Geographic Dissection of the Twitter Network

Juhi Kulshrestha **Farshad Kooti** **Ashkan Nikravesh** **Krishna P. Gummadi**

Max Planck Institute for Software Systems (MPI-SWS), Germany

Abstract

Geography plays an important role in shaping societal interactions in the offline world. However, as more and more social interactions occur online via social networking sites like Twitter and Facebook, users can interact with others unconstrained by their geolocations, raising the question: *does offline geography still matter in online social networks?* In this paper, we attempt to address this question by dissecting the Twitter social network based on users' geolocations and investigating how users' geolocation impacts their participation in Twitter, including their connections to others and the information they exchange with them. Our in-depth analysis reveals that geography continues to have a significant impact on user interactions in the Twitter social network. The influence of geography could be potentially explained by the shared national, linguistic, and cultural backgrounds of users from the same geographic neighborhood.

Introduction

A person's geographical location crucially affects her social connections and interactions in the offline world. People in close geographic proximity have a much higher chance of coming in contact with one another than those who are farther away and so geography plays an important role in shaping social interactions in the real-world. However, as people increasingly adopt online social networking services, interactions become unconstrained by geographic distances, raising the question: *does offline geography still matter in online social networks?*

In this paper, we attempt to address this question by conducting a careful and detailed geographic dissection of the popular Twitter social network. More specifically, we first inferred the geographic locations of over 12 million Twitter users in the dataset described in (Cha et al. 2010) and then analyzed how the users' geolocation affects their participation in the Twitter network, including who they connect to and exchange information with. Our analysis reveals several interesting ways in which geography affects user participation and we highlight a few key findings below.

First, we examined the geographical distribution of Twitter users across different countries. We find a geography-based digital divide, where a small number of countries not only account for a large share of the total user population, but also for an even larger share of "elite" Twitter users—the most active and influential Twitter users. Second, we investigated how users' geolocations affect their social connections. We find that even as users preferentially connect with other users within their own country, more than a third of all social connections are transnational (i.e., they cross national boundaries). Further examination of transnational links shows that users tend to preferentially connect with users in other countries with whom they share geographical or linguistic proximity. Third, analyzing information trade between different countries, we find that more than a third of all tweets are exchanged across national boundaries. Most countries run substantial deficits, consuming more tweets than they produce, and their high deficits are counterbalanced by a small group of countries led by the US, where users run a huge surplus of tweets.

In summary, our findings indicate that offline geography still holds considerable influence over online social interactions. One potential explanation is that even as geographic distances do not matter for communication in the online world, people from the same geographic neighborhood in the offline world tend to share similar national, political, linguistic, and cultural backgrounds, which in turn facilitate greater communication between them. Our findings have potential applications in predicting or recommending social connections for a user as well as in understanding information diffusion over the Twitter social network.

Related work

There is a growing interest amongst researchers to understand how offline boundaries (e.g., geographic, linguistic, national, and cultural boundaries) impact users' interactions in the online world. Some recent studies have analyzed the geographic distribution of Twitter users, albeit on small datasets consisting of tens of thousands of users. (Java et al. 2007) and (Krishnamurthy, Gill, and Arlitt 2008) examined and discovered differences between the properties and growth of the networks of Twitter users in different geographic regions like North America, Europe, South America and Asia-Pacific. More recently, (Takhteyev, Gruzd, and

Wellman 2011) found that geographic distances, national boundaries, and languages hold considerable influence on the formation of social ties on Twitter. (Hong, Convertino, and Chi 2010) studied the differences in usage patterns between different language communities in Twitter. Similar to these prior studies, our current work shows that both linguistic similarity and geographical proximity play a significant role in shaping the users' online interactions. Compared to these previous studies, our current work presents a considerably more detailed study of how geolocations of users impact their participation, connectivity and information exchange with other users, using a significantly larger dataset containing tens of millions of users.

A number of techniques have been explored to infer geolocations of Twitter users. (Hecht et al. 2011) use map APIs to resolve location data provided by the users as part of their profile information. Others have tried to predict the location of users who do not provide their profile information, either based on the location of the users' neighbors in the social graph (Sadilek, Kautz, and Bigham 2012) (Backstrom, Sun, and Marlow 2010), or based on the content of their tweets (Cheng, Caverlee, and Lee 2010). In this work, we rely only on the profile information provided by the users themselves as it is sufficient to infer a considerable fraction of all Twitter users in our dataset.

Understanding how the geolocation of users affects their online behavior has applications in predicting link formation, designing search and recommendation systems for finding local experts and authorities, and studying diffusion of information in the social network. For example, (Liben-Nowell et al. 2005) constructed a model for predicting friendship link formation based on the observation that the probability of forming friendship links is inversely proportional to the geographic proximity and to the number of people who are geographically closer. Similarly, (Toole, Cha, and Gonzalez 2011) used user geolocation and exposure to mass media to develop a model of adoption of innovations on social networks. The findings of our current work have many potential applications as well. However, exploring any specific application of our findings is out of the scope of this work.

Dataset and Methodology

In this section, we first describe the Twitter dataset we used in this study and then discuss the methodology that we used to infer the geographical locations of users.

Twitter dataset: We used the Twitter dataset described in (Cha et al. 2010). The dataset includes the profile information of 51.9 million user accounts and their 1.9 billion follow links, based on the snapshot of the network taken in September 2009. The dataset also contains the 1.7 billion public tweets posted by these users from the launch of Twitter in March 2006 till September 2009.

Inferring users' geolocations: In this study, we focus on inferring location information for Twitter users at the granularity of countries. To this end, we use information from two of their profile fields: the location field and the time zone. The location field is a free-text string entered by the user, while the timezone field is a selection made from a drop-down menu. The timezone entries consist of a location name alongside a UTC offset, which can be used to determine the user's country of residence.

	Bing & Yahoo	Yahoo & time zone	Bing & time zone	At least 2
Overlap	10.58 M	12.24 M	10.19 M	12.86 M
Match	9.78 M (92.4%)	10.85 M (88.7%)	8.99 M (88.2%)	12.22 M (94.5%)

Table 1: *Match between the different sources for location resolution*

Out of the total 51.9 million users, $13,148,002$ (25.3%) users filled in the location field. For these users, we used public APIs provided by both Yahoo Maps[1] and Bing Maps[2] to resolve the free-text string entered by the users into country locations. Out of these 13 million, we were able to resolve the locations of $10,709,638$ (81.5%) users using Bing Maps, and $12,908,671$ (98%) users using Yahoo Maps. Also, out of the total 51.9 million users, $19,365,683$ (37.3%) users provided their time-zone information, which we resolved to the corresponding country.

Previous studies have suggested that location inference using individual map APIs can be error prone (Hecht et al. 2011). So we compared the results obtained using the two map APIs and the timezone, in order to minimize inference errors. Table 1 shows the number of users that were common between the sets of users whose location information was successfully resolved using each of these three sources. We also show the fraction of these overlapping users for whom the inferred locations matched. We find a high agreement in the resolved country name between any two of the three sources.

For our study here, we only considered the set of users for whom the resolved location matched for at least two out of the three sources. The number of such users is $12,220,719$, which accounts for 23.5% of all users in our dataset. These users are distributed across 231 countries and they account for 73.65% of all tweets posted and 37.6% of all social links in the network.

Limitations: Our inference methodology may be biased by the fact that users in different countries might not have the same probability of sharing their location information. In this case, the 23.5% of users for whom we inferred location information might not be a representative sample of the total Twitter user population. Yet another potential source of bias is the fact that our dataset is over two-years old. So some of the analysis results presented here (e.g., the top-10 countries with the most Twitter users) might not accurately reflect the current Twitter network.

Finally, for inferring users' locations we are relying on the users themselves to provide correct location information. (Hecht et al. 2011) showed that 19.5% of the users either entered non-geographic information as the location string in Twitter or the map APIs do not always return correct results. To investigate the effect of these two sources of error on our

[1] http://developer.yahoo.com/geo/placefinder/

[2] http://www.microsoft.com/maps/developers/web.aspx

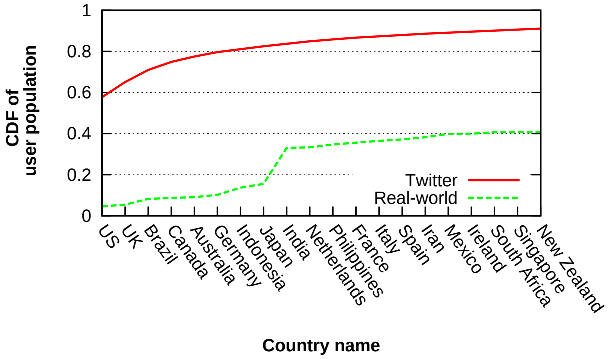

Figure 1: *CDF of Twitter population and world population of 20 countries with the most users.*

location inference, we took a random sample of 1000 users from our final set of 12 million users and manually examined this random set. We looked at the timezone and the location string entered by the user and judged whether these were correctly resolved to the corresponding country. We determined the inference to be correct in 94.7% of cases. In the remaining 5.3% erroneous cases, 4.4% of the users had entered non-geographic location, while for the remaining 0.9% of the users, the country was incorrectly resolved by the map APIs. These numbers give us an estimate of errors introduced due to the unreliability of user-provided location strings and the map APIs. Our error estimates are considerably reduced from the 19.5% reported in (Hecht et al. 2011) due to our requirement that at least two of the three sources (Bing maps, Yahoo maps, and the timezone information) resolve the users' location to the same country.

User populations

In this section, we analyze the adoption of Twitter in different geographical locations around the world. Our analysis is driven by the following three high-level questions:

1. How are Twitter users spread across the world?
2. Is the adoption of Twitter in a country related to the socio-economic status of the country's population?
3. How are the *elite* Twitter users distributed across different countries?

Geo-distribution of Twitter users

The 12 million Twitter users in our dataset, for whom we successfully inferred location information, are spread across 231 countries world-wide. The number of Twitter users varies considerably across the different countries, with only a small number (13) of countries with 100,000 or more users, while a large number (167) of countries have 10,000 or fewer users. Not surprisingly, the top few countries account for a vast majority of the total Twitter population.

In Figure 1, we show the skew in Twitter populations towards a few countries, by plotting the cumulative distribution function (CDF) of the Twitter users from the 20 countries with the most Twitter users. The US, the country with the highest number of users, by itself accounts for 57.7% of the total Twitter population in our dataset. The top 10 countries alone account for 84.9% of the whole Twitter population, while the bottom 80% of countries only account for 2.3%.

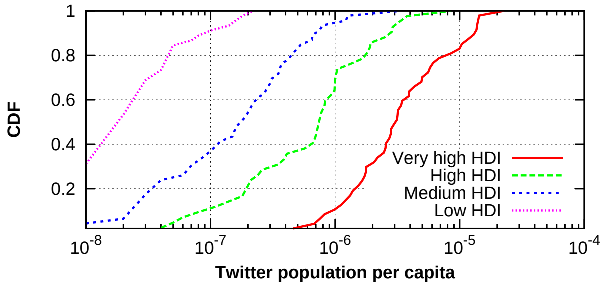

Figure 2: *Distribution of per capita Twitter population of countries grouped by their HDI.*

Interestingly, the top countries account for a significantly higher fraction of the Twitter population than their share of the world population[3] living in those countries (also shown in Figure 1). The difference between the two curves exemplifies the geography-based digital divide[4] in today's world, where users outside of a small number of developed and developing countries have limited reach to online services like Twitter.

Correlation with socio-economic status

Next we studied the correlation between rankings of countries based on per capita Twitter user population and their Human Development Index (HDI)[5], which is a comparative statistic that is based on several factors such as literacy and standards of living.

Figure 2 plots the distribution of per capita Twitter population for countries in four well-recognized categories of HDI, namely, very high human development, high human development, medium human development, and low human development. The figure shows a large difference in per capita Twitter population (Twitter adoption rate) between the four different groups, which suggests a high correlation between HDI and per capita Twitter user population.

So far our analysis of geolocations of Twitter users has been primarily limited to a snapshot of the population in 2009. As Twitter adoption grows world-wide, one would expect the adoption rates to change over time. We analyzed the temporal evolution of Twitter user population[6] by studying several snapshots of the network during the time period from 2006 to 2009. While the number of users in each country increased considerably during this period, our observations about the skew in Twitter service adoption towards a small number of countries hold true at all times.

[3] http://en.wikipedia.org/wiki/List_of_countries_by_population

[4] http://bit.ly/Digital_Divide_Report_World_Economic_Forum

[5] http://en.wikipedia.org/wiki/Human_Development_Index

[6] For more details and visualizations, please visit: http://twitter-app.mpi-sws.org/geo-dissection

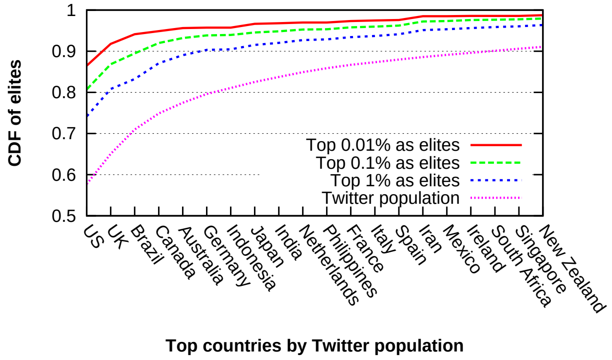

Figure 3: *CDF of number of elites in the 20 countries with the most users.*

Geo-distribution of elite Twitter users

Not all users in the Twitter network are equal. Studies have shown that a small number of Twitter users, *elites*, account for a disproportionately large number of followers and tweets consumed on Twitter; for example (Wu et al. 2011), have shown that roughly 50% of URLs consumed are generated by just 20K elite users (0.05% of all users). Such influential users in the network can be detected using ranking methods such as PageRank or FollowerRank (Kwak et al. 2010). We now focus our attention on the distribution of elite Twitter users across different countries.

The distribution of elite users across the countries is even more skewed than the distribution of Twitter users themselves. Figure 3 plots this bias in the geolocation of elites. For example, if we consider the top 0.1% of users with highest PageRank, then 80.7% of them are in the US, which is much higher than its 57.7% share of the total Twitter population. The ten countries with the most users account for more than 95% of the top 0.1% elites, even though they represent only 85% of the user population. Our results show that the digital divide is even larger amongst the elite users. They also suggest that when building location specific search or recommendation services, global ranking algorithms might not be sufficient as they would ignore local elites—we also need a local ranking scheme.

Network links

In this section, we shift our focus to the social links between Twitter users and investigate how the geolocations of users impact who they follow and who follows them. Specifically, we attempt to answer the following three questions:

1. How important are transnational links? What fraction of follower or following links cross national boundaries? Does this fraction vary from one country to another?
2. Do users preferentially receive followers or follow others from their own country?
3. Do geographical, linguistic, or cultural proximity have an impact on social links between users in different countries? Can we cluster countries based on interconnections between their user populations?

Country	% of Trans-national Followings	% of Intra-national Followings	Twitter Population Share
India	82.28%	17.72%	1.21%
Canada	79.84%	20.16%	3.91%
Australia	78.57%	21.43%	2.62%
Indonesia	73.19%	26.81%	1.46%
UK	69.79%	30.21%	7.33%
Netherlands	62.42%	37.58%	1.16%
Germany	62.26%	37.74%	2.12%
Brazil	32.9%	67.1%	5.9%
Japan	26.41%	73.59%	1.45%
US	18.44%	81.56%	57.74%

Table 2: *Fraction of trans- and intra-national following links for the 10 countries with the most users, ranked by their fraction of transnational followings*

For our analysis in the rest of the paper, we only consider the 100 countries with the most users. The remaining countries have too few users (less than 1000) in our dataset to extract meaningful and representative information. Out of these 100 countries, we excluded 9 countries (Bahamas, Bosnia and Herzegovina, Costa Rica, Cyprus, Iceland, Iran, Israel, Jordan, and Switzerland) as accounts from these countries exhibited spammer-like excessive connectivity with users around the world.

Transnational vs. intra-national links

In Twitter, 35.15% of all social links are transnational, i.e., they connect a follower and a followee that are located in different countries. The percentage increases to 37% when we exclude the US, which accounts for a majority of users and links in the Twitter network. Thus, even as a majority of social links stay within national boundaries, a considerable fraction (more than a third) of all links cross national boundaries, highlighting the global nature of connections in the Twitter social network.

However, the fraction of transnational links varies considerably from country to country. For the 10 countries with the largest Twitter user population, Table 2 shows the fraction of their transnational and intra-national following links along with their share of the Twitter populations. There are two interesting take-aways from this table. First, even amongst the top-10 countries, the fraction of trans-national links varies from as high as 82% in some countries to as low as 18% in others, suggesting that users in some countries seek information from around the world, while those in others look for information primarily from their compatriots. In the former category, we have countries like India, Australia, Canada, Indonesia, and the UK, with more than two-thirds of their following links going to users in other nations. At the other end, users in the US, Japan, and Brazil have more than two-thirds of their links remaining within their national boundaries. Netherlands and Germany lie in the middle with a more even division between national and transnational links. Thus, users in some countries are much more globally connected than others.

Second, comparing the fraction of intra-national links for countries with their share of the Twitter population, we ob-

Country	Closest 5 Followers	Closest 5 Followings
Chile	Argentina, Bolivia, Ecuador, Peru, Uruguay	Argentina, Bolivia, Ecuador, Spain, Uruguay
Egypt	Lebanon, Morocco, Saudi Arabia, Tunisia, UAE	Kuwait, Lebanon, Morocco, Qatar, Saudi Arabia
Japan	China, Hong Kong, South Korea, Taiwan, Vietnam	China, Hong Kong, Jamaica, South Korea, Taiwan
Russia	Belarus, Greece, Latvia, Lithuania, Ukraine	Belarus, Estonia, Greece, Latvia, Ukraine
Spain	Argentina, Bolivia, Ecuador, El Salvador, Uruguay	Argentina, Bolivia, Ecuador,Mexico, Uruguay
US	Australia, Canada, Nepal, New Zealand, Pakistan	Australia, Canada, New Zealand, Singapore, UK

Table 3: *Closest 5 follower and following countries for a few example countries around the world*

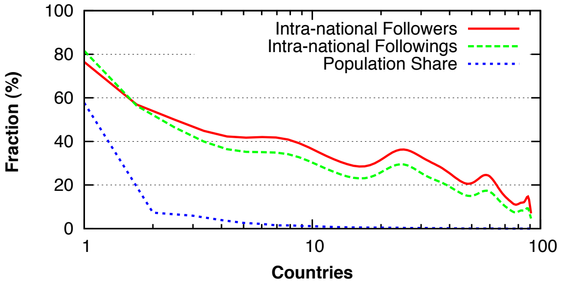

Figure 4: *Fraction of intra-national followers and followings in comparison to the Twitter user population share in different countries, ranked by their Twitter user population.*

serve that there is a significant bias towards following other users from the same country. For example, 37.74% of all users followed by German users are from within Germany itself, even though German users account for only 2.12% of the total Twitter population, which suggests that German users prefer to follow other German users almost 18 times more than users elsewhere. Figure 4 plots the fraction of intra-national followers and followings for the different countries in our dataset along with their share of user populations. The figure shows a clear bias towards intra-national links for users in all the countries. The ratio of the percentage of intra-national links to the percentage of user populations is very high across the different countries; average ratio across all countries for following links is 1085 and for follower links is 756.5. Thus, even as users connect to others globally, they also exhibit significant preference for connecting to local users.

Figure 4 also shows that for most of the countries the percentage of intra-national followers is slightly but consistently higher than intra-national followings, with the US being an exception. The higher percentage of intra-national followers suggests that there is less global demand for information from users in countries outside the US than there is demand for global information from users within those countries. This imbalance could be potentially explained by the relatively large fraction of elite users with large numbers of followers within the US. Users in other countries follow these elite users in the US, but the countries themselves contain few elites, leading to lower demand for follower links from outside of them.

Impact of geography & language

We now focus on transnational connections between pairs of countries. More specifically, we investigate whether transnational links from users in a country are preferentially directed towards other countries that are geographically or linguistically close to the country.

Type of neighbors	Closest 5 followers	Closest 5 followings
Linguistic	37.58 %	38.46 %
Geographic	55.16 %	55.16 %
Continent	74.73%	70.99 %
Ling. or Geo.	72.53 %	73.41 %
Ling. or continent	90.11 %	87.25 %

Table 4: *Percentage of closest follower and following country pairs that share a geographic boundary, or a common language, or lie within the same continent*

To conduct our analysis, for each country, we ranked all other countries based on how *closely* their users followed (or were followed by) users in the other countries. We computed the closeness of a country A with another country B based on the number of links (both followers and followings separately) that go between the countries, normalized by the number of users in country B.

Table 3 shows the top-5 closest follower and following countries for a few countries around the world. We make two observations: first, while the top-5 closest follower and following countries are not the same, there is considerable overlap between the lists. In fact, when we compared the lists of top-10 closest countries according to follower and following links there was, on average, an overlap of 75.8%. Second, for some countries, such as Japan, the closest countries correspond to geographical neighbors in east Asia, while for others, such as Spain, the closest countries are geographically distant countries in South America that share the same language. Thus, both language and geography appear to play a role in determining the connectivity between people in different countries.

We investigated the impact of geography and language by computing the percentage of top-5 closest pairs of countries that are geographical neighbors (share a border or lie within the same continent) or linguistic neighbors (share a common language). Table 4 shows the results for pairs of top-5 closest follower and following countries. The percentages for both closest follower and following countries are similar. They show that a vast majority of closest countries are geographical or linguistic neighbors: 55% of closest pairs of countries share a common border, while 38% share a common language. In fact, 73% of countries share either a boundary or language, indicating that both language and geography influence transnational social links.

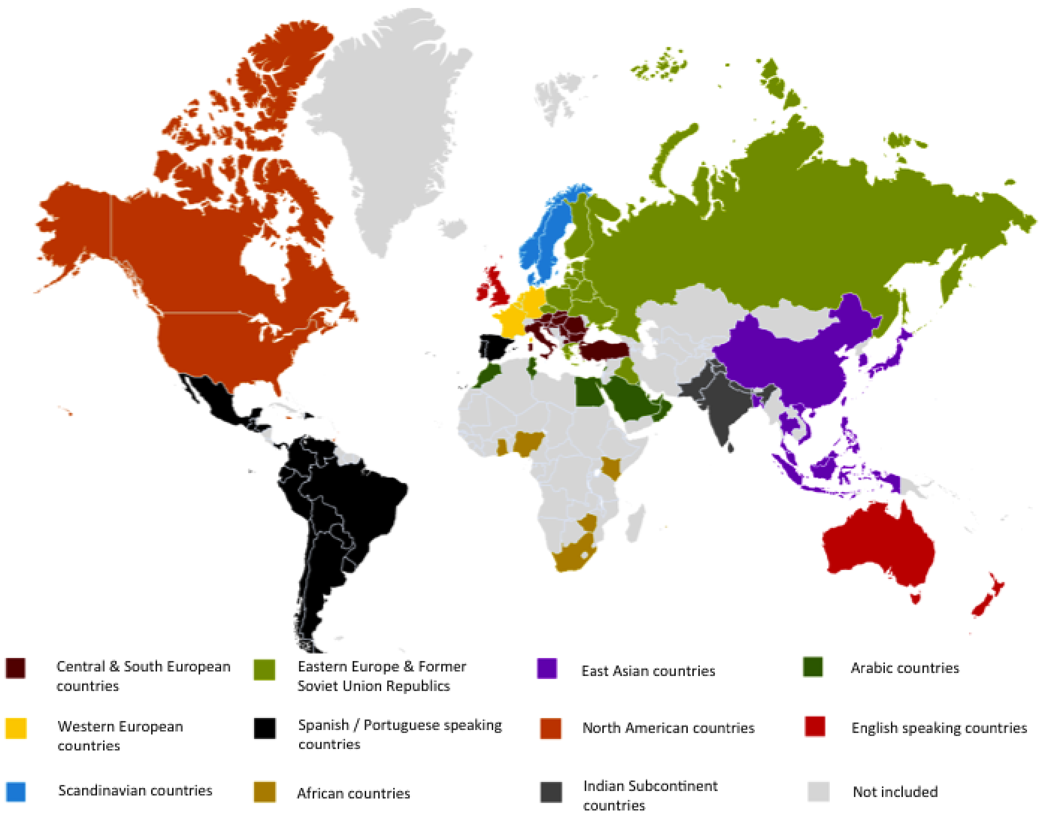
Central & South European countries
Eastern Europe & Former Soviet Union Republics
East Asian countries
Arabic countries
Western European countries
Spanish / Portuguese speaking countries
North American countries
English speaking countries
Scandinavian countries
African countries
Indian Subcontinent countries
Not included

Figure 5: *Groups of countries whose users are closely connected with one another.*

Groups of closely connected countries

We use the closeness rankings discussed in the previous section to create a friendship graph between countries, where each country is connected to its closest 5 follower or following countries. We then applied the Louvain method for community detection (Blondel et al. 2008) to detect closely interconnected groups of countries within the graphs. Figure 5 shows the country groupings resulting from the graph of closest 5 following countries on a world map. We got similar results when we used closest 5 follower countries.

The figure shows that the 91 countries in our dataset fall into eleven distinct groups of countries. These groups correspond strikingly with well-recognized geographic, linguistic, political, and cultural groupings of countries in the offline world. For example, the east Asian countries such as China, Vietnam, Thailand, South Korea, and Japan form a grouping distinct from countries in the Indian sub-continent, such as India, Pakistan, Sri Lanka, and Nepal. Similarly, Arabic speaking countries in Middle East and North Africa, such as Egypt, Tunisia, Morocco, Saudi Arabia, UAE, and Qatar form one group. Interestingly, the western European countries of Spain and Portugal are grouped with Spanish- and Portuguese-speaking countries in South and Latin America, such as Argentina, Brazil, and Mexico. Similarly, the Scandinavian countries of Sweden, Norway, and Denmark form their own group distinct from other western European countries such as France, Germany, Belgium, and Netherlands. While eastern European countries like Poland and the Czech Republic are grouped with countries that were formerly republics of Soviet Union, the central and southern European countries like Austria, Hungary, Greece, and Romania are grouped with Turkey.

The existence of these eleven distinct groupings of countries corresponding to well known national, political, linguistic, and cultural boundaries underscores the importance and influence of these offline factors on societal connections and communications in the online world.

Information trade

In this section, we investigate information exchanged or traded between users in different countries over the Twitter network. The information traded can be measured in terms of number of tweets or URLs (links to web pages) included in the tweets. We only present the results for tweets, but the results for URLs are very similar.

The tweets are being traded between the users via the social links between them. In our analysis, when a user produces a tweet, then all her followers are assumed to be consumers of that tweet.[7] So if a user, who has n followers, tweets t times, then the user effectively produces $n \cdot t$ tweets. Consequently, each of her followers consumes t tweets from her, leading again to a total consumption of $n \cdot t$ tweets

[7] Not every tweet posted might be read by each of the followers, but in the absence of any real data about what fraction of tweets are actually read, we treat all tweets received by a user as consumed by that user.

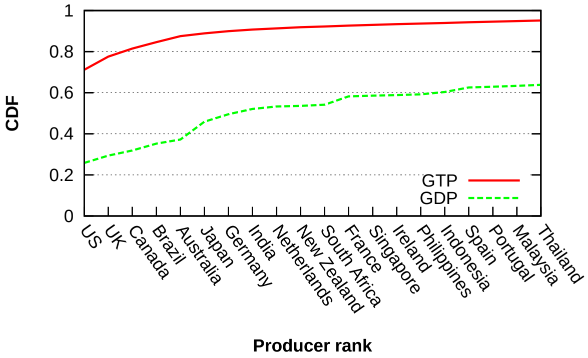

Figure 6: *GTP Share of the top 20 producer countries*

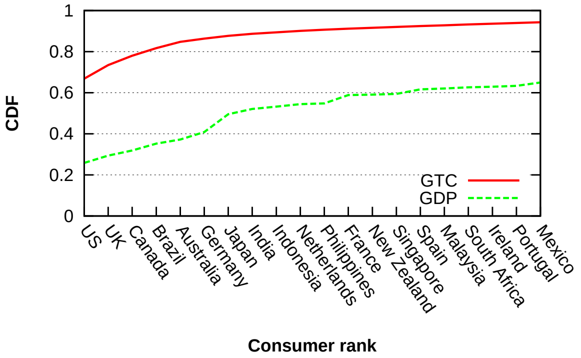

Figure 7: *GTC Share of the top 20 consumer countries*

produced by her. We don't distinguish between tweets and retweets, as less than 5% of the tweets in our dataset are retweets and therefore attributing the production to the original author will only change the results marginally.

To obtain the results in this section, we analyzed 41 million tweets that were posted in a week towards the end of our data crawl period in May 2009.

Production & Consumption

Each country on Twitter can be thought of as both a producer and a consumer of information or tweets. Inspired by the popular economic metric, Gross Domestic Product (GDP), we define two metrics for each country: Gross Tweet Production (*GTP*) and Gross Tweet Consumption (*GTC*). GTP of a country is the total number of tweets produced by all the users of that country, while GTC is the total number of tweets consumed by that country's users.

Figure 6 shows the cumulative GTP of the top 20 producer countries. We observe that the top 10 countries account for 92% of total tweets produced and that this percentage is higher than the population share of the top 10 countries with the most users (85%). The ranking of the top producing countries correlates highly (0.97 correlation coefficient) with the ranking of their Twitter population, i.e., countries with larger Twitter populations produce more tweets, as one might expect. Figure 7 similarly shows the GTC of the top 20 consumer countries. Once again, we see that the top 10 countries account for a very high percentage (90%) of total consumption. The consumption rankings correlate well (0.99 correlation coefficient) with production rankings.

We compare the percentages of GTP and GTC accounted by top countries with their share of GDP in the offline world.[8] In general, the rankings correlate well, with considerable overlap between the world's top economies and the top tweet producing and consuming countries. Figures 6 and 7 also show the cumulative GDP share of the countries in the real-world. We observe a considerably higher imbalance in GTP values compared to GDP: the US alone accounts for 25% of world's GDP, whereas it accounts for 72% of all tweets produced in Twitter. Thus, economic imbalances in the offline world are exaggerated in the online world.

Exports & Imports

In the previous section, we investigated the total information produced and consumed by different countries. In this section, we focus only on the tweets that are *imported* or *exported*—the tweets that crossed national boundaries. We found that 37.54% of produced tweets are traded between nations. In other words, roughly two thirds of tweets are consumed in the same country that they were produced in, but a non-trivial fraction (more than a third) of tweets are traded internationally.

Table 5 shows the percentage of produced tweets that are exported (and consumed tweets that are imported) for the top 10 countries with the most Twitter users. We find that the extent to which countries rely on exports and imports varies considerably across the different countries. Countries like the US, Japan, and Brazil depend on exports and imports considerably less than countries like Canada, UK, Australia, and India. Furthermore, the percentage of exports and imports match well for some countries but not for others. For example, for the UK, Canada, and Australia, the percentages match fairly well. However, for Germany and Indonesia, the percentage of imported tweets far exceeds that of exported tweets. This suggests that some countries might consume far more tweets than they export, a topic which we investigate in greater detail in the next section.

Country	% of produced tweets that are exported	% of consumed tweets that are imported
US	25.01 %	19.78 %
UK	74.38 %	74.02 %
Brazil	26.69 %	42.8 %
Canada	83.74 %	85.82 %
Australia	81.27 %	83.57 %
Germany	58.83 %	71.2 %
Indonesia	59.96 %	89.32 %
Japan	18.58 %	16.09 %
India	78.3 %	83.11 %
Netherlands	60.32 %	62.03 %

Table 5: *Fraction of exported and imported tweets for 10 countries with the most users*

[8] http://en.wikipedia.org/wiki/List_of_countries_by_GDP_(nominal)

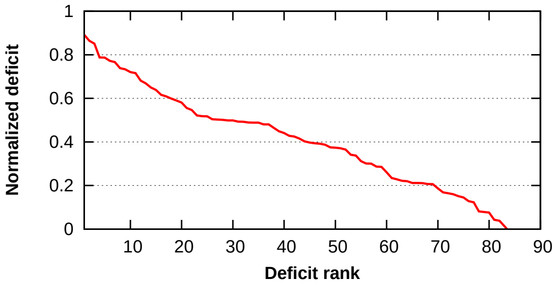

Figure 8: *Normalized deficit of countries running a tweet deficit*

Surplus & Deficits

In the previous section, we analyzed the percentage of produced tweets that are exported (and consumed tweets that are imported) for each country. We observed that some countries are exporting much more than importing and vice versa. In this section, we study the difference between the exports and imports of each country. In other words, we find which countries have a tweet surplus (i.e., number of tweets produced exceeds the number of tweets consumed) and which have a tweet deficit (i.e., the number of tweets consumed exceeds number of tweets produced). To compare the surplus and deficit numbers across different countries with varying tweet productions and consumptions, we normalize surplus tweets by the number of tweets produced and normalize deficit tweets by the number of tweets consumed.

Out of the 91 countries in our dataset, we found that only 8 countries had a tweet surplus, while the rest incurred tweet deficits. The normalized surplus for the US is 7% of all tweets produced, and the normalized surplus is lower than 15% for all the 8 countries with surplus tweets. Note, however, that since the US accounts for nearly 71% of all tweets produced, a 7% normalized surplus for the US translates to a large number of tweets.

Figure 8 shows the normalized deficit for the 83 remaining countries in our dataset. Interestingly, a large number of these countries (54) run normalized deficits that are 33% or larger. This suggests that these countries import considerably more tweets than they export. The list of these high deficit countries includes countries with large user populations like France and Germany. The high tweet deficits of these countries are largely funded by the huge surplus of tweets produced by the US. Thus, US users dominate the global information trade, producing large number of surplus tweets that in turn fund import deficits in other countries worldwide.

Conclusion

In this paper, we attempted to address the question: *does offline geography still matter in online social networks?* To this end, we dissected the Twitter social network based on users' geolocations and investigated how users' geolocations impact their participation in Twitter, their connectivity with other users, and the information they exchange with them. Our in-depth analysis reveals that geography crucially impacts all aspects of the Twitter social network. Specifically, we find that even though users preferentially connect and exchange information with other users from their own country, more than a third of all links and tweets are exchanged across national boundaries. Such transnational links and interactions occur between users in geographically and linguistically proximal countries. Our findings have potential applications in predicting or recommending social connections for a user as well as in understanding global diffusion of information.

Acknowledgements We would like to thank Chloe Kliman-Silver and Alan Mislove for sharing their Bing location resolution results with us. Juhi Kulshrestha is supported in part by the Google European Doctoral Fellowship in Social Computing.

References

Backstrom, L.; Sun, E.; and Marlow, C. 2010. Find me if you can: improving geographical prediction with social and spatial proximity. In *Proc. of the 19th int'l conference on World wide web*.

Blondel, V.; Guillaume, J.; Lambiotte, R.; and Mech, E. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech* P10008.

Cha, M.; Haddadi, H.; Benevenuto, F.; and Gummadi, K. P. 2010. Measuring user influence in twitter: The million follower fallacy. In *Proc. AAAI Int'l Conference on Weblogs and Social Media*.

Cheng, Z.; Caverlee, J.; and Lee, K. 2010. You are where you tweet : A content-based approach to geo-locating twitter users. In *Proc. of the 19th ACM int'l conference on Information and knowledge management*.

Hecht, B.; Hong, L.; Suh, B.; and Chi, E. H. 2011. Tweets from justin bieber's heart: the dynamics of the location field in user profiles. In *Proc. of the 2011 annual conference on Human factors in computing systems*.

Hong, L.; Convertino, G.; and Chi, E. H. 2010. Language matters in twitter: A large scale study. In *Proc. AAAI Int'l Conference on Weblogs and Social Media*.

Java, A.; Song, X.; Finin, T.; and Tseng, B. 2007. Why we twitter: understanding microblogging usage and communities. In *Proc. of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*.

Krishnamurthy, B.; Gill, P.; and Arlitt, M. 2008. A few chirps about twitter. In *Proc. of the first workshop on Online social networks*.

Kwak, H.; Lee, C.; Park, H.; and Moon, S. 2010. What is twitter, a social network or a news media? In *Proc. of the 19th int'l conference on World wide web*.

Liben-Nowell, D.; Novak, J.; Kumar, R.; Raghavan, P.; and Tomkins, A. 2005. Geographic routing in social networks. *Proc. of the National Academy of Sciences of the United States of America* 102(33):11623–11628.

Sadilek, A.; Kautz, H.; and Bigham, J. P. 2012. Finding your friends and following them to where you are. In *Proc. of the fifth ACM int'l conference on Web search and data mining*.

Takhteyev, Y.; Gruzd, A.; and Wellman, B. 2011. Geography of twitter networks. *Social Networks* 34(1):1–25.

Toole, J. L.; Cha, M.; and Gonzalez, M. C. 2011. Modeling the adoption of innovations in the presence of geographic and media influences. *CoRR* abs/1110.0535.

Wu, S.; Hofman, J. M.; Mason, W. A.; and Watts, D. J. 2011. Who says what to whom on Twitter. In *Proc. of the 20th int'l conference on World wide web*.