Journal of Financial Crime

Social media content and sentiment analysis on consumer security breaches

Jianqiang Hao, Hongying Dai,

Social media content and sentiment analysis on consumer security breaches

Jianqiang Hao
First National Bank of Omaha, Omaha, Nebraska, USA, and
Hongying Dai
Department of Health Services and Outcomes Research, Children's Mercy Hospital, Kansas City, Missouri, USA

Abstract

Purpose – Security breaches have been arising issues that cast a large amount of financial losses and social problems to society and people. Little is known about how social media could be used a surveillance tool to track messages related to security breaches. This paper aims to fill the gap by proposing a framework in studying the social media surveillance on security breaches along with an empirical study to shed light on public attitudes and concerns.

Design/methodology/approach – In this study, the authors propose a framework for real-time monitoring of public perception to security breach events using social media metadata. Then, an empirical study was conducted on a sample of 1,13,340 related tweets collected in August 2015 on Twitter. By text mining a large number of unstructured, real-time information, the authors extracted topics, opinions and knowledge about security breaches from the general public. The time series analysis suggests significant trends for multiple topics and the results from sentiment analysis show a significant difference among topics.

Findings – The study confirms that social media monitoring provides a supplementary tool for the traditional surveys which are costly and time-consuming to track security breaches. Sentiment score and impact factors are good predictors of real-time public opinions and attitudes to security breaches. Unusual patterns/events of security breaches can be detected in the early stage, which could prevent further destruction by raising public awareness.

Research limitations/implications – The sample data were collected from a short period of time on Twitter. Future study could extend the research to a longer period of time or expand key words search to observe the sentiment trend, especially before and after large security breaches, and to track various topics across time.

Practical implications – The findings could be useful to inform public policy and guide companies responding to consumer security breaches in shaping public perception.

Originality/value – This study is the first of its kind to undertake the analysis of social media (Twitter) content and sentiment on public perception to security breaches.

Keywords Sentiment analysis, Security breach, Social media monitoring MM

Paper type Research paper

Introduction

Data breaches have been widespread recently, and The Associated Press (2014) ranked security breach among the top three business stories. Target, a multinational retailer,

Journal of Financial Crime
Vol. 23 No. 4, 2016
pp. 855-869

1359-0790
DOI 10.1108/JFC-01-2016-0001

reported over 40 million credit and debit cards hacked just before holiday shopping season in 2013 and an additional theft of 70 million personal records later, followed by breaches at Home Depot, Michaels, Sally Beauty Supply, Neiman Marcus, AOL, eBay, P.F. Chang's Chinese Bistro, Supervalu, Dairy Queen, Jimmy Johns, Kmart, Staples, Bebe Stores and, most recently, Ashley Madison. Breaches exposing consumers' payment data (such as credit card account numbers and bank account numbers) rose from 127 cases in 2009 to 217 cases in 2013, and breaches exposing consumers' personally identifiable information [social security numbers (SSNs), passwords, medical records and tax returns] rose from 475 cases in 2009 to 841 cases in 2013. Breaches exposing online IDs have seen the largest increase, from 22 cases in 2009 to 342 cases in 2013 (Sullivan, 2014). The Federal Trade Commission's (2014) Consumer Sentinel Network (CSN) received over 290,000 complaints related to identity theft in 2013. Government documents/benefits fraud accounted for 34 per cent of reported identity theft, followed by credit card fraud (17 per cent), phone or utilities fraud (14 per cent) and bank fraud (8 per cent).

These security breaches are very costly. For instance, the Target data breach alone has cost financial institutions nearly $500m to replace card and incur other expenses (Berger, 2014). Data breaches and identity theft often cause significant stress and financial loss to consumers. Harrell and Langton (2013) reported that approximately 16.6 million persons were victims of one or more incidents of identity thefts in 2012, which represents 7 per cent of all USA. Residents aged 16 years and older. About 14 per cent of identity theft victims experienced some out-of-pocket losses, and victims reported spending an average of about 9 h clearing up the issues. The direct and indirect losses from identity theft reached a total of $24.7bn in 2012. The Wall Street Journal (Armour, 2015) also reported that medical identity theft, in which someone fraudulently use others' identity to bill for medical services, has been rising from 1.4 million cases in 2009 to 2.4 million cases in 2014. The medical identity fraud has caused overwhelming stress to victims ranging from lost health insurance, unpaid medical bills and diminished credit score to inability to review the medical records because of the law protecting the privacy of the identity thief.

It would be extremely beneficial to conduct real-time analyses, closely monitor public sentiment about various security breaches and further raise awareness about data security. Traditional ways to collect opinion data and disseminate information are slow and expensive. The rising social networks provide researchers and investigators real-time "big data" to detect and monitor consumers' sentiment, opinions and responses to security breaches. There are over 500 million tweets sent at Twitter each day, and some 23 per cent of online adults/19 per cent of entire adult population currently use Twitter. About 36 per cent of Twitter users visit the site daily and another 24 per cent visit a few days a week. Over one billion people use Facebook actively each month and, 71 per cent of adult internet users currently use Facebook (Duggan *et al.*, 2014). Recent advances in "Big Data Analytics" have also made the social media data mining possible (Murthy, 2013; Duggan *et al.*, 2014).

People who have experienced data breach or identity theft might engage in social networking conversations through Twitter or Facebook or simply express their opinions about data breaches or identity theft by tweeting or retweeting related topics. Twitter data and other social media data have been widely used in public opinion polls, for instance, Tumasjan *et al.* (2010) used Twitter sentiment to predict election results.

Twitter messages were also used in many other fields, such as stock market prediction (Bollen *et al.*, 2011), disaster management (Xiao *et al.*, 2015) and students' learning experience (Chen *et al.*, 2014).

The recent use of social networking data (i.e. Twitter and Facebook) in health care has shown that these data could outperform data collected through traditional technologies. For example, Ginsberg *et al.* (2009) show that they can use Google search queries to detect influenza earlier than the Centers for Disease Control and Prevention. Ram *et al.* (2015) used social networking data along with data collected from electronic medical records and air quality sensors and successfully predicted the number of asthma Emergency Department visits with approximately 70 per cent precision. Wong *et al.* (2015) estimated consumer's sentiment related to Affordable Care Act (Obamacare) using social media data from Twitter and found that the estimated sentiment score is positively associated with enrollment in Obamacare at the state level. All these studies show that social media data could provide an effective way to detect, monitor and predict social events.

Little is known about how people view security breaches on social media platforms. To our knowledge, this is the first study conducted to utilize social media as a measure of consumers' sentiment to security breach. This paper seeks to fill the gap by proposing a framework for studying the social media surveillance on security breaches along with an empirical study to shed light on public attitude and concerns reacting to security breaches. The goal is to assess consumer affairs by investigating:

- how consumers react to security breach events in the social media sites;
- how to gauge consumers' sentiment and opinions effectively through text mining; and
- whether we can real-time monitor security breach and analyze relevant topics along with their popularity and sentiment.

Figure 1 summarizes the proposed framework for social media surveillance on security breach.

Data and method

Twitter data collection

Twitter, the fast-growing social networking company, has over 400 million active users across the world and over 500 million tweets per day. Tweets are short messages that are no more than 140 characters in length. It is estimated that Twitter was used by more men than women and by more young adults (18-49 years old) than older adults (50-65 years old). Black, non-hispanic or hispanic are more likely to use Twitter than white, non-hispanic (Duggan *et al.*, 2014). Twitter users can be followed by other twitter users, allowing others to receive and share tweets ("retweets) and, thus, distributing the tweets to a large audience. Twitter users tend to express in-the-moment feelings that reflect users' current experience, making Twitter as an ideal candidate for a real-time surveillance tool.

Twitter provides public access to its data through an advanced programming interface, which yields a random sample of approximately 1 per cent of all tweets in near-real time (Murthy, 2013). Statistical software, R, was used to collect tweet information and analyze the unstructured data in this study:

- We used "twitteR" package to retrieve tweet information related to "data breach" or "identity theft" in August 2015.

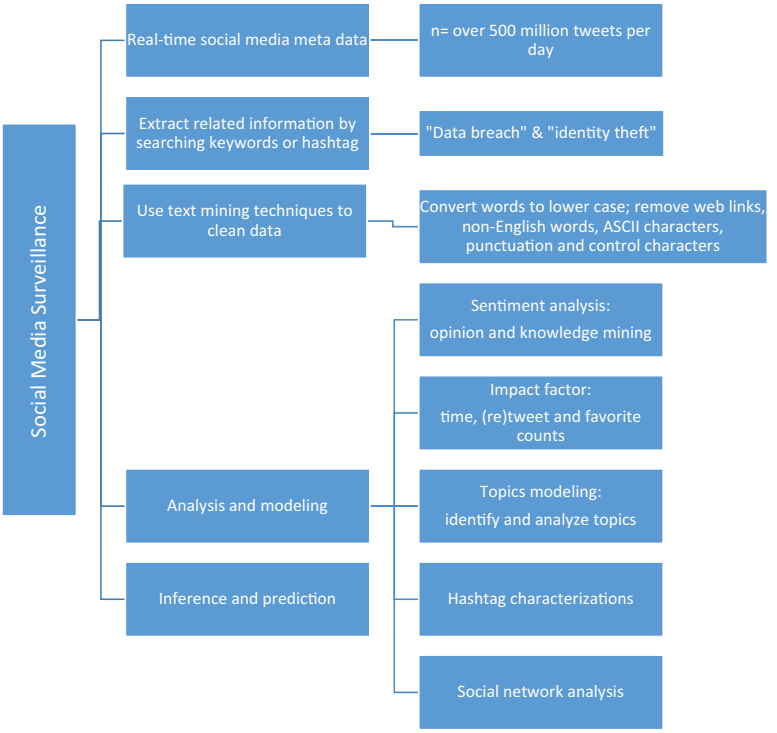

Figure 1.
Proposed framework and scope of social media surveillance on security breach

- The metadata includes text; time when the tweet was sent; user's language; whether a message is a retweet, favorite count and retweet count; and the geocode of latitude and longitude if users choose to enable this feature.
- As tweets are unstructured, free-text information, we further applied text mining techniques to clean the tweets: all words were converted to lower case; non-English words and ASCII characters along with Web links (URLs), punctuation and control characters were removed by using R's regex-driven global substitute.

Sentiment analysis

Sentiment analysis is also called opinion mining (Liu, 2012) and is used to analyze people's opinions, emotions or attitude toward entities such as products, services, events or topics. Unstructured data are common in the sentiment analysis, as people usually express their opinions through words instead of numbers. The sentiment can be represented by:

$$S_{ijk} = f(e_i, t_{ij}, h_{ijk})$$

where e_i represents the i^{th} event:

T_{ij} represents the j^{th} topic of the i^{th} event; and

h_{ijk} represents opinions from the k^{th} holder regarding the j^{th} topic of the i^{th} event.

We start with a pre-established opinion lexicon compiled by Liu, (2012), which includes a list of positive and negative English sentiment words. The data set includes 2,914 positive words and 4,914 negative words. As words may have different meanings in different fields, a modification to the general lexicon is necessary to accurately capture the opinions in the context. For instance, Loughran and McDonald (2011) used a large sample of 10-Ks during 1994-2008 and found that words list developed from other disciplines misclassifies the common words in finance text. Almost three-fourths of negative word counts in 10-K filings based on the Harvard dictionary are typically not negative in a financial context, such as "depreciation", "liability" or "foreign". As a result, Loughran and McDonald developed an alternative words list to better reflect tone in financial context of 10-K filings.

Table I lists the top ten positive and negative words in the tweets related to security breach. Table II presents ten tweet examples with positive or negative sentiment. Clearly, the common opinion lexicon is not sufficient enough to capture the sentiment in security breach. Most of the tweets classified with positive sentiment were false positive, and the true opinions expressed by the users were mostly neutral or even slightly negative. For example, the words "safe", "avid", "good", "protect" or "peace" are typically considered as positive when people express their opinions, but they are more likely to be neutral in the context in selected examples. In addition, "theft" and "breach" are considered as negative words in the general lexicon, and we have to reclassify them in our study as neutral. As a result, we focus on the negative sentiment score in this study to remove the potential false positive from positive words.

To verify the sentiment to security breach, two human judgers reviewed 1,000 randomly selected tweets and further adjusted the opinion lexicon. As a result, we achieved an accuracy of 93 per cent in the sentiment expressed by related tweets.

Results

Sentiment and impact factors

We collected a total of 113,340 tweets with key words "data breach" or "identity theft" from August 1 to August 31, 2015. We randomly reviewed a sample of 1,000 tweets to assess whether the tweets were related to data breach or identity theft (>99 per cent) and whether they were in English (>99 per cent).

Positive words	Count	(%)	Negative words	Count	(%)
Famous	8211	7.2	Fraud	3375	3.0
Trust	4129	3.6	Stolen	3136	2.8
Protection	3327	2.9	Sued	2583	2.3
Protect	2545	2.2	Victims	1920	1.7
Top	1376	1.2	Hack	1893	1.7
Free	1169	1.0	Risk	1710	1.5
Profound	1145	1.0	Threats	1180	1.0
Patient	735	0.6	Catastrophic	1137	1.0
Safe	704	0.6	Criminal	1095	1.0
Leads	674	0.6	Dump	1091	1.0

Table I. Top 10 positive and negative words

No.	Tweets	Positive words	Negative words	Overall sentiment	Negative sentiment
1	Keep your account *safe* from identity fraud & theft by routinely changing passwords. http://t.co/DEKUqP0O2 http://t.co/cExz4Sp445	Safe	Fraud	0	−1
2	*Avid* Life Media faces $578m class action over data breach https://t.co/cvugmCHxUy	Avid		1	0
3	RT @MassBar: How can I *protect* myself from identity theft? Learn how by visiting our #ConsumerLaw Resource Center http://t.co/iQv6XqPCLI	Protect		1	0
4	"ARM YOURSELF WITH ARMOURSTIXx" buy today and save yourself the *headache* of a data breach or *worse* identity theft http://t.co/FRwRC5r4Rp		Headache, worse	−2	−2
5	RT @ID_Analytics: Identity thieves use stolen personal data to get medical treatment and you get *stuck* with the bill http://t.co/zpZPw9W3Yx		Stolen, stuck	−2	−2
6	if you were *outraged* by the @Target data breach but ur making *excuses* for @HillaryClinton - you've gone full *retard.*		Outraged, excuses, retard	−3	−3
7	Want vacation *peace* of mind? Check your credit card's exp. date, and make sure you have identity theft coverage http://t.co/YYJut0BWox	Peace		1	0
8	Ashley Madison data breach may *undermine* privacy for everyone http://t.co/DRwb8Halvn		Undermine	−1	−1
9	Why we need to mitigate the *severe* consequences of identity theft in the US: http://t.co/8eihY7PanH #clientprotection		Severe	−1	−1
10	*Good* #crisiscomms insights from @AndrewPelosi - How Marketers Can *Protect* Their Reputation During a Data Breach http://t.co/MqjZBj7v4G	Good, protect		2	0

Table II.
Sample tweets with positive or negative words

Of 113,533 tweets, 66,301 (58 per cent) tweets contained keyword "data breach" and 47,232 (42 per cent) tweets contained keyword "identity theft". The negative sentiment score for data breach was significantly lower than that of identity theft ($p < 0.0001$).

In addition to sentiment analysis, we also compare the impact factors of tweets calculated using Twitter metadata. Table III summarizes the sentiment score along with other user-level metadata characteristics for data breach and identity theft. These tweets were posted by 53,163 unique users, and the average tweets per user were lower for data breach compared with identity theft (1.9 vs 2.2). However, people were more likely to retweet data breach-related events than identity theft-related events (35.7 vs 28.5). There were over 3.71 million tweets or retweets (371 million given we collected approximately 1 per cent of total tweets) related to data breach or identity theft in the study period.

The negative sentiment score is consistent with the results from the 2012 National Crime Victimization Survey (Harrell and Langton, 2013), which reported that about 36 per cent of identity theft victims reported moderate or several emotional distress as a result of the incident.

Topics

We further classified the tweets into five categories based on the interested topics. The first four topics are related to the latest data breach events, and the last two topics are tied to the common identity theft:

(1) *Carphone*: It refers to the data breach on the Carphone data warehouse located in UK. It was reported that personal details of up to 2.4 million of its customers may have been accessed in a cyber-attack and the attack was made in public on August 8, 2015.

(2) *Ashley Madison*: It refers to the data hack on the commercial website for people seeking extramarital affairs. Over 60 gigabytes worth of data were confirmed to be valid and released on various websites on August 18, 2015.

(3) *Target*: It announced on August 18, 2015, that Target Corp. has reached a settlement up to $67m with Visa over the massive data breach in 2013.

(4) *Medical*: Medical identity theft continues to be one of major concerns and is rising over the time.

(5) *Credit card*: Credit card identity theft ranked second in the CSN and poses significantly financial risk to consumers.

Figure 2 depicts the percentage of tweets from each of these five topics by day. Clearly, the tweets related to the first four latest events were tied to specific dates. For instance, 66.2 per cent of collected tweets on August 8, 2015, were related to "Carphone" data

Event	Negative sentiment score	Total # of tweets	Total # of retweets	Total # of twitter users	Average # of tweets per user	Average # of retweets per tweet
Data breach	−47.3 (SD 67.8)	66,301	2,366,357	35,138	1.9	35.7
Identity theft	−52.5 (SD 79.2)	47,232	1,347,546	21,343	2.2	28.5
Total	−49.5 (SD 72.8)	113,533	3,713,596	53,163	2.1	32.7
p-value	$p < 0.0001$					

Table III. Impact factors and sentiment of tweets regarding data breach and identity theft

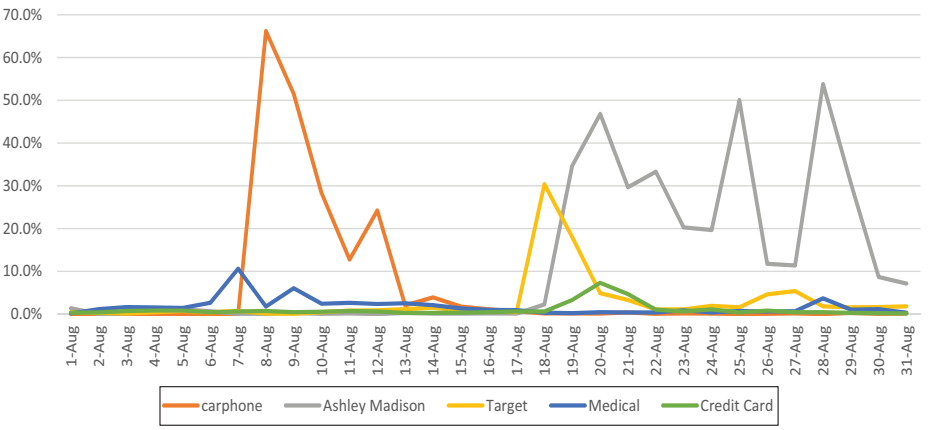

Figure 2.
Per cent of tweets by date for selected topics

warehouse hack, 46.8 per cent of collected tweets on August 20, 2015, included the key word "Ashley Madison" and 30.4 per cent of collected tweets on August 18, 2015, were related to "Target". The Twitter messages highly coincided with the data breach events. The trends of two common topics, "medical" and "credit card", were relatively flat over the time.

The tweets from different topics also exhibited diverse negative words when people conveyed their opinions. Figures 3 depicts the word cloud of negative words from these five topics. The minimum frequency is five, and the font size is proportional to the frequency of words within each category. The word clouds show different high-frequency words among topics. The tweets related to "Ashley Madison" and "Carphone" have large number of negative words, whereas tweets related to "Credit card" have relatively limited negative words, indicating that consumers might have different attitudes to these topics.

Table IV summarizes the impact factors and sentiments from user-level metadata among five selected topics. The public had the lowest sentiment to "Medical" (−102.8), followed by "Ashley Madison" (−69.3) and identity theft related to "credit card" (−42.2). The attack on Carphone data warehouse had the second highest sentiment score of −18.5, next to the topic of "Target" (−13.5). The sentiment scores are significantly different among these topics with a $p < 0.0001$. The sentiment scores suggest that people had very diverse opinions among different topics. They were very concerned on the medical identity theft but less concerned on the hack of Carphone data warehouse. "Target" had the highest sentiment score, possibly because the breach happened over 18 months ago and consumers are less sensitive to such information.

The average tweets per users were slightly different among topics, and the topics of "Carphone" and "Ashley Madison" had the highest tweets per user (1.5 tweets per user), whereas the "Credit card" topic had the lowest (1.2 tweets per user). The average retweets per user were very different: the "Medical" topic had the highest number of retweets per tweet (318.3), and "Ashley Madison" topic had the second highest number of retweets (100.8) per tweet, whereas the "Target" and "Credit" topics had the lowest number of retweets per user (6.2 and 5.2, respectively), suggesting that the public is more interested in the eye-catching news or popular topics.

Notes: (a) carphone; (b) ashley madison; (c) target; (d) medical; (e) credit card

Figure 3.
Word cloud of negative words from five topics

Table IV.
Impact factors and sentiment of tweets for selected topics

Time	Topic	Negative sentiment score	Total # of tweets	Total # of retweets	Total # of twitter users	Average # of tweets per user	Average # of retweets per tweet
Latest event	Carphone	−18.5 (SD 52.6)	9,416	122,903	6,196	1.5	13.1
	Ashley Madison	−69.3 (SD 72.7)	17,753	1,790,194	12,175	1.5	100.8
	Target	−13.5 (SD 37.3)	5,089	24,720	3,963	1.3	4.9
Common event	Medical	−102.8 (SD 96.5)	1,887	600,546	1,503	1.3	318.3
	Credit card	−42.2 (SD 60.0)	1,351	5,628	1,089	1.2	4.2
	p-value	$p < 0.0001$					

Hashtags

A hashtag is a type of label to tag a specific content on social media networks and make it easier for users to search for such information. A hash character (#) is placed in front of a word for users to create and use hashtags. The health-care hashtag project has compiled a list of 7,896 hashtags related to health care from 3,071 contributors to make the use of social media data easy and more accessible for the health-care community (Symplur, 2015).

Table V summarizes the top ten hashtags from each of these five topics and the corresponding frequency; "#news" ranked the second in the first three topics of "Carphone", "Ashley Madison" and "Target"; "#infosec" and "#security" showed up in all five topics; and "#data", "#breach", "#tech" and "cybersecurity" also appeared in multiple topics, suggesting that there might have common social communities to discuss such events and disseminate relevant information.

Social network analysis

Social network analysis (SNA) can be used to investigate the social structure in terms of nodes (i.e. people) and edges (relationships) that connect them. Social communities are believed to have existed in the virtual world, and SNA is important to identify the interactions of people among the social media. We further performed SNA to explore the associations among hashtags in the social communities. We selected the top hashtags identified from previous section and visualized the social structure for hashtags related to security breach and selected topics. Figure 4(a) presents the social structure from the top 20 hashtags related to security breach, and, clearly, there are three distinct social communities: the first one is tied to data, security and infosec; the second one is related to search or investigation, such as "#google", "#FBI" and "#sutherland"; the third one has one hashtag, "#wocket", is related to information related to smart wallet. Figure 4(b) and (c) depicts the social structure for "Ashley Madison" and "Carphone" topics, and we found there existed different social communities. The SNA also provides insights on how information might be disseminated among social communities. For instance, it looks like that "#AshleyMadison" and "#data" might be the center of information spread for tweets related to "Ashley Madison".

Discussion

This study is the first of its kind to analyze the security breaches by using social media data and provided empirical evidences about how consumers respond to security breach events in the social networks. We collected a sample of 113,340 Twitter messages in August 2015 and found that people were concerned about their identity theft or data breach with a −49.5 sentiment score. Our findings suggest that social media data could provide valuable information in understanding public sentiment to security breach.

We further examined five related topics in our collected data and observed significantly different sentiment scores among these topics. The "medical" topics had the lowest sentiment score, whereas the "Target" and "Carphone" topics had the highest sentiment score, suggesting different public opinions on various data breaches or identity theft events. On the other hand, the "medical" topic had the highest retweets per user, indicating that the public is concerned about the medical related identity theft. The recent "Ashley Madison" topic had the second highest number of retweets per user, validating the potential use of social media sites to monitor the public awareness. The

Table V.
Top ten hashtags from selected topics

	Carphone			Ashley Madison			Target	
Rank	Hashtag	# of tweets	% of tweets	Hashtag	# of tweets	% of tweets	Hashtag	# of tweets	% of tweets
1	#carphonewarehouse	1,060	11.3	#ashleymadison	1,163	6.6	#target	163	3.2
2	*#news*	229	2.4	*#news*	313	1.8	*#news*	114	2.2
3	*#technology*	226	2.4	*#tech*	300	1.7	*#infosec*	112	2.2
4	*#infosec*	208	2.2	*#data*	276	1.6	*#databreach*	111	2.2
5	*#data*	168	1.8	*#ashleymadisonhac*	266	1.5	*#breach*	97	1.9
6	*#breach*	159	1.7	*#infosec*	264	1.5	*#security*	97	1.9
7	*#security*	158	1.7	*#breach*	241	1.4	#visa	94	1.8
8	*#tech*	135	1.4	*#privacy*	219	1.2	#business	69	1.4
9	#carphone	88	0.9	*#security*	218	1.2	*#data*	65	1.3
10	#android	84	0.9	*#technology*	194	1.1	*#cybersecurity*	40	0.8

	Medical debt			Credit card	
Hashtag	# of tweets	% of Tweets	Hashtag	No. of tweets	% of tweets
#medical	82	4.3	*#infosec*	83	6.1
#identity	76	4.0	*#security*	81	6.0
#identity	62	3.3	*#breach*	45	3.3
#provider	46	2.4	#credit	24	1.8
#veterans	42	2.2	*#data*	17	1.3
#medical	35	1.9	*#cybersecurity*	14	1.0
#databreach	33	1.7	#10	12	0.9
#infosec	32	1.7	*#databreach*	12	0.9
#cyberse	27	1.4	#anonymous	11	0.8
#medical	22	1.2	#free	10	0.7

Note: *the hashtags in Italic indicates that they showed up in multiple topics

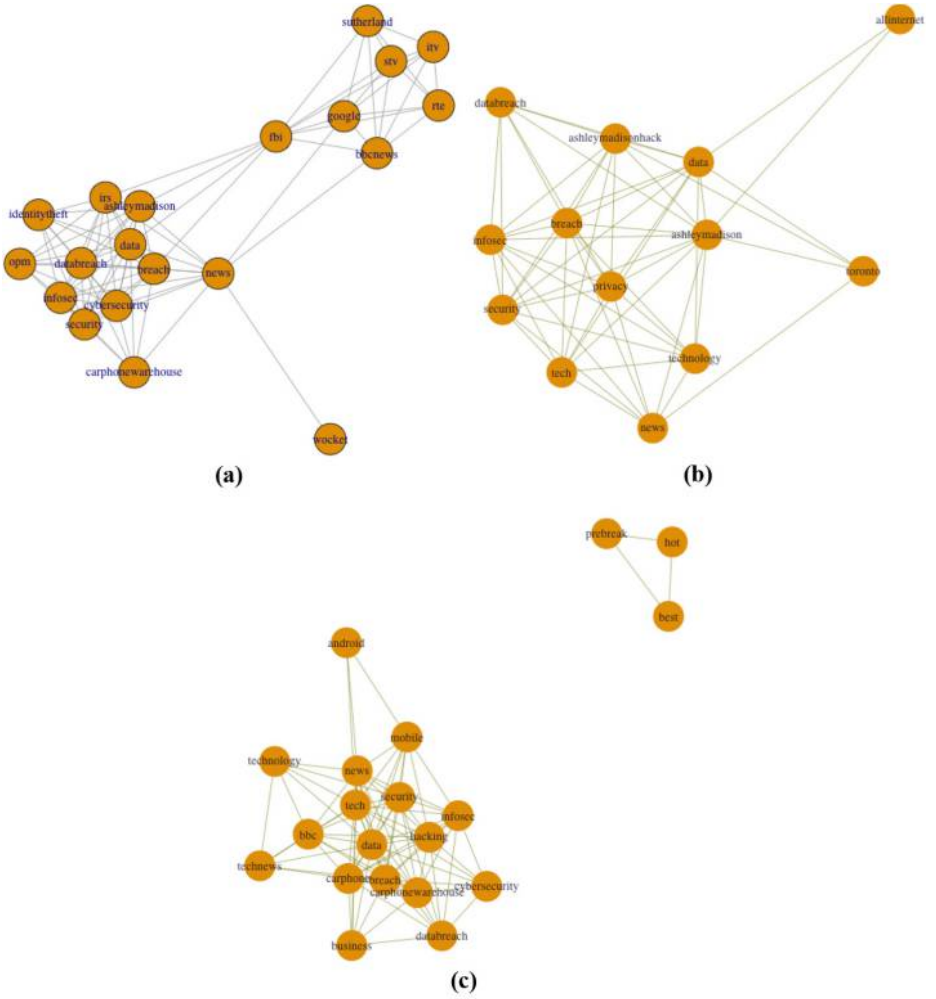

Figure 4.
Social networks of top hashtags

Notes: (a) Top 20 hashtags related to security breach; (b) top 13 hashtags related to ashley madison; (c) top 20 hashtags related to carphone

hashtag analysis found that consumers use multiple common hashtags to disseminate information related to security breach, such as "#infosec", "#security" and "#data". SNA further confirms that there exist distinct social communities in the virtual world. Such information is valuable in consumer affairs to raise awareness and promote information related to security breach.

Our findings are very promising to better understand consumers' sentiment to security breach with the use of social media data. This paper is intended to provide a framework in such studies and call for more researches to expand the use of social media surveillance in security breach. Our findings would be useful to inform public policy and

guide companies responding to consumer security breaches in shaping public perception. Future study could extend the research to a longer period of time or expand key words search to observe the sentiment trend, especially before and after large security breaches, and track various topics across time. Additional research could also focus on whether people's sentiment score could be used to predict the outcome (i.e. financial losses and stress) along with other social economics information and further quantify the social network interactions among social communities (i.e. hashtag).

References

Armour, S. (2015), "How identity theft sticks you with hospital bills", *Rhe Wall Street Journal*.

Berger, D. (2014), "One year after target breach, consumers vulnerable as ever", American Banker no. 179, available at: www.americanbanker.com/bankthink/one-year-after-target-breach-consumers-vulnerable-as-ever-1071750-1.html (accessed 13 September, 2015).

Bollen, J., Mao, H. and Zeng, X.J. (2011), "Twitter mood predicts the stock market", *Journal of Computational Science*, Vol. 2 No. 1, pp. 1-8.

Chen, X., Vorvoreanu, M. and Madhavan, K. (2014), "Mining social media data for understanding students' learning experiences", *IEEE Transactions On Learning Technologies*, Vol. 7 No. 3.

Duggan, M., Ellison, N.B., Lampe, C., Lenhart, A. and Madden, M. (2014), "Social media update 2014", Pew Research Center, Washington, DC.

Federal Trade Commission (2014), "Consumer sentinel network data book for January – December 2013", available at: www.ftc.gov/system/files/documents/reports/consumer-sentinel-network-data-book-january-december-2013/sentinel-cy2013.pdf (accessed 13 September, 2015).

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009), "Detecting influenza epidemics using search engine query data", *Nature*, Vol. 457 No. 7232, pp. 1012-1014, doi: 10.1038/nature07634.

Harrell, E. and Langton, L. (2013), "Victims of identity theft, 2012", NCJ 243779, US Department of Justice, Bureau of Justice Statistics.

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan & Claypool Publishers, San Rafael, CA.

Loughran, T. and McDonald, B. (2011), "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks", *Journal of Finance*, Vol. 66 No. 1, pp. 35-65.

Murthy, D. (2013), *Twitter: Social Communication in the Twitter Age, Digital Media and Society*, Polity, Cambridge.

Ram, S., Zhang, W., Williams, M. and Pengetnze, Y. (2015), "Predicting Asthma-related emergency department visits using big data", *IEEE Journal of Biomedical Health Information*, Vol. 19 No. 4, doi: 10.1109/JBHI.2015.2404829.

Sullivan, R.J. (2014), "Controlling security risk and fraud in payment systems", *Economic Review*, Vol. 99 No. 3, pp. 47-78.

Symplur (2015), "The healthcare hashtag project", available at: www.symplur.com/healthcare-hashtags/ (accessed 13 September, 2015).

The Associated Press (2014), *AP's Biz Editors Rank Top 10 Stories of 2014*, The Associated Press, New York, NY.

Tumasjan, A., Sprenger, T.O., Sandner, P.G. and Welpe, I.M. (2010), "Predicting elections with Twitter: what 140 characters reveal about political sentiment", *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media, Washington, DC*.

Wong, C.A., Sap, M., Schwartz, A., Town, R., Baker, T., Ungar, L. and Merchant, R.M. (2015), "Twitter sentiment predicts affordable care act marketplace enrollment", *Journal of Medical Internet Researc*, Vol. 17 No. 2, p. e51, doi: 10.2196/jmir.3812.

Xiao, Y., Huang, Q. and Wu, K. (2015), "Understanding social media data for disaster management", *Natural Hazards*, Vol. 79 No. 3.

About the authors

Dr Jianqiang Hao is the Managing Director of Customer Data Management at First National Bank. Dr Hao has his PhD in Applied Economics and MS in Statistics, both from the University of Kentucky. He is a CFA (Chartered Financial Analyst) Charterholder. Dr Hao has over 10 years of experience in the areas of big data analytics, social media data mining and credit risk modelling. Dr Hao is also an Adjunct Professor at Bellevue University and teaches economics, finance and risk management courses. Jianqiang Hao is the corresponding author and can be contacted at: jqhao74@yahoo.com

Dr Hongying Dai is an Associate Professor in the Department of Biomedical and Health Informatics at University of Missouri Kansas City, and Senior Biostatistician in the Department of Health Outcomes Research at the Children's Mercy Hospital. She received her PhD in Statistics from University of Kentucky, and her research works focus on statistical epidemiology, social media monitoring, secondary analysis of big data and development of novel statistical methods for medical research.