Discovering Multi-Scale Community Structures from the Interpersonal Communication Network on Twitter

Caglar Koylu

Abstract Despite the controversies of privacy and ethics, spatially-embedded communication data from widespread and emerging online social networks provide an unprecedented opportunity to study human interactions at the global scale. Detecting communities of individuals who live close by and have strong communication among each other is critical for a variety of application areas such as managing disaster response, controlling disease spread, and developing sustainable urban spaces and infrastructure. The ease of long-distance travel and communication have generated a highly complex network of human interactions, in which long-distance and short-distance ties coexist in multiple scales. Also, there is a hierarchical spatial organization in human interaction networks which reflect historic and socio-political borders. Patterns of human connectivity cross these historic and socio-political borders at multiple geographic scales. Therefore, a comprehensive understanding of human interactions necessitates analysis methods to take into account the complexity introduced by the multi-scale nature of human connectivity. This paper employs a spatially-constrained hierarchical regionalization algorithm to reveal multi-scale community structures in the interpersonal communication network on Twitter. The interpersonal communication network was constructed using a year of reciprocal and geo-located mention tweets in the U.S. between August 2015 and 2016. The results strikingly showed nested borders of cohesive regions at multiple scales, which are inherent to human communication patterns in the regional hierarchy of the U.S. Unsurprisingly, people communicated with others that live nearby, and multi-scale regions overlap with administrative boundaries of the states, cultural and dialectal regions, and topographical features. Furthermore, visualization of interregional communication patterns revealed a variety of spatial connectivity patterns such as poly-centricity, hierarchies, and spanning trees. Discovery of such patterns is essential for understanding of the complex social system that is influenced by long-distance ties.

C. Koylu (✉)
Department of Geographical and Sustainability Sciences, University of Iowa, 316 Jessup Hall, Iowa City, IA, 52242, USA
e-mail: caglar-koylu@uiowa.edu

L. Perez et al. (eds.), *Agent-Based Models and Complexity Science in the Age of Geospatial Big Data*, Advances in Geographic Information Science,
DOI 10.1007/978-3-319-65993-0_7

Keywords Community detection • Hierarchical regionalization • Interpersonal communication • Twitter mentions • Geo-social networks

1 Introduction

Despite the controversies of privacy and ethics, in recent years, publicly available data from location-based social networks (LBSN) such as Twitter, Foursquare, Gowalla, and BrightKite have made it possible, for the first time in human history, to examine human interactions at the global scale. One can infer human interactions through various forms of geo-tagged communication data such as text, photo, video, and check-in locations provided by online platforms. Understanding of human communication and social ties is crucial for addressing societal challenges such as managing disaster response, controlling disease spread, and developing sustainable urban spaces and infrastructure.

Previous studies in LBSN have utilized various forms of communication data to analyze the effect of geographic proximity on social interactions [1–3]; and the structural and geographic characteristics of communication networks at the global scale [4–8]. In addition to understanding global characteristics of communication networks, there has been a growing interest in identifying community structures in human mobility and communication networks [9–11]. Findings of these studies across various themes highlight strong resemblance of human communication and mobility patterns, and the constraining effect of administrative boundaries, topographical features, cultural and linguistic variations on human mobility and communication [12]. However, the ease of long-distance travel and communication have generated a highly complex network of human interactions, in which long-distance and short-distance ties coexist in multiple scales. Also, there is a hierarchical spatial organization in human interaction networks which reflect historic, and socio-political borders. Patterns of human connectivity cross these historic and socio-political borders at multiple geographic scales [9, 10, 13–15]. Therefore, a comprehensive understanding of human interactions necessitates methods that take into account the complexity introduced by the multi-scale nature of human connectivity.

This paper employs a spatially-constrained hierarchical regionalization algorithm to reveal multi-scale community structures in the interpersonal communication network on Twitter. The interpersonal communication network was constructed using a year of reciprocal and geo-located mention tweets in the U.S. between Aug. 2015 and 2016. The results strikingly showed nested borders of cohesive regions at multiple scales, which are inherent to human communication patterns in the regional hierarchy of the U.S. Unsurprisingly, people communicated with others that live nearby, and multi-scale regions overlap with administrative boundaries of the states, cultural and dialectal regions, and topographical features. Furthermore, visualization of interregional communication patterns revealed a variety of spatial connectivity patterns such as poly-centricity, hierarchies, and spanning trees.

2 Related Work

2.1 Distance and Social Interactions

Social ties and communication are constrained by distance, and most of them are geographically local [4]. Deville et al. [16] have shown a great similarity between communication and mobility patterns, and explain the spatial dependencies by a scaling relationship using power laws. Similarly, Emmerich et al. [17] analyzed a variety of spatially-embedded networks such as the Internet, power grid, transportation and communication networks, and found that spatial constraints are relevant, and the relationship between topological and geographic distance varies by dimension and scaling factors. Von Landesberger et al. [18] introduced a flow clustering and visualization approach to identify spatiotemporal variation in the mobility and communication patterns from tweets and phone call records. Von Landesberger et al. [18] found similarities in spatiotemporal patterns such as movements and communication directed from/to central locations given a particular cycle (e.g., daily, weekly). McGee et al. [19] analyzed the effect of distance on the strength of ties, and classified Twitter's utility both as a social network of geographically nearby friends, and as a news distribution network of individuals that live far apart. Higher intensity of communication has also been found to be associated with external factors such as gender, demographics, and socio-economic status. By analyzing 30 billion online conversations, Leskovec and Horvitz [6] found that people tend to communicate more with each other when they have similar age, language, and location; and cross-gender conversations are both more frequent and of longer duration than conversations with the same gender.

Different forms of communication data have been analyzed to examine geographic and structural characteristics of human communication. Krings et al. [20] and Lambiotte et al. [21] revealed that the communication intensity between two cities can be estimated as a function of population, distance, and predominant language using phone call records. Barnett et al. [22] also analyzed phone call records and found that the relationship between homophilly and spatial autocorrelation is amplified in places with high density of individuals. Garcia-Gavilanes et al. [23] studied Twitter user mention network, and found that the probability of two user mentioning each other correlates with power distance. Several studies [24–26] have shown similar findings, and revealed that user mentions on Twitter occur between users that are in close geographic proximity. In addition to distance, Garcia-Gavilanes et al. [23] incorporated economic, cultural and social variables to predict the volume of communication flows between countries. Herdagdelen et al. [27] analyzed social, political and geographic characteristics of news-sharing communities on Twitter, and defined social groups based on local, national and global level. By analyzing a large Twitter dataset, Groh et al. [28] found that (1) the social tie strength decreases as expected with increasing spatial distance among users (2) the information value decreases when the tie strength increases; and (3) the value of information is independent from the distance.

2.2 Community Structures in Spatial Networks

In a network, a community is defined as a set of nodes (individuals) in which the density of connections is stronger internally within the community than it is externally with the individuals from different communities [29]. Community detection algorithms without explicit spatial constraints [30] can be applied to identify communities in spatial networks, which may be multi-part (split) in geography. Various modularity-based community detection algorithms have been used to discover community structures in networks of human mobility [31], commuting [32], telephone call records [10], friendship networks [11], twitter [33, 34], and credit card transactions [35]. Communities discovered by these studies are often geographically confined to nearby regions, however, some of them are multi-part in geographic space. To bridge the geographical and network aspects of communities, Croitoru et al. [36] integrated Louvain and density clustering methods to identify and link community structures in the network (cyber) space and geographic space. Similarly, gravity models have been applied in non-spatial and modularity-based community detection algorithms [37] to estimate expected flows as a function of geographic distance, and derive geographically cohesive community structures. Alternatively, one can embed spatial constraints in community detection to partition a spatial network into smaller sets of contiguous nodes or functional regions that are densely connected internally. In this paper, a spatially-constrained hierarchical regionalization algorithm [9] is used to reveal multi-scale community structures in the spatially-embedded reciprocal mention network.

3 Data and Network Extraction

Geo-located tweets in the Contiguous U.S. between Aug. 1, 2015 and Aug. 1, 2016 were collected using the Twitter Streaming API. Location of tweets are available in two different levels of granularity: exact geographic coordinates, or in a descriptive manner by listing of a place name such as a city. Stefanidis [38] reported that 0.5 and 3% of the tweets had precise coordinates over a period of two years prior to 2013, and also highlighted that the use of precise coordinates increased to 16% during events such as Fukushima disaster in Japan. The dataset used in this paper included 14% of the tweets with precise geographic coordinates, which could potentially be attributed to increasing adoption of mobile technology. In this paper, tweets with both exact geographic coordinates and place names that corresponded to an area at least at city scale were used. Therefore, place names that were at the state or country level, which corresponded to 18% of the tweets with place names, were excluded. As a result, the dataset of tweets with exact coordinates and place names that are at least at city level, consisted of 700 million tweets, and 6.6 million users.

Communication between Twitter users is handled through a set of functions. Follower, favorite and retweet functions are useful for modeling information

diffusion, whereas mentions and replies allow users to join conversations on Twitter, wherein direct personal communication could be extracted [36]. A reply is a response to another user's tweet that begins with the @username of the original poster, a mention is a tweet that contains another user's @username anywhere in the body of the message. In a user mention, the tweet includes only the location of the sender who mentions another user (recipient), and a representative location of the recipient in a mention can be derived only if the recipient has at least one geo-located tweet in the sample. Also, since individuals are mobile, locations of tweets from each user are variable across space. In this paper, tweet locations were overlaid with census data (e.g., county boundaries) to identify a home area for each user based on the most frequent tweet location. Another commonly used strategy could be to determine the home location based on tweets posted at night time where individuals are assumed to be home. In this paper, only the reciprocal mention pairs, or in other words, back-and-forth conversations [37] were used while the tweets that were not replied were disregarded.

A data cleaning procedure was performed prior to constructing the geo-located user mention network on Twitter. Using the metadata provided by the Twitter Streaming API, the following tweets and users were filtered out: (1) the tweets authored by non-personal user accounts such as news feeds, weather and emergency reports, and external applications such as Foursquare and Instagram (2) users with more than 3000 followers to prevent any bias caused by a large number of user mentions attracted by a few users, i.e., celebrities [39]. After the cleaning process, the number of tweets decreased to 290 million (42%). Of these 290 million tweets, 221 million (76%) included a user mention. There were 4.7 million users who were mentioned in a tweet at least once.

After the initial data cleaning, the following steps were performed to extract the reciprocal mention network. First, a spatially embedded individual-to-individual reciprocal mention network was constructed by taking into account the tweets of users who both send and receive messages between each other. Of the 221 million mention tweets, 71 million tweets (32%) corresponded to tweets exchanged between users that both users' home county can be located. After further filtering to obtain reciprocal mentions, the number of tweets was reduced to 33 million (46% of geo-located mentions). The individual reciprocal pairs were then aggregated into a county-to-county network by using the most frequent county location for each user. In the county-to-county network, a link illustrates the total number of reciprocal pairs between two counties.

4 Methodology

A spatially-constrained hierarchical regionalization algorithm [9] was employed to reveal multi-scale community structures in the spatially-embedded reciprocal mention network. The regionalization method produces a hierarchy of spatially contiguous regions, where there are more flows within regions than across regions.

First, a modularity measure of connection strength was computed rather than using the raw flow counts (reciprocal pairs) between each pair of locations. This step is necessary to remove the effect of population by calculating the difference between the actual flow and the expected volume of flow for each pair of locations (counties). While a variety of statistical measures can be used to calculate the expected volume of reciprocal pairs, the following formula that is based on an adjusted flow volume was employed.

$$\mathrm{EP}\,(\mathrm{O},\mathrm{D}) = \mathrm{F_O F_D f}\,(\mathrm{O},\mathrm{D}) \,/ \left(\mathrm{F_s^2} - \sum\nolimits_{i=0}^{n} F_i^2\right)$$

where EP (O, D) is the expected number of reciprocal pairs between origin O and destination D, F_O is the number of reciprocal pairs between county O and its connections, F_D is the number of reciprocal pairs between county D and its connections, f (O, D) is the number of reciprocal pairs between county O and county D, F_S is the number of reciprocal pairs between all counties, and $\sum_{i=0}^{n} {F_i}^2$ is used to remove within-county expectations. Finally, modularity of a link O-D is calculated as:

$$\mathrm{MOD}\,(\mathrm{O},\mathrm{D}) = \mathrm{AP} - \mathrm{EP}$$

where AP is actual number of pairs, and EP is expected number of pairs on link O–D. Using this formula, the raw counts of reciprocal pairs were transformed into a county-to-county modularity graph, in which the weight of a link represents the modularity between two counties. If modularity value is positive the link is considered to be above expectation, if the value is negative the link is below expectation. Next, a full-order average linkage algorithm (ALK) [40] was employed to construct a set of spatially contiguous regions. One can find the algorithmic details of the clustering method in [40]. The average linkage algorithm is a clustering method which is used to build a hierarchy of spatially contiguous clusters by iteratively merging the most connected adjacent clusters. The method outputs a spatially contiguous tree, where each edge connects two geographic neighbors and the entire tree is consistent with the cluster hierarchy. Next, each region in the spatially contiguous tree was partitioned into two regions based on an objective function. Partitioning starts downward from the top of the clustering tree by removing edges. To obtain k regions, (k−1) edges must be removed. For example, four edges must be removed from the initial spatially contiguous tree to derive a five-region partition. To derive k regions, a hierarch of k sets of region partitions are obtained. Each of these sets corresponds to a hierarchical level and is embedded in the next higher level of region partition. Given two regions generated at each level of the hierarchy, a fine-tuning procedure [9] was performed to modify the boundaries by moving locations from one region to other to further optimize the objectives. In this paper, two objectives were used: (1) maximizing within-region modularity (2) maximizing compactness for each region. The modularity is the sum of flow-expectation difference for each pair of units inside a region and for all

regions. Different from the original algorithm [9], we used hierarchical expectation by recalculating the marginal flows for the new region division after each edge removal. For example, if an edge removal partitions ten spatial objects into two regions, region A with three and region B with seven; the marginal flows of the three locations in A is recalculated as the marginal flows within A, and the same applies to region B. Therefore, the marginal flows and flow totals of locations in both regions are dynamically updated according to which region they belong to [41]. The compactness of a region was calculated using the Relative Distance Variance [42, 43], which was found to outperform the other measures of compactness [44]:

$$Compactness = \sqrt{\frac{Area}{2\pi\left(\sigma_x^2 + \sigma_y^2\right)}}$$

where Area is the area of the shape, and σ_x^2 and σ_y^2 represent the variance of the distances between the centroid of the shape, and the x and y coordinate pairs that define the boundary of the shape.

5 Results

5.1 Network Characteristics and the Distance Effect

Individual-to-individual reciprocal mention network consisted of 1,539,396 users (nodes) who participated in at least one conversation. There were 2,621,831 undirected edges, where each edge illustrates a reciprocal pair of users who communicated with each other at least once. Despite the extensive filtering process, the reciprocal communication network is still well connected [45]. The largest connected component consisted of 1,271,530 users (83%) and 2,424,224 edges (92%). This means that 83% of the individuals are connected with each other by a varying number of steps, and an individual has 1.9 connections on average. Figure 1 illustrates the cumulative density of reciprocal pairs by geographic distance. While 50% of the reciprocal communication happened within the same county, 77% happened within the same state. This finding agrees with the previous work in that individuals who engage in conversations are strongly constrained by geographic space.

5.2 Multi-scale Community Structures

The individual-to-individual network was aggregated into to county-to-county network of reciprocal communication, and the regionalization algorithm was performed to derive a hierarchy of regions from 1 to 48. The partition with 48 regions

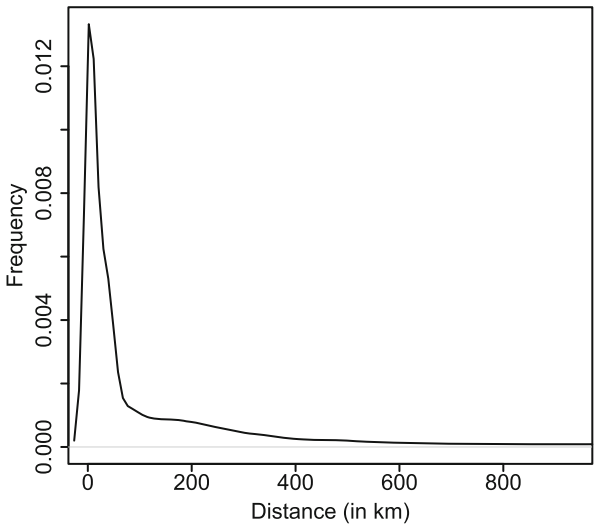

Fig. 1 Frequency of geographical distances among reciprocal pairs. While 50% of the reciprocal pairs were within the same county, 77% were within the same state

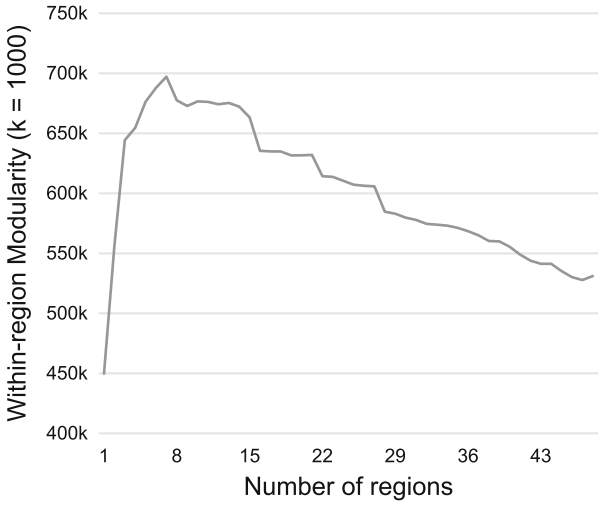

Fig. 2 Total within-region modularity for partitions from 1 to 48 regions in the hierarchy

was selected as the maximum number of regions in the hierarchy in order to compare the data-driven regions to the boundaries of the lower 48 states. The total within-region modularity for region levels from 1 region to 48 regions highlights patterns of communication at multiple scales (Fig. 2). The three-region partition (Fig. 3a) splits the country into East, Central South and Midwest-West divisions. The existence of the eastern region is likely to be influenced by different time zones, which enforce a significant constraint in human communication. The partition with eight regions maximizes the total within-region modularity, and suggests a stable partitioning

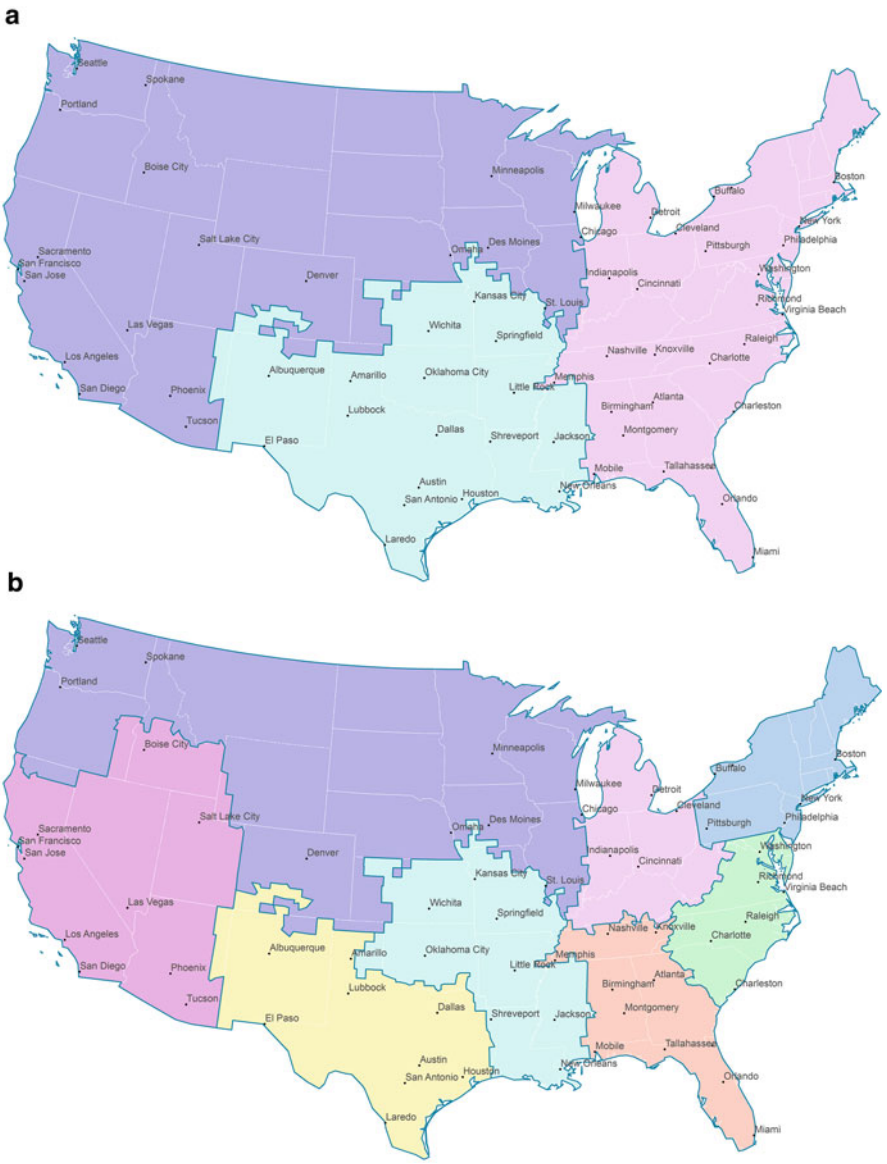

Fig. 3 Hierarchy of interpersonal communication at (**a**) three regions (**b**) eight regions. Partition with eight regions achieves the maximum within-region modularity, and suggests a stable partitioning of the network for the discovery of community structures

of the network for the discovery of community structures (Fig. 3b). Eight-region partition highlights known boundaries as well as unexpected splits that can be explained by socio-economic, cultural and dialectal, and topographical structure of the country. The Northeast region almost exactly matches the designated region by the Census Bureau. This is not surprising as the cultural and political make-up of the Northeast was established long before other regions, and over several centuries. The region was formed by various ethnic groups that were spatially clustered, and tightly connected with each other. On the other hand, the neighboring regions of the Northeast are largely influenced by the natural boundaries such as the Appalachian Mountains and Ohio Valley which act like a physical barrier, and catalyst for human connectivity. Regions in the south were split by the state boundaries of Texas, Tennessee, Louisiana, Alabama, Mississippi and Georgia. The Northwestern region was merged with Midwest, which formed the largest region with a minimal effect of state boundaries. California, Arizona, Nevada, Utah and South of Idaho formed the Western region. Regardless of the diversity in landscape and climate, the Western region contains various racial and ethnic groups that are connected with each other across longer distances.

Figure 4 illustrates 27 regions which were selected based on the most significant drop (slope) in total within-region modularity around the mid-level regions (Fig. 2). This partition highlights previously known splits in regional geography of the U.S. and patches created by metropolitan areas such as Dallas, Los Angeles,

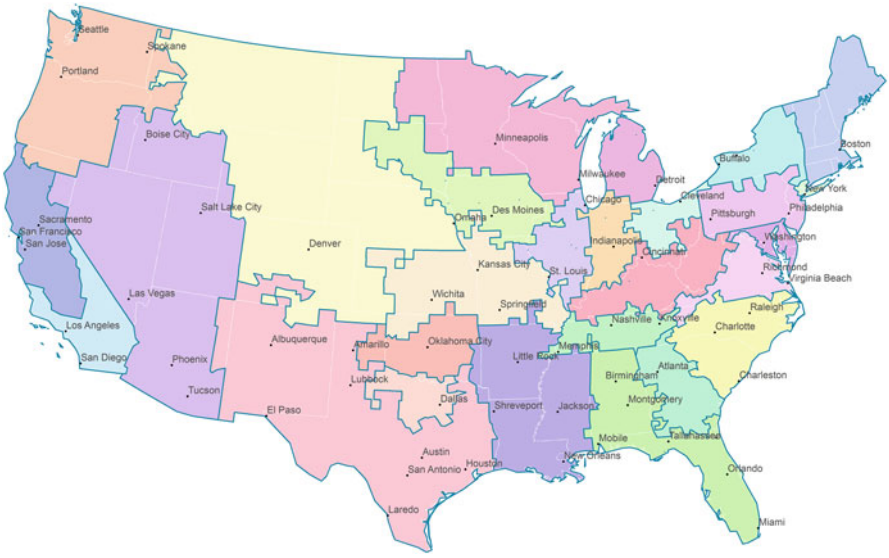

Fig. 4 Interpersonal communication at 27 regions. This partition highlights previously known splits in regional geography of the U.S. such as the division between northern and southern California; Carolinas; Great Lakes region including Minnesota, Wisconsin and Michigan; and patches created by metropolitan areas such as Dallas, Los Angeles, and Washington D.C

and Washington D.C. There are many known splits in this partition such as the division between northern and southern California; Carolinas; Great Lakes region including Minnesota, Wisconsin and Michigan; the combined Kansas-Missouri region centered on the two Kansas Cities and Springfield, Missouri; and the separation of New York City from the rest of New York.

Figure 5a illustrates the partition with 48 regions in order to compare with the boundaries of the lower 48 states of the U.S. While the regions in the east are partitioned into smaller regions, regions in the west are still very large due to lower population, thus, communication sparsity. Figure 5b illustrates the overlap between the state borders and the boundaries of the 48 data-driven regions. The overlap between the state boundaries and 48 data-driven regions was found to be 45%. The states with the most overlap with the region boundaries are Pennsylvania (83%), New Jersey (80%), South Carolina (80%) and Arizona (78%) (Fig. 5b). While some states were split into smaller regions, some were merged to form larger regions that contain multiple states. For example, Texas was split into three regions influenced by the metropolitan cores of Houston, San Antonio, and Dallas. California was split into San Francisco, Central Valley and the rest of California that is pulled by Los Angeles. Florida was split into two regions as a result of the pull effect of the metropolitan areas of Miami, and Northern Florida (i.e., Orlando, and Jacksonville). Small deviations from state borders are caused by the swapping of counties as a result of the pull-effect of a metropolitan core in an adjacent state. Some states were merged to form larger regions that include multiple states. Most of these examples are from the Great Plains. A common characteristic of these regions is the low population density, and thus, less volume of communication.

5.3 Spatial Connectivity Between Regions

Figure 6 illustrates the patterns of spatial connectivity between 48 regions. A modularity threshold of 500 was used to reduce the cluttering and visualize flows that are above expectation (i.e., observed—expected >500). A circle symbol is placed at the population-weighted centroid of a region and the size of the circle is proportional to the within-region modularity. Modularity flows between the regions are represented by flow lines with varying width proportional to the modularity value. Background choropleth map illustrates the region boundaries, and the color value is used to symbolize the density of reciprocal pairs within each region using quantile classification. The structure of flows follow a variety of forms. For example, the Texas Triangle portrays a polycentric pattern, where there are approximately equal strength of connections (flows) between the three metropolitan regions of Houston, San Antonio, and Dallas. On the other hand, connections in California follow a more hierarchical structure, where the hinterland of Los Angeles is tightly connected with the hubs of Central Valley, San Francisco, and Arizona; the connections between these hubs are not as strong. The regions in the East Coast, on the other hand, follow a linear pattern similar to a spanning tree, where each of the

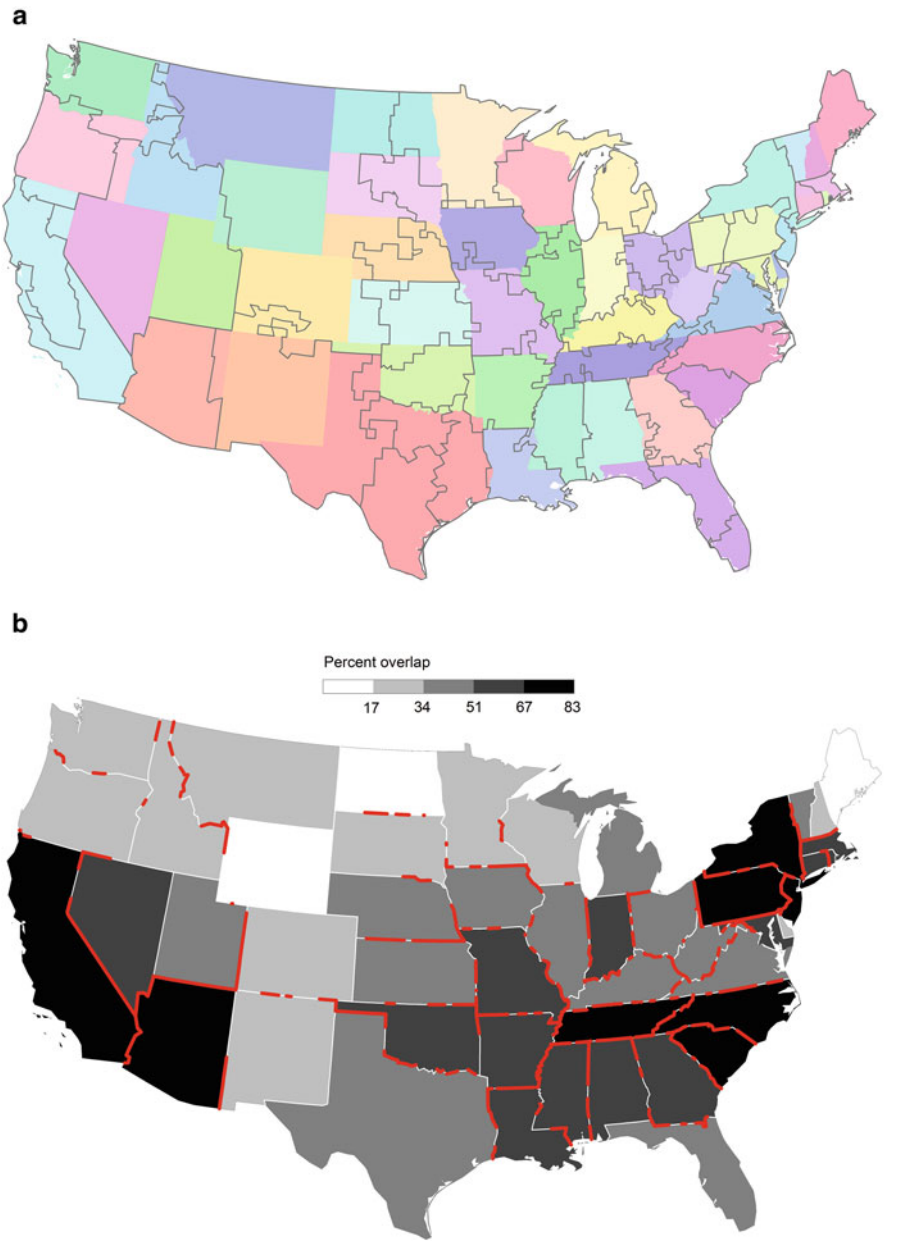

Fig. 5 Comparison of state borders with the boundaries of the 48 data-driven regions of user mention tweets. (**a**) Color-coded areas correspond to the boundary of the states, *black lines* correspond to the boundaries of data-driven regions discovered by the regionalization algorithm. (**b**) *Red lines* illustrate the overlap between the state boundaries and the 48 regions, and the color value symbolizes the percentage of overlap for each state. The overlap between the state boundaries and 48 data-driven regions was found to be 45%

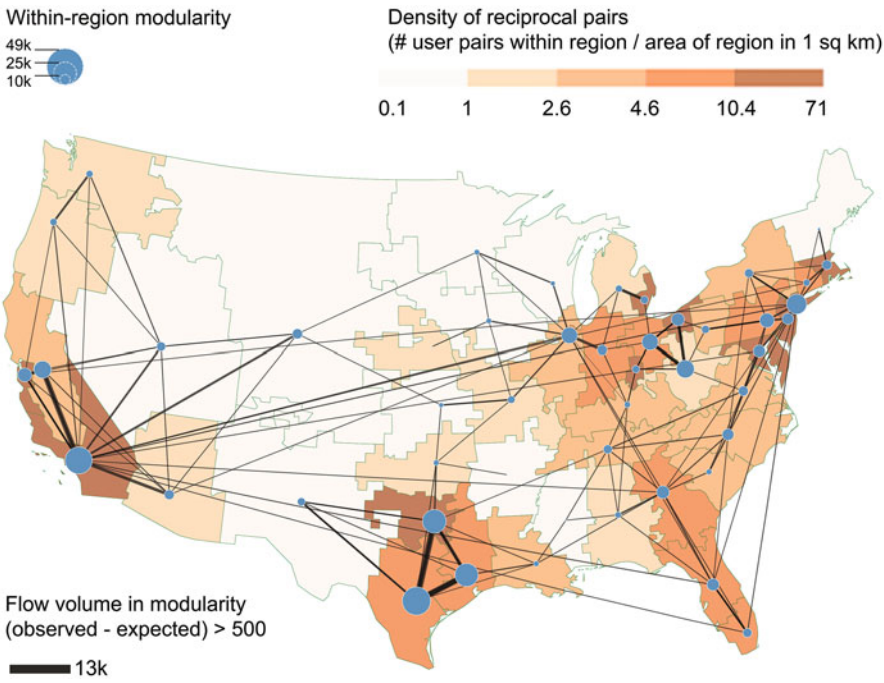

Fig. 6 Reciprocal mentions between 48 regions. A *circle symbol* is placed at the population-weighted centroid based on the number of users within each region, and the size of the circle is proportional to within-region modularity. Modularity flows between the regions are represented by flow lines with varying width proportional to the modularity value. Background choropleth map illustrates the region boundaries, and the color value is used to symbolize the density of reciprocal pairs within each region using quantile classification

regions are strongly connected to one of its close-by neighbors along the east coast. The only exception to this pattern are the big hubs of New York City and New Jersey, which follow a hierarchical pattern. Chicago also follows a hierarchical pattern of connectivity, whereas Cleveland, Columbus and West Virginia follow a polycentric one with strong connections among each other.

6 Discussion and Conclusion

A hierarchical regionalization algorithm was used to identify multi-scale community structures within the interpersonal communication network on Twitter. The results strikingly showed cohesive regions in different scales, which overlap with administrative boundaries of the states, cultural and dialectal regions, and topographical features. Although the regionalization process did not involve state level information, 45% of the state borders overlapped with the data-driven regions,

which is similar to the findings of the previous studies that analyzed a variety of human mobility and communication datasets [9, 13]. Also, the patterns of spatial connectivity between the 48 regions revealed a variety of structural patterns such as poly-centricity, hierarchies, and spanning trees. Discovery of such patterns is essential for understanding of the complex social system that is influenced by long-distance ties.

There are a number of limitations in this study. The first limitation is well-known: demographics of twitter users are not reflective of the general population [46]. Twitter is only a small portion of interpersonal communication which mostly happen in person, through phone calls, text messaging, and video conferencing. However, one can analyze any form of communication data with spatial information in a similar manner without revealing privacy of individuals, and discover community structures in a spatial hierarchy. Although a large volume of geo-located tweets were used, these tweets represent only a sample of all tweets (approximately 1%). Moreover, constrained by opt-in behavior of users for geographic location, a large portion of user mentions was not represented in the datasets used in this study due to the inability to locate all mention pairs. For future work, there is a need to take into account the changing frequency of communication over time. In addition to studying the temporal aspect of the network, there is also a need to examine the semantics of the communication using the content of the tweets. By analyzing the content of the conversations using text mining methods one can understand how online conversations vary based on pairs of users in different locations, and different time periods. Such information can help identify both linguistic and topical variation across regions, and improve our understanding of complex semantics in human communication.

References

1. Backstrom L, Sun E, Marlow C (2010) Find me if you can: improving geographical prediction with social and spatial proximity. In: Proceedings of the 19th international conference on World wide web. pp. 61–70. ACM
2. Mok D, Wellman B, Carrasco J (2010) Does distance matter in the age of the Internet? Urban Stud 47:2747–2783
3. Garcia-Gavilanes R, Mejova Y, Quercia D (2014) Twitter ain't without frontiers: economic, social, and cultural boundaries in international communication. In: Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing, p. 1511–1522. ACM, Baltimore, Maryland, USA
4. Takhteyev Y, Gruzd A, Wellman B (2012) Geography of Twitter networks. Soc Networks 34:73–81
5. Yardi S, Boyd D (2010) Tweeting from the Town Square: Measuring geographic local networks. In: ICWSM
6. Leskovec J, Horvitz E (2014) Geospatial structure of a planetary-scale social network. IEEE Trans Comput Soc Syst 1:156–163
7. Park P, Weber I, Mejova Y, Macy M (2013) The mesh of civilizations and international email flows. In: WebSci 2013 Proceedings. ACM

8. Kylasa SB, Kollias G, Grama A Social ties and checkin sites: connections and latent structures in location based social networks. In: Proceedings of the 2015 IEEE/ACM international conference on advances in social networks analysis and mining 2015, pp. 194–201. ACM
9. Guo D (2009) Flow mapping and multivariate visualization of large spatial interaction data. IEEE Trans Visual Comp Grap 15:1041–1048
10. Sobolevsky S, Szell M, Campari R, Couronné T, Smoreda Z, Ratti C (2013) Delineating geographical regions with networks of human interactions in an extensive set of countries. PLoS One 8:e81707
11. Chen Y, Xu J, Xu MZ (2015) Finding community structure in spatially constrained complex networks. Int J Geogr Inf Sci 29:889–911
12. Ratti C, Sobolevsky S, Calabrese F, Andris C, Reades J, Martino M, Claxton R, Strogatz SH (2010) Redrawing the map of Great Britain from a network of human interactions. PLoS One 5:e14248
13. Thiemann C, Theis F, Grady D, Brune R, Brockmann D (2010) The structure of borders in a small world. PLoS One 5:e15422
14. Slater PB (1975) Hierarchical regionalization of RSFSR administrative units using 1966-69 migration data. Soviet Geog Rev. Transl 16:453–465
15. Grauwin S, Szell M, Sobolevsky S, Hövel P, Simini F, Vanhoof M, Smoreda Z, Barabási A-L, Ratti C (2017) Identifying and modeling the structural discontinuities of human interactions. Sci Rep 7:46677
16. Deville P, Song CM, Eagle N, Blondel VD, Barabasi AL, Wang DS (2016) Scaling identity connects human mobility and social interactions. Proc Natl Acad Sci U S A 113:7047–7052
17. Emmerich T, Bunde A, Havlin S, Li G, Li D (2013) Complex networks embedded in space: dimension and scaling relations between mass, topological distance, and Euclidean distance. Phys Rev. E 87:032802
18. von Landesberger T, Brodkorb F, Roskosch P, Andrienko N, Andrienko G, Kerren A (2016) Mobilitygraphs: visual analysis of mass mobility dynamics via spatio-temporal graphs and clustering. IEEE Trans Vis Comput Graph 22:11–20
19. McGee J, Caverlee JA, Cheng Z (2011) A geographic study of tie strength in social media. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 2333–2336. ACM
20. Krings G, Calabrese F, Ratti C, Blondel VD (2009) Urban gravity: a model for inter-city telecommunication flows. J Stat Mech Theory Exp 2009(07):L07003
21. Lambiotte R, Blondel VD, De Kerchove C, Huens E, Prieur C, Smoreda Z, Van Dooren P (2008) Geographical dispersal of mobile communication networks. Phys A Statis Mech Appl 387:5317–5325
22. Barnett I, Khanna T, Onnela J-P (2016) Social and spatial clustering of people at humanity's largest gathering. PLoS One 11:e0156794
23. Garcia-Gavilanes R, Quercia D, Jaimes A (2013) Cultural dimensions in twitter: time, individualism and power. In: International AAAI conference on weblogs and social media
24. Yamaguchi Y, Amagasa T, Kitagawa H (2013) Landmark-based user location inference in social media. In: Proceedings of the first ACM conference on Online social networks, pp. 223–234. ACM
25. Jurgens D (2013) That's what friends are for: inferring location in online social media platforms based on social relationships. ICWSM 13:273–282
26. Compton R, Jurgens D, Allen D (2014) Geotagging one hundred million twitter accounts with total variation minimization. In: 2014 IEEE international conference on big data (big data), pp. 393–401. IEEE
27. HerdaĞdelen A, Zuo W, Gard-Murray A, Bar-Yam Y (2013) An exploration of social identity: the geography and politics of news-sharing communities in twitter. Complexity 19:10–20
28. Groh G, Straub F, Eicher J, Grob D (2014) Geographic aspects of tie strength and value of information in social networking. p. 1–10. ACM
29. Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99:7821–7826

30. Clauset A, Newman MEJ, Moore C (2004) Finding community structure in very large networks. Phys Rev. E 70:066111
31. Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located Twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41:260–271
32. Nelson GD, Rae A (2016) An economic geography of the United States: from commutes to megaregions. PLoS One 11:e0166083
33. Kallus Z, Barankai N, Szule J, Vattay G (2015) Spatial fingerprints of community structure in human interaction network for an extensive set of large-scale regions. PLoS One 10:e0126713
34. Wang F, Mack EA, Maciewjewski R (2017) Analyzing entrepreneurial social networks with big data. Ann Am Assoc Geog 107:130–150
35. Sobolevsky S, Sitko I, des Combes RT, Hawelka B, Arias JM, Ratti C (2014) Money on the move: big data of bank card transactions as the new proxy for human mobility patterns and regional delineation. In: The case of residents and foreign visitors in Spain. 2014 IEEE international congress on big data (bigdata congress), pp. 136–143
36. Croitoru A, Wayant N, Crooks A, Radzikowski J, Stefanidis A (2015) Linking cyber and physical spaces through community detection and clustering in social media feeds. Comput Environ Urban Syst 53:47–64
37. Gao S, Liu Y, Wang Y, Ma X (2013) Discovering spatial interaction communities from mobile phone data. Trans GIS 17:463–481
38. Stefanidis A, Cotnoir A, Croitoru A, Crooks A, Rice M, Radzikowski J (2013) Demarcating new boundaries: mapping virtual polycentric communities through social media content. Cartogr Geogr Inf Sci 40:116–129
39. Lansley G, Longley PA (2016) The geography of Twitter topics in London. Comput Environ Urban Syst 58:85–96
40. Guo D (2008) Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP). Int J Geogr Inf Sci 22:801–823
41. Newman ME, Girvan M (2004) Finding and evaluating community structure in networks. Phys Rev. E 69:026113
42. Bachi R (1973) Geostatistical analysis of territories. Bull Int Stat Ins 45:121–133
43. Blair D, Biss T (1967) The measurement of shape in geography: an appraisal of methods and techniques. Bulletin of Quantitative Data for Geographers. p 45
44. MacEachren AM (1985) Compactness of geographic shape: comparison and evaluation of measures. Geografiska Ann Ser B Human Geogr 67:53
45. Cogan P, Andrews M, Bradonjic M, Kennedy WS, Sala A, Tucci G Reconstruction and analysis of twitter conversation graphs. In: Proceedings of the First ACM international workshop on hot topics on interdisciplinary social networks research, pp. 25–31. ACM
46. Pavalanathan, U., Eisenstein, J. (2015) Confounds and consequences in geotagged twitter data. arXiv:1506.02275