Spectral Analysis of Multiscale Cultural Traits on Twitter

Chandler Squires  
*MIT*, csquires@mit.edu

Nikhil Kunapuli  
*MIT*, nikhil.k.kunapuli@gmail.com

Yaneer Bar-Yam  
*NECSI*, yaneer@necsi.edu

Alfredo Morales  
*MIT*, alfredo.mo.gu@gmail.com

Follow this and additional works at: [https://orb.binghamton.edu/nejcs](https://orb.binghamton.edu/nejcs)

Part of the Computational Linguistics Commons, Numerical Analysis and Computation Commons, Social Media Commons, and the Sociology of Culture Commons

**Recommended Citation**

Squires, Chandler; Kunapuli, Nikhil; Bar-Yam, Yaneer; and Morales, Alfredo (2022) "Spectral Analysis of Multiscale Cultural Traits on Twitter," *Northeast Journal of Complex Systems (NEJCS):* Vol. 4 : No. 2 , Article 2.  
DOI: [10.22191/nejcs/vol4/iss2/2](https://orb.binghamton.edu/nejcs/vol4/iss2/2)

Available at: [https://orb.binghamton.edu/nejcs/vol4/iss2/2](https://orb.binghamton.edu/nejcs/vol4/iss2/2)

This Article is brought to you for free and open access by The Open Repository @ Binghamton (The ORB). It has been accepted for inclusion in Northeast Journal of Complex Systems (NEJCS) by an authorized editor of The Open Repository @ Binghamton (The ORB). For more information, please contact [ORB@binghamton.edu](mailto:ORB@binghamton.edu).
Abstract

Understanding and mapping the emergence and boundaries of cultural areas is a challenge for social sciences. In this paper, we present a method for analyzing the cultural composition of regions via Twitter hashtags. Cultures can be described as distinct combination of traits which we capture via principal component analysis (PCA). We investigate the top 8 PCA components of an area including France, Spain, and Portugal, in terms of the geographic distribution of their hashtag composition. We also discuss relationships between components and the insights those relationships can provide into the structure of a cultural space. Finally, we compare the spatial autocorrelation of PCA components in the Twitter data to similar components resulting from the Axelrod model. We conclude that properties of Twitter behavior can be framed in the discussion of cultural emergence and collective learning.

1 Introduction

A culture is made of individuals who share a set of traits and own the culture [1]. Cultural traits are transmitted through collective learning and imitation, and may include language, behavioral norms, beliefs, religion or literature [2]. Communication is fundamental for the transmission, persistence and mutation of cultural traits. Without communication imitation or reinforcement dynamics would not be possible. Internet and social media have created mechanisms for the spread of information and behaviors instantaneously over large distances [3]. It represents a current challenge for social sciences to understand how social heterogeneity influences the way collective attention and online information are structured. In this work, we look at geo-located Twitter data in order to characterize and model patterns of cultural regions in the data space.
The emergence and evolution of cultures have been explained using models of collective imitation [2], social innovation [4] and evolutionary dynamics [5]. These models agree that cultures result from processes of transmission, variation and differential replication of cultural traits. Variation may occur due to local reinforcement dynamics that amplify initial differences in traits between areas. The ubiquity of a trait increases its chances of replication and variation. Not all traits have the same scale. Some of them are local and other global to its culture and social system.

Cultures are not homogeneous across scales of resolution. At every level, geographic proximity is linked to cultural similarity. Nearby locations are more likely to share language, food, fashion, religion, and behavioral norms than distant locations are. However, the balance between homogeneity and heterogeneity is not uniform across space. Countries may appear to be culturally uniform in the large scale but, as looking closer, a greater heterogeneity is manifest. Some areas exhibit relatively high homogeneity across large distances, while in other areas, there is substantial heterogeneity even at small distances. To understand the structure of cultures and highlight these differences, we must adopt multi-scale methods for analysis and visualization. Social media provides a rich and large data source to observe collective patterns of behavior [6]. The data generated by human activity contains unstructured patterns of behavior that are distinctive of social groups at multiple scales of observation–from the most local activity up to cities and nations [7].

In this work, we capture multiscale cultural distinctions by applying spectral analysis to hashtag usage on Twitter. We apply Principal Component Analysis [8] to a data space created by aggregating hashtags by geographical locations. The structure of the PCA reveals dominant behaviors in the data such as clusters and trends [9]. We found that the hashtag space reflects the cultural divisions of geographical regions, separating countries from one another in the larger scale, as well as regions and local areas in the smaller scale. These cultures show a trend for local homogeneity and global heterogeneity, despite online platforms having created mechanisms for the spread of traits instantaneously over large distances.

We tie our observations back to models of cultural change by subjecting a snapshot of the Axelrod model [2] to the same method of analysis. Within the context of the above discussion, we note that the Axelrod model describes a value-neutral process of change. The success of a particular variant of a trait is similar to a preferential attachment model [10]. The more sites adopt a given trait, more sites are adjacent to a site with that trait. This produces some widespread cultural variants as well as some highly local cultural variants.

The paper is organized as follows. In section 2 we explain the data collection and processing methods. In section 3 we describe the components of the hashtag space and present our modeling framework. We also compare the model results
with the patterns observed in the data. In section 4 we summarize and present our conclusions.

2 Materials and Methods

Data collection

The tweets used in this analysis were collected from the Twitter Streaming API. Twitter samples from different APIs have been previously analyzed [11], and the Streaming API has been found to be adequate for medium to large samples [12] as the one used in this study. The latitude, longitude, and country code of each tweet are given by the API, and hashtags are identified from the text of the tweet via regex.

Biases in geo-located Twitter users have been previously analyzed [13–15]. Twitter users trend younger, wealthier and urban. However, under-aged individuals are under-represented and wealth is not a significant factor among multiple industrialized cities [16]. Despite biases, the points of view collected from social media about popular issues are consistent with traditional surveying methods [17].

Our analysis centers on the region including France, Spain, and Portugal. We chose this region because of its cultural diversity, hosting multiple languages such as French, Spanish, Portuguese and Catalan, among other regional languages and dialects. According to recent reports, Spain and France are among the top 20 countries that use Twitter the most [18] with penetration rates above 15%, and the penetration rate in Portugal has been reported to be above 30% [19].

We collected thousands of geo-located tweets coming from this region during the period of October 2014. Tweets include metadata that provides the hashtags that are present in the text. Hashtags are keywords people use to identify their tweets with ongoing trends or to include idiomatic expressions in their messages.

Data processing

We first create a lattice of “sites”, each a square of size 1° latitude by 1° longitude over the region of analysis. We considered sites ranging in longitude from the westernmost point in Portugal to the easternmost point in France, and ranging in latitude from the southernmost point in Spain to the northernmost point in France. This results in 180 distinct sites. At each site \(i\) and for each hashtag \(h\), we computed \(h_i\) as the number of unique users who tweeted with that hashtag at least once in the given time period from within the 1° latitude by 1° longitude square northeast of the lattice points.

\[1\]https://docs.python.org/3/library/re.html
Due to the considerable number of hashtags, not all of them were included in the final description of a site’s culture. At each site \(i\), we found the set of the top 50 hashtags, \(H_i\). By sampling the top hashtags per location, we avoid some imbalance issues in the data that would arise if we globally sample for top hashtags. The set of hashtags we considered, \(H\), were those in at least one of these sets, i.e. \(H = \bigcup_i H_i\). \(|H| = 65,335\). We then constructed the matrix \(L \in \mathbb{R}^{180 \times |H|}\), with arbitrary ordering of sites and hashtags, where the entry \(L_{i,h} = h_i\). This description results in over-weighted importance for widely-used hashtags. We increase the sensitivity of our feature vectors to local differences by applying a transform well-known to the information retrieval community named TF-IDF (term-frequency inverse-document-frequency). After applying TF-IDF we obtain a new normalized matrix \(L^{tf}\). TF-IDF transformation avoids other issues from imbalanced samples by measuring not only the number of times hashtags appear, but also their local importance. This transformation adjusts that some hashtags appear more frequently in general and some locations have more hashtags and activity than others [20].

Our final representation of the data comes from performing PCA on \(L^{tf}\). The PCA begins by creating a squared covariance matrix \(V \in \mathbb{R}^{180 \times 180}\) with as many rows or columns as the number of locations. Elements \(V_{ij}\) represent correlations between the respective locations’ hashtag vectors which are provided in the \(L^{tf}\) matrix. Then an eigen-vector decomposition is applied to \(V\). The resulting eigenvectors represent the original data set in a reduced set of dimensions or components that describe most of the variance. In a previous study [16] we analyze the geographical structure of hashtags at the city level using both PCA and LDA topic modeling [21]. The results were consistent between both methods.

A previous study developed by Wood et al [22] apply eigenvector decomposition to matrices representing sentiment analysis on a planetary level. They characterize the distribution of social behaviors, such as sentiment, beyond the overly simplistic average value. The components they obtain explain variance in the data not attributable to regular language use. They define \textit{eigenmood} as the small set of components of a matrix and found that two components were enough to describe the overall behavior. The matrices they use to characterize sentiment include a temporal component that we do not consider. Instead, we aggregate hashtag usage over time and create a single vector that represents the behavior in each location. Also, we apply PCA to the squared covariance matrix \(V\), rather than Singular Value Decomposition on the rectangular \(L^{tf}\) matrix.

For the sake of exposition, we reduce the culture vectors to \(\mathbb{R}^8\) (8 components). We denote the matrix of components to hashtag weights \(C \in \mathbb{R}^{8 \times |H|}\). A new matrix of locations to components weights is denote as \(L^C \in \mathbb{R}^{180 \times 8}\). The elements of that matrix show the weights of each component in each location. We will refer to each column of a matrix by indexing, and each row by indexing in lowercase, e.g.
\( \mathbf{L}^C = (\mathbf{L}_1^C \ldots \mathbf{L}_8^C) = (\ell_1^C \ldots \ell_{180}^C)^T \). We chose 8 components because we are able to decompose the space from widespread hashtags down to very local ones. Further components show local differences in distinct regions.

![Figure 1: Strength of top 8 components.](image)

**Figure 1: Strength of top 8 components.** Panel A shows the first 3 components together, where the value of the R, G, and B channels are equal to the weights (normalized to the range \([0,1]\)) of \( C_1 \), \( C_2 \), and \( C_3 \), respectively. Panels B-I show the strengths of components 1-8, respectively, where \( \mathbf{L}_j^C \) is normalized by \( ||\mathbf{L}_j^C||_\infty \). \( C_1 \) divides Spain from France and Portugal. \( C_2 \) divides North of Spain, coastal France, and parts of Switzerland, from the rest of France and Spain. \( C_3 \) divides Portugal from Spain and France. \( C_6 \) divides Catalonia from the rest of northern Spain. \( C_7 \) divides Barcelona from the islands of Catalonia, and \( C_8 \) divides the islands of Catalonia.

3 Results

Characterization of components

Figure[II] shows the projection of the hashtag vector of each location onto the top 8 components. In panel A, we have colored each location with an RGB code based on the projection to the three main components. Distinct colors indicate that regions coexist in different locations of the hashtag space. A clear distinction of five areas can be noticed. In blue, Portugal stands as a distinct culture. Spain is divided by a yellow patch in the south and a purple patch in the North and the Catalan speaking area. This purple area expands across the south of France as well. Most France stands as a distinct culture in green.

The first components explain most of the variance which is given by language and national borders. The other components explain less of the variance and show local granularity, such as the different parts of a country or the differentiation of
Figure 2: **Pairwise comparison of the strengths of selected components.** Each of the dots represents a location, colored the same as in Panel A of Figure [1] The above scatter plots demonstrate a variety of relationships between components. Panel A displays the tight dependence between $C_1$ and $C_2$. Panel B shows the near independence of $C_2$ and $C_3$ and clearly separates Portugal (in blue) from France (in green), Catalonia and other regions (in purple) and the rest of Spain (in yellow). Panel C shows the irrelevance of $C_7$ and $C_8$ for most locations, while separating Barcelona from the islands near Catalonia along $C_7$, and the islands themselves from each other along $C_8$.

neighboring islands. From panels B-D we show the strength of each component on the map. Blue indicates that the hashtag vectors are in the opposite direction of the component, red shows that the hashtag vectors are in the same direction as the component, and white indicates that the strength of the hashtag vector on the component is irrelevant. The first component in panel B differentiates Southern Spain from the rest, in particular from France and Portugal. The second component differentiates Northern Spain, Catalonia and parts of France from the rest. The third component in panel D differentiates Portugal from the rest, and slightly less, northern Spain from southern Spain and France. Panel H shows a differentiation of Catalonia from the Baleares Islands (which share history and local culture), Panel I shows a differentiation between the different islands, in particular between Majorca and Ibiza.

The tight dependence between $C_1$ and $C_2$ relates use of Spanish to being within Spain. Most locations in Spain lie along $C_1 > 0$ (right to the dashed line in Figure 2). Locations closer to $C_1 \approx 1$ correspond to greater use of Spanish phrases, with negative values for sites in Portugal and France, and low values for sites in Catalonia and the Baleares Islands. This relationship describes the fact that high usage of Spanish phrases is a good predictor of being in Spain, although being in Spain is not as good a predictor for using Spanish phrases.

$C_2$ and $C_3$ are more independent and the interpretation of their relationship in-
Figure 3: **30 strongest hashtags in each component and their average entropies.** The columns at the top of the figure show the weights of the top 30 hashtags in each component in descending order of absolute value, from $C_1$ on the left to $C_8$ on the right. The weights in $C_i$ are normalized by the absolute value of the weight farthest from 0, i.e. $||c_i||_\infty$. $C_1$ and $C_2$ have high positive weights for Spanish phrase and acronyms, such as “bn”, “bd”, and “bt” (short for “buenas noches”, “buenos dias”, and “buenas tardes”, respectively). $C_2$ also has high negative weights for cities in the north of Spain $C_6$, $C_7$, and $C_8$ have high weights for “mallorca”, “ibiza”, and “barcelona” with varying polarities. The line plot at the bottom shows the average spatial entropy of the top 30 hashtags in each of the components above. The red line takes the average over only hashtags with positive weights, the blue line over those with negative weights, and the black line over all 30 hashtags. Translations are provided in the SI.
volves less subtleties. Portugal lays close to the hyperplane defined by $C_2$, but is cleanly separated from sites in Spain and France by $C_3$. Together with the first scatter plot, this plot provides the 3 dimensions that dictate the coloring of Figure 1A. The outliers in $C_7 > 0$ correspond to mainland Catalonia, while the outliers in $C_7 < 0$ correspond to the Baleares Islands. These outliers in both directions show that are different from the rest of locations clustered near $C_7 = 0$. Then the islands Ibiza and Majorca get differentiated from each other along $C_8$.

Figure 3 shows the weights of the top 30 hashtags for each component (top panel), along with the average spatial entropy of those hashtags (bottom panel). The hashtags shown in the top panel have the highest weights in their respective component either in the positive (red) or negative (blue) direction. We measure hashtags’ geographic spread by means of spatial entropy. Spatial entropy is defined per hashtag and averaged by component. We first create a vector per hashtag whose elements represent the number of users at each location who posted that hashtag, normalized by the total number of users. Then, we calculate the vector entropy, which is normalized by the maximum possible value coming from the hypothetical uniform distribution. Wide-spread hashtags have higher spatial entropy than highly localized hashtags that will have a lower spatial entropy.

We calculate the spatial entropy of the hashtags shown in the top panel of Figure 3 and measure the component average. The black curve is the average spatial entropy of the top 30 hashtags for each component. The blue and red curves show the results after splitting hashtags according to their direction in the component. The prevalent hashtags in the first components are geographically more wide-spread than the prevalent hashtags from less significant (and higher order) components. The average spatial entropy exhibits a downward trend as we move from more to less significant components. The highest levels of structure in cultural space reflect differences in the usage of hashtags that are widespread, corresponding to the heterogeneity at the coarsest resolution. They distinguish the usage of distinct languages like Spanish, French, Portuguese, or Catalan. Finer levels of structural detail reflect differences in the usage of hashtags that are more spatially isolated, indicating heterogeneity at increasingly local levels. They represent more semantic and local differences in the same language.

The hashtag space shows cultural and linguistic boundaries. These cultural patterns have a multiscale structure, which differentiates large features like language, down to regional and local features, as the case of Catalonia and the Baleares Islands. We could expect that other representations of any population will follow a similar PCA pattern. First a large differentiation of the most abundant feature, followed by smaller differentiation of the most local feature.
Figure 4: **PCA breakdown of Axelrod simulation.** Panel A shows the first 3 components together, where the value of the R, G, and B channels are equal to the weights (normalized to the range [0,1]) of $C_1$, $C_2$, and $C_3$, respectively. Panels B-I show the strengths of components 1-8, respectively. We simulated the Axelrod model of cultural dissemination for 500,000 time steps, with 5 cultural features and 10 variants per feature. We visualize the resulting grid of cultures in the same way as Figure 1.

Model

We propose a model that generates geocultural structure with similar patterns of spatial heterogeneity across scales based on collective learning and imitation of cultural traits. The model is based on the Axelrod model of cultural dissemination. This model assumes that cultures change when sites adopt variants of cultural traits and imitate each other. In the model, each site is more likely to adopt from sites that are similar to them. This phenomenon is called homophily and creates clusters of behavior that become distinct over time.

Each time-step in the model is defined as follows:

- An active agent $k$ is selected at random.
- One of $k$’s neighbours, agent $r$, is selected at random.
- Agents $k$ and $r$ interact with probability proportional to their cultural similarity, expressed in the distance between their cultural vectors.

The culture of each site can be described by an element of $q^{|F|}$, where $F$ is the set of independent cultural traits (e.g. religion, cuisine, dress), and $q$ is the number of variants associated with each trait, assumed to be equal for all traits. As means of example, we study a realization of the Axelrod model that assumes $|F| = 5$.
and \( q = 10 \). Other values of \( F \) and \( q \) yielded similar results. The selected set of parameters provided results with reasonable computational resources and displayed consistent behaviors.

Analogously to the hashtag data, we then use a TF-IDF transform and PCA to find the components of the simulated sites. We first run the model 500,000 time steps. For this propose, we form a new representation of the feature space where each trait is converted to a one-hot vector of length \( q \) and appending these vectors, giving a new representation of culture in \( \{0, 1\}^{|F|q} \). We then apply correlation or distance metrics among locations to obtain the squared covariance matrix that we can use to apply eigenvector decomposition. The results are analogous to the Twitter case. A set of components condense the information about the structure of the feature space and reveal tendencies of behaviors.

Figure 4 shows the results of the PCA analysis on the model results from a grid of size 20 \( \times \) 20. The colored panel in the left shows the RGB code of each component. Colored regions show areas whose cultural vectors are closer to each other than with the rest and create clusters among the modeled cultural traits. The emergence of colored patches is consistent across multiple model simulations. The patchy structure has an analogy to the patterns observed in Figure 1A. In both cases, the patches emerge due to common behaviors among the feature vectors which are captured by the PCA. In Figure 1A vectors are composed by hashtags while in Figure 4 (left) the vectors are composed by artificial binary vectors.

The smaller panels show the prevalence of cultural traits in the space. Red areas indicate that certain trends are prevalent in those regions and not in the blue areas. Just as the data, the first components reveal large cultural areas. Each of them is differentiated from the rest in the main components. As we explore the higher order components, we find that smaller regions start to differentiate from others. We believe that the prevalence of certain traits over large areas is due to a preferential attachment mechanism. The probability of new adoptions of any trait is proportional to the size of the border of the area it belongs to.

The observation of a trend in cultural structure from global to local scales can be made precise by considering the spatial autocorrelation of each component. One measure of spatial autocorrelation is the Gamma index, defined for component \( j \) as

\[
\Gamma(j) = \Sigma_{i,i'} A_{i,i'} L^C_{j,i} L^C_{j,i'}
\]

Figure 5 shows the autocorrelation \( \Gamma \) indices of the first 8 components (rows) of both the model (left) and the data (right). In both cases, the spatial autocorrelation is higher for the main components and decreases as we consider higher order components. This happens because the main components capture the most prevalent features, while higher order components manifest more local cultural traits. This
Figure 5: **Gamma indices of each component on the left and each simulated component on the right.** The blue lines show the distribution of the Gamma index for 10,000 random permutations of the selected weights, while the red line shows the actual Gamma index. In both the actual data and the model, the Gamma index of each component is many standard deviations away from the mean, and there is a general downward trend in the spatial autocorrelation.
multiscale description of cultures is consistent with language use and other cultural properties of social systems. There are differences between the model and the data autocorrelation values. The decline in the spatial autocorrelation of the data is much sharper than in the actual data. Nevertheless, the similar pattern of behavior suggests that hashtags can be used as proxies for observing cultural borders.

The Axelrod model requires that locations copy each other. In the real world, one location is not an entity that can copy another location. Locations aggregate multiple users that are connected to each other by means of the followers network and by living in the same social system. Hashtags correspond to information that flows across that social network and there are multiple analyses of how and why people tweet [23]. However, we can observe that the geographical spread at larger scales can be explained by complex systems models such as Axelrod with simple interaction rules.

4 Conclusion

In this paper we applied spectral analysis to characterize multiscale cultural traits using spatially-resolved Twitter data. The PCA decomposition of the hashtag space enabled the analysis and visualization of the heterogeneity of online activity at multiple scales. The structure of the hashtag space permits defining features of cultural differences, and to view the relationships between the different scales of order.

We also compared patterns of collective hashtag usage with the Axelrod model for cultural emergence from the complex systems literature. We found a similarity between the patterns of spatial autocorrelation of PCA components from both the model and the Twitter data. This means that both data and model show dominant forms of behavior at larger scales that get increasingly localized as we explore higher order components or behaviors with less variance. These models are discussed in terms of cultural and identity formation due to background learning and imitation.

Acknowledgments

We thank NECSI and MIT for their support during this research.

References

[1] A. Birukou, E. Blanzieri, P. Giorgini, and F. Giunchiglia, “A formal definition of culture,” in Models for intercultural collaboration and negotiation, pp. 1–26, Springer, 2013.
[2] R. Axelrod, “The dissemination of culture: A model with local convergence and global polarization,” *Journal of conflict resolution*, vol. 41, no. 2, pp. 203–226, 1997.

[3] A. J. Morales, V. Vavilala, R. M. Benito, and Y. Bar-Yam, “Global patterns of synchronization in human communications,” *Journal of The Royal Society Interface*, vol. 14, no. 128, p. 20161048, 2017.

[4] R. A. Bentley, M. W. Hahn, and S. J. Shennan, “Random drift and culture change,” *Proceedings of the Royal Society of London. Series B: Biological Sciences*, vol. 271, no. 1547, pp. 1443–1450, 2004.

[5] F. D. Neiman, “Stylistic variation in evolutionary perspective: inferences from decorative diversity and interassemblage distance in illinois woodland ceramic assemblages,” *American Antiquity*, pp. 7–36, 1995.

[6] D. Lazer, A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, et al., “Social science. computational social science.,” *Science (New York, NY)*, vol. 323, no. 5915, pp. 721–723, 2009.

[7] L. Hedayatifar, R. A. Rigg, Y. Bar-Yam, and A. J. Morales, “Us social fragmentation at multiple scales,” *Journal of the Royal Society Interface*, vol. 16, no. 159, p. 20190509, 2019.

[8] H. Abdi and L. J. Williams, “Principal component analysis,” *Wiley interdisciplinary reviews: computational statistics*, vol. 2, no. 4, pp. 433–459, 2010.

[9] N. Eagle and A. S. Pentland, “Eigenbehaviors: Identifying structure in routine,” *Behavioral Ecology and Sociobiology*, vol. 63, no. 7, pp. 1057–1066, 2009.

[10] A.-L. Barabási and R. Albert, “Emergence of scaling in random networks,” *science*, vol. 286, no. 5439, pp. 509–512, 1999.

[11] Y. Wang, J. Callan, and B. Zheng, “Should we use the sample? analyzing datasets sampled from twitter’s stream api,” *ACM Trans. Web*, vol. 9, jun 2015.

[12] J. R. Soler, “Twitter research for social scientists: A brief introduction to the benefits, limitations and tools for analysing twitter data,” *Dígitos: Revista de Comunicación Digital*, vol. 1, no. 3, pp. 17–32, 2017.

[13] M. Duggan and J. Brenner, “The demographics of social media users, Pew Research,” tech. rep., Pew Research, Washington, DC, 2013.
[14] A. Mislove, S. Lehmann, Y.-Y. Ahn, J.-P. Onnela, and N. Rosenquist, “Understanding the demographics of Twitter users,” in *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media*, (Palo Alto, CA), pp. 554—557, The AAAI Press, 2011.

[15] Y. Jiang, Z. Li, and X. Ye, “Understanding demographic and socioeconomic bias of geotagged Twitter users at the county level,” *Cartography and Geographic Information Science*, 01 2018.

[16] A. J. Morales, X. Dong, Y. Bar-Yam, and A. ‘Sandy’ Pentland, “Segregation and polarization in urban areas,” *Royal Society Open Science*, vol. 6, no. 10, p. 190573, 2019.

[17] K. Kalimeri, M. G. Beiro, A. Bonanomi, A. Rosina, and C. Cattuto, “Evaluation of biases in self-reported demographic and psychometric information: Traditional versus Facebook-based surveys,” *arXiv preprint arXiv:1901.07876*, 2019.

[18] P. by Statista Research Department and M. 22, “Twitter: Most users by country,” Mar 2022.

[19] T. Scott, “The 10 most used social networks in portugal in 2022,” Jan 2022.

[20] J. Leskovec, A. Rajaraman, and J. D. Ullman, *Mining of Massive Datasets*. New York, NY, USA: Cambridge University Press, 2nd ed., 2014.

[21] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[22] I. B. Wood, P. L. Varela, J. Bollen, L. M. Rocha, and J. Gonçalves-Sá, “Human sexual cycles are driven by culture and match collective moods,” *Scientific reports*, vol. 7, no. 1, pp. 1–11, 2017.

[23] D. M. Romero, B. Meeder, and J. Kleinberg, “Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter,” in *Proceedings of the 20th international conference on World wide web*, pp. 695–704, 2011.
Supporting Information

Hashtag Translation

Many hashtags shown in Figure 3 correspond to places such as Madrid, Barcelona, Portugal, Algarve, Asturias, Coimbra, Huesca, France, Aveyron, Zaragoza, etc. Others correspond to events like 9n. Others to greetings and support to soccer teams. Below the English translation of some hashtags shown in Figure 3.

- adamyeva: Adam and Eve.
- amarillo: yellow.
- bn, buenasnoches: good night.
- bd, buenosdias: good day.
- cerregabenfica: support to soccer team in Portugal.
- chien: dog.
- dgt: General Transit Direction.
- halamadrid: support to soccer team in Spain.
- mar: sea.
- peque: kid.
- perdu: lost.
- salvemosaexcalibur: Let’s save Excalibur.
- voiture: car.

Dimensional reduction

In Figure 6 we present scatter plots comparing each of the top 8 components with each other. The main components differentiate features that are prevalent across multiple locations, while the higher order components differentiate the most local behavior. The first components explain most of the variance. The higher order components explain less of the variance and show local granularity. $C_1$ differentiates Southern Spain from the rest, in particular from France and Portugal. $C_2$ differentiates Northern Spain, Catalonia and parts of France from the rest. $C_3$ differentiates
Portugal from the rest. $C_4$ and higher components differentiate smaller regions and localities and appear as outliers. An example is given in the differentiation of the Baleares Islands in $C_8$. 
Figure 6: **Pairwise comparison of the strengths of selected components.** Each of the dots represents a location, colored the same as in Panel A of Figure 1. The above scatter plots demonstrate a variety of relationships between components. Panel A displays the tight dependence between $C_1$ and $C_2$. Panel B shows the near independence of $C_2$ and $C_3$ and clearly separates Portugal (in blue) from France (in green) from Catalonia (in purple) from the rest of Spain (in yellow). Panel C shows the irrelevance of $C_7$ and $C_8$ for most locations, while separating Barcelona from the islands of Catalonia along $C_7$, and the islands of Catalonia from each other along $C_8$. 