Validating Volunteered Geographic Information: Can We Reliably Trace Visitors' Digital Footprints?

Jason L. Stienmetz
University of Florida

Daniel Fesenmaier

Validating Volunteered Geographic Information: Can We Reliably Trace Visitors' Digital Footprints?

Introduction

Information technologies (IT), in particular social media and mobile apps, have changed the way in which travelers plan, experience, and share all phases of travel (Gretzel, 2010; Xiang, Choe, & Fesenmaier, 2014). In addition to having dramatic implications for destination marketing and management strategy (Fesenmaier & Xiang, 2014; Gnoth & Jaeger, 2007; Xiang, Wang, O'Leary, & Fesenmaier, 2014), travelers' use of IT is also providing new opportunities for analyzing tourism related phenomena as vast amounts of data describe travelers' journeys (Gonzalez, Lopez, & de la Rosa, 2003; Onder, Koerbitz, & Hubmann-Haidvogel, 2014). Much of this digital trace data are characterized as Volunteered Geographic Information (VGI) as they describe user location in both time and space, typically facilitated by GPS enabled mobile devices. Some popular sources of VGI include photo sharing services such as Instagram and Flickr. Indeed, VGI has been used by scholars to describe tourist behavior such as their use of urban space and the flows or movement patterns of visitors within a destination (e.g. Kádár & Gede, 2013; Vu, Li, Law, & Ye, 2015; Wood, Guerry, Silver, & Lacayo, 2013; Zeng, Zhang, Liu, Guo, & Sun, 2012).

As a low cost and easily accessible source of data is it expected that VGI tools and related methods will continue to increase in popularity among tourism researchers. However, while interest in VGI data for tourism research expands very little discussion has addressed important questions regarding the credibility of VGI data, specifically if VGI data can be used as reliable measures of visitor behavior (Flanagin & Metzger, 2008; Goodchild & Li, 2012) that do not threaten the validity of study findings (Shadish, Cook, & Campbell, 2002). Therefore, the objective of this study is to evaluate the use of VGI data for the purpose of describing patterns of visitor movement within a tourism destination. Using St. Augustine, Florida as a case study, this research quantifies visitor flow networks using both data mined from Instagram and data collected using a traditional online survey methodology, and then conducts a series of statistical analyses to compare results. In doing so, this paper highlights the advantages of using VGI data for tourism research, but also draws attention to potential trappings which must also be addressed.

Literature Review

Travelers engage in numerous forms of social media, some of the most popular being Twitter, Instagram, Flickr, YouTube, Facebook, and TripAdvisor (U.S. Travel Association, 2012), and the mining and analysis of big data generated from Internet activities (i.e. travelers' use of social media) in order to understand human behavior has become increasingly common. For instance, researchers have demonstrated the ability to predict influenza outbreaks based on the frequency of Google search queries (Ginsberg et al., 2008) and the ability to predict political election outcomes based on Twitter messages (Tumasjan, Sprenger, Sandner, & Welpe, 2010). Analyses of big data have also facilitated the monitoring and prediction of changes in consumer behavior for retailers such as Target (Duhigg, 2012). Photo sharing services are particularly promising sources of big data for tourism research related to understanding visitor experiences as the taking and sharing of photos is now a commonly occurring tourist behavior (Onder et al., 2014).

One particular application of VGI is to use geocoded and timestamped social media data to represent the trajectories or paths that tourists "activate" (Zach & Gretzel, 2011) as they move

spatially and temporally through the destination (e.g. Kádár & Gede, 2013; Vu et al., 2015; Wood et al., 2013; Zeng et al., 2012). For example, the trajectories of all social media users within a particular destination can be used to quantify a network structure where nodes represent the unique touchpoints within the destination and ties represent the physical movement of visitors from one touchpoint to another (Stienmetz & Fesenmaier, 2013). In effect, VGI data can be used to create a weighted matrix where each cell value represents the weight of the tie (i.e. how many users took the path) connecting two touchpoints. The weighted matrix can then be used to perform traditional network analysis using software tools such UCInet (Borgatti, 2002), the tnet (Opsahl, 2009) plugin for the statistical software package R (R Core Team, 2013), and Gephi (Bastian, Heymann, & Jacomy, 2009).

Previous research related to tourist experiences, mobility, and movement has relied primarily on survey based data (e.g. Shih, 2006; Zach & Gretzel, 2011), specialized tracking equipment (e.g. Shoval & Isaacson, 2010), and/or qualitative interviews (e.g. Hristov, 2015; Kimbu & Ngoasong, 2013). However, these techniques are not without several disadvantages such as recall and other forms of measurement bias, costs in terms of time and financial resources, and the challenge of defining destination boundaries. With VGI, recall bias is no longer an issue, as photos posted on social media are also accompanied by metadata detailing when and where a photo was taken. Another issue related to recall bias is the very large set of attractions that must be listed in a survey questionnaire in order to serve as recall aides. Such techniques are more effective than unprompted, open-ended or travel diary types of data collection, but again there are inherent limitations by having to maintain a manageable list of places for the destination, and such a list may become politically sensitive or difficult to conceive. VGI data does not face this limitation as data for an entire region can be captured without discrimination to the perceived popularity of the place. Survey based data is also difficult to collect in terms of cost in time and money. Respondents must be recruited, surveys distributed, and data analyzed. Increasingly, online survey methods are becoming more difficult to manage as response rates to surveys continue to decline. The use of incentives such as prize drawings or the purchase of online survey panels are potential solutions, but they add additional costs to a study.

While the mining of VGI requires expertise in computer systems, once a data collection system has been set-up the costs of collecting VGI data are relatively low, especially considering the quantity of data that can be collected. Additionally, much of the VGI data created through social media use is publicly accessible, thereby providing an opportunity to analyze destination level data at a low cost and without requiring the sharing of proprietary traveler data held by individual destination firms. Other key advantages of using VGI data include the relative ease of replicating destination data generated by other scholars and the ability to have a continuous collection of data for use in longitudinal studies (Rogers, 1987)

VGI data, however, are not without limitations. In particular, the argument can be made that VGI data from social media sources are not generalizable or representative of typical tourist behavior. For example, many social media services represent niche markets or are utilized by younger generations whose behaviors could be significantly different than those of older generations. Therefore, it is essential to understand the user profile of any social media service that are used as a source of VGI. For example, as of September, 2015, there were an estimated 77.6 million active Instagram users in the United States (approximately 41% of smartphone users) and 400 million users worldwide (Statistica, 2015). Research of Instagram user behaviors estimates that 18.8% of all Instagram users voluntarily opt-in to share the location of where their photos are taken

(Manikonda, Hu, & Kambhampati, 2014). Interestingly, while over half (55 percent) of Instagram users are between ages 18 and 29 (see Figures 1 and 2), there is an almost equal distribution of household income among users (Statistica, 2015).

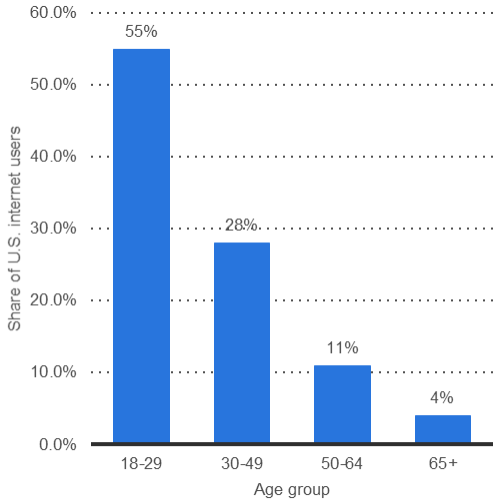

Figure 1. Percentage of U.S. internet users who use Instagram in April 2015, by age group

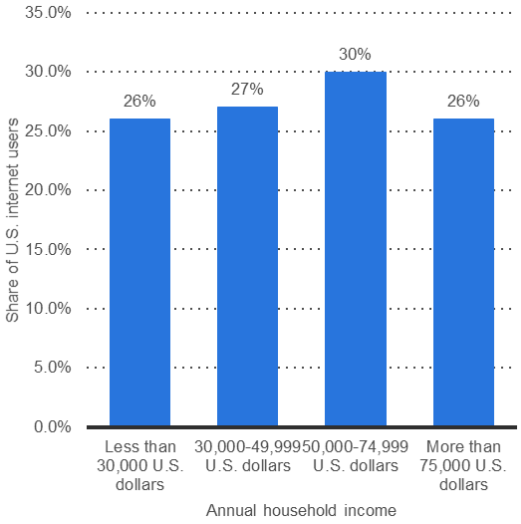

Figure 2. Percentage of U.S. internet users who use Instagram in April 2015, by annual household income

Methodology

In order to test the usefulness of destination networks created through VGI data, a comparison is made between a visitor flow network created with Instagram VGI data and a network created using online survey data for the same destination (St. Augustine, FL). The survey data collection period was between January 2012 and September 2013 and resulted in 6,334 responses. Included in the visitor survey data are the specific places each respondent visited (based on a pre-defined list of 36 attractions provided in the survey questionnaire) within the destination, as well as the demographic characteristics of each respondent. Similarly, the Instagram Application Program Interface (API) was used in October 2015 to create a database of all photos that were geo-located within the boundaries of the destination (St. Johns County, Florida) and posted between January 2010 and October 2015. In total 61,956 Instagram photos were downloaded along with the unique user id associated with each photo, the time the photo was posted, and the latitude and longitude coordinates that were tagged with the photo. Because the survey data did not contain information about sequence of places visited, a weighted, undirected network was created following the procedures used by Stienmetz and Fesenmaier (2015a, 2015b). The process of collecting and analyzing the VGI data to generate tourism paths in a destination followed the methodology of Vrotsou, Andrienko, Andrienko, and Jankowski (2011). In order to compare the survey-based network with the VGI-based network, a 36 node, undirected matrix corresponding to the touchpoints included in the visitor survey were extracted from the Instagram VGI data. In order

to make direct comparisons between the visitor flow network based on survey data, and that of the VGI data, only VGI data posted between January 2012 and September 2013 were used for network creation and analysis.

Results

In total, the travel trajectories (i.e. paths taken within the destination) of 6,334 survey respondents were aggregated to create the survey-based visitor flows network, while the trajectories of 3,553 unique users were aggregated to create the VGI based network. On average, survey respondents reported visiting 3.2 places, while Instagram users took on average 2.94 pictures while at the destination. The visual representations of the networks created by the survey data and Instagram VGI data are shown as Figures 3 and 4. Clear difference is network appearance can be observed. While both networks have a strongly connected core, the periphery of the VGI network is clearly less connected than that the survey network. Side by side comparison reveals that the survey-based network has a higher graph density (.975) compared to the Instagram-based network (.262).

Statistical comparisons of the networks were made at two levels. First, the weighted degree centrality metric was calculated for all nodes in both networks. The average weighted degree centrality for nodes in the survey network was 3,054, and the average weighted degreed centrality for nodes in the VGI network was 72. The large difference in centrality values is primarily a function of the large number of observations in the survey data relative to the VGI data. It is also important to note that because the nodal weighted degree centrality for both networks was monotonically decreasing (i.e. not normally distributed), the Spearman's correlation test was used to determine the association between the centrality metrics of the two networks (Choi, Barnett, & Chon, 2006). Results indicate a very strong correlation ($r=.859$, $p<.001$) between the two networks at the nodal level.

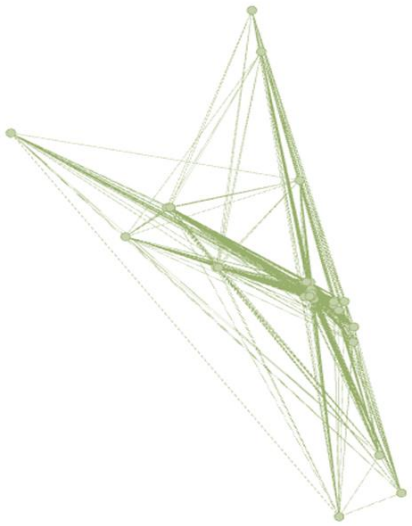

Figure 3: Visualization of Survey-based Network

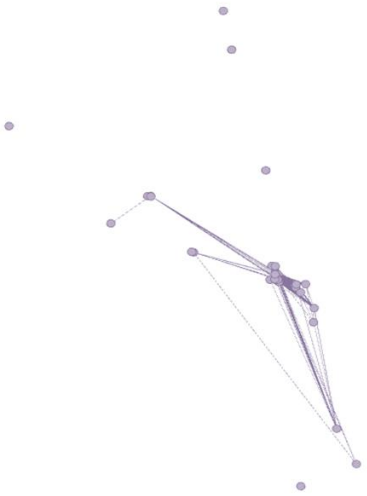

Figure 4. Visualization of VGI-based Network

Quadratic assignment procedure (QAP) matrix correlation was then estimated to compare the structural equivalence of the two networks at a system level. The QAP technique randomly rearranges matrices in order to meet the assumptions of identical and independently distributed observations (Choi et al., 2006; Takhteyev, Gruzd, & Wellman, 2012). Similar, to the comparison of centrality metrics, there is a strong correlation ($r=.747$, $p<.001$) found between the survey-based matrix and the VGI-based matrix.

As a final validity check for the VGI data, the known monthly visitor counts for a prominent attraction (Castillo de San Marcos) within St. Augustine, Florida were compared to the monthly number of unique users taking geotagged photos at the attraction (National Park Service, 2015). Because of the exponential growth in Instagram users since 2012, a natural log (LN) transformation was performed on the number of monthly users observed in the VGI dataset. Figure 5 shows visitation of the attraction based on National Park Service attendance data and the number of monthly unique visitors identified in the VGI data. While both datasets mostly mirror each in terms of season fluctuations in attendance, it is clear that there are also significant differences. For example, the VGI data indicate an overall pattern of linear growth in attendance, while the NPS data does not. In fact, there is only a weak correlation ($r=.11$) between the monthly arrivals of each data set.

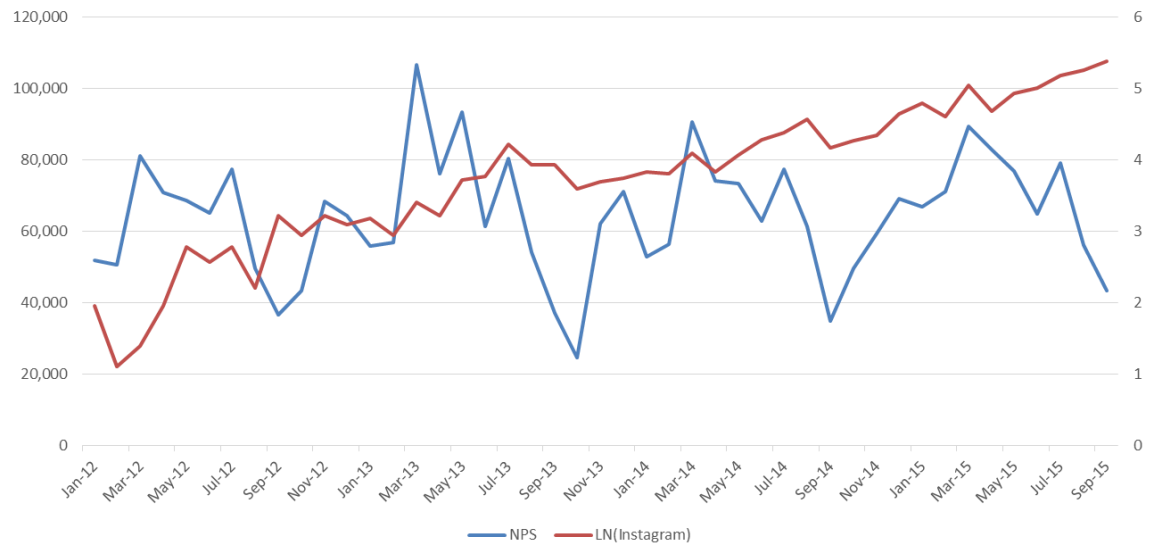

Figure 5: Comparison of VGI-estimated visitation and National Park Service reported visitation

Conclusion and Discussion

VGI data represent new opportunities for tourism researchers to better understand visitor behaviors. As an alternative to data collected through traditional surveys, qualitative interviews, or methods involving specialized equipment, massive amounts of VGI data can be obtained at a relatively low cost and are less susceptible to measurement error due to recall ability or cognitive overload caused by complex survey questionnaires. Indeed, a growing number of tourism scholars are using VGI tools and techniques to describe tourist experiences, mobility, and movement. Importantly, this research has demonstrated that Instagram VGI data can be used as a reliable measure of tourist

behavior. Patterns in the number of places visited and the number of photos taken are comparable, and strong, statistically significant correlations in nodal and system level characteristics were found in the visitor flow networks based on survey and VGI based data. While the quality of VGI data was validated, analysis also reveals important pitfalls to avoid when using VGI data, especially in the context of conducting analysis of destination systems. First, it is apparent that a large number of observations is required in order to obtain sufficient coverage of secondary network structures. As illustrated in this case study, the VGI data with half as many observations as the survey data was unable to detect some of the weaker connections found on the periphery of the destination network. Because weakly connected components have a lower probability of being observed, it is important that sample size be carefully considered when interpreting results. Finally, the number of users of a particular social media must also be considered, especially in the case of studies that involve longitudinal analyses. As illustrated in this case study, the exponential growth in Instagram users must be controlled for in a time series analysis. VGI data, just like the data generated from any other measurement technique, are not perfect and this research reiterates the importance of externally validating all data, regardless of source, to the greatest extent possible.

References

Bastian, M., Heymann, S., & Jacomy, M. (2009). *Gephi: An Open Source Software for Exploring and Manipulating Networks.* Paper presented at the ICWSM.

Borgatti, S. P. (2002). Ucinet for Windows: Software for Social Network Analysis: Harvard, MA: Analytic Technologies.

Choi, J. H., Barnett, G. A., & Chon, B.-S. (2006). Comparing World City Networks: A Network Analysis of Internet Backbone and Air Transport Intercity Linkages. *Global Networks, 6*(1), 81-99.

Fesenmaier, D. R., & Xiang, Z. (2014). Tourism Marketing from 1990 - 2010: Two Decades and a New Paradigm. In S. McCabe (Ed.), *The Handbook of Tourism Marketing* (pp. 549-560): Routledge.

Flanagin, A. J., & Metzger, M. J. (2008). The Credibility of Volunteered Geographic Information. *GeoJournal, 72*(3-4), 137-148.

Gnoth, J., & Jaeger, S. (2007). Destinations as Networking Virtual Service Firms. *International Journal of Excellence in Tourism, Hospitality, and Catering, 1*(1), 2-18.

Gonzalez, G., Lopez, B., & de la Rosa, J. (2003). Smart User Models for Tourism: A Holistic Approach for Personaliaed Tourism Services. *Information Technology & Tourism, 6*(4), 273-286.

Goodchild, M. F., & Li, L. (2012). Assuring the Quality of Volunteered Geographic Information. *Spatial Statistics, 1*, 110-120.

Gretzel, U. (2010). Travel in the Network: Redirected Gazes, Ubiquitous Connections and New Frontiers. In M. Levina & G. Kien (Eds.), *Post-Global Network and Everyday Life* (pp. 41-58).

Hristov, D. (2015). Investigating Dmos through the Lens of Social Network Analysis: Theoretical Gaps, Methodogical Challenges Adn Pracitionier Perspectives. *Advances in Hospitality adn Tourism Research, 3*(1), 18-39.

Kádár, B., & Gede, M. (2013). Where Do Tourists Go? Visualizing and Analysing the Spatial Distribution of Geotagged Photography. *Cartographica: The International Journal for Geographic Information and Geovisualization, 48*(2), 78-88.

Kimbu, A. N., & Ngoasong, M. Z. (2013). Centralised Decentralisation of Tourism Development: A Network Perspective. *Annals of Tourism Research, 40*, 235-259.

Manikonda, L., Hu, Y., & Kambhampati, S. (2014). Analyzing User Activities, Demographics, Social Network Structure and User-Generated Content on Instagram. *arXiv preprint arXiv:1410.8099*.

National Park Service. (2015). Recreation Visitors by Month Castillo De San Marcos Nm.

Onder, I., Koerbitz, W., & Hubmann-Haidvogel, A. (2014). Tracing Tourists by Their Digital Footprints: The Case of Austria. *Journal of Travel Research*.

Opsahl, T. (2009). *Structure and Evolution of Weighted Networks.* University of London, Queen Mary College, London, UK.

R Core Team. (2013). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://R-project.org/

Rogers, E. M. (1987). Progress, Problems and Prospects for Network Research: Investigating Relationships in the Age of Electionic Communication Technologies. *Social Networks, 9*(1987), 286-310.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*: Wadsworth Cengage learning.

Shih, H. Y. (2006). Network Characteristics of Drive Tourism Destinations: An Application of Network Analysis in Tourism. *Tourism Management, 27*(5), 1029-1039.

Shoval, N., & Isaacson, M. (2010). *Tourist Mobility and Advanced Tracking Technologies*. New York: Routledge.

Stienmetz, J. L., & Fesenmaier, D. R. (2013). Traveling the Network: A Proposal for Destination Performance Metrics. *International Journal of Tourism Sciences, 13*(2), 57-75.

Stienmetz, J. L., & Fesenmaier, D. R. (2015a). Estimating Value in Baltimore, Maryland: An Attractions Network Analysis. *Tourism Management, 50*, 238-252.

Stienmetz, J. L., & Fesenmaier, D. R. (2015b). *Measuring Intra-Regional Tourist Behavior: Towards Modeling Visitor Expenditure Dynamics*. Paper presented at the 4th International Conference on Sub National Measurement and Economic Anlaysis of Tourism, Puerto Rico.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter Networks. *Social Networks, 34*(1), 73-81.

Vrotsou, K., Andrienko, N., Andrienko, G., & Jankowski, P. (2011). Exploring City Structure from Georeferenced Photos Using Graph Centrality Measures. In D. Gunopulos, T. Hofmann,

D. Malerba, & M. Vazirgiannis (Eds.), *Machine Learning and Knowledge Discovery in Databases*. Berlin/Heidelberg: Springer.

Vu, H. Q., Li, G., Law, R., & Ye, B. H. (2015). Exploring the Travel Behaviors of Inbound Tourists to Hong Kong Using Geotagged Photos. *Tourism Management, 46*, 222-232.

Wood, S. A., Guerry, A. D., Silver, J. M., & Lacayo, M. (2013). Using Social Media to Quantify Nature-Based Tourism and Recreation. *Sci Rep, 3*, 2976.

Xiang, Z., Choe, Y., & Fesenmaier, D. R. (2014). Searching the Travel Network. In S. McCabe (Ed.), *The Routledge Handbook of Tourism Marketing* (pp. 281-298). Oxon: Routledge.

Xiang, Z., Wang, D., O'Leary, J. T., & Fesenmaier, D. R. (2014). Adapting to the Internet: Trends in Travelers' Use of the Web for Trip Planning. *Journal of Travel Research*.

Zach, F., & Gretzel, U. (2011). Tourist-Activated Networks: Implications for Dynamic Bundling and En Route Recommendations. *Information Technology & Tourism, 13*(3), 229-238.

Zeng, Z., Zhang, R., Liu, X., Guo, X., & Sun, H. (2012). Generating Tourism Path from Trajectories and Geo-Photos. In X. S. Wang, I. Cruz, A. Delis, & g. Huang (Eds.), *Web Information Systems Engineering - Wise 2012* (pp. 199-212). Berlin/Heidelberg: Springer.