

Finding a City's Activity Bubbles in Geotagged Social Media

Dan Tasse, Jason Hong

Human-Computer Interaction Institute
Carnegie Mellon University
Pittsburgh, PA 15224
dantasse@cmu.edu, jasonh@cs.cmu.edu

Abstract

Bill Bishop popularized the idea that we all live in a self-curated world of bubbles in his book, “The Big Sort” (Bishop 2009). People often spend most of their time in completely different places from their neighbors, and this fragmentation of society leads to reduced empathy, policy quarrels, and even violence. People may not even be aware that they are segregating themselves so dramatically. Fortunately, thanks to the abundance of available geotagged social media data, we can easily and cheaply detect these social bubbles and give people a greater insight into their activity patterns. Understanding how we spend our time can be the first step in changing from isolated citizens into a connected community.

Introduction

We increasingly live in bubbles. As incomes and mobility have grown, people have moved more frequently from their hometowns to new cities and neighborhoods full of people like them. This has detrimental effects on our society: neighbors don't know each other and therefore are unable to understand each other's problems. Systemic problems such as racial divides can lead to national crises. For example, the recent protests in Ferguson, Missouri brought to light the fact that the average white person's social network is 91% white (Jones 2014). Even in areas that are largely integrated, people can spend all of their time visiting certain locations. Cities like San Francisco deal with disparate groups, such as young tech workers and long-time residents, living nearby but often not mixing.

Social media services, such as Twitter, Instagram, and Flickr provide a large amount of free, public, geotagged data. Using this data, we can tell where certain users go, and understand what “activity bubble” they spend their

time in. We propose a web-based tool that will reveal the bubbles in a city.

Clustering and Topic Modeling

We've gathered over 9 million tweets from Twitter's public API in Pittsburgh, over about 8 months, and 5 million tweets in San Francisco over 4 months. We also have access to the same data set of Foursquare checkins used by Cranshaw (*et al*). We aim to understand activity bubbles in two ways. First, we will use topic modeling on the Foursquare dataset to find clusters of people in those checkins, inspired by Joseph (*et al*). However, where they used locations as the “documents” and people who visit those locations as the “words”, we plan to use people as the “documents” and locations that they visit as the “words”. The result will be clusters of people instead of clusters of places. We will also repeat the process with our Twitter data, but this adds one more difficulty: tweets are posted with latitude/longitude coordinates instead of location names. Therefore, we will first discretize the data by simply superimposing a rectangular grid onto the city region in question.

Visualization

Two visualizations will be very useful. First, we should be able to visualize each cluster on a map. This is particularly critical because the clusters are abstract (“clusters of people”). To create this visualization, we will simply consider all tweets posted by all users in a cluster, weighted as a fraction of their total tweets. For example, if a user tweets four times, one of which is in a given region, that will add $\frac{1}{4}$ to the region's score. We can then simply plot a heat map over the region.

Another useful visualization will be the fraction of the user's time that is spent in each bubble. This will help users understand the different networks that they travel in, and how diverse their network truly is.

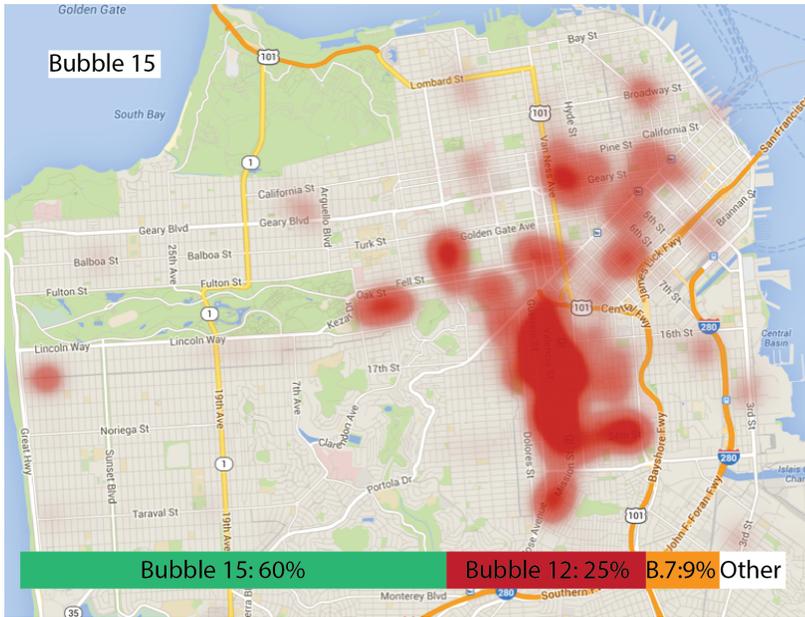


Figure 1 - a mockup. The bar across the bottom shows how much time the user spends in each "bubble".

Maintaining Privacy

Despite the public nature of this data, it remains important for us to maintain users' privacy. Applications like pleaserobme.com (Bilton 2010), which highlights houses whose users have recently posted publicly about being away, show one potential misuse of the data. However, there is nothing about our application that inherently must be invasive, as we only plan to display results of aggregate groups of people.

Effects

Users will be able to tell which cluster or clusters they most resemble, and how closely. They will also be able to see this resemblance visually, through a view of their own activity heat maps. People might also be able to find recommendations for ways they could get outside their bubbles.

In addition, simply finding these clusters could be enlightening. It may be the case that a "Big Sort" as envisioned by Bishop is not happening at all on the individual daily activity level (Abrams and Fiorina 2012) Either way, deeper investigation into where people go within a city will be informative and useful. Furthermore, efforts of this type could help transform the city into a more citizen-powered system, as citizens help shape the different "bubbles" with the social media posts they're already posting.

References

Abrams, S.J. and Fiorina, M.P. "The Big Sort" That Wasn't: A Skeptical Reexamination. *PS: Political Science & Politics* 45, 02 (2012), 203–210.

Bilton, N. Burglars Said To Have Picked Houses Based on Facebook Updates. *The New York Times Bits*, September 12, 2010. http://www.hamiltonnj.com/filestorage/83868/83958/84402/84689/Burglars_Said_to_Have_Picked_Houses_Based_on_Facebook_Updates_-_NYTimes.com.pdf

Bishop, B. *The big sort: Why the clustering of like-minded America is tearing us apart.* Houghton Mifflin Harcourt, 2009.

Cranshaw, J., Schwartz, R., Hong, J. I., & Sadeh, N. (2012). *The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City.* ICWSM.

Jones, R. P. Self-Segregation: Why It's So Hard for Whites to Understand Ferguson. August 21, 2014. <http://www.theatlantic.com/politics/archive/2014/08/self-segregation-why-its-hard-for-whites-to-understand-ferguson/378928/>

Joseph, K., Tan, C.H., and Carley, K.M. Beyond "Local", "Categories" and "Friends": Clustering foursquare Users with Latent "Topics." (2012).

Kwan, M. and Lee, J. Geovisualization of Human Activity Patterns Using 3D GIS : A Time-Geographic Approach. In *Spatially Integrated Social Science: Examples in Best Practice.* 2003.

Zuckerman, E. CHI keynote: Desperately Seeking Serendipity. May 12, 2011. Retrieved from <http://www.ethanzuckerman.com/blog/2011/05/12/chi-keynote-desperately-seeking-serendipity>