer Security Applications Conference*, Tucson, AZ, 418-427.
Wang, G., Chen, H., & Akabashch, H. (2004, March). Automatically detecting deceptive criminal identities. *Communications of the ACM, 47* (3), 71-76.
Yu, B., & Singh, M. (2003, July). Detecting deception in reputation management. *Proc. Conf. Autonomous Agents and Multi-Agent Systems,* Melbourne, AU, 73-80.
Zhou, L, & Zhang, D. (2004, January). Can online behavior reveal deceivers? -- an exploratory investigation

of deception in Instant Messaging. *37th Hawaii Intl. Conf. on System Sciences*, 22-30.

TERMS

bandwidth: Amount of data transmitted per unit time.
cognitive: Psychological phenomena relating to thinking processes as opposed to senses or movement.
cue: A clue to a psychological phenomenon, often nonverbal.
data fusion: Combining evidence for a conclusion from multiple sources of information.
fraud: Criminal deception leading to unjust enrichment.
intrusion-detection system: Software for detecting when suspicious behavior occurs on a computer or network.
IP address: Code numbers designating the computer attached to a network.
killfile: In newsgroups, a list of email names you do not want to read messages from.
polygraph: Electronic device used for measuring human-body parameters in the hope (never proven) of detecting deception.
signature, electronic: A code used to confirm the identity of someone.

ACKNOWLEDGEMENT

This work was supported by the National Science Foundation under the Cyber Trust program.