Emerging Social Media Threats: Technology and Policy Perspectives

R. Chandramouli
Department of Electrical and Computer Engineering
Stevens Institute of Technology
Hoboken, NJ, USA
mouli@stevens.edu

Abstract— **Traditional cyber threats or attacks have targeted information and communication infrastructure that usually result in economic loss. Typically, launching these attacks requires an advanced skill level. Governments around the world have a good understanding of these threats and therefore have put in place many policies to deal with them.**

The rapid growth of social media is giving rise to new types of threats that spill over from the cyber world to real-life. These threats profoundly alter the psychological, social and cultural dynamics of vulnerable social media users. Also, it is becoming increasingly easy even for an average user to exploit social media for malicious purposes. Organizations and governments are finding it difficult to accurately detect, identify, predict, and prevent the malicious exploitation of social media. Quantifying the socio-psychological effect of social media vulnerabilities is another major challenge. Due to these reasons there is a lack of policies to deal with this issue. In this paper, we discuss several challenges in this emerging area, from technologies to policies.

I. Introduction

The Internet is evolving into a medium that is beyond just web search. Social networking, micro blogging, etc. are some of the next generation services that have gained prominence. Users of these services have realtime two-way interaction (e.g., Facebook [1], Myspace [2], Twitter [3]) as well as non real-time communication (e.g., Craigslist [4]).

Facebook [1] has more than 500 million active users, 50% of whom are typically active in any given day. An average user has 130 friends and people spend over 700 billion minutes per month on the site [1]. Facebook is a global phenomenon in that about 70% of its users are outside the United States. According to [5] teen males with ages between 13-17 years are the fastest growing group in Facebook. There are over 4.9 million female Facebook users in the 13-17 years range and about 3.7 million are males between 13-17 years. 54% of teens with ages between 13-14 years have a Facebook page and spend about 19.9 hours online every week. A darker side of the exponential growth of social media is the negative impact it has had on the society at large. Consider the following real-life cases that exemplify this serious problem:

- According to [6], Facebook is leading the way in teenage organized crime and cyber-bullying. This includes harassment, organizing attacks, etc. One of the reasons for this is the ease with which Facebook can be accessed from anywhere using smartphones and other mobile devices.
- In the “Myspace mom” case [7] a 49-year old woman along with two others created a false profile and pretended to be a 16-year-old boy in Myspace. Using this fictitious user account they then sent flirtatious messages to a teenage neighbor Megan Meier and dumped her later. Megan hanged herself in her bedroom closet.
- Douglas Steven French, age 36, is a convicted Portland-area sex offender. He posted a false advertisement on Craigslist for “16 to 19-year-old models” under the title Premier Modeling. A 17-year-old girl victim responded to his advertisement and they started meeting on a regular basis. He allegedly provided the girl with alcohol and illegal drugs and was later arrested. See [8] for details about this and other crimes related to Craigslist.

We broadly categorize Internet based threats as: (a) *hostile intent* and (b) *hostile attack*. An user with hostile intent (e.g., the Premier Modeling case) typically targets or exploits other users’ psychological or emotional state of mind that may ultimately threaten their own physical security or others in the social group. Since hostile intent could be subtle and takes myriad of forms it is challenging to detect it. *How does hostile intent manifest itself on social media?* Is it possible to create a psychological internet profile of a user to differentiate between hostile vs. friendly intent? Clearly finding an answer to this question cuts across ideas and tools from multiple disciplines including cognitive psychology, data mining, digital forensics, network monitoring and national and international governmental policies. Unlike hostile intent, a hostile attack (e.g., denial of service attack) leaves signatures that can be measured and therefore can be detected with current technologies. Hostile attacks on the Internet typically target infrastructure such as web sites, servers, etc.

978-0-615-51608-0/11 ©2011 EWI

> Hello,
> I am looking for caring and honest person to watch my 3 years old daughter while I work, during our stay in the state and a friend introduced craiglist.com to me, I was searching on Childcare Babysitter, when your Ad post pumped up as a Childcare and Babysitter, am interested in your service as my babysitter/Childcare.. My Little daughters is (Mellina) , I would like you to watch over her while i am at work in your Location , I reside in United Kingdom,I will be coming to the States in about two weeks time , Cos I just got a contract with Boss Perfume fashion industry down there, I work as a model and the contract will last for a month, the duration of our service wil be from 10am to 5pm, monday thru friday, I would like to know if she will be taken care of in your childcare centre/Home, or the hotel room where i would be lodging, It's Okay by me, any of this ways.
> My little daughter (Mellina) is 3 years old, I will be waiting patiently for your email indicating the cost of everything,I mean the price for your service,how long you will be available to take care of her,how much u charge per week. she is just three years old as i have said earlier and i will be glad to update you and make the payment in advance to show you how serious i am. So please get back to me as soon as you can if you will be able to handle her and take very good care of her for me.
> Email me back Your's faithfully

Fig. 1. Example of a deceptive email [4].

The paper is organized as follows. In Section II we outline the technological challenges and approaches to detect emerging social media enable threats. Section III discusses some related policy issues. Concluding remarks are provided in Section IV.

II. TECHNOLOGICAL CHALLENGES AND OPPORTUNITIES

In this section we describe some research challenges and opportunities to detect hostile intent in social media. Specifically, we identify deceptive behavior as one indicator of hostile intent and discuss methods to detect deception from text data generated in the context of social media communications.

A. Deception

Deception is defined as the manipulation of a message to cause a false impression or conclusion [9]. Fig. 1 shows an example of a real scam email sent to a Craigslist user. We identify various types of deception in social media, including the following:

1) *Impersonation*: An user creates a false user profile and pretends to be someone else. This could be in the form of falsifying user name, personality, gender[10], age, etc.
2) *Message manipulation:* Messages posted in a user's "wall" in Facebook or tweets in Twitter are intentionally manipulated to provide a false impression. This could include spreading malicious rumors, false propaganda, etc.
3) *Coded language:* To hide the true intent coded language is used so that only a subset of the population in a social group is able to decipher the original meaning of the message. This can be exploited by terrorists for covert communication in social media.

Note that social media is still dominated by text content. In text based media, users with hostile intent often create stories based on imagined experiences or attitudes to hide their true intent. Thus, deception usually precedes a hostile act. But, presenting convincing deceptive stories requires cognitive resources [11] which means deceivers cannot completely hide their true state of mind. Psychology suggests that one's state of mind, such as physical/mental health and emotions, can be gauged by the words they use [12]. Thus, even for trained deceivers, their state of mind may unknowingly influence the type of words they use. However, it is known that human beings have a poor ability to detect deception. Also note that deceptive behavior in face-to-face communication is sufficiently different from Internet based deception. In face-to-face communication we have access to non-verbal cues such as body language, realtime adaptation of stories, etc. These cues are unavailable in Internet based communication.

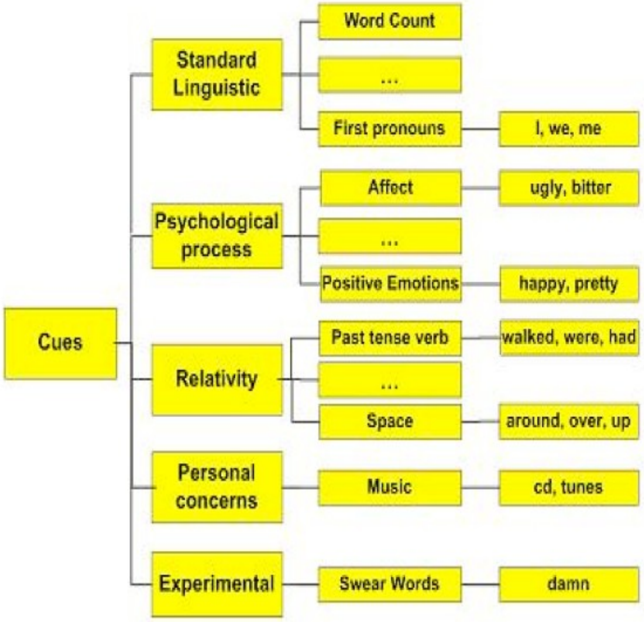

Fig. 2. Psycho-linguistic cues based on LIWC.

B. *Psycho-linguistic Modeling*

Since deception is an indicator of hostile intent we need to develop technologies to detect it. The first step in this process is to identify accurate linguistic deception indicators and model them.

In [12] the following observations have been made about some psycho-linguistic cues that indicate deception in text:

- Fewer first-person pronouns are used as an attempt to dissociate themselves from their words
- Fewer exclusive words are used to keep the deceptive story simple
- Frequency of negative emotions words increase may be due to guilt
- Frequency of active verbs increase as an act of distraction

Therefore to automatically extract linguistic cues from social media based text content we can use software tools such as the Linguistic Inquiry and Word Count (LIWC) [13]. Using LIWC, for each text content, up to 88 output variables can be computed. Some of these variables shown in Fig. 2 include information about the linguistic style, structural composition elements and the frequencies of different linguistic categories.

A major research issue then is to rank these 88 variables as strong, medium and weak indicators of deception. Are there other psycho-linguistic variables that are better indicators? This question needs additional research. Another challenge is to develop an understanding of the intricate relationship between language and culture within the context of deception in social media. A psycho-linguistic theory developed for the English language may not be accurate for other languages. Can machine translators be used in multi-lingual deception? These are important questions that need to be addressed while developing technological solutions to detect hostile intent in social media.

C. *Statistical Modeling, Analysis and Monitoring*

Once the psycho-linguistic cues that are strong indicators of hostile intent are identified the next step is to statistically model and analyze or classify them.

1) *Modeling Issues:* Some of the challenges here include the following:
   - *Scale*: Models have to scale with the text message size. For example, statistical modeling and analysis of tweets [14] are more challenging than regular text messages such as emails. This is because tweets are limited to 140 characters and therefore are very short unlike typical emails.
   - *Robustness*: The collected text data may be incomplete or erroneous due to user privacy settings. The model must be able to withstand such uncertainties and still provide an acceptable level of accuracy.
   - *Incremental modeling*: Once a model has been computed using a training data set then it must be flexible enough to adapt to new data, e.g., new types of deceptive behaviors. But, instead of having to model the new data all over again, the modeling technique must be able to change incrementally. This is important since social media networks generate huge amounts of data every minute.
2) *Statistical Analysis:* There are several opportunities for statistical data analysis in social media security threat mitigation. In the context of deception detection we identify the following problems:
   - *Avatar detection:* We need data mining techniques to detect different avatars of the same user. For example, one user may have different user/screen names on different network sites to conceal their true identity . Then is it possible to identify the user based on correlations in the psycho-linguistic fingerprints derived from his/her blogs, chat sessions, wall posting, tweets, etc.?
   - *Sentiment extraction:* Statistical techniques to accurately estimate user sentiments or moods from what they write on social media is in its nascency. Sentiment extraction from text can be used to predict the psychological state of mind (e.g., happy, depressed, etc.). This can then be used to prevent or mitigate security threats.
3) *Social Network Monitoring:* For statistical modeling, analysis and detection of social media threats it is imperative that the social network be monitored (subjected to privacy policies). A data collection architecture has to be designed. This leads to the following issues:
   - *Coverage*: The number of social media users and the volume of information they generate is enormous and continues to grow rapidly. Heterogeneity in terms of geographical location, language, culture, different government policies etc. further complicate social network monitoring. Monitoring all the social media users and their every conversation is practically impossible. Therefore deciding which users to monitor (based on their psycho-linguistic profiles) to mitigates threats is a challenge.
   - *Metrics*: What kind of metrics are appropriate to search, identify and rank social media threats? Performance of statistical threat detectors are measured by two types of error probabilities, namely, miss and false alarm probability. Miss probability is the percentage of threats that went undetected and false alarm probability is the percentage of falsely detecting a threat. What are acceptable levels for these two types of decision errors? What are the psychological, legal and economic costs incurred due to these errors? Answers to these questions are not obvious.

III. Policy Perspective

A fundamental question is: how much privacy is enough? Social media companies have to balance the need for user privacy with law enforcement needs. For example, the U.S. Department of Justice wanted information about some Twitter users regarding the WikiLeaks case [15]. Privacy policies of companies like Facebook and Twitter have evolved over time, especially as it relates to sharing information with law enforcement agencies. The Electronic Communications Privacy Act (ECPA) in the U.S. sets the parameters for what type of information can be collected in electronic media.

Facebook, in its 2010 policy guide states that falsifying profile information will lead to disabling of the user account. But, checking the veracity of the profile information for each of the several hundred million users is an impossible task. Craigslist allows its users to flag a posting into one of several categories, if they choose to. One of these categories is spam.

Many states in the U.S. have cyberstalking, cyberharassmentand cyberbullying laws [16]. Cyberstalking is the use of the Internet, email or other electronic communications to stalk. This is considered the most dangerous of the three threats. Sanctions range from misdemeanors to felonies. Cyberharassment usually refers harassments using email, blogs or social media sites. Cyberbullying refers to bullying among minors within a school context using the Internet. The sanctions for cyberbullying range from school/parent interventions to misdemeanors and felonies with detention, suspension, and expulsion.

While policies and practices have been defined in the U.S. and many other countries, this is not true globally. This may be because of low Internet penetration, blocking of all or many social media sites, close government monitoring of Internet user activities, etc. But with the growth of cellular networks Internet access is becoming more prevalent and cheaper in many countries. This means that in a few years countries that do not have well defined social media security policies have to rethink this issue to fill the policy gap.

IV. Conclusion

Social media security and threat mitigation leads to several technological and policy issues. Many of these are complex issues since they involve users of different age groups, languages, economic backgrounds, cultures, educational levels, etc. Moreover, the technical and policy challenges have a profound effect on each other. Some approaches that we have outlined in this paper are just a beginning. A holistic solution to this problem can be found only through inter-disciplinary thinking that cuts across academic research, government policies, economics, and human networking.

.References

[1] [Online]. Available: http://www.facebook.com

[2] [Online]. Available: http://www.myspace.com

[3] [Online]. Available: http://www.twitter.com

[4] [Online]. Available: http://www.craigslist.com

[5] [Online]. Available: http://blog.traciscampbell.com/2010/07/19/facebooktweens-teens-statistics-worth-knowing/

[6] [Online]. Available: http://www.examiner.com/internet-in-desmoines/facebook-bullying-and-teen-organized-crime

[7] [Online]. Available: http://www.foxnews.com/story/0,2933,457784,00.html

[8] [Online]. Available: http://craigscrimelist.org/tag/childmolestation/

[9] J. Burgoon and D. Buller, “Interpersonal deception: Ill effects of deceit on perceived communication and nonverbal behavior dynamics,” *Journal of Nonverbal Behavior*, vol. 18, no. 2, pp. 155–184, 1994.

[10] N. Cheng, R. Chandramouli, and K. Subbalakshmi, “Gender identification from e-mails,” *IEEE Symposium on Computational Intelligence and Data Mining*, 2008.

[11] J. Richards and J. Gross, “Emotion regulation and memory: The cognitive costs of keeping one’s cool,” *Journal of Personality and Social Psychology*, vol. 79, pp. 410–424, 2000.

[12] M. Newman, M. L., J. Pennebaker, D. Berry, and J. Richards, “Lying words: Predicting deception from linguistic styles,” *Personality and Social Psychology Bulletin*, vol. 29, pp. 665– 675, 2003.

[13] “Linguistic inquiry and word count.” [Online]. Available: http://www.liwc.net/

[14] X. Chen, R. Chandramouli, and K. Subbalakshmi, “Scam detection in twitter,” *SIAM Text Mining Workshop*, April 2011.

[15] [Online]. Available: http://www.eff.org/deeplinks/2011/01/socialmedia-and-law-enforcement-who-gets-what

[16] [Online]. Available: http://www.ncsl.org/default.aspx?tabid=13495