Social Network Analysis and Mining (2023) 13:15
https://doi.org/10.1007/s13278-022-01017-0

ORIGINAL ARTICLE

# Where do migrants and natives belong in a community: a Twitter case study and privacy risk analysis

**Jisu Kim[1] · Francesca Pratesi[2] · Giulio Rossetti[2] · Alina Sîrbu[3] · Fosca Giannotti[4]**

Received: 22 September 2022 / Revised: 11 December 2022 / Accepted: 13 December 2022 / Published online: 29 December 2022
© The Author(s) 2022

## Abstract

Today, many users are actively using Twitter to express their opinions and to share information. Thanks to the availability of the data, researchers have studied behaviours and social networks of these users. International migration studies have also benefited from this social media platform to improve migration statistics. Although diverse types of social networks have been studied so far on Twitter, social networks of migrants and natives have not been studied before. This paper aims to fill this gap by studying characteristics and behaviours of migrants and natives on Twitter. To do so, we perform a general assessment of features including profiles and tweets, and an extensive network analysis on the network. We find that migrants have more followers than friends. They have also tweeted more despite that both of the groups have similar account ages. More interestingly, the assortativity scores showed that users tend to connect based on nationality more than country of residence, and this is more the case for migrants than natives. Furthermore, both natives and migrants tend to connect mostly with natives. The homophilic behaviours of users are also well reflected in the communities that we detected. Our additional privacy risk analysis showed that Twitter data can be safely used without exposing sensitive information of the users, and minimise risk of re-identification, while respecting GDPR.

**Keywords** International migration · Community detection · Social network · Twitter · Privacy risk assessment · GDPR

## 1 Introduction

Twitter is one of the microblogging platforms that attracted many users. Unlike some of the other platforms, Twitter is widely used to communicate in real time and share news among different users (Kwak et al. 2010). On Twitter, users follow other accounts that interest them to receive updates on their messages, called "tweets". Tweets can include photographs, GIFS, videos, hashtags and polls. Among them, hashtags are widely used to facilitate cross-referencing contents. The tweets can also be retweeted by other users who wish to spread the information among their networks. This involves sometimes adding new information or expressing opinion on the information stated. Despite the limit on maximum 280 characters of tweets,[1] users are able to effectively communicate with others.

But above all, Twitter has become a useful resource for research. Twitter data can be accessed freely through an application programming interface (API).[2] On top of this, the geo-tagged tweets are widely used to analyse real-world

Jisu Kim, Francesca Pratesi, Giulio Rossetti, Alina Sîrbu and Fosca Giannotti have contributed equally to this work.

✉ Jisu Kim
kim@demogr.mpg.de

Francesca Pratesi
francesca.pratesi@isti.cnr.it

Giulio Rossetti
giulio.rossetti@isti.cnr.it

Alina Sîrbu
alina.sirbu@unipi.it

Fosca Giannotti
fosca.giannotti@isti.cnr.it

[1] Max Planck Institute for Demographic Research, Rostock, Germany

[2] Italian National Research Council, CNR-ISTI, Pisa, Italy

[3] University of Pisa, Pisa, Italy

[4] Scuola Normale Superiore, Pisa, Italy

[1] https://developer.twitter.com/en/docs/counting-characters.

[2] https://developer.twitter.com/en/docs/twitter-api.

behaviours. One of fields of research that makes use of geo-tagged tweets is migration studies. Typically, migration studies have relied on traditional data such as census, survey and register data. However, provided with alternative data sources to study migration statistics in the recent period, many studies have developed new methodologies to complement traditional data sources (see for instance, Kim et al. 2020; Hawelka et al. 2014; Zagheni et al. 2014; Mazzoli et al. 2020; Sîrbu et al. 2020). While these studies have successfully shown advantages of alternative data sources, how migrants and natives use and interact on social media has not been fully understood. For instance, *what do migrants/natives talk about? To whom migrants/natives connect to? Do migrants/natives have many followers or friends? Who are the most central users amongst them? Where do migrants and natives belong in a community?* These are the questions that we aim to explore in this work through analysing features and the network of Twitter. In doing so, we expect to discover interests of migrants and natives and evidences for social interaction.

As Twitter has become a popular data source for various types of researches, it has also shed light on the importance of data privacy issues (Mao et al. 2011; De Cristofaro et al. 2012; Garcia et al. 2018; Mahoney et al. 2022). Even if this information is included in the terms and conditions of Twitter, many of the users on Twitter are not aware of the fact that their personal information is freely available to researchers.[3] Furthermore, often, the existing studies have performed their analysis on Twitter data without a privacy risk analysis. We aim to overcome this limitation and provide a privacy risk analysis for ours study.

The first aim of this work is to study the characteristics and behaviours of two different communities on Twitter: migrants and natives. We plan to do so through a general assessment of features of individual users from profiles and tweets and an extensive network analysis to understand the structure of the different communities. For this, we identified 4940 migrant users and 46,948 native users across 174 countries of origin and 186 countries of residence using the methodology developed by Kim et al. (2020). For each user, we have their profile information which includes account age, whether the account is a verified account, number of friends, followers and tweets. We also have information extracted from the public tweets which includes language, location (at country level) and hashtags. With these collected data, we explore how each of the communities utilises Twitter and their interests in both the world- and local-level news using the method developed by Kim et al. (2022). Furthermore, we also explore their social links by studying the properties of the mixed network between migrants and natives. We study centrality, assortativity, and community structure of the nodes in the network.

A second aim is to analyse the risk of re-identification on our data from Twitter by performing a risk assessment methodology proposed by Pratesi et al. (2018). We rely on the requirement of the Data Protection Impact Assessment given in the Article 35 of the General Data Protection Regulation[4] (GDPR) and on the definition of the US National Institute of Standards and Technology,[5] where the privacy risk analysis is defined as "a process that helps organisations to analyse and assess privacy risks for individuals arising from the processing of their data. This focus area includes, but is not limited to, risk models, risk assessment methodologies, and approaches to determining privacy risk factors". We first perform the analysis at the tweet level then, secondly, at the user level. This methodology considers a scenario where an *attacker* could gain access to a dataset and by using some prior knowledge about an individual under attack, the *attacker* can re-identify that specific individual in the dataset. By quantifying the different amount of information that an *attacker* can have about the target, we compute the probability of re-identification of each individual described in the data. Our analysis shows that data can be safely used for the purpose of research without exposing sensitive information of the users, and minimising the risk of re-identification, while respecting the GDPR.

The rest of the article is organised as follows: we begin with related works, followed by Sect. 3 on data and the identification strategy for labelling migrants and natives on Twitter. Section 4 focuses on statistics on different features of Twitter, and Sect. 5 deals with analysis of the different networks. Section 6 presents methods and results for privacy risk analysis on Twitter data. We then conclude the paper in Sect. 7.

## 2 Related works

### 2.1 Characteristics of users on Twitter

Many studies exist that analyse different networks on micro-blogging platforms. Twitter is one of the platforms that has been studied extensively as it enables us to collect directed graphs unlike Facebook for instance. We can study various types of relationships defined by either a friendship (followers or friends[6]), conversation threads (tweets and retweets)

[3] https://www.pewresearch.org/internet/2021/11/15/the-behaviors-and-attitudes-of-u-s-adults-on-twitter/pdl_11-15-21_twitter-0_2/.

[4] https://gdpr-info.eu/.

[5] https://www.nist.gov/itl/applied-cybersecurity/privacy-engineering/collaboration-space/focus-areas/risk-assessment.

[6] Followers are users that follow a specific user and friends are users that a specific user follows. https://developer.twitter.com/en/docs/twitter-api/v1/accounts-and-users/follow-search-get-users/overview.

or semantics (tweets and hashtags). Performing network analysis on these allows us to study properties, structures and dynamics of various types of social relationships.

One of the first quantitative studies on topological characteristics of Twitter, and its role in information sharing is Kwak et al. (2010). From this study onward, many have found distinguished characteristics of Twitter's social networks. According to the study, Twitter has a "non-power-law follower distribution, a short effective diameter, and low reciprocity". The study showed that unlike other microblogging platforms that serve as mainly social networking platforms, Twitter acts as a news media platform where users follow others to receive updates on others' tweets. A further study of the power of Twitter in information sharing and role of influencers is Cha et al. (2010). The authors focused on three different types of influence: indegree, retweets and mentions of tweets. They found that receiving many in-links does not produce enough evidence for influence of a user, but the content of tweets created, including the retweets, mentions and topics, matters equally. The same authors extended the work to observe information spreaders on Twitter based on certain properties of the users which led to a natural division into three groups: mass media, grassroots (ordinary users) and evangelists (opinion leaders) (Cha et al. 2012). Furthermore, by looking at the six major topics in 2009 and how these topics circulated, they found different roles played by each group. For example, mass media and evangelists play a major role in spreading new events despite their small presence. On the other hand, grassroots users act as gossip-like spreaders. The grassroots and evangelists are more involved to form social relationships.

Studies that appear in the latter years focused on characteristics on Twitter networks and properties in various scenarios, e.g. political context, social movements, urban mobility and more (see for instance Xiong et al. 2019; Radicioni et al. 2021; Grandjean 2016), implementing various analytic tools such as social network analysis and machine learning algorithms. For instance, Grandjean (2016) studied the network of followers on Twitter in the digital humanities community and showed that linguistic groups are the main drivers to formation of diverse communities. Another example is Bello-Orgaz et al. 2017 where the authors study the discussion communities on vaccination on Twitter. In this work, on top of the social data analysis, they performed a comparative assessment of various group-based community detection algorithms to study how various communities discuss the topic and to analyse how each community is "opining about vaccination". Through the analysis of communities, authors discovered "groups of similar users opining about vaccines". They found the most influential users, described how users interact with each other, and their collective behaviours, showing that the opinions formed on Twitter can influence the decision-making process of vaccination in some cases. Most importantly, through the communities detected, they observed that the influential users are mostly part of the pro-vaccination movement.

Our work contributes to the same line of these works. But unlike any precedent works, here we explore new types of communities that, to the best of our knowledge, have not yet been explored, i.e. migrants and natives. The labelling of migrants and natives enables us to discover distinguished characteristics of these two population groups and discover communities and influential users within the Twitter network.

## 2.2 Migration and social networks

Various definitions of migration exist from international organisations but also from countries depending on their legal grounds and nature of migrants that they receive. For instance, an international migrant is a "a person who moves to a country other than that of his or her usual residence for a period of at least a year",[7] according to the definition set by the United Nations. The Organisation for Economic Co-operation and Development (OECD), on the other hand, defines an international migrant based on the "ground of the place of birth (foreign-born) or of the citizenship (foreigners)." More precise definitions are applied to define other types of migration such as refugees, asylum seekers and internal migrants based on the context in which the migration is decided to take place.[8] In this work we define a migrant as a person who has a residence different from nationality, so the definition closer to that of the OECD. We estimate residence and nationality based on Twitter data.

In this work, we focus on the social networks of international migrants on Twitter. Social networks of migrants play an essential role in facilitating their lives in the destination country. It is one of the sources for information that immigrants have higher barriers to compare to the locals or to the immigrants that have migrated earlier. Being socially connected in the host society can provide job opportunities, informal insurance, social support and more (see for instance, Foster and Rosenzweig 2001; Blumenstock et al. 2019; Bloch et al. 2008; Gërxhani and Kosyakova 2020; Munshi 2003; McKenzie and Rapoport 2010; Granovetter 2018, 1983). While it is difficult for immigrants to be socially integrated with locals immediately, immigrants tend

[7] Recommendations on Statistics of International Migration, Revision 1(p.113). United Nations, 1998.

[8] A refugee is a "person who has fled war, violence, conflict or persecution and has crossed an international border to find safety in another country" according to the definition of United Nations High Commissioner for Refugees (UNHCR). An asylum seeker is "An asylum-seeker is someone whose request for sanctuary has yet to be processed" (.Idem)

to easily get acquainted with other immigrants that have migrated earlier thanks to backgrounds, and language that they share together (Munshi 2003). In this line of research, several studies exist using traditional data sources such as survey, and census data to examine the role of social networks in international migration from various perspectives and outcomes to the society (Comola and Mendola 2015; Munshi 2014; Krishnan and Sciubba 2009; Rauch 1999).

In Munshi (2003), the authors studied the Mexican migrants in the US labour market to study whether the network improves the labour market outcomes for its members. Using a survey data conducted since 1982, they captured the variation of the migration patterns and labour market outcomes over a long period of time. The network here is captured by the "proportion of the sampled individuals who are located at the destination country in any year". This work showed a significant improvement in labour market outcomes due to the social networks among Mexican migrants in the US labour market, highlighting the importance of social networks also in a modern economy.

In the work of Comola and Mendola (2015), authors investigate the internal structure of the Sri Lankan immigrants' social network living in Milan based on a survey conducted by the authors themselves between Dec. 2011 and Feb. 2012. The survey consists of 105 male Sinhalese immigrants, older than 18 years old. From their detailed internal structure of the immigrants' social network, they find heterogeneous and differentiated patterns of within interactions across immigrants according to the network function. They observed that ties have been observed between immigrants that come from nearby localities in Sri Lanka and between immigrants that arrived in similar times or have migrated earlier. Moreover, authors find that material supports mainly are exchanged between family ties while geographical proximity at origin no longer plays a role to this network function. Interestingly, the ties formed with immigrants that arrived earlier only have a significant role in helping the newcomers to find jobs.

The benefits of the social network are not only limited to immigrants themselves but also to the society as a whole. In the work of Gould (1994), he studies the effect of the immigrants' ties (estimated by the size of the immigrants in the destination country) to the bilateral trades between the origin and destination country. He finds that immigrants' ties to their origin country increases bilateral trades with the destination country. The effect is shown to be stronger for exports than for imports. The two main mechanisms for this effect are information and preference effects where the information effect tells us that thanks to the information that immigrants have about their origin country, information barrier between locals and immigrants is reduced. The information includes various factors such as the language, and the local contacts that immigrants have back home. On the other hand, the preference effect tells us that the presence of immigrants in the destination country induces additional demand for good and services produced in the country of origin.

Although existing studies have provided many evidences for significant role of social network for immigrants, most of the studies employ size of the immigrants' diaspora as a proxy for the effects of social networks or have studied a small sample of individual immigrants' network without considering how they interact with the locals. Our work adds new perspective to this line of research by considering also how immigrants interact with the locals (or natives) in addition to observing how they interact with other immigrants on Twitter. Through this process, we provide also a richer characterisation of various social interactions types between immigrants themselves but also among and between natives.

### 2.3 Data privacy issues

One of the important aspects of employing social media data concerns the ethical dimension of processing personal data that includes sensitive information but also the possibility to describe individual users' activities both online and in-person. As shown in the work of Sîrbu et al. (2020), various aspects of migration phases can be studied and it also highlights the legal requirements and constraints of using social media in such case. For example, the authors underline that a solid understanding of ethical and legal values is required to create an actual ethical and legal framework composed of infrastructural, organisational and methodological principles and measures. Only with these premises, we are allowed to make full use of the functionalities and capabilities that big data can offer us while at the same time allowing them to respect fundamental rights and accommodate shared values, such as privacy, security, safety, fairness, equality, human dignity and autonomy (Sîrbu et al. 2020).

Focusing on the privacy aspects only, a number of works dealing with privacy issues in social media data have been published. Privacy in data from social media has been studied from various angles, and several privacy leaks have been identified over the years, from information that can be inferred from hashtags and posts to network structure and metadata information. One of the first examples was (Mao et al. 2011), where authors focused on characterising the nature of tweets' topics, such as vacation and travel plans or tweeting under the influence of alcohol. In this paper, the authors well exemplified the privacy leaks that can occur by looking at the text of tweets, and the information that can be automatically inferred (for example, burglars who may automatically receive alerts about vacation messages or law enforcement that may receive alerts about drunk driving). Another example of leaks due to the content of tweets is given in Keküllüoğlu et al. (2020), where authors studied how protected accounts behave (e.g. like, retweet) relative

to public accounts. They found that protected accounts were more actively "liking" tweets than public accounts. This could be due to the assumption that protected status provides sufficient protection for their accounts. Furthermore, De Cristofaro et al. (2012) proposed a privacy-enhanced variant of Twitter, where cryptographic protocols allow users to tweet hiding their personal information (including tweeter–follower relationships and hashtags, thus, interests) to the provider, thus making current revenue strategies (e.g. targeted advertising) very difficult to realise. Then, in Buccafurri et al. (2015), Buccafurri et al. analysed data from different social networks (i.e. Twitter and Facebook), generalising users behaviour in different social networks and studying the friend overlap in the two networks, with the aim to pair couples of accounts such that the accounts of each pair belong to the same user. Recognised users on different social networks can lead to privacy leaks as there is a possibility that even an anonymised network information can be linked to real identity of the user.

There is a lot of interest in using Twitter data as a proxy for studying mobility behaviour. Indeed, as highlighted in Hu et al. (2021), Huang et al. (2020), Li et al. (2021); Parrish et al. (2020), social media data facilitate access to (near) real-time human mobility in an active, less privacy-concerning manner, compared to the mobility records collected from mobile devices (Hu et al. 2021). Moreover, they are extensive (covering large spatial areas), easily accessible and at low cost (Huang et al. 2020). In Li et al. (2021), authors argued that the less privacy-concerning nature of social media could be attributed to the user sharing settings: for instance, Twitter allows users to determine whether to share content to the public, whether to reveal locations, and what levels of accuracy to be revealed. Even if we believe that often these settings are not exploited at their best by users, it is undeniable that social media data are surely less intrusive since it is the user that actively decides to share some information. Li et al. (2021) also analysed the challenges offered by various types of data, i.e. mobile devices, connected vehicles and social media. For example, mobile phones data have a huge adoption rate w.r.t. connected vehicles, but it offers very poor accuracy in positioning, while for social media data the accuracy of geo-information varies across different platforms, user settings and mobile devices. Parrish et al. (2020) compared the use of social media data with the adoption of cross-sectional survey and census data, also from the point of view of privacy concerns.

Regarding a quantitative analysis of privacy, Garcia et al. (2018) evaluated the predictability of users' location by using only information given by friends of the user that joined Twitter before the user did, i.e. predicting positions for users that had not joined Twitter yet. Finally, in Gao et al. (2019), authors focus on the issues of user identity de-anonymization and location exposure, investigating the effectiveness of geomasking techniques for protecting the geo-privacy of active Twitter users who frequently share geotagged tweets in their home and work locations. By analysing over 38,000 geotagged tweets of 93 active Twitter users in three US cities (Los Angeles, Washington DC, and Madison), the papers analyse a two-dimensional Gaussian masking technique with proper standard deviation settings that is found to be effective to protect user's location privacy while sacrificing (i.e. geomasking) geospatial analytical resolution. Authors proved that their proposed method performs better than the randomization related to small-distance (i.e. within 2 km).

Our work is complementary to these papers since, unlike Garcia et al. (2018), we quantify the risk of being re-identified from past activities (i.e. publishing tweets), while, unlike Gao et al. (2019), our focus is on the temporal side of every single tweet and on the label assigned to the individual as a whole.

## 3 Data and labelling strategy

### 3.1 Data

The dataset used in this work is similar to the one used in Kim et al. (2022). We begin with Twitter data collected by Coletto et al. (2017) to obtain the first seed of users, from which we extract all geo-tagged tweets from August 2015 to October 2015 published from Italy, resulting in a total of 34,160 individual users (that we call first layer users). We then searched for their friends, i.e. other accounts that first layer users are following which added 258,455 users to the dataset (called second layer users). We further augmented our data by scraping also the friends of the 258,455 users. The size of the data grew extensively up to about 60 million users. To ensure sufficient number of geo-tagged tweets, all of these users' 200 most recent tweets were also collected. Although our data collection starts from the data extracted in 2015, as we collected the 200 most recent tweets, the time windows of tweets have been updated to 2018, allowing us to observe frequency of tweets in 2018 to be able to determine the country of residence in the labelling process. To synthesise the dataset, we focus on a subset of these users for whom we have their social network, and which have published geo-located tweets. This results in total of 200,354 users from the first and second layers with some overlaps present among the two layers.

### 3.2 Labelling migrants and natives

The strategy for labelling migrants and natives originates from the work of Kim et al. (2020). It involves assigning a country of nationality $C_n(u)$ and a country of residence $C_r(u)$ to each user $u$, for the year 2018. The definition of a migrant is "a person who has the residence different from the nationality", i.e. $C_n(u) \neq C_r(u)$. The strategy to assign a user's residence requires observing the number of days spent in different countries in 2018 through the time stamps of the tweets. In other words, the country of residence is the location where the user remains most of the time in 2018. To assign nationality, we analyse the tweet locations of the user and user's friends. In this work, we took into account the fact that tweet language was not considered important in defining the nationality as found in the study of Kim et al. (2020). Thus, the language was not considered here as well. By comparing the labels of country of residence and the nationality, we determined whether the user was a migrant or a native in 2018.

Some users could not be labelled since the procedure outlined in Kim et al. (2020) only assigns labels when enough data are available. As a result, we identified nationalities of 197,464 users and the residence 57,299 users. Among them, the total number of users that have both the nationality and residence labels are 51,888. Most importantly, we were able to identify 4940 migrant users and 46,948 natives from our Twitter dataset. In total, we have identified 163 countries of nationalities for natives. The most present countries are the USA, Italy, Great Britain and Spain in terms of nationality. This is due to several factors. First because Twitter's main users are from the USA. Second, we have large number of Italian nationalities present due to the fact that we initially selected the users whose geo-tags were from Italy. Overall, we have identified 144 countries of nationalities and 169 countries of residences for the migrants. In terms of migration patterns, it is interesting to also remark from our data that the US and UK have significant number of in and outgoing links. In addition, France and Germany have mainly in-coming links.

Here, we emphasise that through our labelling process we do not intend to reflect a global view of the world's migration patterns but simply what is demonstrated through our dataset. However as it is also shown in the work of Kim et al. (2022), the predicted data correlate fairly with official data when looking at countries separately. For instance, when comparing predicted data with Italian emigration data of AIRE,[9] we observed a correlation coefficient of 0.831 for European countries and 0.56 for non-European countries. When compared with Eurostat data on European countries, the correlation coefficient was 0.762. This provides us the confidence to employ this dataset to analyse characteristics of different communities through Twitter.

## 4 Twitter features

In this section we look at the way migrants and natives employ Twitter to connect with friends and produce and consume information.

### 4.1 Origin and destination attachment index

A first analysis concentrates on the types of information that users share, from the point of view of the country where the topics are discussed. In particular, we compute two indices developed by Kim et al. (2022): Origin Attachment (*OA*) and Destination Attachment (*DA*), which describe how much users concentrate on topics from the nationality and residence country, respectively. We compute the two indices for both migrants and natives; obviously, for natives the residence and nationality are equal and thus the two indices coincide.

To compute *OA* and *DA*, we first assign nationalities to hashtags by considering the most frequent country of residence of natives using the hashtags. A few hashtags are not labelled, if their distribution across countries is heterogeneous (as measured by the entropy of the distribution). The *OA* is then computed for each user as the proportion of hashtags specific to the country of nationality. Similarly, the *DA* is the proportion of hashtags specific to the country of residence. Thus, the *OA* index measures how much a user is interested in what is happening in his/her country of nationality and the *DA* index reflects how much a user is interested in what is happening in his/her country of residence.

As shown in Fig. 1, the indices clearly behave differently for the two groups: migrants and natives. Similar to Kim et al. (2022), we observe that migrants have, on average, very low level of *DA* and *OA*. When looking at natives, this index distribution is wider and has an average of 0.447 which is surely higher than the average of migrants. Without a doubt, this shows that natives are more attached to topics of their countries, while migrants are generally less involved in discussing the topics, both for the origin and destination country. However, we observe that a few migrant users do have large *OA* and *DA* showing different cultural integration patterns, as detailed in Kim et al. (2022). At the same time, some natives show low interest in the country's topics, which could be due to interest in world-level topics rather local-level topics.

[9] Anagrafe degli italiani residenti all'estero (AIRE) is the Italian register data.

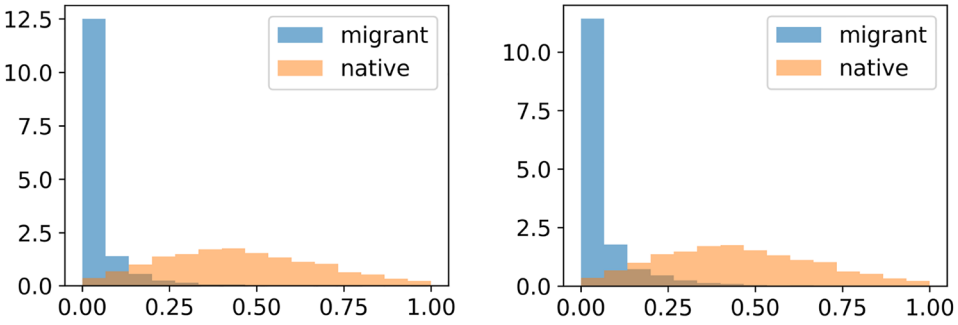


**Fig. 1** Distribution of DA and OA for migrants (in blue) and natives (in orange) (Color figure online)

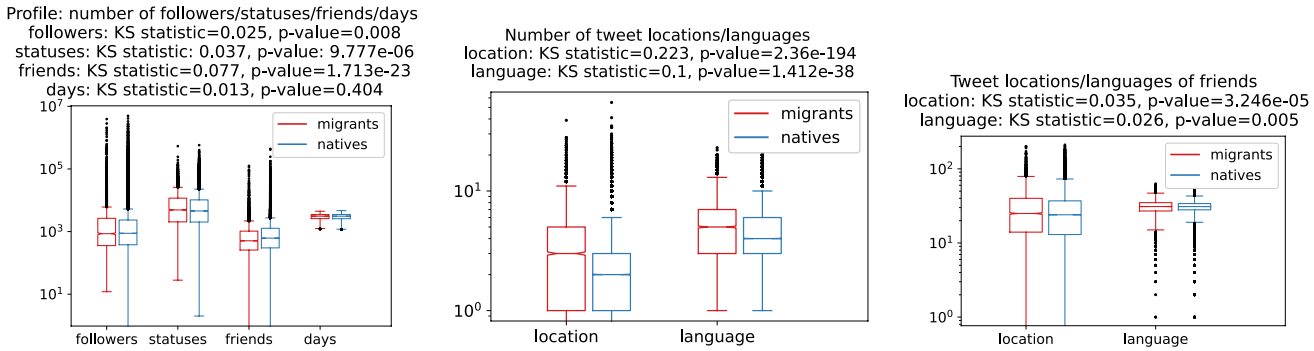


**Fig. 2** Left: Distributions of profile features: number of followers, tweets published (statuses) and friends and number of days since the account was created until 2018, respectively. Centre: Distribution of tweet locations and languages. Right: Distribution of tweet locations and languages of friends

## 4.2 Profile information

Can we find any distinctive characteristics of migrants and natives from the profiles of users? Here, we look at public information provided by the users themselves on their profiles. We examine the distribution of profile information and perform Kolmogorov–Smirnov (KS) test to compare the distributions for migrants and natives. On the profile, various information is declared by the users themselves such as the joined date, location, bio, birthday and more. We begin by looking at the age of the Twitter accounts from the moment they created their accounts till 2018, as shown in Fig. 2. We observe that migrants and natives have similar shape of distributions, providing information that there is no earlier or later arrival of one group or another on Twitter. The KS test with high $p$-value of 0.404 also confirms that the two distributions are indeed very similar. The other criteria we study show some differences. First, we generally observe that natives have slightly more friends than migrants. On average, migrants follow about 1160 friends and 1291 friends for the natives. We can also see from Fig. 2 that the range of this number is much wider for the natives, ranging from 0 to maximum of 436,299, whereas for the migrants, this range ends at 125,315. The KS test yields a $p$-value of $1.713e^{-23}$, confirming that the two distributions are different. Secondly, we observe that the migrants have a larger number of followers. On average, migrants have 10,972 followers versus 7022 followers for natives (KS $p$-value of 0.008). This tells us that there are more users on average that are waiting to get updates on migrant users tweets. Interestingly, when it comes to the number of tweets (statuses) that users have ever tweeted since the account was created, the number is about 9% higher for the migrants than the natives: average values of 9836 for migrants and 9016 for natives, $p$-value of $9.777e^{-06}$.

We also look at the number of accounts that are classified as verified accounts. The verified accounts are usually well-known people such as celebrities, politicians, writers, or directors and so on. Indeed when looking at the proportion of verified accounts, we observe that this proportion is higher among migrants than natives which partly explain also the higher number of followers and tweets for this group. To be more specific, 5% of the users' accounts are verified accounts among migrants and 3.7% of the accounts are verified accounts among natives.

## 4.3 Tweets

Tweets also provide useful information about user behaviour. We are interested in the locations (country level) and

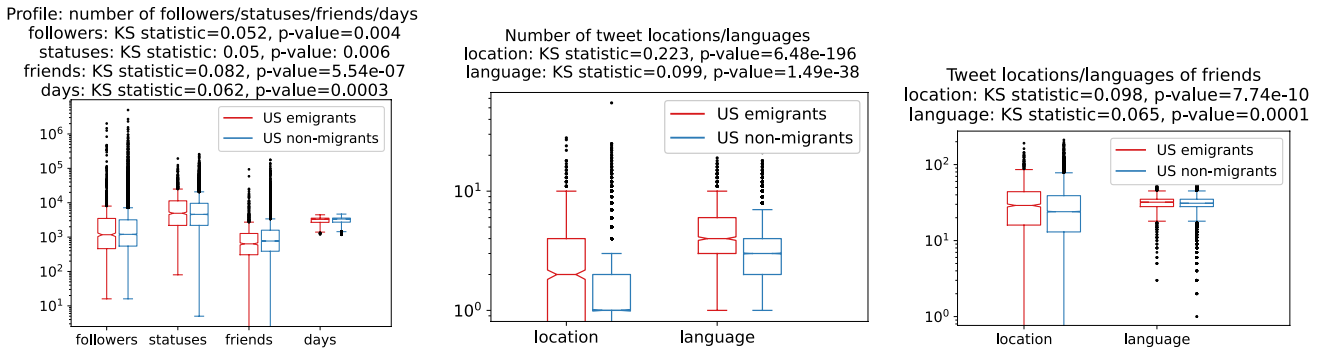


**Fig. 3** Left: Distributions of profile features: number of followers, tweets published (statuses), and friends and number of days since the account was created until 2018, respectively, for the US emigrants and non-migrants. Centre: Distribution of tweet locations and languages for the US emigrants and non-migrants. Right: Distribution of tweet locations and languages of friends for the US emigrants and non-migrants

languages a user employs on Twitter. Hence, we look at the number of languages and locations that appear in the users' 200 most recent tweets and computed also the KS statistics to compare the differences between the distributions of migrants and natives. As shown in Fig. 2 on the left, we note that migrants tweet in a wider variety of languages and locations. The two distributions for migrants and natives are different from each other as the KS tests show low $p$-values; $2.36e^{-194}$ for location and $1.412e^{-38}$ for language.

Since we possess network information, we also studied the tweet language and location information for a user's friends. In Fig. 2 on the right, the two distributions show smaller differences among natives and migrants, compared to figure on the left. However, the $p$-value of the KS test tells us that the distributions are indeed different from one another, where the $p$-value for location and language distribution for migrants and natives is $3.246e^{-05}$ and 0.005, respectively. Although the differences are small, we observe that the friends of migrants tweet in more numerous locations than those of natives, with average of 29.6 for migrants and 27.4 for natives. However, although the two distributions are different from each other from the KS $p$-value, the actual difference between average values is very small in the case of the number of languages of friends. In fact, the average for migrants is 30.22 and 30.43 for natives. These numbers indicate that the migrants have travelled in more various places and hence write in diverse languages than the natives. The friends of migrants tend to have travelled more also. However, no large differences were observed for the number of languages that friends can write in for both migrants and natives.

#### 4.3.1 Case study of the US emigrants and non-migrants

In this section, we investigate whether the differences between migrants and natives that we observed so far have been due to biased samples of migrants. To be more specific, it is possible that the differences observed previously have been due to the fact that migrants on Twitter are highly skilled and professional than the natives. Hence, the differences with natives could be due to this. In order to address this, we compare the characteristics of the US emigrants that have left the USA to live abroad, with those of US non-migrants, i.e. individual users that do not change their usual country of residence. We therefore replicate Fig. 2 to compare these two groups. We compare these two groups of population for two main reasons. First of all, our data cover a significant number of in and out-going links from the USA. Secondly, it has been studied that American emigrants have high level of education and are working in education or are running their own business.[10] These characteristics of American emigrants make it an ideal population group to be compared with the general population. The comparison of these two groups of population will enable us to verify whether the differences observed previously have been due to the fact that migrants on Twitter are genuinely different from natives or only due to the selection bias.

As shown in Fig. 3, we observe patterns similar to those on the general population. To be more precise, we generally observe that non-migrants have slightly more friends than migrants. On average, non-migrants have 1588 friends, whereas emigrants have 1325 friends. We also see that the range of this number is much wider for the non-migrants ranging from 0 to maximum of 177,471. On the other hand, we observe that emigrants have a larger number of followers with an average number of 12,733 followers for emigrants and 7164 for non-migrants. When it comes to the number of tweets that users have ever tweeted since the account was

[10] https://www.internations.org/expat-insider/2021/us-americans-working-abroad-40180.

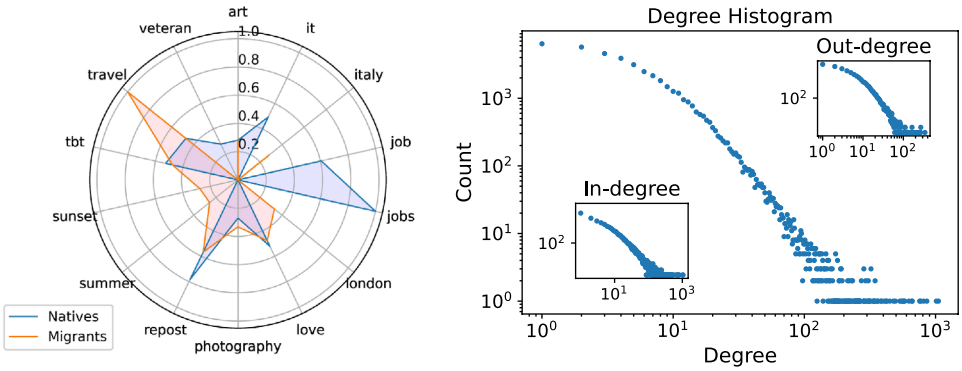


**Fig. 4** Left: Top 14 hashtags used by migrants and natives. Right: Degree distribution of the network

created, on average, the number is about 9525 for the emigrants and 8392 for non-migrants. In addition, we see that emigrants have tweeted in a wider variety of locations and languages than non-migrants. Here the difference is even larger than before. We also analysed the number of locations and languages spoken by alter users, and we observe the difference equally as before where alters of emigrants speak and tweet in more numerous locations than those of non-migrants. Friends of emigrants tweet in minimum of three different languages, whereas friends of non-migrants tweet in one language only. Lastly, the number of verified accounts among emigrants is also higher compared to non-migrants as well. The observed characteristic differences between the US emigrants and non-migrants re-confirm us that migrants have higher popularity level with higher number of followers, tweets and verified accounts for this group. The KS tests all yield significant $p$-values confirming that the compared distributions are different.

#### 4.3.2 Popular hashtags

What were the most popular hashtags used by natives and migrants in 2018? In Fig. 4 we display the top 10 hashtags used by the two communities, together with the number of tweets using those hashtags, scaled to [0, 1]. We observe that natives and migrants share some common interests, but they also have differences. For instance, some of the common hashtags between natives and migrants are #tbt, #love and #art. Other hashtags such as #travel and #repost are in the top list, but the usage of these hashtags is much higher in one of the groups than the other. For instance, the hashtag #travel is much more used by migrants than the natives. This is interesting because the number of tweet locations of migrants also reflects their tendency to travel, more than natives. Followed by the hashtag #travel, migrants also used other hashtags such as #sunset, #photography, #summer and hashtags for countries which show their interests in travelling. On the other hand, natives are more focused on hashtags such as #job, #jobs and #veteran.

## 5 Network analysis

In this section, we perform social network analysis on the social graph of our users to examine the relationships between and within the different communities, i.e. migrants, and natives. Initially, our network consisted of 45,348 nodes and 232,000 edges. We however focus on the giant component of the network which consists of 44,582 nodes and 231,372 edges. Each node represents either a migrant or a native, and the edges are directed and represent friendship on Twitter (in other words, our source nodes are following the target nodes). Since we have migrants and natives labels, our network allows us to study the relationship between migrants and natives.

### 5.1 Properties of the network

In this section, we start by looking at density, reciprocity and shortest path length for the network and then study node centrality including degree distribution. The average density score of our network tells us that on average each node is connected to other 5.2 nodes. The reciprocity coefficient is low and indicates that only 23.8% of our nodes are mutually linked. This is normal on Twitter as most of the users follow celebrities, but the other way around does not happen in many cases. Within the network, the average shortest path length is 2.42, which means we need on average almost 3 hops to receive information from one node to another.

We also compute 7 measures of centrality. The measures include all-, in- and out-Degree (Fig. 4) plus Closeness, Betweenness, Pagerank and Eigenvector centrality measures. As shown in Fig. 4, the degree distribution follows a power-law distribution with alpha equal to 2.9. This means that a minority of the nodes is highly connected to the rest of the nodes. As for the rest of the centrality measures, we observe that most of the users have low centrality, while a small number of users show higher centrality values. This is true for all measures, however for closeness, the number of users who show higher centrality is larger than for the other measures. This means that many users are well-embedded in

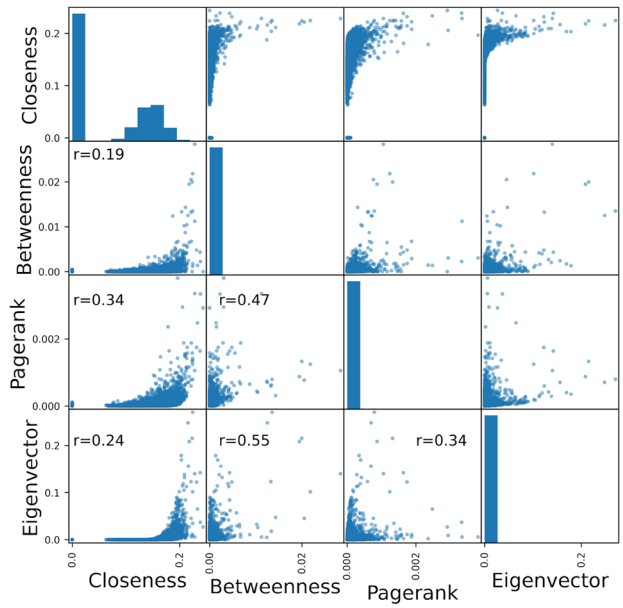


**Fig. 5** Correlation between different centrality measures for network. We computed Closeness, Betweenness, Pagerank and Eigenvector measures, respectively

the core of the network, and are in a good position to receive information. We also compute the correlation between different centrality measures as shown in Fig. 5. First of all, we observe a positive relationship among all measures, which is expected, as it means that users who are central from one point of view are also central from another. The Betweenness and Eigenvector centrality measures correlate the most ($r = 0.55$). This tells us that users that serve as a bridge between two parts of graphs are also likely to be the most influential user in the network. On the other hand, Betweenness and Closeness centrality measures have the lowest correlation with $r = 0.19$. However, the scatterplot shows that those few users who have larger Betweenness also have a large Closeness. The low correlation is determined by the fact that a large majority of users show almost null Betweenness; however, Closeness is heterogeneous among this group. A similar observation can be made for the relation between Closeness on one side and Pagerank and Eigenvector centrality on the other: high Pagerank and Eigenvector centralities always correspond to high Closeness; however, for users with low Pagerank and Eigenvector centrality the Closeness values vary.

When checking the labels, in terms of migrant or native, of the most central users, we see that in general these are mostly natives. To be more specific, we observe that among the top 8–10 users are natives. In other words, most of the nodes have majority of in- and out-going links directed to natives' accounts. This is somewhat expected since in our network only 10% of users are migrants. However, we note that a migrant user is always in the top 3 in Closeness, Pagerank and Eigenvector centrality measures. This tells us that this migrant user has a crucial influence over the network around itself but also beyond its connections.

## 5.2 Assortativity analysis

We now focus on measuring assortativity of nodes by different attributes of individuals, i.e. migrants or natives, country of residence and country of nationality. Assortativity tells us whether the network connections correlate in any way with the given node attributes (Newman 2002). In other words, it tells us whether the nodes in the network tend to connect with other similar nodes. It typically ranges between $-1$ and 1. A value of 1 means nodes always connect with nodes with the same attributes, i.e. full homophily, while $-1$ means nodes tend to connect with nodes with different attributes. In our case this analysis allows us to infer whether and in what measure the network topology follows the nationality or residence of the users, or whether the migrant/native status is relevant when building online social links.

We begin with global assortativity measures, which give one assortativity score for the entire network. First, the degree assortativity coefficient of $-0.054$ shows no particular homophily behaviour from the point of view of the node degree. That means high-degree nodes do not link with other high-degree nodes. However, when we measure the assortativity by different attributes, we observe different results. When looking at the coefficient by the country of residence, the score of 0.54 shows a very good homophily level. The score improves slightly when we examine the behaviour through the attributes of country of nationality (0.6). These values tell us that nodes tend to follow other nodes that share same country of residence and country of nationality, with a stronger effect for the latter. However, when looking at the coefficient by the migrant/native label, we observe no particular correlation (0.033).

The global assortativity scores are susceptible to be influenced by the size of the data and the imbalance in labels, which is our case especially for the migrant/native labels. Therefore we continue to examine the assortativity at local level, allowing us to overcome the possible issues at global level. We thus compute the scores based on an extension of Newman's assortativity introduced by Rossetti et al. (2021), called *conformity*. In Fig. 6 we show the distribution of node-level conformity of migrants and natives, for the three attributes (nationality, residence and migrant/native label). We observe different behaviour patterns for migrants and natives. Specifically, we see that migrants tend to display lower homophily compared to natives, when looking at the conformity of nodes by country of residence. This tells us that migrant users tend to consider less the country of residence when following other users. Instead, most natives tend

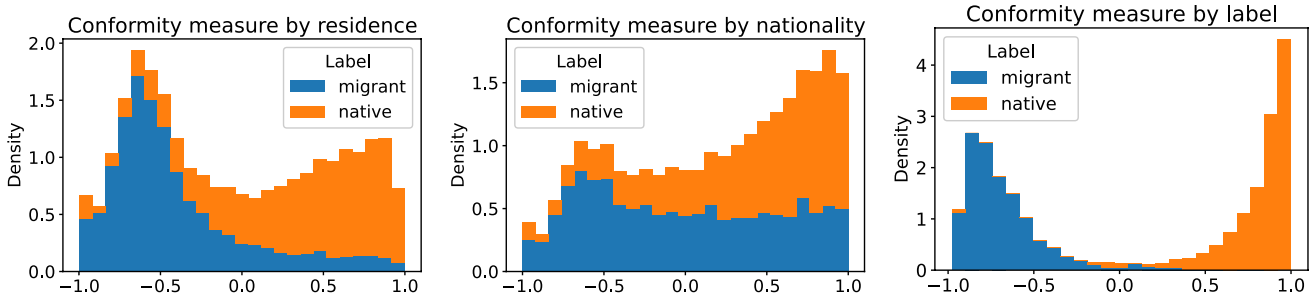


**Fig. 6** Stacked histogram of conformity measures: from left, we have conformity measure by residence, by nationality and by migrant/native label. Please note that the histograms are stacked; therefore, there is no overlap between the plot bars. Blue indicates migrants and orange indicates natives

to connect with users residing in the same country. When looking at nationality, this effect is less pronounced. While natives continue to display generally high homophily, with a small proportion of users with low values, migrants show a flatter distribution compared to the nationality. Again, a large part of migrants show low homophily; however, a consistent fraction of migrant users shows higher nationality homophily, as opposed to what we saw for the residence. This confirms what we observed at global level: there is a stronger tendency to follow nationality labels when creating social links. As for the conformity of nodes by migrant/native labels, we observe that migrants and natives clearly have distinctive behaviours. While natives tend to form connections with other natives, migrants tend to connect with natives as well, resulting in negative conformity values for migrant users. The observed values could also be due to the fact that migrants are only about 10% of our users so naturally many friends will be natives (from either residence, nationality or other country). This result is different from what we observed at global level and confirms that the global conformity score was influenced by the size of the data and the imbalance in labels.

## 5.3 Community detection

Since we are interested in where migrants and natives belong in a community on Twitter, we performed an unsupervised analysis using three different community detection algorithms, focusing on alternative topological/semantic characteristics of our data. The selected algorithms are: Eva (Citraro and Rossetti 2020), Leiden (Traag et al. 2019) and Infomap (Rosvall and Bergstrom 2008). The first, an extension of the well-known Louvain algorithm (Blondel et al. 2008), is designed to identify a highly modular partition (e.g. a partition whose communities are internally densely connected) composed by nodes sharing the same set of labels—thus guaranteeing profile homogeneity within each cluster. Leiden, conversely, has been proposed to address some limitations of the Louvain algorithm: as its predecessor, it is tailored to maximise partition modularity; however, conversely, from Eva, it does not provide any guarantee of semantic coherence of the identified clusters. Finally, Infomap searches for communities that minimise the description length of the partition. In doing so, it implicitly optimises the partition conductance identifying clusters that are well-separated from each other. The selected algorithms allow us to perform our analysis assuming different community definitions, providing a multidimensional perspective on the mesoscale network structure and its relation to the studied phenomenon. It is worth noting that, since community detection is an ill-posed problem, each possible algorithmic choice to address such a task is arbitrary: therefore, we aim to provide a multifaceted descriptive analysis without making claims on which have to be considered the optimal partition, focusing on three different approaches.

For each algorithm's induced clustering, we compute entropy scores for each attributes we have in the network (i.e. nationality, residence and migrant/native labels). Specifically, for each community $c$ we define a dictionary where we store $P_c$ the distribution of the nationalities (or residence, or migrant/native labels) of the users belonging in the community $c$. Hence, $P_c$ is a vector where for each community $c$ we have $P_c(n)$, the fraction of users with nationality $n$. Provided with this probability distribution, we compute the normalised entropy for each community following Eq. 1, where $|P_c(n)|$ is the cardinality of the dictionary $P_c(n)$, i.e. the number of nationality countries (or residence, or migrant/native labels) in each community.

$$H(c) = \frac{-\sum_n P_c(n) \log P_c(n)}{\log(|P_c(n)|)} \tag{1}$$

### 5.3.1 Eva

For Eva, we first iterate through $\alpha$ values from zero to one to find an optimal value. The parameter $\alpha$ controls the modularity, i.e. the separation into clusters that are tightly connected

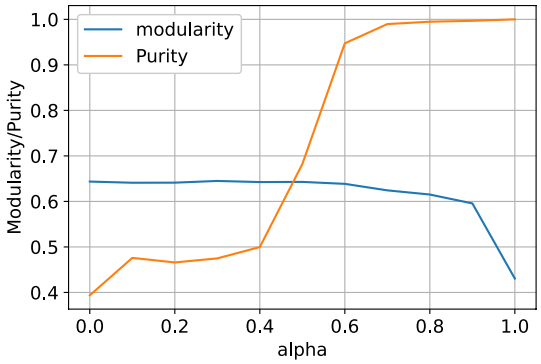


**Fig. 7** Modularity-Purity as a function of $\alpha$: The blue line represents modularity scores over different values of $\alpha$, and orange line represents purity scores over different values of $\alpha$ (Color figure online)

and well separated, and the purity, i.e. the composition of each module in terms of node labels. For instance, a low value of $\alpha$ would generate high modularity and low purity. As shown in Fig. 7, the intersection of the two criteria is around $\alpha = 0.5$. The number of communities increases suddenly when $\alpha$ is greater or equal to 0.6, varying from 78 when $\alpha = 0.5$ to 439 communities when $\alpha = 0.6$. The modularity criterion, however, does not vary much along different $\alpha$ values. Hence, we set the parameter to 0.5 to favour both criteria.

With the defined parameter for Eva, we obtained a total of 78 communities, with the top 10 covering about 72.42% of the users as shown in Fig. 12. Looking at the entropy distributions for three different labels in the first row of Fig. 8, we note that patterns of the distributions are very similar. The $p$-values of KS tests confirm that indeed the three distributions do not have significant differences ($p$-value of 0.91 when comparing nationality and residence labels, 0.97 when comparing residence and migrant/native labels and 0.81 when comparing nationality and migrant/native labels). Many of the communities have entropy scores of 0 which tells us that they are homogeneous, containing one majority country label. These communities are, however, mostly small, with less than 10 users. The average entropy score for each label is 0.25 for nationality label, 0.28 for residence label and 0.26 for migrant/native label.

In Table 2, we show the composition of the top 10 largest Eva communities, by showing pairs of the top 3 most frequent country labels of nationality and residence. Table 1 shows also the proportion of migrants and natives in each

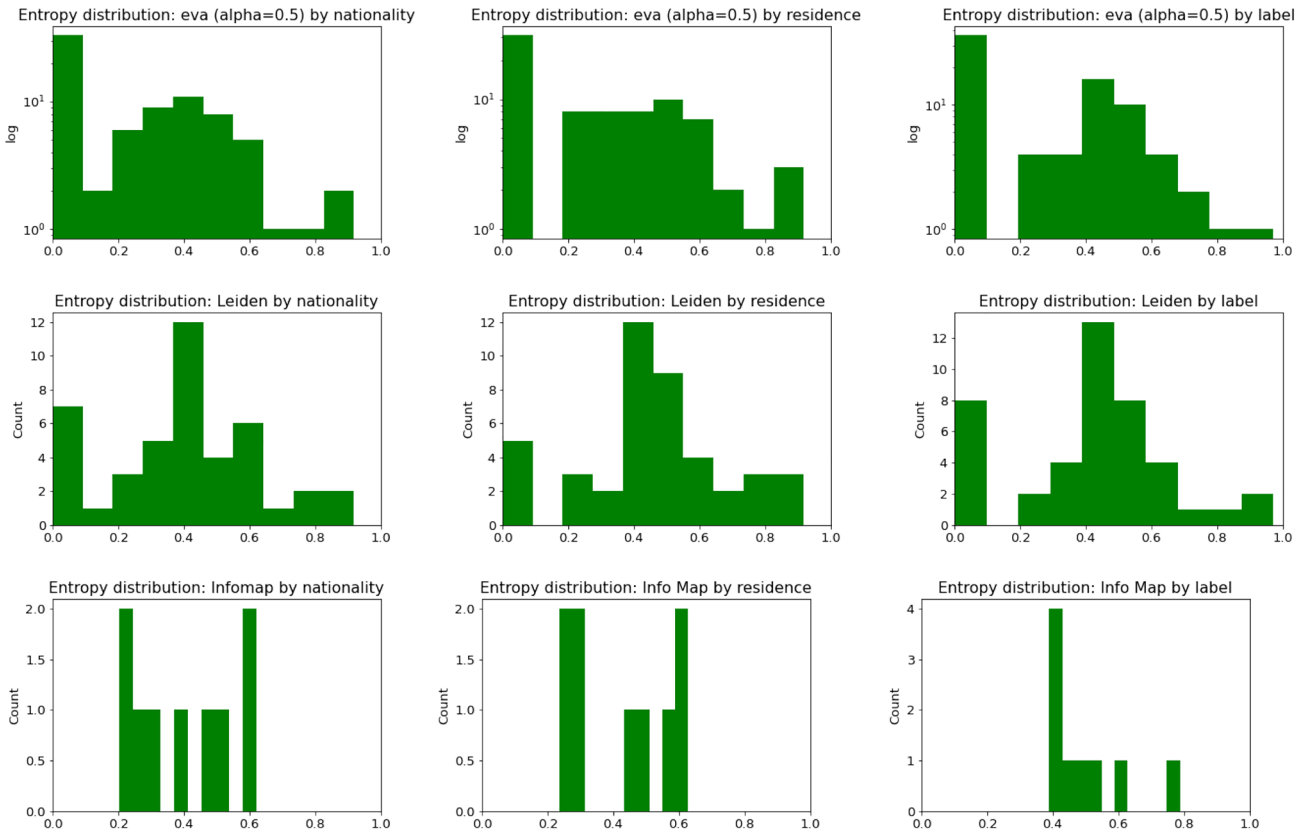


**Fig. 8** Entropy distribution for three labels: *nationality, residence, migrant/native label* for Eva, Leiden and Infomap community detection algorithms

**Table 1** Proportion of migrants and natives in each community for Eva, Leiden and Infomap community detection algorithms

| | Eva | | Leiden | | Infomap | |
|---|---|---|---|---|---|---|
| | Migrants (%) | Natives (%) | Migrants (%) | Natives (%) | Migrants (%) | Natives (%) |
| 1 | 12.08 | 13.66 | 10.57 | 12.73 | 64.89 | 66.1 |
| 2 | 8.45 | 9.26 | 9.72 | 10.74 | 11.52 | 13.12 |
| 3 | 9.28 | 8.84 | 9.55 | 9.96 | 7.99 | 6.29 |
| 4 | 8.96 | 7.47 | 10.62 | 8.63 | 3.89 | 4.12 |
| 5 | 3.55 | 7.74 | 6.79 | 7.55 | 3.07 | 3.77 |
| 6 | 7.23 | 6.83 | 4.63 | 7.74 | 2.43 | 2.91 |
| 7 | 5.24 | 5.38 | 8.45 | 6.9 | 2.07 | 1.84 |
| 8 | 6.04 | 5.06 | 3.75 | 4.54 | 1.78 | 1.06 |
| 9 | 4.46 | 4.03 | 3.53 | 4.12 | 2.34 | 0.77 |
| 10 | 3.7 | 4.22 | 3.04 | 3.83 | | |

community, over the total number of migrants/natives in our data. We observe that all of the nationality and residence country labels are the same which means that natives are the majority in each community. Most of the users are coming from Italy, USA, Great Britain, as also observed previously. The largest community contains 13.52% of the nodes. This community is quite homogeneous in terms of nationality and residence labels as the entropy scores are relatively low (0.2 for nationality label and 0.23 for residence label). In the same table, we also observe that the majority of users in this community is composed of Italian natives (80.17%), followed by small percentage of Americans (3.33%) and English (1.81%). In general, most communities have a strong majority group (over 50% of the users), with the exception of cluster 8, that is as mix of Mexico, US and Colombia users. From the second to fourth largest communities, we observe that the USA and Great Britain labels are always the most frequent country labels. However different from the largest community, the proportions of these country labels are lower, signalling that other country labels are also present in these communities. Indeed, the entropy scores are higher here (0.34, 0.45, 0.41, respectively, for nationality labels and 0.38, 0.47 0.45, respectively, for residence labels). In part, this is due to the proportion of migrants in these communities where for instance, the third largest community includes 9.28% of migrants from the data as shown in Table 1. Note also that the entropy scores are slightly higher for residence labels.

In Table 3, we observe the top three migrant groups in each community (nationality and residence labels). We see that, in terms of migrants, the first community contains mainly Italian natives residing in the USA. The second group are Italians living in Great Britain, followed by American natives residing in Italy. In other communities, we also remark that migrants mostly share the same majority of nationality labels as the one that we observed in Table 2 for natives. This can be referred back to the findings from the previous section that the users tend to connect with other users that share the same nationality more than the country of residence.

#### 5.3.2 Leiden

Through the Leiden algorithm we identified 43 communities, with the top 10 covering about 76.22% of the users as shown in Fig. 13. Observing the second row of Fig. 8, we note, similar to Eva, that the three distributions are very similar to each other. The KS tests yield a $p$-value of 0.62 when comparing nationality label to residence label, 0.8 when comparing residence label to migrant/native label and 0.94 when comparing nationality to migrant/native label, confirming that the three distributions are identical. However, compared to Eva, the average of entropy values is higher here (0.4 for the nationality label, 0.46 for the residence label and 0.41 for the migrant/native label). This is due to the fact that Leiden has captured a smaller number of communities compared to Eva. These communities are larger and probably have grouped together smaller communities that were present in Eva.

Despite the higher entropy scores, the most frequent country labels are quite similar to Eva as observed in Table 2. Specifically, Leiden also shows that Italy, USA and Great Britain are the most frequent country labels. We also observe that the main group in the first five communities is identical with Eva. Also, as seen with Eva, the largest community is the most homogeneous community, composed of mainly Italian natives and a small proportion of natives from the USA and Great Britain. From the second to the sixth communities, we observe relatively lower proportions of American natives compared to Eva and larger proportions of the second most frequent country labels. For instance, in the fourth community, we observe that 12.17% of users are from Germany. For this community the entropy value is high (0.49), indicating heterogeneity of country labels. This is also due to the fact that this community also contains most of the migrants (10.62%) from the data as shown in

**Table 2** List of nationality (Nat) and residence (Res) labels of top 10 largest communities for Eva, Leiden and Infomap community detection algorithms

| | Eva Nat:Res (% within the community) | Leiden Nat:Res (%) | Infomap Nat:Res (%) |
|---|---|---|---|
| (1) Eva: 13.52% | IT: IT (80.17) | IT: IT (82.3) | US: US (43.68) |
| Leiden: 12.53% | US: US (3.33) | US: US (2.88) | GB: GB (12.85) |
| Infomap: 65.23% | GB: GB (1.81) | GB: GB (1.58) | CA: CA (4.1) |
| (2) Eva: 9.18% | US: US (65.56) | US: US (54.35) | IT: IT (77.79) |
| Leiden: 10.65% | GB: GB (4.54) | CA: CA (11.1) | US: US (4.43) |
| Infomap: 13.25% | IT: IT (3.37) | GB: GB (6.95) | GB: GB (2.02) |
| (3) Eva: 8.88% | US: US (52.5) | US: US (46.82) | MX: MX (24.78) |
| Leiden: 9.93% | GB: GB (7.86) | GB: GB (27.66) | CL: CL (13.03) |
| Infomap: 10.47% | CA: CA (4.8) | IT: IT (4.72) | US: US (12.13) |
| (4) Eva: 7.61% | US: US (54.97) | US: US (42.18) | ES: ES (74.69) |
| Leiden: 8.81% | GB: GB (9.52) | DE: DE (12.17) | IT: IT (2.73) |
| Infomap: 3.75% | AU: AU (5.54) | GB: GB (11.58) | US: US (2.57) |
| (5) Eva: 7.36% | US: US (77.77) | US: US (59.3) | BR: BR (75.91) |
| Leiden: 7.49% | CA: CA (3.35) | GB: GB (8.21) | US: US (5.39) |
| Infomap: 2.85% | GB: GB (2.5) | IT: IT (3.87) | IT: IT (2.18) |
| (6) Eva: 6.87% | GB: GB (67.86) | US: US (46.45) | TR: TR (82.55) |
| Leiden: 7.45% | US: US (6.35) | GB: GB (21.94) | US: US (2.27) |
| Infomap: 1.85% | IE: IE (3.27) | AU: AU (9.15) | GB: GB (1.25) |
| (7) Eva: 5.37% | US: US (54.01) | MX: MX (23.94) | ID: ID (50.54) |
| Leiden: 7.05% | GB: GB (17.25) | US: US (13.91) | MY: MY (25.63) |
| Infomap: 1.35% | CA: CA (5.89) | CL: CL (12.03) | SG: SG (6.85) |
| (8) Eva: 5.15% | MX: MX (30.68) | US: US (68.36) | RU: RU (51.5) |
| Leiden: 4.47% | US: US (10.49) | GB: GB (7.48) | UA: UA (6.69) |
| Infomap: 1.2% | CO: CO (9.97) | IT: IT (3.72) | BY: BY (6.19) |
| (9) Eva: 4.32% | ES: ES (72.52) | ES: ES (78.41) | KW: KW (26.72) |
| Leiden: 4.08% | US: US (3.43) | US: US (2.2) | SA: SA (15.93) |
| Infomap: 0.05% | MX: MX (2.03) | MX: MX (1.65) | AE: AE (15.2) |
| (10) Eva: 4.17% | BR: BR (73.75) | BR: BR (78.96) | |
| Leiden: 3.76% | US: US (7.1) | US: US (6.02) | |
| | IT: IT (2.58) | BR: US (1.49) | |

**Table 3** List of nationality (Nat) and residence (Res) labels of migrants in top 10 largest communities for Eva, Leiden and Infomap community detection algorithms

| | Eva Nat:Res (% within the community) | Leiden Nat:Res (%) | Infomap Nat:Res (%) |
|---|---|---|---|
| (1) Eva: 13.52% | IT:US (0.7) | IT: US (0.68) | US:GB (0.39) |
| Leiden: 12.53% | IT:GB (0.61) | IT: GB (0.66) | US:CA (0.27) |
| Infomap: 65.23% | US:IT (0.48) | IT: FR (0.41) | GB:US (0.24) |
| (2) Eva: 9.18% | US:GB (0.54) | CA: US (0.53) | IT:US (0.73) |
| Leiden: 10.65% | US:FR (0.24) | US: CA (0.36) | IT:GB (0.57) |
| Infomap: 13.25% | US:IT (0.24) | US: MX (0.27) | IT:FR (0.41) |
| (3) Eva: 8.88% | US:GB (0.38) | US: CA (0.23) | MX:US (0.9) |
| Leiden: 9.93% | CA:US (0.33) | GB:ES (0.23) | US:MX (0.63) |
| Infomap: 10.47% | US:CA (0.3) | GB:FR (0.2) | US:CO (0.31) |
| (4) Eva: 7.61% | US:GB (0.62) | US: GB (0.59) | ES:US (0.55) |
| Leiden: 8.81% | US:CA (0.62) | US:CA (0.55) | US:ES (0.49) |
| Infomap: 3.75% | CA:US (0.355) | US:DE (0.38) | ES:IT (0.44) |
| (5) Eva: 7.36% | US:IT (0.34) | US:GB (0.45) | BR:US (1.57) |
| Leiden: 7.49% | US:MX (0.24) | US:ES (0.21) | US:BR (1.03) |
| Infomap: 2.85% | US:GB (0.21) | GB:US (0.21) | BR:US (1.49) |
| (6) Eva: 6.87% | US:GB (0.59) | US:GB (0.27) | TR:US (1.02) |
| Leiden: 7.45% | GB:IT (0.46) | US:IT (0.24) | TR: ES (0.47) |
| Infomap: 1.85% | GB:FR (0.39) | US:AU (0.21) | TR: DE (0.39) |
| (7) Eva: 5.37% | GB:US (0.46) | MX: US (0.76) | ID:US (0.6) |
| Leiden: 7.05% | US:FR (0.33) | US: MX (0.51) | ID:JP (0.48) |
| Infomap: 1.35% | US:IT (0.29) | US:CO (0.35) | ID:AU (0.48) |
| (8) Eva: 5.15% | MX:US (1.04) | US: IT (0.35) | RU:ES (0.6) |
| Leiden: 4.47% | US:MX (0.65) | US: AU (0.3) | RU:US (0.6) |
| Infomap: 1.2% | CO:US (0.39) | US:CA (0.3) | RU:IT (0.6) |
| (9) Eva: 4.32% | ES:IT (0.47) | ES:US (0.49) | KW:US (1.72) |
| Leiden: 4.08% | ES:GB (0.32) | ES:IT (0.44) | KW:JO (1.47) |
| Infomap: 0.05% | ES:US (0.32) | ES:DE (0.38) | US:KW (1.47) |
| (10) Eva: 4.17% | BR:US (1.56) | BR:US (1.49) | |
| Leiden: 3.76% | US:BR (0.97) | BR:PT (0.3) | |
| | BR:FR (0.32) | BR:GB (0.18) | |

Table 1. Indeed, when looking at Table 3, we observe that this community mainly contains American emigrants residing in Great Britain, Canada and Germany.

It is interesting to observe the tenth largest cluster that matches the tenth in Eva where Brazilian natives are the majority (see Table 2). In terms of migrants, Brazilian nationals residing in the USA are the third most frequent group, while with Eva, Italian natives are the third group. In Table 3, we remark that, indeed, the top migrant groups in this community are Brazilian emigrants residing in the USA, Portugal and Great Britain. Again, we remark that

most migrants share the same nationality labels within each community which is typically the main native group in that community.

#### 5.3.3 Infomap

With the Infomap algorithm, we identified 9 communities, which is much lower than the previous two algorithms, resulting thus also in very large communities. The largest community alone covers 66% of the users as shown in Fig. 14. Compared to the distributions of previously presented algorithms, the entropy values shown in the last row of Fig. 8 look very different. Here, we observe that the lowest entropy score is 0.2 for the nationality label, 0.24 for the residence label and 0.39 for migrant/native label, indicating very few pure clusters. The average of entropy scores for each label is relatively high (0.4 for nationality label, 0.42 for residence label and 0.49 for migrant/native label) which also tells us that there is more variety of country labels within each community. Same as previous algorithms, Infomap also shows similar entropy distribution patterns across labels, which indicates that cluster composition is similar across labels. The KS tests confirmed that they are indeed statistically similar to each other ($p$-value of 1 when comparing nationality and residence labels, 0.35 when comparing residence label and migrant/native label and 0.35 when comparing nationality label and migrant/native label).

Compared to the two previous algorithms, Infomap has quite a different list of most frequent country labels for each community. It has identified the USA, Italy, Mexico, Spain, Brazil, Turkey, Indonesia, Russia and Kuwait, respectively, as the most frequent community country/residence labels. Different from Eva and Leiden, the first largest community here is composed of mainly Americans, British and Canadians. We then have the second largest community that is mainly composed of Italians, Americans and British. The first largest community, however, has high entropy values, 0.47 for nationality and 0.48 for residence, telling us that country labels are quite heterogeneous. From Table 3, we remark that most of the migrants are from American natives living in Great Britain and Canada but also British natives living in the USA. The second largest community, on the other hand, has entropy scores of 0.22 for nationality and 0.26 for residence, which allows us to remark that this community contains mainly Italian natives. Table 3 also confirms that migrants are mainly Italian natives residing in different countries, which explains the higher entropy value for residence. Indeed, almost 65% of migrants belong to the first largest community as shown in Table 1. Same as Leiden, communities with the largest proportions of migrants are grouped with the US country label and the second largest proportions of migrants are grouped with Italy.

Interestingly, communities seven to nine are communities of users from countries of geographical proximity. For instance, community seven has grouped Indonesia, Malaysia and Singapore together. These communities, however, have high entropy scores of 0.41, 0.52, 0.59, respectively, for nationality and 0.43, 0.55 and 0.63, respectively, for residence. Moreover, we also observe that higher percentages of migrants are present in these communities as shown in Table 1. Table 3 indicates that these three communities are composed of migrants that mostly share the same nationality country labels but heterogeneous residence country labels. We, therefore, observe that communities tend to form based on the nationality label regardless of the community detection algorithms.

## 6 Privacy risk analysis

We now turn to the privacy risk analysis. We begin by performing a risk assessment methodology proposed by Pratesi et al. (2018), where a privacy risk evaluation of human mobility data is performed, considering a scenario where a Service Developer asks a Data Provider for data to develop an analytical service. The Data Provider must guarantee the right to privacy of the individuals whose data are recorded. Then the Data Provider queries its dataset to: (i) identify potential additional information (the so-called background knowledge) that an attacker might have about his/her target; (ii) simulate the attack based on the background knowledge, computing the privacy risk values for every individual in the dataset; (iii) select the dataset with the best privacy-utility trade-off; (iv) if necessary, apply a privacy risk mitigation method (e.g. generalisation, randomisation, suppression) on that dataset; and (v) deliver the sanitised dataset to a third party. In the current case, since we have access to a set of data related to a large number of individuals, we mimic the Data Provider's actions, simulating what could happen if the data we collect and use is exposed consequentially to an attack.

In Pratesi et al. (2018), and consequently in this paper, the risk of re-identification has been used as a measure of privacy risk (Samarati and Sweeney 1998), where they create a scenario where an *attacker* (sometimes called *adversary*) gains access to a dataset and, using some background knowledge about an individual under attack, they try to re-identify that specific individual in the dataset. The background knowledge represents both the kind and quantity of information known by the adversary. We use $b$ to indicate the specific background knowledge (e.g. the fact that a user posted a tweet at a specific moment) and $B_h$ to indicate a set of background knowledge of size $h$ (e.g. $B_2$ represents all the possible couples of tweets posted by

an individual). The need of having both $b$ and $B_h$ is given by the fact that, in order to provide the maximum protection, we extract the worst-case scenario. In the real-world situation, it is, of course, difficult to imagine what would be the actual knowledge of an adversary; however, we can establish a reasonable[11] quantity of information that they could know about the target, and compute the probability of re-identification of all the possible combinations of that specified size.

Let $\mathcal{D}$ be a database, $D$ a dataset derived from $\mathcal{D}$ (e.g. an aggregated data structure on time and/or space), and $D_u$ the set of records representing a user $u$ in $D$, the probability of re-identification is defined as follows.

**Definition 1** (*Probability of re-identification* Pratesi et al. 2018) Given an attack, a function *matching(d, b)* indicating whether or not a record $d \in D$ matches the background knowledge $b$, and a function $M(D, b) = \{d \in D | matching(d, b) = True\}$, we define the *probability of re-identification* of an individual $u$ in dataset $D$ as: $PR_D(d = u|b) = \frac{1}{|M(D,b)|}$ that is the probability to associate record $d \in D$ to individual $u$, given background knowledge $b$.

Note that $PR_D(d{=}u|b) = 0$ if the user $u$ is not in $D$. Since each background knowledge $b$ has its own probability of re-identification, we define the risk of re-identification of an individual as the maximum probability of re-identification over the set of possible background knowledge:

**Definition 2** (*Privacy risk* (Pratesi et al. 2018)) The risk of re-identification (or privacy risk) of an individual $u$ given a set of background knowledge $B_k$ is her maximum probability of re-identification $Risk(u, D) = \max PR_D(d = u|b)$ for $b \in B_k$. It has the lower bound $\frac{|D_u|}{|D|}$ (a random choice in $D$), and $Risk(u, D) = 0$ if $u \notin D$.

An individual is hence associated with several privacy risks, each for every background knowledge of an attack.

We point out that each attack assumes the adversary gains access to the dataset. Performing an attack means finding a set $C$ of possible matches for a target, given a certain background knowledge. The probability of re-identification of the user $u$ is $\frac{1}{|C|}$. A greater number of candidates imply better privacy protection.

In the following, we simulate two kinds of attacks: the first, in which we analyse the tweet(s) of each user, i.e. our raw data, and the second, in which we control the risk of partially transformed data, i.e. at the individual level.

### 6.1 Attack model at the Tweet level

Keeping only the relevant information for a specific application is compliant with the GDPR data minimisation principle, one of the guiding principles of the EU Regulation. Indeed, Article 5 of the GDPR states that personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')". Here, the relevant information for our applications does not include the tweet text, which can be promptly deleted, while for each tweet we keep only the meta-information, i.e. tags, geo-localisation and timestamp.

**Definition 3** (*Tweets History*) The tweets history of a user $u$ is the set of tweets posted by $u$:

$$T^u = t_1^u, t_2^u, ..., t_n^u$$

where $n$ is the number of tweets of the user $u$ in the considered time window.

**Definition 4** (*Minimum Tweet Format*) The data format of each tweet is the set of metadata related to that tweet:

$$t_i^u = time_i, [geo_i], [hashtag_i^1, [hashtag_i^2, [..., hashtag_i^m]]]$$

where $t_i^u$ is the i-th tweet of user $u$, $time_i$ is the time (typically in the timestamp format, i.e. "yyyy-MM-dd HH:mm:ss") in which $u$ posted $t_i^u$, $geo_i$ is the (potentially unspecified) location (despite location can have several format in Twitter can be mapped in a physical place, with a latitude and a longitude associated to it), and $hashtag_i^1$ - $hashtag_i^m$ are the potential hashtag(s) used in the tweet text.

Since geo-located tweets are only a minority (i.e. in our dataset around 75% of tweets are not geo-located), and hashtags are usually adopted in bunches, we focus on the time for our analysis. However, similar consideration and attack simulation can also be done considering the other dimensions.

It is important to note that, even if the tweets come with a timestamp, considering the time with a detail level of seconds is unnecessary for our analyses. For this reason, we should establish the essential level of detail suitable for the service realisation. In our case, this could be at the day level, which is the minimum temporal granularity required. However, for experimental purposes, we show what happens with different levels of detail. Indeed, as said before, generalisation is one of the possible (and, in addition, one of the simpler and more straightforward) strategies to reduce the

[11] The reasonableness, appropriateness and adequacy concepts are often cited in the GDPR with reference to the likelihood to re-identify individuals (Recital 26), technical measures put in place (Recital 78 and Article 25(1)), and safeguards to mitigate privacy risks (Recital 156 and Article 6(4)).

privacy risk, and it has often given good results in terms of the utility of the analysis. Keeping only the relevant information is compliant with the GDPR data minimisation principle, while empirical results have proven in different contexts (Monreale et al. 2014) the effectiveness of this approach to guarantee both privacy and utility.

At this point, we are ready for the first phase of our experiment: computing the risk of re-identification of each Twitter user, given the hypothesis that the adversary knows part of his/her history and, in particular, the timestamp(s) related to that portion of history. We expect that in this first phase, a vast majority of users can be re-identified easily given this very specific knowledge. However, we would also like to study how this risk decreases if we generalise the temporal dimension.

**Definition 5** (*Attack on Tweets*) The attacker uses the background knowledge $b$ on the user $u$ to match all the set of items that include $b$. Given $D(u_i)$ the set of items of the user $u_i \in \mathcal{D}$, the candidate set is computed as $C = \{u_i | b \subseteq D(u_i)\}$.

To clarify the previous attack, we report an example and a toy dataset in the following.

***Example 1*** Suppose that the dataset is composed of the following data:

$user, timestamp, ...$
$user_1, 2020-02-24 \quad 10:32:55, ...$
$user_1, 2020-02-27 \quad 20:25:05, ...$
$user_2, 2020-02-24 \quad 10:55:55, ...$
$user_3, 2020-02-10 \quad 22:32:43, ...$
$user_3, 2020-02-30 \quad 11:12:12, ...$

An adversary (Carol), knowing that her target posted a tweet on 24 February 2020 at 10:32, will be able to recognise $user_1$ her target, i.e. the only one with the cited timestamp. Indeed, in this case, we have $|C| = 1$, which corresponds to a probability of re-identification of 100%.

However, if we generalise the time at the hour level, the dataset would become:

$user_1, 2020-02-24 \quad 10:**:**, ...$
$user_1, 2020-02-27 \quad 20:**:**, ...$
$user_2, 2020-02-24 \quad 10:**:**, ...$
$user_3, 2020-02-10 \quad 22:**:**, ...$
$user_3, 2020-02-30 \quad 11:**:**, ...$

And in this case, even if she has the same background information, Carol cannot be sure if her target is $user_1$ or $user_2$ since both posted at 10 of the selected day. So, in this example, we have $|C| = 2$, which corresponds to a probability of re-identification of 50%.

Suppose we generalise the time again at the month level. In that case, we have that $|C| = 3$ (i.e. a probability of re-identification of 33%, having Carol's target indistinguishable from $user_1$, $user_2$, and $user_3$ since all three users have at least one tweet on February 2020.

Our attack first gives us the probability of re-identification posing $b$ equals to the fact that the target posted a tweet on 24 February 2020, at 10:32. Then, we must do the same reasoning for the second tweet (i.e. the one posted on 27 February 2020 at 20:25:05) in order to compute the risk of simulation of all the $b \in B_1$, respectively 100%, 100%, and 33% when we consider the time in seconds, hours or months. At this point, we can compute the risk of re-identification for each time granularity, as the maximum of each couple, i.e. $\max(100\%, 100\%) = 100\%$ for the time in seconds, $\max(50\%, 100\%) = 100\%$ for the time in hours and $\max(33\%, 33\%) = 33\%$ for the time indicated in months.

Once we define the attack, we can simulate it for every user in the dataset. For this attack, we rely on the data extracted at the first iteration, i.e. our core users, regardless they are migrants or natives. In total, we analyse 350,549 tweets written by 1761 users. Since the analyses described in previous sections do not need raw data, we decide to cancel immediately all the unnecessary information about the remaining users, following the data minimisation principle reported at the beginning of this section. However, we maintain all the relevant information of core users specifically to perform the attack simulation described in this Section, as representative of the whole dataset.

In Fig. 9 we report the results of our experiments, supposing that the adversary knows 1 tweet of his/her target. The plot shows the cumulative distributions of Twitter users (y-axis) with a risk of re-identification less than or equal to a certain value (x-axis), varying the temporal granularity used to store (and process) the tweets.

We do not report the result obtained using the actual timestamp because, as one can see, even with the generalisation to minutes (the full yellow square), all the users are completely re-identifiable. This means that each user has at least one tweet posted in a unique moment w.r.t. the considered dataset. Of course, with the whole Twitter audience, this will probably not be valid anymore: usually, in large set of users, they are more likely to have similar behaviours, and this leads to a well-known "hidden in the crowd" mechanism. In this particular case, it is unlikely that a minute without at least two tweets could pass. However, since the extracted datasets are inevitably limited w.r.t. the complete collection of existing tweets, the result shown in the figure also represents a piece of important evidence that

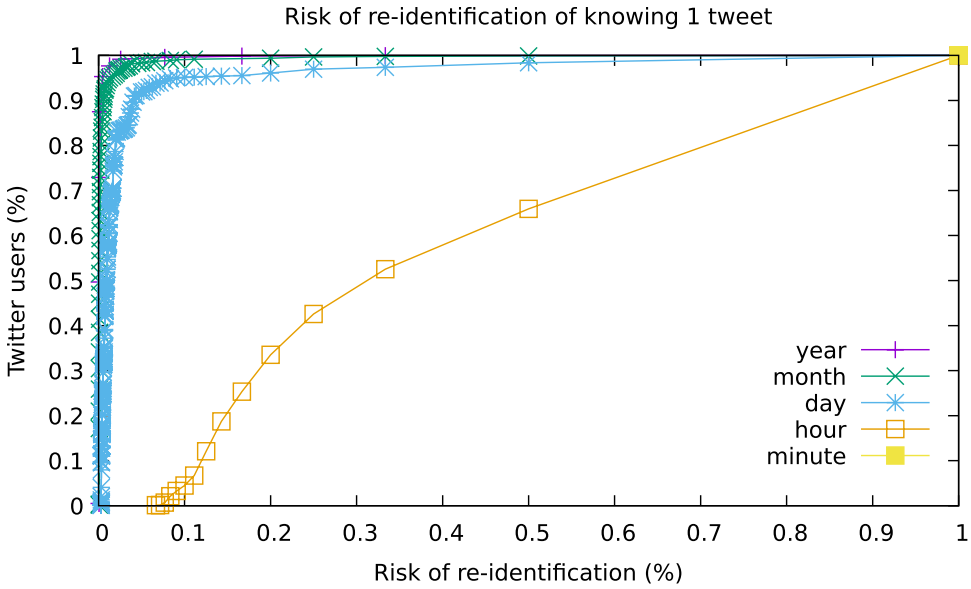


**Fig. 9** Simulation of the attack at the tweet level, varying the level of detail of the time of the tweet; in the key, we report the granularity considered in each attack

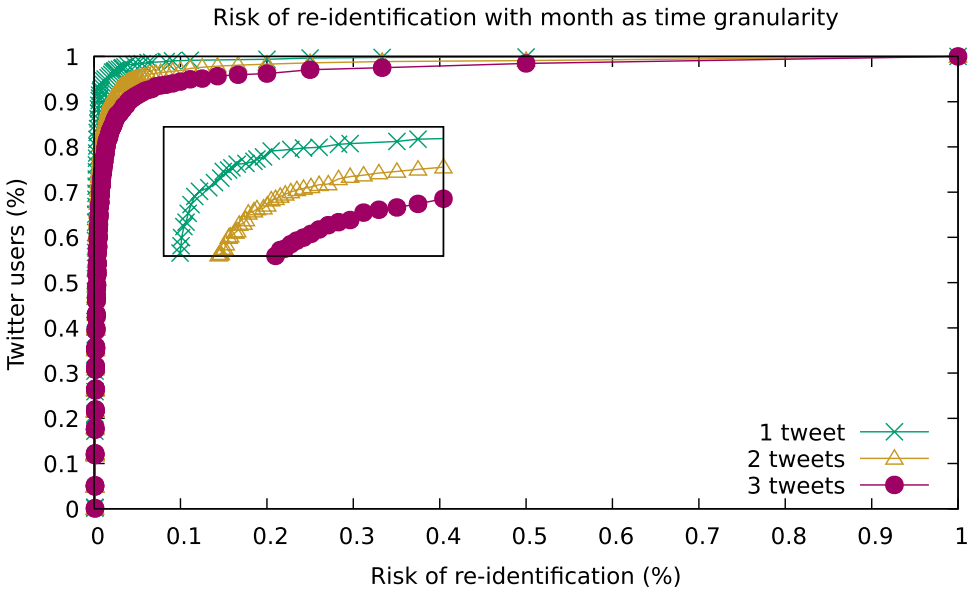


**Fig. 10** Simulation of the attack at the tweet level, fixing the time granularity to months and varying the number of tweets known by the adversary, which are reported in the key

emphasises the need to use only the right amount of necessary information.

If we generalise the time to hours (empty orange squares), we have around 35% of individuals that are surely re-identifiable, while, on the opposite side, 35% of users have a maximum risk of 20%.

However, the situation profoundly changes when we move to day (the light blue stars). In this case, we have 95% of individuals that have a maximum risk rate equal to 10%, which corresponds to a set $C$ of matching users that are indistinguishable from the target of the attack of size 10. This means that for a very large majority of users, there is a chance of 1 out of 10 to be re-identified knowing that they posted on a certain day. As said before, this is a maximum risk; indeed, for 1,430 users (around 80% of the total), this risk goes down to 2% (i.e. $|C| = 50$).

Not surprisingly, the risk lowers again when we pass to months (green crosses): 99% of individuals have a maximum risk of 10%, while for 95% of users, the risk is below 2%, and 93% of individuals have a risk lower than 0.1% ($|C| = 100$). Lastly, given the fact that risk values are already very low, passing from months to years barely affects the privacy risk. In addition, a generalisation to year could compromise the quality of our analyses. For both reasons, storing and processing tweets with this temporal granularity is not convenient.

However, as we saw before, we can also see what happens if we vary the quantity of knowledge known by the

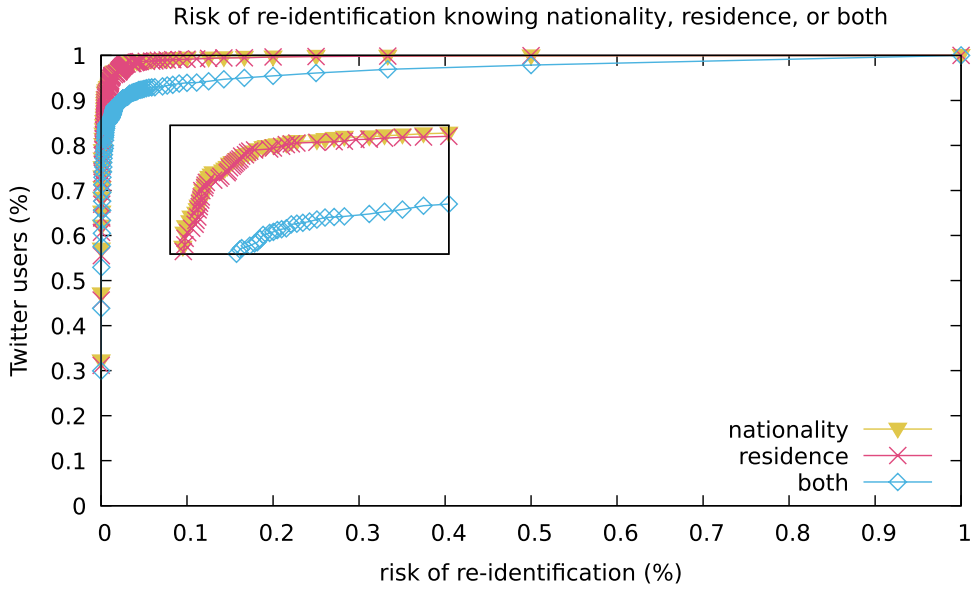


**Fig. 11** Simulation of the attack at the user level, with the knowledge of the country of nationality and/or the country of residence

adversary, i.e. the number of tweets. In Fig. 10, we report an example of this analysis, in which we chose to fix the time granularity at the month level. We again report the risk when we consider all the possible background knowledge of 1 tweet (the same green crosses of Fig. 9), but we also enlarge this knowledge to 2 and 3 tweets. We recall that we are showing what happens in the worst-case scenario, i.e. if the adversary knows the more uncommon combination of 2 (or 3) tweets posted by each user. As one can see, the risk knowing *any* 3 tweets (full eggplant circles) is essentially similar to the one associated to days in the previous plot (Fig. 9): 94% of individuals have a maximum risk equal to 10% ($|C| = 10$) and 85% of users have a risk lower than 2% (i.e. $|C| = 50$). These results prove that using months as the time level for tweets represents a good trade-off choice, providing adequate privacy guarantees while still permitting to perform analyses as the one describes in the previous sections.

## 6.2 Attack model at the user level

According to what we introduced in Sect. 6.1 about minimum data format, once we have computed the nationality and the country of residence of each Twitter user, the information about single tweets is no longer necessary. Indeed, we can decide to store, process and potentially share (with our collaborators or with a broader audience if our risk evaluation process gives good results in terms of privacy) only a more generalised version of the dataset. Indeed, a user can be labelled with two fundamental attributes, i.e. the nationality and the country of residence, which permit to classify him/her as a migrant or a native, and consequentially to pursue the analyses described in Sect. 5.

Thus, it becomes interesting to repeat the previous reasoning, computing each user's privacy risk, supposing that an adversary knows the country of nationality of his/her target, his/her country of residence or even both the previous information. In this case, the probability of re-identification and the privacy risk defined in Sect. 6 coincide since we cannot establish a quantity of background knowledge owned by the adversary, but it can be represented only by the knowledge of a single attribute. So, in this case, our hypothesis is that the attacker knows the nationality and/or the country of residence of the target and tries to re-identify him/her in the generalise version of the dataset, in which we lose the detail of single tweets. In Fig. 11, we illustrate the result of the simulation of this attack with a cumulative distributed function as in Sect. 6.1.

Contrarily to Sect. 6.1, here we have the possibility to perform the simulation on the whole set of users since data are generalised; thus, privacy risk is significantly lower. For this attack, we compute the risk of re-identification of 36,757 users.

As one can see, the privacy risk is essentially the same when the attacker knows one of the available information separately (full sand triangles and cranberry crosses): in both cases, around 93% of users have a maximum risk of 1% (this means a chance of 1 out of 100 to re-identify the target correctly in the dataset). However, when the adversary knows both the countries of residence and the nationality of his/her target, the privacy risk slightly increases: the empty sky blue diamonds indicate that around 94% of individuals have a maximum risk of re-identification of 10% (1 out of 10), while approximately 90% have a risk lower than 2% ($|C| = 50$). Again, even if it is hard to be seen in the plot, 85% of users have a maximum risk of 1% ($|C| = 100$), 78% of the total number of users in the

**Fig. 12** Communities defined by Eva: top 8 communities are highlighted by different colours: In the order of the size of communities, purple, green, blue, black, orange, red, blue-green, light red (Color figure online)

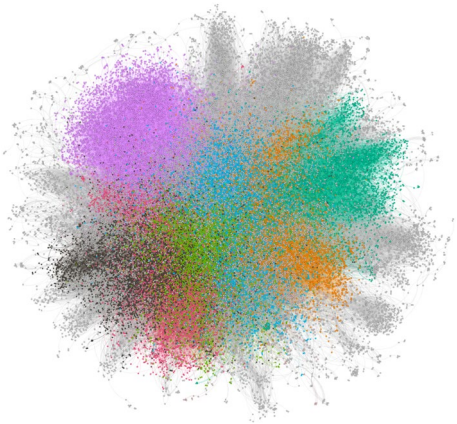

**Fig. 13** Communities defined by Leiden: top 8 communities are highlighted by different colours: In the order of the size of communities, purple, green, blue, black, orange, red, blue-green, light red (Color figure online)

dataset have a maximum risk of $0.\bar{3}\%$ ($|C| = 300$), and for more than 50% of users the risk is even lower than 0.04% (i.e. $|C| = 2500$). These results qualify our dataset to be treated as anonymised and enforce our decision to release it, respecting the Twitter policies.[12]

[12] https://developer.twitter.com/en/developer-terms/agreement-and-policy, accessed on 25 June 2022.

**Fig. 14** Communities defined by Infomap: top 8 communities are highlighted by different colours: In the order of the size of communities, purple, green, blue, black, orange, red, blue-green, light red (Color figure online)

These results are also in line with the expectation due to previous analysis (as in Monreale et al. 2014 or Pratesi et al. 2017), and they are certainly compliant with the values that guided the writing of the GDPR.

## 7 Discussion and conclusions

We studied the characteristics of two different communities, migrants and natives, observed on Twitter. Analysing profiles, tweets and network structure of these communities allowed us to discover interesting differences. More precisely, we were able to answer the following research questions: *Do migrants/natives have many followers or friends? What do migrants/natives talk about? To whom migrants/natives connect to? Who are the most central users amongst them? Where do migrants and natives belong in a community?*. In the respective order of the questions, we observed that migrants have more followers than friends. They also tweet more often and in more various locations and languages. This is also shown through the hashtags, where the most popular hashtags used among migrants reflect their interests in travels. We believe that these characteristics of migrants are due to the fact that they are more likely to travel and meet more people during their travel. Hence, their tweet locations, languages and the context itself reflect this on Twitter. Furthermore, we detected that Twitter users tend to be connected to other users that share the same nationality more than the country of residence. This tendency was relatively stronger for migrants than for natives. Also both natives and migrants tend to connect mostly with natives.

This suggests that migrants feel more connected and closer to their own nationality also in online social networks. The amount of connection to natives could be a new indicator for measuring a level of social integration, taking into account the fact that immigrants may feel closer to the ones that share the same nationality than the locals of the host society at an early stage of immigration. On the other hand, the gradual shift in the composition of social connections to include more locals may indicate a higher level of social integration. The homophilic behaviours of users are also well reflected in the communities that we detected. We saw that natives tend to be the main composition in the top ten largest communities. On the other hand, we found that migrants tend to be part of large communities that are mainly composed of American or Italian natives. In other smaller communities, migrants tend to be grouped with either the users with the same nationality or with immigrants in their country of origin. Additionally among the three community detection algorithms that we computed (Eva, Leiden and Infomap), we observed that Eva and Leiden detect similar community patterns. Conversely, Infomap identified larger size of communities that show different patterns. In particular, we observed a geographic proximity of users in some of the communities. Furthermore, through our simulation studies of re-identification risks, we saw that the risk gets significantly reduced as we generalise details of tweets' timestamps. This finding continues to be valid even if we increase the quantity of information known by an adversary. In a scenario more suited to our setting, we repeated the same analysis without the knowledge of tweets but including the country of residence and/or origin labels of our users and observed that the risk of re-identification is still very low even if an adversary knows more about their target.

We believe that our work can be useful to study the integration process of immigrants allowing researchers and policy makers to measure how these two groups of population integrate, and communicate to each other in an online setting in the host society. As we also provide insights into topics of the discussions and influential users on Twitter, it could enable researchers and policymakers to target specific topics and individuals to facilitate the communication channels between these two groups. Additionally, the composition of communities that we studied can be useful to understand which communities should be targeted to increase diversity of its group members and/or to integrate with other communities. Furthermore, our re-identification risk analysis provides a useful guideline for researchers using digital trace data. The risk of ethics and privacy issues have always been raised in related researches that deal with digital trace data, but, to the best of our knowledge, no technical analysis has been done to measure the level of re-identification risk in social media. Through this study, we provided different scenarios where the risk of re-identification can vary from very high to very low. This would provide other researchers with a guideline on how to deal with data privacy issues when working with Twitter data, helping to find the suitable trade-off for a specific analysis. For our purposes, we found that tweet timestamps restricted to monthly level represent a good choice, providing adequate privacy guarantees while still allowing us to perform social network analysis similar to the first part of this work. Moreover, and most importantly, these analyses produce additional evidence that minimising the information used is a simple yet valid way to mitigate privacy risks when we are dealing with real personal data. Finally, thanks to this quantification of privacy risk, we are able to make our data publicly available as well, removing the users whose risk of re-identification was too high (i.e. greater than 20%).

Our work, however, suffers from a few drawbacks. First of all, the Twitter population is different from the general population. It has been shown that the USA. Twitter population is younger and more educated than the overall US adult population.[13] This work, hence, suffers from selection bias and as mentioned previously, we do not intend to generalise the findings of this work as only a small sample of individual Twitter data was used. Secondly, although we have observed that the difference between migrant and native users persists even when we compare the US emigrants with the US non-migrants, we cannot rule out the fact that the difference may be also driven by other possible factors. It is well known in the literature that migrants are "not randomly selected from population of the source countries" (Borjas et al. 2019). It is hence possible that the differences that we observed are in part due to the self-selection of migrants. For instance, the usage of Twitter for migrants may be more by highly educated and professional migrants than that of natives. Nevertheless, we believe that by aggregating the individual level data, we were able to extract information that is worthwhile to be investigated further. To this extent, we simply intend to present what is demonstrated through our dataset. In spite of this drawback, we were able to observe interests, usages of Twitter and social interactions between migrants and natives thanks to the availability of the Twitter data.

In the future, it would be interesting to exploit further some of the findings of this work. For instance, we can observe how central users in the network are spreading culture or information throughout the network and how effective are the spreading/communication channels initiated by these central users. Additionally, based on the network composition we have observed, it is possible to investigate strong

[13] https://www.pewresearch.org/internet/2019/04/24/sizing-up-twitter-users/.

and weak ties in the network to study network supports for migration settlement (Blumenstock et al. 2019).

**Author contributions** JK, AS, GR and FP conceptualized the study; JK and GR collected the data; JK, AS, GR, FP and FG helped in methodology and writing; JK, FP, GR, AS contributed to formal analysis and investigation; FP was involved in privacy risk analysis; FG, and AS acquired the funding; all authors read and approved the final manuscript.

**Funding** Open Access funding enabled and organized by Projekt DEAL. This work was supported by the European Commission through the Horizon2020 European projects "SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics" (grant agreement no 871042) and "HumMingBird—Enhanced migration measures from a multidimensional perspective" (grant agreement no 870661).

**Availability of data and materials** Data without Twitter IDs and without network information are available on https://doi.org/10.6084/m9.figshare.19348058.v2. According to the analysis described in Sect. 6.2, we removed the users whose risk of re-identification was too high (i.e. greater than 20%).

## Declarations

**Conflict of interest** The authors declare that they have no conflict of interest.

**Ethical approval** All the procedures on use, process and storage of data were reviewed by the SoBigData Board for Operational Ethics and Legality.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

## References

Bello-Orgaz G, Hernandez-Castro J, Camacho D (2017) Detecting discussion communities on vaccination in twitter. Futur Gener Comput Syst 66:125–136

Bloch F, Genicot G, Ray D (2008) Informal insurance in social networks. J Econ Theory 143(1):36–58

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008

Blumenstock JE, Chi G, Tan X (2019) Migration and the value of social networks

Borjas GJ, Kauppinen I, Poutvaara P (2019) Self-selection of emigrants: theory and evidence on stochastic dominance in observable and unobservable characteristics. Econ J 129(617):143–171

Buccafurri F, Lax G, Nicolazzo S, Nocera A (2015) Comparing twitter and facebook user behavior: privacy and other aspects. Comput Hum Behav. https://doi.org/10.1016/j.chb.2015.05.045

Cha M, Haddadi H, Benevenuto F, Gummadi PK et al (2010) Measuring user influence in twitter: the million follower fallacy. Icwsm 10(10–17):30

Cha M, Benevenuto F, Haddadi H, Gummadi K (2012) The world of connections and information flow in twitter. IEEE Trans Syst Man Cybern Part A Syst Hum 42(4):991–998

Citraro S, Rossetti G (2020) Identifying and exploiting homogeneous communities in labeled networks. Appl Netw Sci 5(1):1–20

Coletto M, Esuli A, Lucchese C, Muntean CI, Nardini FM, Perego R, Renso C (2017) Perception of social phenomena through the multidimensional analysis of online social networks. Online Social Netw Media 1:14–32

Comola M, Mendola M (2015) Formation of migrant networks. Scand J Econ 117(2):592–618

De Cristofaro E, Soriente C, Tsudik G, Williams A (2012) Hummingbird: privacy at the time of twitter. In: 2012 IEEE Symposium on security and privacy, pp. 285–299. IEEE

Foster AD, Rosenzweig MR (2001) Imperfect commitment, altruism, and the family: evidence from transfer behavior in low-income rural areas. Rev Econ Stat 83(3):389–407

Gao S, Rao J, Liu X, Kang Y, Huang Q, App J (2019) Exploring the effectiveness of geomasking techniques for protecting the geoprivacy of twitter users. J Spat Inf Sci 19:105

Garcia D, Goel M, Agrawal AK, Kumaraguru P (2018) Collective aspects of privacy in the twitter social network. EPJ Data Sci 7(1):1–13. https://doi.org/10.1140/epjds/s13688-018-0130-3

Gërxhani K, Kosyakova Y (2020) The effect of social networks on migrants' labor market integration: a natural experiment. Technical report, IAB-Discussion Paper

Gould DM (1994) Immigrant links to the home country: empirical implications for us bilateral trade flows. Rev Econ Stat 76:302–316

Grandjean M (2016) A social network analysis of twitter: mapping the digital humanities community. Cogent Arts Humanit 3(1):1171458

Granovetter M (1983) The strength of weak ties: a network theory revisited. Sociol Theory 1:201–233

Granovetter M (2018) Getting a job: a study of contacts and careers. University of Chicago press, Chicago

Hawelka B, Sitko I, Beinat E, Sobolevsky S, Kazakopoulos P, Ratti C (2014) Geo-located twitter as proxy for global mobility patterns. Cartogr Geogr Inf Sci 41(3):260–271

Hu T, Wang S, She B, Zhang M, Huang X, Cui Y, Khuri J, Hu Y, Fu X, Wang X, Wang P, Zhu X, Bao S, Guan W, Li Z (2021) Human mobility data in the covid-19 pandemic: characteristics, applications, and challenges. Int J Digit Earth. https://doi.org/10.1080/17538947.2021.1952324

Huang X, Li Z, Jiang Y, Li X, Porter D (2020) Twitter reveals human mobility dynamics during the covid-19 pandemic. PLoS ONE. https://doi.org/10.1371/journal.pone.0241957

Kekülüoğlu D, Magdy W, Vaniea K (2020) Analysing privacy leakage of life events on twitter. In: WebSci

Kim J, Sîrbu A, Giannotti F, Rossetti G, Rapoport H (2022) Origin and destination attachment: study of cultural integration on twitter. EPJ Data Sci 11(1):1–20

Kim J, Sîrbu A, Giannotti F, Gabrielli L (2020) Digital footprints of international migration on twitter. In: International symposium on intelligent data analysis, pp. 274–286. Springer

Krishnan P, Sciubba E (2009) Links and architecture in village networks. Econ J 119(537):917–949

Kwak H, Lee C, Park H, Moon S (2010) What is twitter, a social network or a news media? In: Proceedings of the 19th international conference on world wide web, pp. 591–600

Li X, Xu H, Huang X, Guo CA, Kang Y, Ye X (2021) Emerging geodata sources to reveal human mobility dynamics during covid-19 pandemic: opportunities and challenges. Comput Urban Sci. https://doi.org/10.1007/s43762-021-00022-x

Mahoney J, Le Louvier K, Lawson S, Bertel D, Ambrosetti E (2022) Ethical considerations in social media analytics in the context of migration: lessons learned from a horizon 2020 project. Research Ethics, 17470161221087542

Mao H, Shuai X, Kapadia A (2011) Loose tweets: an analysis of privacy leaks on twitter. In: Proceedings of the 10th Annual ACM Workshop on Privacy in the Electronic Society, pp. 1–12

Mazzoli M, Diechtiareff B, Tugores A, Wives W, Adler N, Colet P, Ramasco JJ (2020) Migrant mobility flows characterized with digital data. PLoS ONE 15(3):0230264

McKenzie D, Rapoport H (2010) Self-selection patterns in mexico-us migration: the role of migration networks. Rev Econ Stat 92(4):811–821

Monreale A, Rinzivillo S, Pratesi F, Giannotti F, Pedreschi D (2014) Privacy-by-design in big data analytics and social mining. Eur Phys J Data Sci. https://doi.org/10.1140/epjds/s13688-014-0010-4

Munshi K (2003) Networks in the modern economy: Mexican migrants in the us labor market. Q J Econ 118(2):549–599

Munshi K (2014) Community networks and the process of development. J Econ Persp 28(4):49–76

Newman ME (2002) Assortative mixing in networks. Phys Rev Lett 89(20):208701

Parrish R, Colbourn T, Lauriola P, Leonardi G, Hajat S, Zeka A (2020) A critical analysis of the drivers of human migration patterns in the presence of climate change: a new conceptual model. Int J Environ Res Public Health. https://doi.org/10.3390/ijerph17176036

Pratesi F, Monreale A, Trasarti R, Giannotti F, Pedreschi D, Yanagihara T (2018) PRUDEnce: a system for assessing privacy risk versus utility in data sharing ecosystems. Trans Data Priv 11(2):139–167

Pratesi F, Monreale A, Giannotti F, Pedreschi D (2017) Privacy preserving multidimensional profiling. In: International conference on smart objects and technologies for social good (Goodtechs). https://doi.org/10.1007/978-3-319-76111-415

Radicioni T, Saracco F, Pavan E, Squartini T (2021) Analysing twitter semantic networks: the case of 2018 italian elections. Sci Rep 11(1):1–22

Rauch JE (1999) Networks versus markets in international trade. J Int Econ 48(1):7–35

Rossetti G, Citraro S, Milli L (2021) Conformity: a path-aware homophily measure for node-attributed networks. IEEE Intell Syst 36(1):25–34

Rosvall M, Bergstrom CT (2008) Maps of random walks on complex networks reveal community structure. Proc Natl Acad Sci 105(4):1118–1123

Samarati P, Sweeney L (1998) Generalizing data to provide anonymity when disclosing information (abstract). In: PODS, p. 188

Sîrbu A, Andrienko G, Andrienko N, Boldrini C, Conti M, Giannotti F, Guidotti R, Bertoli S, Kim J, Muntean CI et al (2020) Human migration: the big data perspective. Int J Data Sci Anal 11:341

Traag VA, Waltman L, Van Eck NJ (2019) From louvain to leiden: guaranteeing well-connected communities. Sci Rep 9(1):1–12

Xiong Y, Cho M, Boatwright B (2019) Hashtag activism and message frames among social movement organizations: Semantic network analysis and thematic analysis of twitter during the# metoo movement. Public Relat Rev 45(1):10–23

Zagheni E, Garimella VRK, Weber I, State B (2014) Inferring international and internal migration patterns from twitter data. In: Proceedings of the 23rd international conference on world wide web, pp. 439–444