Using Digital Trace Data to Identify Regions and Cities

Christa Brelsford*
brelsfordcm@ornl.gov
Oak Ridge National Laboratory
Oak Ridge, TN

Rudy Arthur
R.Arthur@exeter.ac.uk
University of Exeter
Exeter, UK

Gautam Thakur
thakurg@ornl.gov
Oak Ridge National Laboratory
Oak Ridge, TN

Hywel Williams
H.T.P.Williams@exeter.ac.uk
University of Exeter
Exeter, UK

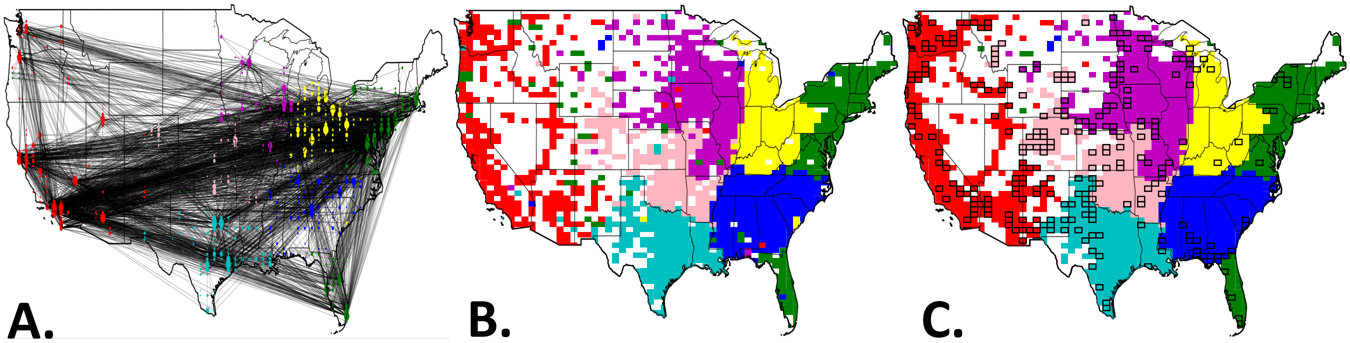

Figure 1: We identify the major regional delineations of the United States, and find them broadly comparable to US census Regions and Divisions. Panel A shows the raw communication based social network, including only edges which represent more than 500 tweets between pairs of nodes. Panel B shows the raw results from the community detection algorithm for all grid cells with greater than 10 tweets. Panel C shows final results when detected communities have been spatially smoothed.

ABSTRACT

A greater understanding of human dynamics as they play out in both physical space and through inter-personal communication is vital for the design and development of intelligent and resilient cities. Physical context provides insight into the space-time distribution of population and their activity patterns, while interpersonal communication can now be measured at the population scale through digital interactions. In this work, we propose a novel method to discover these dynamics. We use a dataset of 72 million tweets to develop a spatially embedded network of communication, and then use community detection algorithms to explore regional and urban delineation in the United States. We compare these results to US census regions and economic and infrastructural networks. We find that the broad spatial delineation of communities and subcommunities is consistent with United States regions, states, and major metropolitan areas. We describe how these methods could be extended to generate a measure of social regions that can be consistently applied anywhere there is a sufficiently rich data source. A deeper understanding of urban social structure measured by spatially embedded communication networks can enable a better understanding of the interactions between urban social and physical contexts. This, in turn, may enable urban managers and policy makers to identify strategies for supporting urban resilience.

CCS CONCEPTS

• **Applied computing → Law, social and behavioral sciences**; *Sociology*; • **Networks**;

KEYWORDS

Networks, Communities, Cities, Digital Trace Data, Twitter

ACM Reference Format:
Christa Brelsford, Rudy Arthur, Gautam Thakur, and Hywel Williams. 2019. Using Digital Trace Data to Identify Regions and Cities. In *2nd ACM SIGSPATIAL International Workshop on Advances on Resilient and Intelligent Cities (ARIC'19), November 5–8, 2019, Chicago, IL, USA.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3356395.3365539

1 INTRODUCTION

What makes a city? What makes a neighborhood? How do communities simultaneously generate and respond to changes in their physical context? Ever-larger shares of the global population now

ARIC'19, November 5–8, 2019, Chicago, IL, USA

ACM ISBN 978-1-4503-6954-1/19/11…$15.00
https://doi.org/10.1145/3356395.3365539

use digital means to enhance and facilitate communication with their real world communities. The digital traces left behind from these communications enable large scale quantitative assessment of what communities exist, descriptions of their spatial properties, and estimations of how they co-evolve with their physical and geo-political context. Current methods for urban and regional delineation center around remote sensing of land surfaces such as impervious surfaces or buildings, spatial analysis of transportation infrastructure, and census-based descriptions of infrastructure and commuter-sheds. Spatially embedded networks of communication generated from digital trace data allow for detection of communities across multiple spatial scales. This is a new and rapidly expanding area of research, with substantial scope for exploration of novel computational methods and science questions regarding social processes, particularly in urban environments. We can explore the spatial and temporal dynamics of community evolution, and measure both the degree of geographic cohesion of these communities as well as explore social teleconnections. These measurements of community dynamics are potentially globally applicable and updatable on an arbitrary timescale, and could may enable a better understanding of how social phenomena worldwide might interact with physical systems or geo-political processes.

The multi-modal nature of social media data permits a variety of previously opaque or separated processes to be observed and linked. For example, spatial networks can be linked to content of communications; teleconnections and physical connections between regions, cities and neighborhoods can be identified and characterized; census data can be linked to social media users by spatialization and demographic inference; and the co-evolution of social processes with geo-physical processes and events can be observed.

Early work looking at spatially embedded social communication networks primarily used cell phone call records. Across a broad range of methods, this literature has thus far found that spatially embedded networks of social communication are highly localized, for cell phone calls [4, 7–9], tweets [10], and facebook friendships [3]. Further, the literature thus far finds that most cases where there are significant geographic discontinuities in identified communities can be explained through the effect of well understood processes: major historical migrations, infrastructure connectivity, or boundary effects due to language differences, social fault lines, or physical or infrastructural impediments. More recently, this tight geographic cohesion observed in almost all studies of spatially embedded networks of social communication has led to the use of these networks to infer regions through network clustering algorithms [1, 12, 13]. However, this literature has not yet used these spatially embedded communication networks to explore urban and regional delineation in the United States, the problem to which we now turn.

2 DATA

We utilized publicly available Tweets using Twitter's Streaming API serviceand selected tweets containing geo-tagged information. Twitter provides geo-tagging that can be based on an exact location or assigned through a pre-selected nearby Twitter place, or both. Twitter Places corresponds to a neighborhood and represented via a bounding box with an array of latitude and longitude coordinates that define the locality of tweet. We setup the data collection for a period of 10 months, between Oct' 2018 through Jun' 2019. The final curated data set consists of 72 million de-identified geo-coded tweets downloaded from the PlanetSense data enclave [11].

3 METHODS

We first isolate Tweets containing place tags. "Places are specific, named locations with corresponding geo coordinates. They can be attached to tweets by specifying a place_id when tweeting. Tweets associated with places are not necessarily issued from that location but could also potentially be about that location."[1]. Place tag information is unique to each tweet: it is not the same as the location optionally described in a user profile. When a user opts into location services their tweets contain place-tags whose co-ordinates are determined from GPS, wifi and cell tower data[2]. Place tags can represent a business, neighborhood, region, or other point of interest.

It was found in [2] that geo-tags, a point based description of user location, represent a very different type of tweet and Twitter user: authors of geo-tagged tweets are primarily bots and users sharing posts from other social media. Consistent with this finding, Twitter is deprecating the precise geo-tag feature for most categories of tweets because it is so little used[3]. Thus, for twitter data going forward, place tags represent the best available measurement of the location of individual users, as they choose to represent it, and so we use place-tags rather than geo-tags.

We further require that each of the tweets in our collection is either a mention or a reply by checking the meta-data associated with each tweet. For every user we identify the grid cell(s) they most commonly tweet from and assign that cell (or cells) as their 'home' location. When a user in cell i mentions or replies to a user in cell j we add 1 to the weight of the link connecting cells i and j. If a user in grid cell i mentions a user k for whom we have no geographic information, we discard that mention. If a user's home is spread across M cells and they mention a user whose home is spread across N cells we add a fraction of a mention to the link between all the cells (there are MN links which are incremented by a fraction $1/MN$).

Once the aggregated network has been constructed the Louvain method [5], a standard stochastic community detection algorithm to identify clusters of densely connected nodes. Since the Louvain method is stochastic we restart it 100 times and choose the clustering which produces the highest modularity score. The modularity score is computed as described in [5], and measures the density of edges inside communities to edges outside communities. The Louvain method is selected for two reasons: it is very fast, and it determines the appropriate number of clusters automatically, instead of requiring selection in advance. We find this approach typically places the majority of grid cells in spatially contiguous clusters.

This approach typically gives contiguous regions which closely correspond to known administrative and social divisions, see Fig 1 panel B. We can see in this figure that within the clusters we sometimes have isolated cells which are assigned to a different cluster than their surrounding cells e.g. the blue cluster in the

[1] https://dev.twitter.com/overview/api/places

[2] https://support.twitter.com/articles/78525#

[3] https://twitter.com/TwitterSupport/status/1141039841993355264

south-east contains some red, green and yellow cells. We also have many cells in sparsely populated regions where we have no data. For some purposes we would like to smooth the regions so they are as coherent as possible, even though the data indicates the assignment is not optimal in terms of modularity. We also would like to assign cluster membership to cells where we have no data. Our spatial smoothing consists of 1) identifying the cells which are assigned to a cluster but are not attached to the largest contiguous polygon of that cluster, 2) identifying the unassigned cells which have at least 4 assigned neighbours, 3) iteratively reassigning these cells to have the same label as the majority of their neighbours until there is no further change possible. This results in the smoothed picture in Fig 1, panel C.

To look at subregions we follow [9]. After identifying regions at the national scale we apply the methodology described above to tweets originating in and mentioning users from each region separately, see e.g. Fig 3 for an example of this hierarchical clustering approach applied to the red cluster in Fig 1.

4 RESULTS

The results presented here are based on the dataset described in section 2. The national level individual scale network constructed contains 3.2 million individual users and 22.9 million relevant @mentions. Once these users are binned into the 64x64 uniform grid shown in Fig 1, this becomes network with 1,558 nodes, which has a mean degree of 474 and a mean weighted degree of 24,000. We use a 64x64 grid to balance computational time and map resolution. We find that coarser grids give very similar results.

The results we find for the full dataset are quite similar when the same analysis is run on a 2 million tweet randomly sampled subset. We still identify seven major regions, with the largest communities in essentially the same location. The communities of the broad mid-western region (pink, cyan and yellow) have somewhat different boundaries than those shown on the full dataset, but together cover generally the same area as the yellow, pink and magenta communities from the full dataset shown in Fig 1B.

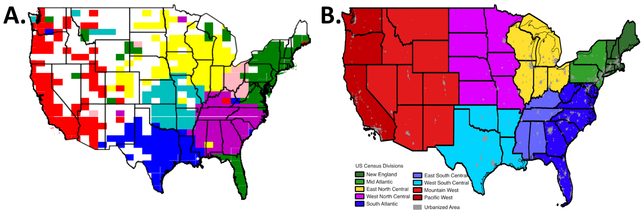

Figure 2: Panel A shows the raw results for a sampled dataset consisting of 2 million tweets drawn from the full dataset. Panel B shows the regions and divisions of the US census.

4.1 National Scale

At the national scale shown in Fig 1, we find seven regions that are broadly consistent with the nine United States Census divisions, Fig 2B. The community we identify in the western US is shown in red in Fig 1. The US census West region (in red in Fig 2B) consists of overlapping states with the addition of Wyoming and Colorado. The Northeast region we find is shown in green in Fig 1 and includes the only robust geographic non-continuity: Florida. The US census Northeast region (in green in Fig 2B) consists of the same green states, but excludes Florida and the states in the Washington DC metro region: Maryland, Delaware, and Virginia. These four states are all included in the South Atlantic division of the South region. The East South Central division and South Atlantic divisions (light and dark blue, respectively) of the US census together roughly correspond to the South region (dark blue) with exceptions for the states that were included in the Northeast cluster. Other communities we detect have similarly close correspondence with US census divisions and regions- with differences between the two of no more than a few states.

The regions we identify are thus generally consistent with one of the most commonly used descriptors of socially meaningful divisions of the United States. Note that while US census divisions are defined based on state boundaries, there is no such constraint for the communities we detect. Nonetheless, Pennsylvania and Missouri are the only two states that clearly are split across detected communities. This suggests that the communities we detect via twitter are identifying communities with meaningful geopolitical content.

4.2 Western United States

The western United States is sparsely populated aside from the coastal regions, and this is reflected in the density of twitter data. Fig 3 shows the detected communities in the Western US. In California, identified communities are broadly consistent with major metropolitan regions and their hinterlands: San Diego is teal, Los Angeles is green, the interior agricultural regions of Bakersfield and Fresno are in pink, and the San Francisco and Sacramento metro regions are combined in red. The states of Washington and Oregon make up the brown community. The yellow cluster includes most of Arizona (Fig 3B). The purple cluster is notable for how closely it follows the interstate highway system (Fig 3C). The blue cluster contains the cities of Santa Fe, Las Vegas, and Reno.

4.3 Southern United States

Fig 4 shows the Southern region of the US. The detected network structure (Fig 4B), shows that the largest clusters are well aligned with major urban areas, and the communities we identify generally center around an urban area and it's within-state hinterlands. The two major exceptions to this outcome are the cities of Memphis, Tennessee and Charlotte, North Carolina. Memphis sits on the border between Tennessee and Mississippi, with some of the metro area located in Mississippi. About 65% of the city's residents are African American, nearly half of the statewide population [6]. By contrast, Mississippi's population is currently 40% African American [6]. From a demographic perspective, Memphis' population is more consistent with Mississippi than it is with Tennessee, and so it's reasonable that twitter users in Memphis communicate more with Mississippi residents than they do with Tennessee residents. Similarly, a substantial share of the Charlotte, NC MSA is located in South Carolina. As a result, it's reasonable that Charlotte and its suburbs are included in the cluster that includes most of South Carolina.

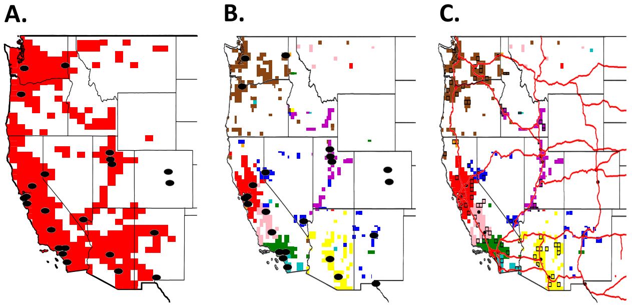

Figure 3: Communities within the Western US. Panel A shows the national scale community. Panel B shows the communities detected within panel A, as well as major US metropolitan areas. Panel C additionally shows the interstate highway system, overlaid on the smoothed detected communities, in order to demonstrate that identified communities are consistent with basic geographic delimiters.

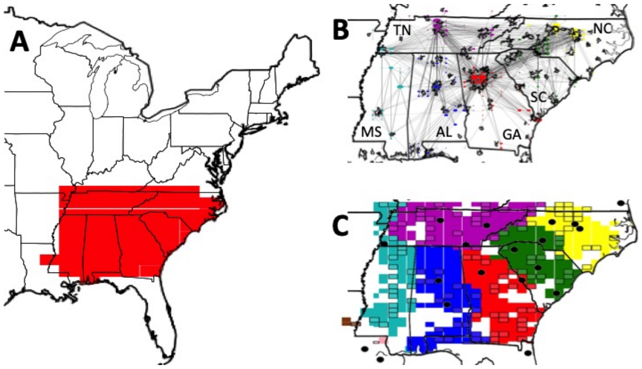

Figure 4: Communities within the Southern US. Panel A shows the identified community at the national scale. Panel B shows the network structure for the community shown panel A, as well as major US metropolitan areas. Panel C shows the smoothed communities.

5 DISCUSSION AND CONCLUSIONS

The results presented here provide suggestive evidence that communities detected in place tagged tweets are representative of real world social relationships both across the United States as a whole and also at the scale of major urban areas.

We assess the accuracy these results by considering both their computational robustness and their external validity. To test the computational robustness of these results, we explore changes in detected communities in response to different randomly selected sub-samples of the core dataset and to variations in the underlying geographic grid, as briefly shown in Fig 2. To consider the external validity of these results, we compare the detected communities to known social, physical, and economic characteristics that we expect influence community structure, as shown in Figs 2-4. The concordance we observe between the twitter based communities and known geo-political regions, infrastructure systems, and demographic characteristics suggests that these communities have external validity. These results could also be more formally compared to other related interaction networks such as physical infrastructure, as hinted at in Fig 3C, trade connections via transportation routes and weights, and other economic input output tables at the city and regional levels. In a further extension, these networks could also be used to describe the spread of ideas, measured by hashtags, memes, or other viral content.

ACKNOWLEDGMENTS

Research sponsored in part by the Laboratory Directed Research and Development Program of Oak Ridge National Laboratory, managed by UT-Battelle, LLC, for the US Department of Energy under contract DE-AC05-00OR22725.

REFERENCES

[1] Rudy Arthur and Hywel TP Williams. 2019. The human geography of Twitter: Quantifying regional identity and inter-region communication in England and Wales. *PloS one* 14, 4 (2019), e0214466.

[2] Rudy Arthur and Hywel TP Williams. 2019. Scaling laws in geo-located Twitter data. *PloS one* 14, 7 (2019), e0218454.

[3] Michael Bailey, Rachel Cao, Theresa Kuchler, Johannes Stroebel, and Arlene Wong. 2018. Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives* 32, 3 (Aug. 2018), 259–280. https://doi.org/10.1257/jep.32.3.259

[4] Vincent D Blondel, Adeline Decuyper, and Gautier Krings. 2015. A survey of results on mobile phone datasets analysis. *EPJ data science* 4, 1 (2015), 10.

[5] Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. 2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment* 2008, 10 (2008), P10008.

[6] United States Census Bureau. 2010. *2010 Summary File 1: Total Population.* Technical Report. Washington D.C.

[7] Paul Expert, Tim S Evans, Vincent D Blondel, and Renaud Lambiotte. 2011. Uncovering space-independent communities in spatial networks. *Proceedings of the National Academy of Sciences* 108, 19 (2011), 7663–7668.

[8] Carlo Ratti, Stanislav Sobolevsky, Francesco Calabrese, Clio Andris, Jonathan Reades, Mauro Martino, Rob Claxton, and Steven H Strogatz. 2010. Redrawing the map of Great Britain from a network of human interactions. *PloS one* 5, 12 (2010), e14248.

[9] Stanislav Sobolevsky, Michael Szell, Riccardo Campari, Thomas Couronné, Zbigniew Smoreda, and Carlo Ratti. 2013. Delineating geographical regions with networks of human interactions in an extensive set of countries. *PloS one* 8, 12 (2013), e81707.

[10] Monica Stephens and Ate Poorthuis. 2015. Follow thy neighbor: Connecting the social and the spatial networks on Twitter. *Computers, Environment and Urban Systems* 53 (2015), 87–95.

[11] Gautam S. Thakur, Budhendra L. Bhaduri, Jesse O. Piburn, Kelly M. Sims, Robert N. Stewart, and Marie L. Urban. 2015. PlanetSense: A Real-time Streaming and Spatio-temporal Analytics Platform for Gathering Geo-spatial Intelligence from Open Source Data. In *Proceedings of the 23rd SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '15).* ACM, New York, NY, USA, Article 11, 4 pages. https://doi.org/10.1145/2820783.2820882

[12] Qi Wang, Nolan Edward Phillips, Mario L Small, and Robert J Sampson. 2018. Urban mobility and neighborhood isolation in America's 50 largest cities. *Proceedings of the National Academy of Sciences* 115, 30 (2018), 7735–7740.

[13] Junjun Yin, Aiman Soliman, Dandong Yin, and Shaowen Wang. 2017. Depicting urban boundaries from a mobility network of spatial interactions: a case study of Great Britain with geo-located Twitter data. *International Journal of Geographical Information Science* 31, 7 (2017), 1293–1313.

A ONLINE RESOURCES

Code used to perform the analysis and generate these figures is open source and available on github at https://github.com/seda-lab/USmap.