Social Media Location Intelligence: The Next Privacy Battle - An ArcGIS add-in and Analysis of Geospatial Data Collected from Twitter.com

Weidemann, C. and Swift, J.
Team - UNIGIS@Los Angeles
Spatial Sciences Institute, Spatial Sciences Institute, University of Southern California, Los Angeles CA, USA, Phone: (213) 740-5910, Fax: (213) 740-9687, E-mail: Cdweidem@usc.edu

Abstract
While GIScience professionals are spatially aware and literate to geospatial terminology and options of use, the majority of the general public is not conscious of the location intelligence they disclose when using social media outlets. This article outlines the integration of a unique technical application with GIScience and the subsequent effect location-based data can have on one's personal privacy, security, and web-presence. A innovative new ArcGIS add-in is introduced that captures location data from Twitter, harvests ambient location information, analyzes the captured data to provide personal location intelligence, and visualizes possible areas of interest. In addition, the article examines the results and trends discovered through use of this add-in and how these same techniques can be applied to other social media outlets and analytical intents. Finally this article discusses potential opportunities to educate and inform the general public more about their social media location privacy.

1. Introduction

1.1 Social Media Privacy

Anyone that reads the news, watches T.V., or converses about social media is aware of the privacy concerns that users are exposed to. Social media entities like Facebook, Google+, and MySpace store countless pieces of personal information about their users. Not only do they store the information provided to them through registration and status updates, but also the information users grant them access to through things like internet browser cookies, search history, and even e-mail conversations. All of this information is stored securely in good faith and used to "target" a user with appropriate advertising. Of course social media users should be concerned with much more than just what their social media providers have access to. Lini (2012) found that 35 percent of hiring managers, across a wide array of disciplines, have rejected an applicant based purely on information they found online. They also found that the amount of Human Resource departments now screening new hires through social media searches has risen 38.4 percent since 2008. The screening doesn't end with employers. Thieves use social media as a tool for gathering intelligence and for picking their victims. In 2008 United States vice presidential candidate Sarah Palin had her e-mail account stolen after a thief was able to gather intelligence from her social media outlets and web searches. The National Foundation for Credit Counseling has found that these types of social media tactics and online identity thefts are ground zero for credit card identity thieves (Benda, 2010). Loeffler (2012) argues that privacy laws need to be updated to reflect evolving social media trends to protect against these types of actions. These thefts are made possible through the innocent sharing of information through social media. For example, most Internet users don't stop to think that a mother's maiden name, which is a common secondary security question online, can be gathered from social media relationships. It is also common for users to share, seemingly harmless photographs of their families or homes. To the ill intentioned, this is yet more information that could be used against the social media user.

1.2 Background

Over the last several years extensive research has been done on the need for privacy constraints within social networking (Stefanidis et al., 2011 and Friedland and Sommer, 2010). However very

limited research has even been done on the subcategory of location privacy within social networks. Recent research has been done on spatial data collection from social media and alterative research has been done on location based cyber stalking, but no article has specifically focused on social media location privacy concerns or methods for systematically mining potential location privacy threats. An example of this research is presented by Stefanidis et al., (2011), who proposes system architecture for capturing geospatial information from social media as volunteer geographic information, and presents methods for analyzing social media streams for event "Hot Spots". This work produced evidence that social media analysis can be used as an alert and to monitor real world events. In addition, this study revealed the potential for harmful use of social media data, and also suggested that social media streams be considered as "pseudo" volunteer geographic information. In another study attempting to raise awareness of cybercasing, Friedland and Sommer (2010) scrutinize the activity of using publically available geo-information in conjunction with geo-tagged pictures and videos to infer real world situational awareness for questionable purposes. The authors focus on geo-tagged pictures and video for the source of geospatial information instead of social media outlets. They argue that most submitters of the geo-tagged media are unaware of the location information they're publishing as metadata and suggest steps need to be taken on mobile devices to decrease privacy concerns and increase awareness. Additionally there has also been significant research accomplished concerning opportunities to increase location privacy through computational awareness. For example Krumm (2009) discusses opportunities for situational based restrictions on location data. The author argues that while most people do not seem to comprehend the potential negative consequences of sharing location data, system designers should be responsible and instinctively protect users' privacy through location anonymizing algorithms.

1.3 Social Media Location Privacy

This article argues the idea that relationships and photographs shared through social media should be the least of a user's concerns. Instead social media users should concern themselves with their social location privacy and the potential risk they place themselves in due to that sharing. This study advances this concern by building upon recent research and arguments made by several authors discussed in the background review on this topic, including Stefanidis et al., 2011 and Friedland and Sommer, 2010. The study reported in this article is unique in that it specifically investigates opportunities for location data mining through the Twitter social media network, with the intent of informing users of potential negative consequences realized through innocent location sharing. A new analysis tool built as an Esri ArcGIS add-in is introduced which facilitates the collection of location data from one of the Internet's top social media outlets, Twitter.com, and then interpolation of the location privacy data. This article presents a new way that Geographic Information Systems (GIS) can be utilized to collect and harvest location intelligence from social media and other emerging Internet sensors. While this specific exercise attempts to shed light on the ease with which personal location data can be collected and thus on privacy concerns, these same techniques can be adapted to collect and analyze other data and trends in social media. In turn such value-added information can inform decision makers, educate the masses, and raise awareness of social media providers regarding weaknesses as well as strengths in the privacy options offered to consumers.

2. Twitter.com

2.1 Twitter.com Activity

'Tweets', '@ replies', and 'DMs' are all slang for short form communication being relayed through Twitter.com. Twitter.com is one of the Internets largest websites with 500,000,000 users (Lunden, 2012). On a daily basis they process on average 200 million short form messages. These messages are published to the entire Internet through Twitter.com. Individuals tweet about anything and everything; general communication dialog, social invitations, and personal updates are among the most common. Users divulge their geographic location through location based status Tweets, check-ins, and even through the text in their posts. Through naive intentions, many social media users publically post positioning information that can be used to derive physical addresses and locations. Stefanidis et al., (2011) refers to this emerging data type as ambient geospatial information.

2.2 Tweet Locations

Twitter.com provides locations for most published tweet. The user, through privacy settings, defines the accuracy of this location data. A user can be as inaccurate as defining an incorrect region of the world or as accurate as allowing GPS coordinates to

be embedded in the tweet. Twitter also allows for a user to tweet from a specified "place". When using Twitter.com's Streaming API, a geographic region is defined as an area of interest. This applies a filter to the Twitter Stream that only returns tweets that fall within that defined area. Therefore by nature, any returned result is accompanied with coordinates. This proves to be very beneficial from an analysis standpoint because no results have to be omitted due to the lack of location. The data collection for this project was defined to the contiguous United States. The United States was selected to ensure the largest sample set. While the Netherlands has the highest Twitter user penetration, at 26.8 percent of its population actively using Twitter (Azevedo, 2011), the United States still accounts for the majority of Twitter activity, at 50.88 percent (Sysomos Inc, 2010).

3. Twitter2GIS Add-in

3.1 Overview of Twitter2GIS

An inventive new ArcGIS add-in, Twitter2GIS, has been developed for this research, which builds up the ideas about cybercasing, discussed by Friedland and Sommer (2010), and ambient geospatial information discussed by Stefanidis et al., (2011). This add-in uses the Twitter.com application programming interface (API) to collect a stream of tweets from either a user-specified geographic region or a user-specified Twitter profile. Geo-tagged tweets are mapped and stored while the remaining tweets are analyzed for ambient location data then geocoded using the Google Geocoding API. The resulting spatial data is archived and queried off of a dictionary of terms also compiled as part of this project, which are known to divulge potential location privacy data. Additionally, the archived spatial data is analyzed for other trends in location, time, content, etc. The results of all analyses depict possible personal points of interest, tweet density trends, ambient data locations, and a profile of location privacy information unique to each user.

3.2 Add-in Methodology

The inventive Twitter2GIS analysis model for this project was developed with the intent of discovering social media personal location insecurities. Therefore it has two process flows: 1) discover ambient spatial data 2) find trends in all available spatial data. This unique application discovers ambient spatial data by looking for keywords and phrases that are known to divulge location. For instance the term "at home", when found in a tweet, has a high probability of actually being sent from their physical residence. The context of such a Twitter post is found in Figure 1.

Figure 1: Example of ambient information in tweet: "at home"

When a user tweets ambient spatial information, such as in the example shown in Figure 1, the analysis tool is able to either more accurately determine the user's physical location or, as in the example above, define characteristics about their published location. The second model process evaluates the frequency of words in tweets and discovers trends in time/location and location intensity. This is most effective when the dataset includes a large series of tweets from individual users or a series which covers popular or highly mentioned events. It is assumed that a handful of tweet samples are not enough to build a logical location profile for individual users. Within the 15 million collected tweets, successful location profiles were realized with users that had greater then 30 geolocated tweets over a 60-day period of time. While trends were visible with some users with as little as 6 geolocated tweets over their lifetime, it was determined that the completeness of the profile is directly related to the total number of sampled tweets for a user. The current analysis model looks for events through the popularity of words and location, therefore its ability to profile events is dependent on twitter activity at the time of collection.

Further research can be done on these thresholds to improve event profiling. Nevertheless, these two processes combined form a very powerful, innovative social media data analysis tool. Such a tool can inform business leaders, educate and inform the public about their privacy. The downside is that unfortunately a tool like this can also provide opportunities for criminal misuse of this data.

3.3 Twitter2GIS Development

The Twitter2GIS add-in application is built using the Python programming language and utilizes ArcPy for geoprocessing (ESRI, 2011). Additionally the new add-in employs third party libraries to help digest the Twitter stream. In total, ten "modules" were created that gather the data from Twitter, then parse, convert and pass it to a user defined geocoding API, import it into ArcMap for geoprocessing, and lastly analyze the data for trends. The code is executed through a custom add-on menu located on the users ArcGIS toolbar. The Twitter2GIS add-in flows through the typical stages of a spatial analysis: data collection or sampling, data preparation, data analysis, and displaying results. This process is depicted in the flow chart, Figure 2.

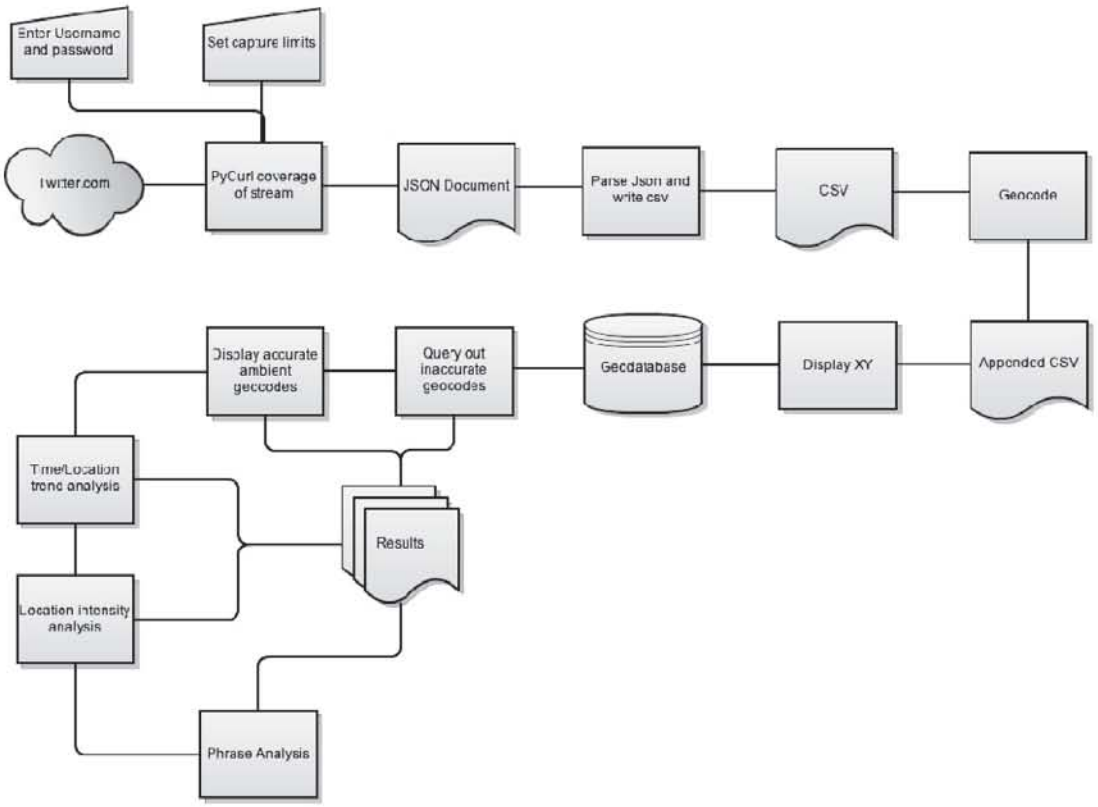


Figure 2: Twitter2GIS process flow

3.4 Twitter2GIS Limitation

The Twitter2GIS add-in has two accepted limitations: dependence on third party Python library and third party geocoding API's. The add-in relies on PyCurl (Kjetilja, 2006) to digest the incoming Twitter.com Streaming API. ArcGIS add-ins cannot register third party libraries for the user. Therefore all library dependencies must be installed on the client machine prior to use. Third party geocoding API's are used to define coordinates for ambient geospatial data. Due to limits in terms of service for many geocoding API's, the user of Twitter2GIS is responsible for managing their acceptable use of each geocoding API.

4. Findings

While Twitter does not publish any statistics on its users' location privacy practices, through the use of Twitter2GIS a user is able to analyze well beyond 15 million location enabled tweets. As discussed above in section 2.2, while Twitter's Streaming API does return a location for each result, the accuracy of that location is controlled by the user (Twitter, 2013). Figure 3 presents the density distribution of all 15 million collected tweets. The dark red areas represent high Tweet density, which primarily correlates to traditional college towns across the United States as well as large metropolitan centers. Figure 4 shows the total number of location enabled tweets collected during each of the seven data sampling sessions conducted as part of this study, in comparison to the total number of tweets broadcast during those same time periods. On average about 3.5 percent of tweets during the collection period were location enabled. The trend follows the day of the week, the curve peaking in at the 6th sampling event because it represents a busy Friday evening. It is important to note that the Twitter2GIS samples follow a very similar trend to the total tweets, indicating there is a direct correlation between the number of total tweets and the number of users Tweeting their location. Lastly Figure 5 provides the estimated number of street level accuracy or better tweets received in comparison to the total number of location enabled tweets, which averaged 23.5 percent. To summarize these results, during this study's sampling periods on average 0.8 percent, or roughly 4 million Twitter.com users, were divulging their current physical location through GPS coordinates or other active location monitoring. Additionally, 2.2 percent of all tweets, which equates to an average 4,400,000 tweets a day, were providing substantial ambient location data in the text of their tweets, measured as a Google geocoding accuracy level return of 4 or higher. These numbers indicate that Twitter users share enough spatial information on Twitter.com to make it worthwhile for third parties to harvest this information for whatever purpose they desire.

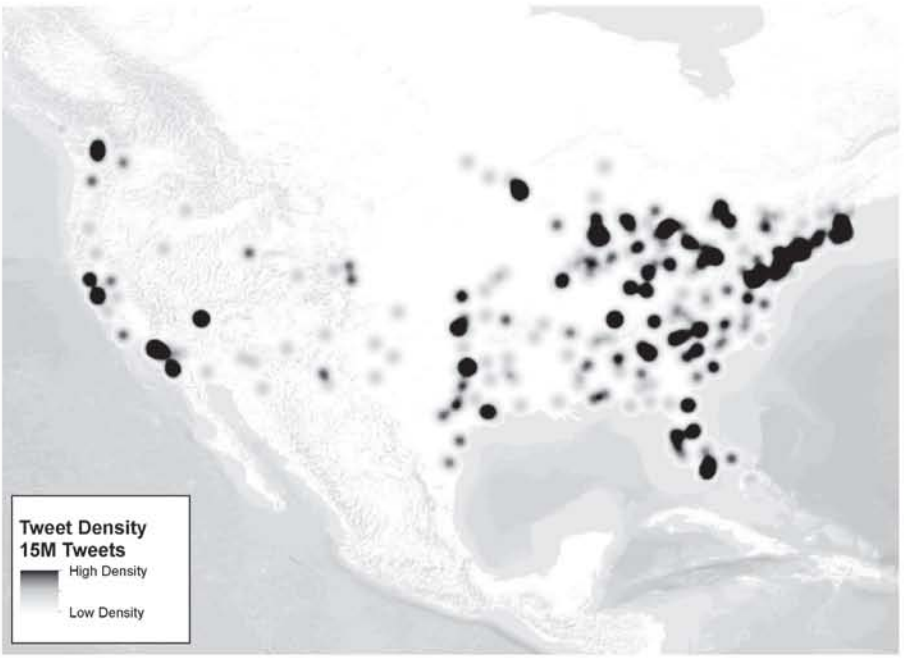
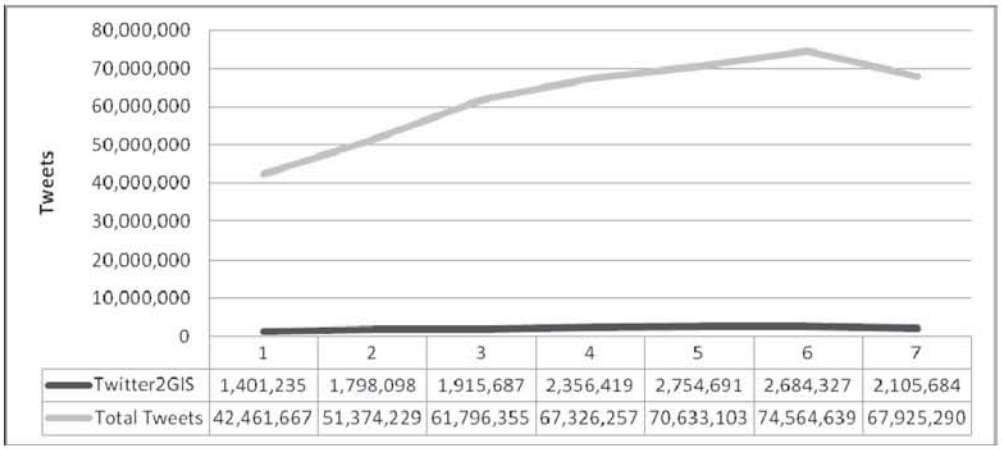
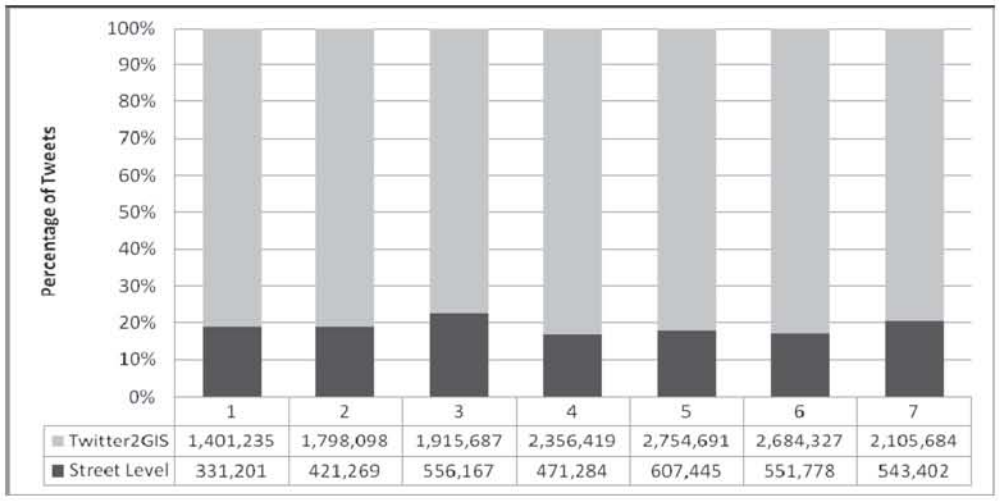


Figure 3: A heat map showing the density of tweets of all location enabled tweets during the collection period



	1	2	3	4	5	6	7
Twitter2GIS	1,401,235	1,798,098	1,915,687	2,356,419	2,754,691	2,684,327	2,105,684
Total Tweets	42,461,667	51,374,229	61,796,355	67,326,257	70,633,103	74,564,639	67,925,290

Figure 4: Comparison of location enabled tweets during all seven collection periods to the total number of registered tweets during that same period



	1	2	3	4	5	6	7
Twitter2GIS	1,401,235	1,798,098	1,915,687	2,356,419	2,754,691	2,684,327	2,105,684
Street Level	331,201	421,269	556,167	471,284	607,445	551,778	543,402

Figure 5: Compares all Twitter2GIS collected tweets to collected Tweets which have an estimated accuracy of street level or greater

5. Conclusion

In reality, though the 15 million tweets collected for this research seem significant, it's tiny in comparison to the total 72 billion tweets that are expected in the year 2013. These tweets only represent those that were geolocated within the United States during the seven data sampling periods, which greatly reduces the randomness of tweet content. At the very least, data collection should be continued in the near future over longer time periods and the same analysis completed to increase the confidence of the findings. Twitter2GIS could be extremely effective in helping the general public audit their own social location presence as well as educate them on location privacy.

This tool also represents a great educational opportunity for GIScience professionals, since it enables tweet collection and analysis in a very simple and straightforward way. While GIScience professionals consciously understand the importance of location, the general public does not usually grasp the potential risks when using common social media applications like Twitter. The lack of location privacy provides those with ill will much more valuable information than even credit card numbers and bank statements; a potentially continuous, real-time record of the physical location of the social media user, which in turn can provide opportunities for criminals to induce harm.

Thus the use of this tool strongly supports the argument that the general public needs to become more educated on their location privacy, or lack thereof. It is apparent that further research needs to be completed on this topic, in these specific areas: long term location profiling of tweets, group profiling, the potential for habitat movement simulation, and location prediction modeling. While Twitter only stores and allows access to six months of location data for any one individual, an archiving system could be built to access and store location data for individuals for long term profiling. Allowing for a larger dataset may lower the location profiling requirement thresholds. Much like herds of animals that are tracked over a period of time, a similar study should be done to investigate if it is possible to map a Twitter user's immediate social network relating to both their location publishing habits as well as their physical movements. A study could also be done to determine the interaction of individuals living in proximity of a given location, and also compare those living in similar types of environments – urban, suburban, etc. Analogous to TIGMOD (Ahearn et al., 2001), the results of such a study could be used to build human habitat coverage areas and ultimately simulate movement patterns for individual users. The data from such an analysis could in turn be used as a solid basis for Twitter data criteria selection and gathering activities and thus generation of detailed and predictive user profiles.

References

Ahearn, S. Smith, J. Joshi, A. and Ding, J. 2001, TIGMOD: An Individual-Based Spatially Explicit Model for Simulating Tiger/Human Interaction in Multiple use Forests, *Ecological Modelling*, Volume 140, Issues 1–2, 81-97, ISSN 0304-3800, 10.1016/S03043800(01)00258-7.

Ahn, G. Shehab, M. Squicciarini, A. 2011, Security and Privacy in Social Networks, Internet Computing, IEEE, Vol.15, No.3, 10-12, May-June 2011.

Azevedo, H., 2011, The Netherlands Ranks #1 Worldwide in Penetration for Twitter and Linkedin. comScore Press Release April 26, 2011. Retrieved from: http://www.comscore.-com/Insights/Press_Releases/2011/4/The_Netherlands_Ranks_number_one_Worldwide_in_Penetration_for_Twitter_and_LinkedIn

Benda, D. 2010, Sharing on Social Media Invites ID theft. The Herald-times (Bloomington, Ind.),E.8.

ESRI, 2011, The ArcPy Site Package, Retrieved fromhttp://help.arcgis.com/en/arcgisdesktop/10.0/help/-index.html#/What_is_ArcPy/000v000000-v7000000/

Friedland, G. and Sommer, R., 2010, Cybercasing the joint: On the Privacy Implications of Geotagging. Retrieved from: http://www.icsi.-berkeley.edu/pubs/networking/cybercasinghotsec10.pdf

Google, 2013, Google Maps Web Services Documentation, Geocoding API V2. Retrieved from: https://developers.google.com/maps/documentation/geocoding/v2/

Kjetilja, 2006, Pycurl — A Python Interface to the cURL library. Retrieved from http://pycurl.-sourceforge.net/doc/pycurl.html

Krumm, J., 2009, A Survey of Computational Location Privacy. Personal Ubiquitous Comput. 13, 6 (August 2009), 391-399. DOI=10.1007/s00779-008-0212-5 http://dx.doi.org/10.1007/s00779-008-0212-5

Lini, S. K., 2012, Employers Eyeing Twitter, Social Media. Privacy Journal, 38(5), 1-7. Retrieved from http://search.proquest.com/docview/101-1004350?accountid=14749

Loeffler, C., 2012, Privacy Issues in Social Media. The IP Litigator: Devoted to Intellectual Property Litigation and Enforcement, 18(5), 12-18. Retrieved from http://search.proquest.com-/docview/-1082016549?accountid=14749

Lunden, 2012, Twitter May Have 500M+ Users But Only 170M Are Active, 75% On Twitter's Own Clients. TechChurch, Retrieved from http://techcrunch.com/2012/07/31/twitter-may-have-500m-users-but-only-170m-are-active-75-on-twitters-own-clients/

Ruiz, C., Freni, D., Bettini, C. and Jensen, C. S., 2011, Location-Related Privacy in Geo-Social Networks, Internet Computing, *IEEE* , Vol.15, No.3, 20-27, May-June 2011.

Stefanidis, A., Crooks, A. and Radzikowski, J., 2011. Harvesting Ambient Geospatial Information from Social Media Feeds. *GeoJournaldoi*:10.1007/s10708-011-9438-2

Sysomos Inc, 2010, Exploring the use of Twitter aAround the World. Retrieved from: http://www.sysomos.com/insidetwitter/geography/