

# Twitter ain't Without Frontiers: Economic, Social, and Cultural Boundaries in International Communication

Ruth Garcia-Gavilanes  
Universitat Pompeu Fabra  
Barcelona  
ruth.garcia@upf.edu

Yelena Mejova  
Yahoo Labs  
Barcelona  
ymejova@yahoo-inc.com

Daniele Quercia  
Yahoo Labs  
Barcelona  
dquercia@acm.org

## ABSTRACT

With the advent of Twitter and other lightweight social-networking services, one might think that it is easier than ever to maintain geographically dispersed, weaker social ties. By contrast, in this study we show that the international Twitter communication landscape is not only still largely predetermined by physical distance, but that it also depends on countries' social, economic, and cultural attributes. We describe a study of an international Twitter mention network of 13 million users across over 100 countries. We show that the *Gravity Model*, which hypothesizes that the flow between two areas is proportional to their masses (which we approximate using internet penetration) and inversely proportional to the distance between them, is correlated ( $r = 0.68$ ) with the international communication flow. Using this model, along with other social, economic, and cultural variables, we predict the communication volume at *Adjusted  $R^2$*  of 0.80, with trade, language and racial intolerance especially impacting communication. We discuss the implications of these barriers to communication in the contexts of collaborative work, software design, and recommendation systems.

## Author Keywords

Communication; Social Media; Culture

## ACM Classification Keywords

J.4 Social and Behavioral Sciences: Sociology

## INTRODUCTION

The rise of the Internet and social networks have lead some researchers to hypothesize that “distance is dead” [6] or is not longer important to make social contacts. At the very conception of online networking pundits predicted the loosening of the “grip of geography” [7], foreseeing the strengthening of the bonds between people with the same interest in different parts of the world, and globalization of both the workforce and the scope of governmental considerations. Nevertheless, empirical studies have shown that distance still matters in online communication, including email [32, 25] and instant

messages [23], with these new modes of communication reinforcing the strong ties we make in person. However, recently other factors were shown to mediate the effect of distance, including language, air travel frequency [34], and culture [32]. For instance, countries sharing cultural features have a higher affinity in international email exchanges, and can be effectively clustered into “civilizations”, as suggested by Samuel Huntington in “The Clash of Civilizations” [16].

Finding whether Internet users are trapped in socio-economic or cultural “bubbles”, despite the supposed freedom and multi-cultural nature of the web, is a first step to identifying the blind spots in our communication. Specifically, cultural dimensions have long been studied by sociologists. One way to measure cultural values, as they relate to personal behavior, is using Hofstede’s culture indexes [14]. Available for many countries around the world, they characterize authority relations, the relationship between individual and society, gender roles, and social and environmental uncertainties. We bring these into the realm of social media analysis by relating the international communication flows in Twitter to the extent to which countries share these cultural characteristics and various other country-specific attributes.

Recent wide adoption of Twitter has fostered a global network of relatively weak ties based on user interests. As defined by Granovetter [13], the strength of a tie is “a combination of the amount of time, the emotional intensity, the intimacy (mutual confiding), and the reciprocal services which characterize the tie”. Since Twitter messages are short (140 characters), and are broadcast publicly, easy to both read and to ignore, Twitter provides a perfect platform for the establishment of weak ties. Also, the connections need not be reciprocal, and users are free to ‘follow’ (subscribe to) any other user with a public profile in order to see the posts, or status updates, of that user in their timeline. Furthermore, users are free to contact others (being their followers or not) by simply mentioning their users names. The value of such characteristics make Twitter a useful tool for exploring communication in online social media, going beyond the strong ties of personal e-mails or Facebook. Indeed, we find major differences between the importance of economic and cultural factors in Twitter communication as compared to e-mail, as described in [32].

Here, we explore how various factors (distance, social, economic, and cultural dimensions) shape the cross-country communication through the lightweight social networking services. Specifically, we address two questions: (1) To what extent does distance determine the informal communication

of users from different nations? and (2) To what extent do social, economic and cultural factors mediate/impede this communication? To tackle these questions, we study user mentions among 13 million geolocated users during a 10 week period from March to May 2011. Using this data, covering 111 countries, along with country-specific statistics gathered from outside sources (CIA, World Banks and World Values Survey), we make two main contributions:

- We employ the *gravity model* [20], which uses node population and physical distance, to construct a baseline communication network, and test to which extent it estimates cross-country Twitter communication. We use the Haversine distance between two countries and two population proxies: country population and the Internet penetration, with the latter showing moderate correlation with the number of mentions and retweets of 5,932 pairs of countries ( $r = 0.68$  for unique mentions, and  $r = 0.66$  for unique retweets).
- We build a regression model that uses economic, social and cultural country attributes, along with the gravity model to predict communication volume between pairs of countries. We find that the complete model performs well with  $Adjusted R^2 = 0.80$ , illustrating the importance of social economic and cultural variables in bilateral online communication.

We conclude by discussing the design implications of these findings in the realms of collaborative work, software design, and recommendation systems.

## RELATED WORK

The emergence of Internet social platforms has enabled, fostered, and recorded social networks of an unprecedented scale. Containing social connections, virtual communities, and the content they produce and share, these platforms enable researchers to study social phenomena, extending the reach of conventional sociological studies [19].

A number of studies have used confidential communication to examine the social connections between individuals across the world. A well known study by Leskovec & Horvitz [23] uses the private messages to build a “planetary scale” social network of 180 million nodes, and examines social phenomena, such as Milgram’s “6 degrees of separation” [36] (finding that, indeed, the users of the service had an average path length of 6.6).

Specialized communication has also been considered. A community of travelers on CouchSurfing.com was studied by Lauterbach *et al.* [21], who attempted to predict the trust the users display toward one another. They show that, among more personal variables (such as whether the users have met in person), whether users are from the same country affects the chances of one user vouching for another. Olson *et al.* [27] carried out empirical studies of remote work, both in the field and in the laboratory, concluding that distance impacts the quality of end result, regardless of the technology used. More recently, Takhteyev [33] discussed examples of successful collaboration over long distances by looking at how several

cultural and geographic constraints were negotiated in the face of increasingly “global” knowledge and technology.

Across social media, geographical distance has been shown to play a major role in human connections. Scellato *et al.* [29] show that, among the users in Brightkite, Foursquare, and Gowalla communities, 40% of links are made in a radius of under 100km. Similar results were found of a 2,852-user sample of the Twitter network in 2009 by Takhteyev *et al.* [34] with 35% of links being under 100km but they also find that other variables, such as the commonalities in language and the extent of air travel to be more predictive of Twitter communication than physical distance. They speculate that air travel may stand as “a proxy for other kinds of pre-existing connections between places, which in turn influence formation of electronic ties”. Inspired by this, we examine social, economic, and cultural factors in international Twitter communication. However, a marked difference between these previous studies and one described here is the network construction process. Instead of using follower or followee edges (subscriptions), which do not necessarily imply active communication or attention [38] with 25% of Twitter users never tweeting at all [2], we use geolocated user mentions in nearly 3 billion posts.

Though first, we use the *Gravity Model* to capture the effect of distance. Inspired by Newtonian physics, it models the importance of physical distance in communication between two populated nodes, using a proportion of the population sizes of the two nodes to the distance between them [20]. For example, it has been applied to modeling road and airline networks [1, 17], phone calls [20], and flows of passengers in a London metro system [31].

We also build on the seminal work by Hofstede [14] to measure cultural differences. In 1980, he conducted a survey of IBM employees from 55 different countries, examining a wide range of cultural values. Specifically, we use six cultural indexes: the power-distance index (PDI), individualism-collectivism (IDV), masculinity-femininity (MAS), uncertainty avoidance (UAI), indulgence-restraint (IVR), and the long term orientation index (LTO). Hofstede’s cultural dimensions have also been used to study differences in social media usage, for example previous work has found that culture of a country is associated with the way people use Twitter [10].

Although popular in cross-cultural sociology and psychology literature, these measures have largely not been used in internet-mediated communication studies. A notable exception is an email communication study by State *et al.* [32] who examine the extent to which inter-national communication flows according to the civilization, as defined by Samuel Huntington in *The Clash of Civilizations* [16]. By including geographic, economic and cultural factors in their regression model (including the first four of Hofstede’s cultural indexes), they show that the membership in the same Huntington civilization to nearly double the pairwise communication density, increasing it by factor of 1.941. However, since the notion of civilization encompasses both geographic and religious attributes of the countries, we find it unsuitable in our aim of separating geography from culture.

# of geolocated users	13,139,763
# of tweets (with mentions)	2,924,398,138
# of mentions	534,868,476
# of unique mentions	258,534,246
# of countries with > 1K users	111
# of country pairs with complete predictors	481

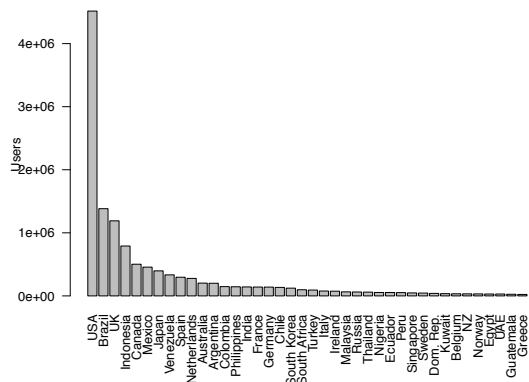
**Table 1: Summary of the dataset.** We identified the geolocation of more than 13M users but considered only the countries with more than 1K users, which represents more than 90% of our sample users. In total, we obtained 481 country-pairs with no missing attribute values for regression analysis.

In summary, to the best of our knowledge, we present a previously unattempted study of international Twitter communication which combines cultural information with geographic, economic, and social features, using a variety of outside sources from the CIA and World Bank. Unlike e-mail, Twitter mention graph goes beyond strong ties of interpersonal communication, potentially breaking down barriers of distance and culture. Also, unlike the previous studies on Twitter which use subscriptions instead of tweet content [28, 34], this study focuses on the active conversation and attention beyond the user’s immediate follower/followee network. Finally, we use the gravity model and a variety of other predictors to build a communication model based solely on data independent of the particulars of Twitter dataset.

## DATASET

We model communication across countries in Twitter by observing mentions and retweets by users in one country involving users from another. Similar to [32], we say that a communication is established from country  $a$  to  $b$  when a message is posted by a user from  $a$  mentioning a user from  $b$ . A “mention” consists of any Twitter username preceded by the *at* symbol (@). So, for example, if user @Maria located in Spain creates a post “@BarackObama is the president of the United States,” we know two things: a) BarackObama received a notification in his account (although unlikely to answer) and b) a communication was made from Spain to USA. The interpretation of this phenomena consists of both conversation and attention, in that, mentions and retweets may be used in a conversation between users, but may also signify an awareness of another user (as with BarackObama and the notification he received in the previous example). Thus, in this paper, when we refer to the number of mentions or retweets as “communication”, we do so loosely.

The data was collected as follows. We first randomly selected 55K users who tweeted at least once on March 2011 and obtained their profile information. From this information, we selected users with out-degree and in-degree in the range of [100, 1000] and crawled their corresponding followee network (for a user  $u$ , it is all users who  $u$  is following). This choice was made in order not to exceed the limit of the API calls. It also has the added benefit of filtering away less legitimate (e.g., spam) users, since, according to [22], the majority of spam users tend to have out-degree and in-degree outside the range of [100, 1000]. Also [35] show that 89% of users following spam accounts have fewer than 10 followers. So,



**Figure 1: Number of users per the country in the sample (showing top 40 countries).**

while we can not guarantee that all users in our dataset are not spammers, previous studies indicate that our sample will indeed have a higher probability of containing legitimate users.

We then proceeded to collect all of the tweets posted by the original 55K users as well as their followees during 10 weeks starting from the second half of March 2011. We also collect all tweets containing a mention of any user of our sample (i.e., identified by @username) and the user profile of who posted these tweets.

We continue by finding the geolocation of each user via the location field entered in their profiles. Often these locations are either strings specified by the users themselves or GPS coordinates coming from their mobile devices. We then map these locations into  $(long, lat)$  points (using Yahoo! Place-Maker<sup>1</sup> for user-specified strings), resulting in 13 million geo-located users.

To alleviate any bias due to the selection of seed users and obtain representative samples, we only consider the countries with more than 1,000 users in our sample.

Figure 1 shows sample sizes, with the typical skew across the countries with USA having by far the largest share of Twitter users, followed by Brazil, United Kingdom, and Indonesia. Seven out of the top 10 countries in our sample overlap with the top 10 countries by site traffic in 2011<sup>2</sup> and we also find a *Pearson correlation* of 0.72 to the corresponding logarithm of the number of internet users in 2011 reported by the US’s Central Intelligence Agency<sup>3</sup> (CIA). The discrepancy can be attributed to our sampling method which favors users who are mentioned, thus promoting Mexico, Venezuela, and Netherlands in our ranking, excluding Germany, India, and Australia which appear in the traffic ranking.

Table 1 shows statistics about the final dataset, including the number of geolocated users and their tweets, and mentions

<sup>1</sup><http://www.programmableweb.com/api/yahoo-placemaker>

<sup>2</sup><http://www.marketinggum.com/twitter-statistics-2011-updated-stats/>

<sup>3</sup><https://www.cia.gov/library/publications/the-world-factbook/rankorder/2153rank.html>

	Sample Size	$\gamma$	Internet Penet.	$\gamma$	Country Population	$\gamma$
Mentions	0.915	0.83	0.670	0.42	0.489	0.84
Unique mentions	0.919	0.83	0.679	0.43	0.501	0.84
Retweets	0.911	0.88	0.676	0.49	0.505	0.92
Unique retweets	0.904	0.87	0.663	0.48	0.492	0.91

**Table 2: Pearson correlation between observed Twitter interactions and gravity model estimations using three different population masses ( $N = 5392$  country pairs) and adjusted distance exponent ( $\gamma$ )**

found in those tweets. We count “unique” mentions per user, summing the number of unique accounts mentioned by each. On average, for each user-user conversation, there is one duplication, since it is common to mention a specific user more than once (same holds true for unique retweets).

With the geolocation information, we can now analyze the communication across countries. We do this by mapping the country of the mentioned users to the countries of those who posted the tweets, obtaining a country to country graph. Since we are interested in measuring the flow of information between countries and not the direction of it, we obtain an undirected graph of the inter-country communication by adding the bilateral number of mentions and retweets between a pair of countries. Furthermore, we discard self-loop edges since we are interested in communication across countries, not within. This resulted in 5,392 country-country pairs.

Finally, to tackle our hypotheses and objectives, we obtain geographic, social, economic, and cultural features of these countries. We collect the number of direct flights between each of the countries<sup>4</sup>, as well as the spoken languages in each country, as reported by the CIA. Additional social, economic and cultural indicators came from the WorldBank API for R<sup>5</sup>. Each of the variables is explained in the *Social, Economic & Cultural Predictors* section. Since data was missing for some of the countries, we excluded the records with no values. This gave us a total of 481 pairs with complete information for each predictor variable in our model.

## GRAVITY MODEL

### Definition

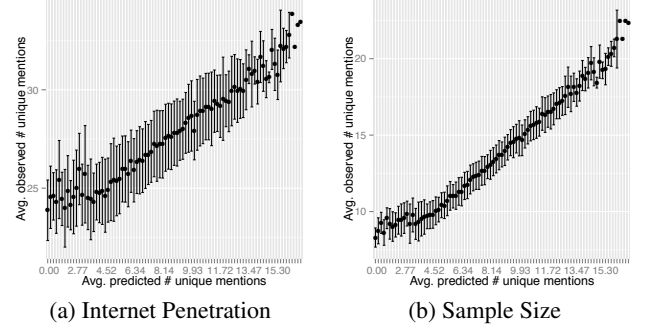
In its simplest formulation, the Gravity Model posits that the gravitational interaction between two places is proportional to their mass and inversely proportional to the distance between [41] and it takes the form of:

$$I_{1,2} = k \frac{p_1^\alpha * p_2^\beta}{d_{1,2}^\gamma} \quad (1)$$

where  $I_{1,2}$  is the volume of interaction between communities 1 and 2,  $k$  is a constant,  $p_1$  and  $p_2$  refer to the “population

<sup>4</sup><http://openflights.org/data.html>

<sup>5</sup><http://www.r-chart.com/2010/06/world-bank-api-r-package-available.html>



**Figure 2: Unique mentions versus gravity model using (a) internet penetration and (b) sample size, with standard deviations of unique mentions. The country pairs are first binned by estimated flow, then we plot the mean estimated flow in each bin vs. the mean observed flow of the edges in each bin. The error bars show the standard deviation of the observed flows in each bin.**

mass” (that is, community size) of communities 1 and 2, and  $d_{1,2}$  refers to the distance between these communities. The exponents  $\alpha$ ,  $\beta$ ,  $\gamma$  and the scaling factor  $k$  are adjustable parameters chosen to fit the data modeled. The pure gravity model is retained if the population exponents ( $\alpha$  and  $\beta$ ) are 1 and the distance exponent ( $\gamma$ ) is 2; but the formula allows the exponents to be adjusted to finely tune the data being modeled.

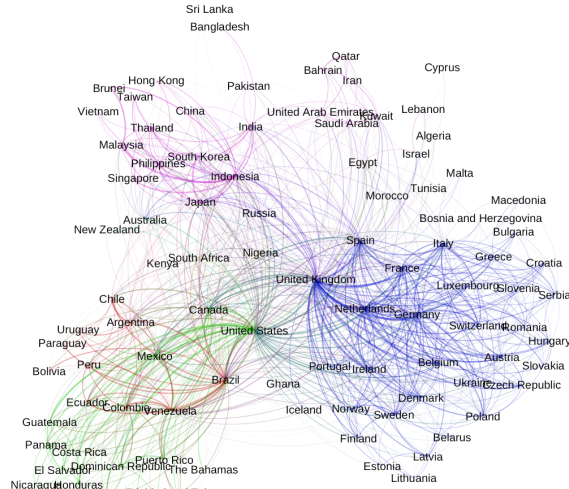
It has been previously shown that the gravitational model is applicable to various phenomena such as telecommunication, email and transportation flow between countries, cities and within cities [30]. Since the gravity model can be used to account for any interaction or flow from one place to another, we apply it to estimate the volume of Twitter traffic between two countries and adjust the  $\gamma$  exponent to better fit our data.

We employ several alternative proxies for the “population mass” in the gravity model: (i) sample size, (ii) internet penetration and (iii) country population; and as proxy for distance, we use the Haversine distance<sup>6</sup> (distance between two points on a sphere). We examine the correlation with Twitter bilateral interactions, measured as a) the number of mentions, b) unique mentions, c) retweets, and d) unique retweets. For the case of re-tweets (cases c and d), we counted only the original authors of a tweet ignoring all other mentions.

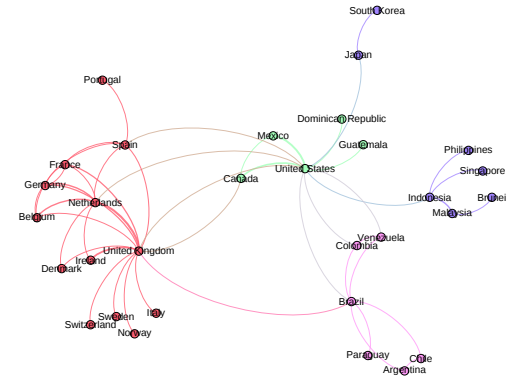
Table 2 shows the Pearson correlation between Twitter interactions and gravity model estimations using three different estimates for the population mass, along with the best values of  $\gamma$  for each. Sample size produces the strongest correlation with all four measures, at  $r = 0.919$  with unique mentions, with no significant difference in communication flow across countries between re-tweets and mentions. In the following regression experiments, to make sure the dependent variable is not related to the predictors, instead of using sample size we use Internet penetration as a proxy for population.

In Figure 2, we show the distribution of observed unique mentions vs. estimated mentions flows, using (a) internet penetra-

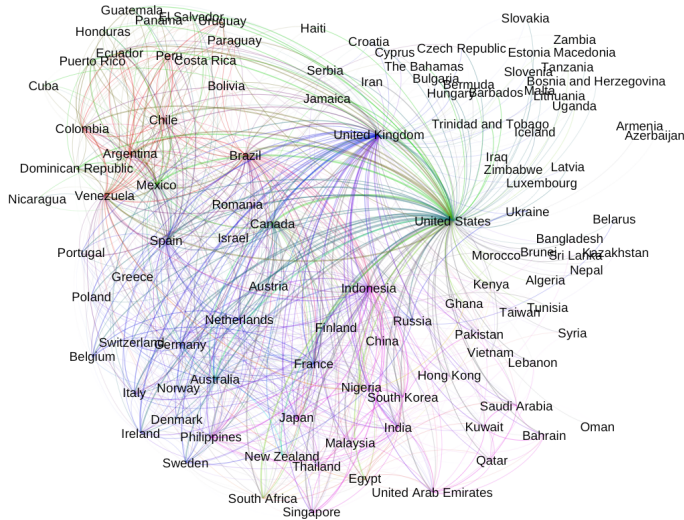
<sup>6</sup><https://github.com/linkedin/datafu>



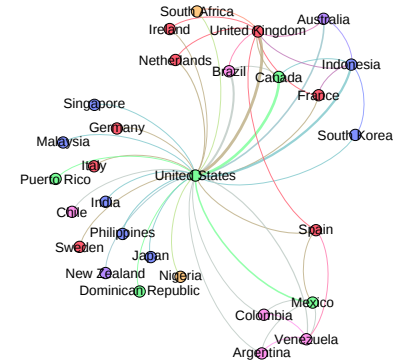
(a) Gravity Model



(a) Gravity Model



(b) Unique Mentions



(b) Unique Mentions

**Figure 3: Cross-country communication network, 1000 most prominent edges, color-coded by continent.**

tion and (b) sample size as proxies for population. We see that both versions of gravity model provide estimates which correlate well with observed Twitter interactions. Sample size provides the best estimation, but the standard deviation tends to increase as communication increases.

Finally, we visualize the mention networks induced by these two measures by selecting the top 1000 strongest edges (Figure 3) and top 50 edges (Figure 4). The nodes are positioned using force-directed algorithm using log-transformed edge weights, and colored according to the continent on which they reside. The countries in gravity model network are largely clustered according to their geography, with most populated countries (Brazil, US, UK) connecting them at the center. The network built using unique mentions also has UK and US as central hubs, however the geographical differences between

**Figure 4: Cross-country communication network, 50 most prominent edges, color-coded by continent.**

the countries are less pronounced. Now, Spain is much closer to Mexico and its South American peers in language. In larger mentions network, countries with a smaller Twitter use are also pushed out into their own group in the upper right, including countries from Africa, Eastern Europe, and Middle East. Although the major players remain the same (partially because we are using our sample size in the gravity model), the geographic separations are less pronounced in the Twitter-induced network. Next, we examine the extent to which international communication is explained using features other than physical distance.

## SOCIAL, ECONOMIC & CULTURAL PREDICTORS

The high correlation of retweets and mentions with the gravity model testifies to the importance of distance. Nevertheless, the standard deviation of the observations seen in Figure

2 show that there are other factors to take into account when studying cross country communication. We observe an acute tendency to overestimate communication flows. This is especially the case for European countries which are very close together and have a large number of people online, including Germany, The Netherlands, Poland, and United Kingdom. Same is true for China and Japan. Similarly, communication between countries which are far apart, such as United States and United Kingdom, and United States and Australia, is underestimated. This behavior suggests that the model does not take into account important information, such as culture and other international connections. We now proceed to study 16 variables that we hypothesize will impact communication. These variables are classified into social, economic and cultural.

#### *Economic Indicators*

Does difference in income divide people? Despite the fact that the so called “liberation technologies” have and continue to alter information propagation across countries during crises, the boundaries separating high- from low- income countries affect the daily real world interactions between people, and therefore affect interactions online [24]. In fact, where income differences are bigger, social distances are bigger and social stratification more important [40]. We use predictors (in American dollars) that account for economy described as follows.

*Income:* We take the GDP per capita for each pair of countries and multiply them. A high product stands for the combination of two wealthy nations.

Similarly, the trade relationships between countries has been shown to be affected by ease of communication [5]. For this reason, we assume the trade between two countries should also be taken as a predictors of communication and we do so under three perspectives (metrics obtained from the World Bank)<sup>7</sup> :






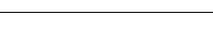










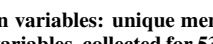
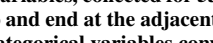
*Export importance:* We propose a metric that measures the importance of the exports between two countries with regard to the overall exportations of both countries – we add their pair-wise exports and normalize by the total sum of all their exports.

*Trade intensity index:* The value of trade between two countries on the basis of their importance in the world trade. This metric is defined as the share of one country’s exports going to a partner divided by the share of world exports going to the partner. To obtain one value per country pair, we multiply their corresponding trade intensity index.

*Trade market share of total exports:* It measures how much of the world import demand is covered by the country’s exports. Similarly, to obtain a single value per country pair, we multiply the share of total exports for each country.

#### *Social Indicators*

We present four social variables expected to affect information flow related to migration and air travel.

	Distribution	Max
Unique Mentions		58,214,512
Gravity Model		2,731,487
<b>Economic Variables</b>		
Income		802,604.5
Exports		0.35
Trade Intensity		395.8
Trade Market Share		92.6
<b>Social Variables</b>		
Routes		6.68
Emigration		0.83
Migration		0.05
Migration Rate		39.9
<b>Cultural Variables</b>		
Language		1
Intolerance		86
Power Distance		82
Individualism		84
Masculinity		90
Uncertainty avoidance		104
Long Term Orientation		88
Indulgence vs. Restrain		97

**Table 3: Statistics of regression variables: unique mentions (dependent variable) and 17 independent variables, collected for 5392 country pairs. The distributions begin at zero and end at the adjacent maximum. Language and income group are categorical variables converted to numeric factors. There are 481 pairs having values for all the predictive variables.**

The term “transnational migrants” refers to the extent to which immigrants keep cross-border ties when sharing political or religious ideas as well as maintaining cross-border activities of travel, remittance flow and telephone communication with their home-country. These interlocking networks across national boundaries are even more evident with people from border-free travel zones where individuals can work and live in a different country and travel regularly to their home-countries without major bureaucratic barriers.

We propose four migration metrics :

*Net migration rate:* the difference between the number of persons entering and leaving a country during the year (per 1000 persons). A positive value indicates an excess in immigration and a negative number an excess in emigration. We calculate the absolute difference between these values for each country.

*Emigration:* obtained by summing the number of emigrants from one country to the other divided by the total number of emigrants for both countries. *Migration:* obtained by summing the number of emigrants from one country to the other divided by the total.

<sup>7</sup><http://wits.worldbank.org/>

*Direct flights:* The availability of direct flights has been proven to mediate distance when measuring social interactions [34]. Besides simplifying the process of travel of immigrants to their home-country, it fosters interactions between tourists, visitors and business partners. We consider the number of direct flights between a pair of countries and hypothesize that

#### Cultural Indicators

One of the most prominent efforts to measure cultural differences is Hofstede’s series of surveys [15] administered to the residents of over 70 countries. In his original experiment he administered opinion surveys to IBM employees in over 50 countries. In the following decades his research was expanded to other populations and countries. Using factor analysis, Hofstede identified six main cultural dimensions: 1) Power distance (*PDI*) the level of acceptance of unequal distribution of power, 2) Individualism (*IDV*) the degree to which individuals are integrated into groups, 3) Masculinity (*MAS*) the distribution of emotional roles between the genders, 4) Uncertainty avoidance index (*UAI*) a society’s tolerance for uncertainty and ambiguity, 5) Long term orientation (*LTO*), importance to tradition and the future and 6) Indulgence vs. restraint (*IVR*), describes hedonistic behaviors. We take the absolute value of the difference between the cultural indexes of a pair of countries to measure the cultural differences between their inhabitants.

To this, we have also added *Racial intolerance* as one more dimension that can strongly affect communication between people from different countries. Researches have shown that there is a causal relationship between well-being, economic freedom and tolerance [3]. Using the “World Values Survey”, racial intolerance was measured in more than 80 countries by asking participants what kinds of people they would not want as neighbors answered and calculating the percentage of those who answered “people of a different race” option. We calculate intolerance as the maximum percentage reported by the survey between a pair of countries: the highest intolerance will determine the level of communication with people from the other country.

Finally, we add *Language* to this category because it defines a culture, through the people who speak it and what it allows speakers to say. Many immigrants and tourists choose to travel to places where they can communicate, as we tend to establish social ties with people who can speak the same language. The CIA provides a rank ordering of spoken languages per country. We set the binary variable *language* to 0 if there is no common language between two countries and to 1 if there is.

#### REGRESSION

To verify the predictability power of the gravity model, as well as the economic, social and cultural variables (summarized in Table 3), we run a regression analysis, and build a model to predict the normalized volume of mentions between the countries. To avoid an excess of variables versus data points, we only consider the pairs with no missing values (resulting in 481 pairs). Finally, to account for the violations of normality exhibited by the distributions in Table 3, every variable is log transformed and then standardized.

We define communication strength as the communication volume between two countries, as measured by the number of unique user mentions between users of two countries. We choose this measure as it encompasses both conversations and unsolicited mentions of users. We normalize the raw unique mention volume between countries to a scale of  $\{0, 1\}$  in order to represent the communication flow strength of a pair of countries in comparison to the rest. The transformation was made by:

$$s_{i,j} = \frac{m_{i,j} - \min_m}{\max_m - \min_m} \quad (2)$$

where  $s_{i,j}$  is a normalized mention volume,  $m_{i,j}$  is the number of unique mentions from  $i$  to  $j$  and vice versa, and  $\min_m$  and  $\max_m$  are the minimum and the maximum observed unique mentions between any pair of countries in the dataset.

We use multiple linear regression to predict our dependent variable. Consequently, we model communication strength as a linear combination of the predictive variables and the gravity model:

$$cs_{i,j} = \alpha + \beta_1 G_{i,j} + \beta_2 R_{i,j} + \beta_3 D_{i,j} + \epsilon_{i,j} \quad (3)$$

where  $cs_{i,j}$  is the communication strength between the  $i$ -th and  $j$ -th country,  $G_{i,j}$  is the gravitational model variable,  $R_{i,j}$  is the vector of remaining predictive variables (classified into social, economic and cultural),  $D_{i,j}$  represents the pairwise interactions between all the predictors, and  $\epsilon_{i,j}$  is the error term.

**Multicollinearity.** Before applying the model, we check for multicollinearity among the model’s variables. We employ Variance Inflation Factors (VIF), which measure the extent to which errors of the estimated coefficients are inflated by the existence of correlation among the predictor variables in the model [37]. We detected two groups of variables for which VIF was high (above 4): one dealing with trade: Trade Intensity at 7.3 and Trade Market Share at 12.3, and with migration: Migration at 50.2 and Emigration at 48.6. One way of eliminating multicollinearity is to remove one of the violating predictors. Thus, we exclude Trade Intensity and Emigration from the analysis, with the resulting model showing VIFs of under 4.

#### RESULTS

When predicting the normalized communication volume, the complete model fits the data very well, with an *Adjusted*  $R^2 = 0.80$  at  $p < 0.001$  and a standard error of 0.087. This error interval is of less significance for countries with communication close to 1. Note that if we use sample size as a proxy for population, the full model achieves *Adjusted*  $R^2 = 0.95$ , suggesting over-fitting. As mentioned in *Gravity Model* section, we instead use Internet population in order to remove the effects of our sampling method. As part of a regression routine, we have looked for normality of residuals in a QQplot and observed that they follow approximately a



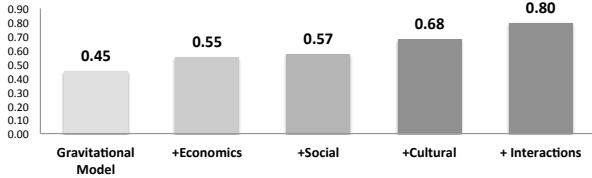


Figure 5: Adjusted  $R^2$  as new dimensions are added to the model. Modeling interactions between dimensions results in substantial performance boost.

normal distribution except for the top and bottom of the line. These outliers are understandable in this situation and therefore should not be considered as evidence for instability of the model [11].

Figure 5 summarizes the model’s performance for communication volume broken down by four predictor groups. There are notable gains when adding economic and cultural predictors to the model, but it is the interaction term that is responsible for boosting the performance to  $Adjusted R^2 = 0.80$ .

Figure 6 visualizes the predictive power of the four dimensions as part of communication volume. For this figure, we have not included interactions in order to analyze each dimension individually (recall that we have controlled for the multicollinearity). The weight of a dimension is calculated by summing the coefficients of the variables belonging to it as in [12]. As described in more detail later, cultural factors (most prominently language) and the gravity model play the largest role (also, Figure 5 shows that indeed cultural features are more important in boosting the results).

Next, we add interaction factors to our model. Table 4 presents the coefficients of the top 12 predictive variables ordered by a) beta coefficient and b)  $t$ -value. Trade, cultural dimension of Masculinity vs. Femininity (MAS), and gravity model and its combinations show the highest significant coefficients. Gravity model alone, as well as in combination with the economic variable of trade, exports, and cultural variable of language is high on the coefficient ranking. However, by the magnitude of the coefficient (at 0.165) Trade Market Share proves to be an even better predictor. Among the cultural variables, we see MAS to have the highest coefficient, followed by intolerance with a negative coefficient at  $-0.054$ . Language in combination with the gravity model proves to be a more significant predictor than language alone.

If we consider the  $t$ -values, which signify a variable’s importance in the presence of other variables (the left column of Table 4), we find three significant combinations of cultural attributes. The most significant is the interaction between intolerance and the cultural dimension of Long-Term Orientation (LTO) ( $t$ -value = 3.66) and intolerance and Uncertainty Avoidance (UAI) ( $t$ -value = 2.83). The dimensions of LTO and UAI are linked to tradition, nationalism, and the fear of the unknown [14]. For example, studies show that in Japan (ranking high in LTO and UAI), studies show that people avoid communication with non-Japanese for fear of failing to understand and interact with strangers from different

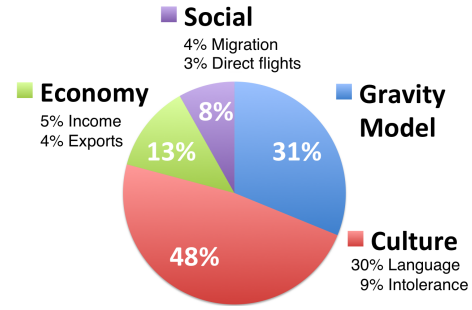


Figure 6: The predictive power of the four dimensions with three most important variables. A dimension’s weight is computed by summing the absolute values of the coefficients belonging to it.

cultures which reflects the way of how strangers are treated [8]. UAI and LTO combined with the intolerance variable, although not explicitly studied by Hofstede, shows that they are indeed related. Also, language, in combination with Masculinity vs Femininity (MAS) ( $t$ -value = 2.57) is more significant than language alone, suggesting its importance in the cultural domain.

The most prominent economic factor is trade market share of total exports – the share of the world’s import that is covered by the two countries’ exports – with a coefficient of 0.165 – eclipses the direct measure of income groups (at  $-0.03$  not included in the top 12 predictive features), showing trade to be a better indicator of communication than per-capita GDP. Trade agreements are organized over various historic events and through geo-political considerations, thus it is interesting to see them play such an important role in determining everyday online communication. A connection between political climate and communication would be an enticing potential future direction of this research.

Finally, the importance of language (here, considered a cultural feature) is mitigated when we add the interactions, with trade and gravity model playing a more important role than language. This is interesting if we refer to recent studies on cooperative work in software [33] where it was found that English is almost always the working language of such communities, even if their mother tongue is not English and hence reducing the importance of common language other than English.

Figure 7 compares the model’s prediction to communication volume. The figure shows a higher accuracy at high communication volumes with worse performance as the communication decreases. In the next section we discuss these results and look at some of the most difficult to predict cases. Finally we look at practical significance of the findings and its limitations.

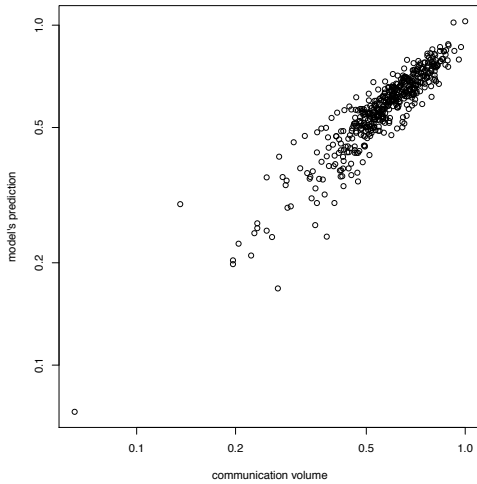
## DISCUSSION

Is distance dead? We show that no, it is still predictive of international communication in Twitter, but cultural and socio-economic factors, especially that of language, also play an important role. Linguistic and physical separation has been



Variable	$\beta$	t-value	p-value	Variable	$\beta$	t-value	p-value
Trade Market Share	0.165	3.90	***	Gravity Model	0.072	7.17	***
Exports	-0.151	-1.48		Gravity Model x Trade Market Share	-0.067	-4.67	***
Exports x Language	-0.110	-1.45		Trade Market Share	0.165	3.90	***
MAS	-0.102	-2.76	**	Gravity Model x Language	0.060	3.70	***
Gravity Model x Exports	0.098	2.73	**	Intolerance x LTO	0.022	3.66	***
Gravity Model	0.072	7.17	***	Migration x PDI	0.041	3.24	**
Language	-0.070	-1.70	.	Trade Market Share x Exports	0.016	2.95	**
Gravity Model x Trade Market Share	-0.067	-4.67	***	Gravity Model x MAS	0.031	2.93	**
PDI	0.061	1.63		Intolerance x UAI	0.023	2.83	**
Gravity Model x Language	0.060	3.70	***	MAS	-0.102	-2.76	**
Intolerance	-0.054	-2.11	*	Gravity Model x Exports	0.098	2.73	**
Income group x Migration Rate	-0.051	-2.41	*	Language x MAS	0.042	2.57	*

**Table 4: The top 12 predictive variables in the final model (including interaction factors) ordered by beta coefficients (columns 1-4) and *t*-value (columns 6-9). The gravity model was calculated by using internet penetration as a proxy for population. Significance: \*\*\*  $p < 0.0001$ , \*\*  $p < 0.001$ , \*  $p < 0.01$ , .  $p < 0.05$ .**



**Figure 7: Observed unique mention volume versus the model’s predictions.**

considered a major obstacle in international communication and collaboration. In 2000 Olson *et al.* [27] argued that distance impacts the effectiveness of collaborative work, with language, trust and cultural differences endanger the quality of project results, despite technological enablement of international communication. However, we show that the language barrier is strongest in combination with cultural factors dealing with intolerance and the fear of the unfamiliar. Finding a common culture, thus, may present a way of overcoming language barriers. For instance, a recent study by Takhteyev *et al.* [33] describes successful international collaborative projects in Open Source software development. This takes place, authors argue, when contributors follow an agreed “common” culture and communicate mostly in English. For example, *Lua*, a programming language developed in Brazil and used in the development of several well-known projects such as *World of Warcraft* and *Angry Birds*, was adapted by the global collaborative circles, such that the manuals were in English rather than in Portuguese, fostering widespread international partnerships. To improve col-

laboration in a culturally-diverse setting, Kittur *et al.* [18] propose several strategies, including observing behavior of other workers, electing leaders, and passing knowledge to others. Figure 4 (b) shows that many of the strongest ties lie between countries with different native language, such as United States and Japan, Indonesia and South Korea, Spain and United Kingdom, which, although geographically remote, may be connected by common cultural attributes. Thus, our findings show that finding a common culture could be an important barrier which software designers, in particular those who mean to enable international conversation and collaboration, must overcome.

Similarly, State *et al.* [32], find the concept of *civilization* – countries that share the same religion and continent – having a strong positive effect on the private email exchanges. If one ranks the significant coefficients of their model, one finds Colonial Link and Huntington’s Civilization as the top and third most important predictors, respectively. Language, physical distance, and population size also appear at the top (as second, fourth, and fifth in the ranking). For Twitter communication, we show that even though people can subscribe to the majority of users without authorization or reciprocity (unlike the definition of links in [32]), active interactions (through mentions) are still aligned along culture and physical distance.

### Practical Implications

Our findings have several implications for the design of social media software. First, we find language and culture to be substantial barriers to Twitter communication. It is noted by Nardi *et al.* [26] that people have to adjust to the technology from other cultures. For instance, Japanese have adapted their writing style to the horizontal typewriter-style word processors, and spelling of words in languages having letters not included in the standard English keyboard has been adjusted accordingly. In the communication network shown in Figure 4, the strongest links are with the United States – the country in which Twitter originates – and the culture of which would drive the design of the software. However, Twitter is already making an effort to diversify its service to embrace

non-English languages by providing support for a variety of character sets and automatic translation<sup>8</sup>.

One of the major cultural factors we found impacting communication is intolerance, implying that the users of countries associated with higher intolerance would be less likely to communicate with other nations. As discussed by Borning & Muller [4], designers must not assume that some cultural views and values, including those on gender, age, and speech, are universally held [9]. Some of these values could be learned by gathering user behavior data using built-in tools and interaction logs, combined with the cultural attributes extrapolated from user location information and other personalized data, providing potentially better culture-aware services.

However, using a user's cultural preferences, we may want to instead enrich her experience and broaden the reach of information. Including economic, cultural, and social factors, we plan to enhance the recommendations presented to the user. In particular, we'll consider the task of finding people that will most likely re-tweet posts and reach larger audiences [39] not only locally, but internationally.

### Limitations

Although we attempt to reduce the effect of multicollinearity in our model and exclude some of the variables, it is impossible to find a complete, and yet altogether independent set of real-world variables. Further, a different selection of country-specific variables would likely somewhat change the observed results. For example, the importance of trade market share may imply that certain economic and trade policies may also be important, as well as official social policies and visa requirements between countries. Also, due to unavailability of some of the predictive variables, the data set for regression was quite limited compared to original data. The missing values are a source of potential selection bias. This happens because the data extracted from the World Bank and CIA does not distinguish between countries that do not report their trade statistics and country pairs with no bilateral trade (resulting in zero values). A more complete dataset would expand the scope and accuracy of such analysis.

Even though Twitter may not be a representative sample of the world's population, our study shows barriers even among the relatively more well-off, technically savvy communities. As a new kind of light-weight, public communication, it is a platform which encourages weak ties between its users. More longitudinal studies are needed to determine whether the development and change of these barriers can be detected in other, more personal or established, means of communication like e-mail and texting.

Finally, we have not considered in our regression inputs that capture how automative or repressive a countries's regime is, nor the imposition of internet censorship<sup>9</sup>.

<sup>8</sup><http://gigaom.com/2013/07/03/twitter-enables-translation-on-tweets-from-high-profile-users-amid-egyptian-turmoil/>

<sup>9</sup>Twitter was banned in Iran and China before 2011

### CONCLUSION

Will distance, economic and social constraints impact communication online forever? Or will eventually ubiquitous internet access, new social platforms and globalization open the door for unrestricted communication between countries despite their economic and social differences? The awareness of these boundaries is prerequisite in our understanding of the kinds of information residents of these countries are likely to consume, and of the constraints on the world-wide information propagation.

In this paper we study online communication and attention as measured in Twitter mentions – a platform which provides an easy way for its users to maintain “weak social ties”. We find that, similarly to more personal communication, these weak ties are best predicted using both distance, as well as cultural and socio-economic factors.

We are continuing this line of investigation. In this study, we have unveiled the dynamics of flow between countries, but have not considered directionality and the focus of international attention. Of particular interest would be a deeper study of the language used among residents of different countries as well as the topics discussed in their interactions. By tracking these topics over time, we could detect major shifts in public attention and opinion, especially around crises or other major events.

We are also looking to invert the focus of this research, and attempt to predict socio-economic factors of populated areas, such as city districts, via their online communication. For this, we will attempt to increase the international coverage of our data by including sites like Weibo (an equivalent of Twitter and China) in our dataset.

### ACKNOWLEDGMENTS

This research is partially supported by European Community's Seventh Framework Programme FP7/2007-2013 under the ARCOMEM and Social Sensor projects, by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037 “Social Media”, and by Grant TIN2009-14560-C03-01 of the Ministry of Science and Innovation of Spain.

### REFERENCES

1. Barrat, A., Barthelemy, M., Pastor-Satorras, R., and Vespignani, A. The architecture of complex weighted networks. *Proceedings of the National Academy of Sciences of the United States of America* 101, 11 (2004), 3747–3752.
2. Beevolve. An exhaustive study of twitter users across the world. <http://www.beevolve.com/twitter-statistics/>, January 2012.
3. Berggren, N., and Nilsson, T. Does economic freedom foster tolerance? Working Paper Series 918, Research Institute of Industrial Economics, 2012.
4. Borning, A., and Muller, M. Next steps for value sensitive design. In *Proceedings of ACM annual conference on Human Factors in Computing Systems*, (2012), 1125–1134.

5. Hutchinson, W. K. Does ease of communication increase trade? commonality of language and bilateral trade. *Vanderbilt University Department of Economics Working Papers 0217*, June 2002.
6. Brun, J.-F., Carrère, C., Guillaumont, P., and De Melo, J. Has distance died? evidence from a panel gravity model. *The World Bank Economic Review* 19, 1 (2005), 99–120.
7. Cairncross, F. The death of distance: How the communications revolution is changing our lives. Harvard Business Press, 2001.
8. Duronto, P. M., and Nakayama, S. Japanese communication. avoidance, anxiety, and uncertainty during initial encounters. *Journal of East Asian Affairs* (2005), 22–49.
9. Ess, C., and Sudweeks, F. On the edge: Cultural barriers and catalysts to it diffusion among remote and marginalized communities. *New Media & Society* 3, 3 (2001), 259–269.
10. García-Gavilanes, R., Quercia, D. and Jaimes, A. Cultural Dimensions in Twitter: Time, Individualism and Power. In *Proceedings of the The 7th International AAAI Conference on WebLogs and Social Media (ICWSM)* (2013).
11. Gelman, A., and Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*, 1 ed. Cambridge University Press, Dec. 2006.
12. Gilbert, E., and Karahalios, K. Predicting tie strength with social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)* (2009).
13. Granovetter, M. S. The strength of weak ties. *American journal of sociology* (1973), 1360–1380.
14. Hofstede, G. Culture's consequences: International differences in work-related values, vol. 5. *Sage Publications, Incorporated*, 1980.
15. Hofstede, G., Hofstede, G. J., and Minkov, M. *Cultures and Organizations: Software Of The Mind*. McGraw-Hill USA, 2010.
16. Huntington, S. P. The clash of civilizations? *Foreign affairs* (1993), 22–49.
17. Jung, W.-S., Wang, F., and Stanley, H. E. Gravity Model in the Korean Highway. *EPL (Europhysics Letters)* 81, 4 (2008).
18. Kittur, A., Nickerson, J. V., Bernstein, M., Gerber, E., Shaw, A., Zimmerman, J., Lease, M., and Horton, J. The future of crowd work. In *ACM Proceedings of the 2013 conference on Computer supported cooperative work (CSCW)* (2013), 1301–1318.
19. Kleinberg, J. The convergence of social and technological networks. *Communications of the ACM* 51, 11 (2008), 66–72.
20. Krings, G., Calabrese, F., Ratti, C., and Blondel, V. D. Urban gravity: a model for inter-city telecommunication flows. *Journal of Statistical Mechanics: Theory and Experiment* (2009).
21. Lauterbach, D., Truong, H., Shah, T., and Adamic, L. Surfing a web of trust: Reputation and reciprocity on couchsurfing. com. In *the 12th IEEE International Conference on Computational Science and Engineering (CSE)* (2009).
22. Lee, K., Eoff, B., and Caverlee, J. Seven months with the devils: A long-term study of content polluters on twitter. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).
23. Leskovec, J., and Horvitz, E. Planetary-scale views on a large instant-messaging network. In *Proceedings of the ACM 17th international conference on World Wide Web*, (2008).
24. Meier, J. P. Do "Liberation Technologies" Change the Balance of Power Between Repressive States and Civil Society? PhD thesis, The Fletcher School of Law and Diplomacy, 2011.
25. Mok, D., Wellman, B., and Carrasco, J. Does distance matter in the age of the internet? *Urban Studies* 47, 13 (2010), 2747–2783.
26. Nardi, B. A., Vatrpu, R. K., and Clemmensen, T. Comparative informatics. *Interactions* 18, 2 (2011), 28–33.
27. Olson, G. M., and Olson, J. S. Distance matters. *Human-Computer Interaction* (2000), 139–179.
28. Scellato, S., Mascolo, C., Musolesi, M., and Latora, V. Distance matters: geo-social metrics for online social networks. In *the Proceedings of the 3rd Conference on Online Social Networks* (2010).
29. Scellato, S., Noulas, A., Lambiotte, R., and Mascolo, C. Socio-spatial properties of online location-based social networks. In *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media* (2011).
30. Smith, C., Quercia, D., and Capra, L. Anti-gravity underground? In *the 2nd Workshop on Pervasive Urban Applications (PURBA)* (2012).
31. Smith, C., Quercia, D., and Capra, L. Finger on the pulse: identifying deprivation using transit flow analysis. In *Proceedings of the ACM 16th Conference on Computer Supported Cooperative Work* (2013).
32. State, B., Park, P., Weber, I., Mejova, Y., and Macy, M. The mesh of civilizations and international email flows. Tech. Rep. No. arXiv: 1303.0045, 2013.
33. Takhteyev, Y. *Coding Places: Software Practice in a South American City*. The MIT Press, 2012.
34. Takhteyev, Y., Gruz, A., and Wellman, B. Geography of twitter networks. *Social Networks* 34, 1 (2012), 73–81.

35. Thomas, K., Grier, C., Song, D., and Paxson, V. Suspended accounts in retrospect: an analysis of twitter spam. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference* ( 2011).
36. Travers, J., and Milgram, S. An experimental study of the small world problem. *Sociometry* (1969), 425–443.
37. University, T. P. S. Stat 501 - regression methods, 2013.
38. Venkataraman, M., Subbalakshmi, K. P., and Chandramouli, R. Measuring and quantifying the silent majority on the internet. In *the 35th IEEE Sarnoff Symposium* (SARNOFF) (2012).
39. Wang, B., Wang, C., Bu, J., Chen, C., Zhang, W. V., Cai, D., and He, X. Whom to mention: expand the diffusion of tweets by @ recommendation on micro-blogging systems. In *Proceedings of the 22nd ACM International Conference on World Wide Web* (2013).
40. Wilkinson, R., and Pickett, K. The Spirit Level: Why Equality is Better for Everyone. *Penguin Books*, 2010.
41. Zipf, G. K. The p 1 p 2/d hypothesis: on the intercity movement of persons. *American sociological review* 11, 6 (1946).