FINAL PROJECT WRITE-UP

Mapping Twittsburgh: Visualizing Twitter Data & Neighborhood Demographics

Visit the final project at: tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

FRAMING THE PROJECT

Social media, a relatively new phenomenon, has rapidly captured the attention of researchers in a variety of fields. While there are occasions where this attention may be unwarranted (Munro and Manning, 2012), data captured from social media outlets such as Twitter have the potential to provide novel insights in fields like sociology and communications and to provide a breadth of new application areas for researchers in network analysis and machine learning.

One area of particular interest is in the study of the city, where the field of urban computing (Zheng et al., 2011) has sprung up to take advantage of geo-tagged social media data (as well as a myriad of other types of data), in an effort to uncover an understanding of how humans live, work and play within the city. Work in the area of urban computing has led to a variety of endeavors that consider how one can use novel data sources to redefine the "neighborhood." The concept of a neighborhood —a community within the city—is qualitatively easy to define, while a more quantitative description of the function, culture and boundaries of neighborhoods has been much more difficult to come by.

In this research project, we use handcrafted data from PGHSNAP of the neighborhoods within the city of Pittsburgh to sidestep the problem of defining neighborhood boundaries. Instead, our focus is on how social media data, in particular geo-tagged Tweets, can be used in combination with census data to provide a visualization that suggests cultural and functional assets of different neighborhoods within the city of Pittsburgh.

Our particular interest in visualization stems from our belief that while social media has been used in novel ways to present interesting understandings of neighborhoods, the true definition of a neighborhood requires domain-specific knowledge that only residents of cities can currently provide. We hypothesized that by providing a visualization to enable residents of Pittsburgh to make sense of the characteristic demographics and term-based representations of the neighborhoods within the city, we might uncover novel observations that could be used to understand cultural and functional phenomenon within and across the city of Pittsburgh.

To this end, we create a map-based visualization showing relationships between popular terms in location-tagged tweets, the demographics of Pittsburgh neighborhoods and what "residents" of a neighborhood tend to tweet about versus what "non-residents" tweet about when they are in a

given neighborhood. It was our hope that with this information, map viewers familiar with the city would be able to not only provide insight into social media behavior and demographics within neighborhoods, but also to expand in order to describe possible patterns across neighborhoods as well.

It was thus our intention to produce a tool that could be viewed by a large number of people within Pittsburgh and to generate novel and interesting observations from the crowd. Unfortunately, we found that the most difficult and by far the most time-consuming part was developing a visual language to display both demographic and social media data on the same map in an intuitive way. Because of that challenge, a large part of this paper describes how we eventually came to a solution, how we used what we learned in the course to do so, and what we drew from our own analysis of our final product. In the interest of suggesting possible merits of the tool we created, we add our own hypotheses generated from the visualization as well as a single user study, and we suggest ways in which our work could be improved in the future.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

RELATED WORKS
FROM THE COURSE

We were very fortunate to pull a lot of resourceful information from the various visualization research papers that were provided in class this semester. Some of the readings we focused on the most included "Perception in Visualization" (Healey, C.), "Parallel Sets: Visual Analysis of Categorical Data" (Bendix, F., Kosara, R., & Hauser, H.), "Making Sense of Large Network Data" (Chau, D. Kittur, Hong, J., Faloutsos, C.), "Balancing Systematic and Flexible Exploration of Social Networks" (Perer, A. & Shneiderman, B.) and "Graph Visualization & Navigation" (Herman, I. Malençon, G. & Marshall, M.S.). The readings and class discussions provided extensive insights on how to formulate our visualization and on how we could manipulate the network data to best suit our project.

Healey's paper about perception in visualization touched on human perception and how color is a common feature that is used in many visualization designs. When faced with determining how to depict different regions in Pittsburgh, as well as how to compare Pittsburgh regions side-by-side, color was the first method that came to mind. Healey goes into detail about perceptual balance as well as distinguishability. By using proportional opacities of yellow to depict population size of Pittsburgh neighborhoods, as well as a solid colors of blue and red to depict neighborhoods that were being compared, we found that our color scale produced a perceptually uniform difference in color and that every color was equally distinguishable from all the others; no one color was easier or harder to identify (Healey, 2009). Bendix, Kosara, and Hauser also contributed to visualization research, but more along the lines of analyzing visualizations of categorical data. Given that our data referred to location in Pittsburgh (neighborhood) and the nature of the tweet (emotion or feeling), this was very applicable to our project. Bendix et al., delve into interactive visual analysis and explain that there are two main requirements for a visualization technique. The latter requirement states that there needs to be a "powerful, user-friendly,

and user-driven interaction scheme" (Bendix, et al., 2005). When asked to test our visualization, a user that had no previous knowledge of our project or the course found it very easy and intuitive to navigate and understand our visualization. The on-screen instruction as well as the changing and interactive menus are intended to keep the user engaged with and intrigued by the trends that are present in the visualized data.

Herman et al. also provided valuable insight on zooming and panning, given the mapping feature of our visualization. Zooming is one of the basic interaction techniques of information visualizations. Since the resolution and color depth of a display often limits the maximum amount of information, zooming is a crucial technique to overcome this limitation. There are three different zooming techniques: geometric, fisheye and semantic. Geometric zooming can be commonly seen in map applications and is most relevant to our project. It allows the user to specify the scale of magnification, then increases or decreases the magnitude of an image by that scale. This allows the user to focus on a specific area; information outside of this is then generally discarded. Given the nature of our Pittsburgh map in our visualization, though, discarding is less common due to the size of the region.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

Although aesthetics and functionality are important aspects of visualization, a focus on how graphs and networks work is also fundamental in a successful visualization. Chau, Kittur, Hong and Faloutsos analyze a program called Apolo and how it successfully makes sense of large network data by combining user interaction and machine learning. The key takeaways that are most related to our project are that Apolo was able to build on a large body of research that was aimed at understanding and supporting how people gain insights through visualization. Apolo, as opposed to many systems that focus on providing overviews, adopts a bottom-up sensemaking approach aimed at helping users construct their own landscapes of information (Chau et al., 2011). Capitalizing on the mutual topic of network data usage, Perer and Shneiderman delve into social network analysis (SNA), which has emerged as an impactful means for evaluating the importance of connections in networks. Visual representation of social networks is important to understand the network data and convey the result of the analysis. Lots of analytic software has modules for network visualization. Exploration of the data is done through displaying nodes and ties in various layouts, and attributing colors, size and other advanced properties to nodes. Visual representations of networks may be a powerful method for conveying complex information, but care should be taken in interpreting node and graph properties from visual displays alone, as they may misrepresent structural properties better captured through quantitative analyses.

FROM OUTSIDE THE COURSE

There has been a significant amount of work in the field of urban computing under a variety of guises, including but not limited to human mobility modeling (González et al., 2008), geospatial social networks (Butts et al., 2012; Hipp et al., 2012), and ubiquitous computing (Cranshaw et al., 2012). This work draws on,

of course, decades (or perhaps more appropriately centuries) of work in urban sociology and the study of the city.

Much of this work is relevant to our current project, but of primary interest are those focusing on neighborhoods and demographics in social media data. Cranshaw et al., define neighborhoods as areas of the city that provide a diverse array of functions to a specific section of the population. They show that in a spatially constrained graph of foursquare venues, one can use shared users across venues to provide a representation of neighborhoods within the city of Pittsburgh. In contrast, Noulas et al., use foursquare data to define neighborhoods as regions of the city having similar functions, and show that these types of neighborhoods can also be inferred from social media data.

These works, and others like them using "big data" to infer neighborhood structure (Hipp et al., 2012; Yuan et al., 2012), suggest that an interesting avenue of work in urban computing is to better understand what a neighborhood is, and how it can be defined in ways that are more representative of culturally and socially homogenous regions within the city. Many of these works, however, assume that these definitions are observable from social media data. We believe that incorporating crowd-sourced input beyond those that are by-products of potentially biased social processes (Tang et al., 2010) may be necessary to understand neighborhoods within the city and how different neighborhoods inter-relate.

Furthermore, we believe that a distinction is necessary in the literature between what a neighborhood means to those who live there as opposed to what it means to those who visit it. For example, our dataset showed (in work not presented) that people who were non-residents of East Liberty tended to tweet only from the park, whereas "residents" tweeted from, as one would expect, the residential areas of the city. Our visualization was geared towards seeing if language, in addition to location of these tweets, reflected these differences.

While interest in the definition of neighborhoods is growing within the urban computing domain, computational socio-linguists have found that census data combined with geo-tagged social media provides an interesting way to study how demographics correlate with the evolution of language across space and time (Eisenstein et al., 2012, 2011, 2010). These works suggest that demographics are still an important aspect of social behavior across social media. However, conclusions are often made at a relatively high level, suggesting, for instance, that the diffusion of new terminology across the US was based on racial similarities across cities.

Regardless, these works suggest that demographic differences are observable in terminology used in social media. By combining demographic data and social media with a map into a single visualization, we hoped that viewers could perform sensemaking on a complex dataset and provide more socially interesting, detailed hypotheses on the relationships between space, sociality, terminology and demographics within Pittsburgh.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

DEVELOPING THE VISUALIZATION

TWITTER DATA: HANDLING AND VISUALIZING

The social media dataset that we use is from Twitter. We obtained approximately 120,000 geo-tagged tweets from Pittsburgh and the surrounding areas using TweetTracker (Kumar et al., 2011), a tool that allows one to easily set up a bounding box to capture tweets in a specific area. Given a set of tweets, our visualization required that we present a) interesting keywords of tweets occurring in b) different neighborhoods, and c) differentiating between tweets of "residents" and "non-residents".

In order to parse out keywords from tweets, we use a tokenizer developed specifically for tweets (Gimpel et al., 2011). The next step was to determine the neighborhood in which the tweet occurred. In order to do so, we use the R package maptools (Bivand and Lewin-Koh, 2013) to perform a spatial join on the location of each tweet and the neighborhood data obtained from PGHSNAP, described below. Having determined the neighborhood that each tweet was from, we could now determine the "home" neighborhood of each user in our dataset.

We first remove all users having less than 5 tweets, as we suspected that any fewer would result in a highly biased view of a user's home, while any higher of a threshold removed too much of the data. Following in the line of previous work (Cho et al., 2011), the neighborhood in which a user had the most geo-tagged tweets was considered her "home neighborhood". Though this definition of home is liable to produce some noise, the heuristic has been shown to be reasonably reliable, and thus seemed like a good foundation for this project.

As a result of these steps, we now have a set of keywords for each neighborhood, where this set of keywords is split by tweets made by "residents" of that neighborhood and those made by residents of other neighborhoods ("non-residents"). In order to determine the terms most relevant to a given subset (i.e. resident or non-resident) of a given neighborhood, we treat the set of associated keywords as a document in the form of a bag of words. We have 2N documents, where N is the number of neighborhoods. Using this collection of documents, we apply tf-idf weighting to all terms in the corpora, and select the top five terms in each subset of each neighborhood for visualization.

DEMOGRAPHIC DATA: HANDLING & VISUALIZING

Data from 90 Pittsburgh neighborhoods was included in this visualization. We retrieved the neighborhood boundaries and a variety of demographic measures from the publicly available database at PGHSNAP, a map and neighborhood data initiative orchestrated by the Pittsburgh Department of City Planning that draws on 2012 US census data. The data points we selected to incorporate in our visualization included population density, education levels, median income and home price, ethnicity, crime rates, and age.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

In our visualization, each demographic can be selected from a dropdown menu. Every neighborhood is then overlaid with an opacity that reflects its weighted value based on the current data selection, generating a heat map. Specific values corresponding to each neighborhood are displayed in the corner of the browser window when the cursor hovers over that neighborhood. As detailed in the user study below, this interaction supports the user's ability to suggest inferences based on a third dimension of input.

IMPLEMENTATION DETAILS

We relied on several open-source JavaScript libraries to implement the map and interactions, expressly:

> *OpenStreetMaps*, a worldwide cartography database with graphical representation provided by CC-BY-SA.
>
> *Leaflet.js*, a JavaScript library for manipulating map boundaries.
>
> *Jquery.js*, a JavaScript library that supports event-handling and element styling.

The neighborhood data from PGHSNAP was formatted using JSON, and is called when the webpage is loaded. Leaflet.js interprets the coordinates in the file and draws the outlines and color fills of each neighborhood. Different user events, such as selecting an item from the demographic list or clicking a region of the map, trigger jQuery functions that update the visualization. Although our technical implementation could scale to include additional cities or regions, this expansion would require a new assembly of geographic, demographic and Twitter data to support its functions.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

RESULTS: A QUALITATIVE ANALYSIS OF THE VISUALIZATION

In order to get some feedback from people outside of our group, input was solicited from a convenience sample consisting of a close relation ("the user"). We opened up the website on her laptop and asked her to see if she could find anything interesting using the map. The first questions posed related heavily to some errors in the visualization itself, namely:

> The drop-down menu that allowed the user to select which demographics to view on the map was missing any explanation.
>
> The bottom comparison box for the tweets from a given neighborhood was cut off due to the lack of screen real-estate.
>
> The size of a word in the comparison boxes was not explained.

The first two issues were fixed in the final version up on the web, however, the user did a good job of guessing the meaning of the word sizes, and so this was left as is. After fixing these issues, we re-opened the website on her machine and asked the same question. One comment of interest provided by the user is discussed here, regarding the neighborhood Lincoln-Lemington-Belmar which the user came across via clicking around on the map. The user stated

(paraphrased), "I know Lincoln-Lemington-Belmar is heavily African American and dangerous – why were people talking about Waterworks?" This comment suggests three points of interest.

First, the user noted the demographic make-up of Lincoln-Lemington-Belmar without making use of the tool in the visualization that could have explicitly given her this information. This suggests that we could have done a better job of drawing attention to this feature of the visualization. Interestingly, however, the user was hesitant to allow the above comment to be used in this report, stating that, "If I was wrong, it'd sound like I was really racist!" Implicitly, this claim suggests that having raw data available in the visualization for users to stake their claims on may lead them to be more comfortable making claims about neighborhoods that they have previously stereotyped.

At the same time, this data can serve as a means to cue users to observe discontinuities between neighborhood boundaries, demographics and language used on Twitter in a given area. Though using her own mental model of Lincoln-Lemington-Belmar, the user was drawn to make a statement about this area precisely because of a perceived disconnect between the words deemed relevant by Twitter ("Waterworks") and the demographics of the area. A variety of possible conclusions could be made here–we consider three in particular.

First it seems that, in accordance with the findings of (Cranshaw et al., 2012), social media data may serve to suggest a better representation of neighborhoods than even handcrafted data grounded in census demographics. This would seem to be particularly the case if one buys the Hipp et al., argument that neighborhoods should be defined with explicitly, as opposed to implicitly, social characteristics. Second, it seems that there is some grounding for the expectation that resident's mental models of their city are a powerful tool to grasp anomalies and intricacies of the city that may not be available with the current tools (i.e. census and social media data) used to define neighborhoods. Finally, as the user observed that "resident" terms aligned more closely with her mental model of Lincoln-Lemington-Belmar than "non-resident" terms from Twitter, there is some evidence that differentiating actions of "regulars" in a neighborhood from that of "visitors" may provide interesting conclusions beyond those made by simply aggregating over all social media data accumulated in a given area.

CONCLUSION

As we look back to our original hypothesis, we can clearly see the merits of combining multiple rich datasets in a map-based visualization in order to facilitate sensemaking. Although we did not get a chance to explore all of the possible uses and insights that can be gained from the final visualization, we certainly got a flavor of the potential value and impact of the project. Social media data appears to be a powerful indicator of the sense of a neighborhood, particularly when it is categorized according to "resident" and "non-resident"

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

status, and we are enthused about the far-reaching effects of visualizations like ours. We see a significant opportunity in this area of helping people make sense of neighborhood data to challenge or enhance their mental models of a certain location and make new connections and conclusions about the relationships between various datasets. An interactive, map-based visualization seems to be a good start for working toward this end goal.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University

REFERENCES

Bendix, F., Kosara, R., & Hauser, H. (2005). Parallel sets: visual analysis of categorical data. *In Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on* (pp. 133-140).

Bivand, R., Lewin-Koh, N., 2013. maptools: Tools for reading and handling spatial objects.

Butts, C.T., Acton, R.M., Hipp, J.R., Nagle, N.N., 2012. Geographical variability and network structure. Soc. Networks 34, 82–100.

Chau, D., Kittur, Hong, J., Faloutsos, C. (in press). Making Sense of Large Network Data: Combining Rich User Interaction and Machine Learning. To appear in *CHI 2011.*

Cho, E., Myers, S.A., Leskovec, J., 2011. Friendship and mobility: user movement in location-based social networks, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11. ACM, New York, NY, USA, pp. 1082–1090.

Cranshaw, J., Schwartz, R., Hong, J.I., Sadeh, N., 2012. The Livehoods Project: Utilizing Social Media to Understand the Dynamics of a City, in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media, ICWSM '12. AAAI.

Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P., 2010. A latent variable model for geographic lexical variation, in: Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, EMNLP '10. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1277–1287.

Eisenstein, J., O'Connor, B., Smith, N.A., Xing, E.P., 2012. Mapping the geographical diffusion of new words. arXiv:1210.5268.

Eisenstein, J., Smith, N.A., Xing, E.P., 2011. Discovering sociolinguistic associations with structured sparsity, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1, HLT '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 1365–1374.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A., 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments, in: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2, HLT '11. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 42–47.

González, M.C., Hidalgo, C.A., Barabási, A.-L., 2008. Understanding individual human mobility patterns. Nature 453, 779–782.

Healey, C. Perception in visualization. (2009).

Herman, I., Melançon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1), 24–43.

Hipp, J.R., Faris, R.W., Boessen, A., 2012. Measuring "neighborhood": Constructing network neighborhoods. Soc. Networks 34, 128 – 140.

Kumar, S., Barbier, G., Abbasi, M.A., Liu, H., 2011. Tweettracker: An analysis tool for humanitarian and disaster relief, in: Fifth International AAAI Conference on Weblogs and Social Media, ICWSM.

Munro, R., Manning, C.D., 2012. Short message communications: users, topics, and in-language processing, in: Proceedings of the 2nd ACM Symposium on Computing for Development, ACM DEV '12. ACM, New York, NY, USA, pp. 4:1–4:10.

Noulas, A., Scellato, S., Mascolo, C., Pontil, M., 2011. Exploiting semantic annotations for clustering geographic areas and users in location-based social networks. Proc Smw'11.

Perer, A., & Shneiderman, B. (2006). Balancing systematic and flexible exploration of social networks. *IEEE Transactions on Visualization and Computer Graphics*, 693–700.

PGHSNAP. City of Pittsburgh, Department of City Planning, 2012. Web. 07 Apr. 2013. <http://www.pittsburghpa.gov/dcp/snap/>.

Tang, K.P., Lin, J., Hong, J.I., Siewiorek, D.P., Sadeh, N., 2010. Rethinking location sharing: exploring the implications of social-driven vs. purpose-driven location sharing, in: Proceedings of the 12th ACM International Conference on Ubiquitous Computing, Ubicomp '10. ACM, New York, NY, USA, pp. 85–94.

Yuan, J., Zheng, Y., Xie, X., 2012. Discovering regions of different functions in a city using human mobility and POIs, in: Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12. ACM, New York, NY, USA, pp. 186–194.

Zheng, Y., Liu, Y., Yuan, J., Xie, X., 2011. Urban computing with taxicabs, in: Proceedings of the 13th International Conference on Ubiquitous Computing. pp. 89–98.

Visit the final project at:
tinyurl.com/mapTwittsburgh

David Baboolall, BHCI
Robyn Hammond, MDes
Kenny Joseph, PhD
Ian Todhunter, MHCI

Sensemaking
Spring 2013
Carnegie Mellon University