CHI Article Review of

“I Know Where You Live: Inferring Details of People’s Lives by Visualizing Publicly Shared Location Data”

Ilaria Liccardi, Alfie Abdul-Rahman, Min Chen

John Maloney

ITEC 435 001

Abstract

The article I selected, titled “I Know Where You Live: Inferring Details of People’s Lives by Visualizing Publicly Shared Location Data” investigates and tests how accurately a person could infer the functional location of social media users when posting with a geolocation tag. The study involves geo-location data collected from the social media profiles of 27 participants in the Boston, MA area, and tests the inferences of 45 participants located in England, UK.

# Introduction

This area of study initially piqued my interest because social media security is a relatively unexplored topic that is difficult to address considering the rate of advancement in modern technology. Social media is often used to connect and communicate with friends, for example, on what restaurant they are eating at or where they are travelling to. Although communicating this information with one’s friends is harmless, the nature of the Internet creates potential security flaws. The world-wide web allows for many computers to be interconnected and widely accessible. Any web user that provides personally identifiable information online could be leaving themselves vulnerable to attackers that can exploit this location data and discover further information about a person (Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel., 2013). Obtaining a user’s location data can aid an attacker in inferring that user’s activity patterns and the structure of their friendship networks (N. Eagle, A. Pentland, and D. Lazer., 2009). Also, in some cases location data is being collected without the users’ knowledge. Many companies target advertising based on a profile that can be created using data from geotags broadcasted on social media sites (H. Xu, H.-H. Teo, B. C. Y. Tan, and R. Agarwal., 2009).

# Study Methodology

The aim of the research was to investigate how easily a casual observer (without specialized skills or tools that require high technical ability) can infer the type and/or relative function of a given geo-location point by visualizing location data in a simple and easily replicable manner (Liccardi, I., Abdul-Rahman, A., & Chen, M., 2016). This study presents a controlled quantitative user experimental design to reach a conclusion. The experiment aimed to produce statistically significant results in order to draw a generalization about the whole user population.

In their hypothesis, Liccardi, Abdul-Rahman and Chen sought out to prove that users of social media can infer with relatively high accuracy the functional type (i.e., home, work, leisure or transport) of another user based on just three pieces of geo-location data from tweets. Other questions they answered with the data’s conclusion were “How do different data representations impact participants’ ability to infer functional location? Does the accuracy of inference depend on different types of location? Does increasing data density improve or impede the accuracy of inference? How easy is it to discover locational information that is private to people? Does one need a large dataset to do so?” (Liccardi, I., Abdul-Rahman, A., & Chen, M., 2016).

In this quantitative study, the independent variables involved were the visual representations of the data (textual, static and animated visualizations) and the amount of data (1, 3 or 5 day(s). The dependent variable was the functional location type (home, work, leisure and transport). These variables are appropriate for the content of the study and provide enough conditions for participants to select from without introducing unnecessary context that could skew the results. The researchers implemented a within group design for this study to test all participants on all conditions corresponding to the geo-location data to achieve the most accurate results for the population. The experiment could have used a between-subjects design, testing some participants on one specific functional location, for example only tweets with geotags posted at work, to collect more precise results identifying users at that functional location, however this design would require a much larger testing population in order to produce accurate results of the whole user population (Between-Subjects vs. Within-Subjects Study Design, 2018). In this study, the within subjects design is a better fit because the study can draw conclusions based on a smaller group of participants and compare the results of each case.

Due to complexities caused by the multiple variables being tested, there are many challenges involved in this study that had to be approached. One limitation of the experiment was the data owners daily routine. Travel routines create a pattern in the data that can cause testers to create an inference based on prior information rather than the data that is being tested. Another limitation of this test was data representation. In order to conclude that the participants could accurately infer the functional location of data owners without advanced technical skills or software, this test had to be presented with only pen and paper. A further limitation was learning effect vs. confounding effect.(Smith, P. G., & Day, N. E., 1984). Geo-location data tagged in an area familiar to a tester could skew the results based on their prior knowledge of that specific location. In order to create a task list of identification tasks, participants in the study were given a pre-test with questions regarding their educational and demographic information to confirm that their background would not cause outliers in the data.

Before testing the hypothesis, those orchestrating this study had to create a real-life location dataset. By recruiting 230 people in the Boston, MA area using Twitter with location sharing mode on, tweets were collected over a period of three weeks. After assessing an ideal population for the study, 87 users were selected to submit their location data for testing, which was narrowed down to 27 participants. The locational data was then organized into four categories (home, work, transport and leisure). The data owners were given tests on their daily routines and other habitual information, so as to control the independent variables in the study. Once the data was gathered, another population of 45 participants was selected from the Cambridge, UK area and given a similar demographic and habitual test. In order to stay consistent, the study was conducted in a computer lab with identical machines to avoid any variation in screen size and computer speed. Participants were trained and asked to identify to the best of their ability the functional location which was tagged in the Twitter post represented by either textual, static or visual geo-locational data.

# Results

The results of the data collected using the standard T-test technique displayed participants that correctly matched the functional location to the that of the actual data in animated, textual and static formats of geo-locational data. The T-test was optimal for the study because it allows for easy comparison between the different types of location data (One Sample T-Test., 2019). The average results of each of these tests were 66.3%, 56.1% and 66.4%, respectively, with the overall mean being 62.9%. The data found concluded that transport was found to be identified more accurately than the rest, work was discovered more accurately than home, while leisure was more complicated to identify. Overall, the study indicated that textual visualization can lead to correct identification over 50% of the time. The results showed that three different affinities for data visualization were present and participants were either textually or visually inclined. The results of the experiment supported the hypothesis that without expansive knowledge or tools, a person could accurately identify the functional location of a social media user by using geo-locational data the user posted online. The general methodology of this study is solid in that it takes into account the many factors that could skew data, however this test could be improved by expanding the population of both participants who provided their own geo-locational data and of participants who were being tested, to achieve more precise data.

References

Between-Subjects vs. Within-Subjects Study Design. (2018, May 13). Retrieved from <https://www.nngroup.com/articles/between-within-subjects/>

H. Xu, H.-H. Teo, B. C. Y. Tan, and R. Agarwal. 2009. The Role of Push-Pull Technology in Privacy Calculus: The Case of Location-Based Services. J. of Management Inf. Syst. 26, 3 (Dec 2009), 135–174.

Liccardi, I., Abdul-Rahman, A., & Chen, M. (2016). I Know Where You Live. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems - CHI 16*. doi:10.1145/2858036.2858272

N. Eagle, A. Pentland, and D. Lazer. 2009. Inferring friendship network structure by using mobile phone data. Proc. of National Academy of Sciences 106, 36 (2009), 15274–15278.

One Sample T-Test. (2019). Retrieved from https://www.statisticssolutions.com/manova-analysis-one-sample-t-test/

Smith, P. G., & Day, N. E. (1984). The Design of Case-Control Studies: The Influence of Confounding and Interaction Effects. *International Journal of Epidemiology,13*(3), 356-365. doi:10.1093/ije/13.3.356

Y.-A. de Montjoye, C. A. Hidalgo, M. Verleysen, and V. D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. Scientific Reports 3, 1376 (2013).