ON CYBERSECURITY, CROWDSOURCING, AND SOCIAL CYBER-ATTACK

By **Rebecca Goolsby, Ph.D.**, Office of Naval Research

Social media is responsible for much positive change in the world. But these new tools can be used by bad actors to foment strife and undermine stability, as seen during violent incidents in the Assam state of northeast India in July 2012. Cybersecurity efforts must take into account the growing potential for cyber-attack using social media, where hoax messages are incorporated into a stream of otherwise legitimate messages, and understand how quickly mobile apps and text services can disseminate false information. Authorities and volunteers must develop a healthy skepticism about information derived from these systems and new research and tools are needed to facilitate the self-policing of social media.

Civil unrest and social media have become indelibly linked with the advent of the Arab Spring. And yet, the history of civil unrest and the expansion of public expression have a much longer history. As Nate Silver observed, the invention of the printing press itself helped to spur civil conflicts, religious wars, and ethnic contests:

> The instinctual shortcut that we take when we have "too much information" is to engage with it selectively, picking out what we like and ignoring the remainder, making allies of those who have made the same choices and enemies of the rest. …Martin Luther's Ninety-five Theses were not that radical; similar sentiments had been debated many times over. What was revolutionary…is that Luther's theses…were reproduced at least three hundred thousand times by Gutenberg's printing press.[1]

Social media's amazing capability to spread information at extremely high volumes and velocities is

in many ways an exponential magnification of the printing press because of the media's two-way, interactive quality. Social media are "cooler" than McLuhan's "cool" medium of television.[2] Their multi-way communications vectors are even more engaging than television as they leverage social trust and social connectivity. But like Guttenberg's printing press, which provided a springboard into the Enlightenment—and the expansion of knowledge, truth, and rationality—social media also provide avenues for darker intentions and possibilities.

Social Cyber-Attack in Assam, India

Ethnic tensions in India between Hindu and Muslim populations have been an enduring source of conflict since the early Muslim expansion into the Sind in the 700s. These tensions became more evident to the Western world after the partition of Pakistan in 1947. The increase in size of Muslim populations in India continues to provide issues of conflict. Recently, the conflicts expanded into the digital world. It began with a brutal attack in the northeastern Indian state of Assam in late July of 2012. Hindu members of the indigenous Bodo tribe clashed with Bengali-speaking Muslim settlers and migrants. More than 300,000 refugees were relocated to a heavily guarded camp; their houses were burned, and 78 people were reported dead.[3,4]

In urban centers away from Assam, Muslims from the area also reported being targeted for attack based on their distinctive facial features. In cities as large as Bangalore, as well as smaller urban areas such as Pune, Chennai, and Mysore, riots and smaller clashes broke out.[5] And then, something slightly new was added.

Texts and photographs warning of renewed attacks began circulating in urban centers, leading many young people to flee these centers. Train stations were swamped, and refugee camps swelled. Urban residents received on their telephones texts and doctored photographs of an alleged riot in progress, an event that was all too possible given recent upheavals in the region. The photographs, some from friends and other people known to the people who received them, were actually subtly altered photographs of other riots.[6] Several arrests were made, including one of an individual alleged to have sent more than 20,000 messages.[7] People receiving these messages often sent them on to friends or published tweets about them. Although no riot was actually taking place in Assam, the circulation of the photographs helped to convince people that such an upheaval was indeed happening and resulted in a panicked mass exodus.[8] Thus, Twitter and other social networks were implicated in facilitating the spread of rumors that led to panic.

Videos, websites, and tweets uncovered by a research team at Arizona State University (but now blocked by YouTube, Facebook and Internet service providers) showed gruesome images of mass deaths in other contexts (e.g., an earthquake in Tibet in 2008, a cyclone in Myanmar also in 2008).* Many messages passed off these images as deaths from the recent attacks on Muslims in Assam.

---

* This is original research conducted by the author, Huan Liu of Arizona State University, and his graduate students Pritam Gundecha, Shamanth Kumar, and Fred Morstatter. The author is greatly indebted to Liu, Gundecha, Kumar and Morstatter, who not only researched the background for this discussion, but also conducted the original datamining to find the Twitter, YouTube, and other social media references related to the Assam social cyber-attack.

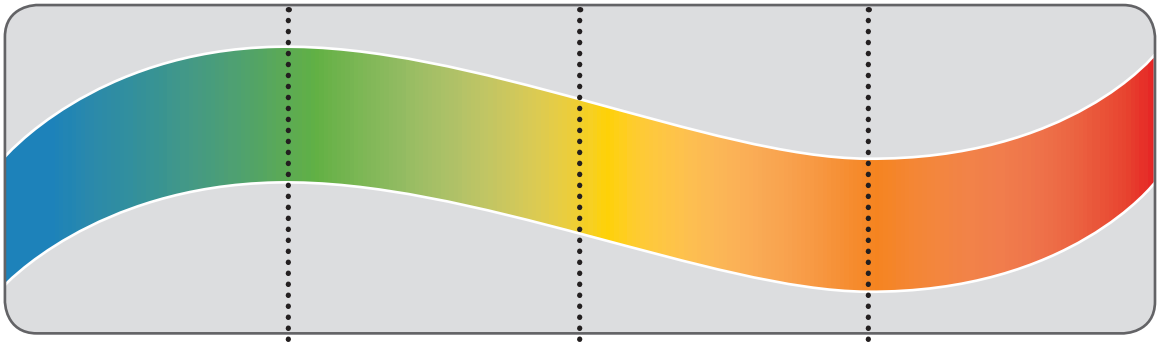
At this writing, it is unclear how organized these perpetrators were, given the distributed nature of the attack. In some ways, one might speculate that this was the work of an ad hoc "hate crowd"—a hastily formed, informal network of friends and strangers.[9]

Passed around by socially trusted networks, these messages, were disseminated rapidly and in bulk, as evidenced by the individual alleged to have sent 20,000 messages during the unrest.[10] Amid this inundation, news organizations were unaware of the scope of the problem and did not provide sufficient coverage at the time to help squelch the misinformation. Civil authorities were likewise late in discovering the panic, not recognizing it until it had already passed the point of no return.

Although the positive social impacts of social media are well-known, social media have already entered a new chapter. They are becoming sources of inflammatory information and disinformation. Like the conflicts sparked by Guttenberg's invention, significant real-world impacts of a more troubled nature are beginning to arise from them. The creation of hoaxes, hate speech, and other attempts at crowd manipulation and exploitation reveal the darker side of the social media phenomenon; the targeted "social-cyber attack" is rapidly coming of age.

Information sharing has a spectrum of social impact, ranging from the "clean" and humanitarian efforts, such as disaster relief, coordination

FIGURE 1. Uses of digital communication for crowd effects



CRISIS RESPONSE

Disaster relief

Humanitarian assistance

Crisis monitoring

PROMOTES:

Order and discourse

Cooperative behavior

Information sharing

COMMUNITY DIALOGUE

Anti-propaganda

Rumor squelch

Community outreach

PROMOTES:

Discussion expansion

Spread of verifiable information

INFLUENCE

Propaganda

Rebellion cry

Hate messages

PROMOTES:

Special point of view

Bandwagon effects

Conflict and argument

Mass protests

SOCIAL CYBER-ATTACK

Crowd manipulation

Hysteria propagation

PROMOTES:

Chaotic mass behavior

Escalation of rumor

Confusion, panic, and violence

of humanitarian activities, and the promulgation of truthful information, to darker topics, such as hate speech and crowd manipulation (Figure 1). The cleanest type of messaging seeks to bolster social order, relieve suffering, and promote positive social bonds. Counter-messaging—the refutation of bad information, lies, and mischief—is a bit grey, colored by the propaganda that it seeks to refute. Its objectives are a bit more biased, promoting a particular "story" against the claims of others who seek to use deceit or misrepresentation to disseminate their views. Propaganda of every kind attempts to rally the base or influence others, making it greyer still, with the objective of swaying others toward a particular agenda. Hoaxes and scare-mongering campaigns seek to subvert public order, then generate and exploit the resulting chaos so as to benefit or gain in some way. Their creation is something of a black art.

Social cyber-attack is not a new phenomenon. In the late 1980s and early 1990s, USENET groups, forums, and bulletin boards suffered the problem of "flame wars" instigated through "trolling," new social behaviors and attacks that were designed to destroy nascent virtual communities by stirring up conflict. "Trolls" commonly attempted to reveal hidden divergences among community members. For example, a troll would pose provocative questions or post extreme viewpoints and opinions on controversial issues, sometimes to manipulate

Definitions

Botnets (also known as zombie armies or zombie botnets): A group of computers controlled by a piece of software that forces the computers to obey the commands of the software, such as sending emails. The user whose computer is infected is often unaware that a botnet has taken over portions of the input/output processes of the computer.

Distributed denial of service attack (also known as DDOS): An attack, often on a website or servers of an agency, in which millions of requests overwhelm the input/output processes to service the requests, leading to server failure. This attack is usually distributed over thousands or tens of thousands of computers driven by "botnet" technology.

"Cyber-hoodlums": A slang term for hackers, who form loose communities trying to exploit or manipulate social media crowds.

MMS: Multimedia service, video or photo sent through short message service, available only through smart phones.

Robot Twitter accounts: Accounts that are created by computer code to send the same or highly similar messages onto the Twitter platform, polluting the information stream with messages that appear to come from many different people (rather than just one person). Typical uses are to inflate support for an idea or person and inflate trending statistics.

SMS: Short message service, a text message delivered from a mobile phone to other mobile phones.

All social media users need to develop a healthy skepticism about the messages that they receive, learn to check sources, and refine their skills of discernment.

the crowd to attack the troll and other times to promote the posting of divergent viewpoints that would lead members of the community to attack one another. The resulting argument over something inconsequential to the main discourse of the community would destroy the trust and consensus built up over the main issues. Thus, a troll might post pornography onto a discussion about free speech and let the community members fight about whether to throw the person out of the group or a troll might rail against vaccines in a group discussing infant health. Hot topics, deep visceral concerns, false assertions, and irrelevant tangents were the hallmark of these altercations, which became known as "flame wars." They destroyed many a small virtual community and damaged many a large one.

Trolls enjoyed the attention, the excitement of the war, and the ability to manipulate the sentiments and emotions of the crowd. Like arsonists, trolls would appear only to start these wars again and again, primarily for the entertainment value of watching the community turn upon itself. The Internet maxim (or meme) "Do not feed the trolls" refers to the only solution for trolling: ignoring the trolls and paying no heed whatever to their attempts to bait a conversational trap. Today, trolls are considered mildly amusing, as Internet communities have become more inured to their attempts to cause chaos.

Cyber-Attack in the Age of Social Media

The evolution of social media has created new opportunities for "trolling," bringing it into the real world of ethnic division and social unrest.

Social cyber-attacks are of two kinds: (1) pre-meditated, which are designed to create an excited signal in a social network, often under false pretenses, so as to benefit from the chaos and upheaval; and (2) opportunistic, which take advantage of an existing excited social network signal and, by manipulating it through various means, derive benefit. In the Assam case, the attack was both pre-mediated and opportunistic. The "signal"—news of the arrest of the riot organizers from several weeks earlier—would spark the bulk of expected messages. People interested in the problems in Assam would be expected to send messages about this news, so the signal would be a "natural" excited one.

By jumping into an expected or "natural" excited signal, the hoax messages—the "false" signal—could hide among the stream of natural messages and be accepted, perhaps even if it was difficult to determine *precisely* the origin of those messages. This represents a sophisticated sort of attack. It is clear from the Assam incident that malefactors are becoming sufficiently familiar with the social dynamics of virtual communities to successfully launch such attacks. More experimentation is no doubt under way by myriad actors with intentions to exploit this

medium, and the cell phone, the weak security link in these attacks, is the least defended device in the mix.

Morphed images have been used to create fear, disgust, and chaos before, particularly in the Middle East, where one group or another alters images to suggest that police brutality, mob violence, or other acts occurred in a particular place at a given time (when, in fact, the pictures were from a different place, time, and situation). Savvy social media enthusiasts know how to use "reverse image search" to find the true origins of photos and are often skeptical of images found on the Internet. New entrants into the world of social media are not aware of these capabilities, however, and they can be readily fooled, as was the case in Assam.

The use of multimedia service (MMS) mass texting, where images are sent to cell phones rather than through the Internet directly is an interesting addition to these social media attacks. Details are scant, but it is possible that social media played a role in the social cyber-attack in Assam; social media enthusiasts often link their phones and emails to their accounts without availing themselves of privacy shields and sometimes allow third-party apps (software programs) to access their information. Further, many people allow these apps to receive information about people in their social networks. Some apps allow the app creators to push content, which would be a good way to seed a snowball of interconnecting links. The research conducted at Arizona State University was unclear about whether any of this actually happened in Assam; however, smart phone security is vulnerable to malware and other kinds of attacks.[11] This is admittedly simple conjecture, but it is a disturbing possibility.

Through social media, hate speech proliferates, reaching hordes of interested mischief makers who are comfortably anonymous and hard to track. Social cyber-attack as a means to bully and trick others, as well as to sow uncertainty in tense situations, is not going to go away. It is no longer a matter of finding "the one person behind all this," as malcontents, trolls, and malicious actors are legion. In addition, they are connected in loose cyber-communities and technically capable. Moreover, Robot Twitter accounts and other zombie systems can extend the reach of individuals. When these techniques are shared among like-minded anarchists and zealots, the capability of a small minority is magnified. They are thus able to pump up their apparent numbers and spread around the risk of getting caught.

The potential for oppressive governments to use cyber-attack techniques is certainly an issue that deserves further study. The perpetrators and organizers of these "hate crowds" can remain anonymous, hiding behind both real people and automated zombie botnets, which minimize the risk of discovery. As a result, the capability to develop an extensive cadre of cooperating "cyber-hoodlums" is growing while the discovery and repercussion become more difficult. In the Assam case, for example, authorities arrested a number of Indian citizens, but other perpetrators based outside of the country remained beyond the law's reach.

Conclusion

Just as the printing press was a key factor in the development of the Enlightenment, social media show promise of producing positive and meaningful change in the world. However, the darker aspects of their capabilities should not be ignored. Disaster responders, humanitarian relief workers, and a new breed of digital volunteers who provide technical support during and after a crisis may

need to be particularly cautious. They often interact in social media streams that are high in volume and velocity, with great uncertainty and chaos—the very characteristics that attract malefactors.

All social media users need to develop a healthy skepticism about the messages that they receive, learn to check sources, and refine their skills of discernment. New technologies that assist users to better protect themselves are one part of the solution. Social media watchdog groups also play a role in the education of the user community, spreading the word about hoaxes, scams, and attacks very quickly and widely—if only users will pay attention.

It would be more profitable to try to discover clever ways of determining who benefits from these social cyber-attacks and how they benefit, both politically and economically, and connect the dots from beneficiary to crowd. Further, policymakers should consider new national and international sanctions against botnet operators whose creations result in serious conflict.

New technologies to facilitate self-policing of social media vulnerabilities, such as improved reverse search engines and new techniques and methods to discover scams, hoaxes, and exploitations, would also be of great benefit. A crowd that is able to carefully self-police itself is the absolute best defense. Government cannot be everywhere, but the crowd certainly is.

Notes

1. Nate Silver, *The Signal and the Noise* (New York: Penguin Press, 2012), 4.
2. Marshall McLuhan, *Understanding Media: The Extensions of Man* (1964). Available at http://beforebefore.net/80f/s11/media/mcluhan.pdf. Accessed January 9, 2013.
3. BBC, "How the Assam Conflict Creates a Threat to All India" (August 20, 2012). Available at http://www.bbc.co.uk/news/world-asia-india-19315546. Accessed December 27, 2012.
4. Jim Yardley, "Panic Seizes India as a Region's Strife Radiates," *The New York Times* (August 17, 2012). Available at http://www.nytimes.com/2012/08/18/world/asia/panic-radiates-from-indian-state-of-assam.html?pagewanted=all&_r=0. Accessed December 27, 2012.
5. Zee News, "Northeast Issue: Exodus Subsides in Bangalore, No Let Up in Chennai, Pune." Available at http://zeenews.india.com/news/nation/north-east-peoples-exodus-continues_794432.html. Accessed January 2, 2013.
6. The Times of India, "Doctored MMS Clip Provoked Attackers: Cops." Available at http://articles.timesofindia.indiatimes.com/2012-08-14/india/33200173_1_mms-clip-kondhwa-attacks-on-northeast-students. Accessed January 3, 2013.
7. Northeast Today, "Man Sends More Than 20,000 Hate Messages, Held in Bangalore" (August 22, 2012). Available at http://www.northeasttoday.in/national-news/man-sends-more-than-20000-hate-messages-held-in-bangalore/. Accessed January 9, 2013.
8. Pritam Gundecha, Zhuo Feng, Huan Liu, Pritam Gundecha, "Searching for Provenance Data in Social Media" (Paper to be presented at the Social Computing and Behavior Prediction Conference, April 2013).
9. Indo-Asian News Network, "India Raises Social Media Misuse with Pakistan, Situation Normal in the South: Roundup" (August 19, 2012). Available at http://www.india-forums.com/news/national/431157-india-raises-social-media-misuse-with-pakistan-situation-normal.htm. Accessed January 2, 2013.
10. Northeast Today, "Man sends more than 20,000 hate messages, held in Bangalore." http://www.northeasttoday.in/national-news/man-sends-more-than-20000-hate-messages-held-in-bangalore/. August 22,2012. Accessed 9 Jan 2013.
11. Amy Gahran, "Mobile Phone Security: What Are the Risks?"CNN Tech (June 7, 2011). Available at http://articles.cnn.com/2011-06-17/tech/mobile.security.gahran_1_android-app-android-phone-apple-s-app-store?_s=PM:TECH. Accessed January 9, 2013.

DR. REBECCA GOOLSBY is a Program Officer with the Office of Naval Research. An anthropologist by training, she directs research in computational social science and social networks.

Email: **Rebecca.Goolsby@navy.mil**
Website: **http://www.onr.navy.mil/en/Media-Center/Fact-Sheets/Social-Cultural-Behavioral-Sciences.aspx**

The views expressed in this policy memo are those of the author and do not reflect the opinions of the Office of Naval Research.

EDITORS: Lea Shanley and Aaron Lovell, Wilson Center

THE WILSON CENTER, chartered by Congress as the official memorial to President Woodrow Wilson, is the nation's key non-partisan policy forum for tackling global issues through independent research and open dialogue to inform actionable ideas for Congress, the Administration and the broader policy community.

Science and Technology Innovation Program

One Woodrow Wilson Plaza
1300 Pennsylvania Ave. NW
Washington, DC 20004-3027
(202) 691-4000

COMMONS LAB advances research and non-partisan policy analysis on emerging technologies that facilitate collaborative, science-based and citizen-driven decision-making. New tools like social media and crowdsourcing methods are empowering average people to monitor their environment, collectively generate actionable scientific data, and support disaster response.

http://CommonsLab.wilsoncenter.org

http://bit.ly/CommonsLabVideo

@STIPCommonsLab

/CommonsLab

http://bit.ly/CommonsLabReports