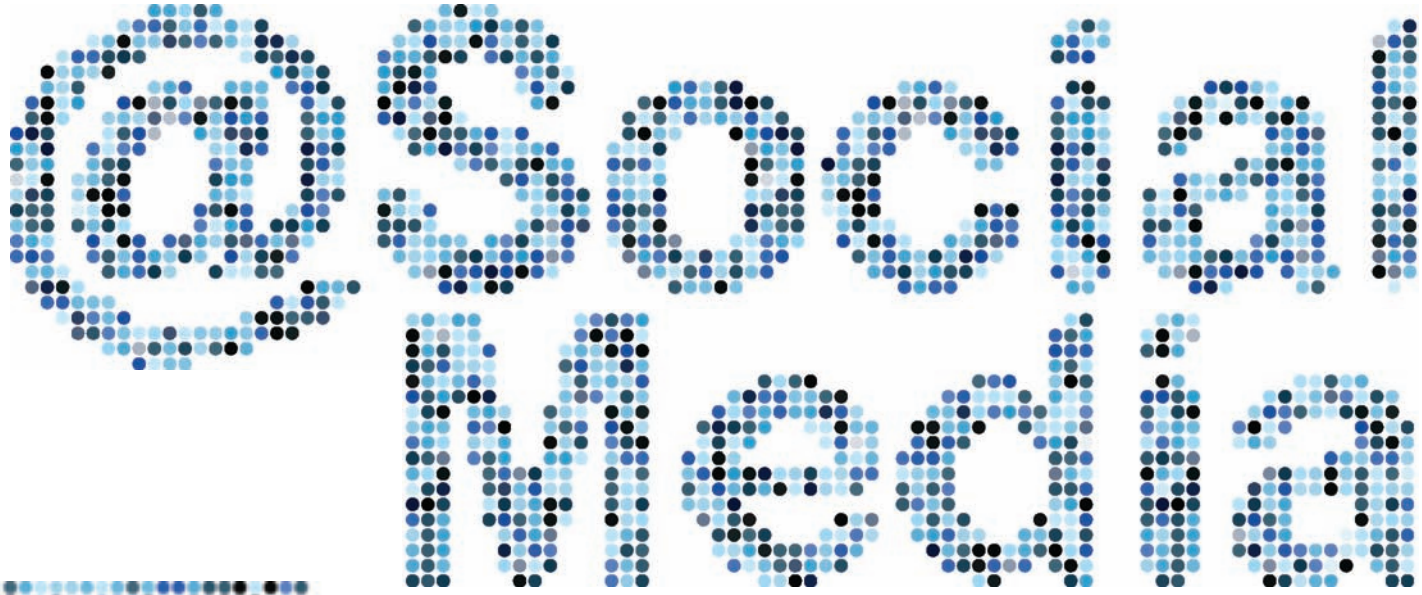

Using Twitter and Facebook to understand how networks—and people—behave

For most people, social network websites such as Facebook and Twitter provide easy real-time updates on the activities and opinions of friends and celebrities. But for computer scientists, these sites provide large data sets that enable new research possibilities. Two professors at McCormick are using social networking to establish new platforms for Internet communication and security.

Alok Choudhary. *Photo by Sally Ryan.*

#DATAMINING: ALOK CHOUDHARY

Which celebrities had the most influence on Twitter was the furthest thing from Alok Choudhary's mind when he started his work with data mining—the process of gleaning patterns from large data sets. Initially, he set out to analyze research on protein interactions, which are the basis for every process in cells. A database of protein interactions could help improve the understanding of diseases and lead to possible treatments.

To do this, Choudhary, the John G. Searle Professor and chair of electrical engineering and computer science, began searching through thousands of articles in scientific journals and developed data mining software that could identify specific outcomes from the papers. The software used such processes as high-dimensional clustering, classification, and deviation detection, which businesses have used to understand customers' preferences and help enable recommendation tools (like the one used by Amazon.com).

But Choudhary and PhD student Ramanathan Narayanan wanted to extend their research beyond a static network into real-time data mining. What could an evolving network tell them about the nature of news and information today? So they turned to the ultimate real-time data behemoth, Twitter, where users offer up their diatribes, opinions, news flashes, and announcements in 140 characters or less.

"The tweets are so small and so numerous that they make an easily accessible large database of real-time information," Choudhary says. "We wanted to know how people followed opinions on important topics." Choudhary and Narayanan developed software for pattern recognition that could identify and measure sentiment—which would come in handy to determine whether tweets were positive or negative—and they set up a website: pulseofthetweeters.com. They began using data mining, social network analysis, and sentiment analysis to determine top influencers on Twitter and how tweeters felt about top trends.

At pulseofthetweeters.com, users can search for a topic (such as Justin Bieber), judge user opinions (83 percent positive), and see which tweeter is most influential (fellow pop star Demi Lovato). But trends and influencers aren't always that straightforward: For the Chilean miner rescue, talk show host Conan O'Brien was the most influential tweeter. For the Haitian earthquake, the biggest influencer was singer Adam Lambert.

"We can determine who is helping to shape opinions," Choudhary says. "Our real-time analysis can determine public sentiment on any given topic quickly, because 15 minutes later another topic or event might be a top trend."

Choudhary and his colleagues have developed a specialized algorithm to rank influence; to qualify as an influencer, a tweeter must actively tweet about the topic and have a following that subsequently tweets about it, too. "Influence determines the value of communication," he says. That, he says, could prove especially valuable in marketing. "If a company wanted to target its communication, it could figure out who its influencers are and see how its message or brand is received."

Earlier this year, Choudhary and his team released a list of the top Twitter trends of 2010. Lady Gaga, Mel Gibson, and Justin Bieber were among the most tweeted-about people, and national news organizations NPR News, the *New York Times,* Time.com, CNN, and the *Wall Street Journal* were often top influencers on politics and world affairs. Top sports topics included Lebron James, Brett Favre, Michael Vick, Wimbledon, and Manchester United, and in the trending topic #thankful, in which people tweet about what they are thankful for, Bieber was the most influential.

The website has gotten major media attention, but Choudhary says he is most interested in continuing to develop new techniques in data mining—including a research project that determines the impact of climate change around the world. "We didn't start out expecting to be looking at these things," he says. "We wanted to discover patterns in data, and popular culture is a major part of Twitter and the data we mine."

Visit McCormick's Facebook page:
facebook.com/mccormickengineering

Follow McCormick on Twitter:
twitter.com/nu_mccormick

STATUS: SPAM

Someone has a crush on you. Get free ringtones. Check out this cool video. It's easy: Just click here.

For the 500 million Facebook users worldwide, these wall posts are a common stain on the fabric of social networking: malicious spam that redirects users to sites that ask for personal information or install viruses onto unsuspecting users' computers. But just how prevalent are these posts? How do they work? Where do they come from?

Those are the questions that Yan Chen, associate professor of electrical engineering and computer science, and his collaborators at the University of California, Santa Barbara, set out to answer when they conducted the first study that quantifies the extent of malicious content and compromised accounts in a large online social network.

Analyzing more than 187 million Facebook wall posts, the team found 200,000 malicious messages with embedded URLs, more than 70 percent of which linked to phishing sites that ask users for their passwords or other personal information. Their research results could help programmers design techniques to automatically detect online social spam.

Chen's team used data gathered from the Facebook walls of 3.5 million user accounts. By "crawling" the user sites of eight regional networks (Egypt, Los Angeles, London, Monterey Bay, New York City, Russia, Santa Barbara, and Sweden) from April to June of 2009, the researchers were able to download users' publicly available wall posts from the last year and a half. They narrowed their search to the messages containing URLs—about 2 million.

Researchers then sorted the posts based on destination URL or strong textual similarity, with the assumption that similar spam posts would come from the same spam campaign. They found about 200,000 posts were embedded with malicious URLS. They analyzed the posts for two distinguishing features: distributed coverage and "bursty" nature. Distributed coverage refers to the number of users who send wall posts. "Bursty" describes the small time intervals between consecutive wall posts; most spam campaigns involve coordinated action by many accounts.

Using third-party tools to assess the malice of URLs in their dataset, Chen and his team found that approximately 70 percent of malicious wall posts direct the victim to a phishing site. About 35 percent of malicious wall posts direct victims to sites laced with viruses. Chen also found that the vast majority of those wall posts came from existing, hacked accounts. "It's much easier to create a fake account, but attackers who hack into existing accounts can have a higher rate of success because there is a level of trust among real friends," he said.

The most popular ploy was a message that said someone had a crush on the user. Tempting, no doubt, but Chen urges Facebook users to stop and think before they click. "Don't trust a suspicious wall post even if it's from your friends," he says. "And alert your friends immediately."

Yan Chen

The attacker usually has control of the account for a short period of time—about 80 percent of malicious accounts are active for less than an hour. Most malicious wall posts are posted at 3 a.m.—when most users are asleep.

So how can online social networks fight back against spammers? Facebook has started trying to eliminate fake accounts by launching a new feature allowing users to reject friend requests as "don't know." Facebook collects this information to identify and remove spammers.

Spamming on Facebook highlights one of the major communication issues of our age: because it's so difficult to categorize and understand the huge volume of traffic on the Internet, it's difficult to design security procedures to protect users. Chen and his research group are on the front lines of this battle. They have previously analyzed attacking strategies of spammers and designed intrusion detection and prevention systems for networks. His group takes two approaches to improving the reliability and security of the Internet: designing network-based intrusion detection and prevention systems to combat large-scale attacks and creating new protocols and architecture to improve the reliability and security of the Internet.

"A lot of security breaches can be stopped by patches that were released months or years ago, but users do not pay attention," says Chen. "We want to develop a networking-based approach that we can deploy at routers and gateways so we can protect users automatically."

Real-time detection of spammers that compromise social networking accounts is still far off, however. "We need much more research," Chen says. "There's no good solution yet, and attackers are becoming a lot more powerful. They have their own mature society—their own forums, banks, and markets—and they often prey on security breaches that have already been fixed with patches. We need more people to be aware of security problems in order to stop them."

That sort of large-scale impact is ultimately what drives a computer scientist like Chen, who says, "I ultimately hope that this research can have a direct impact on society's well-being." M Emily Ayshford