You are where you E-mail: Using E-mail Data to Estimate International Migration Rates*

Emilio Zagheni
Max Planck Inst. for Demographic Research
Konrad-Zuse-Str. 1, Rostock, Germany
zagheni@demogr.mpg.de

Ingmar Weber
Yahoo! Research Barcelona
Av. Diagonal 177, Barcelona, Spain
ingmar@yahoo-inc.com

ABSTRACT

International migration is one of the major determinants of demographic change. Although efforts to produce comparable statistics are underway, estimates of demographic flows are inexistent, outdated, or largely inconsistent, for most countries. We estimate age and gender-specific migration rates using data extracted from a large sample of Yahoo! e-mail messages. Self-reported age and gender of anonymized e-mail users were linked to the geographic locations (mapped from IP addresses) from where users sent e-mail messages over time (2009-2011). The users' country of residence over time was inferred as the one from where most e-mail messages were sent. Our estimates of age profiles of migration are qualitatively consistent with existing administrative data sources. Selection bias generates uncertainty for estimates at one point in time, especially for developing countries. However, our approach allows us to compare in a reliable way migration trends of females and males. We document the recent increase in human mobility and we observe that female mobility has been increasing at a faster pace. Our findings suggest that e-mail data may complement existing migration data, resolve inconsistencies arising from different definitions of migration, and provide new and rich information on mobility patterns and social networks of migrants. The use of digital records for demographic research has the potential to become particularly important for developing countries, where the diffusion of Internet will be faster than the development of mature demographic registration systems.

*We thank Josh Goldstein for his helpful suggestions throughout the development of this study. The manuscript also benefited from the comments of James Raymer, our colleagues at MPIDR and Yahoo! Research, the participants to the Alp-Pop meeting (2012) and to the Demography seminar of CEDEPLAR (2011). We thank four anonymous referees for their constructive comments. This research was partially supported by MPIDR and by the Torres Quevedo Program of the Spanish Ministry of Science and Innovation, co-funded by the European Social Fund, and by the Spanish Centre for the Development of Industrial Technology under the CENIT program, project CEN-20101037, "Social Media" http://www.cenitsocialmedia.es/

WebSci 2012, June 22–24, 2012, Evanston, Illinois, USA.

Author Keywords

Demographics, Migration, Mobility, E-mail data

ACM Classification Keywords

J.4 Social and Behavioral Sciences; H.4.3 Communications Applications

General Terms

Experimentation, Human Factors

INTRODUCTION

International migration is an important driver of demographic growth in many countries [13], and a major source of uncertainty in demographic projections carried out by the United Nations [14]. Migrations have relevant social, economic, and environmental consequences that are felt for decades in both sending and receiving countries. Although there is growing interest in quantifying international migration flows, available statistics are largely inexistent, inconsistent across countries, or outdated [5, 12].

In this article, we use an innovative approach to evaluate global migration rates. We estimate age and gender-specific migration rates using data extracted from a large sample of anonymized Yahoo! e-mail messages. For each e-mail message in the sample, we know the self-reported age and gender of the anonymized sender, as well as the IP address assigned to the device used to send the e-mail. Every machine connected to the Internet is assigned an IP address, which is a numerical label that is typically given to Internet service providers within blocks related to geographic regions. Therefore, IP addresses can be used to identify the geographic region (e.g., city, county, or country) from where the device is connected to the Internet. The actual IP addresses were not used in this study: they were mapped to locations and then removed. User names were also removed and replaced by large random integers.

The available data set allows us to track, with a reasonable level of accuracy, the movements of millions of anonymized e-mail users worldwide. Our work differs from previous research on the use of digital records to estimate mobility patterns in two main ways. First, the geographic scale of our work is global. Previous research focused on mobility within a country, or relatively small regions like cities. Second, we evaluate cross-border movements for a fairly long time frame (about 2 years), compared to the studies in the

literature. This allows us to make inferences on changes of residence and international migration, in addition to estimates of short-term mobility.

The classic sources of information on international flows of migrants are demographic registration systems. This form of data collection suffers huge problems of underestimation because migrants tend not to register after they move. Moreover, for those who do register, there is often a long lag between the actual change of residence and the registration. In addition, in most developing countries, registration systems are not even in place. The work that we present in this article has the potential to transform the way in which migration statistics are compiled.

This article is organized as follows. The next two sections give some background information, discuss the literature and describe the data that we used. Then a section on methods provides details on our estimation approach and the associated challenges, in particular selection bias. After that, there is a section that presents our main results regarding age and gender-specific migration rates and recent trends in human mobility. The article ends with a discussion of our results, their relevance, the limitations of our work and future directions for research in this area.

BACKGROUND AND RELATED WORK

There are two main types of data and methods that have been developed to estimate human mobility and international migration. First, classic demographic data sources (e.g., censuses, registration systems, and sample surveys) have been used to estimate stocks and flows of migrants. Second, the increasingly availability of geo-located digital records has made possible the development of new approaches to evaluate short-term localized mobility or tourist paths. In this section, we briefly review the main contributions of the two research fields.

International migration is a central research area in demography. However, data on flows of migrants, in particular by age and gender, are almost inexistent. When some data exist, they are largely inconsistent across countries because of different definitions of migration and because of different methods of data collection. For example, 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. [20]

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. [1, 6, 12] More often, data on flows are not available at all. In some cases, flows can be roughly estimated indirectly, by evaluating differences over time in migration stocks obtained from census data. [18] To analyze historical trends, classic demographic techniques are the best available tools. However, there are a number of limitations to evaluate current trends. In particular, under or over estimation and delays. There may be a considerable lag between events like change of residence, and their registration. In some cases, people may not register at all (e.g., undocumented migrants) or they may be simultaneously registered in two different cities (e.g., in some countries people may have an economic incentive not to unregister from the place of origin). This generates bias in migration estimates. In addition, it often takes a long time to compile migration statistics. Therefore, there is a long period of time between data collection and publication of statistics.

Digital records are becoming an increasingly important source of information to study human mobility at a global scale [11], with large implications for social sciences and epidemiology. Data from a large bill-tracking website (`wheresgeorge.com`), and from trajectories of trackable items recorded at the website `geocaching.com`, have been used to infer statistical regularities about long distance human mobility. [4] More often, geo-located records have been used to evaluate spatial mobility at city or regional level. For instance, Ferrari et al. [10] analyzed urban patterns, for the city of New York, from Twitter data. Routine whereabouts have also been extracted from applications that allow people to share their locations with friends (e.g., from Google Latitude [9] and Foursquare [15]). Mobile phone data for a small sample of people have been tracked over a period of 9 months to understand frequent mobility patterns [3]. Geo-tagged pictures in Flickr [7, 8] and recommendations posted on the social network service for travelers CouchSurfing [16] have proved useful to reconstruct and improve travel itineraries for tourists.

The literature on demographic methods has largely ignored the recent developments in Web data mining. In this article, we show that the statistical analysis of digital records may generate new information for migration studies and provide insightful contributions to the field of demography.

DATA

For this study, we used four types of data. First, geographic information, over time, for a large sample of Yahoo! e-mail messages. Second, self-reported demographic information of Yahoo! users. Third, migration rates for 11 European countries that Eurostat gathered from national statistical agencies. Fourth, international statistics on Internet penetration rates by age and gender.

We analyzed a large sample of Yahoo! e-mail messages sent between September 2009 and June 2011. For each message, we know the date when it was sent and the geographic location from where it was sent. In addition, we could link the message with the person who sent it, and with the user's de-

mographic information (date of birth and gender), that was self reported when he or she signed up for a Yahoo! account. We estimated the geographic location from where each e-mail message was sent using the IP address of the user. All IP addresses were removed after they were mapped to locations, and all user names were mapped to integers using a non-invertible hash function. All analysis was done in aggregate, and individual users were not identified nor identifiable. Only information for users who were at least 14 years old was used. There may be some uncertainty in estimates of geographic locations at the city level. However, IP addresses provide extremely accurate estimates of the country from where messages were sent.

The data went through a cleaning process in order to discard spammers and users for whom we had only a small number of observations (e-mail messages sent), or only observations covering a very limited temporal frame. Spammers were identified using a Yahoo!-internal trust score assigned to each e-mail user. In addition to spammers, we discarded users for whom we did not have at least one observation in each of 9 distinct months. After the preprocessing steps described above, our data set contained data for about 43 million users. About half of these users registered in the US. For most European countries and large developing or middle-income countries, such as India and Brazil, the sample size is between 500,000 and 2 million users.

Although it is likely that some users reported false information about their age and gender, we believe that the large majority of users reported true and accurate information. Internal verification of the data, and consistency checks based on independent sources, such as data on web searches, confirmed that the data on the demographics of users is highly reliable [21].

Eurostat collects data from European national statistical agencies and publishes data on migration flows by age.[1] The data sources are typically administrative records or national surveys. Although the definitions of migration are not completely consistent across countries, emigration is usually intended as long-term migration. Therefore, emigration denotes the action of changing the usual residence for a period of at least 12 months. Data from Eurostat, available for 2009 for 11 countries, are used to validate the age profile of our estimates based on e-mail data.

In order to correct for selection bias, we developed a model informed by statistics on Internet penetration rates by age and gender. These data are obtained from the UNECE Statistical Database on Internet use by age and gender.[2] The database provides information on the proportion of Internet users, over a period of 3 months, by age group and gender, for a large number of countries.

[1]Data on migration flows are available at: http://epp.eurostat.ec.europa.eu/portal/page/portal/population/data/database

[2]http://w3.unece.org/pxweb/dialog/varval.asp?ma=02_GEICT_InternetUse_r&path=../database/STAT/30-GE/09-Science_ICT/&lang=1&ti=Internet+use+by+age+and+gender

METHODS

Estimation of emigration profiles by age and gender

The IP address reveals the geographic location of the user (e.g., city, county and country) when he or she sends e-mail messages. At the country level, this localization is highly accurate and is often used to restrict access to certain types of content from countries with incompatible copyright laws. At finer granularities the localization is noisier, but is still used for things like localized advertising. To the extent that the user regularly sends e-mail messages, it is in principle possible to track his or her changes of residence. For example, if a user regularly sends most of his or her e-mails from France for 6 months, and then regularly sends most of his or her e-mails from the United States for 6 months, we could infer that the user has changed residence over the course of the year from France to the United States.

Our main goal is to analyze the mobility patterns of Yahoo! users in order to estimate population-level age and gender-specific migration rates, and to evaluate changes over time in mobility. The process involves 3 main steps. First, we choose a definition of migration and estimate relevant rates for Yahoo! users accordingly. Second, we formulate a model to infer rates at the population level. This involves adjusting estimates in order to correct for the selection bias associated with different Internet penetration rates. When the Internet penetration rate is very high, we expect the population of Yahoo! users to be highly representative of the entire population. When the Internet penetration is low, we expect the population of Yahoo! users to over-represent the fraction of the population with higher education and income of urban areas. Third, we validate our profile of age-specific rates against Eurostat data.

We define migration as a change of usual residence between the period from 09-2009 to 06-2010, and the period from 07-2010 to 06-2011. A change of residence can be estimated to the extent that there is a large number of observations, over a rather long period of time, and that the ratio between the variability in the locations visited by the user, and the frequency of messages sent from the usual residence, is not too high. We estimate that the user moves his or her residence from country A to country B if 3 conditions are met. First, country A is the country from where the highest frequency of messages were sent (the modal country) in the first portion of the sample (between 09-2009 and 06-2010); Second, country B is the modal country for the second portion of the sample (between 07-2010 and 06-2011). Third, when we sample with replacement a large number of messages sent by the user, in more than 80% of the cases, country A is the modal country in the first portion of the sample, and country B is the modal country in the second portion of the sample. This last condition, inspired by bootstrap techniques, indicates that the country of usual residence is selected only when there is a high level of confidence that the user spent most of his or her time in the country considered. In other words, if the number of messages sent from the modal country during a particular period of time is not much larger than the number of messages sent from a different country, during the same period of time, we would be highly uncertain about

which of the two countries is the country of usual residence for the user. If the uncertainty is too high (in less than 80% of the resampled messages we would obtain the same country as the country of usual residence), then we prefer not to make inferences on the usual residence and potential change of residence for the user considered.

Age and gender-specific out-migration rates for a particular country are estimated as the fraction of two quantities. On the numerator, there is the number of users that, according to the rule previously described, changed residence from the country considered to any other country in the world. On the denominator, there is the number of users who had their usual residence in the reference country during the first period (of the two periods considered).

Selection bias correction

The profiles by age and gender are adjusted in two ways. First, they are rescaled: the entire profile is shifted upward or downward by a factor that makes estimates for European countries match the order of magnitude of annual age-specific rates provided by Eurostat for the year 2009. Second, the estimated number of migrants, by age group and gender, is multiplied by a correction factor to adjust for over-representation of more educated and mobile people in groups for which the Internet penetration is low. The correction factor, CF is:

$$CF = \frac{p_{gac}(e^{-k} - 1)}{(e^{-kp_{gac}} - 1)}$$

where p_{gac} is the Internet penetration rate for gender g, age group a, and country c. k is a parameter that measures the intensity of the impact of lower Internet penetration rates on the selection of the more mobile people in the sample of users. For a large number of countries, data on Internet penetration rates, by age and gender, are available from the UNECE Statistical Database on Internet use. Therefore, the correction factor is in practice a function of k. Figure 1 shows the assumed relationship between the correction factor and Internet penetration rates for different choices of the parameter k. If the Internet penetration rate is equal to 1, meaning that everybody uses Internet, then the correction factor is equal to 1. In other words, no correction is needed. The underlying assumption is that when Internet penetration is very high, then the population of Yahoo! users is highly representative of the entire population. As Internet penetration decreases, the correction factor becomes smaller: a correction factor less than 1 multiplies the observed number of migrants in the Yahoo! population to adjust for the fact that Yahoo! users may come from a selected population that is more mobile than the average population. The parameter k determines how fast the selection bias increases with decreases in Internet penetration rates. Low values of k imply that the selection bias is small, even at very low Internet penetration rates. Conversely, high values of k mean that the selection bias increases substantially when Internet penetration rates become lower.

In summary, we introduce two parameters, a level parameter l that rescales profiles up or down, and a shape parameter k,

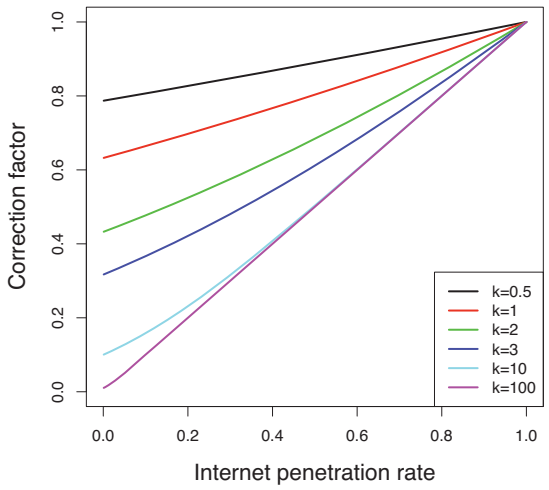

Figure 1. Illustrative relationship between the correction factor for selection bias and Internet penetration rates for different choices of the parameter k.

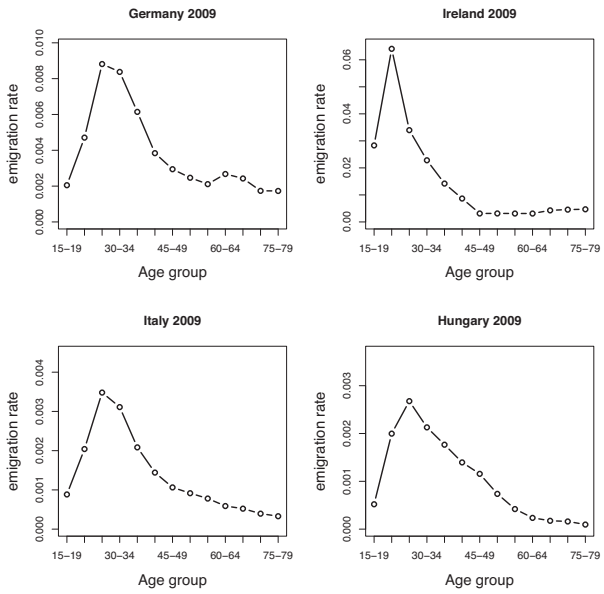

Figure 2. Age profiles of emigration rates for 4 European countries gathered from Eurostat using data from national statistical agencies. Data refer to the year 2009. Different scales are shown for each country.

that determines the effect of low Internet penetration rates on the selection bias of the sample. The parameters are estimated by calibrating our preliminary estimates for 11 European countries against data on age-specific emigration rates published by Eurostat for the same countries, for the year 2009. Figure 2 shows the age profiles of emigration rates published by Eurostat for 4 European countries. The 4 profiles show the typical qualitative features of migration rates: young adults have the highest migration rates. Past the peak,

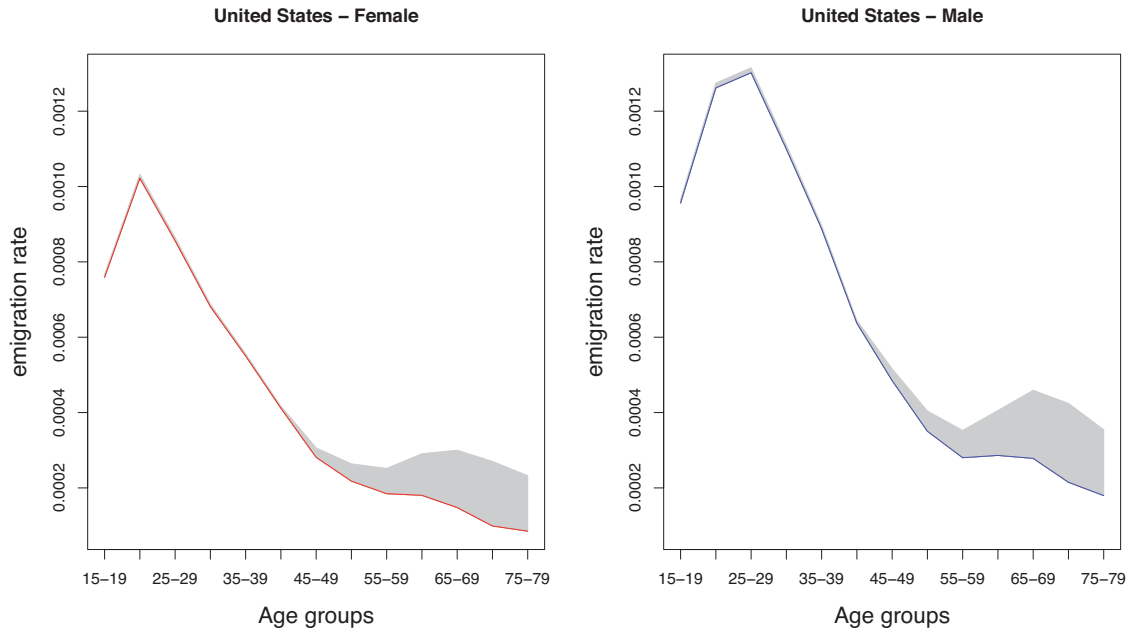

Figure 3. Estimates from e-mail data (2009-2011) of age and gender-specific emigration rates for the US. The gray area represents the size of the bias correction.

migration rates tend to decrease with age. The pace at which they decrease depends on the type of migration processes in the country (e.g., whether migration is mostly related to students' mobility or job opportunities abroad). In some countries, there is a second, but smaller, peak in migration rates around the age at retirement (people often move to places with milder weather, when they exit the labor force).

We chose the parameters l and k that maximize the likelihood of observing the Eurostat estimates of migration rates for 11 European countries.[3] The parameter l is equal for all countries and all age groups: it simply multiplies all the rates up and down to match the order of magnitude of Eurostat data. k determines the extent to which the observed Yahoo! migrants for each age group and sex are corrected (downward) to account for selection bias related to Internet penetration rates. We have Eurostat data on age-specific emigration rates for 11 countries. We consider these rates as if they were the 'true' emigration rates, and the equivalent of a success rate in a binomial process, where 'success' means migration. For the 11 countries considered, given the Eurostat success rate, and the Yahoo! sample size for each age group, we can compute the mean and the variance of the number of migrants that we would expect to observe in the Yahoo! sample if the Eurostat rates were the 'true' population rates.[4] Since the sample sizes are very large, we can approximate the distribution of the number of expected migrants with a normal distribution (with mean and variance given by the binomial distribution previously described). Then, for different choices of the parameters l and k, we evaluate the probability density of observing the 'corrected' number of Yahoo! migrants given the Eurostat rates. In other words, we obtain the probability of observing the data, given the parameters, that is the likelihood of the parameters. We chose the values of the parameters that maximize the likelihood, or, in other words, the ones that minimize deviations from the Eurostat estimates. We then used these estimated parameters for all countries.

[3]Denmark, Germany, Ireland, Spain, Italy, Hungary, Netherlands, Austria, Portugal, Finland and Sweden.

[4]For a specific age group, if we let n be the Yahoo! sample size, and p be the Eurostat migration rate, the expected number of migrants in the Yahoo! sample would be $n \times p$ and the variance $n \times p \times (1-p)$.

RESULTS

Estimates of emigration rates

We estimated emigration rates, by age and gender, for a large number of countries, using the methods that we described in the previous section. Here, we show and discuss the results for two cases that are particularly instructive: The United States and the Philippines.

Figure 3 shows our estimates, based on e-mail data, of age and gender-specific profiles of emigration rates for the United States, during the period for which we have data (Sept. 2009 - June 2011). The solid lines are the estimates obtained after correcting for selection bias. The gray area represents the size of the bias correction. Without any correction for bias, the point estimates would be at the upper end of the gray area. For the youngest age groups, Internet use is almost universal and the selection bias is estimated to be almost inexistent. For older age groups, there is some uncertainty regarding the extent of the peak in international migration at retirement age.

We want to make two important observations about the

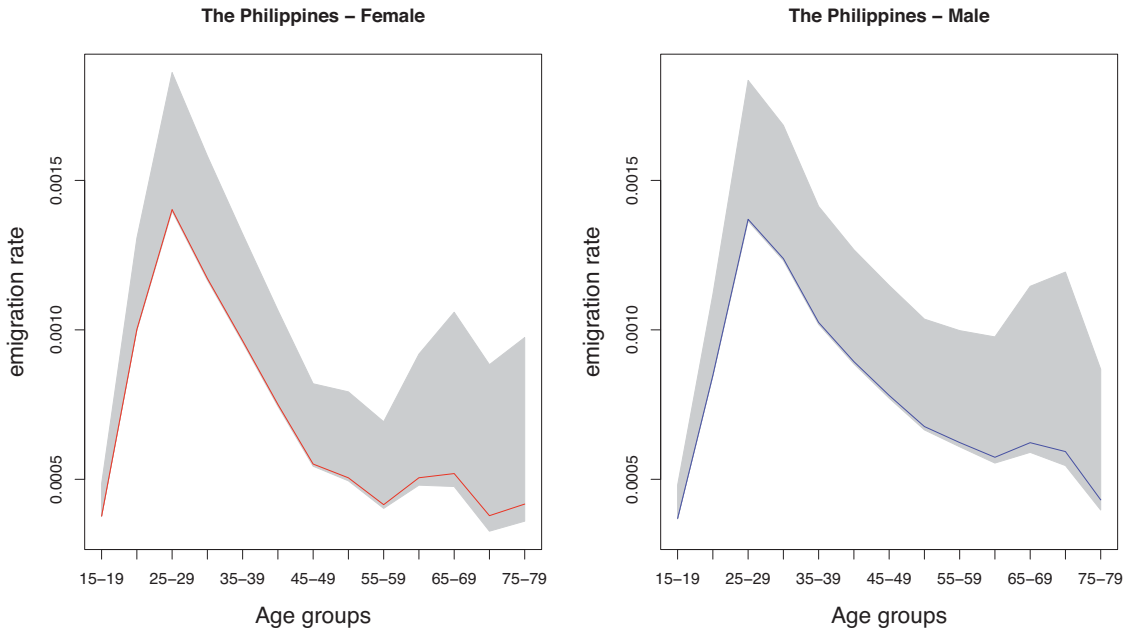

Figure 4. Estimates from e-mail data (2009-2011) of age and gender-specific emigration rates for the Philippines. The gray area represents the sensitivity of the estimates to the choice of the parameter for selection bias correction.

sources of uncertainty for these estimates. First, the sample sizes are so large that standard errors are very close to zero. For the United States, the figures were obtained using a sample of about 6 million males and 7.5 million females. Therefore, all the uncertainty is related to bias. Second, in Figure 3 we show only the bias associated to differentials in Internet penetration rates by age and sex. However, there is a potential source of bias also in the level of our estimates. The level parameter l is calibrated against European data, which are treated as 'ground truth' for European countries. In reality, registration systems suffer problems of underestimation. Thus, it is possible that we may slightly misestimate the migration rates levels.

It is relevant to observe that the age profiles of our estimates conform to regularities in age profiles of migration rates described in the literature [17, 19], and to qualitative features of Eurostat estimates (see, for instance, Figure 2). This is true not only for the United States, but for almost all the countries we analyzed. Our approach has some limitations regarding the estimation of unbiased migration levels, but allows us to evaluate relative differences in migration rates. For instance, we see that in the United States international mobility tends to be higher for young males than for young females. This is novel information, because in the United States there is not a system that keeps track of people who move to other countries. As a country of in-migration, most of the statistics collected in the United States are about people who move into the country. However, data on out-migration are also very important to study phenomena such as return or circular migration.

For the specific case of the United States, we also looked into the sub-population of those who spent some time in Mexico, given the relevance of the Mexico-US border. We observed that the majority of those who changed residence from Mexico to the United States either spent some time in the United States before actually moving to the United States, or went back to visit Mexico after moving their residence to the United States. The individuals that are more mobile across the Mexico-US border are those in their 30s. People in their 50s and older tend not to visit Mexico during the months immediately after they have moved to the United States.

The United States is a case where estimates with relatively low uncertainty can be obtained. That is true also for most countries with high Internet penetration rates (e.g., European countries). Our methods can also be applied to developing countries. There would be more uncertainty in the estimates, but also high potential, since the spread of the Internet may be faster than the development of registration systems. Figure 4 shows our estimates of age and gender-specific emigration rates for the Philippines. The solid lines are the estimates obtained after correcting for selection bias. The gray areas represent the sensitivity of our estimates to changes in the parameter (k) that regulates the intensity of the impact of penetration rates on selection bias. The value of k that generates the solid lines is 20. The gray areas show the range of possibilities when the parameter is between 5 and 35. For countries with high Internet penetration rates, the choice of the parameter k affects the estimates of profiles very little. For countries with low penetration rates, the impact is large and reflects our uncertainty about the importance of the selection bias. One of the reasons that explains why we have larger uncertainty for developing countries is

that our model is calibrated against data for European countries, that is countries with high Internet penetration rates. Data for countries with low Internet penetration rates would have a large leverage on the estimate of k. However there is no such data to calibrate our model. At the same time, estimates for countries with high Internet penetration rates are not very sensitive to changes in k, whereas estimates for countries with low Internet penetration rates are. Since we only have statistics from European countries, the likelihood function with respect to the parameter k tends to be fairly flat, meaning that we have rather high uncertainty. As we improve our models and introduce additional data for the calibration process, we expect to be able to provide more accurate statements on the levels and shapes of profiles for middle-income and developing countries. For instance, the US Census "American Community Survey"[5] provides information on where the respondent was living a year before the survey. That gives us information on age and gender-specific flows from the Philippines to the United States, that could be used to calibrate the overall flows from the Philippines to all other countries.

Despite the limitations related to potential bias in our estimates, the case of the Philippines is particularly interesting for comparisons of migration rates by gender. The mobility rates of females and males are very similar. This is related to the rising number of female migrants in Asian countries since the 1990s, as a consequence of the increased demand for domestic workers in the developed world. [2] Our estimates capture this trend and give us a picture of the relative importance of female migration, with respect to male migration at one point in time. In the next section, we will discuss more in details what our data can tell us regarding trends over time.

Mobility rates over time

In this section, we show some results about changes in mobility rates over time. In other words, we address the following general questions: Are people moving across borders more or less than at the end of 2009? Are there differential trends by age and gender? These substantive questions are relevant to understand the impact of economic downturns or upturns on short-term mobility (e.g., seasonal migration), and on tourism. Moreover, by looking at differences in rates over time, instead of levels, the issue of selection bias becomes much less relevant.

We split our data set in 7 trimesters (Sept-Nov '09; Dec'09-Feb'10; ... ; Mar-May '11). For each trimester, we evaluated the rate of mobility by age, gender and country of residence, using an approach analogous to the one described in the section on methods, but without bias correction. For example, if the user sent most of his or her messages from country A during the first trimester and then most of his or her messages from country B during the second trimester, we estimated that the user was residing in country A during the first trimester and moved his or her residence to country B during the overall period of 6 months. For two consecutive trimesters, we only considered those users for whom we had at least 3 e-mail messages for each trimester.

We compared the rates of mobility for the same trimesters in two different years. This allowed us to contrast rates of mobility computed with the same method, and to evaluate whether mobility increased or decreased, net of seasonal effects. Figure 5 shows rates of international mobility, by age and gender, for the United States. The rates refer to the same two trimesters (as reference starting time) for consecutive years. They provide some information on the extent to which people residing in the United States went abroad, at least for relatively short periods of time. It is interesting to observe that the age profile of short-term mobility is qualitatively very similar to the one of longer term changes of residence, that we showed in Figure 3. When we consider people residing in the United States from September to November, we see a large increase in mobility rates between 2009 and 2010. When we consider people residing in the United States from December to February, we see a slightly negative decrease in mobility between 2010 and 2011. Overall, mobility is estimated to be larger in 2011 than in 2009. People who are between 25 and 35 years old represent the group with the largest changes in mobility. The general trend of increasing international mobility since 2009, with rapid increase in 2010, is consistent with data on the number of international air passengers, provided by the Bureau of Transportation Statistics.[6]

There may be several reasons behind changes in international mobility for the United States. The economic crisis of 2008 might have reduced the number of international tourists and business trips. At the same time, as people experienced difficulties in selling their own houses, the number of internal and international relocations declined. As the global economy picked up again in 2010, a rebound in mobility rates occurred.

Table 1 shows the relative change in mobility rates for a number of countries in our sample. The symbol ++ indicates that mobility increased for both trimesters considered; - - indicates that mobility decreased for both trimesters; +- indicates that mobility increased during the first trimester considered and decreased during the second trimester considered. For each trimester and country, the sign is in red for the gender that experienced the higher relative increase or the lower relative decrease in mobility. For example, in Spain mobility has increased during both trimesters, for both genders. During the first trimester considered we observed a higher relative increase for males than for females. During the second trimester considered we observed a higher relative increase for females. For Germany, we estimated a decrease in mobility for both genders during the first trimester considered. The relative decrease was smaller for females than for males.

To evaluate changes in mobility we tested several measures. First, we used population-level estimates of mobility rates, by sex and for each trimester. Then, we analyzed only the subset of adult population. Finally, we built an index of mo-

[5] http://www.census.gov/acs

[6] www.bts.gov

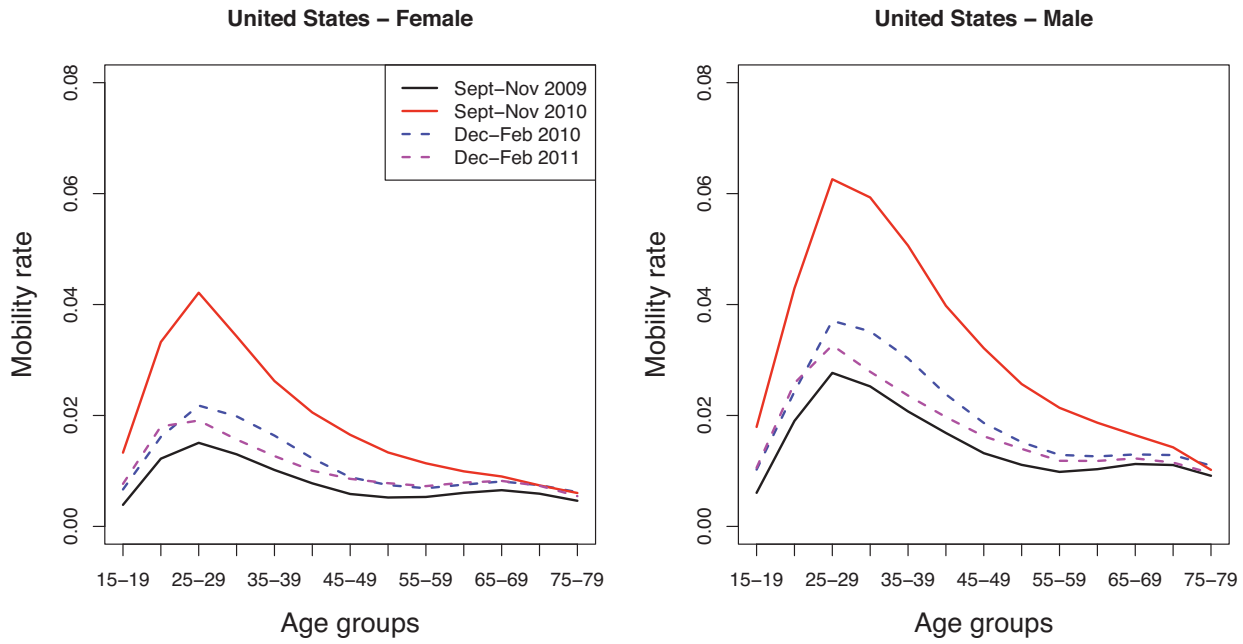

Figure 5. Rates of international mobility, by age and gender, for the United States. The mobility rates for the same two trimesters are estimated for consecutive years to evaluate change over time in mobility rates, net of seasonal effects.

Country	Relative change in mobility (2009-2011)	
	Female	Male
Ireland	++	++
Spain	++	++
Italy	++	++
Brazil	++	++
India	++	++
Nigeria	++	++
Romania	++	++
Russian Federation	++	++
United States	+-	+-
Australia	-+	-+
Japan	-+	-+
Germany	-+	-+
France	-+	-+
China	+-	+-
Switzerland	- -	- -
Taiwan	- -	- -

Table 1. Relative change in mobility rates for selected countries (2009-2011). "++" indicates that mobility increased for both trimesters considered; "- -" indicates that mobility decreased for both trimesters; "+-" indicates that mobility increased during the first trimester considered and decreased during the second trimester considered. For each trimester and country, the sign is in red for the gender that experienced the higher relative increase or the lower relative decrease in mobility.

bility for each country based on the sum of age-specific mobility rates, in order to standardize for differences in the population age structure of Yahoo! users across countries. The value of the relative change is sensitive to the measure used, but the sign of the change (positive or negative) is not sensitive to the choice of the measure. Therefore we presented the signs of change (Table 1), for which we have a higher level of confidence (compared to point estimates). We observed that the increase in international mobility since 2009 has been a global phenomenon. For most countries in our sample, international mobility increased for both trimesters considered. There are a number of countries where an increase in one trimester was followed by a decrease in the next, or vice versa. However, only in a very small number of countries, such as Switzerland and Taiwan, we observed a decrease in mobility for both trimesters.

We estimated that the pace of increase in mobility has been higher for females than for males for the large majority of the countries in our sample. We did not see a consistent trend in changes by age group. In some countries, like Mexico, mobility increased more for people in the age group 15-25. For the United States, it was the age group 25-35 the one that experienced the highest rates of change. For other countries, the change in rates was more homogeneous across age groups.

DISCUSSION AND FUTURE DIRECTIONS

In this article, we presented an innovative approach to evaluate global migration and mobility rates using digital records. The results are encouraging: we could estimate migration rates that are consistent with the ones published by those few countries that compile migration statistics. By using the same method for all geographic regions, we obtained country statistics in a consistent way, and we generated new information for those countries that do not have registration

systems in place (e.g., developing countries), or that do not collect data on out-migration (e.g., the United States).

Substantively, we documented a global trend of increasing mobility. We observed that mobility is growing at a faster pace for females than males. The rate of increase for different age groups varies across countries. Methodologically, we addressed the issue of selection bias for samples such as the one of Yahoo! users. Given the huge sample size, the standard errors of our estimates are close to zero. Bias is the main source of uncertainty. We proposed a method to correct our estimates for selection bias.

Our interdisciplinary work is relevant both for demographic research and for Web science. The use of geo-located digital records has the potential to transform the way in which certain types of demographic data are collected and interpreted. At the same time, statistical inference based on non-representative samples of the population may generate challenges for Web data mining. Tools that have been developed by historical demographers to analyze convenience samples (e.g., incomplete parish registries or historical tax registration systems) may potentially become relevant for the study of Web data sources.

Digital records that provide information on the geographic location of users over time may help us address a series of research questions. Geo-located records can be used, for example, to estimate the short and medium-term consequences of environmental disasters on mobility. Consider an earthquake or the nuclear crisis in Japan. We could keep track of the mobility pattern of those who resided in the Fukushima region before the crisis. To what extent did people leave for a short time and then went back? To what extent did people relocate far away? Where did people move to? These questions can be addressed very effectively using digital records, which would provide the most timely estimates of mobility, in case of future environmental crises. A census or a traditional survey would take months and a large amount of resources.

The availability of geo-located records could improve our understanding of the role of social networks in migration processes. By keeping track of those who moved, we could evaluate how clustered migration processes are. For instance, we could assess the extent to which people from a specific county in Mexico tend to move almost exclusively to a specific county in the United States. That is presumably related to the strength of social networks across borders, and to the number of visits that migrants make to their home towns. Analogously, we could evaluate whether sending a high proportions of e-mail messages to a particular country (which is a proxy for having a strong social network in the country) is related to the decision of actually moving to the country.

There is a large potential in publicly available data sources like Twitter posts. Data from the online microblogging service combine information on geographic locations with content of blog posts that could be used for sentiment analysis. In addition to mobility or migration rates, we could evaluate sentiments pro or against migration for different geographic areas. This would help us understand how sentiments change near an international border or in regions with different migration rates and economic conditions. The public nature of Twitter 'tweets' make them particularly appealing because they could be used to produce statistics on mobility rates on a regular basis and in a fully transparent way. Similar analyses could be done using e-mail data, which provide a huge amount of information both on mobility and on sentiments. However, the private nature of e-mail data limits opportunities for independent reproduction of results. It is more and more important to develop models for data sharing between private companies and the academic world, that allow for both protection of users' privacy and private companies' interests, as well as reproducibility in scientific publishing.

Analyzing e-mail data is challenging. E-mail use varies with age, sex, and level of Internet penetration in the country. The demographics of e-mail use may be different for different providers. Some users may have multiple accounts with a number of providers, and the frequency with which they use one or the other may change over time. Some accounts may be shared by more than one user. In this article, we thoroughly addressed the problem of selection bias correction. We believe that it is important to use statistical tools to produce estimates that correct for the fact that the group of users in the sample, although very large, is not representative of the entire population. We think that more work in this area should be done. The creation of a 'panel' of users is a promising direction for future research. For those countries with a large number of users, it may be relevant to select a group of users who are active over the entire period of study and for whom we have reliable demographic and geographic information over time. For these users, mobility could be evaluated over a continuous interval of time, instead of fixed periods (e.g., trimesters), that may be problematic in certain circumstances. Statistical weights to correct for underrepresentation of specific demographic groups or geographic areas could be generated with methods analogous to the ones used in standard sample surveys.

With this article, we wanted to identify some of the opportunities and challenges for research at the intersection of demographic analysis and Web data mining. What we addressed so far is only the tip of the iceberg. We showed some descriptive results and discussed some promising areas for the future. We believe that new data and new methods are the sources of theoretical advances. Demography has been a data-driven discipline since its birth, about 350 years ago. In 1661, John Graunt published the very first life table after compiling data that had been collected for 'marketing' purposes', that is to evaluate the number of potential customers (i.e., live people by age) in London. Since then, data collection and development of formal methods have sustained most of the major advances in our understanding of population processes. Today, Web science and demographic research may give fresh impetus to each other.

REFERENCES

1. Abel, G. Estimation of International Migration Flow Tables in Europe. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2010) 173:797-825.

2. Asis, M. When Men and Women Migrate: Comparing Gendered Migration in Asia. *Proceedings of the Meeting on "Migration and Mobility and how this Movement Affects Women"*. United Nations Division for the Advancement of Women (DAW) 2004.

3. Bayir, M.A., Demirbas, M., and Eagle, N. 2009. Discovering Spatiotemporal Mobility Profiles of Cellphone Users. *World of Wireless, Mobile and Multimedia Networks & Workshops, 2009. WoWMoM 2009. IEEE International Symposium on a* (2009) 1-9.

4. Brockmann, D., and Theis, F. Money Circulation, Trackable Items, and the Emergence of Universal Human Mobility Patterns. *Pervasive Computing, IEEE*. (2008) 7(4):28-35

5. Cohen, J.E., Roig, M., Reuman, D.C., GoGwilt, C. International Migration beyond Gravity: A Statistical Model for Use in Population Projections. *Proceedings of the National Academy of Sciences of the U.S.A* (2008) 105(40):15269-74.

6. De Beer, J., Raymer, R., Van der Erf, R., and Van Wissen, L. Overcoming the Problems of Inconsistent International Migration Data: A New Method Applied to Flows in Europe. *European Journal of Population* (2010) 26(4):459-481

7. De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., and Yu, C. Constructing Travel Itineraries from Tagged Geo-temporal Breadcrumbs. *Proceedings of the 19th International Conference on World Wide Web* (2010) 1083-1084

8. De Choudhury, M., Feldman, M., Amer-Yahia, S., Golbandi, N., Lempel, R., and Yu, C. Automatic Construction of Travel Itineraries Using Social Breadcrumbs. *Proceedings of the 21st ACM Conference on Hypertext and Hypermedia* (2010) 35-44

9. Ferrari, L. and Mamei, M. Discovering Daily Routines from Google Latitude with Topic Models. *2011 IEEE International Conference on Pervasive Computing and Communications (PERCOM Workshop)* (2011) 432-437.

10. Ferrari, L., Rosi, A., Mamei, M., and Zambonelli, F. Extracting Urban Patterns from Location-based Social Networks. In *Proc. LBSN '11 ACM SIGSPATIAL*, ACM Press (2011)

11. Hui, P., Mortier, R., Piorkowski, M., Henderson, T. and Crowcoft, J. Planet-scale Human Mobility Measurement. In *Proc. HotPlanet '10*, ACM Press (2010)

12. Kupiszewska, D. and Nowok, B. Comparability of Statistics on International Flows in the European Union. (2008) Pp. 41-71 in *International Migration in Europe: Data, Models and Estimates*, J. Raymer, F. Willekens, Eds. (John Wiley and Sons Ltd., Chichester, UK).

13. Lee, R. The Outlook for Population Growth. *Science* (2011) 333:569-573.

14. National Research Council, Commission on Behavioral and Social Sciences and Education, *Beyond Six Billion: Forecasting the World's Population*, J. Bongaarts, R. A. Bulatao, Eds. (National Academies Press, Washington, DC, 2000).

15. Noulas, A., Scellato S., Mascolo, C. and Pontil, M. An Empirical Study of Geographic User Activity Patterns in Foursquare. *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (2011) 570-573

16. Pultar, E. and Raubal, M. A Case for Space: Physical and Virtual Location Requirements in the CouchSurfing Social Network. *Proceedings of the 2009 International Workshop on Location Based Social Networks* (2009) 88-91.

17. Raymer, J. and Rogers, A. Applying Model Migration Schedules to Represent Age-specific Migration Flows. (2008) Pp. 175-192 in *International Migration in Europe: Data, Models and Estimates*, J. Raymer, F. Willekens, Eds. (John Wiley and Sons Ltd., Chichester, UK).

18. Rogers, A., and Raymer, J. Origin Dependence, Secondary Migration, and the Indirect Estimation of Migration Flows from Population Stocks. *Journal of Population Research*. (2005) 22(1):1-19.

19. Rogers, A. and Castro, L.J. *Model Migration Schedules*. (1981) RR-81-30. Laxenburg: International Institute for Applied Systems Analysis.

20. Stillwell, J., Boden, P. and Dennett, A. Monitoring Who Moves Where: Information Systems for Internal and International Migration. (2011) Pp. 115-140. in *Population Dynamics and Projection Methods* , J. Stillwell, M. Clarke, Eds. Springer.

21. Weber, I. and Castillo, C. The Demographics of Web Search. *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2010) 523-530.