Multidimensional tie strength and economic development

scientific reports

Check for updates

OPEN

Multidimensional tie strength and economic development

Luca Maria Aiello[1,2]✉, Sagar Joglekar[3] & Daniele Quercia[3,4]

The strength of social relations has been shown to affect an individual's access to opportunities. To date, however, the correspondence between tie strength and population's economic prospects has not been quantified, largely because of the inability to operationalise strength based on Granovetter's classic theory. Our work departed from the premise that tie strength is a unidimensional construct (typically operationalized with frequency or volume of contact), and used instead a validated model of ten fundamental dimensions of social relationships grounded in the literature of social psychology. We built state-of-the-art NLP tools to infer the presence of these dimensions from textual communication, and analyzed a large conversation network of 630K geo-referenced Reddit users across the entire US connected by 12.8M social ties created over the span of 7 years. We found that unidimensional tie strength is only weakly correlated with economic opportunities ($R^2 = 0.30$), while multidimensional constructs are highly correlated ($R^2 = 0.62$). In particular, economic opportunities are associated to the combination of: (i) knowledge ties, which bridge geographically distant groups, facilitating the knowledge dissemination across communities; and (ii) social support ties, which knit geographically close communities together, and represent dependable sources of social and emotional support. These results point to the importance of developing high-quality measures of tie strength in network theory.

The strength of social relations has been shown to affect an individual's access to innovation[1], access to economic opportunities[2], life expectancy[3], and happiness[4]. According to Granovetter's classic theory about tie strength[5], information flows through social ties of two strengths. First, through weak ties. These ties, despite being used infrequently, bridge distant groups that tend to posses diverse information, facilitating the knowledge dissemination across communities. Second, information also flows through strong ties. These ties, by being used frequently, knit close communities together, and represent dependable sources of social and emotional support.

To date, however, the correspondence between tie strength and population's economic prospects has not been quantified, largely because of the inability to operationalize tie strength based on Granovetter's conception. Typically, network studies operationalize strength with indicators like frequency or volume of contact[6]. Eagle et al. did so by studying the relationship between the structure of a national communication network and access to socio-economic opportunity[7]. They found that network diversity was associated to opportunities, but communication volume or number of contacts was not. The prospect that tie strength is not a unidimensional construct ranging from weak to strong but might be multidimensional is broadly consistent with theoretical and experimental work by Marsden and Campbell[6] and Wellmann and Wortley[8]. It is also consistent with Granovetter's original operationalization of strength as *"a (probably linear) combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie."*[5] These indicators have been repeatedly found to be only weakly related to frequency of contacts[6,7]. Therefore, network studies using frequency of contacts to model strength are capturing only one aspect of the linkages among individuals.

Our work departed from the premise that tie strength is a unidimensional construct, built upon work on social psychology starting from Granovetter's conception of tie strength, and identified and validated ten fundamental *dimensions* of social relationships[9,10]. In previous work, we showed that these ten dimensions correspond to how people perceive and categorize most of their own social relationships[9], and we built a state-of-the-art NLP tools to infer the presence of these dimensions from textual communication[10]. In this work, we used these tools to analyze a large conversation network of geo-referenced Reddit users across the entire US (~13M ties). Then, going back to Eagle et al.'s work and borrowing their methodological framework[7], we were able to test whether the structure of a national communication network (in particular, its tie diversity) was related to access to socio-economic opportunities, and whether switching from a unidimensional notion of tie strength to a multidimensional one would improve explanatory power. We found that tie diversity measured on the networks of knowledge exchange

[1]IT University of Copenhagen, 2300 Copenhagen, Denmark. [2]Pioneer Centre for AI, 2100 Copenhagen, Denmark. [3]Nokia Bell Labs, CB30FA Cambridge, UK. [4]CUSP, King's College London, WC2R2LS London, UK. ✉email: luai@itu.dk

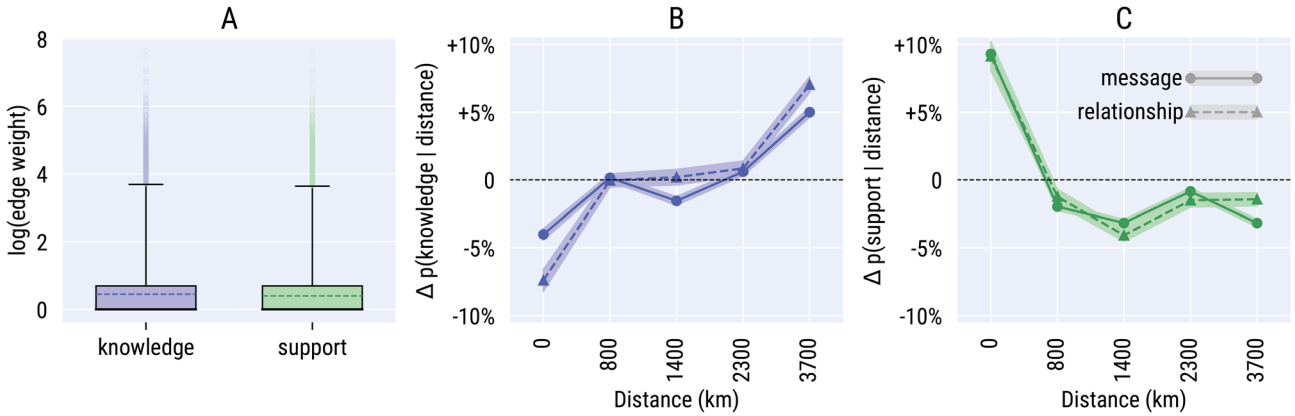

Figure 1. (**A**) Boxplots of the weight distributions of ties that exchanged at least one message of *knowledge* or one message of *support*, on a logarithmic scale. Boxes represent the two mid quartiles of the distributions, with the median marked with a dashed line. The whiskers show the 99th percentiles of the distributions. (**B**,**C**) Percent change $\Delta p(d|l)$ of the probability that a dimension d is expressed by a social tie spanning a geographical distance l, compared to random chance. The change is estimated by comparing the real data with distance measurements on 50 instances of a null model that reshuffled user locations at random; the average values along with their 95% confidence intervals are reported. Distances are discretized in five bins, each containing the same number of social ties. Bins are labeled with the median distance of the ties inside that bin. The 'zero distance' bin contains almost exclusively pairs of users who live in the same state. Two types of measurements are presented: (i) at the level of social relationships, where each social tie is counted once regardless of its weight, and (ii) by performing a distance measurement for each individual message, thus effectively weighting more pairs of users who communicated frequently.

and social support correlates much more strongly with economic development ($R^2 = 0.62$) than diversity measured on a network simply weighted on frequency of interactions ($R^2 = 0.30$).

In line with Granovetter's conception of tie strength, we found that knowledge ties and social support ties: are hardly distinguishable solely based on frequency of interaction; have opposite geographic distribution (knowledge ties are global, spanning longer geographical distances, while social support ones are local, typically staying in the same state); and both contribute to economic opportunities (states with higher GDP per capita are characterized by both global access to knowledge and local access to support). These results point to the importance of developing multidimensional measures of tie strength in network theory to better reflect the nature of human relationships that social links ought to model.

Results

From a set of 65M comments posted on Reddit by 1.3M users between the years of 2006 and 2017, we extracted the social interactions of all Reddit users that we could geo-reference at the level of the 51 US states using high-accuracy heuristics validated in previous work (see "Methods"). In Reddit, conversations develop over discussion *threads*. If user i commented over either a submission or a comment of another user j, we considered that i sent a *message* to j, as it is common practice when studying Reddit conversation networks[11]. We created a directed communication graph $\mathcal{G}$ (U, E) to model such exchange of messages. The set of nodes U contains all the geo-referenced Reddit users in our dataset. Two users i and j are connected by a directed edge $(i, j, w(i, j)) \in E$ if user i sent at least one message to user j. The edge weight $w(i, j)$ represents the frequency of contacts and it is equal to the total number of messages sent. In total, the graph contains 630K nodes and 12.8M edges. The distribution of node degree and link strength is shown in Fig. SI1.

By applying our social dimensions classifier to the corpus of messages, we identified the subset of messages that express a social dimension d (see "Methods" for details). In particular, we focused on the dimensions of *knowledge exchange* and *social support* (respectively, *knowledge* and *support* for short). Other dimensions are discussed in Supplementary Information). The classifier ranked the messages according to their likelihood of containing expressions of a given social dimension; we marked with dimension d only the top 1% of messages from the likelihood ranking of d (we discuss results with looser thresholds in Supplementary Information, Fig. SI2). Out of these smaller sets of messages, we constructed *dimension-specific communication graphs* $\mathcal{G}_d$ using the same procedure we adopted for building the overall communication graph $\mathcal{G}$. Such dimension-specific graphs capture only one type of social interaction each; for example, the *knowledge* graph $\mathcal{G}_{knowledge}$ contains only edges formed by knowledge-exchange messages, and edge weights encode the number of knowledge-exchange messages flowing between the two endpoints. The dimension-specific graphs contain roughly 1% of the edges of the full communication graph and between 16 to 23% of its nodes, depending on the dimension (see Table 2). The networks of *knowledge* and *support* include 20% and 21% of all nodes, respectively. The edges of $\mathcal{G}_{knowledge}$ and $\mathcal{G}_{support}$ overlap only slightly: around 2% of the edges of each graph are also present in the other.

By having a sample of edges annotated with both social dimensions and weight, we were able to look into the relationship between frequency of contacts, *knowledge*, and *support*. The typical weight of edges connecting users who exchange knowledge is not dissimilar from the typical weight of those providing support. Figure 1A

Predicting GDP per capita from:											
Population density				**Diversity on full communication graph**				**Spatial diversity on dimension-specific graphs**			
Feature	***β***	**SE**	***p***	**Feature**	***β***	**SE**	***p***	**Feature**	***β***	**SE**	***p***
α (intercept)	0.310	0.045	0.000	α (intercept)	− 0.035	0.108	0.747	α (intercept)	0.1943	0.061	0.003
Pop. density	0.636	0.113	0.000	Pop. density	0.565	0.174	0.002	Pop. density	0.4713	0.116	0.000
				$D_{spatial}$	0.243	0.151	0.116	$D_{spatial}^{knowledge}$	1.0327	0.164	0.000
								$D_{spatial}^{support}$	− 0.5549	0.154	0.001
Durbin–Watson stat. = 1.982		$R^2_{adj} = 0.26$		Durbin–Watson stat. = 2.082		$R^2_{adj} = 0.30$		Durbin–Watson stat. = 2.069		$R^2_{adj} = 0.62$	

Table 1. Linear regressions to predict GDP per capita of US states from: (left) population density only; (center) spatial diversity computed on the full communication graph; (right) spatial diversity computed on dimension-specific communication graphs. Population density is added as a control variable in the latter two models. Adjusted R^2 and Durbin–Watson statistic for autocorrelation (values close to 2 indicate no autocorrelation) are reported. The contribution of individual features to the models is described by their *beta*-coefficients, standard errors (SE) and *p*-values.

compares the weight distribution of edges connecting users who exchanged knowledge with the weight distribution of edges connecting those who exchanged support. A two-sample Kolmogorov–Smirnov test (a statistic to measure the distance between two distributions) indicated that the two distributions, albeit statistically different, are very similar: $KS = 0.03\,(p = 0.0)$ on a range from 0 (indicating identical distributions) to 1 (maximum difference). This comparison exposes the inherent limit of quantifying tie strength with the mere frequency of interactions to adequately qualify the nature of social relationships.

In Reddit conversations, the main difference between *knowledge* and *support* ties does not lie in their strength but in their geographic span. The probability of creating *knowledge* ties increases with the geographical distance between the two endpoints, while the probability of creating *support* ties drops with distance (Fig. 1B,C). This is consistent with theoretical expectations. Knowledge production on the Web follows Pareto's law: a restricted number of experts create and spread information to a vast audience[12]; consequently, knowledge ends up being locally scarce[13] and needs to travel longer distances to reach multiple communities. In past studies, a similar pattern was detected for the communications within large corporations, where geographically distant ties were estimated to be more effective conduits for knowledge flow[14,15]. The opposite trend holds for *support*. Geographical distance impacts significantly people's ability to provide both material and emotional support[16]. Despite computer-mediated communication has grown the opportunities for providing remote support[17], people have an innate sense for local attachments and an economic advantage to foster them[18], which might be why *support* appears more rarely in long-distance relationships[8].

Last, we tested if dimension-specific graphs are more indicative of economic development than the full communication graph. We did so by borrowing the experimental setup by Eagle et al.[7], who studied the network of phone calls among residents of England and measured the *spatial and social diversity* ($D_{spatial}$) for each of nearly 2,000 regional exchanges in the country. $D_{spatial}$ captures the diversity of areas that the residents of a given area communicate with, and they found it to be correlated with the Index of Multiple Deprivation—a composite score of social and economic development based on UK census data. They also tested the robustness of their results with an alternative measure of diversity D_{social} that captures the diversity of people connected to the residents of a given area. We reproduced Eagle et al.'s experimental setup and ran an Ordinary Least Squares linear regression (OLS) to predict per-capita Gross Domestic Product (GDP) of US states in the year 2017[19] from the spatial diversity at state-level computed on (i) the full communication graph ($D_{spatial}$) and (ii) the two dimension-specific communication graphs ($D_{spatial}^{knowledge}$, $D_{spatial}^{support}$). Results for D_{social} are highly aligned with those for $D_{spatial}$, and we discuss them in Supplementary Information. We focused on 44 states for which Reddit penetration is sufficient and aligned with the population distribution (see "Methods"), however we found qualitatively similar results when considering all states (see Supplementary Information, Table SI3). Regressions models with different combinations of social and spatial diversity are presented in Tables SI1 and SI2.

In Table 1 we compare three linear regressions models: one based on population density only (a validated predictor of economic growth[20]), one using spatial diversity on the full graph with links weighted based on frequency of interaction, and one using the two spatial diversity scores calculated on the graphs of *knowledge* and *support*. The model based on the selected social dimensions is 138% more accurate than the density-only baseline, while the model based on the full communication graph is only 15% more accurate. To check whether the difference in performance is due to the selection of *knowledge* and *support* ties or just to the smaller sample considered, we ran a regression using a random sample of ties as small as the number of *knowledge* ties, and obtained the worst fit (R^2_{adj} of approximately 0.1, see Supplementary Information).

In the regression model with the social dimensions, the coefficient for *knowledge* diversity is positive and the one for *support* diversity is negative. People living in areas characterized by superior economic outcomes access novel information that is not available locally by establishing a diverse set of global interactions, which is in agreement with the *weak tie* pillar of Granovetter's theory. Residents of states with highest per-capita GDP draw their social support mostly from local connections, in agreement with the *strong tie* pillar of the theory. The effect size of *knowledge* is stronger (almost double) than the effect size of *support*, which indicates that the process of knowledge exchange is the primary correlate of economic development, and the network of support compounds

over it. A linear regression including other social dimensions is discussed in Table SI4, but the interplay between *knowledge* and *support* is more predictive than any other combination of dimensions.

Discussion

In agreement with Granovetter's theory, we found that economic development at the level of US states is associated to the abundance of global ties that carry factual knowledge and with the abundance of local ties providing social support. This finding is compatible with the established notion of innovation being fueled primarily by novel information flowing from diverse regions of the social network, and secondarily by an adequate support network to favor the re-elaboration of those ideas locally. This perspective enriches the corpus of experimental evidence about the existence of a trade-off between seeking novel information and building tight networks of support[13,21,22]. We showed that geographical regions generally experience that trade-off but the regions that achieve high economic success are those that have both global outreach of knowledge exchange and local networks of support.

In contrast with a variety of network science studies, we provided evidence that frequency of contacts might not be a good proxy for tie strength: network diversity calculated on a weighted social network is weakly associated to economic development at state level. Moreover, our results challenge the equivalence between weak ties and knowledge flow, at least for the case of Reddit. Interestingly, we found that *knowledge* and *support* ties differ in terms of their geographical span, with *knowledge* ties being far-reaching, and *support* ties being local.

The ability of measuring directly these two aspects of social interaction that are postulated by Granovetter's theory to be drivers to innovation enhances the predictive and descriptive power of network models. Strikingly, narrowing down the analysis to a small subset of messages that express either *knowledge* or *support* yields a predictive performance that is as much as double of that of models used in previous research that considered only frequency of contacts[7].

The ability of decomposing relationship data into interpretable social constituents opens up ample avenues of exploration in social network analysis. Studying how different social dimensions are instantiated by different anatomic patterns of social networks such as their community structure or the centrality of their actors might be a promising research direction. Also, this work showed the association of *knowledge* and *support* with GDP, but other social dimensions may well explain other socio-economic outcomes such as health or quality of life.

Both our data and methods suffer from limitations that future work may address. Unlike the work by Eagle et al., upon which our experimental setup was based[7], our study relies on social network data that covers only a small sample of the population; this was a necessary sacrifice in order to gain the crucial ability to analyze the content of social interactions.

Among all the social platforms from which we could have collected conversational text, we selected Reddit because its richness of information and variety of social interaction types. Other popular platforms (e.g., Facebook, Twitter) either authorize data collection exclusively from volunteer users[23] or expose data APIs that may be limited by volume, temporal scope, and known sampling biases[24]. On the contrary, Reddit allows for the collection of the full conversation history between any pair of users, and includes metadata useful for their characterization, such as geo-localization[25]. Also, Reddit's etiquette, credit system, and topic-oriented subreddits encourage social participation for purposes that are akin to real-life social networks[26], such as socialization, entertainment, and information exchange[27], while naturally disincentivizing practices that disproportionately favor status-seeking, which are prominent in platforms such as Twitter and Facebook[28,29]. As a result, Reddit's comment threads enjoy properties that are typical of human conversations, such as the high topical coherence of successive messages in a thread[11,30]. Because of these desirable properties, Reddit has been the platform of choice for hundreds of quantitative and qualitative studies on social behavior in the last ten years[31]. Furthermore, the anatomy and dynamics of the Reddit conversation network exhibit properties that are in line with those of most social networks[32–34], which speaks to the potential of our findings to generalize to other contexts. These properties include broad distributions of the node degree and of the frequency of most user activities[35–37] (see also Fig. SI1), marked community structure[38], assortativity[36], and burstiness of interactions[39]. Nevertheless, Reddit user base is biased towards males (64%) and young adults (36% in the age range 18–29, 22% in the range 30–49), and our study focuses entirely on US residents[40]; therefore, replicating our analysis to multiple conversation networks is in order to corroborate the robustness of our results.

Within Reddit, our perspective on the ecosystem of social interactions is restricted by our focus on the physical space. In particular, the communication graphs include only a sample of all the existing edges, namely those that connect users whose geo-locations could be estimated. This entails three main biases. First, the majority of interactions are left out of the picture, thus potentially reducing the predictive and descriptive power of our models. Second, the social links we considered were not randomly sampled, as they connect users who self-selected themselves to join geo-salient subreddits. Last, the limited resolution of the user spatial location (state-level) affected our ability to perform a finer-grained geographic analysis (e.g., at city level). To address these biases, future work ought to consider social systems where a larger portion of users can be geo-referenced at a finer geographic resolution.

Even if our social dimensions classifiers were trained on Reddit data and were shown to achieve high accuracy (see "Methods"), their output is not error-free. To improve both precision and recall, a systematic error analysis and a fine-tuning of the model with additional training data would be in order. The ten social dimensions, albeit more comprehensive than any existing model, do not exhaustively map all the possible elements that define social interactions. The concepts that these social dimensions encode are rather broad and encompass a rich spectrum of nuances. The main goal of this work was to go beyond simple frequency of contacts as a proxy for tie strength, offering well-founded interaction archetypes that could be explored and refined in the future.

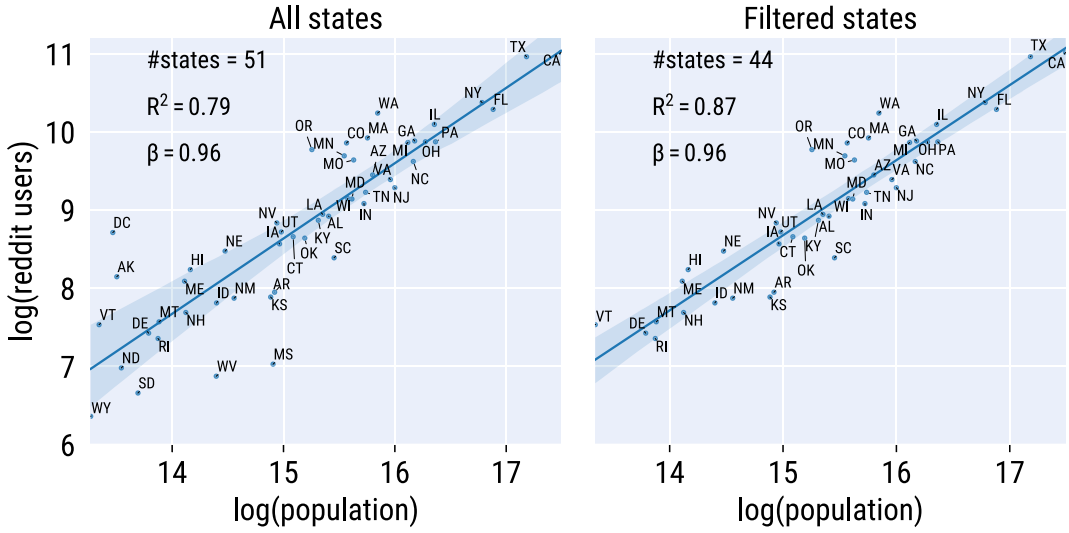

Figure 2. Relationship between population and number of Reddit users across US states. The best linear fit is shown, together with its slope β and the R^2 coefficient to measure the goodness of fit. On the left, all states are included. On the right, the states whose Reddit penetration was too low or was not proportional to the population of residents were removed.

Methods

Reddit data collection. Reddit is a public discussion website particularly popular in the United States where half of its user traffic is generated. Reddit is structured in an ever-growing set of independent *subreddits* (1.2M at the time of writing) dedicated to a broad range of topics[25]. Users can post new *submissions* to any subreddit, and other users can add *comments* to submissions or to existing comments, thus creating nested conversation *threads*.

The vast majority of Reddit submissions and comments since 2007 is publicly available through the `pushshift.io` API[41]. For the purpose of this study, we gathered the content created in two temporal windows: from 2007 until the end of 2012, and for the whole year of 2017. The findings presented in the "Results" section were obtained using the data from these two windows jointly, but having at hand two collections from distinct time periods allowed us to study how data recency affects the ability to predict the desired outcome (see Supplementary Information, Fig. SI3). In total, we collected 65M comments from 1.3M users.

We restricted our study to users whom we could geo-reference at the level of US States. Although Reddit does not provide explicit information about user location, we used a location-estimation heuristic proven to be effective in previous work[42]. We first identified 2,844 geo-salient subreddits related to cities or states in the United (https://www.reddit.com/r/LocationReddits/wiki/faq/northamerica). We assigned a user to a state if (i) they posted at least n submissions or comments in subreddits related to that state, and (ii) 95% or more of their comments and submissions posted to geo-salient subreddits were done in subreddits related to that state. The findings presented earlier were obtained with $n = 3$; in Supplementary Information (Fig. SI4) we discuss results obtained by varying this threshold. Overall, we found 632k users who are likely to be located in one of the 51 US states. The number of users per state ranges from less than 1k (Wyoming) to 61k (California). In total, these users posted 16.2M comments in total (9.8M in 2007–2012, and 6.4 in 2017).

Filtering states by Reddit penetration. States in which the number of Reddit users is not proportional to the number of residents might distort the representation of social communication patterns that actually take place in those states. To identify such cases, we proceeded as follows. We first plotted the census population in 2017 against the number of Reddit users, across states (Fig. 2, left). We then obtained the best linear fit of the data and calculated the residuals between the number of Reddit users and the predicted value according to the linear fit. Last, we calculated the distribution of residuals and removed states whose residuals were more than 1 standard deviation away from the average of the distribution. Those included two states whose Reddit user base was higher than what one would expect based on their population (DC and AK) and two for which it was lower (MS and WV). In addition, we removed three outlier states whose Reddit penetration was lowest (less than 1000 users), which left us with a total of 44 states (Fig. 2, right).

Social dimensions from textual conversations. Social science research proposed several categorizations of constitutional sociological dimensions that describe human relationships[8,51,52]. By surveying such extensive literature, Deri et al.[9] compiled one of the most comprehensive categorizations to date, which identifies ten main *dimensions* of social relationships (Table 2). This theoretical model is rather exhaustive in that most relationships are accurately defined by appropriate combinations of the ten dimensions—Deri et al. showed it by asking hundreds of volunteers to write down keywords that described their relationships and found that all of

Dimension	Description	% Nodes in $\mathcal{G}_d$
Knowledge	Exchange of ideas or information; learning, teaching[43]	0.20
Support	Giving emotional or practical aid and companionship[43]	0.21
Power	Having power over the behavior and outcomes of another[44]	0.17
Status	Conferring status, appreciation, gratitude, or admiration upon another[44]	0.22
Trust	Will of relying on the actions or judgments of another[45]	0.23
Romance	Intimacy among people with a sentimental or sexual relationship[46]	0.22
Similarity	Shared interests, motivations or outlooks[47]	0.21
Identity	Shared sense of belonging to the same community or group[48]	0.17
Fun	Experiencing leisure, laughter, and joy[49]	0.21
Conflict	Contrast or diverging views[50]	0.16

Table 2. The social dimensions of relationships surveyed by Deri at al.[9]. The last column reports the fraction of nodes of the full communication graph $\mathcal{G}$ that are included in each dimension-specific graph $\mathcal{G}_d$. The fraction of nodes in the last column is not exclusive, because nodes can be found in multiple dimension-specific graphs. Our work focused mainly on the dimensions of *knowledge* and *support*.

them fitted into the ten dimensions. The ten social dimensions are frequently expressed through conversational language and, most importantly, these verbal expressions can be captured with computational tools.

We infer the social dimensions from Reddit messages using the NLP model proposed by Choi et al.[10], which comes with a publicly-available python implementation (http://www.github.com/lajello/tendimensions). Given a textual message m and a social dimension d, the model estimates the likelihood that m conveys d by giving in output a score from 0 (least likely) to 1 (most likely). Rather than using a multiclass classifier, the model includes ten independently-trained binary classifiers C_d, one per each dimension. This choice was driven by the theoretical interpretation of the social dimensions[9], as any sentence may potentially convey several dimensions at once (e.g., a message expressing both trust and emotional support). Each classifier is implemented using a Long Short-Term Memory neural network (LSTM)[53], a type of Recurrent Neural Network (RNN) that is particularly effective in modeling both long and short-range semantic dependencies between words in a text, and it is therefore widely used in a variety of NLP tasks[54]. Like most RNNs, LSTM accepts fixed-size inputs. This particular model takes in input a 300-dimension embedding vector of a word, one word at a time for all the words in the input text. Embedding vectors are dense numerical representations of the position of a word in a multidimensional semantic space. Such representations are learned from large text corpora. This model uses GloVe embeddings[55] learned from Common Crawl, a text corpus containing 840B tokens.

The dimensions classifiers C_d were trained using about 9k sentences that were manually labeled by trained crowdsourcing workers. Most of these sentences were taken from Reddit, which makes it the ideal platform to apply the model on. In their experiments, Choi et al. reported very high classification performance which averages to an Area Under the Curve (AUC) of 0.84 across dimensions, and specifically 0.82 for *knowledge* and 0.83 for *support*. AUC is a standard performance metric that assesses the ability of a classifier to rank positive and negative instances by their likelihood score, independent of any fixed decision threshold. The AUC of a random classifier is expected to be 0.5, whereas the maximum value is 1.

Given in input a message m, the classifier outputs a score $s_d(m)$ that expresses the likelihood that message m contains dimension d. In practice, the classifier estimates a score for each sentence in m and returns the maximum score, namely: $s_d(m) = \max_{sentence \in m} s_d(sentence)$. By using the maximum score, we considered a message as likely to express dimension d as its most likely sentence, thus avoiding the dilution effect of the average. This reflects the theoretical interpretation of the use of the social dimensions in language[9]: a dimension is conveyed effectively through language even when expressed only briefly.

To conduct our analysis, we binarized the classifier scores $s_d(m)$ using an indicator function that assigns dimension d to m if $s_d(m)$ is above a certain threshold θ_d:

$$d(m) = \begin{cases} 1, & \text{if } s_d(m) \geq \theta_d \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

We used dimension-specific thresholds because the empirical distribution of the classifier scores s_d varies noticeably across dimensions (see Fig. 3, left), which makes the use of a fixed common threshold unpractical. We made a very conservative choice of θ_d as the value of the 99th percentile of the distribution of the classifier score s_d, thus favoring high precision over recall. This effectively reduces the number of messages to 1% of the total and the number of edges to slightly more than 1% of the total. In Supplementary Information (Fig. SI2, right), we experimented with different percentiles, starting from the 75th.

As a result of this procedure, a comment could end up being labeled with multiple dimensions. To measure the extent to which pairs of dimensions are related, we computed the Spearman rank cross-correlation matrix of the classifier scores of all dimension pairs across all messages (Fig. 3, right). Some pairs of dimensions such as *status*, *trust* and *support* occur more frequently together, but overall the ten dimension model exhibits a fairly high degree of orthogonality. To make sure that the ten dimension classifier is not capturing simply the sentiment of the text, we correlated the dimensions scores with the scores from Vader, a simple yet widely-used sentiment analyzer[56]. The correlations were all very low except for a negative correlation with the conflict dimension.

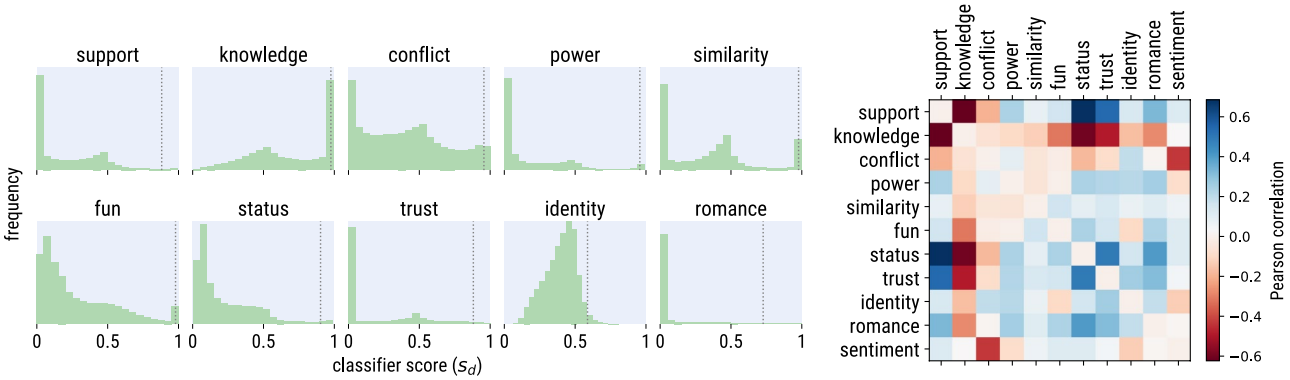

Figure 3. Left: frequency distributions of the classifier scores s_d for all dimensions. The dotted vertical lines mark the values of the 99th percentile of each distribution. Right: cross-correlation matrix of the classifier scores of all dimension pairs across all messages, plus a simple measure of text sentiment.

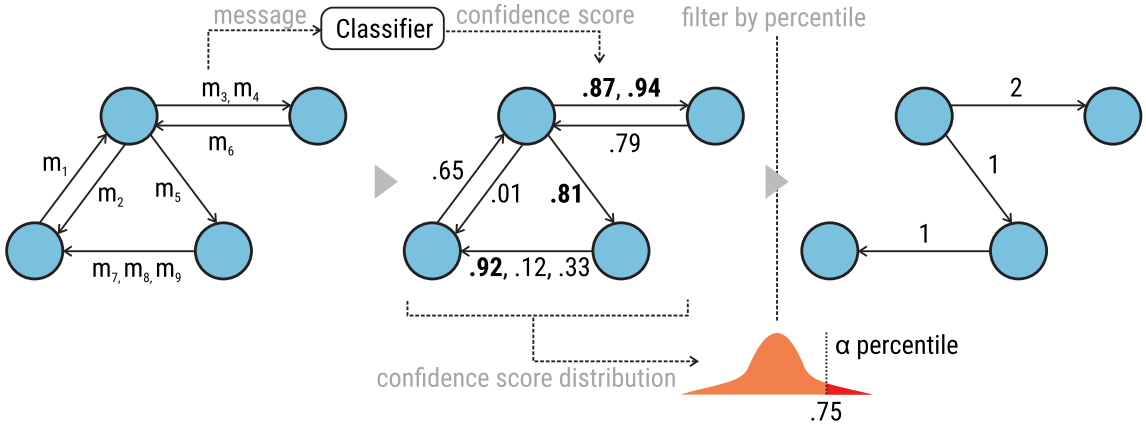

Figure 4. Example of how a dimension-specific conversation multigraph $\mathcal{G}_d$ is built. First, the text classifier for dimension d is applied to all messages and outputs scores that are proportional to the likelihood of a message containing dimension d. Then, for each dimension individually, a score threshold is determined based on a selected percentile α in the overall score distribution. In the illustrated example, the value corresponding to the α percentile is 0.75. Last, only the edges with the messages that pass that threshold are kept; the messages are counted to compute the edge weight.

Communication graphs. In Reddit, conversations develop over discussion threads. If user i commented over either a submission or a comment of another user j, we considered that i sent a *message* to j. We created a directed communication graph $\mathcal{G}\ (U, E)$ to model such exchange of messages. The set of nodes U contains all the geo-referenced users in our sample. We connected two users i and j with a directed edge $(i, j, w(i, j)) \in E$ if user i sent at least one message to user j. The edge weight $w(i, j)$ represents the ties strength and it is equal to the total number of messages sent. Enforcing a minimum threshold on edge weights for them to be included in the communication graph improved the results, likely because it filters out "occasional" interactions that do not provide a strong signal about the type of social relationships. We used the optimal threshold of $w(i, j) \geq 4$; in Supplementary Information (Fig. SI2, left) we present results with different thresholds.

By labeling each message according to the ten social dimensions, we could extract dimension-specific conversation graphs $\mathcal{G}_d$, namely a subgraph of $\mathcal{G}$ created using only the messages that contain dimension d. We built such subgraph using the procedure illustrated in Fig. 4. Given a message m, we computed its classifier score $s_d(m)$, which is proportional to the likelihood of m containing expressions of dimension d. We kept only the messages whose likelihood is higher than a dimension-specific threshold: $s_d(m) \geq \theta_d$. In practice, we assigned to θ_d the value of the 99th percentile of the empirical distribution of $s_d(m)$ values, which effectively retains only 1% of the messages. Given such a heavy filtering, we did not enforce a threshold on edge weights. Based on this reduced sets of messages, we constructed a new dimension-specific graph $\mathcal{G}\ (U_d, E_d)$ that was effectively a subgraph of the original communication graph where an edge $(i, j, w_d(i, j)) \in E_d$ encoded the fact that user i sent $w_d(i, j)$ messages conveying dimension d to user j. When messages were labeled with multiple dimensions, they contributed equally to multiple dimension-specific subgraphs.

Computing diversity of interactions. Eagle et al.[7] define two measures of diversity: social D_{social} and spatial $D_{spatial}$. In practice, the two metrics are highly correlated, hence in the main Results we report findings for $D_{spatial}$. In Supplementary Information, we discuss findings for both diversity measures.

Given a user i, we first calculated the proportion of the total number of messages that i sent to j, namely:

$$p_{ij} = \frac{w(i,j)}{\sum_{j=1}^{k} w(i,j)}, \tag{2}$$

where k is the total number of i's social contacts on the communication graph $\mathcal{G}$. In telephone network, the strength of a tie was measured as the total call duration, whereas we measured it as the total number of messages. We then calculated the normalized Shannon entropy of those proportions:

$$D_{social}(i) = \frac{-\sum_{j=1}^{k} p_{ij} \cdot \log(p_{ij})}{\log(k)}. \tag{3}$$

The dimension-specific social diversity was computed with an analogous formula, but taking into account only the edges in the dimension-specific graph $\mathcal{G}_d$:

$$p_{ij}^{d} = \frac{w_d(i,j)}{\sum_{j=1}^{k_d} w_d(i,j)}, \tag{4}$$

$$D_{social}^{d}(i) = \frac{-\sum_{j=1}^{k_d} p_{ij}^{d} \cdot \log(p_{ij}^{d})}{\log(k_d)}, \tag{5}$$

where k_d is the total number of i's social contacts on the dimension-specific graph $\mathcal{G}_d$. To compute the spatial diversity $D_{spatial}$, we first calculated the proportion of total volume of messages exchanged by user i with any other users living in area a:

$$p_{ia} = \frac{\sum_{j \in U_a} w(i,j)}{\sum_{j=1}^{k} w(i,j)}, \tag{6}$$

where A is the total number of areas and $U_a \subset U$ is the subset of users living in area a. We then computed the spatial diversity as the normalized entropy of the p_{ia} proportions:

$$D_{spatial}(i) = \frac{-\sum_{a=1}^{A} p_{ia} \cdot \log(p_{ia})}{\log(A)}. \tag{7}$$

The same formulation is applied to the dimension-specific graphs:

$$p_{ia}^{d} = \frac{\sum_{j \in U_a} w_d(i,j)}{\sum_{j=1}^{k_d} w_d(i,j)}, \tag{8}$$

$$D_{spatial}^{d}(i) = \frac{-\sum_{a=1}^{A} p_{ia}^{d} \cdot \log(p_{ia}^{d})}{\log(A)}. \tag{9}$$

Last, we computed the diversity values at area level by averaging the diversity scores of users living in the same area:

$$D_{social}(a) = \frac{\sum_{i \in U_a} D_{social}(i)}{|U_a|}; D_{social}^{d}(a) = \frac{\sum_{i \in U_a} D_{social}^{d}(i)}{|U_a|} \tag{10}$$

$$D_{spatial}(a) = \frac{\sum_{i \in U_a} D_{spatial}(i)}{|U_a|}; D_{spatial}^{d}(a) = \frac{\sum_{i \in U_a} D_{spatial}^{d}(i)}{|U_a|} \tag{11}$$

Linear regression. Linear regression is an approach for modeling a linear relationship between a dependent variable (GDP, in our experiments) and a set of independent variables (diversity measures), and it does so by associating a so-called β-coefficient with each independent variable such as the sum of all independent variables multiplied by their respective β-coefficients approximates the value of the dependent variable with minimal error. Specifically, we used an Ordinary Least Squares (OLS) regression model to estimate the coefficients such that the sum of the squared residuals between the estimation and the actual value is minimized. The diversity metrics given in input to the regression were approximately normally distributed and bounded in the interval [0,1] (see Fig. SI5)

Modeling geographical span. To study the dependency between geographical space and social dimensions, we estimated the conditional probability $p(d|l)$ of a dimension d occurring in conversations characterized by a given *geographic span* (or length) l. Specifically, we considered the set $E@l$ of all edges in the conversation graph $\mathcal{G}$ that connect users at geographic distance l, and the subset of those edges $E_d@l$ that belong to the dimension-specific graph $\mathcal{G}_d$. We then computed the conditional probability as the number of dimension-specific edges over the total number of edges at distance l, namely: $p(d|l) = \frac{|E_d@l|}{|E@l|}$.

Because activity and connectivity are not uniformly distributed across states, the probability $p(d|l)$ alone could yield a biased view of the interplay between interactions and space. To understand why, consider a scenario in which most of the users are concentrated in one single state. In such a scenario, all users would be constrained to interact mostly with people from that state, and the resulting spatial patterns will be just reflecting the underlying activity and spatial distributions rather than being indicative of explicit user choices. To account for this, we discounted $p(d|l)$ by a probability $p_{null}(d|l)$ computed on randomized data. In particular, we generated a random *null* model by randomly reshuffling the locations across users. By doing so, we preserved both the connectivity properties of the conversation network and the population distribution across states, yet destroying the original relationship between social links and spatial locations. Finally, we computed a normalized score $\Delta p(d|l) = \frac{p(d|l)}{p_{null}(d|l)} - 1$, which measures the % change of the probability of interaction compared to what it is expected by chance. To obtain the conditional probability associated to individual messages rather than social links, we also computed an alternative version of $\Delta p(d|l)$ that considers each message as an individual edge in the graph, thus effectively weighting more pairs of individuals who communicated often.

Since we could geo-reference users at state-level only, we approximated the span of a social link between two users to the length of the straight line connecting the geographic centroids of their states. Given the relatively limited spatial resolution of such a definition, we were bound to a coarse partitioning of distances. Effectively, we divided the set of edges in quintiles based on their geographic span distribution, thus obtaining five equally-sized distance bins, the first of which contains almost exclusively interactions among people in the same state ($l = 0$).

Data availability

We made all the data used in this study publicly available. The data consists of: (1) individual messages scored with the ten dimension classifier and the identifiers of the sender and receiver; (2) estimated location of the users in the communication graph; (3) aggregated data at state-level reporting the diversity metrics. The DOI of the publicly accessible data is https://doi.org/10.6084/m9.figshare.19918231. The pre-trained social dimensions classifier is available at http://www.github.com/lajello/tendimensions.

Received: 29 August 2022; Accepted: 12 December 2022
Published online: 21 December 2022

References

1. Rogers, E. M. *Diffusion of Innovations* (Simon and Schuster, 2010).
2. Granovetter, M. The impact of social structure on economic outcomes. *J. Econ. Perspect.* **19**, 33–50 (2005).
3. Holt-Lunstad, J., Smith, T. B. & Layton, J. B. Social relationships and mortality risk: A meta-analytic review. *PLoS Med.* **7**, e1000316 (2010).
4. Fowler, J. H. & Christakis, N. A. Dynamic spread of happiness in a large social network: Longitudinal analysis over 20 years in the Framingham heart study. *BMJ* **337**, a2338 (2008).
5. Granovetter, M. S. The strength of weak ties. In *Social Networks* 347–367 (Elsevier, 1977).
6. Marsden, P. V. & Campbell, K. E. Reflections on conceptualizing and measuring tie strength. *Soc. Forces* **91**, 17–23 (2012).
7. Eagle, N., Macy, M. & Claxton, R. Network diversity and economic development. *Science* **328**, 1029–1031 (2010).
8. Wellman, B. & Wortley, S. Different strokes from different folks: Community ties and social support. *AJS* **96**, 558–588 (1990).
9. Deri, S., Rappaz, J., Aiello, L. M. & Quercia, D. Coloring in the links: Capturing social ties as they are perceived. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work and Social Computing, CSCW* 1–18 (ACM, 2018).
10. Choi, M., Aiello, L. M., Varga, K. Z. & Quercia, D. Ten social dimensions of conversations and relationships. In *Proceedings of the Web Conference, WWW* (ACM, 2020).
11. Choi, D. *et al.* Characterizing conversation patterns in reddit: From the perspectives of content properties and user participation behaviors. In *Proceedings of the ACM Conference on Online Social Networks* 233–243 (2015).
12. Baeza-Yates, R. & Saez-Trumper, D. Wisdom of the crowd or wisdom of a few? an analysis of users' content generation. In *Proceedings of the 26th ACM Conference on Hypertext and Social Media* 69–74 (2015).
13. Aral, S. & Van Alstyne, M. The diversity-bandwidth trade-off. *Am. J. Sociol.* **117**, 90–171 (2011).
14. Reagans, R. & McEvily, B. Network structure and knowledge transfer: The effects of cohesion and range. *Admin. Sci. Q.* **48**, 240–267 (2003).
15. Bell, G. G. & Zaheer, A. Geography, networks, and knowledge flow. *Organ. Sci.* **18**, 955–972 (2007).
16. Mok, D. *et al.* Did distance matter before the internet?: Interpersonal contact and support in the 1970s. *Soc. Netw.* **29**, 430–461 (2007).
17. Hampton, K. & Wellman, B. Long distance community in the network society: Contact and support beyond Netville. *Am. Behav. Sci.* **45**, 476–495 (2001).
18. Mesch, G. S. & Manor, O. Social ties, environmental perception, and local attachment. *Environ. Behav.* **30**, 504–519 (1998).
19. US Bureau of Economic Analysis. Economic estimates for year 2017 (2017). https://apps.bea.gov/histdata/.
20. Bettencourt, L. M. The origins of scaling in cities. *Science* **340**, 1438–1441 (2013).
21. Aral, S. The future of weak ties. *Am. J. Sociol.* **121**, 1931–1939 (2016).
22. Rajkumar, K., Saint-Jacques, G., Bojinov, I., Brynjolfsson, E. & Aral, S. A causal test of the strength of weak ties. *Science* **377**, 1304–1310 (2022).
23. Lambiotte, R. & Kosinski, M. Tracking the digital footprints of personality. *Proc. IEEE* **102**, 1934–1939 (2014).
24. Morstatter, F., Pfeffer, J., Liu, H. & Carley, K. Is the sample good enough? Comparing data from twitter's streaming API with twitter's firehose. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 7, 400–408 (2013).
25. Medvedev, A. N., Lambiotte, R. & Delvenne, J.-C. The anatomy of reddit: An overview of academic research. In *Dynamics On and Of Complex Networks* 183–204 (Springer, 2017).
26. Anderson, K. E. Ask me anything: What is reddit? Library Hi Tech News (2015).

27. Moore, C. & Chuang, L. Redditors revealed: Motivational factors of the reddit community. In *Proceedings of the 50th Hawaii International Conference on System Sciences* (2017).
28. Park, N., Kee, K. F. & Valenzuela, S. Being immersed in social networking environment: Facebook groups, uses and gratifications, and social outcomes. *Cyberpsychol. Behav.* **12**, 729–733 (2009).
29. Phua, J., Jin, S. V. & Kim, J. J. Uses and gratifications of social networking sites for bridging and bonding social capital: A comparison of Facebook, Twitter, Instagram, and Snapchat. *Comput. Hum. Behav.* **72**, 115–122 (2017).
30. Weninger, T., Zhu, X. A. & Han, J. An exploration of discussion threads in social news sites: A case study of the reddit community. In *2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2013)* 579–583 (IEEE, 2013).
31. Proferes, N., Jones, N., Gilbert, S., Fiesler, C. & Zimmer, M. Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Soc. Media Soc.* **7** (2021).
32. Newman, M. E. & Park, J. Why social networks are different from other types of networks. *Phys. Rev. E* **68**, 036122 (2003).
33. Leskovec, J., Backstrom, L., Kumar, R. & Tomkins, A. Microscopic evolution of social networks. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 462–470 (2008).
34. Ugander, J., Karrer, B., Backstrom, L. & Marlow, C. The anatomy of the Facebook social graph. arXiv preprint arXiv:1111.4503 (2011).
35. Weninger, T. An exploration of submissions and discussions in social news: Mining collective intelligence of reddit. *Soc. Netw. Anal. Min.* **4**, 1–19 (2014).
36. Cauteruccio, F., Corradini, E., Terracina, G., Ursino, D. & Virgili, L. Investigating reddit to detect subreddit and author stereotypes and to evaluate author assortativity. *J. Inf. Sci.*https://doi.org/10.1177/016555152097986 *(2020).*
37. Baowaly, M. K., Kibirige, G. W. & Singh, B. C. Co-comment network: A novel approach for construction of social networks within reddit. *Computación y Sistemas* **26**, 311–323 (2022).
38. Soliman, A., Hafer, J. & Lemmerich, F. A characterization of political communities on reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media* 259–263 (2019).
39. Wang, C., Ye, M. & Huberman, B. A. From user comments to on-line conversations. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 244–252 (2012).
40. Dixon, S. Distribution of Reddit users, and percentage of U.S. adults who use Reddit. In Statista.com (2022).
41. Baumgartner, J., Zannettou, S., Keegan, B., Squire, M. & Blackburn, J. The pushshift reddit dataset. arXiv preprint arXiv:2001.08435 (2020).
42. Balsamo, D., Bajardi, P. & Panisson, A. Firsthand opiates abuse on social media: Monitoring geospatial patterns of interest through a digital cohort. In *Proceedings of the World Wide Web Conference, WWW* 2572–2579 (ACM, 2019).
43. Fiske, S. T., Cuddy, A. J. & Glick, P. Universal dimensions of social cognition: Warmth and competence. *Trends Cogn. Sci.* **11**, 77–83 (2007).
44. Blau, P. M. *Exchange and Power in Social Life* (Transaction Publishers, 1964).
45. Luhmann, N. *Trust and Power* (Wiley, 1982).
46. Buss, D. M. *The Evolution of Desire: Strategies of Human Mating* (Basic Books, 2003).
47. McPherson, M., Smith-Lovin, L. & Cook, J. M. Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.* **27**, 415–444 (2001).
48. Tajfel, H. *Social Identity and Intergroup Relations* (Cambridge University Press, 2010).
49. Argyle, M. *The Psychology of Happiness* (Routledge, 2013).
50. Tajfel, H., Turner, J. C., Austin, W. G. & Worchel, S. *An Integrative Theory of Intergroup Conflict* (Organizational Identity, 1979).
51. Fiske, A. P. The four elementary forms of sociality: Framework for a unified theory of social relations. *Psychol. Rev.* **99**, 689–723 (1992).
52. Spencer, L. & Pahl, R. *Rethinking Friendship: Hidden Solidarities Today* (Princeton University Press, 2006).
53. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
54. Sundermeyer, M., Schlüter, R. & Ney, H. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association* (Interspeech, 2012).
55. Pennington, J., Socher, R. & Manning, C. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP* 1532–1543 (Association for Computational Linguistics, 2014).
56. Hutto, C. J. & Gilbert, E. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Weblogs and Social Media, ICWSM* 216–225 (AAAI, 2014).

Acknowledgements

LMA acknowledges the support from the Carlsberg Foundation through the COCOONS project (CF21-0432). The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author contributions

LMA conceived the experiments, conducted the analysis, and wrote the manuscript. SJ collected the data and revised the manuscript. DQ conceived the experiments and wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-26245-4.

Correspondence and requests for materials should be addressed to L.M.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.