Studying Inter-National Mobility through IP Geolocation*

Bogdan State
Stanford University
Stanford, CA, U.S.
bstate@stanford.edu

Ingmar Weber
Qatar Computing Res. Inst.
Doha, Qatar
ingmarweber@acm.org

Emilio Zagheni
Queens College - CUNY
New York, NY, U.S.
emilio.zagheni@qc.cuny.edu

ABSTRACT

The increasing ubiquity of Internet use has opened up new avenues in the study of human mobility. Easily-obtainable geolocation data resulting from repeated logins to the same website offer the possibility of observing long-term patterns of mobility for a large number of individuals. We use data on the geographic locations from where over 100 million anonymized users log into Yahoo! services to generate the first global map of short- and medium-term mobility flows. We develop a protocol to identify anonymized users who, over a one-year period, had spent more than 3 months in a different country from their stated country of residence ("migrants"), and users who spent less than a month in a country different from their country of residence ("tourists"). We compute aggregate estimates of migration probabilities between countries, as inferred from a user's location over the observed period. Geolocation data allow us to characterize also the pendularity of migration flows – i.e., the extent to which migrants travel back and forth between their countries of origin and destination. We use data regarding visa regimes, colonial ties, geographic location and economic development to predict migration and tourism flows. Our analysis shows the persistence of traditional migration patterns as well as the emergence of new routes. Migrations tend to be more pendular between countries that are close to each other. We observe particularly high levels of pendularity within the European Economic Area, even after we control for distance and visa regimes. The dataset, methodology and results presented have important implications for the travel industry, as well as for several disciplines in social sciences, including geography, demography and the sociology of networks.

Categories and Subject Descriptors

H.3.5 [**Information Storage and Retrieval**]: Online Information Services; J.4 [**Social and Behavioral Sciences**]: Sociology

*We are grateful to Yahoo! Research and the Barcelona Media Foundation for their support. State was an intern and Weber an employee of Yahoo Research! during the course of this research. State's work was supported by the Joan Butler Ford Stanford Graduate Fellowship. Additional material is available at: https://sites.google.com/site/studyinginternationalmobility/

WSDM'13, February 4–8, 2013, Rome, Italy.

Keywords

mobility, IP addresses, migrations, tourism

1. INTRODUCTION

International migration is a major source of population growth in several regions of the world, in particular in developed countries, where fertility has fallen under replacement level [23]. Migration processes are complex phenomena that have relevant consequences on societies, economies, culture and the environment. Despite the importance of migration dynamics, our understanding of global human mobility is still quite limited. As a matter of fact, migrations account for a very large source of uncertainty in population projections carried out by the United Nations [9]. One of the reasons for the lack of understanding of migration processes is the limited availability of data on international migration flows. For a substantial number of developing countries, data on migration flows are not collected at all. For developed countries with mature demographic registration systems, data on migration flows are often inconsistent across countries, since different countries collect data for different purposes and thus use different definitions of migration [22].

In this article, we analyze human global mobility using an innovative data source. We track the geographic locations from where anonymized users of a large provider of Internet services (Yahoo!) log in over time. We use geolocation data (from IP addresses) for a period of one year to build a data set on global mobility patterns. The data that we gather provide the most comprehensive and up-to-date picture of human mobility across the world. The data set that we generated is unique for three main reasons. First, migration flows are easily comparable across countries, since we use the same definition of migration consistently. Second, data from digital records show the latest trends, compared to official migration statistics, that are typically published with a considerable lag from the collection period. Third, the nature of our data allows us to analyze mobility on a continuous scale between tourism and longer-term changes of residence.

There is a growing body of literature on the use of geolocated data for the analysis of human mobility. This article complements existing studies in several different ways. For the first time, a data set that covers most countries in the world is used to estimate country-to-country flows of migrants and tourists. Previous work has focused on either tourist itineraries or migration rates from a specific country to the rest of the world. Our work goes beyond a descriptive analysis of human flows across borders: we develop a statistical model to predict flows and to understand the major determinants of migration and tourism patterns.

After a brief review of relevant related work, we provide a thorough explanation of how we built the data set on global mobility. Then we perform a statistical analysis of the new data set. Some of

the results confirm classic theories of migrations. Some others open new questions for social sciences and challenge traditional models of migration processes.

2. RELATED WORK

The study of human mobility is inherently an interdisciplinary field. This area of research has been the domain of demographers, geographers, statisticians, sociologists and economists, often working in teams. The recent availability of geographic information (through IP addresses) from where users log into Internet services has made possible unprecedented developments in the study of human mobility. Data mining of geo-located digital records has increasingly become central to the analysis of human movements at a global scale [21], with large implications for social sciences.

There are three main lines of literature in Web data mining that are relevant for our work. A first set of studies has used geolocated records from various sources to evaluate spatial mobility at a city or regional level. For instance, Ferrari et al. [16] analyzed urban patterns, for the city of New York, from Twitter data, whereas Rinzivillo et al. [35] used GPS traces to analyze regional mobility patterns. Routine whereabouts have also been extracted from applications that allow people to share their locations with friends (e.g., from Google Latitude [15] and Foursquare [29]). Mobile phone data have been tracked to understand frequent mobility patterns [6] and regularities in spatio-temporal trajectories of human mobility [19]. The applications of these types of analysis range from identification of urban routines [16] to evaluation of mobility trajectories in case of disasters [24], characterization of traffic jams [17] and inference on social ties from co-occurrence in time and space [10].

A second set of studies has focused on the analysis of tourists, with the goal of improving tourist itineraries or targeting tourists with ad hoc ads. Geo-referenced pictures in Flickr [12, 13], cellphone network data [18] and recommendations posted on the social network service for travelers CouchSurfing [34] have proved useful to reconstruct and improve travel itineraries for tourists.

A third group of studies has looked into longer distance and medium- to long-term movements. For instance, data from a large bill-tracking website, and from trajectories of traceable items, have been used to infer statistical regularities about long distance human mobility [7]. More recently, telecommunications data was used to track changes in individual's social networks as a result of internal migrations in Portugal [31]. Geolocated e-mail data have been used to infer emigration rates for a number of countries [39]. This last paper is the main inspiration for this article. In our work, we complement previous analyses by looking into global mobility (instead of local or regional patterns), by evaluating country-to-country migration flows (instead of overall emigration rates from selected countries), and by providing a statistical model for prediction and analysis of flows of migrants and tourists.

The work that we present in this study is relevant not only to the field of Web data mining, but also to the disciplines of demography, geography and sociology. International mobility is a central research area in social sciences. However, data on flows of migrants, in particular by age and gender, are almost inexistent. When some data exist, they are typically inconsistent across countries because of different definitions of migration and because of different methods of data collection.[1] Data on migration stocks (e.g., number of foreign born residents in a country) often come from censuses and are used to produce summary statistics, like the net migration rate of a country (the difference between immigrants and emigrants, over a period of time, per 1,000 residents). Data on migration flows are more sparse. For European countries, there has been a large effort to harmonize migration data. Thus, for some countries, there are data on flows reported by both the sending and the receiving country. These data have allowed the development of methods intended to generate comparable statistics on flows [2, 11, 22]. More often, data on flows are not available at all. In some cases, time series of migration flows can be roughly estimated indirectly, by evaluating differences over time in migration stocks obtained from census data [3, 36]. The lack of reliable and comparable data on migration flows is one of the main reasons behind the high level of uncertainty in models of global migration.

Web data mining of geo-referenced records has proven useful to provide new and relevant information for migration studies [39]. Statistical analysis of digital records is becoming increasingly influential in disciplines like demography and sociology. Our article complements a descriptive analysis of current global human flows with a statistical model whose results provide a test of classic theories of migration [8] in the contemporary world.

3. DATA SET

Our dataset was extracted from IP address data recorded from the logins of an initial sample of over 100 million anonymized users of Yahoo! Web services, over the period of one year (July 2011 - July 2012). Random numbers were used as user IDs. IP addresses were matched against the latest version of the GeoCityLite database provided by MaxMind[2]. Each user was thus linked with the country from where he/she logged in. Note that though IP-based geolocation has been found to be very noisy at the city level, it is generally considered reliable at the country level [33, 20, 37].

The data underwent a data-cleaning protocol that followed a number of rules. Each user's data were partitioned into "spells"[3] according to the country from where they logged in. A spell was defined as the total amount of time during which a user recorded logins only from within the same country. In addition to imposing the constraint that a user only logged in from the same country for the entire duration of the spell, we kept in our sample only those users whose geographic location, as given by city-level data extracted from the MaxMind database, varied in a plausible manner. Thus, the logins of a user within the same spell were not allowed to move in space at a speed higher than 150 km/h.[4] Our method builds on previous work on inferring location and mobility patterns using series of logins for the same individual over time [32]. Adopting this rule allowed us to remove from the data a great deal of noise that is typically associated with IP geolocation data, oftentimes as a result of users employing proxy servers to connect to the Internet.

Two other rules were implemented when extracting the dataset: we only kept in our dataset those users for whom the cumulative spell duration was of at least 300 days out of 365[5]. As a further

[1] Certain countries consider a migrant a person who moves his or her residence for at least 6 months, other countries use a threshold of 3 months or 1 year. Some countries collect data only on out-migration, some others only on in-migration. Some national statistical offices use data from registration systems, others from sample surveys, etc. [38].

[2] http://www.maxmind.com/app/geolite

[3] A spell is understood as the length of time during which a user is assumed to be in one state.

[4] We believe this method is accommodating of most instances of air travel within the same country, given that most individuals are not expected to log in immediately before or after a flight, but at most at a few hours' distance from the moments of take-off and landing.

[5] As a spell is defined as a series of logins from within the same country, the cumulative spell duration is the total number of days during the year for which we can say with relative certitude that the user was in the country, rather than somewhere else. Between

safeguard against noisy geolocations due to proxy servers we eliminated from our sample those individuals who recorded a location spell that lasted less than a day in duration (i.e., those users who went back and forth from the same country during a day, and logged in twice in the same country, but only for a few hours).

A a result of our data-cleaning procedure we obtained a sample of order 10^8 users, each of whom experienced an average of 1.12 spells of contiguous logins from one country. The average length of a spell was 306.00 days, whereas the average cumulative length of all spells was 341.76 days, out of 365 days possible.[6] On average, a user in our sample logged in more than 100 times. Out of all anonymized users, 96.68% spent (tracked) time in only one country, 3.10% spent time in two countries, 0.20% in three countries, and 0.03% spent time in four countries or more.

3.1 Defining Migration Events

Although the concept of migration seems intuitive at first blush, the problem of defining migration is not without its complications. Despite the fact that the United Nations provide recommendations on statistics of international migration [26], there is no universally-accepted definition of what a migrant is. Different countries and statistical agencies often use different definitions of the concept. There have been some efforts to harmonize migration statistics, particularly in Europe (see, for instance [14]). However, most international migration statistics are inconsistent, if existent at all.

The definition of migration used in this paper derives from the observation that a person who migrates during a given year typically spends a considerable amount of time in at least two countries during that year. From this consideration we define a migrant an individual who spends at least 90 days in exactly two countries during the observed timespan of one year. The timespan is not necessarily contiguous: thus we consider a migrant an individual who spends 45 days in country A, followed by 90 days in country B, followed by another 45 days in country A. Our operational definition aims to capture the *process* of migration rather than the *state* of being a migrant. An individual is considered to be potentially participating in migration when he or she moves between countries. The individual may move only once and follow the traditional conceptualization of being an emigrant from the home country or an immigrant to the destination. In this case the individual experiences one migration event. Alternatively, the individual may move between the country of origin and the one of destination multiple times, possibly over the course of a number of years. This type of individual is constantly engaged in the process of migration, and the reality of this increasingly-common pattern is oftentimes inadequately described by classical definitions of migration. Thus we allow individuals to spend their minimum 90 days in each country in multiple spells. A person is thus considered a migrant even if he or she returns to the home country for part of the year, as long as he or she meets the threshold of spending a minimum of 90 days in exactly two countries. This protocol generated a subsample of hundreds of thousands of migrants, about 3% of all individuals who spent more than three months in at least one country.

Our protocol identified individuals moving between countries, but gave us no information about the directionality of a move. To infer the most likely direction of a move we used the individuals' reported country of residence as obtained from the Yahoo! User database from April 2012.[7] The matching protocol further reduced our sample to a size of 223,344 individuals, as we dropped those individuals for whom no home country could be identified (6%), those emigrants whose reported home country was in neither of the two identified countries in which they had spent a considerable amount of time (17%), as well as those individuals whose reported age was lower than 15 or larger than 75 (2%).

3.2 Identifying Short-Term Mobility

Geolocation data allow us to identify not only geographic patterns consistent with international migration, but also movement between countries indicative of short-term mobility ("tourism"). Official definitions of tourism are broad and include "people traveling to and staying in places outside their usual environment for not more than one consecutive year for leisure, business or other purposes" [30]. In this paper, we refer to (international) tourism as short-term trips taken to a particular country, by a person who has a primary residence in a different country. More specifically, we considered tourists those individuals who spent more than three months in exactly one country (considered their primary country of residence), and who spent spells of less than a month in at least one other country (inferred to be the country visited by the individual). As in the case of migration, to be considered a tourist an individual could spend non-contiguous amounts of time in one country, though the cumulative time spent in the visited country could not exceed one month overall. We identified a sample of millions of individuals, about ten times larger than the sample of migrants.

3.3 Normalization Procedure

Our final dataset consists of the conditional probabilities of short- and medium-term mobility out of each country. We defined a mobility flow (migration or tourism) as the total number of individuals from one country (the "origin") who spent at least a certain amount of time in a different country (here referred to as the "destination") during the observed period. We normalized each count against the total numbers of individuals from the country of origin who engaged in short- or medium-term mobility, respectively. Edges (migration flows) corresponding to either 0% or 100% of outgoing flows from a country were dropped (since the log-odds are undefined). The final dataset consists of two weighted directed graphs of migration and tourism flows between countries, where for each node the outgoing edge-weights sum to 1.

4. RESULTS

4.1 Global Inter-Country Mobility

Figure 1 shows inferred worldwide migration patterns according to conditional probabilities of migration. These flows quantify the likelihood that a migrant from one country go to another one. Thus, for each country of origin, the conditional probabilities all sum to one. The lines used to represent migration flows encode three characteristics. Intensity encodes the magnitude of the conditional probability. Line type communicates the number of times individuals travel back-and-forth between two countries during the course of the year. As all individuals undergo migration during the interval of interest, they participate in at least one

two spells the user may be traveling, or simply not logging into his or her account. We imposed this constraint to keep in our sample those users for whom highly consistent location data was available, while leaving some room for inconsistent location data between spells due to international travel.

[6]Our interval included 366 days, but logs for one day were unavailable.

[7]For some individuals in our sample it is possible that this data may have been obtained after the occurrence of the migration event. Given that there is oftentimes considerable latency in users' updating their website profile information, we do not expect this potential delay in data gathering to impact our conclusions in a substantive way.

movement between the two countries. Solid lines on the map indicate those migrant flows where individuals are less than half as likely to return to their home country during the year (where the overall number of travels between the two countries is less than 1.5). Dashed lines indicate migrant flows where individuals have a moderate likelihood of between- country travel (between 1.5 and 2 between-country movements). Dotted lines indicate highly "pendular" migration flows, where it is quite common for individuals to show interspersed spells of time spent in both origin and destination country, with an average of more than 2 movements per individual per year. Color is used to indicate the direction of migrant flows, which are shown as arcs that transition from black to red when moving from the country of origin to the migrants' destination. Thus, a country with a lot of red arcs is mainly a country of immigration while one with mainly black arcs is a country of emigration.

At the global level, data on conditional probabilities of migration show the persistence of migration patterns dominated by geography, language and economics. The United States dominates among global migration destinations, as it is the top destination for 58 (44%) out of the 132 countries with at least 50 migrants represented in the dataset. The United States is followed by Great Britain and France, which represent the top migration destinations for 10 and, respectively, 20 countries in our sample. India and Australia complete the top-five, each figuring as the top migration destination for 7 countries. The findings for short-term stays (broadly classified as tourism) reveal a relatively similar structure as for migration. The United States is the top destination for short-term stays for users coming from 57 (32%) of the 176 countries with more than 50 persons reporting short-term stays in the sample. France is second for short-term stays, being the top destination for 23 countries, followed by India with 13, Spain with 10 and Great Britain with 9.

Comparing Figures 1 and 2 reveals rather similar patterns of both short- and medium-term human mobility, with only a few noticeable differences. The most striking pattern is a web connecting all countries to the United States, followed by a smaller, though still noticeable tendency for many countries to be strongly connected with their former colonial metropolises - France, Spain and England, a trend from which Portugal however seems to be excepted. Another trend is the emergence of regional hubs. India, China, Australia, Brazil and Argentina, as well as South Africa are emerging as both migration and tourist destinations. Given that these countries have been experiencing above-average growth during the past decade, it stands to reason that economic development is a driving factor of choice of both migration and tourist destinations. We provide a more explicit examination of the role of economics in Section 4.3.

4.2 Regional Patterns of Migration

Figures 3 - 8 show migration flows in Europe, Africa, Latin America, the Middle East, South-East Asia and North America, respectively. Within Europe we observe movement from East to West, the main preferred destinations being the U.K., Germany, France, Spain and Italy. Europe likewise shows highly-pendular migration flows, which are represented with dashed and dotted lines, as is evident from the profusion of such lines on the map. Because the map only shows flows originating in countries with at least 50 users that migrate, and given that most African countries do not cross this threshold, there are few visible flows in Africa (portrayed with part of the Middle East in Figure 4). The exceptions are flows from Liberia to Lebanon, as well as from Botswana and Zimbabwe to South Africa. Latin America (Figure 5) shows a broad pattern of regional migration gravitating towards Brazil and Argentina, and, to a lesser extent, towards Mexico and Colombia. There is an interesting observable flow of people from Haiti to the Dominican Republic, as well as from Haiti to Jamaica, possibly as a result of the humanitarian crisis caused by the country's major earthquake in 2010. Another striking migration pattern connects Cuba to Venezuela, otherwise a country of emigration. A local pattern is visible in Central America, with migration flows from Belize, Honduras and El Salvador to Guatemala, as well as from Honduras to Nicaragua and from Nicaragua to Panama. The South East Asian region (Figure 7) is dominated by India, China and Australia, the main poles towards which migrant flows gravitate. India and Saudi Arabia are the centers of migration flows for the Middle East region (Figure 6). The United States attracts most of the migration flows in North America (Figure 8).

Figure 3: Conditional Probabilities of Migration in Europe Represented flows originate in countries with at least 50 total observed migrants. Line intensity represents conditional propensity of migrating to destination country. Solid lines show flows with less than 1.5 trips between origin and destination; dashed lines indicate between 1.5 and 2 trips per year, and dotted lines show over 2 trips per year. Line thickness indicates observed size of migrant flow.

4.3 Determinants of Mobility

Given the large number of observations obtained from our dataset it is possible to test empirically for the extent to which multiple factors determine mobility choices across the world. To do so we matched our dataset with three other data sources. From [27] we obtained data on colonial ties between countries, travel visa regimes, whether two countries share language and geographic location, Purchasing Power Parity-Adjusted GDP, and the volume of bilateral trade between two countries. From [28] we obtained a measure of distance between countries, weighted by the spatial distribution of their populations. Our analysis also employed current (2011) per-capita GDP figures from the World Bank [1].

Our dataset consists of conditional probabilities of migration and tourism for pairs of origin and destination countries. Because most countries generate repeated observations both as origin and as destination, a typical linear regression (e.g. Ordinary Least Squares)

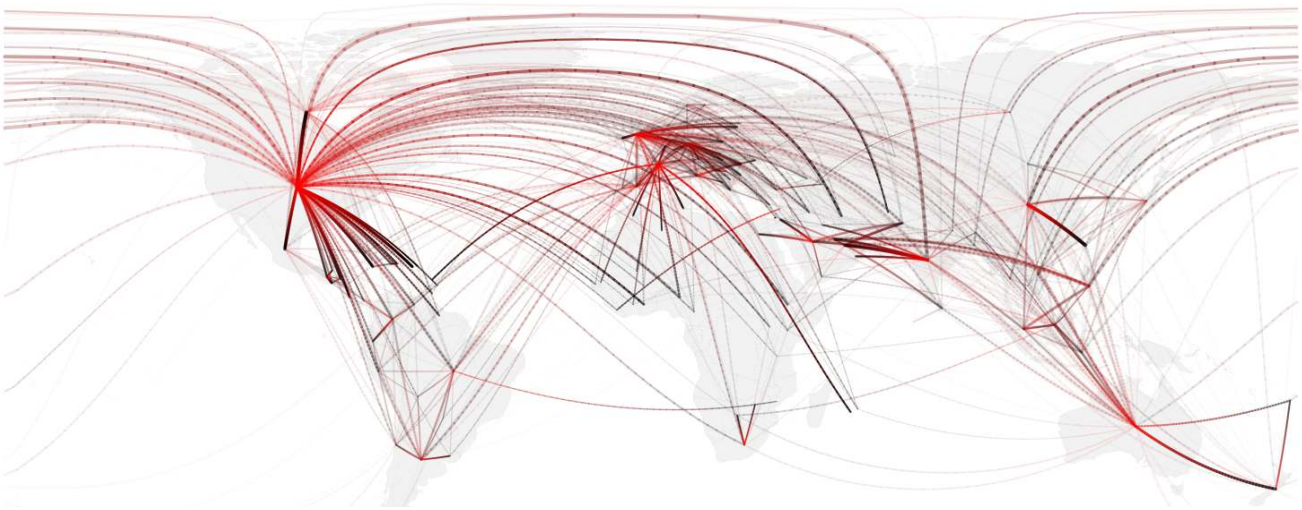

Figure 1: Conditional Probabilities of Migration
Represented flows originate in countries with at least 50 total observed migrants. Line intensity represents conditional propensity of migrating to destination country. Line thickness indicates observed size of migrant flow.

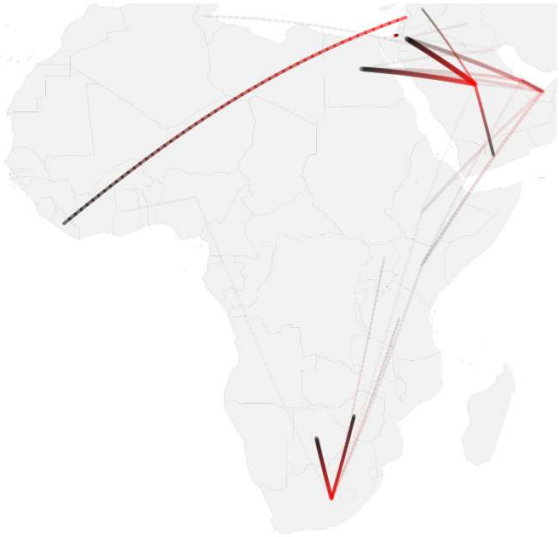

Figure 4: Conditional Probabilities of Migration in Africa

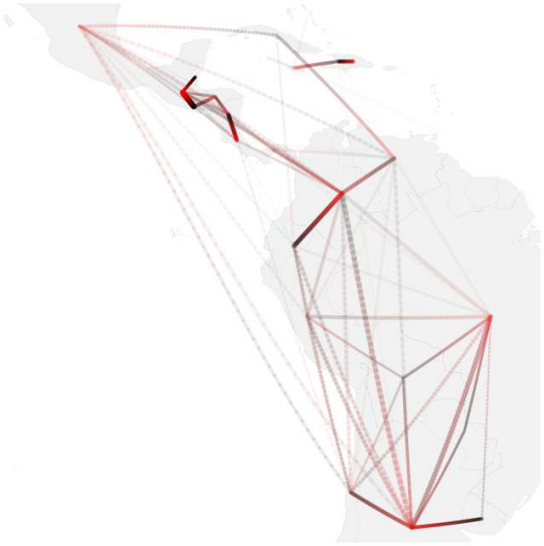

Figure 5: Conditional Probabilities of Migration in Latin America

would lead to biased results due to non-independence of observations. Treating observations about the same country as independent would overweight the amount of information available in them. To correct for this potential source of bias, we use a Linear Mixed Effects model [25], specified as:

$$y = XB + ZU + \epsilon,$$

with y a vector of responses, X the data matrix, B the fixed-effects coefficients, U a vector of "random effects" partitioned according to a classification of the data given by the incidence matrix Z and ϵ, the error component. This specification improves on the typical linear regression since it allows us to partition the error term in each regression into a systematic error component ("random effect"), due to the country of origin, and a random effect, due to the destination, in addition to the fixed effects given by the predictors.

Given that we are trying to predict a conditional probability, the distribution of which is highly skewed towards zero (as most mobility flows are minor ones), we transform the probability into its log-odds ratio, which we use as the Dependent Variable in our analysis.[8] We fit the model using the R package lme4[5]; tests of statistical significance for the estimated coefficients were calculated using MCMC methods provided by the languageR package[4].

Results for the regression analysis are presented in Table 1. The Dependent Variable being log-odds, the coefficients can be interpreted the same way as for a logistic regression. Thus we can see that having a colonial tie to the destination country increases the

[8]The analysis presented here does not include those cases where the observed probability is either 0 or 1, as the log-odds would be undefined.

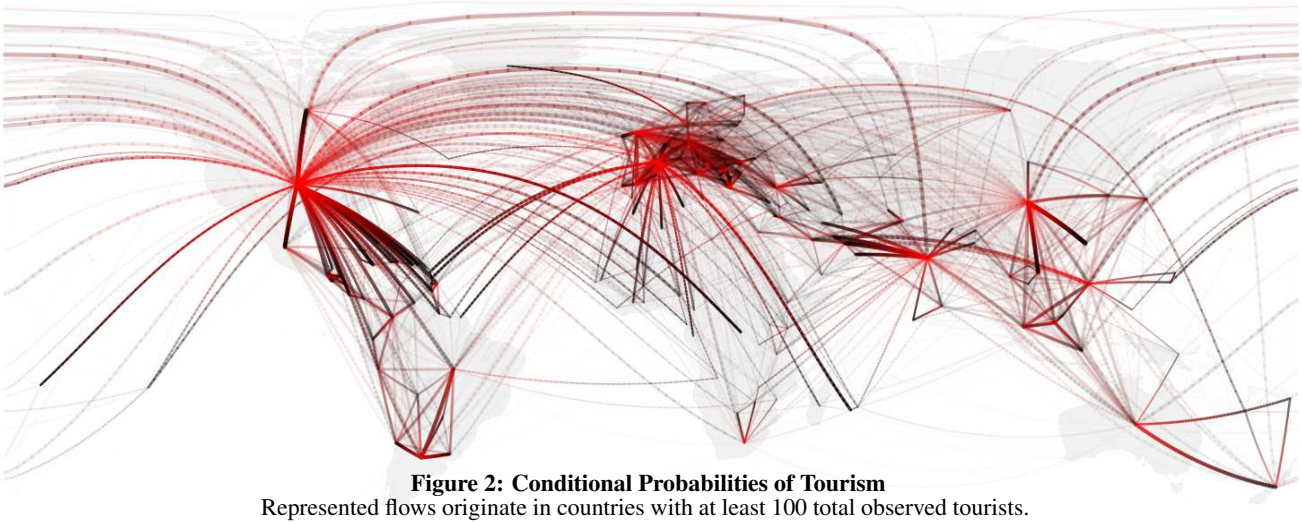

Figure 2: Conditional Probabilities of Tourism
Represented flows originate in countries with at least 100 total observed tourists.

Figure 6: Conditional Probabilities of Migration in the Middle East

Figure 7: Conditional Probabilities of Migration in South-East Asia and Australia

conditional log-odds of migration by 1.369, which is equivalent to an increase in the odds by a factor of 3.93. Individuals in our sample appear to have close to 4 times higher odds of migrating to a country to which their country of origin has a colonial tie (outside of the British Commonwealth), all else being equal. The respective effect for tourism is $3.23(e^{1.174})$. Commonwealth membership likewise has an effect, a smaller one, on the mobility log-odds, which increase by .345 and .310 in the case of tourism and migration, respectively.

Our analysis reveals that barriers to mobility are more salient in the case of short-term stays than they are in the case of longer-term migrations. A travel visa requirement has an effect comparable (same order of magnitude) to that of a Commonwealth tie, decreasing the log-odds ratio by .236 in the case of migrations, and .410 in the case of tourism. Common language facilitates migration: its presence increases the migration log-odds by .424, but its effect is even more poignant in the case of tourism, the log-odds of which are increased by .846. A one unit increase in the log-weighted distance measure yields a decrease of .686 in the migration log-odds, and of 1.128, roughly twice as much for tourism. Finally, being the same geographic region has a relatively minor effect (with distance accounted for) on the migration log-odds (.117), but an effect three times as large in magnitude (.387) on the log-odds of tourism.

The model presented in Table 1 also explores the influence of the economy in the country of destination – a "pull" factor – on the patterning of mobility. Interestingly, the destination's level of economic development, measured in absolute terms as well as relative to the country of origin's level, seems to have an effect of similar magnitude on short-term as well as long-term mobility. Each additional thousand U.S. dollars in the destination country GDP increases the log-odds ratio of migration by 0.058, and the log-odds of tourism by .095. The ratio between the destination and origin GDP has a smaller, seemingly second-order effect on the mobility log-odds: each unit increase in the ratio adds .007 to the migration log-odds and .004 to the log-odds of short-term mobility. Finally, bilateral trade between origin and destination, measured as

Table 1: Mixed Effects Regression of Conditional Log-Odds of Migration and Tourism

	Migrations			Tourism		
Fixed Effects	Coef	(S.E.)	T-stat.	Coef	(S.E.)	T-stat.
Intercept	-0.355	(0.273)	-1.303	1.934^{***}	(0.267)	7.248
Colonial Tie (non-Commonwealth)	1.369^{***}	(0.117)	11.703	1.174^{***}	(0.103)	11.369
Commonwealth Tie	0.345^{***}	(0.084)	4.114	0.310^{***}	(0.061)	5.082
Visa Required	-0.236^{***}	(0.044)	-5.321	-0.410^{***}	(0.031)	-13.323
Common Language	0.424^{***}	(0.058)	7.270	0.846^{***}	(0.043)	19.465
Log-weighted Distance	-0.686^{***}	(0.027)	-25.609	-1.128^{***}	(0.019)	-58.058
Same Region	0.117^{*}	(0.050)	2.361	0.387^{***}	(0.036)	10.722
Destination GDP / Origin GDP	0.007^{***}	(0.001)	6.300	0.004^{***}	(0.001)	5.570
Destination GDP (PPP-adjusted)	0.058^{***}	(0.010)	5.504	0.095^{***}	(0.016)	5.906
Bilateral Trade	0.035^{***}	(0.003)	10.875	0.032^{***}	(0.003)	11.313
Random Effects	Variance	(S.D.)		Variance	(S.D.)	
Country of Origin	.573	(.757)		.627	(.792)	
Country of Destination	.947	(.973)		1.532	(1.238)	
Residual	.911	(.955)		1.248	(1.117)	
Counts	N	Origins	Destinations	N	Origins	Destinations
	4,641	144	123	9,236	157	125
Log-Likelihood	Model	Baseline	McFadden R^2	Model	Baseline	McFadden R^2
	-6,793	-10,481	.352	-14,671	-23,820	.477
Mean Prediction Error	Model	Baseline	Ratio	Model	Baseline	Ratio
	.290	.370	.784	.265	.457	.580

Source: Mobility Log-Odds extracted from observations of Yahoo! users. Predictors from [27] and [28]. Mean Prediction Error calculated with Training and Test datasets stratified according to country of origin, each containing half of observations in a country of origin. *:, p < .05, **: p < .01, ***: p < .001.

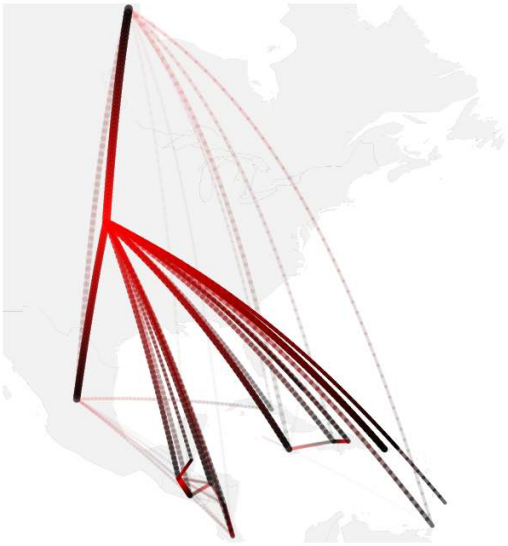

Figure 8: Conditional Probabilities of Migration in North America

a proportion of the origin's internal trade flows, and indicating the overall strength of economic ties between the two countries has an almost identical effect on the log-odds of migration and tourism (.035 and .032, respectively, for every unit increase in the proportion). The importance of economic ties for short-term mobility may seem counter-intuitive, but only if one loses sight of the fact that such mobilities may include business trips, educational travel, as well as short-term employment – all arguably influenced by the destination's economy – in addition to holiday travel.

In addition to fixed effects discussed above, the model includes country-specific random effects, which account for the systematic error components due to every origin and destination country present in the sample. Because the log-odds used as the Dependent Variable in the model are based on conditional probabilities of migration, the random effects for origin and destination countries have different interpretations. Similar to an entropy measure, the origin-specific random effect accounts for the lopsidedness in a country's mobility patterns. While the conditional probabilities of mobility originating in one particular country all sum to one, their logits have a slightly different behavior. The mean log-odds is relatively high for countries with few observed mobility destinations, such as the Republic of Congo or Swaziland, the top two countries of origin in terms of their random effects. At the other end of the spectrum are countries with a large number of observed mobility destinations, as is the case with India and the Philippines for migration.

Destination country random effects offer a more intuitive measure of surprise of a mobility destination, insofar as a country's popularity is not accounted by the fixed effects. The United States is the most "surprising" migration destination, in light of the explanatory variables included in the model. In context, the result is not counter-intuitive: the United States is known to be a popular migration destination for most countries, even though it imposes wide-spread visa restrictions, has few colonial ties, and is separated from most of the world's countries by a considerable distance. The estimated random effect for the United States is 2.56, which translates into an estimated 13-fold increase in the odds ratio of migrating to the United States, compared to the prediction based on the fixed effects. Likewise exceptional is the case of China for tourism, which yields an increase in the log-odds of 4.66. Exponentiated this quantity would predict an 105-fold increase in the odds of visiting China for a short time. This surprising increase should be seen in context however: with a large diaspora spread across the world, China is visited every year by a far larger number of individuals living very far away than distance would predict. The pattern of surprising tourist destinations being those with large and widespread diasporas is likewise supported by the second- and third- highest random effect, associated with India (3.68) and the

Philippines (2.82). Interestingly enough, the effect associated with the United States in the case of tourism (2.72, the fourth-highest) is similar in terms of magnitude to that associated with the country for migration log-odds.

The models show high explanatory power for all of the independent variables. All fixed effect coefficients are significant at the .001 level, with the exception of the coefficient associated with two countries being in the same region (significant at the .05 level), and of the intercept (not significant) in the case of migration. To provide a measure of the model's explanatory power we compute McFadden's R^2 measure, defined as 1 minus the ratio between the log-likelihood of the model and that of a baseline model. Our baseline ("null") model is set up conservatively: it is specified as a Linear Mixed Effects model which includes the same random effects, but only an intercept term for the fixed-effects portion. This baseline provides us with a measure of the extent to which the fixed-effects included in our model explain the variation in the Dependent Variable, the results being .352 for migrations and .477 for tourism.

The analysis was done using the whole data set. To ensure that the improvements in goodness-of-fit we observed were not merely caused by having more degrees of freedom and overfitting, we evaluated the model performance in a prediction setup, separately for migration and tourism, as follows. First, we looked at all source countries which had at least four target countries. Then, for each source country, we split all (source, target) pairs into train and test sets in a balanced, 50-50 manner. For all the train pairs we then fitted a single model to describe the log-odds as described above. To the pairs in the test set the fitted model was applied to obtain log-odds, and these log-odds L were then converted back to probabilities P according to $P = \exp(L)/(1.0 + \exp(L))$. These transition probabilities were then grouped by source country and re-normalized to 1.0. Re-normalization was required as (i) only half the transitions were present in the test set and (ii) a model fitted to log-odds ratios does not guarantee normalization to begin with. The predicted probabilities P_p obtained in this manner were then compared to the observed probabilities P_o and the difference measured according to $\|P_p - P_o\|_1 /2$, where the division by 2.0 guaranteed that the error was between 0. 0 for identical probability distributions and 1.0 for orthogonal ones. These errors are shown at the bottom of Table 1.

4.4 Pendular Migration

Europe stands out among world regions due to the extent to which migrations are circular on the continent. Migrants within Europe made an average of 2.52 trips between home and destination country during the year, whereas the next region in terms of mobility - the Americas - registers 0.72 fewer trips per person per year (t = 44.62, dF=39,904, p < .001). The next in terms of mobility is Australia and Oceania, with .26 fewer trips per year than the Americas (t=8.42, dF=1,567, p<.001), followed by Asia, .09 lower than Australia and Oceania (t=2.85, dF=1,479, p=.004). The continent with the least pendularity in intra-continent migrations is Africa, with .09 fewer trips than Asia (t=2.31, dF=569, p=.02).

To discover patterns in the pendularity of world migrations we perform regression analysis of the mean number of trips undertaken between countries (Table 3). We use the same Linear Mixed Effects specification, with random effects on origin and destination countries. In addition to the predictors included in Table 1 we included a dummy variable to account for both origin and destination countries being in the European Economic Area, a region in which our analysis reveals a great deal of pendularity.

The fixed effects estimates reveal no statistically-significant effect for colonial ties, travel visa regimes, having a common language, or the extent of commercial ties. It appears that these factors, while important for the initial decision of migration, lose salience when it comes to influencing migrants' opportunities to return temporarily to their country of origin. Distance has a great deal of influence on pendularity, however: each unit increase in the log-weighted distance measure yields an estimated -.4 reduction in the mean number of trips migrants undertake between two countries. With distance accounted for, having both origin and destination countries in the European Union increases by .322 the mean number of trips, suggesting an important effect of European integration efforts on human mobility.

Table 2: Mean Number of Trips between Origin and Destination Countries, within the same continent

Continent	Mean No. Moves	N
Europe	2.52	25,859
Americas	1.80	35,923
Australia and Oceania	1.54	1,388
Asia	1.45	44,254
Africa	1.36	550

There is likewise a small *negative* effect of both countries being in the same region (-.098). Given that travel between countries in the same geographic region is expected to be easier, this effect appears counter-intuitive. Likewise odd is the effect of the ratio of the destination and origin per-capita GDPs (-.002 for each unit increase). Though small, both effects are significant at the 5% level and demand an explanation. While a thorough explanation is beyond the scope of this paper, we hypothesize that these effects may be due to highly-skilled migrants. We believe that this migrant group is likely to make up a higher proportion of people migrating across regions, as well as between countries of similar GDP levels (especially from one developed country to another). Since these migrants typically command higher salaries and have greater access to flights between their countries of origin and destinations, the potentially greater share of highly-skilled migrants moving across regions and between similar-GDP countries would account for these unexpected effects. A likewise-small, though intuitive effect is presented by the destination country's per-capita GDP: for each additional thousand dollars in the destination's GDP, migrants to that country are likely to undertake .006 more trips between origin and destinations.

Some interesting patterns appear when considering the regression's random effects. Four out of the lowest five origin countries in terms of their random effects (which are, in decreasing order, Egypt, Malta, Tunisia, Yemen, Algeria) are Middle-Eastern countries that experienced turmoil during the recent Arab Spring. Originating in any of these countries was estimated to reduce a migrant's mean number of trips by between 0.36 (for Egypt) and 0.26 (for Algeria). Conversely, four out of the top five random effects (Croatia, Russia, the Czech Republic, Slovenia, and the Netherlands) are from Eastern European countries that are typically a source of economic migrants. Random effects for these countries range between .74 for Croatia and .41 for the Netherlands. There is less of a discernible pattern in terms of destination-specific random effects. Destination countries that consistently reduce pendularity beyond the fixed effects' predictions are Ireland, Slovenia, Latvia, Singapore and Ghana. The (negative) magnitude of their effects ranges between -.20 and -.12. The most pendularity-increasing destinations are, in descending order, Ukraine, Kazakhstan, Hungary, the Czech Republic and Slovakia. Their random effects are between .23 and .32. A potential explanation for the first two countries'

high levels of mobility is their proximity to Russia: many highly-mobile, economic migrants move between Russia (and other former Soviet republics) and Ukraine and Kazakhstan. The existence of large cross-border national minorities between the Czech Republic and Slovakia, and Slovakia and Hungary potentially accounts for the high level of pendularity created by these destination countries.

Even when adding a variable for within-EEA migration[9], the predictors have less explanatory power for pendularity than they have for the conditional odds of migration (McFadden's R^2 is .275, and the test Mean Squared Error decreases by only 16% against a baseline model including random effects, when the regression is trained on a stratified dataset containing 50% of all observations and tested on the rest of the observations).

Table 3: Mixed Effects Regression of Mean Number of Trips between Origin and Destination

	Migrations		
Fixed Effects	Coef	(S.E.)	T-stat.
Intercept	4.868^{***}	(0.182)	26.786
Colonial Tie (non-Comm.)	-0.104	(0.094)	-1.104
Commonwealth Tie	-0.035	(0.061)	-0.569
Visa Required	0.005	(0.033)	0.143
Common Language	-0.061	(0.044)	-1.379
Log-weighted Dist.	-0.400^{***}	(0.021)	-19.467
Same Region	-0.098^{*}	(0.039)	-2.509
Both in EEA	0.322^{***}	(0.056)	5.712
Dest. / Origin GDP	-0.002^{*}	(0.001)	-2.246
Dest. GDP (PPP-adjusted)	0.006^{***}	(0.002)	2.899
Bilateral Trade	-0.002	(0.003)	-0.608
Random Effects	Variance	(S.D.)	
Country of Origin	.573	(.757)	
Country of Destination	.947	(.973)	
Residual	.911	(.955)	
Counts	N	Origins	Destinations
	4,641	144	123
Log-Likelihood	Model	Baseline	R^2
	-5,668	-7,819	.275
Mean Squared Error	Model	Baseline	Ratio
	.709	.825	.859

Source: Mobility Log-Odds extracted from observations of Yahoo! users. Predictors from [27] and [28]. Mean Prediction Error calculated with Training and Test datasets stratified according to country of origin, each containing half of observations in a country of origin. *: p < .05, **: p < .01, ***: p < .001. McFadden R^2 measure used.

5. CONCLUSIONS

In this article, we presented an innovative method to estimate global flows of migrants and tourists using geo-located logins to Yahoo! Web services. For the first time, we estimated country-to-country flows in a consistent way both for developed and developing countries. Our work shows the unifying power of the Internet. Official migrations statistics are based on inconsistent definitions across countries. The Web brings all countries into a single dimension: once a definition of migration/tourism is chosen, then estimates comparable across countries can be obtained in a straightforward way. The methods that we discussed, and the data set that we generated have the largest potential in developing countries, where no other reliable data sources on migration flows exist.

[9]This was an insignificant predictor when added to models in Table 1, results not reported.

At the global level, we showed the persistence of migration patterns dominated by geography, language, history and economics. The findings for short-term movements (broadly classified as tourism) reveal patterns relatively similar to the ones observed for migration. The United States is the global center in the network of migration flows. We also observed the emergence of regional hubs of migration, like India and China, and a tendency for many countries to connect with their colonies. Individuals in our sample have almost four times higher odds of migrating to a country if they have a colonial tie. A travel visa requirement has an effect on mobility of the same order of magnitude as the one of a Commonwealth tie (but with opposite sign).

The dataset that we produced allows for the analysis of "pendularity" of migration, a phenomenon that is becoming more and more prominent, but that is understudied because of lack of data. We observed a high level of pendular, or circular, movements within the European Union. We also noted that countries that experienced turmoil during the recent Arab Spring tended to have very low levels of pendularity. In other words, people who left those countries had low probabilities of returning, at least for short visits.

Our study opens new and exciting opportunities for interdisciplinary research at the intersection of Web data mining and social sciences. Our study also comes with limitations that open up challenges for future research. Perhaps the most important problem that needs to be addressed is selection bias: individuals observed in our sample may or may not be representative of the entire population for their respective countries. For this study, two facts reassure us about the quality of our dataset and results. First, the regression analyses performed on all the estimated flows give results that are consistent with key findings in the sociological literature. Second, the conditional probabilities of migration that we estimated are consistent with the ones obtained by Guy Abel from census data for the period 1990-2000 [3]. We compared the top five destination countries for each country in our data set with the tables available in the companion website of [3]. Our estimates refer to the period 2011-2012, whereas the published tables are for the period 1990-2000. We observed striking similarities as well as new developments in migratory routes. More than 40% of the top five destinations for each country in Abel's data set are in the top five destinations for the respective countries in our world estimates based on Yahoo! data. The value is even higher (about 50%) if we consider only countries in the developed world. In some cases we observed a persistence of migration routes. For instance, the top destination countries for Italy in [3] are Germany, US, Serbia, France and Spain. All of these countries, except for Serbia, are in the top five destinations for Italy in our data set. Flows between Italy and Serbia at the end of the 1990s might have been related to the beginning and end of a period of political instability in the former Yugoslavia. For the US, we observed a change in traditional routes of migration. In the 1990s, the five top destinations from the US were Mexico, Germany, France, UK and Israel. In our data set, the top five destinations countries from the US are Mexico, China, India, Canada and the Philippines. The new trend is probably related to economic integration and return migration.

Combining geolocation data with social-network data and demographic information opens a series of exciting research questions that we intend to explore in future work. Sociologists of migration have long developed rich theories of migrant identities, immigrant integration, ethnic networks, immigrant entrepreneurship, etc. While these studies have been restricted to small-sample research, the increasing worldwide Internet usage opens promising new avenues to advance our understanding of human mobility.

6. REFERENCES

[1] *World development indicators*. World Bank (data.worldbank.org), 2011.

[2] G. Abel. Estimation of international migration flow tables in europe. *J R Stat Soc Ser A-G*, 173:797–825, 2010.

[3] G. Abel. Estimating global migration flow tables using place of birth data. In *European Population Conference*, 2012.

[4] R. H. Baayen. *languageR: Data sets and functions with "Analyzing Linguistic Data: A practical introduction to statistics"*., 2011. R package version 1.4.

[5] D. Bates and D. Sarkar. Linear mixed-effects models using s4 classes. *See http://cran. r-project. org/web/packages/lme4/index. html*, 2006.

[6] M. Bayir, M. Demirbas, , and N. Eagle. Discovering spatiotemporal mobility profiles of cellphone users. In *WoWMoM*, pages 1–9, 2009.

[7] D. Brockmann and F. Theis. Money circulation, trackable items, and the emergence of universal human mobility patterns. *PERCOM*, 7(4):28–35, 2008.

[8] S. Castles and M. Miller. *The age of migration: International population movements in the modern world*. Palgrave Macmillan, 2003.

[9] N. R. Council. *Beyond Six Billion: Forecasting the World's Population*. National Academies Press, 2000.

[10] D. J. Crandall, L. Backstrom, D. Cosley, S. Suri, D. Huttenlocher, and J. Kleinberg. Inferring social ties from geographic coincidences. *PNAS*, 107(52):22436–22441, 2010.

[11] J. De Beer, R. Raymer, R. Van der Erf, and L. Van Wissen. Overcoming the problems of inconsistent international migration data: A new method applied to flows in europe. *Eur J Popul*, 26(4):459–481, 2010.

[12] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Automatic construction of travel itineraries using social breadcrumbs. In *HT*, pages 35–44, 2010.

[13] M. De Choudhury, M. Feldman, S. Amer-Yahia, N. Golbandi, R. Lempel, and C. Yu. Constructing travel itineraries from tagged geo-temporal breadcrumbs. In *WWW*, pages 1083–1084, 2010.

[14] J. DeWaard and J. Raymer. The temporal dynamics of international migration in Europe: Recent trends. *Demogr Res*, 26(21):543–592, 2012.

[15] L. Ferrari and M. Mamei. Discovering daily routines from google latitude with topic models. In *PERCOM*, pages 432–437, 2011.

[16] L. Ferrari, A. Rosi, M. Mamei, and F. Zambonelli. Extracting urban patterns from location-based social networks. In *GIS-LBSN*, pages 9–16, 2011.

[17] F. Giannotti, M. Nanni, D. Pedreschi, F. Pinelli, C. Renso, S. Rinzivillo, and R. Trasarti. Unveiling the complexity of human mobility by querying and mining massive trajectory data. *VLDB*, 20(5):695–719, Oct. 2011.

[18] F. Girardin, F. Calabrese, F. Fiore, C. Ratti, and J. Blat. Digital footprinting: Uncovering tourists with user-generated content. *PERCOM*, 7(4):36–43, 2008.

[19] M. Gonzalez, C. Hidalgo, and A. Barabasi. Understanding individual human mobility patterns. *Nature*, 453:779–782, 2008.

[20] B. Gueye, S. Uhlig, and S. Fdida. Investigating the imprecision of ip block-based geolocation. In *PAM*, pages 237–240, 2007.

[21] P. Hui, R. Mortier, M. Piorkowski, T. Henderson, and J. Crowcoft. Planet-scale human mobility measurement. In *HotPlanet*, 2010.

[22] D. Kupiszewska and B. Nowok. *International Migration in Europe: Data, Models and Estimates*, chapter Comparability of Statistics on International Flows in the European Union, pages 41–71. John Wiley and Sons Ltd., 2008.

[23] R. Lee. The outlook for population growth. *Science*, 333:569–573, 2011.

[24] X. Lu, L. Bengtsson, and P. Holme. Predictability of population displacement after the 2010 haiti earthquake. *PNAS*, 109(29):11576–11581, 2012.

[25] R. McLean, W. Sanders, and W. Stroup. A unified approach to mixed linear models. *Am Stat*, pages 54–64, 1991.

[26] U. Nations. Recommendations on statistics of international migration. Technical Report Statistical Papers Series M, No. 58, Rev.1, Statistics Division, Department of Economic and Social Affairs, United Nations, New York, 1998.

[27] E. Neumayer. Unequal access to foreign spaces: how states use visa restrictions to regulate mobility in a globalized world. *T I Brit Geogr*, 31(1):72–84, 2006.

[28] E. Neumayer. On the detrimental impact of visa restrictions on bilateral trade and foreign direct investment. *Appl Geogr*, 31(3):901–907, 2011.

[29] A. Noulas, S. S., C. Mascolo, and M. Pontil. An empirical study of geographic user activity patterns in foursquare. In *ICWSM*, pages 570–573, 2011.

[30] W. T. Organization. *Collection of tourism expenditure statistics*, 1995. Technical Manual No. 2.

[31] S. Phithakkitnukoon, F. Calabrese, Z. Smoreda, and C. Ratti. Out of Sight Out of Mind–How Our Mobile Social Network Changes during Migration. In *Third international Social Computing (SocialCom)*, pages 515–520. IEEE, 2011.

[32] A. Pitsillidis, Y. Xie, F. Yu, M. Abadi, G. M. Voelker, and S. Savage. How to tell an airport from a home: techniques and applications. In *Hotnets*, pages 13:1–13:6, 2010.

[33] I. Poese, S. Uhlig, M. A. Kaafar, B. Donnet, and B. Gueye. Ip geolocation databases: unreliable? *CCR*, 41(2):53–56, 2011.

[34] E. Pultar and M. Raubal. A case for space: Physical and virtual location requirements in the couchsurfing social network. In *LBSN*, pages 88–91, 2009.

[35] S. Rinzivillo, S. Mainardi, F. Pezzoni, M. Coscia, D. Pedreschi, and F. Giannotti. Discovering the Geographical Borders of Human Mobility. *Kuenstliche Inteligenz*, 26(3):253–260, 2012.

[36] A. Rogers and J. Raymer. Origin dependence, secondary migration, and the indirect estimation of migration flows from population stocks. *Journal of Population Research*, 22(1):1–19, 2005.

[37] Y. Shavitt and N. Zilberman. A geolocation databases study. *JSAC*, 29(10):2044–2056, 2011.

[38] J. Stillwell, P. Boden, and A. Dennett. *Population Dynamics and Projection Methods*, chapter Monitoring Who Moves Where: Information Systems for Internal and International Migration, pages 115–140. Springer, 2011.

[39] E. Zagheni and I. Weber. You are where you email: Using e-mail data to estimate international migration rates. In *WebSci*, pages 497–506, 2012.