Pers Ubiquit Comput
DOI 10.1007/s00779-013-0658-y

ORIGINAL ARTICLE

Understanding the locality effect in Twitter: measurement and analysis

Ruben Cuevas · Roberto Gonzalez · Angel Cuevas · Carmen Guerrero

Received: 17 November 2012 / Accepted: 15 February 2013

Abstract Twitter is one of the most popular applications in the current Internet with more than 500 M registered users across the world. In this paper, we conduct a comprehensive analysis to understand the geographical characteristics of Twitter using cross-community mining techniques. Specifically, we study the *locality* level shown by the three main elements of Twitter, namely users, relationships and information flow. For this purpose, we rely on a dataset including the geolocation information of more than 17, 100 and 3.5 M users, relationships and tweets, respectively. Our main findings are: (1) most of the Twitter users perform their activity from an area of at most few hundred kms covering few cities within a unique country; (2) the location (i.e., country), and in particular factors such as language or Twitter popularity within a country, dictates the level of locality in the relationships of users and Twitter conversations originated in that country. The combination of these factors reveals the presence of four types of country locality profiles that we carefully analyze and compare in the paper.

Keywords Twitter · Measurements · Locality · Geographical communities · User behavior

R. Cuevas (✉) · R. Gonzalez · A. Cuevas · C. Guerrero
Universidad Carlos III de Madrid, Avda. Universidad, 30, 28911 Leganés, Madrid, Spain
e-mail: rcuevas@it.uc3m.es

R. Gonzalez
e-mail: rgonza1@it.uc3m.es

A. Cuevas
e-mail: acrumin@it.uc3m.es

C. Guerrero
e-mail: guerrero@it.uc3m.es

1 Introduction

Twitter [1] is a microblogging system created in 2006 by Jack Dorsey and Biz Stone. Twitter's users can post text messages of upto 140 characters named *tweets*. Furthermore, a given user, for example Bob, registered in the system can *follow* any other user in the system, for example Alice. We then refer to Bob as an Alice's *follower* and Alice as a Bob's *friend*. This *friend* → *follower* relationship (or link) lets Bob to visualize every *tweet* posted by Alice. Twitter has rapidly attracted a large number of users and become one of the most successful platforms for both social interactions and information diffusion. For instance, it currently counts with more than 500 M registered users of which 140 M are active and more than 340 M tweets are uploaded every day to the system [2, 3]. The great success of Twitter has attracted the research community that has recently started to investigate different aspects of Twitter [4–9].

In this paper, we focus on understanding the geographical properties of the main elements of Twitter, namely users, *friend* → *follower* relationships and information flow. Toward this end, we leverage the concept of cross-community mining (CCM) recently proposed by Guo et al. [10]. Basically, CCM consists of studying (and exploiting) interactions between communities from the *physical* and the *virtual* world. We leverage this novel concept in order to understand the impact of (well-established) geo-cultural-political (GCP) communities from the physical world in the geographical properties of a *virtual community* such as Twitter. We define a GCP community as a group of users who share common cultural (e.g., language, gastronomy, popular sports or celebrities) and political properties and are geographically close. Examples of GCP communities are a set of users within a specific region or country.

Published online: 17 April 2013

In particular, the main goal of our investigation is understanding the level of *locality* for the three aforementioned elements of Twitter. We refer to *locality* as the phenomenon that makes the activity and/or relationships of a user in Twitter to remain local within its GPC community. For instance, a user presents a high locality if she performs most of her activity from just few nearby locations. In addition, a user presents a high locality level in her relationships if her followers belong to the same GPC community and thus are located close to her. Finally, an information flow is highly localized when most of the tweets of that flow are posted by members from the same GPC community.

Understanding the effect that factors from the physical world (e.g., language or culture) have at the user, relationship and information flow levels are essential in order to depict a detailed and comprehensive model of the online behavior of Twitter users. In addition, understanding the locality properties of Twitter have a direct application in other fields such as social marketing, information diffusion modeling or design of future infrastructures to provide social media services.

In order to conduct our research, we have performed an extensive measurement study to collect the geolocation information of more than 17 M Twitter users and 250 K Twitter conversations including 3.5 M tweets that allows us to map users and tweets to a specific GPC community. Note that in most of our analysis we define a GPC community as the users within a country, since this group matches perfectly the definition of a GPC community.

Using this dataset, first, we study the locality at the user level by computing the number of locations from where a user post tweets in the system and the distribution of the user activity across these locations. Second, we study the locality at the relationship level in Twitter. For this purpose, we compute the geolocation of the origin and destination for more than 100 M *friend* $\rightarrow$ *follower* relationships in Twitter and conduct a careful comparative analysis based on the origin GPC community (i.e., country) of these relationships. Finally, we analyze the locality of more than 250 K Twitter conversations formed by an original tweet and all its associated retweets. The analysis of Twitter conversations allows us to investigate how localized is the flow of information in the system. Specifically, we study the percentage of retweets that stay local within a GPC community (i.e., country) for every conversation.

The main contributions of this paper are:

- A high performance tool to collect relevant geographical information associated with Twitter users and tweets.
- We observe that Twitter users present a significant locality level. Specifically, around 75 % of Twitter users perform their activity from a single country on an area including few ($\leq$5) cities and an even smaller ($\leq$2) number of regions. Furthermore, a Twitter user performs a non-negligible fraction of her activity from every one of its associated locations, although one of them (i.e., the main location) is significantly more used than the others.
- Factors such as language and local popularity of Twitter drive the online behavior of users within different GPC communities (i.e., countries). This behavior is defined by the observed locality at the relationship and information flow levels. Specifically, the combination of these factors allows to define four distinguishable profiles among the 14 analyzed countries:
  - *USA:* USA is the country where Twitter is (by far) more popular. Around half of the users and *friend* $\rightarrow$ *follower* links in our dataset belong to USA. This prominent presence of American users in Twitter leads to a high level of locality at both relationships and flow of information levels in USA.
  - *Local profile:* This profile includes all those non-English speaking countries where Twitter is relatively popular. These countries typically present a high locality level at both relationship and information flow levels.
  - *Shared profile:* This profile includes all those countries where Twitter is less popular among the studied ones. This low popularity leads the users within these countries to keep only one third of their relationships local and make the Twitter conversations generated in these countries to be spread in other ones. In particular, more than 50 % of the original tweets generated in France (a country belonging to this profile) are exclusively retweeted outside the country.
  - *English profile:* This profile includes all English speaking countries (apart from USA) within our dataset. The predominance of US population in Twitter produces a surprising *external locality* phenomenon in these countries. That is, the major portion of the relationships originated in these countries are destined for USA. The same reason leads to a low locality level at the information flow.
- A user's popularity (i.e., number of followers) has generally a clear impact on the locality associated with her relationships. The more popular a user is, the less localized her relationships are. However, the popularity factor is modulated by the user's GPC community (i.e., country). For instance, popularity does not affect much to the locality of relationships of users located at highly localized countries such as Brazil. Surprisingly,

popularity has a much lower impact in the locality of Twitter conversations (i.e., information flow) that seems to be mostly dictated by the GPC community (i.e., country) where the conversation was originated.

The rest of the paper is organized as follows. Section 2 describes our measurement methodology and the datasets used for the analysis. Sections 3, 4 and 5 analyze the locality in Twitter at the user, relationship and information flow levels, respectively. Section 6 briefly discusses few examples to demonstrate the usefulness of the obtained results. Finally, Sect. 7 presents the related work and Sect. 8 concludes the paper.

2 Measurement methodology and datasets

The main objective of our measurement study is to retrieve the geographical location of a large number of Twitter users and tweets in order to map them into GPC communities and analyze the locality properties of Twitter. In this section, we describe our measurement methodology and infrastructure as well as the data cleaning process used to achieve this goal. Furthermore, we present the datasets used in the analysis conducted in the rest of the paper.

2.1 Measurement methodology

Twitter provides access to the information of users and tweets through different APIs [11]. Specifically, in this study, we use the REST API and the STREAMING API. First, the REST API provides the profile information associated with a user. This information includes (among other attributes) the list of followers, the list of friends and a location tag that indicates the geographical location of the user. Moreover, the REST API allows to collect all the tweets posted by a given user. Second, the STREAMING API receives as input a given term and provides as response all the tweets including that term since the instant the query was issued. Therefore, using the STREAMING API, we are able to collect a large number of tweets. In addition, Twitter offers to its users a Tweet Geolocation Service. This service allows users to publish a tweet along with the GPS coordinates from where the tweet was posted.

Using the described tools offered by Twitter, we are able to collect meaningful data to analyze the geographical properties of the main elements of Twitter as follows:

- *User's geographical properties*: Using the REST API, we gather the tweets from a large number of users who have the Tweet Geolocation Service active. Using the GPS coordinates of these users' tweets, we can infer the geographical locations from where these users utilize the system.
- *Relationship's geographical properties*: We gather the location of a user and all its followers from their profiles location tag. Since we know the geographical location of both points of a given relationship, that is, friend and follower, we can analyze the geographical properties of that relationship. We repeat this process for a large number of users so that we can obtain meaningful conclusions.
- *Information flow's geographical properties*: Using the STREAMING API, we are able to collect original tweets with an associated geographical location. For each one of these tweets, we collect all its retweets that also have an associated geographical location. For this purpose, we rely on both the STREAMING and the REST API. With this information, we can understand the geographical properties of a flow of information formed by an original tweet and its retweets.

2.2 Measurement infrastructure

The maximum number of queries to the Twitter REST API allowed by Twitter is 350 per hour per IP address/user-id.[1] In order to speed up the data collection process, we have developed a master-slave distributed measurement architecture to query the REST API. This architecture counts with 1 master and 20 slaves located in different virtual machines on top of two physical machines. The master indicates to each slave the user-ids to be monitored. Moreover, each slave is configured with a different IP address and user-id and can then perform 350 queries per hour to the Twitter REST API. Therefore, this distributed measurement architecture lets us to perform upto 7 K queries per hour. All the information collected by the slaves is stored into a redundant centralized database.

In addition, the Twitter STREAMING API offers a best effort service, and then, in those periods in which the system is overloaded, it may provide just a subset of all the tweets associated with a given term. In order to collect a large number of popular terms and reduce the impact of the best effort service, we use 5 different virtual machines with different IP addresses to query the Twitter STREAMING API.

2.3 Data filtering

The user's location tag is an open and non-mandatory attribute in the user's profile where the user can write any text. Hence, it is not homogeneous across users (e.g., New

[1] In the past Twitter gifted whitelisted accounts which were allowed to perform up to 20 K queries per hour. Unfortunately, these whitelisted accounts are anymore available.

York can appear as NY, NYC, New York City, etc.) and non-existing or meaningless in some cases.

To address this problem, we have implemented a module in our measurement tool to filter those users who do not provide or provide a meaningless location in their location tag. Furthermore, this module uses the Yahoo geolocation API [12] to homogenize the users location. In particular, this tool provides as output the city, region/state and country associated with the input location. For instance, Yahoo geolocation API maps all those users indicating NY, NYC, New York City, etc. as their location to a unique location: New York City (city), NY (state) and USA (country).

2.4 Dataset description

Using the methodology and data filtering described above, we have generated the following datasets that constitute the basis for the analysis conducted in the rest of the paper:

- *Relationships dataset*: We have crawled the profile of 2 M Twitter users randomly selected from [8]. After filtering and homogenizing the data, the final dataset includes a total of 973 K geolocated friends, 16.5 M geolocated followers for those friends and more than 100 M *friend* → *follower* relationships.
- *Users location dataset*: This dataset is formed by 140 K users from the relationship dataset that have the Tweet Geolocation Service active, have a meaningful location tag and have posted at least 5 tweets including GPS coordinates.
- *Tweets dataset*: This dataset is formed by more than 250 K Twitter conversations including more than 3.5 M tweets. We refer to a Twitter conversation as the set of tweets formed by an original tweet and its associated retweets. Note that for each of the 3.5 M tweets, we have its associated location obtained either from the GPS coordinates of the tweet or from the location tag of the user who posted the tweet.

3 Twitter users' locality

Our goal in this section is characterizing the locality properties associated with the activity of Twitter users. For this purpose, we define the concept of *coverage area*. We define the coverage area as the geographical location (or set of locations) from where a Twitter user performs her activity. The activity of a Twitter user is divided into two major tasks: posting (producing) and reading (consuming) tweets. Although the coverage area from where these two tasks are performed may not be perfectly correlated at a low granularity level (e.g., specific address from where both activities are performed), it is reasonable to think that the location of both types of activity is highly correlated when we consider larger geographical areas such as a city or a country. Therefore, we assume that the set of geographical locations (e.g., city, country) from where a user either post or read tweets accurately defines the coverage area of this user.

To the best of our knowledge, there is any proposed technique that allows to retrieve the location from where a large number of Twitter users consume tweets. However, the methodology described in Sect. 2 enables us to collect the location from where hundreds of thousands users post their tweets. Therefore, in this section, we use our *Users Dataset* to characterize the coverage area of Twitter users. For this purpose, we first explore the geographical distance between the location tag provided by a user and the geolocation coordinates associated with her tweets. Second, we map the GPS coordinates of a user's tweets to different GCP communities (i.e., country, region/state and city). Finally, we analyze the fraction of tweets that a user posts from the different locations that form the user's coverage area.

3.1 Geographical distance of the coverage area

For each user in our *Users Dataset*, we consider the location tag as the user's reference location. We compute the distance between the location specified in the location tag and the location defined by the GPS coordinates for each one of the user's tweets. Figure 1 presents the CDF of the median distance between the location tag and the location of the different tweets for a user. The result shows that a large fraction of users (>70 %) typically post their tweets in a range of less than 100 Km from the location indicated in their profile. This suggests that: (1) the location tag can be safely used as an accurate location for a major portion of Twitter users. (2) A major fraction of Twitter users shows a coverage area in the order of few hundred kms.

3.2 Geopolitical composition of the coverage area

In this subsection, for each users within our *User Dataset*, we map the GPS coordinates of all her tweets to different GCP communities with different granularity, namely countries, regions[2] and cities.

Figure 2 shows the distribution of number of cities, regions and countries from where users within our *Users Dataset* send tweets. Note that the box represents the 25, 50

[2] We define a region as a GCP community smaller than a country and larger than a city. For instance states in USAor Germany or administrative regions in France.

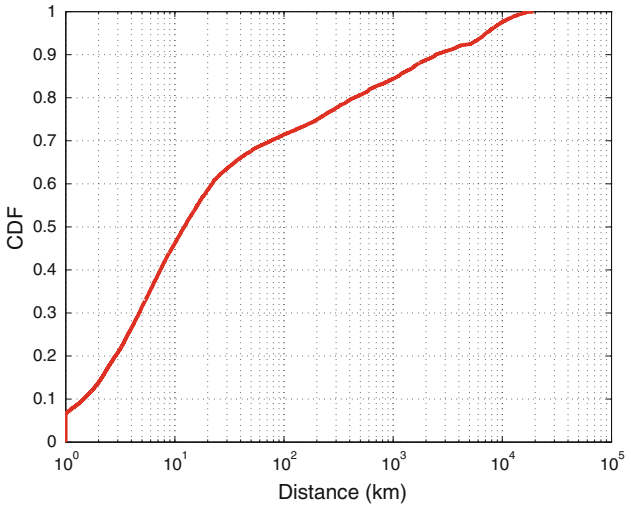

Fig. 1 Median distance between the user's location tag and the user's tweets GPS coordinates

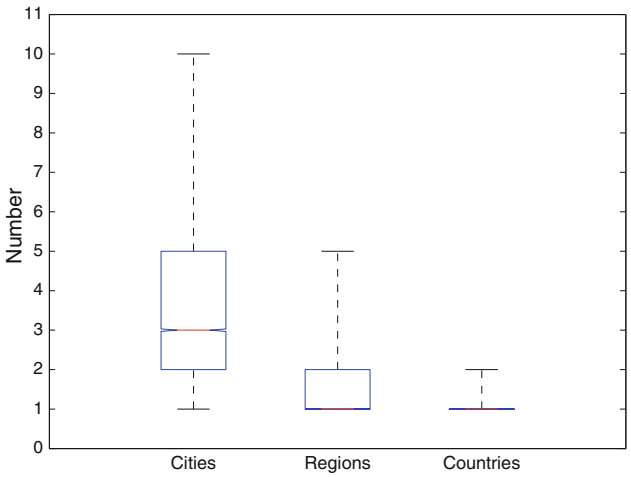

Fig. 2 Distribution of number of cities, regions and countries from where users send tweets

and 75 percentiles and the two external bars represent the 5 and 95 percentiles for the considered metric.

The obtained results show that the coverage area of Twitter users is formed by 3 cities in median, whereas just 25 % of users send tweets from more than 5 cities. Furthermore, if we consider carefully the other two more coarse metrics, we observe that 75 % of users send their tweets from just one or two regions and a single country.

3.3 Distribution of user's activity across different locations

In the previous subsections, we have analyzed the coverage area of Twitter users. Specifically, we have analyzed its size and the number of countries, regions and cities included in each user's coverage area. However, in order to fully characterize the locality associated with users' activity, it does not suffice with knowing from how many locations (e.g., cities) they perform their activity (i.e., post tweets), rather we need to analyze what is the fraction of the activity performed from each location. We address this issue in this subsection.

Figure 3 shows the cumulative percentage of users (y axis) who send at least x % of their tweets (x axis) from outside their main location using three types of GCP communities with different granularity: city, region and country. We group users by the number of associated locations (n) in the following groups: 2 locations, 3 locations, 4 locations, 5 locations, more than 5 locations and *global* that includes all users in our dataset. Note that the group of users with an unique location ($n = 1$) is not included in the figure since they send all their tweets from that single location.

Let us focus first on the *global* group that includes all the users. The results show that 90 % of users send all their tweets from a single country. This percentage shrinks to 60 and 20 % for regions and cities, respectively.

If we now consider the other groups, the results reveal two important observations: (1) the main location is significantly more used by the user than the other ones. For instance, 50 % of users post at least 50 % of their tweets from the main location for all groups (excepting for $n > 5$) and all types of locations (city, region or country); (2) in general, the users do not tweet from sporadic locations, rather they tweet from locations that they visit frequently. We refer to a sporadic location as that one that the user visit just one (or few times) and from where she posts just few tweets (e.g., during a business trip). Note that if these sporadic locations were common, their presence would influence more to those groups having larger values of n. Then, the separation between the curves should become significantly smaller as we increase the number of locations.

3.4 Summary

The obtained results suggest that around 3/4 of Twitter users perform their activity from a relatively reduced coverage area within a country that covers few hundred km including few ($\leq$5) cities and an even smaller ($\leq$2) number of regions. Hence, these results reveal that the area from where most Twitter users perform their activity is highly localized. In addition, our study on the activity distribution across users' locations reveals that there is typically a predominant location (city, region or country) from where the user post a significant portion of her tweets. At the same time, users seem to rarely post tweets from "sporadic" locations. Finally, our analysis reveals that the user's location tag accurately define the location of a user (at least at the country level).

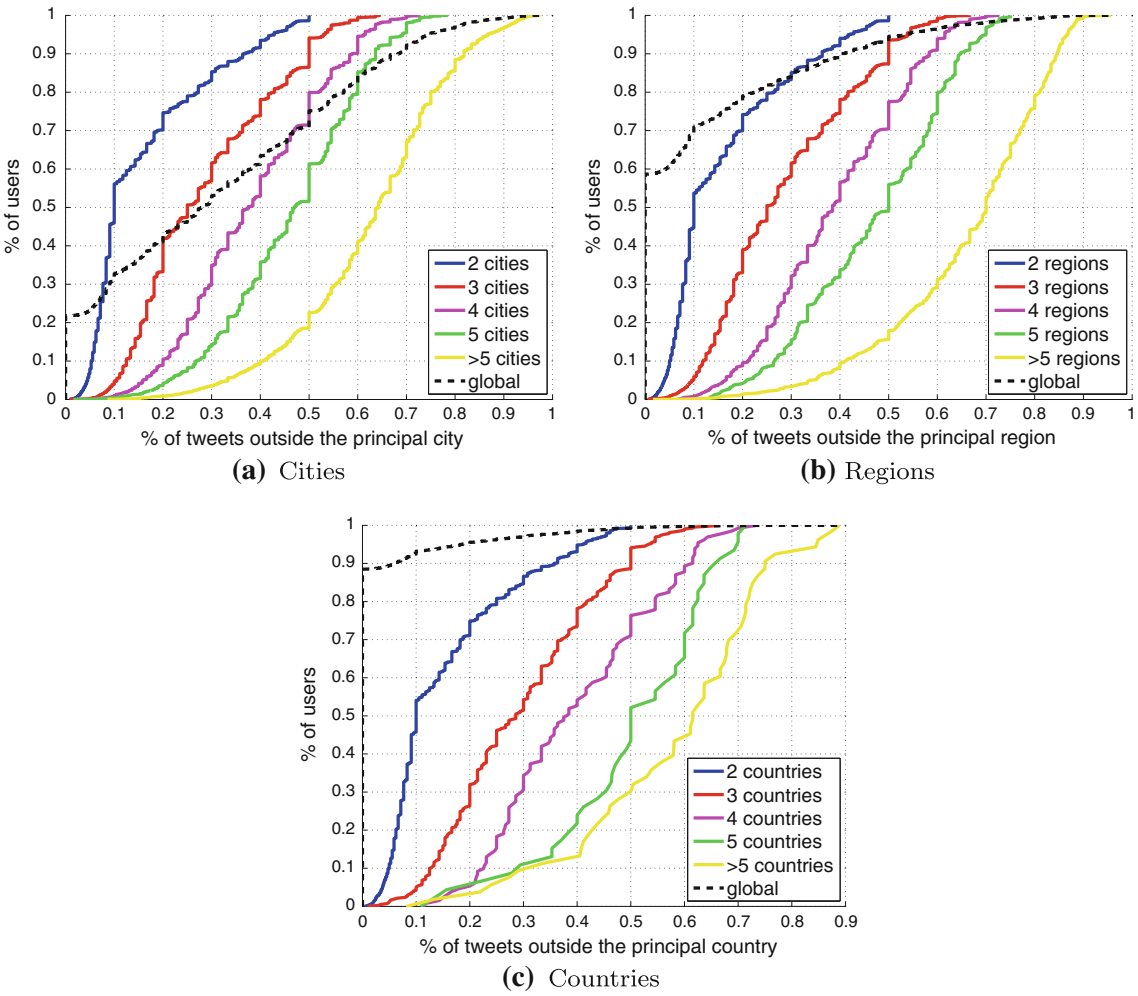

Fig. 3 Percentage of users versus percentage of tweets sent from a different location than the principal one (city, region and country)

4 Twitter relationships' locality

In this section, we study the geographical properties associated with Twitter relationships, that is, *friend* → *follower* links. For this purpose, we rely on our *Relationship Dataset* that includes more than 100 M relationships in which both friend and follower have a location tag.

In order to perform the analysis, we group the friends in our dataset by country. We have selected the country criteria since it perfectly matches the concept of GPC community, that is, friends having a close geographical location, a similar cultural profile and the same language. Furthermore, as we have demonstrated in the previous section, we can map a user to a country with a very low error probability.[3]

[3] We could perform the same analysis using GPC communities at different granularities (e.g., regions or cities). However, as we will see, our analysis based on countries reveals important insights, then we leave the analysis with other GPC communities for future work.

We first study the demographic composition of our dataset. Then we characterize the geographical properties of the Twitter relationships by carefully studying the fraction of intra and inter *friend* → *follower* relationships for the most relevant countries in our dataset.

4.1 Twitter demographics

Table 1 shows the number of friends, the number of followers and the number of originated and received *friend* → *follower* links for the top 14 countries in our dataset that are those that contribute more than 100 K users. Note that overall these 14 countries are responsible for around 85 % of all the friends, followers and relationships within our *Relationships Dataset*. Furthermore, USA is clearly a predominant country in Twitter responsible for around half of the friends, followers and links. Among the other countries we observe two clear profiles from a language perspective. On the one hand, we have those countries

Table 1 Contribution of the top 14 countries to the *Relationships Dataset*, sorted by the number of originated friend → follower links

Country	Language	Friends (num/%)	Followers (num/%)	Originated friend → follower links (num/%)	Received friend → follower links (num/%)
USA	EN	528 K/54.24	7.37 M/44.59	60.1 M/59.82	57.1 M/56.84
UK	EN	70.6 K/7.27	987 K/1.41	7.18 M/7.15	6.94 M/6.90
BR	PO	61.7 K/6.34	1.81 M/10.94	6.46 M/6.42	6.74 M/6.70
CA	EN/FR	39.4 K/4.05	565 K/3.42	4.74 M/4.72	4.55 M/4.53
AU	EN	20.3 K/2.09	232 K/1.40	2.50 M/2.48	2.40 M/2.38
DE	DE	21.7 K/2.23	331 K/2.00	2.02 M/2.01	2.26 M/2.25
IN	IN/EN	18.8 K/1.93	442 K/2.67	1.28 M/1.28	1.52 M/1.51
NL	NL	14.9 K/1.53	334 K/2.02	1.22 M/1.22	1.26 M/1.25
ES	SP	8.7 K/0.89	277 K/1.68	0.90 M/0.89	904 K/0.90
FR	FR	10.8 K/1.11	232 K/1.41	0.82 M/0.82	840 K/0.84
ID	ID	12.1 K/1.24	862 K/5.22	0.64 M/0.64	1.09 M/1.09
MX	SP	5.5 K/0.56	234 K/1.41	0.55 M/0.55	657 K/0.65
IT	IT	7.1 K/0.73	159 K//0.96	0.49 M/0.48	637 K/0.63
JP	JP	6.9 K/0.71	192 K/1.16	0.48 M/0.48	597 K/0.59
TOP 14	–	827 K/85.00	13.37 M/80.31	89.9 M/88.95	88.06 M/87.08
ALL	–	973 K/100	16.53 M/100	100.5 M/100	100.5 M/100

whose official (or co-official) language is the English such as USA, Canada, UK, India and Australia. On the other hand, we find those countries with a different official language than English such as Brazil, Spain, Germany, France, Italy, Indonesia, Japan and The Netherlands. Finally, it is worth to note the presence of developing countries such as Brazil, India and Mexico in the list. This is mainly due to the big population of these countries that enables to contribute a large number of users, but also indicates the interest of their population on new social ways of communication such as Twitter.

Once we have analyzed the basic demographics of our dataset, in the rest of the section we focus on analyzing the fraction of intra- and inter-country relationships for each one of the top 14 countries. For this purpose, we rely on both the GPC community information (i.e., user's country) and the geographical distance of the *friend* → *follower* links.

4.2 Geopolitical analysis

For each *friend* → *follower* link within our *Relationships* dataset, we identify the country of the friend and the follower involved in the relationship. This allows us to study the destination of all the relationships originated in a given country. In particular, we perform a twofold analysis. First, we study the aggregate percentage of relationships generated in a country that go to different destinations. We refer to this analysis as *link-level* analysis. However, the behavior of unpopular users might not be well captured in such analysis since those popular users are the ones responsible for a larger portion of the relationships generated in a country. Therefore, in the second part of our analysis, we study the percentage of links associated with each individual user who go to different destinations. We refer to this analysis as *user-level* analysis.

4.2.1 *Link-level analysis*

For each one of the top 14 countries, we compute the percentage of *friend* → *follower* links originated in the country that: (1) remain within the country, (2) go to USA (predominant country) and (3) go to a different country other than USA. Figure 4 depicts the obtained results that show the presence of significantly different behaviors across the studied countries. Specifically, we can distinguish the next four different profiles:

USA: due to its predominant role, it has to be considered as a separated profile. It keeps more than 70 % *friend* → *follower* relationships local. This is consequence of first, the predominance of USA users in Twitter and second, the strong local culture (e.g., sports, music, TV, etc) of USA.

Local profile: This is formed by a group of countries that keep local a higher number of links than those going to USA or other countries. This is *Local* > *USA* & *Local* > *Other* in Fig. 4. This profile includes Brazil, The Netherlands, Indonesia, Germany and Spain. All these countries have an official language different than English and present a relatively high popularity for Twitter. Moreover, we found also some noticeable differences within the group. On the one hand, Brazil is the country showing the highest locality in our dataset with almost 80 % of local links. This

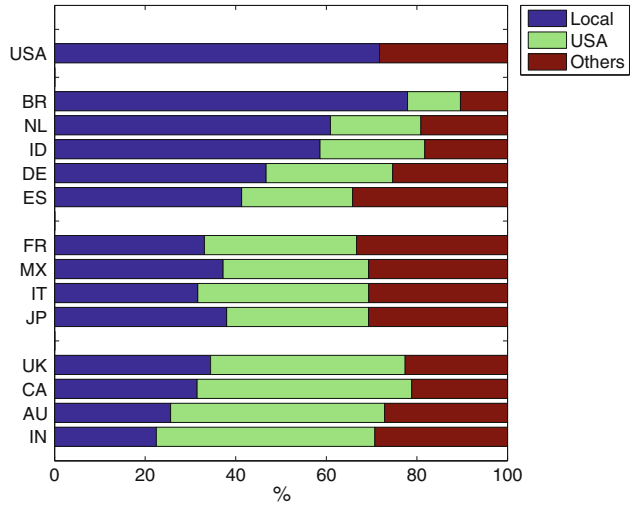

Fig. 4 Percentage of *friend* → *follower* relationships originated in each one of the top 14 countries that remain local, go to USA or go to another country different than USA

is because it is a big country with a strong local culture and the spoken language (Portuguese) is not very spread. Just other countries, not very representative in Twitter, such as Portugal use Portuguese. On the other hand, we have Spain whose local links are reduced to 41 %, since now many relationships (>20 %) are established with Latin-America. Note that Spain shares a common language with most south and central American countries.

Shared profile: This group is formed by those countries that distribute their *friend* → *follower* links roughly equally among those that remain local, those that go to USA and those that go to other countries. This profile includes France, Mexico, Italy and Japan that are those countries where Twitter is less popular among the studied ones.

English profile: This group is formed by all those countries from our dataset where English is the official or a co-official language (apart from USA): UK, Canada, Australia and India. In addition, all these countries are members of the Commonwealth of Nations. Language becomes the major driver to define the geographical properties of the links originated in these countries. The demographic predominance of USA (another English speaking country) produces that the major fraction of links originated in the countries within this group are destined to USA (e.g., 48 % in the case of India and 47 % in the case of Australia and Canada). We refer to this phenomenon as *External locality*. Furthermore, a lower but also important portion of links stay local (e.g., 34 % for UK and 31 % for Canada) and the rest are shared mainly with other English speaking countries.

In summary, the results reveal that there are three main drivers that define the locality profile for the friend → follower relationships originated in a specific country, namely the language and culture of the country and the local popularity of Twitter. The combination of these factor highlights the presence of four different profiles.

4.2.2 *User-level analysis*

Again for this analysis, we group the users per country and consider the top 14 countries. For every friend in a specific country, we calculate the fraction of *friend* → *follower* links that stay local within the country, go to USA and go to another country different than USA. Due to space limitations, in this paper, we present results for one representative country per each defined profile above. Specifically, we consider the country with the largest number of users from each profile. These countries are: Brazil for the local profile, France for the shared profile, UK for the English profile and USA since it represents a unique profile. Note that the described experiments have been conducted for every country within each profile and the obtained results lead to similar conclusions to those presented in this paper.

Figure 5 depicts density diagrams in which the x axis represents the percentage of *friend* → *follower* links that remain local and the y axis represents the percentage of *friend* → *follower* links that go to either USA (Fig. 5a–c) or another country (Fig. 5d–g) for each individual user within each analyzed country.

The results show clear differences across the studied countries. First, we can observe that the intra-country locality grows in the following order: BR (locality Profile) > USA > FR (Shared Profile) > UK (English Profile, presenting an external locality phenomenon). Specifically, most of the Brazilian users have between 80 and 100 % of internal followers, whereas in USA we observe a slightly lower intra-country locality where users present a percentage of local followers between 70 and 90 %. Looking at the European countries, we observe a higher level of localization in France where the vast majority of users show between 40 and 80 % of local followers, whereas the UK presents a less concentrated diagram where the percentage of local followers per user ranges from 20 to 80 %. Moreover, we observe how the remote followers of UK are more concentrated in USA, whereas French users tend to have a balanced presence of followers in USA compared with other countries.

4.3 Distance-based analysis

The previous subsection has demonstrated the presence of clearly differentiated profiles across the studied countries. In this subsection, we use geographical distance associated with *friend* → *follower* links instead of GPC communities information (i.e., user's country) in order to validate and extend our previous observations. Due to space limitations,

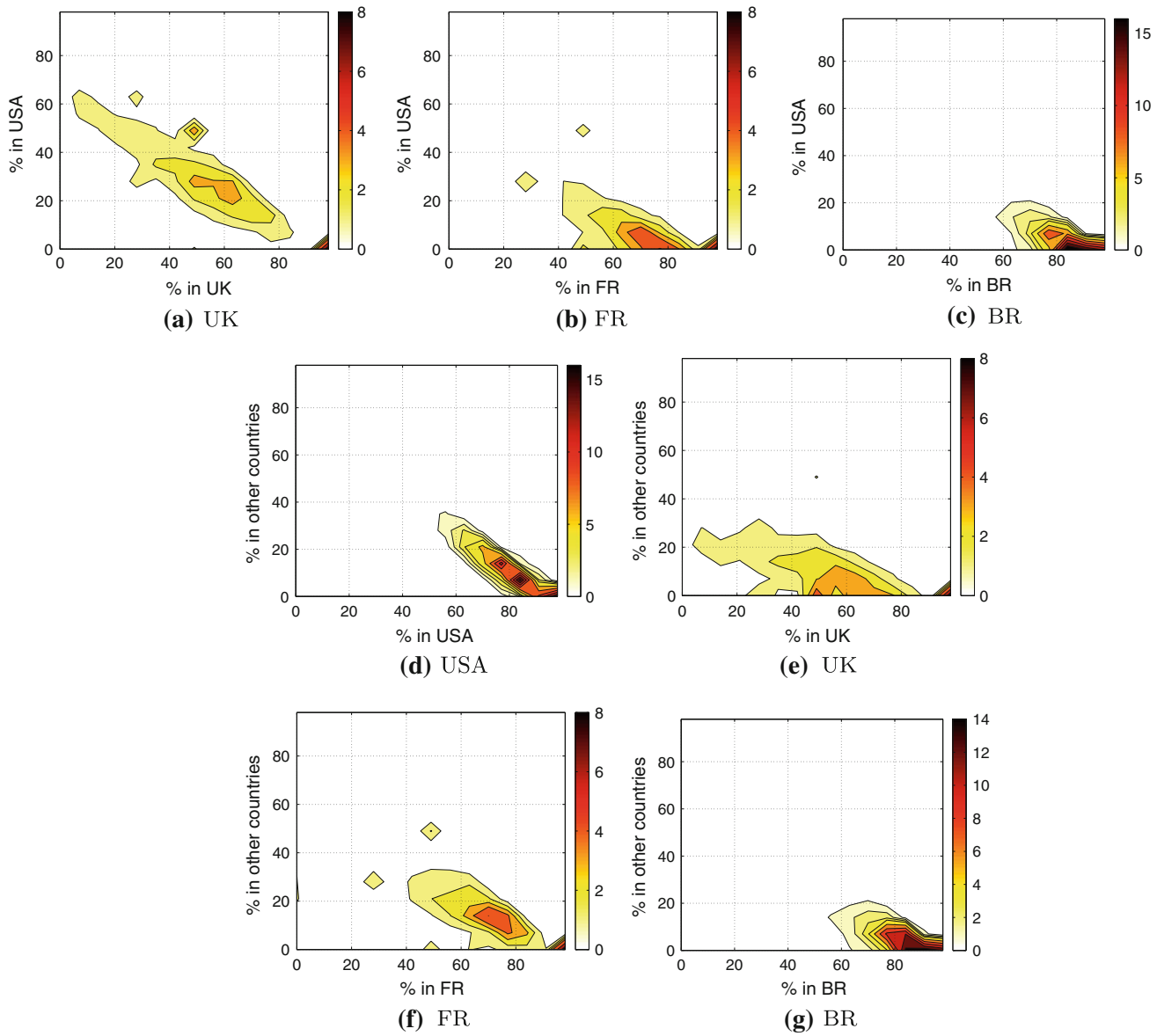

Fig. 5 Percentage of *friend* → *follower* relationships that remain local versus those that go to USA (*top*) or to another country (*bottom*) for each individual user within the following countries: USA, UK, France and Brazil

we again present results for one representative country per profile that are Brazil, France, UK and USA. We have repeated the experiments for the rest of top 14 countries and we conclude that the overall observations presented in this paper are generally valid.

As in the previous subsection, we perform a twofold analysis: link- and user-level analyses. The link-level analysis considers separately each individual link originated in a specific country. As mentioned before this makes that popular users have a major impact in the observed results than unpopular users since the former contribute more links. In order to perform the user-level analysis, we have to calculate a distance metric that characterizes the typical distance from a friend to its followers. To this end, we compute the user-level distance as the median of all the *friend* → *follower* distances associated with a friend.

Figure 6 presents the distribution of *link-level* and *user-level* distances for each one of the analyzed countries. In addition, Table 2 shows the analytical distribution that best fit the empirical link- and user-level distribution for each country. In particular, we have used a power-law fitting technique [13], and in those cases where the distribution has two differentiated parts (i.e., UK, FR and BR), we have applied the fitting technique separately for each part. Finally, we have computed a Kolmogorov–Smirnov test [14] for each empirical/analytical distributions pair and confirmed the accuracy of all the presented analytical distributions.

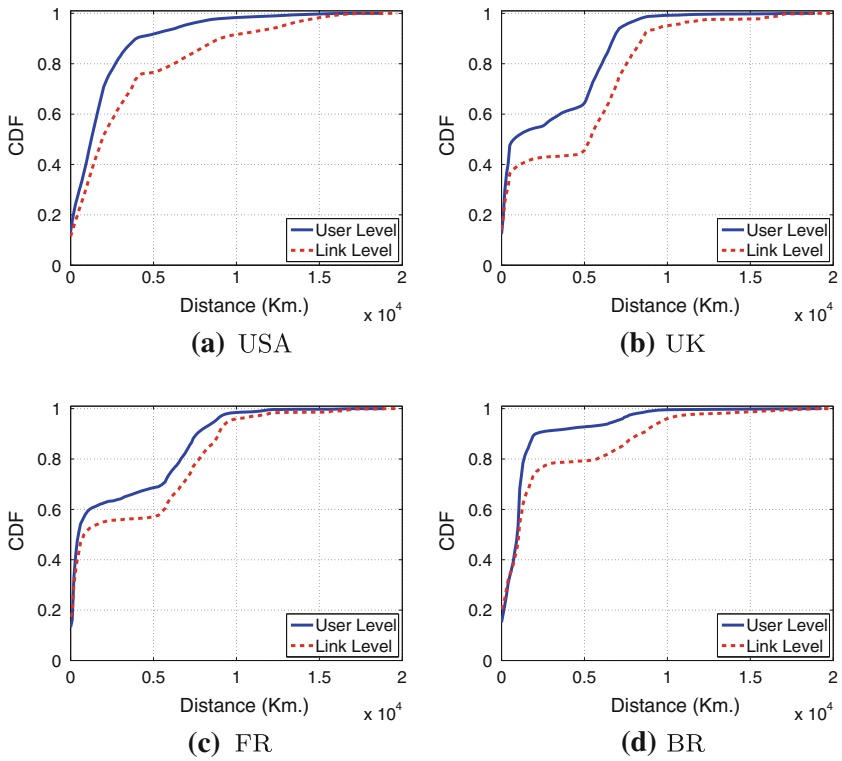

Fig. 6 Distribution of user- and link-level distances for USA, UK, France and Brazil

We observe that around 90 % of USA users have a typical user-level distance to its followers ≤4,000 km that defines the intra-country boundary for most relationships originated in USA. This intra-country locality effect is even more impressive in Brazil where 90 % of the users have a user-level distance ≤2,000 km, when the limit for most intra-country relationships is also about 4,000 km. If we analyze UK, it shows, a clear bi-polar distribution that validates the observation done by our geopolitical analysis. Around 60 % of links have an associated link-level distance over 5,000 km that correspond to cross-continental links from which a major portion goes to USA. Furthermore, around 40 % of links have an associated link-level distance of few hundreds km that correspond to local relationships. If we focus now on France, 60 % of its links have an associated link-level distance shorter than 1,000 km. Several neighbor countries such as Belgium, Switzerland,[4] The Netherlands, Italy and Germany are located within this distance range. Hence, this 60 % of links are divided into intra-country relationships and inter-country relationships with followers located in neighbor countries. In addition, around 1/3 of the French users present a user-level distance to its followers between 5,500 and 9,500 km, which mostly represents the followers population in USA. Therefore, our distance-based analysis validates the observations done during our geopolitical-based analysis and the presence of four different profiles.

Finally, we observe that every country shows a higher locality (more skewed curve) at the user-level than at the link-level. This suggests that unpopular users tend to have a more localized followers population than popular users. In order to confirm this hypothesis, we group the users by its popularity[5] (i.e., number of followers) and for each group, we calculate the median for user- and link-level distances. Figure 7 shows the obtained results. In general, we observe that our hypothesis is correct since more popular users typically present a larger user-level distance and their relationships show a higher link-level distance. However, we observe significant differences among the analyzed countries that are worth to discuss. USA shows a quasi-linear correlation between popularity and locality. The higher the popularity is the longer are the user's *friend* → *follower* links. Contrary, Brazil users show a high intra-country locality (median distances around 1,000 km) that is almost independent of their popularity (i.e., the curve is almost flat). Finally, we can observe a clearly denoted bi-polarity in UK and France. In UK, those unpopular users with less than 100 followers present a clearly marked

[4] Note that French is co-official language in both Belgium and Switzerland.

[5] We group the users in the following popularity buckets as function of the number of followers: [1–50], [51–100], [101–500], [501–1,000], [1,001–5,000], [5,001–10,000], [10,001–50,000], [50,001–100,000], [100,001–500,000] and a last bucket including all those users having >500 K followers.

Table 2 Power-law parameters for the distribution of user- and link-level distances for USA, UK, France and Brazil

Country	Distance limit (km)	User level		Link level	
		α	x_{min}	α	x_{min}
USA	All	27.45	15K51	26.86	16K09
UK	≤5 K	1.89	221.72	2.32	659.72
UK	>5 K	7.32	7K04	4.94	6K33
FR	≤5.5 K	2.01	295.91	2.36	540.70
FR	>5.5 K	6.93	6K98	16.58	8K2
BR	≤6 K	4.00	1K16	4.20	1K79
BR	>6 K	7.02	7K22	5.38	8K42

For those distribution having two differentiated parts we present specific parameters for each part

intra-country locality, whereas the popular users show an *external* locality phenomenon with most of its followers in other continents (mainly USA). In France, we observe the same bi-polar phenomenon, but the transition happens for 1,000 rather than 100 followers.

4.4 Summary

The geopolitical- and distance-based analyses conducted in this section have revealed important insights into the geographical properties of *friend* → *follower* relationships in Twitter. The combination of language, culture and Twitter popularity has a clear influence in the locality level of the users' relationships in different countries. Indeed, these factors produce the presence of four different country profiles that we have thoroughly discussed along the section. Furthermore, the conducted user- versus link-level distance analysis has demonstrated that locality and popularity are generally inversely proportional. However, the level of correlation varies across countries.

The insights revealed on this section demonstrate that the user's GCP community (i.e., country) clearly impacts its relationships. Moreover, we have showed how the combination of factors such as language and local Twitter popularity produces interesting interactions between different GCP communities (e.g., external locality phenomenon).

5 Twitter information flows' locality

The goal of this section is understanding the level of locality existing in the information flow in Twitter. For this purpose, we use our *Tweets Dataset* that includes more than 250 K Twitter conversations. We first compare the locality level observed in the conversations generated in different GCP communities. Again in this section, we use GCP communities formed by users within a country. Afterward we study how the popularity of Twitter conversations influences their level of locality.

5.1 Locality of Twitter conversations in different countries

Figure 8 shows the CDF of the percentage of retweets done from a different country than that one where the conversation was originated. The figure shows results for all the conversations in our datasets (All) as well as conversations originated in USA, Brazil, France and UK (representative countries of each profile defined in Sect. 4). Let us first analyze the aggregate behavior by looking at the curve associated with "All" conversations. We observe that in general, Twitter conversations show a low locality. Specifically, just 10 % of the conversation remains local within a country whereas more than 20 % of the conversations have all the retweets in different countries than the country associated with the original tweet. If we focus now on different countries, as expected, we observe very different behaviors. On the one hand, USA and Brazil show a higher locality level compared with the aggregate trend represented by "All". Specifically, the conversations originated in Brazil present the highest locality level (70 % of the conversation present at least 70 % of local retweets) clearly above the level shown by conversation generated in USA. On the other hand, the conversations generated in UK and France show a locality level below than the aggregate trend. In the case of UK, the fact that English is a widespread language and the predominance of USA in the number of users ease that conversations originated in UK rapidly move outside the country. France shows a surprisingly low locality since more than half of the conversation originated in France have all its retweets outside France. This seems to be a consequence of the low popularity of Twitter in the country.

5.2 Influence of popularity in the locality of Twitter conversations

We have divided the conversations in the four following groups based on their number of retweets (r): $r < 10$, $10 \leq r < 50$, $50 \leq r < 100$ and $r \geq 100$.

Figure 9 shows the CDF of the percentage of retweets done from a different country than that one where the conversation was originated for the defined popularity groups. Furthermore, we add the curve including all the conversations (All) for reference. We observe that the different distributions are relatively close to each other. This suggest that the influence of the popularity of conversations in their locality is small. Only those conversations with >100 retweets present a relatively significant

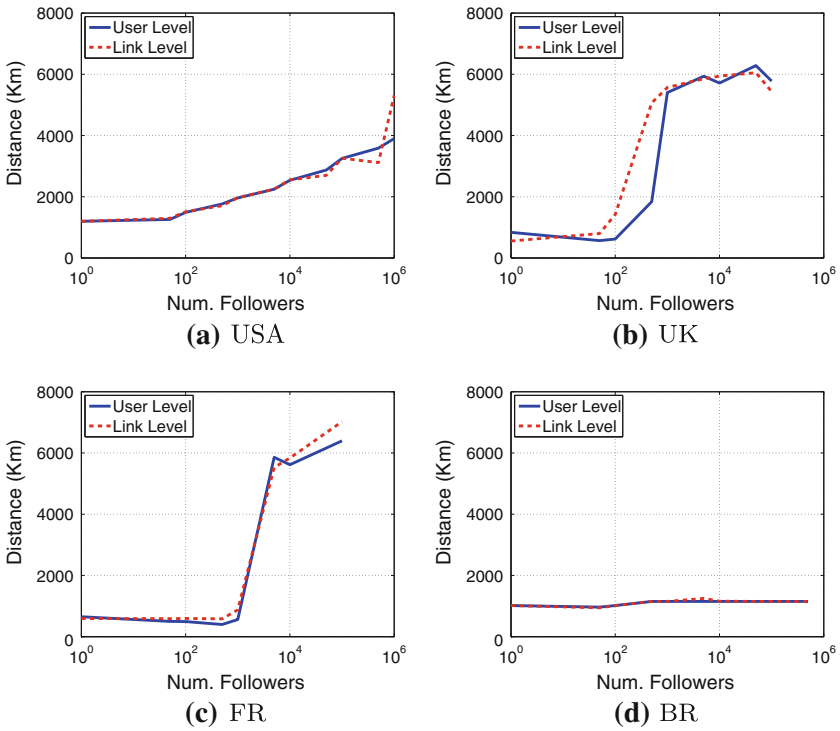

Fig. 7 Median link- and user-level distance as function of the users' popularity for USA, UK, France and Brazil

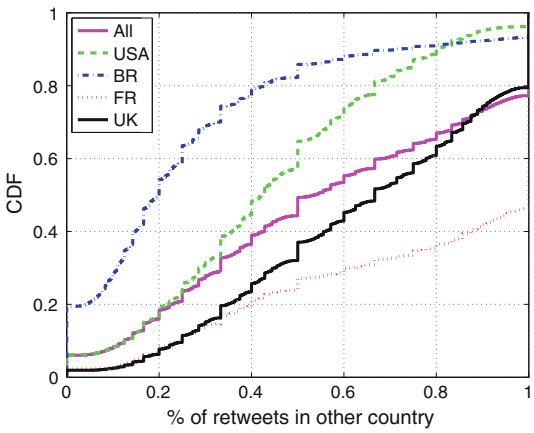

Fig. 8 CDF of the percentage of retweets posted from a different country than the original tweet for "All" conversations and conversations originated in USA, UK, France and Brazil

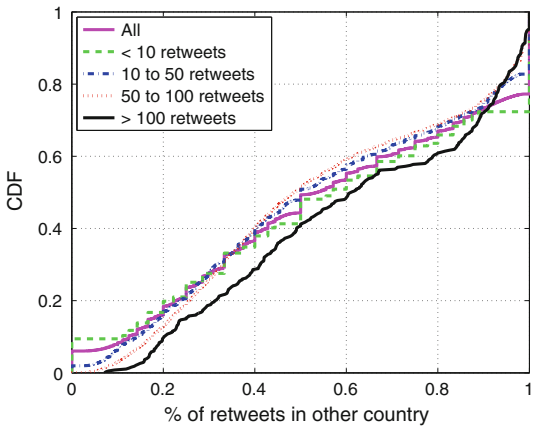

Fig. 9 CDF of the percentage of retweets posted from a different country than the original tweet for different popularity levels of the conversation

lower locality than the other groups what is an expected result.

5.3 Summary

In this section, we have studied the level of locality of more than 250 K Twitter conversations. First, we have observed that Twitter conversations present a rather low locality since just 10 % of them remains fully local within a country. Furthermore, our results reveal that the origin GCP community (i.e., country) of the conversation have a much higher impact on the locality level than the popularity of the conversation. Indeed, the low impact of the conversation popularity in its locality level is a surprising result, since as occurred in the case of relationships we expected that locality level of Twitter conversations were correlated with their popularity.

Finally, the analysis of individual countries shows that the locality levels associated with relationships and conversations (i.e., information flow) are clearly correlated for a country. Therefore, we can conclude that drivers such as language or Twitter popularity determine the overall level

of locality observed at both the relationship and the information flow level.

6 Practical applicability of the obtained results

In this section, we briefly discuss some practical applications for which the obtained results are of high interest. Note that our intention is not to present an exhaustive list but a few representative examples to demonstrate the great usefulness of the obtained results.

Online social marketing: Given the overwhelming success of OSNs in the recent years, companies have started to use them as a channel to advertise their products and services. However, contrary to traditional broadcast media such as TV or radio, there are still not clear (or successfully proven) marketing strategies in OSNs. The results obtained in this paper provide important insights into be considered in such strategies. For instance, localized online social marketing campaigns are potentially more successful in countries within the Local Profile (e.g., Brazil) than in those experiencing an external locality phenomenon (e.g., India). Furthermore, other factors such as language or popularity of the specific OSN seem to play a key role that must be considered by marketing specialists. For instance, our results reveal that in those countries where Twitter is less popular the information seems to rapidly flow outside of the country what is an indication that local marketing campaigns may not be efficient in those countries.

OSN infrastructure: Scalability is a major issue for OSN providers [15] who have to continuously upgrade their infrastructure in order to satisfy the ever increasing demand of the offered services. Several researchers have proposed the utilization of distributed infrastructures to deal with the scalability problems of the currently centralized infrastructures of OSN provides [15–18]. The design of distributed infrastructures to provide OSN services faces several challenges such as: selecting the location of the distributed servers (or datacenters), designing efficient caching/prefetching algorithms (i.e., in which server to cache/prefetch the content of which user), defining effective replication algorithms. The results obtained in this paper provide relevant information for the design of a distributed infrastructure for Twitter in particular and other OSN systems in general. First, our study of the Twitter demographics across countries can be used as reference for the placement of servers (or datacenters) so that they are deployed close to a major portion of the Twitter population. Second, the analysis of geographical properties associated with Twitter users reveal that most users present a reduced mobility in the range of few hundreds kms. Therefore, if the distributed infrastructure is formed by servers (or datacenters) at regional level (or any other more coarse granularity, e.g., country or continent levels), we could associate users to a specific server within the distributed architecture with a high level of accuracy reducing the overhead produced by algorithms to dynamically assign users to servers. Furthermore, the proposed measurement methodology can accurately identify the location of the followers for a specific user. Therefore, it can be used in order to implement sophisticated caching/prefetching algorithms to decide the server(s) where the information generated by a user should be stored.

Other applications: In addition to the two previous examples, the insights revealed in our study are also of high interest in other fields. Specifically, it has been shown that language, cultural or popularity factors have a direct impact in the behavior of Twitter users, and thus, they must be included as part of new defined community detection algorithms or models for information diffusion in OSNs.

7 Related work

Twitter measurements: Several previous works have exploited the different APIs offered by Twitter in order to collect data and describe different characteristic of the system. Krishnamurthy et al. [7] performed one of the initial measurement studies on Twitter collecting data for 100 K users. The authors report basic characteristics of the system such as the correlation between number of followers and friends of a given user or the distribution of Twitter users per continent. Afterward, Kwak et al. [8] collected the complete *friend* → *follower* Twitter graph including 41.7 million users at the moment of the study. The authors analyze the properties of the graph topology as well as some other social aspects of Twitter such as the users influence. Also in the field of users influence Cha et al. use a large dataset in order to analyze the dynamics of user influence across topic and time in Twitter. Finally, some other studies [5, 6, 9] focus on understanding social aspects of the Twitter system.

Geographical properties in location-based OSNs: Location-based OSNs (LbOSNs) are an specific type of OSNs where users share their location through *check-ins* in different places. Several studies have recently analyzed geographical-related properties of these applications. Scellato et al. [19] analyzed the distance of social links established between users for three different LbOSN (Foursquare, Gowalla and Brightkite). They conclude that the three systems exhibit 40 % of links below 100 km. Some other works [20, 21] leverage data from different LbOSNs to study human mobility patterns. Cho et al. [20] show that most users' movements happen in a short distance range. Furthermore, Noulas et al. [21] demonstrate that the density of places as well as the range of available

places at a given distance play an important role in the human mobility patterns within a city. Finally, a most recent work by Allamanies et al. [22] study the link formation phenomenon in a LbOSN (Gowalla) to understand how geographical distance and social factors affect the creation of new links in this type of networks. Twitter and LbOSNs are OSNs of different nature, therefore, the results from the previously discussed works are likely to not apply in Twitter. Furthermore, in our paper, we look at the geographical properties of the three main elements of an OSN (users, relationships and information) rather than focusing in a single one as the previous papers do. Finally, we use CCM techniques to provide insights into how factors from the physical world (e.g., language or culture) affect the geographical properties of the aforementioned elements.

Locality in large-scale systems: Locality is an important aspect to be considered in the design of most of the large-scale Internet applications. Having it into consideration may help to improve the system design and performance of distributed systems. Some examples demonstrate it for the case of p2p file-sharing applications [23–25], p2p live-streaming applications [26] or OSNs such as Facebook [27]. Although Twitter has significantly different characteristics to p2p applications and slightly different to Facebook, considering the *locality* phenomenon in the system design may help to improve the performance and also the data storage procedure [28] of Twitter.

Locality in Twitter: The recent work by Kulshrestha et al. [29] makes a geographical dissection of the Twitter network with the goal of investigating how users geolocation impacts in their participation in Twitter. This work is partially focused on what we described in our paper as *User Locality*. There are also other interesting works on the exploitation of the location information in Twitter and other OSNs to improve content distribution and evolution algorithms in real systems [30–32]. However, to the best of our knowledge, our study is the first one that performs a comprehensive study of the locality at the user, relationship and information flow levels in Twitter. Furthermore, we leverage novel CCM techniques in order to understand the factors from the physical world (e.g., language or culture) that influence the online behavior of Twitter users.

8 Conclusion

In this paper, we use a large-scale dataset including the geolocation information of more than 17 M users and 3.5 M tweets to perform a comprehensive analysis of the geographical locality properties of Twitter users, relationships and information flow.

Our users' locality analysis reveals that Twitter users present a high locality profile since they typically perform their activity from few cities separated at most few hundreds kms within the same country. Furthermore, for the analysis of relationships' locality, we have formed meaningful geo-cultural-political (GPC) communities in which we group users per country. Our results demonstrate that factors directly associated with the user's GPC community (i.e., country) drive her behavior in Twitter. These factors include language and local popularity of Twitter within the country. Interestingly, these same factors dictate the locality level associated with Twitter conversations originated in a country. For instance, we have demonstrated that countries with a different language than English and where Twitter is popular, such as Brazil, present a high locality level at both relationships and information flow level. Furthermore, the clear predominance of USA on Twitter demographics influences the Locality of relationships originated in other English speaking countries such as UK or Canada, that show an interesting external locality phenomenon since the major fraction of the relationships generated in these countries are destined to USA. Finally, countries where Twitter shows a (relatively) low popularity, such as France, present a high unlocalized profile since just 1/3 of the relationships and 2 % of the conversations remain local within the country.

The presented results are a step forward on our understanding of the online behavior of Twitter users and more importantly the factors that influence such behavior. Furthermore, these results have applications in multiple fields such as: (1) they reveal important parameters to be considered in the definition of community detection algorithms for OSNs, (2) our observations have a direct impact on the area of social marketing, since they clearly differentiate several country profiles in which marketing should be addressed in a different way (e.g., countries with a high level of locality are more suitable for local marketing campaigns), (3) these results constitute a basic element to understand the information diffusion in social media and (4) our conclusions can be considered in the design of future infrastructure to provide social media services. In particular, they are interesting in the design of distributed solutions.

Acknowledgments The authors would like to thank anonymous reviewers for their valuable feedback. This work has been partially supported by the European Union through the FP7 TREND (257740) and eCOUSIN (318398) Projects, the Spanish Ministry of Economy and Competitiveness through the eeCONTENT project (TEC2011-29688-C02-02) and the Regional Government of Madrid through the MEDIANET project (S-2009/TIC-1468).

References

1. Twitter. http://www.twitter.com
2. Techcrunch. http://techcrunch.com/2012/07/30/analyst-twitter-passed-500m-users-in-june-2012-140m-of-them-in-us-jakarta-biggest-tweeting-city/
3. Twitter blog. http://blog.twitter.com/2012/03/twitter-turns-six.html
4. Cha M, Haddadi H, Gummadi PK (2010) Measuring user influence in Twitter: the million follower fallacy. In: Proceedings of AAAI ICWSM
5. Honeycutt C, Herring SC (2009) Beyond microblogging: conversation and collaboration via Twitter. In: Proceedings of Hawaii international conference on system sciences
6. Java A, Song X, Finin T, Tseng B (2007) Why we Twitter: understanding microblogging usage and communities. In: Proceedings of WebKDD/SNA-KDD '07
7. Krishnamurthy B, Gill P, Arlitt M (2008) A few chirps about Twitter. In: Proceedings of WOSN '08
8. Kwak H, Lee C, Park H, Moon S (2010) What is Twitter, a social network or a news media? In: Proceedings of WWW'10
9. Zhao D, Rosson MB, How and why people Twitter: the role that micro-blogging plays in informal communication at work. In: Proceedings of ACM GROUP '09
10. Guo B, Zhang D, Yu Z, Zhou X (2012) Hybrid sn: interlinking opportunistic and online communities to augment information dissemination. In: Proceedings of the 9th IEEE international conference on ubiquitous intelligence and computing (UIC12), Fukuoka, Japan
11. Twitter API Documentation. http://dev.twitter.com/doc
12. Yahoo Geo Technologies. http://developer.yahoo.com/geo/
13. Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703
14. Massey FJ (1951) The kolmogorov-smirnov test for goodness of fit. J Am Stat Assoc 46(23):68–78
15. Xu T, Chen Y, Zhao J, Fu X (2010) Cuckoo: towards decentralized, socio-aware online microblogging services and data measurements. In: Proceedings of the 2nd ACM international workshop on hot topics in planet-scale measurement, HotPlanet '10, pp 4:1–4:6, ACM, New York, NY, USA. doi:10.1145/1834616.1834622
16. Buchegger S, Schiöberg D, Vu LH, Datta A (2009) Peerson: P2p social networking: early experiences and insights. In: Proceedings of the second ACM EuroSys workshop on social network systems, SNS '09, pp 46–52. ACM, New York, NY, USA. doi:10.1145/1578002.1578010
17. Shakimov A, Varshavsky A, Cox LP, Cáceres R (2009) Privacy, cost, and availability tradeoffs in decentralized osns. In: Proceedings of the 2nd ACM workshop on online social networks, WOSN '09, pp 13–18, ACM, New York, NY, USA. doi:10.1145/1592665.1592669
18. Kryczka M, Cuevas R, Guerrero C, Yoneki E, Azcorra A (2010) A first step towards user assisted online social networks. In: Proceedings of the 3rd workshop on social network systems, p 6, ACM
19. Scellato S, Noulas A, Lambiotte R, Mascolo C (2011) Socio-spatial properties of online location-based social networks. Proc ICWSM 11:329–336
20. Cho E, Myers S, Leskovec J (2011) Friendship and mobility: user movement in location-based social networks. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining, pp 1082–1090. ACM
21. Noulas A, Scellato S, Lambiotte R, Pontil M, Mascolo C (2012) A tale of many cities: universal patterns in human urban mobility. PLoS One 7(5):e37, 027
22. Allamanis M, Scellato S, Mascolo C (2012) Evolution of a location-based online social network: analysis and models. In: Proceedings of ACM internet measurement conference (IMC 2012)
23. Choffnes DR, Bustamante FE, Taming the torrent: a practical approach to reducing cross-isp traffic in peer-to-peer systems. In: Proceedings of ACM SIGCOMM '08
24. Cuevas R, Laoutaris N, Yang X, Siganos G, Rodriguez P (2011) Deep diving into bittorrent locality, booktitle = Proceedings of IEEE INFOCOM'11
25. Xie H, Yang YR, Krishnamurthy A, Liu Y, Silberschatz A (2010) P4P: provider portal for applications. In: Proceedings of ACM SIGCOMM '08
26. Picconi F, Massoulié L (2009) ISP friend or foe? Making P2P live streaming ISP-aware. In: Proceedings of IEEE ICDCS'09
27. Wittie MP, Pejovic V, Deek L, Almeroth KC, Zhao BY (2010) Exploiting locality of interest in online social networks. In: Proceedings of ACM CoNEXT '10
28. Pujol JM, Erramilli V, Siganos G, Yang X, Laoutaris N, Chhabra P, Rodriguez P (2010) The little engine(s) that could: scaling online social networks. In: Proceedings of ACM SIGCOMM '10
29. Kulshrestha J, Kooti F, Nikravesh A, Gummadi K (2012) Geographic dissection of the Twitter network. In: Proceedings of the international AAAI conference on weblogs and social media, ICWSM '12
30. Scellato S, Mascolo C, Musolesi M, Crowcroft J (2011) Track globally, deliver locally: improving content delivery networks by tracking geographic social cascades. In: Proceedings of the 20th international conference on world wide web, WWW '11, pp 457–466, ACM, New York, NY, USA. doi:10.1145/1963405.1963471
31. Scellato S, Mascolo C, Musolesi M, Latora V (2010) Distance matters: geo-social metrics for online social networks. In: Proceedings of the 3rd conference on online social networks, WOSN '10, pp 8–8, USENIX Association, Berkeley, CA, USA. http://dl.acm.org/citation.cfm?id=1863190.1863198
32. Tang J, Musolesi M, Mascolo C, Latora V (2009) Temporal distance metrics for social network analysis. In: Proceedings of the 2nd ACM workshop on online social networks, WOSN '09, pp 31–36, ACM, New York, NY, USA. doi:10.1145/1592665.1592674