Explaining International Migration in the Skype Network: The Role of Social Network Features

Riivo Kikas
University of Tartu
riivokik@ut.ee

Marlon Dumas
University of Tartu
marlon.dumas@ut.ee

Ando Saabas
Microsoft
ando.saabas@skype.net

ABSTRACT

In recent years, several new ways have appeared for quantifying human migration such as location based smartphone applications and tracking user activity from websites. To show usefulness of these new approaches, we present the results of a study of cross-country migration as observed via login events in the Skype network. We explore possibility to extract human migration and correlate it with institutional statistics. The study demonstrates that a number of social network features are strongly related to net migration from and to a given country, as well as net migration between pairs of countries. Specifically, we find that percentage of international calls, percentage of international links and foreign logins in a country, complemented by gross domestic product, can be used as relatively accurate proxies for estimating migration.

Categories and Subject Descriptors

J.4 [**Social and Behavioral Sciences**]: Sociology; H.3.5 [**Online Information Services**]: Web-based services

Keywords

social networks; human mobility; migration;

1. INTRODUCTION

International human mobility and migration are determinants of demographic changes, which in turn are correlated with socio-economic shifts [5]. Traditionally, human migration has been studied using datasets gathered by institutional statistical organizations, which rely in large part on official government registries. These datasets are often not fresh, they are coarse-grained (typically yearly and at the level of entire country populations), and are in some cases incomparable across countries [4].

The increasing amounts of available Web and social network usage data makes it possible to analyze human migration from new perspectives. A number of studies have been reported on human mobility based on Web log files associating users' website visits to IP addresses, from where an approximate physical location can be determined using geolocation databases. For example, Yahoo E-mail users were tracked for one year by login locations and the data was used to analyze inter-country migration [10].

Mobility datasets extracted from online service usage stand out for their scale (typically worldwide), freshness and fine level of granularity, and are thus a fertile ground for understanding the determinants of mobility and migration. And while online mobility datasets suffer from biases stemming from the fact that only a subset of the overall population uses such services, and errors due for example to IP hiding and proxying, they do not on the other hand suffer from other cross-country biases and discrepancies found in traditional migration datasets, where different data collection and estimation methods are applied for different countries.

This paper reports the results of a study on cross-country medium-to-long-term mobility as observed from locations of user logins in the Skype network. The study analyzes the following questions:

RQ1. To what extent can we explain the net migration of a country (as observed in the Skype network) from social network features aggregated at a country level and/or from country-level socio-economic indicators?

RQ2. To what extent can we explain the net migration between a given pair of countries from social network features aggregated at a country level and/or from country-level socio-economic indicators?

To address RQ1, we study correlations between net migration rate as observed in Skype and potentially relevant features derived from the Skype network (e.g. fraction of international links in a country) and from external data (e.g. economic attributes such as GDP). We then build a regression model to explain net migration of a country based on these features.

In a similar vein, to address RQ2, we build regression models to explain net migration between pairs of countries based on economic, social and geographic variables, as well as social network features, and we analyze the relative predictive power of social network variables versus socio-economic and geographical variables in this context.

The rest of the paper is structured as follows. Section 2 describes the dataset employed in this study and provides a descriptive statistical analysis of its properties. This section also compares the dataset at hand against institutional statistics. Section 3 addresses the research questions for-

SIdEWayS'15, September 1, 2015, Guzelyurt, TRNC, Cyprus.
© 2015 ACM. ISBN 978-1-4503-3798-4/15/09 ...$15.00.
DOI: http://dx.doi.org/10.1145/2806655.2806658.

mulated above. Finally, Section 4 compares this study with related work and Section 5 summarizes the findings and outlines directions for future studies.

2. DATASET AND DESCRIPTIVE ANALYSIS

In this section we discuss the datasets employed, the definition of migration adopted and general observations on the observed migration in the network.

2.1 Datasets

The main dataset (DS0) used in this study consists of aggregated login data for each anonymized user in a uniformly random sample of about 15 million Skype users. For each user and for each month, the dataset tells us how many days the user logged in from particular country during the month in question. The dataset covers logins for the sampled users between 2007 and 2011 (included). There are in total 497,637,658 month-country data points and on average for each user we have 33 month-country pairs. Login data is available for each user from January 2007 or from the month of the user's account creation, whichever is earlier. Thus the length of user login history varies from one user to another, from one month to five years.

Country information for each anonymized user was obtained by IP address geocoding. Geocoding is generally considered accurate at the country level, but is not 100% correct all the time. The geocoding service treats some regions as a separate units, although in some context they might be considered as single country, for example in the official world bank datasets. We removed such cases manually – the majority of which are small colonial islands of France and The Netherlands (Guadeloupe, Martinique, Netherlands Antilles, Reunion), Pacific Ocean islands, but also Palestinian territory.

After geocoding at the country level, the login time series per user allows us to determine which countries a user visited during a given month as well as the country where the user spent the largest number of days during a month. On the other hand, it does not allow us to track the exact sequence of countries visited by a user during a month.

We also make use of two additional datasets: dataset DS2 consists of the Skype graph as of November 2011. This dataset provides us the entire graph of links in the Skype network, with a link representing the fact that one user has the other user in their contact list.

In addition, we use the calling volume between countries, which is obtained by recording origin and destination of a random subset of calls during some short period. This dataset captures number of calls made between each pair of countries.

Datasets DS1 and DS2 are anonymized and users are identified by hashed identifiers. No other data, such as age, gender etc is associated with user data. The call volume dataset is aggregated on country level and does not provide details about exact pairs of calls.

2.2 Definition of migration

The dataset DS0 does not provide us directly which users are migrants nor when they migrated. Thus, we need to adopt a definition of migrant based on observed countries of login. We hereby define a user as a migrant, if they have been in one country for at least five consecutive months and in another country for at least five consecutive months. Setting these time limits prevents counting longer holidays or business visits as migration events.

Note that after migration, an individual might return for short visits to the originating country or to third countries, which in our data shows up as logins from two or more countries during a single month. We offset such short-term after-migration visits by defining a notion of country of residence for each month. Specifically, we adopt a sliding-window definition of country of residence as follows: The residence country of a user on a given month M is the country where the user had more login-days during the period M to $M+3$ months. Thus in the case of short-term trips, the residence country for the month would still be the home country provided that the individual uses Skype regularly when coming back home. In the case of long-term movements however, the new target country becomes country of residence when the movement occurs. The choice of a 3-months sliding window is intended to offset the effects of short-term travels (up to six weeks).

Henceforth, when we use the term migration, we refer to the long term movement (at least five months) and we use the term mobility for short visits up to four months.

2.3 Limitations of the dataset and threats to validity

Before further analysis of the dataset, we clarify below two limitations of the dataset that ought to be kept in mind when interpreting the results.

2.3.1 *Identification of initial country of residence*

In the dataset at hand the ground truth regarding user's real place of residence at the time of creating their Skype account is not directly available. Thus, if a user moves from one country to another, and stays in the target country for a period exceeding the five-months threshold, then said user is assumed to have migrated from the source to the target country. However, it might be that the person was returning from a visit abroad and registered a Skype account while abroad. In this case, the user is merely returning from a visit to their home country rather than migrating. This situation cannot be identified based on the available data.

2.3.2 *Adoption bias*

Skype is not used equally in every country or among different age groups. Multiple factors, such as Internet availability and presence of competing products influence Skype adoption. Estimated migration volumes might thus be biased towards countries with higher Skype adoption rates. Correction for adoption bias might be needed [13], but we do not apply such correction in this study as the goal is not to quantify real migration, but to identify relations between observed migration in the network and internal and external variables.

2.4 General migration statistics

After extracting movement patterns, we see that 58% of users are stationary (i.e. always connected from a single country) while 20% connected from two different countries and 10% from three countries. Based on the above definition of migration event, 15% of users migrated at least once, representing about 2.25 million users. This can be contrasted against the percentage of users who traveled at least once (short or long-term), which is in the order of 42%. There are very few extreme cases of users who migrated after every possible migration threshold, resulting in up to 13 different migration events.

We measured a low correlation (Pearson correlation 0.42) between world bank migration data and Skype migration data. When focusing on Europe however, we see that the net migration in Skype is correlated (Pearson correlation 0.75) with the migration information from EUstat (Figure 1). Lithuania and Latvia stand out as outliers, with Skype data reporting higher net migration from these countries than EUStat, suggesting that Skype data might overestimate the migration volume or some migration observed in Skype for these countries might not be reflected in officially recorded migration statistics.

3. MODELING MIGRATION

To address the two research questions formulated in Section 1, we build regressions models to estimate the net migration rate of a country and migration between pairs of countries from a range of features, including social network features extracted from datasets DS0 and DS1 and from World Bank socio-economic indicators as described below.

3.1 Input features

We consider two types of features: (i) features extracted from the social network (DS1) and from the IP-geocoded login activity; and (ii) socio-economic features extracted from external sources.

Table 1 lists the features in the first category. The table contains both independent variables and dependent variables that will be used as target for estimation. To account for the difference between country sizes, most of the features are normalized by the sedentary population in the country, meaning the number of people who did not migrate out of the country. We only consider countries with a sedentary population of at least 1000 users.

From datasets DS1 and DS2, we extracted the following additional features (independent variables):

- *Outgoing social links* (openness): Fraction of social links going outside of the country divided by social links inside the country. This indicated how international is the country. This variable, ranging from 0 to 1 also indicates how much Skype is used inside the country to some extent.

- *International calls*: Ratio of calls going outside the country divided by all calls from the country. This include audio calls, video calls and calls to landlines. This feature again indicates where the communication is targeted to, inside or outside. The dataset about calls is obtained using sampling during a smaller time period, where as the social structure is complete representation of the network at that time (2011).

Socio-economic features are extracted from World Bank data[1], downloaded in February 2013. We extracted from this dataset features that may directly or indirectly cause human mobility, such as low GDP drives people to leave for countries with better living conditions. The extracted features are mirrored from those used in [10] to explain migration between pairs of countries.

Table 1: Features extracted from dataset DS0 for each country.

Independent variables	**Description**
Sedentary population	Number of users who did not migrate out of the country.
Visited From	Number of short-term visits of residents of the country to another country divided by sedentary population.
Visited From Unique	Number of unique residents of the country who visited another country divided by sedentary population.
Visited To	Number of short-visits from residents to another country.
Local logins	Number of logins made by residents.
Foreign logins	Number of logins made by foreign visitors.
Dependent variables	**Description**
Migrated From	Number of users who moved away from the country (long-term move) divided by sedentary population.
Migrated To	Number of users who moved to the country (long-term move) divided by sedentary population.
Net migration	Number of people who migrated to the country minus number of people who migrated out of the country (*Migrated To - Migrated From*) divided by sedentary population.

Table 2: Pearson correlation between features.

Feature 1	**Feature 2**	**Correlation**
Net migration	Foreign logins	0.597
Net migration	Frac. of Int'l. links	0.482
Net migration	Frac. of Int'l. calls	0.240

In general, there is low correlation between every pair of features. Table 2 shows the three pairs of features with the highest correlation. The highest correlation is between foreign logins and net migration volume, reflecting the fact that migrants use the service to call back to their home country.

3.2 Modeling country net migration

We use ordinary least squares linear regression to build the model. The target variable was net migration rate in Skype. We build the model for different countries and feature subsets to avoid so called target leakage, for example that net migration in Skype can be correlated with number of visitors.

The results of the linear model can be seen in Table 3. The obtained R^2 values for the EU country subset indicate that the migration might be predictable. The high values of goodness of fit can be explained with European countries being similar and the Skype adoption can be more uniform. In the global context, there is definitely more variance across the population sizes and Skype adoption rates.

The variables derived from the social graph, namely fraction of international links and calls, seem to have some relevance in the migration context as they have large weights assigned by the model. The assumption is that in countries with large fraction of foreign links, Skype is mostly used by foreigner moving there and eventually moving back. To an-

[1] `http://data.worldbank.org/`

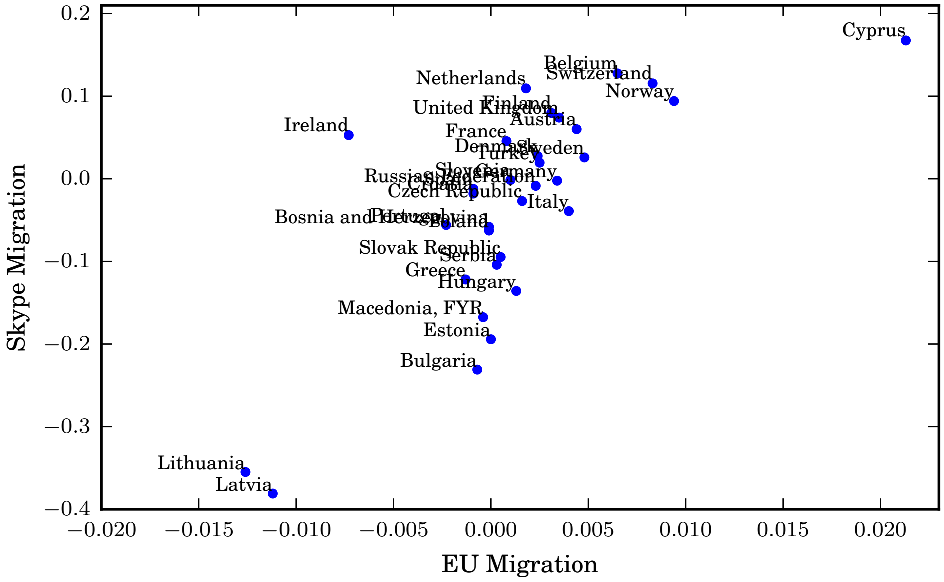

Figure 1: EUstat migration rate (migrants per 1000 people) versus net migration in Skype (net migrants divided by stable population)

Table 3: Model for predicting net migration. Values are feature coefficients, numbers in parentheses are standard errors.

	Dependent variable:	
	Net migration	
	All countries	EU countries
Broadband Internet subscribers per 100	0.023 (0.062)	-0.007^{*} (0.004)
GDP growth annual	0.125 (0.105)	0.003 (0.007)
GDP per capita current US	0.0001^{**} (0.00003)	0.00000^{***} (0.00000)
Internet users per 100	−0.015 (0.031)	0.001 (0.002)
Mobile cellular subscriptions per 100	−0.003 (0.013)	0.001 (0.001)
log(Population according World Bank)	−0.166 (0.430)	0.021 (0.041)
log(Sedentary population)	0.630 (0.680)	−0.052 (0.049)
Visited from	−0.228 (1.565)	0.046 (0.113)
Visited to	−0.084 (0.682)	0.002 (0.046)
Visited from persons	0.00000 (0.00000)	0.00000 (0.00000)
Foreign logins	75.466^{***} (20.298)	1.220 (1.735)
Fraction of International links	22.200^{**} (8.532)	1.359^{**} (0.486)
Fraction of International calls	-10.671^{***} (3.712)	-0.578^{**} (0.212)
Constant	-18.581^{**} (8.762)	−0.638 (0.504)
Observations	92	32
R^2	0.616	0.862
Adjusted R^2	0.552	0.763
Residual Std. Error	2.777 (df = 78)	0.063 (df = 18)
F Statistic	9.617^{***} (df = 13; 78)	8.661^{***} (df = 13; 18)
Note:		*p<0.1; **p<0.05; ***p<0.01

Table 4: Linear model for migration volume between countries. Features in bold are derived from social network data.

	Dependent variable:
	migrations
Import from A->B	0.00000** (0.00000)
Export from A->B	0.00000*** (0.00000)
Bilater trade or taxation treaty	−0.127*** (0.033)
Number of Skype messages	0.089*** (0.029)
Number of Social links	0.580*** (0.031)
Call count	0.085*** (0.017)
Target GDP (per capita)	0.00001*** (0.00000)
Source GDP (per capita)	0.00000** (0.00000)
GDP difference	−0.027* (0.015)
Common language	−0.056 (0.042)
Log weighted distance	0.058*** (0.015)
Shared border	−0.021 (0.037)
Same region	−0.006 (0.032)
Colonial link	0.180** (0.079)
Common civilization (Russett)	−0.053* (0.032)
Same commonwealth	0.289*** (0.055)
Visa	−0.137*** (0.029)
Constant	−3.972*** (0.227)
Observations	1,532
R^2	0.791
Adjusted R^2	0.788
Residual Std. Error	0.444
F Statistic	336.610
Note:	*p<0.1; **p<0.05; ***p<0.01

swer our RQ1, we can say that social features are significant variables in the net migration estimation model.

3.3 Modeling bilateral inter country migration

People migrate to specific countries and select their target based on some criteria, such as existing relationships in the country, better economic situation or geographical closeness. To better understand this process, we built a model to understand determinants of migration volume between pairs of countries and another model for predicting the actual volume.

For each pair of countries, a set of features was extracted that describe the two countries. Here, the features describe relationships between two countries and not a single country as in the last section, therefore there is a difference in the feature set. We used the same features as in [10] and added specific features derived from the Skype social data. The target for the prediction was the migration volume between pair of countries. In the training data, we only kept those pairs where the migration volume was larger than 100 persons and such that all additional data were available.

First, we fit a linear model to understand if there is some effect to the target variable. The model (Table 4) can be considered relatively good with the R^2 score of 0.791. The linear model was fit on the whole data.

To understand the relative explanatory power of the selected features, we built a RandomForest [1] regression model using the same feature set. Random forest is able identify to non-linear relationships in contrast to ordinary linear regression. The dataset was split into two equal parts: the training and testing set. Training and test samples were stratified with respect to target country. This means that for each target country, half of the examples were in training and half of the examples in the testing set. This ensures that one training example for each target country is available in training data and at the same time, not all information about one country is held in the training part.

Table 5: Random forest performance on different feature subset.

Features	$\mathbf{R^2}$	**Median Error**
All + source/target country	0.84	0.23
All	0.84	0.23
Social only	0.83	0.26
Social + trade + GDP	0.83	0.24
Geography + GDP	0.19	0.53
Geography + GDP + trade	0.49	0.55

The results (Table 5) show that most important features are social features (number of social links, number of Skype messages and calls) between pair of countries. These features determine the migration volume the best. Geographic features do not give much information, indicating that migration process is not constrained by geography in the Skype network. To answer RQ2, we can conclude that social network features are the best determinants for cross-country migration volume compared to all other features.

4. RELATED WORK

Most of human migration studies have been carried out using mobile phone data [2] and are thus restricted to national scale as mobile phone carriers operate in a single country. Other examples, where data from online social networks has been studied include Foursquare [8] for city-level human movement. Twitter data [12, 3] has been used to quantify migration in global scale, and user attention patterns in Yahoo Memetracker have been show to have correlation with migration inside Brazil [11].

Global human migration has been studied from Yahoo E-mail data [13, 10] where authors concluded that migration can be estimated from e-mail communication logs. Zagheni et al. [13, 14] estimates net migration from selected countries from socio-economic features. Their analysis incorporates age and Internet adoption bias. They also analyze migration between Philippines and USA and conclude that e-mail communication datasets can potentially be used for analysis of migration between pairs of countries. State et al. [10] examine closer the question of estimating migration between pairs of countries from e-mail data. The socio-economic features described in this work are mirrored from those in [10] and original data is from [7, 6]. However, in our analysis we added social network and communication features (e.g. social links, calls, messages) and we find that these features increase explanatory power of the models.

State et. al [9] have also studied how the online professional network LinkedIn data reflects migration of high-skilled workers. They conclude that LinkedIn data can reveal important information about movement of highly educated workers, which is a relevant aspect of global migration.

5. CONCLUSIONS AND FUTURE WORK

In this work we explored user mobility and migration in Skype. We see that Skype users travel around the world and movement destinations can be explained to some extent by demographic and economic data. We built models to estimate net migration of a country and migration volume between pairs of countries. The results indicate that Skype data can give meaningful input in terms of global human mobility and migration, especially in the EU context. In addition, one of our main contribution is that we utilized information from social networks structure and found that it can yield in better understanding how people move in the real world between countries.

The future work should take into account more fine grained data, i.e., user location with a day resolution and on city level. Also, availability of directed communication data between users would add more meaningful aspects to the study by enabling us to study better what happens to Skype usage during travel, migration or to understand if communicating causes travel or travel causes communication.

Acknowledgments. This research is supported by Microsoft Skype Labs and ERDF via the Software Technology and Applications Competence Centre (STACC). The authors acknowledge the valuable input and comments of Adriana Dumitras and Anna Leontjeva.

6. REFERENCES

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Marta C. Gonzalez, Cesar A. Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *Nature*, 453(7196):779–782, June 2008.

[3] Bartosz Hawelka, Izabela Sitko, Euro Beinat, Stanislav Sobolevsky, Pavlos Kazakopoulos, and Carlo Ratti. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, 41(3):260–271, 2014.

[4] Dorota Kupiszewska and Beata Nowok. *Comparability of Statistics on International Migration Flows in the European Union*, pages 41–71. John Wiley & Sons, Ltd, 2008.

[5] Ronald Lee. The outlook for population growth. *Science*, 333(6042):569–573, 2011.

[6] Eric Neumayer. Unequal access to foreign spaces: how states use visa restrictions to regulate mobility in a globalized world. *Transactions of the Institute of British Geographers*, 31(1):72–84, 2006.

[7] Eric Neumayer. On the detrimental impact of visa restrictions on bilateral trade and foreign direct investment. *Applied geography*, 31(3):901–907, 2011.

[8] Anastasios Noulas, Salvatore Scellato, Renaud Lambiotte, Massimiliano Pontil, and Cecilia Mascolo. A tale of many cities: Universal patterns in human urban mobility. *PLoS ONE*, 7(5):e37027, 05 2012.

[9] Bogdan State, Mario Rodriguez, Dirk Helbing, and Emilio Zagheni. Migration of professionals to the u.s. In LucaMaria Aiello and Daniel McFarland, editors, *Social Informatics*, volume 8851 of *LNCS*, pages 531–543. Springer International Publishing, 2014.

[10] Bogdan State, Ingmar Weber, and Emilio Zagheni. Studying inter-national mobility through ip geolocation. In *Proceedings of the Sixth International Conference on Web Search and Data Mining*, WSDM'13, Rome, Italy, 2013. ACM.

[11] Carmen Vaca-Ruiz, Daniele Quercia, Luca Maria Aiello, and Piero Fraternali. Tracking human migration from online attention. In *Citizen in Sensor Networks*, pages 73–83. Springer International Publishing, 2014.

[12] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, and Bogdan State. Inferring international and internal migration patterns from twitter data. In *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, pages 439–444, Republic and Canton of Geneva, Switzerland, 2014. International World Wide Web Conferences Steering Committee.

[13] Emilio Zagheni and Ingmar Weber. You are where you e-mail: using e-mail data to estimate international migration rates. In *Proceedings of the 3rd Annual ACM Web Science Conference*, WebSci '12, pages 348–351, New York, NY, USA, 2012. ACM.

[14] Emilio Zagheni and Ingmar Weber. Demographic research with non-representative internet data. *International Journal of Manpower*, 36(1):13–25, 2015.