


Language, demographics, emotions, and the structure of online social networks

Kristina Lerman¹  · Luciano G. Marin¹ · Megha Arora² ·
Lucas H. Costa de Lima³ · Emilio Ferrara¹ · David Garcia⁴

Received: 8 August 2017 / Accepted: 7 October 2017 / Published online: 31 October 2017
© The Author(s) 2017. This article is an open access publication

Abstract Social networks affect individuals' economic opportunities and well-being. However, few of the factors thought to shape networks—culture, language, education, and income—were empirically validated at scale. To fill this gap, we collected a large number of social media posts from a major US metropolitan area. By associating these posts with US Census tracts through their locations, we linked socioeconomic indicators to group-level signals extracted from social media, including emotions, language, and online social ties. Our analysis shows that tracts with higher education levels have weaker social ties, but this effect is attenuated for tracts with high ratio of Hispanic residents. Negative emotions are associated with

Electronic supplementary material The online version of this article (doi:[10.1007/s42001-017-0001-x](https://doi.org/10.1007/s42001-017-0001-x)) contains supplementary material, which is available to authorized users.

✉ Kristina Lerman
lerman@isi.edu

Luciano G. Marin
luciano.gallegos@gmail.com

Megha Arora
marora@andrew.cmu.edu

Lucas H. Costa de Lima
lhenriquecl@dcc.ufmg.br

Emilio Ferrara
ferrarae@isi.edu

David Garcia
garcia@csh.ac.at

¹ USC Information Sciences Institute, Marina del Rey, USA

² Carnegie Mellon University, Pittsburgh, USA

³ Federal University of Minas Gerais, Belo Horizonte, Brazil

⁴ Complexity Science Hub Vienna and Medical University of Vienna, Vienna, Austria

more frequent online interactions, or stronger social ties, while positive emotions are associated with weaker ties. These results hold for both Spanish and English tweets, evidencing that language does not affect this relationship between emotion and social ties. Our findings highlight the role of cognitive and demographic factors in online interactions and demonstrate the value of traditional social science sources, like US Census data, within social media studies.

Keywords Social media · Social ties · Emotions · Demographics

Introduction

Humans have evolved large brains, in part to handle the complex cognitive demands of social interactions [1]. The social structures resulting from these interaction confer numerous fitness advantages. Scholars distinguish between two types of social relationships: those representing strong and weak ties. Strong ties are characterized by high frequency of interaction and emotional intimacy that can be found in relationships between family members or close friends. People connected by strong ties share mutual friends [2], forming cohesive social bonds that are essential for providing emotional and material support [3, 4] and creating resilient communities [5]. In contrast, weak ties represent more casual social relationships, characterized by less frequent, less intense interactions, such as those occurring between acquaintances. By bridging otherwise unconnected communities, weak ties expose individuals to novel and diverse information that leads to new job prospects [6] and career opportunities [7, 8]. Online social relationships provide similar benefits to those of the offline relationships, including emotional support and exposure to novel and diverse information [9–12].

How and why do people form different social ties, whether online or offline? Of the few studies that addressed this question, Shea and collaborators examined the relationship between emotions and cognitive social structures [13], i.e., the mental representations individuals form of their social contacts [14]. In a laboratory study, they demonstrated that subjects experiencing positive affect—emotions such as happiness and joy—were able to recall larger and more diverse social contacts than those experiencing negative affect, e.g., sadness. In other words, positive affect was more closely associated with weak ties and negative affect with strong ties in cognitive social structures. The cost of maintaining strong ties is higher than for maintaining weak ties, as they involve higher frequency of interaction, but also their associated benefits are higher, such as emotional support that manifests in the social sharing of negative emotional experiences [4, 15]. As a consequence, frequency of interaction along social ties is positively associated with stronger emotional intensity and negative emotional expression [16].

In addition to psychological factors, social structures also depend on the participants' demographic characteristics, including socioeconomic status [17], culture. A study, which reconstructed a national-scale social network from the phone records of people living in the UK, found that people living in more prosperous regions formed more diverse social networks, linking them to others

living in distinct communities [18]. On the other hand, people living in less prosperous communities formed less diverse, more cohesive social structures. In addition, culture plays a prominent role in shaping the structure of social interaction, but only recently few studies focused on how culture, as well as spoken language, affect online interactions [19, 20]. The link between social structures and place has led researchers examine the role of neighborhoods in shaping communities [21–23].

The present paper examines how group-level psychological, socioeconomic, demographic factors, and language, affect the structure of online social interactions. We restrict our attention to interactions on the Twitter microblogging platform as a first approximation to measuring macroscopic signals through digital traces. We collected a large body of geo-referenced text messages, known as tweets, from a large US metropolitan area. We linked these tweets to US Census tracts through their locations. Census tracts are small regions, on a scale of city blocks, that are relatively homogeneous with respect to population characteristics, economic status, and living conditions. Some of the tweets also contained explicit references to other users through the ‘@’ mention convention, which has been widely adopted on Twitter for conversations. We used mentions to measure the strength of social ties of people tweeting from each tract. Using these data, we study group-level relationships between social ties, the demographic characteristics of the tract, and the emotions expressed by people tweeting from there. We separately measure emotions expressed in English- and Spanish language tweets, enabling us to additionally explore the impact of language on emotions and social tie formation. In addition, people tweeting from one tract often tweeted from other tracts. Since geography is a strong organizing principle, for both offline [24, 25] and online [26–28] social relationships, we measured the spatial diversity to correct for its effect on social network structures in our statistical analyses.

This article illustrates how digital trace data can be a complement to previous studies in online social networks and characterize group-level relationships between the structure of online interactions in urban places and their demographic and socioeconomic characteristics. While unfit to analyze emotions, demographics, and social structures at the individual level, our methods link general properties of the structure of online interactions in groups with their aggregated levels of positive affect. Groups of people who express happier emotions, regardless of language, interact with a more diverse set social contacts, which puts them in a position to access, and potentially take advantage of, novel information. As our social interactions increasingly move online, understanding—and unobtrusively monitoring—online social structures at a macroscopic level is necessary for ensuring equal access to the benefits of social relationships. Although many important caveats exist about generalizing these results to offline social interactions, our work highlights the value of linking social media data to traditional data sources, such as US Census, to drive novel analysis of online behavior and online social structures.

Results

Of the roughly 2000 tracts in Los Angeles (LA), we collected tweets from 1700 tracts. The “Methods” section describes our approach to measuring emotional expression of these tweets. We observe systematic differences between emotions expressed in the tweets posted from different places, the language of the tweets, and the structure of online social interactions, as well as spatial mobility. By combining these data, we created one set with 688 tracts that had at least 15 tweets in both Spanish and English from which we could measure emotional expressions. In spite of that, some of these tracts did not have all demographic or socioeconomic variables available from the US Census and were ignored in further analysis. After cleaning the sample of tracts, as explained more in detail in the “Methods” section, the number of analyzed tracts is 539, comprising a total of more than 28 thousand tweets.

Our regression analysis approach, explained in the “Methods” section, applies a mixed effects model with random intercepts per tract group, to correct for spatial autocorrelations. Regression models include a control term of spatial diversity to correct for the negative correlation with tie strength shown on the left panel of Fig. 1. In this section, we explore these differences and their association with the demographic and cultural characteristics of places from which people tweet.

Demographics and social ties

We initially explored the bivariate correlations between tie strength per tract (see Methods) and the demographic variables measured in the census. The center and right panels of Fig. 1 show, respectively, a negative correlation between tie strength and education (Pearson’s $\rho = -0.338$, CI $[-0.26, -0.41]$) as well as a positive correlation between tie strength and the ratio of Hispanic residents (Pearson’s

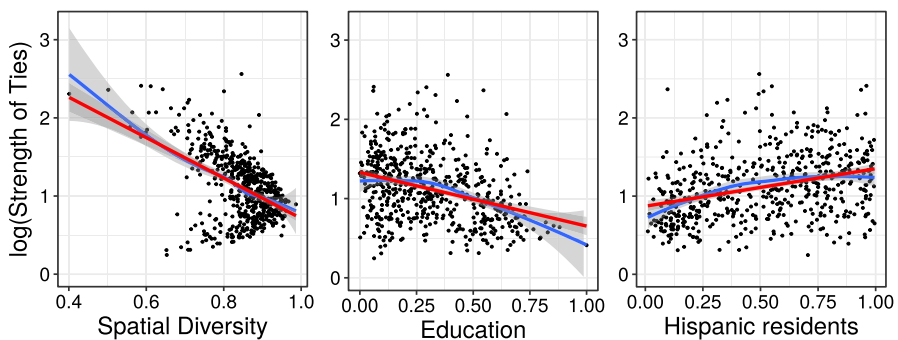


Fig. 1 Tract tie strength as a function of spatial diversity and demographic factors. Black dots show the empirical points, red lines show linear fits, blue lines non-parametric local fits, and shaded regions show prediction errors of model fits. Average tie strength is negatively correlated with spatial diversity, which we include as a control variable in our analysis. Tie strength shows a negative correlation with education, measured as the ratio of residents with a bachelor’s degree, and a positive correlation with the fraction of Hispanic residents

Table 1 Fit results for the mixed effects models of log(tie strength) as a function of demographic variables, controlling for spatial diversity and including random intercepts per tract group

	Model 1	Model 2	Model 3
Education	−0.536*	−1.146***	−1.122***
HispanicRatio	0.255*	−0.013	
Education*hispanicRatio		2.091***	2.062***
spatialDiversity	−2.513***	−2.513***	−2.513***
Age	0.003	0.001	0.001
Intercept	3.105***	3.273***	3.258***
BIC	459.622	452.470	444.097
Var: tractGroup (Intercept)	0.036	0.035	0.035
Var: Residual	0.089	0.087	0.087

*** $p < 0.001$, ** $p < 0.01$,* $p < 0.05$

$\rho = 0.33$, CI [0.25, 0.403]). While this initially indicates that tracts with more educated but fewer Hispanic inhabitants have weaker social ties, we contrast these observations against incremental regression models to verify that our observations are not due to confounds across demographic variables.

Table 1 reports the regression results for the logarithm of tie strength as a linear combination of education and the ratio of Hispanic residents, including control terms for age and spatial diversity. The three models include incremental terms that explore the role of education and Hispanic ratio including interaction terms, as both variables are strongly colinear (Pearson's $\rho = -0.865$, CI [−0.885, 0.842]). The analysis shows a negative association between education and tie strength. On average, tracts with a higher ratio of residents with a bachelor's degree have weaker social ties in Twitter.

The positive relationship between the ratio of Hispanic residents appears in the model without interactions (Model 1) but adding interaction terms shows its dependence on education (Model 2). The positive marginal effect of the ratio of Hispanic residents is not significant, but the interaction term with education is significant and positive. This has two implications: first, that the positive association between the ratio of Hispanic residents and tie strength is only present for higher levels of education, and second that the negative effect of education in tie strength is counterbalanced by the positive interaction with the ratio of Hispanic residents. This way, for ratios of Hispanic residents above 0.6, we can expect a positive relationship between education and tie strength. This is further evidenced by the fact that the best model in terms of Bayesian Information Criterion is the model with an interaction term but no direct effect of the ratio of Hispanic residents on tie strength (Model 3).

We repeated the fit of the full model replacing the education variable with two of its strongest correlates: the employment rate and the logarithm of the median household income. These two models, presented in detail in the Supplementary Material, do not show significant effects on the logarithm of tie strength, neither directly nor through interactions with the ratio of Hispanic residents. This highlights the role of education in the result, which is clearly not a confound with the economic variables of income and unemployment.

Language and emotion

As a preliminary step before the study of emotions and tie strength, we surveyed the correlations between emotion measurements across languages. Figure 2 shows the relationships between average normalized values of affect for English and Spanish language tweets. There is a significant positive correlation (Pearson's $\rho = 0.27$, CI [0.20, 0.35]) between the mean valence of English language tweets and the mean valence of Spanish language tweets from the same tract. Arousal values are only weakly correlated (Pearson's $\rho = 0.12$, CI [0.03, 0.2]). In contrast, there is no significant correlation between positive values of sentiment in English and Spanish or negative values of sentiment in the two languages. This absence of strong correlations motivates the analysis of affect in more than one language, as emotional experiences and social network structures might vary across ethnicities in Los Angeles.

The average value of valence across all tracts is 5.78 for English language tweets and 5.71 for Spanish language tweets, which corresponds to slightly positive affect

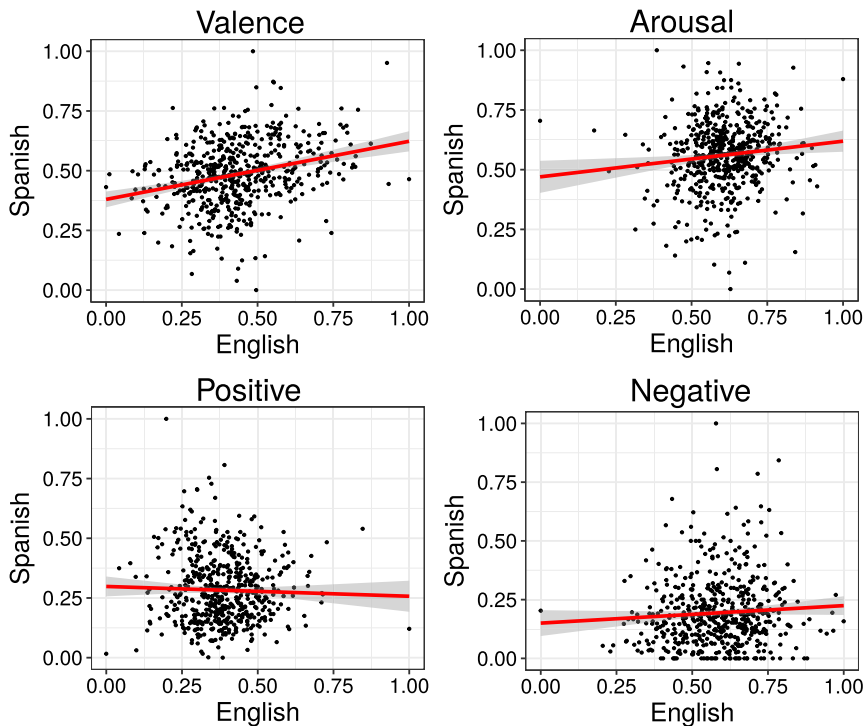


Fig. 2 Comparison of average sentiment scores per tract in Spanish and English. Each point in the scatter plot represents a tract, with the x-axis showing the mean normalized value of affect measured from the English language tweets, and y-axis showing mean normalized affect value measured from the Spanish language tweets posted from the same tract. The measured affect corresponds to valence (top-left) and arousal (top-right) sentiment measured by the WKB and GISB lexicons for the English and Spanish language tweets, respectively, and positive (bottom-left) and negative (bottom-right) measured by SentiStrength. Lines show linear fits, and shaded regions show prediction errors of model fits

with respect to the neutral point of 5.0, in line with emotional expression in other media [29]. As expected, the measurements of valence are positively correlated with the measurements of positive affect ($\rho = 0.50$ in English and $\rho = 0.30$ in Spanish) and negatively correlated with the measurements of negative affect ($\rho = -0.55$ in English and $\rho = -0.21$ in Spanish). This illustrates the relative dependence between the dimensional representations of valence-arousal (VA Model) and positive–negative (PN Model), which capture emotional life in different ways. In the further analysis, we apply these representations in two parallel models for each language, to have a double test of the hypothetical relation between tie strength and emotions.

Emotion and social ties

Finally, we study the relationship between emotions expressed in tweets, the language in which they are written, and strength of social interactions in the analyzed tracts. Despite the overall positive affect of tweets, there is a negative correlation between the strength of social ties and the variables of valence and positive affect. Figure 3 shows the scatter plots of tie strength versus affect variables in both languages. Three patterns can be observed: (1) valence and positive affect are negatively correlated with average tie strength in both languages, (2) arousal appears to have little to no correlation with average tie strength, and (3) negative affect appears to be positively correlated with average tie strength in English but not in Spanish.

We evaluate the above observations in two regression models for each language, one using valence and arousal (VA Model) and another using positive and negative affect (PN Model). The main results of these fits are reported on Table 2, with an additional model controlling for demographic effects reported in the Supplementary Material accompanied by model diagnostics and robustness tests. The regression results verify that, for both languages, valence and positive affect are negatively correlated with average tie strength, while no effect can be observed for arousal in either of the languages. This is consistent with the hypothesis that positive emotions are more likely to be shared with weak contacts, while negative experiences are chosen to be shared through stronger ties. This result is in line with theories of social regulation of emotions [4] and with previous results in protest movements that showed that online negative emotions were associated with stronger collective action [30]. This appears also as a positive relationship between negative affect and tie strength in English, but is inconsistent with the absent pattern for Spanish, which points to the opposite direction but without consistent significance.

Discussion

The availability of large scale, near real-time data from social media sites such as Twitter brings novel opportunities for studying online behavior and social interactions at an unprecedented spatial and temporal resolution. By combining Twitter data with US Census, we were able to study how the socioeconomic and

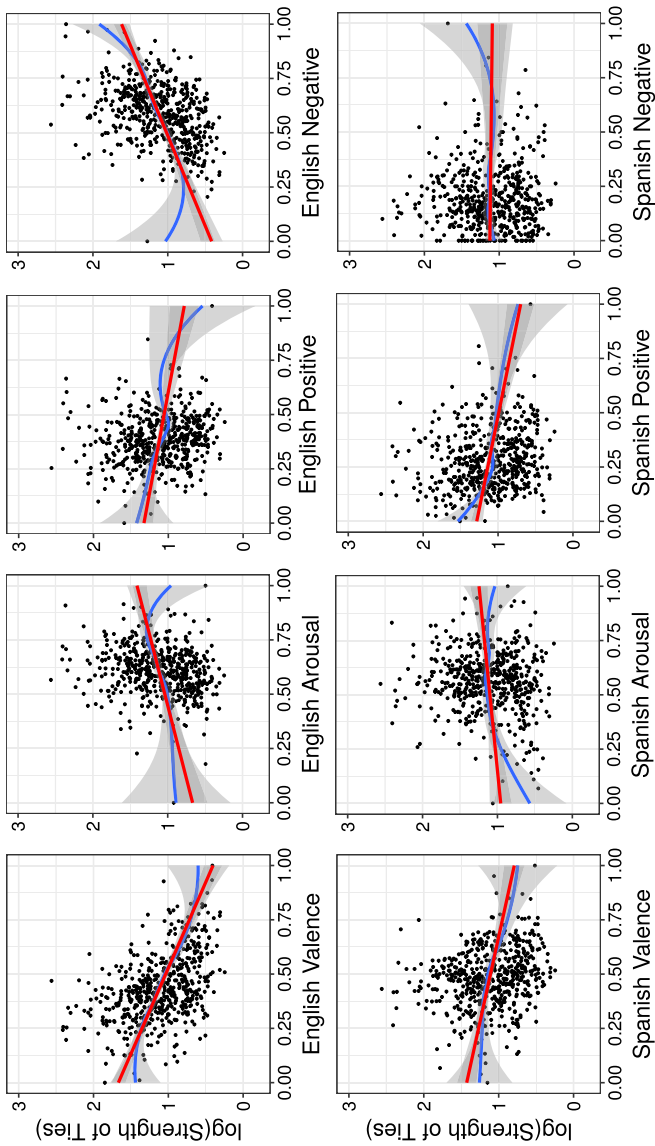


Fig. 3 Tract tie strength as a function of sentiment averages. Each scatter plot shows the logarithm of the average strength of ties in the tracts versus the four sentiment averages in English (top) and Spanish (bottom). Linear fits are depicted in red, non-parametric local regression fits are shown in blue, and shaded regions show prediction errors of model fits

Table 2 Fit results for the mixed effects models of log(tie strength) as a function of sentiment values (valence and arousal in VA Model, positive and negative in PN Model), including random intercepts per tract group

	English		Spanish	
	VA Model	PN Model	VA Model	PN Model
(Intercept)	1.593***	0.618***	1.387***	1.312***
Valence	−1.237***		−0.598***	
Arousal	0.103		0.03	
Positive		−0.517***		−0.602***
Negative		1.194***		−0.111
BIC	514.5	546.3	620.7	620.5
Var: tractGroup (Intercept)	0.008	0.006	0.018	0.031
Var: Residual	0.134	0.144	0.155	0.142

*** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$

demographic characteristics of residents of different census tracts are related to the structure of online interactions. Sentiment analysis of tweets in English and Spanish languages originating from a tract revealed a link between emotional expression and tie strength of Twitter users at the group level.

Our findings are broadly consistent with results of previous studies carried out in offline settings, and also give new insights into the structure of online social interactions. We found that, in line with previous research and theoretical arguments, Twitter users express more positive emotions in areas with weaker ties, while negative emotions are more salient where ties are stronger. We find a lack of correlation with arousal that is consistent with a general pattern in which sentiment analysis techniques do not seem to capture the subjective experience of arousal [31]. We find that at an aggregate level, areas where Twitter users form stronger social ties have lower levels of education, but that this effect interacts with the ratio of Hispanic residents in the opposite direction. Since weak ties are believed to play an important role in delivering strategic, novel information, our work identifies education as a main correlate with the presence of weak ties and their associated novel information.

Our results highlight the social component of culture: Hispanic cultures share collectivist values and are less individualist than anglo-saxon cultures (for example, Mexico scores 30 and the US 91 in the individualism scale of Hofstede [32]). This provides an explanation for the stronger links of tracts with higher number of Hispanic residents (interacting with education), as their online network structures reflect their shared values [33, 34]. However, this manifestation of shared values in digital traces is subject to appear only in areas where education levels are higher, as they also have higher levels of penetration of online social media.

Some important considerations limit the interpretation of our findings. First, our methodology for identifying social interactions may not give a complete view of the social network of Twitter users. Our observations were limited to social interactions initiated by users who geo-reference their tweets. This may not be representative of

all Twitter users posting messages from a given tract, if systematic biases exist in what type of people elect to geo-reference their tweets. For demographic analysis, we did not resolve the home location of Twitter users. Instead, we assumed that characteristics of an area, i.e., of residents of a tract, influence the tweets posted from that tract. Other subtle selection biases could have affected our data and the conclusions we drew [35]. It is conceivable that Twitter users residing in more affluent areas are less likely to use the geo-referencing feature, making our sample of Twitter users different from the population of LA county residents.

Recognizing this limitation, our conclusions only apply at the group level and not at the level of individual behavior of LA residents. In the same vein as how Twitter data can be used to identify group effects on heart disease mortality [36], our analysis identifies relations between properties of groups of people. While recent research opens the possibility to how to reweight Twitter metrics across demographic sections [37], evaluating the external validity of social media metrics requires an interdisciplinary effort beyond the scope of this contribution from data science. Further research can build on our work, combining social media data with standard social science resources, such as surveys, questionnaires, and the census, achieving a more complete picture of social interaction in our current Digital Society.

Methods

Data

Los Angeles (LA) County is the most populous county in the USA, with almost 10 million residents. It is extremely diverse both demographically and economically, making it an attractive subject for research. We collected a large body of tweets from LA County over the course of 4 months, starting in July 2014. Our data collection strategy was as follows. First, we used Twitter's location search API to collect tweets from an area that included Los Angeles County. We then used Twitter4J API to collect all (timeline) tweets from users who tweeted from within this area during this time period. A portion of these tweets were geo-referenced, i.e., they had geographic coordinates attached to them. In all, we collected 6M geo-tagged tweets made by 340K distinct users.

We localized geo-tagged tweets to tracts from the 2012 US Census.¹ A tract is a geographic region that is defined for the purpose of taking a census of a population, containing about 4000 residents on average, and is designed to be relatively homogeneous with respect to demographic characteristics of that population. We included only Los Angeles County tracts in the analysis. We used data from the US Census to obtain demographic and socioeconomic characteristics of a tract, including the mean household income, median age of residents, percentage of residents with a bachelor's degree or above, as well as racial and ethnic composition of the tract.

¹ American Fact Finder (<http://factfinder.census.gov/>).

Emotion analysis

We apply sentiment analysis [38], i.e., methods that process text to quantify subjective states of the author of the text, to measure happiness or subjective well-being of Twitter users. Two recent independent benchmark studies evaluate a wide variety of sentiment analysis tools in various social media [39] and Twitter datasets [40]. Across social media, one of the best performing tools is SentiStrength [41], which also was shown to be the best unsupervised tool for tweets in various contexts [40].

English language analysis SentiStrength quantifies emotions expressed in short informal text by matching terms from a lexicon and applying intensifiers, negations, misspellings, idioms, and emoticons. We use the standard English version of SentiStrength.² To each tweet in our dataset, quantifying positive sentiment $P \in [+1, +5]$ and negative sentiment $N \in [-1, -5]$, consistently with the Positive and Negative Affect Schedule (PANAS) [42]. SentiStrength has been shown to perform very closely to human raters in validity tests [41] and has been applied to measure emotions in product reviews [43], online chatrooms [44], Yahoo answers [45], Youtube comments [46], and social media discussions [47]. In addition, SentiStrength allows our approach to be applied in the future to other languages, like Spanish [30, 48], and to include contextual factors [49], like sarcasm [50].

Beyond positivity and negativity, meanings expressed through text can be captured through the application of the semantic differential [51], a dimensional approach that quantifies emotional meaning in terms of valence and arousal [52]. The dimension of *valence* quantifies the level of pleasure or evaluation expressed by a word, while *arousal* measures the level of activity induced by the emotions associated with a word. Research in psychology suggests that a multidimensional approach is necessary to capture the variance of emotional experience [53], motivating our application of this approximation of two dimensions that goes beyond simple polarity approximations. The state of the art in the quantification of these dimensions is the lexicon of Warriner, Kuperman, and Brysbaert (WKB) [54]. The WKB lexicon includes scores in the three dimensions for more than 13,000 English lemmas. We quantify valence and arousal in a tweet by first lemmatizing the words in the tweet, to then match the lexicon and compute mean values of the valence and arousal as in [55]. While this method is not the most accurate, it provides high coverage for Twitter data [39], and allows a multidimensional representation of emotions that is not frequent in mainstream sentiment analysis. In our dataset, for example, the lexicon matched terms in 82.39% of the tweets.

The Fig. 4 shows word clouds of tweets from a tract with one of the highest average valence and one from a tract with a lower average valence. The words themselves are colored by their valence, with red corresponding to high and blue to low valence words. Despite seemingly small differences in average tract valence, the words depicted in the word clouds are remarkably different in the emotions they convey. The “happy” tract has words such as “beach”, “love”, “family”,

² <http://sentistrength.wlv.ac.uk/>.

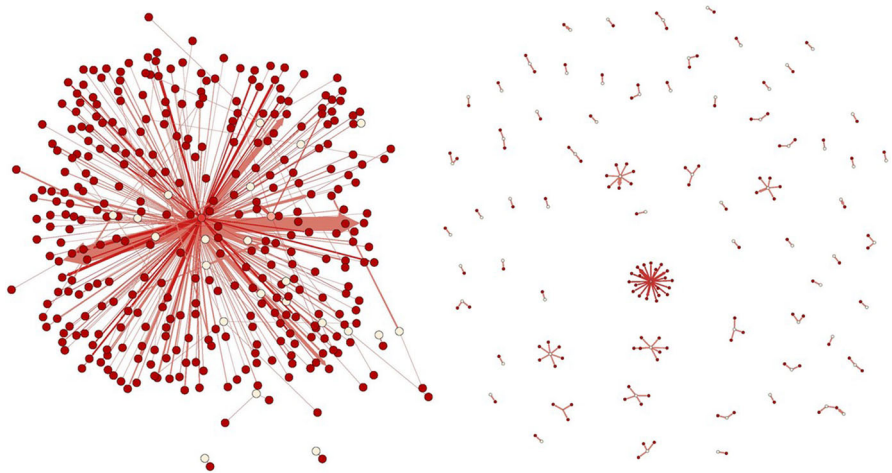


Fig. 5 Mentions graphs of two different tracts showing (left) strong ties (average tie strength $S_i = 7.33$) and (right) weak ties ($S_i = 1.08$). Tweeting users are represented as white nodes, while mentioned users are red nodes. Users who tweet and are mentioned are pink in color. The width of an edge represents the number of mentions

figures and celebrities also using this convention. We use mention frequency along a tie as a proxy of tie strength, drawing upon multiple studies that used frequency of interactions as a measure of tie strength [6, 57, 58]. In contrast to other measures, such as clustering coefficient, it does not require knowledge of full network structure (which we do not observe).

Tie strength For each tract, we create a mention graph with users as nodes and an edge from user A to user B if A mentions B in her tweets. Using this graph, the average social tie strength per tract is defined as

$$S_i = \frac{\sum_{j=1}^{k_i} w_j}{k_i} \quad (1)$$

where w_j is the weight of the j th edge (i.e., the number of times user A mentioned user B), and k_i is the total number of distinct users mentioned in tract i .

We do not have complete knowledge of network structure, since we only observe the tweets of users who geo-referenced their tweets, and not necessarily the tweets of mentioned users. However, even in the absence of complete information about interactions, average tie strength captures the amount of social cohesion and diversity [59]. Figure 5 illustrates mention graphs from two tracts with very different tie strength values. High tie strength (Fig. 5a) is associated with a high degree of interaction and more clustering [2]. In contrast, low tie strength is associated with a sparse, more diverse network with few interconnections (Fig. 5b). In the analysis presented in this article, we apply a logarithmic transformation to S_i , reducing its skewness. After this transformation we identified three outliers that could be mapped to tracts with very few tweets, which we removed for the statistical analysis. For

completeness, we repeated the analysis including these three outliers, leading to very similar results and model diagnostics that we report in the Supplementary Material.

Spatial diversity Geography and distance are important organizing principles of social interactions, both offline [24, 25] and online [27, 28, 58]. While most social interactions are short-range, long-distance interactions serve as evidence of social diversity [18]. In this paper, we use the movement of people across tracts as evidence of the spatial diversity of their social structures. Following Eagle et al. [18], we measure spatial diversity of places from which people tweeting from a given tract also tweet from, using Shannon's Entropy ratio, as

$$D_i = \frac{-\sum_{j=1}^{n_i} p_{ij} \log(p_{ij})}{\log n_i}, \quad (2)$$

where n_i is the number of tracts from which users who tweeted from tract i also tweeted from, and p_{ij} is the proportion of tweets posted by these users from tract j such that

$$p_{ij} = \frac{T_{ij}}{\sum_{j=1}^{n_i} T_{ij}}, \quad (3)$$

where T_{ij} is the number of tweets that have been posted in tract j by the users who have tweeted from both tract i and j .

Thus, spatial diversity is a ratio that compares the empirical entropy of data with its expected value in the uniformly distributed case. As a consequence, a high spatial diversity value for a tract suggests that people tweeting from that tract split their tweets evenly among all the tracts they are tweeting from. In contrast, a low value implies that people tweeting from that tract concentrate their tweets in few tracts.

Regression models Our statistical analysis applies mixed-effects regression models of the logarithm of the average tie strength in each tract. To control for spatial correlations, we model a random intercept for each tract group, which are captured by the tract prefixes of the census.⁶ These tract groups are formed by the division of earlier tracts into subtracts, capturing spatial autocorrelations more sensitive to demographic features than pure geographic methods that ignore urban and administrative barriers. We fit models using the `lmer` function of the `lmer` package, specifying a model with tract groups as random intercepts.

The left panel of Fig. 1 shows the existence of a sizeable negative correlation (Pearson's $\rho = -0.47$) between tie strength and spatial diversity. Tracts that bring together people who also tweet from different places have weaker ties, while tracts with more concentrated user groups have stronger ties. To exclude this pattern from our demographic and emotion analysis, we include a linear term of spatial diversity in each regression model. We further include an additional control term with the average age of residents in the census, to control for the possible age effect in the intensity of social links as manifested in Twitter.

⁶ https://www.census.gov/geo/reference/gtc/gtc_ct.html.

For the demographics model, we perform three fits to survey the interaction between education levels and the ratio of Hispanic residents. For the analysis of emotions, we fit each model twice: first a mixed-effects model using tract groups as random effects and sentiment variables as fixed effects, and second a model that takes as dependent variable the residuals of the demographics model and regresses them against the emotion variables. This way, we verify that the results of the emotion model fits are robust to the role of demographic factors in tie strength.

To verify the validity of our fit results, we perform regression diagnostics that are reported in the Supplementary Material. These verify that multicollinearity is weak (moderate Variance Inflation Factors), that residuals are roughly normally distributed, that no pathological correlations exist between residuals and independent variables, and that no pattern of heteroscedasticity is present. These verifications support the validity of the inferences that evidence the conclusions of this analysis.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Dunbar, R. I., & Shultz, S. (2007). Evolution in the social brain. *Science*, 317(5843), 1344–1347.
2. Granovetter, M. (1973). The strength of weak ties. *The American Journal of Sociology*, 78(6), 1360–1380.
3. Putnam, R. D. (2000). *Bowling alone: The collapse and revival of American community*. New York: Simon & Schuster.
4. Rimé, B. (2009). Emotion elicits the social sharing of emotion: Theory and empirical review. *Emotion Review*, 1(1), 60–85.
5. Sampson, R. J., Raudenbush, S. W., & Earls, F. (1997). Neighborhoods and violent crime: A multilevel study of collective efficacy. *Science*, 277(5328), 918–924.
6. Granovetter, M. (1983). The strength of weak ties: A network theory revisited. *Sociological Theory*, 1(1), 201–233.
7. Burt, R. (1995). *Structural holes: The social structure of competition*. Cambridge: Harvard University Press.
8. Burt, R. S. (2004). Structural holes and good ideas. *American Journal of Sociology*, 110(2), 349–399.
9. Aral, S., & Van Alstyne, M. (2011). The diversity-bandwidth trade-off. *American Journal of Sociology*, 117(1), 90–171.
10. Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In Proceedings of the 21st International Conference on World Wide Web (pp. 519–528). ACM.
11. De Meo, P., Ferrara, E., Fiumara, G., & Provetti, A. (2014). On Facebook, most ties are weak. *Communications of the ACM*, 57(11), 78–84.
12. Kang, J. H., & Lerman, K. (2017). Effort mediates access to information in online social networks. *ACM Transactions on the Web (TWEB)*, 11(1), 3:1–3:19.
13. Shea, C. T., Menon, T., Smith, E. B., & Emich, K. (2015). The affective antecedents of cognitive social network activation. *Social Networks*, 43, 91–99.
14. Krackhardt, D. (1987). Cognitive social structures. *Social Networks*, 9(2), 109–134.
15. Sutcliffe, A., Dunbar, R., Binder, J., & Arrow, H. (2012). Relationships and the social brain: Integrating psychological and evolutionary perspectives. *British Journal of Psychology*, 103(2), 149–168.

16. Niedenthal, P. M., & Brauer, M. (2012). Social functionality of human emotion. *Annual Review of Psychology*, 63, 259–285.
17. Messer, L. C., Laraia, B. A., Kaufman, J. S., Eyster, J., Holzman, C., Culhane, J., et al. (2006). The development of a standardized neighborhood deprivation index. *Journal of Urban Health*, 83(6), 1041–1062.
18. Eagle, N., Macy, M., & Claxton, R. (2010). Network diversity and economic development. *Science*, 328(5981), 1029–1031.
19. Ronen, S., Gonçalves, B., Hu, K. Z., Vespignani, A., Pinker, S., & Hidalgo, C. A. (2014). Links that speak: The global language network and its association with global fame. *Proceedings of the National Academy of Sciences*, 111(52), E5616–E5622.
20. Schich, M., Song, C., Ahn, Y. Y., Mirsky, A., Martino, M., Barabási, A. L., et al. (2014). A network framework of cultural history. *Science*, 345(6196), 558–562.
21. Logan, J. R. (2012). Making a place for space: Spatial thinking in social science. *Annual Review of Sociology*, 38, 507–524.
22. Reardon, S. F., Matthews, S. A., O'Sullivan, D., Lee, B. A., Firebaugh, G., Farrell, C. R., et al. (2008). The geographic scale of metropolitan racial segregation. *Demography*, 45(3), 489–514.
23. Lee, B. A., Reardon, S. F., Firebaugh, G., Farrell, C. R., Matthews, S. A., & O'Sullivan, D. (2008). Beyond the census tract: Patterns and determinants of racial segregation at multiple geographic scales. *American Sociological Review*, 73(5), 766–791.
24. Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 32(4), 425–443.
25. Barthélemy, M. (2011). Spatial networks. *Physics Reports*, 499(1), 1–101.
26. Quercia, D., Capra, L., & Crowcroft, J. (2012). The social world of twitter: Topics, geography, and emotions. In Proceedings of the 6th International AAAI Conference on Weblogs and Social Media (ICWSM), Vol. 12, pp. 298–305.
27. Liben-Nowell, D., Novak, J., Kumar, R., Raghavan, P., & Tomkins, A. (2005). Geographic routing in social networks. *Proceedings of the National Academy of Sciences of the United States of America*, 102(33), 11623–11628.
28. 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.
29. Garcia, D., Garas, A., & Schweitzer, F. (2012). Positive words carry less information than negative words. *EPJ Data Science*, 1(1), 1.
30. Alvarez, R., Garcia, D., Moreno, Y., & Schweitzer, F. (2015). Sentiment cascades in the 15M movement. *EPJ Data Science*, 4(1), 1–13.
31. Garcia, D., Kappas, A., Küster, D., & Schweitzer, F. (2016). The dynamics of emotions in online interaction. *Royal Society Open Science*, 3(8), 160059.
32. Hofstede, G. (1984). *Culture's consequences: International differences in work-related values* (Vol. 5). Thousand Oaks: SAGE Publications.
33. 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.
34. Kayes, I., Kourtellis, N., Quercia, D., Iamnitchi, A., & Bonchi, F. (2015). Cultures in community question answering. In Proceedings of the 26th ACM Conference on Hypertext and Social Media, pp. 175–184. ACM.
35. Tufekci, Z. (2014). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In International AAAI Conference on Web and Social Media. <http://www.aaai.org/ocs/index.php/ICWSM/ICWSM14/paper/view/8062>
36. Eichstaedt, J. C., Schwartz, H. A., Kern, M. L., Park, G., Labarthe, D. R., Merchant, R. M., et al. (2015). Psychological language on twitter predicts county-level heart disease mortality. *Psychological Science*, 26(2), 159–169.
37. Barbera, P. (2016). Less is more? How demographic sample weights can improve public opinion estimates based on twitter data. NYU Working Paper.
38. Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1–2), 1–135.
39. Ribeiro, F. N., Araújo, M., Gonçalves, P., Gonçalves, M. A., & Benevenuto, F. (2016). Sentibench—a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1), 1–29.

40. Abbasi, A., Hassan, A., & Dhar, M. (2014). Benchmarking twitter sentiment analysis tools. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)
41. Thelwall, M., Buckley, K., & Paltoglou, G. (2012). Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1), 163–173.
42. Watson, D., Clark, L. A., & Tellegen, A. (2013). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 3(6), 1063.
43. Garcia, D., & Schweitzer, F. (2011). Emotions in product reviews—empirics and models. In Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third International Conference on Social Computing (SocialCom)
44. Garas, A., Garcia, D., Skowron, M., & Schweitzer, F. (2012). Emotional persistence in online chatting communities. *Scientific Reports*, 2
45. Kucukunc, O., Cambazoglu, B. B., Weber, I., & Ferhatosmanoglu, H. (2012). A large-scale sentiment analysis for Yahoo! answers. In Proceedings of the Fifth ACM International Conference on Web Search and Data Mining
46. Garcia, D., Mendez, F., Serdült, U., & Schweitzer, F. (2012). Political polarization and popularity in online participatory media: An integrated approach. In: Proceedings of the First Edition Workshop on Politics, Elections and Data
47. Ferrara, E., & Yang, Z. (2015). Quantifying the effect of sentiment on information diffusion in social media. *PeerJ Computer Science*, 1, e26.
48. Vilares, D., Thelwall, M., & Alonso, M. A. (2015). The megaphone of the people? Spanish SentStrength for real-time analysis of political tweets. *Journal of Information Science*, 41(6), 799–813.
49. Thelwall, M., Buckley, K., Paltoglou, G., Skowron, M., Garcia, D., Gobron, S., Ahn, J., Kappas, A., Küster, D., & Holyst, J. A. (2013). Damping sentiment analysis in online communication: Discussions, monologs and dialogs. In International Conference on Intelligent Text Processing and Computational Linguistics (pp. 1–12). Springer, Berlin, Heidelberg.
50. Rajadesingan, A., Zafarani, R., & Liu, H. (2015). Sarcasm detection on twitter: A behavioral modeling approach. In Proceedings of the Eighth ACM International Conference on Web Search and Data Mining
51. Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1964). *The measurement of meaning*. Champaign: University of Illinois Press.
52. Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3), 273–294.
53. Fontaine, J. R., Scherer, K. R., Roesch, E. B., & Ellsworth, P. C. (2007). The world of emotions is not two-dimensional. *Psychological Science*, 18(12), 1050–1057.
54. Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, 45(4), 1191–1207.
55. González-Bailón, S., Banchs, R. E., & Kaltenbrunner, A. (2012). Emotions, public opinion, and us presidential approval rates: A 5-year analysis of online political discussions. *Human Communication Research*, 38(2), 121–143.
56. Stadthagen-Gonzalez, H., Imbault, C., Sánchez, M. A. P., & Brysbaert, M. (2016). Norms of valence and arousal for 14,031 Spanish words. *Behavior Research Methods*, 49(1), 111–123.
57. Onnela, J., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., et al. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, 104(18), 7332–7336.
58. Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking gross community happiness from tweets. In Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp. 965–968. ACM
59. Gonçalves, B., Perra, N., & Vespignani, A. (2011). Modeling users' activity on Twitter networks: Validation of Dunbar's number. *PLoS One*, 6(8), e22656.