TwitterHitter: Geovisual Analytics for Harvesting Insight from Volunteered Geographic Information

Jeremy J. D. White[1], Robert E. Roth[2]

[1]Department of Geography
University of Wisconsin-Madison
550 N. Park Street
Madison, WI 53706
Email: jeremy@blueshirt.com

[2]GeoVISTA Center, Department of Geography
Penn State University
302 Walker Building
University Park, PA 16802
Email: reroth@psu.edu

1. Introduction

Here, we introduce TwitterHitter, an application that leverages GIScience and information visualization techniques to harvest spatiotemporal insights from microblogging generally and Twitter.com specifically. *Microblogging* describes a one-to-many form of computer-mediated communication that allows individuals or organizations to broadcast brief messages about their status, interests, and opinions via the web (Grace, Zhao, and Boyd 2010). Microblogging is a faster method of information sharing than traditional blogging, leading to more frequent updates, and is a broader method of communication than text messaging or email, leading to updates containing content intended for public viewing (Starbird et al. 2010). Microblogging offers informative and timely information that may be useful in a variety of application domains, including emergency response (Longueville, Smith, and Luraschi 2009; Starbird and Palen 2010), epidemiology and public health (Brownstein, Freifeld, and Madoff 2009; Scanfeld, Scanfeld, and Larson 2010), mobile education (Ebner and Schiefner 2008), news monitoring and recommendation (Grinev et al. 2009; Phelan, McCarthy, and Smyth 2009), and workplace coordination/collaboration (Zhao and Rosson 2009; Zhang et al. 2010).

Twitter, launched in 2006, is a popular service that combines microblogging with social networking to allow *twitterers* to post 140 character or less *tweets*, either publicly or to a restricted set of *followers* (McFedries 2007). Developers can mash-up custom applications using the Twitter API (Makice 2009), gaining access to recent public tweets and their attributes. We are particularly interested in the geospatial information provided through the Twitter API, which includes time zone, residence of the twitterer, and latitude/longitude (when twitterers opt-in to this service on GPS-enabled mobile phones). Hughes and Palen (2009) estimate that nearly 70% of public tweets are or could be georeferenced; our initial tests have shown that 10% of public tweets already have lat/long coordinates.

2. Context

This work directly relates to two emerging research thrusts within GIScience: volunteered geographic information and geovisual analytics. *Volunteered geographic information* (VGI) describes the collection and maintenance of geospatial information by citizens that are not acting in their professional capacity (Elwood 2008a, 2008b).

With Twitter, these citizens are the source of this information, acting as sensors in the landscape that reveal and explain the changing conditions surrounding them (Goodchild 2007a, 2007b). VGI is becoming a cultural phenomenon, as people are embracing a 'geo-lifestyle' by making their own location, or the location of things important to them, explicit through emergent technologies (Field 2009).

Geovisual analytics (GVA) describes the use of visual, map-based interfaces to geocomputational methods to support human reasoning (Thomas et al. 2005; Andrienko et al. 2007). The goal of GVA is to make sense of a large collection of information in order to identify and organize relevant evidence, weight this evidence against competing hypotheses, and then determine the appropriate course of action (Pirolli and Card 2005). GVA differs from prior work in GIScience in that it generates insights about information that concomitantly is voluminous, spatiotemporal, multivariate, multi-scalar, heterogeneous, and uncertain—all qualities exhibited by the VGI produced by microblogging (Starbird et al. 2010).

3. TwitterHitter

TwitterHitter (Figure 1) is a desktop application developed in the Microsoft.NET framework that allows users to retrieve all tweets and their attributes that match a user-defined query and to store this record set in an Access database. While TwitterHitter can be used in isolation as a way to gather information, it is most useful when implemented as a node in the MapNodes (Figure 2) framework (White 2010), which allows users to apply quickly a combination of spatial statistics, geocomputational processes, and output visualizations to the collected tweets using a visual programming interface. In the following, we focus on two primary output visualizations: (1) a linked map-timeline view that can plot the tweets of single individual or group of individuals and (2) an extended network graph view for visualizing connections among individuals in a region.

4. Scenario: Crime Analysis

To demonstrate the potential utility of TwitterHitter, we describe how it could support the functions of crime analysis. Many large police departments employ trained analysts that manage and interpret crime incident datasets to support law enforcement (Getis et al. 2000). Crime analysts rarely interact with patrolmen or detectives working in the field, however, making it difficult for them to interpret these analytical results in context (O'Shea and Nicholls 2003). Twitter is a supplementary stream of information that a crime analyst can use to support his or her work. Boba (2005) describes five forms of crime analysis: criminal investigative analysis, intelligence analysis, tactical analysis, strategic analysis, and administrative analysis. Applications of TwitterHitter to each is considered below.

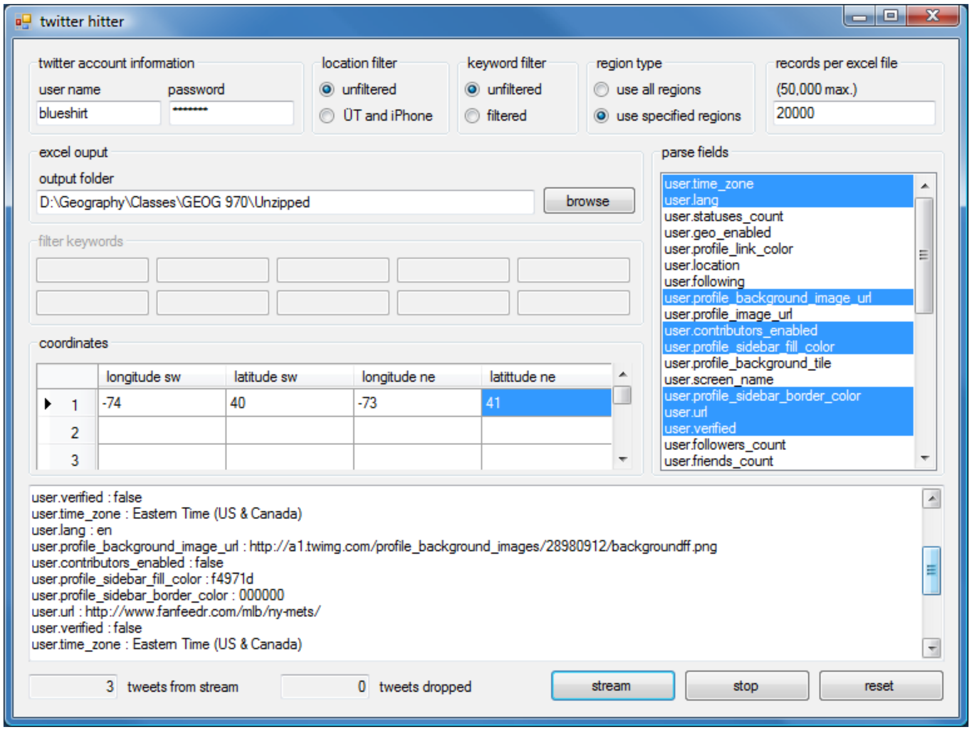

Figure 1: TwitterHitter

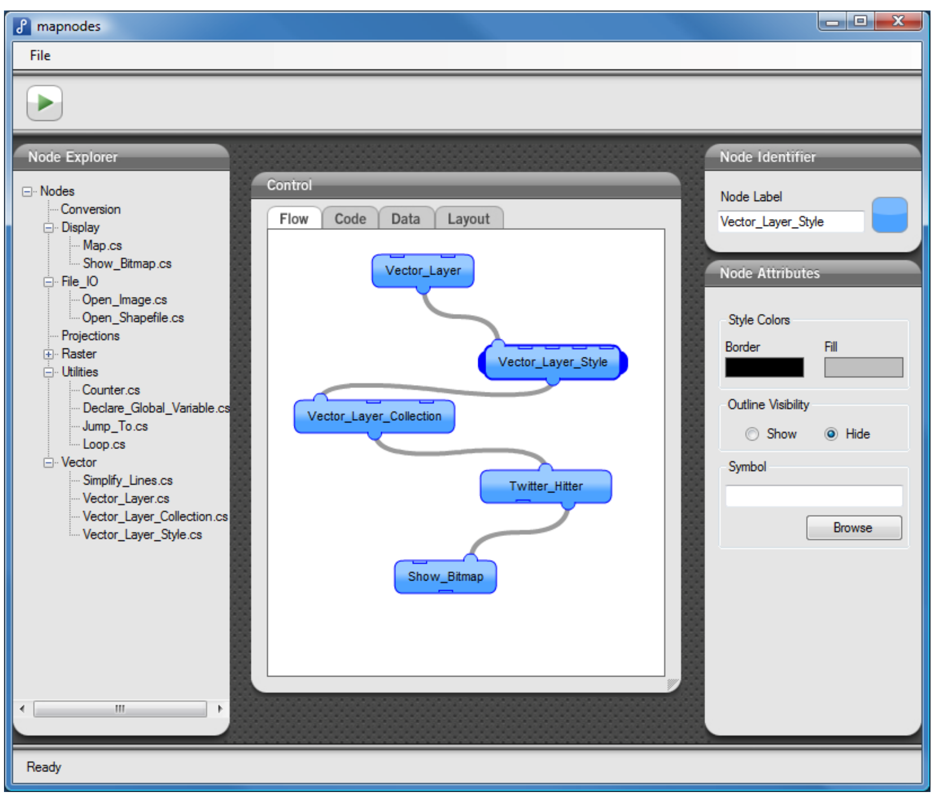

Figure 2: TwitterHitter as a component of the MapNodes framework

4.1 Criminal Investigative Analysis

Criminal investigative analysis describes the process of collecting and analyzing information about a single crime series to develop offender profiles and identify potential suspects (Rossmo and Velarde 2008). Here, analysts can enter the name and known aliases for a suspect in TwitterHitter, retrieve a spatiotemporal record of their activity, and plot a linked map-timeline view of their recent activity on Twitter (Figure 3). Analysts can also generate a directed geographic network graph (Weaver et al. 2007) of the suspect's known associates (i.e., Twitter friends), centered upon the individual suspect. The focus on individual suspects in criminal investigative analysis makes the use of TwitterHitter an 'all-or-nothing' technique, as the suspect must be 'active' for TwitterHitter to be useful (Hughes and Palen 2009); however, several researchers have been successful in extracting meaning from individual streams (e.g., Starbird et al. 2010).

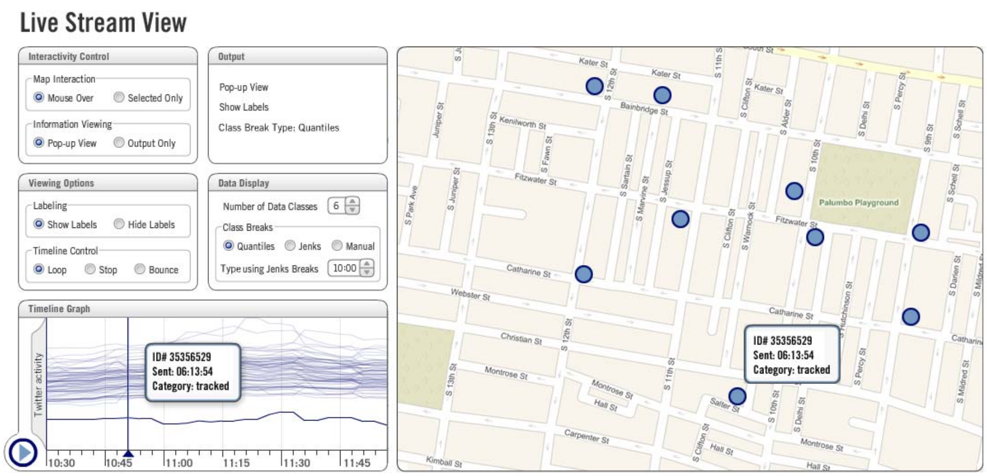

Figure 3: Individual linked map-timeline

4.2 Intelligence Analysis

Intelligence analysis investigates relationships among suspected offenders to uncover key players in crime syndicates (Innes, Fielding, and Cope 2005). An extended network graph of twitterers within a region can be constructed to represent all potential connections (Figure 4). From this, analysts can use the principle of mutual awareness to segment the extended graph into communities according to keyword themes (Lin et al. 2006). There are several statistical techniques available to characterize an identified community, such as the HITS algorithm (Kleinberg 1999), which identifies authorities within a large community (i.e., key players), and the modularity metric (Clauset, Newman, and Moore 2004), which the describes the degree of total connectivity within the community. An example of the use of comprehensive network graphs to analyze tweets is provided by Java et al. (2007).

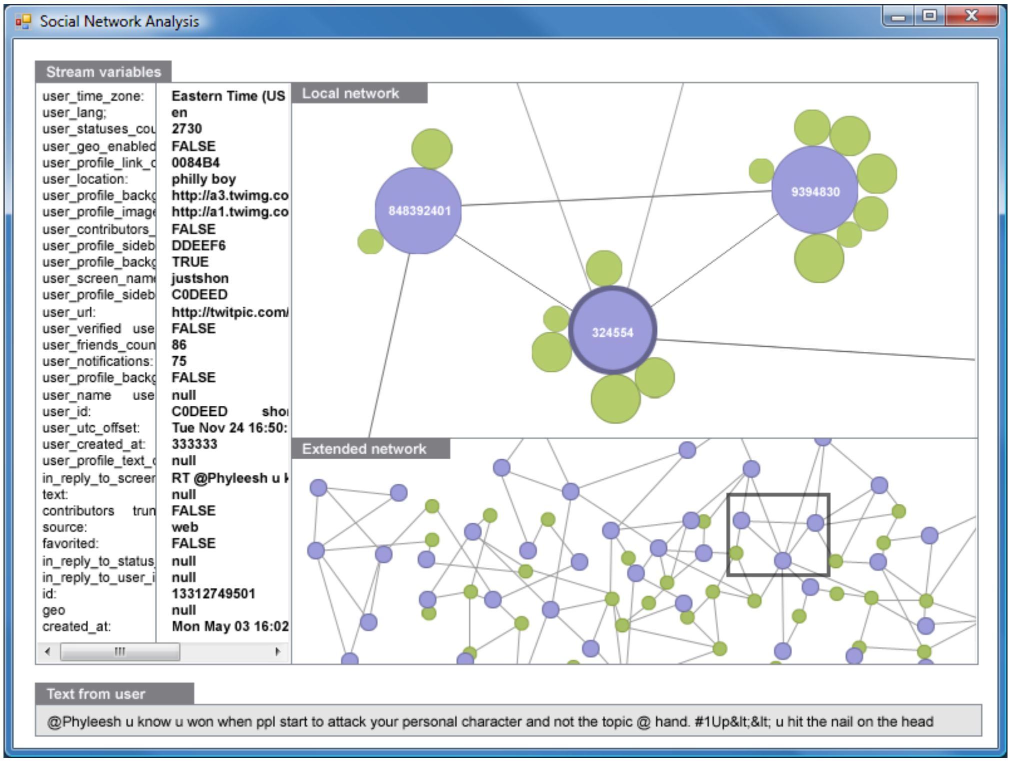

Figure 4: Extended network graph

4.3 Tactical Crime Analysis

Tactical crime analysis describes that the application of crime analysis in reaction to a recent crime spike (Bruce 2008). The potential of Twitter for tactical crime analysis is particularly intriguing, as the content of tweets can help analysts understand changing conditions before officers can collect and report information from their beat. Hughes and Palen (2009) demonstrated that the number of tweets generally corresponds to the severity of the event. A linked map-timeline view can be generated showing the volume of tweets that match a set of keywords as a heat map, allowing analysts to identify spatiotemporal hotspots (Figure 5). Difference tag clouds (Stryker, Turton, and MacEachren 2008) showing the change in frequency of words between two time slices can be generated by police beat to show the recent change in character of these neighborhoods, providing insight into the best tactical approach to combat the spike.

Figure 5: Heat map of relevant tweets over time

4.4 Strategic Crime Analysis

Strategic crime analysis describes the application of crime analysis for understanding long-term crime patterns (Goldstein 1979). Here, analysts can apply multivariate statistical analysis, such as geographically weighted regression (Cahill and Mulligan 2007), to their datasets to understand the etiology of the criminal activity, with the collected tweets or their attributes as potential explanatory variables in the analysis. Twitter also provides a partial solution to one of the most significant problems in crime analysis: determining the denominator when calculating crime rates. Twitter generally shows where people are located throughout the day, providing an approximation of population density with which to standardize the frequency of crime incidents (Abrol and Khan 2010).

4.5 Administrative Crime Analysis & Ethical Considerations

Finally, *administrative crime analysis* supports the business functions of law enforcement agencies and presents crime analysis findings to government officials and citizens (Reuland 1997). Twitter provides a supplementary feed of evidence to crime analysts—who are trained to handle private information—when making sense of criminal activity. With regards to administrative crime analysis, it is an open question if tweets can be used similarly as evidence in the courtroom, or, further, in the court of public opinion. Individual-based analysis, as described above, also may be cause for privacy concerns, although there is long-standing precedent of law enforcement accessing personal information that has not been publicly volunteered (while tweets are part of the public record); for a recent example in GIScience, see Schmitz et al. (2009). It is essential that GIScientists, who have an intimate knowledge of the importance of place, are active not only in pushing out new GVA technologies using VGI, but also active in the ethical debate concerning how such technologies should (or should not) be used.

References

Abrol, S., and L. Khan. 2010. TWinner: Understanding news queries with geo-content using Twitter. Paper read at Geographic Information Retrieval, at Zurich, Switzerland.

Andrienko, G., N. Andrienko, P. Jankowski, D. Keim, M.-J. Kraak, A. MacEachren, and S. Wrobel. 2007. Geovisual analytics for spatial decision support: Setting the research agenda. *International Journal of Geographical Information Science* 21 (8):839-857.

Boba, R. 2005. Crime analysis defined. In *Crime analysis and crime mapping*, 5-18. Thousand Oaks, CA: Sage.

Brownstein, J. S., C. C. Freifeld, and L. C. Madoff. 2009. Digital disease detection—Harnessing the web for public health surveillance. *The New England Journal of Medicine* 360 (21):2153-2157.

Bruce, C. W. 2008. Police strategies and tactics: What every analyst shoud know, 11: International Association of Crime Analysts.

Cahill, M., and G. Mulligan. 2007. Using geographically weighted regression to explore local crime patterns. *Social Science Computer Review* 25 (2):174-193.

Clauset, A., M. E. J. Newman, and C. Moore. 2004. Finding community structure in very large networks. *Physical Review E* 70 (6):#066111.

Ebner, M., and M. Schiefner. 2008. Microblogging - more than fun? Paper read at IADIS Mobile Learning Conference, at Algarve, Portugal.

Elwood, S. 2008a. Geographic Information Science: New geovisualization technologies - emerging questions and linkages with GIScience research. *Progress in Human Geography* 33:256-263.

———. 2008b. Volunteered geographic information: Key questions, concepts and methods to guide emerging research and practice. *GeoJournal* 72:133-135.

Field, K. 2009. Editorial: Cartographic Twitterings. *The Cartographic Journal* 46 (2):59-61.

Getis, A., P. Drummy, J. Gartin, W. Gorr, K. Harries, P. Rogerson, D. Stoe, and R. Wright. 2000. Geographic information science and crime analysis. *URISA Journal* 12 (2):7-14.

Goldstein, H. 1979. Improving policing: A problem-oriented approach. *Crime & Delinquency* 25 (2):236-258.

Goodchild, M. F. 2007a. Citizens as sensors: The world of volunteered geography. *GeoJournal* 69:211-221.

———. 2007b. Citizens as voluntary sensors: Spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research* 2:24-32.

Grace, J. H., D. Zhao, and D. Boyd. 2010. Microblogging: What and how can we learn from it? Paper read at CHI, at Atlanta, GA.

Grinev, M., M. Grineva, A. Boldakov, L. Novak, A. Syssoev, and D. Lizorkin. 2009. Sifting micro-blogging stream for events of user interest. Paper read at Research and Development in Information Retrieval\, at Boston, MA.

Hughes, A. L., and L. Palen. 2009. Twitter adoption and use in mass convergence and emergency events. Paper read at Information Systems for Crisis Responses and Management (ISCRAM), at Gothenburg, Sweden.

Innes, M., N. Fielding, and N. Cope. 2005. 'The Appliance of Science?': The Theory and Practice of Crime Intelligence Analysis. *British Journal of Criminology* 45 (1):39.

Java, A., T. Finin, X. Song, and B. Tseng. 2007. Why we Twitter: Understanding microblogging usage and communities. Paper read at WebKDD, Workshop on Web Mining and Social Network Analysis, at San Jose, CA.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46 (5):604-632.

Lin, Y.-R., H. Sundaram, Y. Chi, J. Tatemura, and B. Tseng. 2006. Discovery of blog communities based on mutual awareness. Paper read at WWW06 Workshop on Web Intelligence.

Longueville, B. D., R. S. Smith, and G. Luraschi. 2009. "OMG, from here, I can see the flames!": A use case of mining Location Based Social Networks to acquire spatiotemporal data on forest fires. Paper read at Location Based Social networks, at Seattle, WA.

Makice, K. 2009. *Twitter API: Up and Running*. Sebastopol, cA: O'Reilly Media.

McFedries, P. 2007. Technically Speaking: All A-Twitter. *IEEE Specturm* 44 (10):84.

O'Shea, T. C., and K. Nicholls. 2003. Crime analysis in America: Findings and recommendations, 30. Washington, D.C.: Office of Community Oriented Policing Services, U.S. Department of Justice.

Phelan, O., K. McCarthy, and B. Smyth. 2009. Using Twitter to recommend real-time topical news. Paper read at Recommender Systems, at New York, NY.

Pirolli, P., and S. Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. Paper read at International Conference on Intelligence Analysis, at McLean, Va.

Reuland, M. M. 1997. *Information management and crime analysis: Practitioners' recipe for success*. Washington, D.C.: Department of Justice.

Rossmo, D. K., and L. Velarde. 2008. Geographic profiling analysis: principles, methods and applications. In *Crime Mapping Case Studies: Practice and Research*, eds. S. Chainey and L. Tompson, 35-43: John Wiley & Sons, Ltd.

Scanfeld, D., V. Scanfeld, and E. L. Larson. 2010. Dissemination of health information through social networks: Twitter and antibiotics. *Infection Control and Epidemiology* 38 (3):182-188.

Schmitz, P. M. U., S. Riley, and J. Dryden. 2009. The use of mapping time and space as a forensic tool in a murder case in South Africa. Paper read at International Cartographic Conference, 15-21 November, at Santiago, Chile.

Starbird, K., and L. Palen. 2010. Pass it on?: Retweeting in mass emergency. Paper read at Information Systems for Crisis Responses and Management (ISCRAM), May 2-5, at Seattle, WA.

Starbird, K., L. Palen, A. L. Hughes, and S. Vieweg. 2010. Chatter on *The Red*: What hazards threat reveals about the social life of microblogged information. Paper read at Computer Supported Cooperative Work, at Savannah, GA.

Stryker, M., I. Turton, and A. M. MacEachren. 2008. Health GeoJunction: Geovisualization of news and scientific publications to support situation awareness. Paper read at GIScience 2008, 23 September, at Park City, Utah.

Thomas, J. J., K. A. Cook, A. Bartoletti, S. Card, D. Carr, J. Dill, R. Earnshaw, D. Ebert, S. Eick, R. Grossman, C. Hansen, D. Jones, K. Joy, D. Kasik, D. Laidlaw, A. MacEachren, C. Plaisant, B. Ribarsky, J. Stasko, M. STone, A. Turner, M. Ward, D. White, P. C. Wong, D. Woods, B. Wright, B. Fisher, B. Hetzler, D. Peuquet, M. Whiting, and P. Whitney. 2005. *Illuminating the path: The research and development agenda for visual analytics*. Los Alametos, CA: IEEE CS Press.

Weaver, C., D. Fyfe, A. Robinson, D. Holdsworth, D. Peuquet, and A. M. MacEachren. 2007. Visual exploration and analysis of historic hotel visits. *Information Visualization* 6:89-103.

White, J. D. 2010. Introduction to MapNodes, a modular approach to GIS. In *Annual Meeting of the Association of American Geographers*. Washington, D.C.: AAG.

Zhang, J., Y. Qu, J. Cody, and Y. Wu. 2010. A case study of micro-blogging in the enterprise: Use, value, and related issues. Paper read at CHI, at Altanta, GA.

Zhao, D., and M. B. Rosson. 2009. How and why people Twitter: The role that micro-blogging plays in informal communication at work. Paper read at Supporting Group Work, at Sanibel Island, FL.