2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)

From Migration Corridors to Clusters: The Value of Google+ Data for Migration Studies

Johnnatan Messias
Universidade Federal
de Minas Gerais
Belo Horizonte, Brazil
johnnatan@dcc.ufmg.br

Fabricio Benevenuto
Universidade Federal
de Minas Gerais
Belo Horizonte, Brazil
fabricio@dcc.ufmg.br

Ingmar Weber
Qatar Computing
Research Institute
Doha, Qatar
iweber@qf.org.qa

Emilio Zagheni
University of Washington
Seattle, USA
emilioz@uw.edu

***Abstract*—**

Recently, there have been considerable efforts to use online data to investigate international migration. These efforts show that Web data are valuable for estimating migration rates and are relatively easy to obtain. However, existing studies have only investigated flows of people along migration corridors, i.e. between *pairs of countries*. In our work, we use data about "places lived" from millions of Google+ users in order to study migration 'clusters', i.e. groups of countries in which individuals have lived sequentially. For the first time, we consider information about more than two countries people have lived in. We argue that these data are very valuable because this type of information is not available in traditional demographic sources which record country-to-country migration flows independent of each other. We show that migration clusters of country triads cannot be identified using information about bilateral flows alone. To demonstrate the additional insights that can be gained by using data about migration clusters, we first develop a model that tries to predict the prevalence of a given triad using only data about its constituent pairs. We then inspect the groups of three countries which are more or less prominent, compared to what we would expect based on bilateral flows alone. Next, we identify a set of features such as a shared language or colonial ties that explain which triple of country pairs are more or less likely to be clustered when looking at country triples. Then we select and contrast a few cases of clusters that provide some qualitative information about what our data set shows. The type of data that we use is potentially available for a number of social media services. We hope that this first study about migration clusters will stimulate the use of Web data for the development of new theories of international migration that could not be tested appropriately before.

I. Introduction

Advances in our understanding of demographic processes have historically been the result of a graceful dance between new theories and new data. In some areas of demographic research, e.g., the study of mortality and fertility, large-scale data collections that include censuses, vital registration systems, and surveys have profoundly enhanced our knowledge of population dynamics. On the other hand, concerning migration studies, lack of data and issues related to cross-country harmonization of existing sources have drastically limited our ability to test theories [1, 2].

Web data have features that are qualitatively different from existing traditional sources and that can be leveraged to evaluate migration theories and their predictive power. In this article, we present a study of migration systems that relies on Google+ data. More specifically, we analyze the extent to which the frequency of people who have lived in three distinct countries is related to bilateral migration flows for pairs of countries. We particularly focus on country triads that occur more or less often than expected given only the data for pairwise flows. The analysis that we present in this article is only possible because our data set of places where Google+ users have lived allows us to evaluate the relative frequencies of triadic groups of countries in which users have lived. This type of information is typically not available in traditional demographic sources which only track movement between pairs of countries.

International migration systems are clusters of countries that are characterized by large exchanges of people and by related feedback mechanisms that connect the countries in terms of flows of goods, capital, information, and ideas. These systems typically persist over time [3]. One mainstream empirical approach for identifying migration systems is to assess changes over time in bilateral flows of migrants for all countries [4, 5]. This approach is problematic partly because reliable data on bilateral flows for a large number of countries, and over time, are not available. In addition, "the trouble with this approach is that the system becomes little more than a summary of flows." [6]

We argue that lack of data constrains the definition of migration systems to a summary of flows. However, with better data, such as self-reported "places lived" that are typically available for a number of social media sources, we can deepen our understanding of migration systems. With the additional knowledge of migration clusters, individual migration corridors are no longer observed independently, yielding a higher level knowledge of migration patterns.

To illustrate that bilateral migration flows (expressed as pairs of countries in which people have lived) are not sufficient to predict more complex migration clusters (triads of countries in which people have lived), Table I provides a simplified example. In the hypothetical situation there are two scenarios, each with four migrants. Both scenarios generate the same distribution of bilateral flows, each occurring exactly once. But they differ in the migration clusters that are observed. Similarly, other scenarios can easily be constructed where

IEEE/ACM ASONAM 2016, August 18-21, 2016, San Francisco, CA, USA 978-1-5090-2846-7/16/$31.00

either all possible clusters or no cluster at all are observed while, again, the distribution of bilateral migration flows is identical.

		Countries Lived In				Bilateral Flows
		A	B	C	D	
Scenario 1	M1	x	x	x		(A,B), (A,C), (B,C)
	M2	x			x	(A,D)
	M3		x		x	(B,D)
	M4			x	x	(C,D)
Scenario 2	M1		x	x	x	(B,C), (B,D), (C,D)
	M2	x	x			(A,B)
	M3	x		x		(A,C)
	M4	x			x	(A,D)

TABLE I: Two toy scenarios for four countries and four migrants illustrating that observing migration corridors is not sufficient to study migration clusters. In both cases, each of the six possible migration corridors is observed exactly once. However, the first scenario features the migration cluster (A,B,C) whereas the second features (B,C,D).

In this paper, we contribute to the literature about migration systems and show how new Web data can be used in the context of classic theories of migration. At the same time, the opportunities opened up by new data and Web science are likely to stimulate the development of new theories that could not be appropriately tested before.

This article is organized as follows. In Section II we provide a review of the relevant literature. Section III describes the data set of Google+ users that we analyzed. Section IV presents our baseline model to estimate triadic groups of countries from bilateral flows. Section V discusses those triads in which the frequency of people who have lived in all three countries is substantially higher or lower than what we would expect based on bilateral flows. The last section summarizes our results and offers some concluding remarks.

II. Related Work

The study of human migration relies on accurate and up-to-date information that is often not available. Traditional demographic sources include censuses, population registers and sample surveys. These data have been extremely useful for improving our understanding of migration processes. However, they are far from perfect. Reliable migration statistics, in particular estimates of flows of migrants, are not directly available for a number of countries. Thus these quantities are often estimated indirectly. For example, Abel and Sanders developed an approach to estimate the minimum sizes of international bilateral flows that are consistent with available estimates of stocks of foreign-born people [7].

The recent availability of geo-located Web data has stimulated the development of new approaches to study international migration. For example, Zagheni and Weber [8] and State *et al.* [9] estimated international migration rates using IP-geolocated data of millions of anonymized Yahoo users' logins. These studies showed that it is feasible to estimate international migration rates, at a large scale, from logins to a website. They also pointed to important challenges related to the fact that the sample is not representative of the underlying population, and offered methodological contributions to deal with selection bias [8, 10].

Zagheni *et al.* [11] and Hawelka *et al.* [12] have used geo-located Twitter tweets data to estimate international migration rates and trends. They showed that estimates of international mobility rates are consistent with statistics about tourism [12]. When no official statistics are available for calibration, an approach based on the 'difference-in-differences' technique proved useful to reduce bias in the data and to estimate trends [11]. Moreover, Twitter geo-located data have a lot of potential for helping us understand the relationship between internal and international migration.

State *et al.* [13] looked into LinkedIn data to investigate trends in international labor migration. They estimated changes in residence, over time, for millions of users, based on their educational and professional histories reported on the LinkedIn website. They found that, conditional on being an international migrant with college education, the probability of choosing the United States as the destination decreased during the period from 2000 to 2012. This is partially related to the rise of migration corridors in East Asia and the dot-com bubble, as well as the great recession in the United States.

Recently, Kikas *et al.* [14] used data from the voice and video call service Skype to study international migration and its relationship to social network features. They found that the percentage of international calls, the percentage of international links and foreign logins in a country, together with information about GDP, could be used to produce relatively accurate proxies of migration rates.

Network theory has been widely used to explain international migration [3]. The main idea is that interpersonal ties that link people in origin and destination countries reduce the costs and risks of migration and increase the expected returns to migration. The network theory of migration is very powerful. However, the lack of comprehensive data about social network connections among countries limit our ability to test and refine theories that explain migrations in terms of networks.

In this paper we contribute to this area by looking at a previously untapped type of data source. We consider the countries people have lived in. This information can only be obtained from data on migration histories, which are typically not available in sample surveys. When some data exist, they are usually collected only for small regions of a country. Data about countries in which people have lived are potentially available for a number of social media services. To our knowledge, nobody has used this type of information to contribute to our understanding of international migration in the context of networks. We thus hope that our paper may stimulate more research in this area.

III. Google+ Dataset

We used the dataset of all Google+ profiles that was collected by Magno et al. [15] between March 23 and June

1, 2012. The data set contains information for 160,304,954 Google+ profiles.

For this article we focus on data about international migration. More specifically, we extract the Google+ field "places lived" ("Places where I lived"). In this field, users list places in the world where they have lived. The items in the list are free text which means that (i) different languages are used ("United States" vs. "Estados Unidos"), (ii) different variations are used within the same language ("São Paulo" vs. "Sampa"), and (iii) locations of different geographic granularities occur ("Brazil" vs. "Minas Gerais" vs. "Belo Horizonte"). Google+ automatically performs geo-coding and maps the free text entries to co-ordinates on Google Maps. For our study, we used these co-ordinates and mapped them to countries. In total, our sample includes 22,578,898 (14%) users with a geo-mapped location.

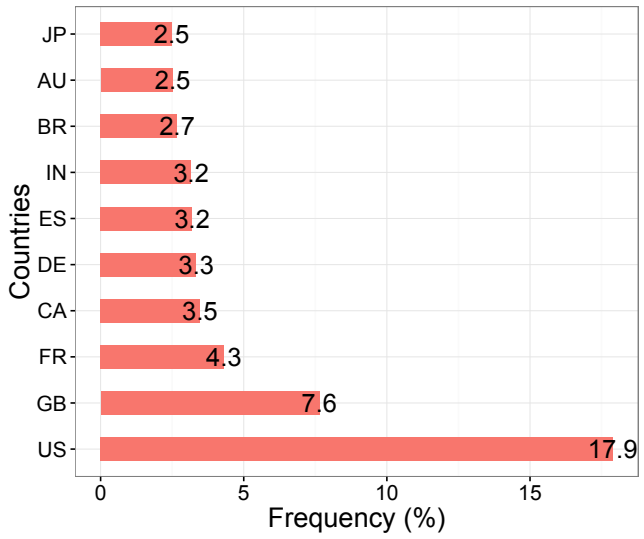

Fig. 1: Fraction of top 10 countries, in terms of number of users, in our data set.

The "places lived", unfortunately, do not come in chronological order, e.g., either the first or the last location might indicate the user's country of origin. It is therefore impossible to tell if a user who lived both in the US and in India moved from India to the US or the other way around. Though this is an obvious limitation, our main analysis is centered around *sets of countries where subsets of users have lived in*. In particular, we look at users who have lived in triples of countries (A,B,C) without distinguishing their order.

As our study is about international migration, we only considered the subset of users who have lived ("places lived") in at least two distinct countries. We refer to this group of users as *migrants*. Our dataset includes 1,958,656 migrants. Users who lived in USA correspond to 17.9% of the data set, while for GB the fraction is 7.6% (see Figure 1). In terms of the number of distinct countries users have lived in, (i) 1,565,803 have two countries in their list, (ii) 271,142 have three, (iii) 69,129 have four countries, and (iv) 52,582 have at least five.

In order to avoid data sparsity issues for countries with very few migrant users, we only considered countries that have at least 1,000 people who have lived there. There are 192 such countries.

For each migrant user, we extracted all the pairs and triples of valid countries they lived in. For example, if a user has lived in countries {BR, FR, HU}, then we would generate the set of country pairs {(BR, FR), (BR, HU), (FR, HU)} as well as the triple (BR, FR, HU). Countries in pairs and triples are listed in alphabetical order to have a canonical form, but no chronological order is implied. For each pair and triple we count how often it occurs among our 1.96M migrant users. In the following, we will also refer to country pairs found in our data as "migration corridors", and to country triples as "migration clusters". Our analysis looks at how the counts for the migration corridors relate to the corresponding clusters. In particular, we are interested in finding and explaining counts for migration clusters which are unusually high or low, given the counts of the contributing migration corridors.

Aiming at allowing reproducibility we release our migration dataset to the research community. The dataset is available at http://www.dcc.ufmg.br/~fabricio/migration-dataset/.

IV. Expected Migration Flows

Our data set enables us to identify clusters of countries that are connected through people who have lived in all of them at some point. We then assess whether the frequency of particular clusters in our data set is higher or lower than what we would expect purely on the basis of frequencies of pairwise connections between countries (number of users who have lived in two countries). For example, if we observe certain migration flows among the pairs of countries (UK, USA), (India, USA), and (India, UK), respectively, intuitively one could expect that the number of Google+ users that lived in the cluster (India, UK, USA) is somehow proportional to these bilateral flows. We want to investigate just how strong this proportionality is and, in particular, which factors are linked to over- or under-proportionate counts of particular migration clusters. In other words, our general goal is to identify and study cases where observed counts of people who have lived in three countries are higher or lower than expected. By 'expected', we mean the counts that one would predict if one only knew data for bilateral migration flows, i.e., pairs of countries in which users lived in.

Here we present our approach to define the expected migration flow of a cluster. For simplicity, we only consider cluster sizes of three countries. However, our methodology easily generalizes to larger cluster sizes, though data sparsity quickly becomes a limiting factor for tuples of more than three countries. We formulate the comparison of "more or less than expected" as a ranking comparison task. Concretely, we rank clusters both (i) according to a function associated to the pairwise counts and (ii) according to their actual frequencies in our Google+ data. The relative difference in the positions between the predicted and observed rankings is then our measure of interest.

Note that the functional dependency between the pair and triple counts is not a priori clear and would depend heavily on assumptions of how migrants move. As we are interested in *discovering* such patterns, we try to avoid overly specific modeling assumptions and, instead, experiment with four different formulas to see which gives the best match between the predicted and observed rankings. All these four formulas merely (i) are symmetric in the three edges, i.e., there is no "first" or "second" edge, and (ii) their predicted frequency of triples increases with increases in the individual pairwise counts.

- Ranking 1 $\sim freqAB + freqAC + freqBC$
- Ranking 2 $\sim freqAB * freqAC * freqBC$
- Ranking 3 $\sim \min(freqAB, freqAC, freqBC)$
- Ranking 4 $\sim \min(freqAB, freqAC, freqBC) * mean(freqAB, freqAC, freqBC)$

where $freqAB$, $freqAC$, $freqBC$ are the frequencies of migrations flows among the three pairs of countries of a cluster (A, B, C).

Intuitively, as the observed summed counts of the pairs in a triangle increase, the corresponding observed triple counts should also increase. This is why we included (freqAB + freqAC + freqBC) in our baseline 'Ranking 1'. The model 'Ranking 2' is inspired by approaches to the study of migration flows known as gravity models [16]. These models explain flows between origin and destination countries as proportional to the product of their sizes and inversely proportional to their distances. Here we consider that the effect of distance on triples of countries where users lived in is implicitly accounted for by the number of users who have lived in the respective pairs of countries. 'Ranking 2' is appealing because it is intimately connected to a class of models, gravity models, that have been used quite successfully by migration scholars. For our specific situation, however, it is also clear that the *minimum* value of the three pairwise counts plays an important role as, trivially, the triple count is upper bounded by the minimum of the three pairwise counts. In other words, when we consider a system of three countries, the maximum number of people who have lived in all three countries cannot be larger than the minimum value of the number of people who have lived in only two of the three countries. To take this dependency into account, we also included $\min(freqAB, freqAC, freqBC)$ in our baseline 'Ranking 3'. The model 'Ranking 4' is a further extension that adds to 'Ranking 3' by including the average size of the pairwise frequencies. The intuition is that the larger the migration system, the higher the probability that people who have lived in two countries might have been attracted to a third country as well.

In order to measure the extent to which these rankings produce accurate results, we compare them with the ground truth data from Google+. Table II shows the correlation of these rankings with the ground truth ranking according to two well-known rank correlation measures: Kendall and Spearman rank correlation coefficients [17]. We can see that Ranking 4 yields the best prediction of the actually observed Google+ cluster ranking, using only information from pairs of countries. In the rest of the paper we refer to this ranking as the *expected ranking*.

TABLE II: Performance of ranking formulas

Description	Kendall	Spearman
Ranking 1	0.350	0.498
Ranking 2	0.546	0.737
Ranking 3	0.502	0.689
Ranking 4	**0.565**	**0.754**

The creation of an expected ranking from pairs of countries enables us to gain some insights about how countries are integrated in terms of people who have lived in all of them. For example, in our data set, 1,077 people have lived in Great Britain (GB), Malaysia (MY), and Singapore (SG). This number, freq(GB,MY,SG), is substantially larger than what we would expect from the counts of users who have lived in two of these countries: freq(GB,MY)=5,552; freq(GB,SG)=6,642; freq(MY,SG)=7,242. This means that within this group of countries, users who have lived in two of them have a relatively high probability to have lived in the third country. In this situation, the observed value for the cluster is *higher than expected*. Conversely, when we consider the cluster formed by Great Britain (GB), the Philippines (PH), and the United States (US), we observe that a similar number of users (1,022) have lived in all the three countries. However the pairwise frequencies are substantially higher: freq(GB,PH)=3,179; freq(GB,US)=152,976; freq(PH,US)=24,599. In this case a large number of users have lived either in the Great Britain and the US, or in the Philippines and the US. However, only a small proportion of these users have lived in all the countries. The observed number of users who have lived in the three countries is lower than what we expected based on pairwise frequencies. We refer to this situation as *lower than expected*.

In the next section we formulate a classification problem where we investigate the discriminative power of additional features, such as a shared language, colonial link, distance, to differentiate clusters.

V. Explaining Deviance from Expectation

Our next step is about identifying a set of features related to migration clusters. The aim is to investigate their relative discriminatory power to distinguish clusters that are ranked higher than, lower than, or as expected. First, we present a definition for three classes.

A. Classes of Clusters

We rank the triples by how much their actual frequency ranking differs from the expected one. We then divide this ranking into five strata, each containing 20% of the data.

Based on this division, we consider the following three cluster classes.

- **As expected**: We consider as expected or close-to-expected the center 20% of the clusters with the expected and actual ranks approximately equal.
- **Higher than expected**: We consider as higher-than-expected those clusters that appear in the top 20% on the positive side.
- **Lower than expected**: We consider as lower-than-expected those clusters that appear in the top 20% on the negative side.

Thus, our approach neglects 40% of the data, which corresponds to the folds that appear in between these three cluster classes we considered. For the observations that we do not consider, there is much more uncertainty associated to potential differences in ranking.

B. Features

Migration patterns depend on a multitude of factors. The goal of our analysis is to understand which type of features (derived from the triads), e.g., geographical or historical, either lead to or inhibit the formation of migration clusters. This type of analysis is impossible with traditional data sources which only record pairwise migrations independently.

- **Common Civilization**: A recent study [18] has found empirical evidences, from online data, that eight culturally differentiated civilizations can be identified, as theoretically posited by Huntington [19], with the divisions corresponding to differences in language, religion, economic development, and spatial distance. We operationalized it as a single numeric score, with values 0, 2, or 3, that represent the number of countries (None, 2 out of 3, and All) in the triad of countries with common civilization. The same approach of assigning a single integer to a triple was used for Common Colonial Link, Common Language, and Visa Requirement.
- **Geographic Distance**: The distance among countries represents a physical barrier for migration. For each cluster we consider as features the average distance among the pairs, as well as the maximum and minimal distances between the pairs of countries within the cluster. The distances were obtained from the geolocation[1] (latitude, longitude) of the center of the mass of each country. Thus, the distance between countries is calculated by the spherical distance, considering the earth curvature. Another geographic related feature is the common region, which represents the main continental regions in which countries are grouped.
- **GDP**: The gross domestic product (GDP) is one of the primary indicators used to gauge the size of a country's economy. It represents the total dollar value of all goods and services produced over a specific time period. For each cluster we consider as features the average GDP among the pairs, as well as the maximum and minimum GDP between a pair of countries within the cluster.
- **Common Colonial Link**: This feature aims at capturing if two countries share a colonial past.
- **Common Language**: This feature aims at assessing if two countries share the same language.
- **Visa Requirement**: Visa requirement may represent another barrier for migration.

[1]http://opengeocode.org/download/cow.txt

Figure 2 and Figure 3 show the cumulative distribution function for features *minimum distance* and *maximum GDP* for the three cluster classes, respectively. We can note that 75% of the pairs of countries within the cluster higher-than-expected are within 2,000 Km in distance, whereas only around 27% of the pair of countries within the cluster lower-than-expect are within this same distance. Similarly, we can note that 50% of the pairs of countries within the cluster close-to-expected have GDP lower than 88 (hundreds of billions of USD), a higher value in comparison with the other cluster classes (49% for higher-to-expected and 82% for lower-than-expected).

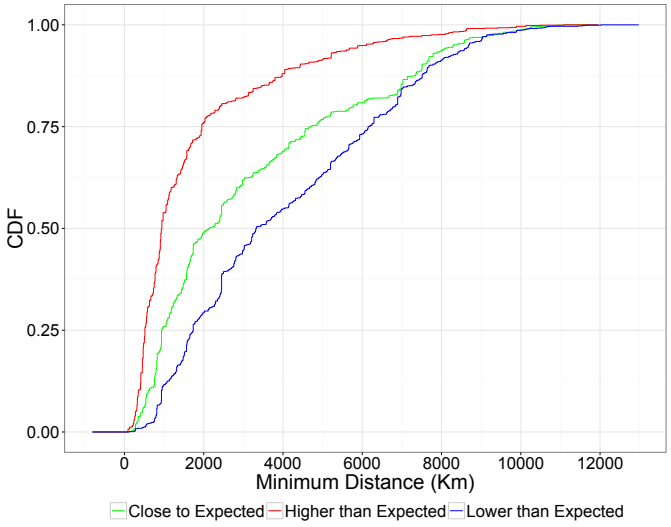

Fig. 2: Cumulative distribution function (CDF) for the feature minimum distance for the three cluster classes

Figure 4 shows the difference between the ground truth and the expected ranking considering four features that account for common factors among countries. Particularly, we show the amount of countries (out of 3, because of the triad) within each cluster class with common civilization, common language, common colonial link, and common region. We can see interesting trends here. For example, we can note triads in the cluster of higher than expected tend to have more countries with common civilization than the rest. We can also note a similar trend for common region and common language. On the other hand, colonial link shows a very similar distribution for all three classes. In the next section we provide a rank for these features in terms of their discriminative power to distinguish among classes.

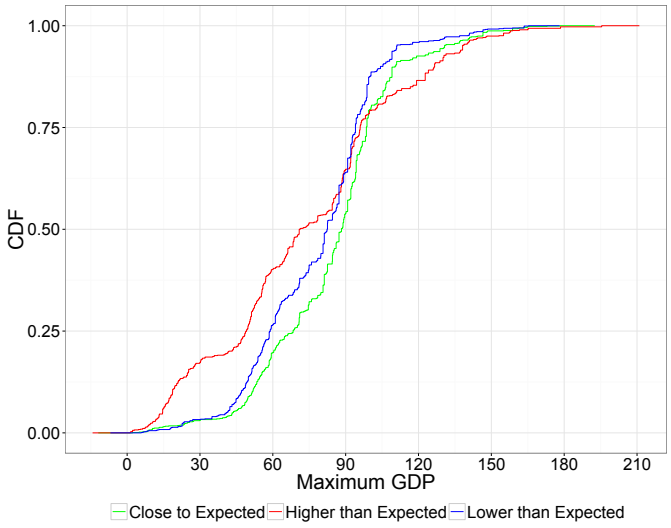

Fig. 3: Cumulative distribution function (CDF) for the feature maximum GDP for the three cluster classes. GDP values are expressed in hundreds of billions of USD

C. Assessing Feature Importance

We assessed the relative power of the features considered in discriminating one cluster class from the others by independently applying two well-known feature selection methods, namely, information gain and χ^2 (Chi Squared) [20]. Table III shows the ranking of the most important features for differentiating the three classes (higher-than-expected, close-to-expected, lower-than-expected). We note that the four geographic distance features appear on the top of the table, followed by all the features related to GDP.

Though the observation that geographic vicinity leads to migration clusters seems obvious, it is worth pointing out that it is not. As the geographic vicinity already increases the pairwise migration counts, it is implicitly already accounted for in the expected ranking of migration clusters. So what is observed here is a "supra-linear" type of effect that is not predicted by the pairs alone.

TABLE III: Ranking of most important features for differentiating the three classes (higher-than-expected, close-to-expected, and lower-than-expected), presented by the IG (Information Gain) Ranking and the χ^2 (Chi-Squared) Ranking.

Description	IG Rank	IG Value	χ^2 Rank	χ^2 Value
Min Distance	1	0.231	1	984.742
Max Distance	2	0.180	3	767.547
Common Region	3	0.178	2	780.458
Avg Distance	4	0.173	4	745.858
Max GDP	5	0.102	5	474.392
Avg GDP	6	0.089	6	408.225
Min GDP	7	0.070	7	312.460
Common Civ.	8	0.033	8	147.838
Common Visa	9	0.017	9	80.004
Com. Col. Link	10	0.0001	10	0.679

D. Illustrative Cases

In the previous section we attempted to summarize, in a quantitative way, the key features that discriminate various classes of countries according to our definition. Here we discuss some examples that offer a more qualitative understanding of what we observed in the data. More specifically, we present a couple of cases in which the observed number of people who have lived in all three countries is higher than what we would have expected based on pairs of flows. We will then discuss a couple of cases for which the opposite is true.

Consider the United Arab Emirates, India and Singapore. In our dataset, 805 users have lived in all the three countries. 17,584 users have lived in the United Arab Emirates and India. 7,665 users have lived in India and Singapore. A lower number of users, 1,970, have lived in the United Arab Emirates and Singapore. Based on pairs of flows, we would expect that a relatively low number of users have lived in all three countries. In fact our original ranking model 4 would rank this triple at place 682. However, in our Google+ dataset the actual ranking is number 200. About 40% of the users who have lived in Singapore and in the United Arab Emirates have also lived in India. This indicates that in addition to the large communities of Indians in Singapore and in the United Arab Emirates, there is also a sizable unexpected community of users who have been in all the three countries and who strengthen interpersonal networks across these countries.

Similarly, when we consider the cluster Spain, France, and Italy, we would expect to observe less people who have been in all three countries than what we actually find in the data. 2,322 users have lived in all the three countries; 15,455 have lived in Spain and France; 11,230 have lived in France and Italy; 9,628 have lived in Spain and Italy. Based on the flows for pairs of countries, our ranking model would have expected the triple to rank number 111, when in fact it ranked number 36 in our data set. This example might be related to the context of European integration that lowers the cost of moving to countries within the Union. Moreover, these countries are close in terms of distance, with languages that are relatively similar. In addition, interpersonal networks may be strong enough to make the cost of moving across these countries relatively low. Overall, we observe that a substantial fraction (more than expected) of the people who have lived in two of these countries, have also lived in the third one.

The situation is quite different for the cluster composed of Brazil, Mexico, and the US. In our Google+ dataset, 14,593 users have lived in Brazil and Mexico; 46,784 users have lived in Brazil and the US; 67,065 users have lived in Mexico and the US. Although these pairs of flows are quite substantial, only 1,386 users have reported living in all the three countries. Brazil, Mexico, and the US have strong bilateral connections, but they do not seem to be integrated within a larger cluster in a demographic sense, meaning that people typically migrate only along one of the corridors. Our ranking model would have expected this triple to rank number 12 based on bilateral flows. Instead it ranked number 80 in the actual Google+ data.

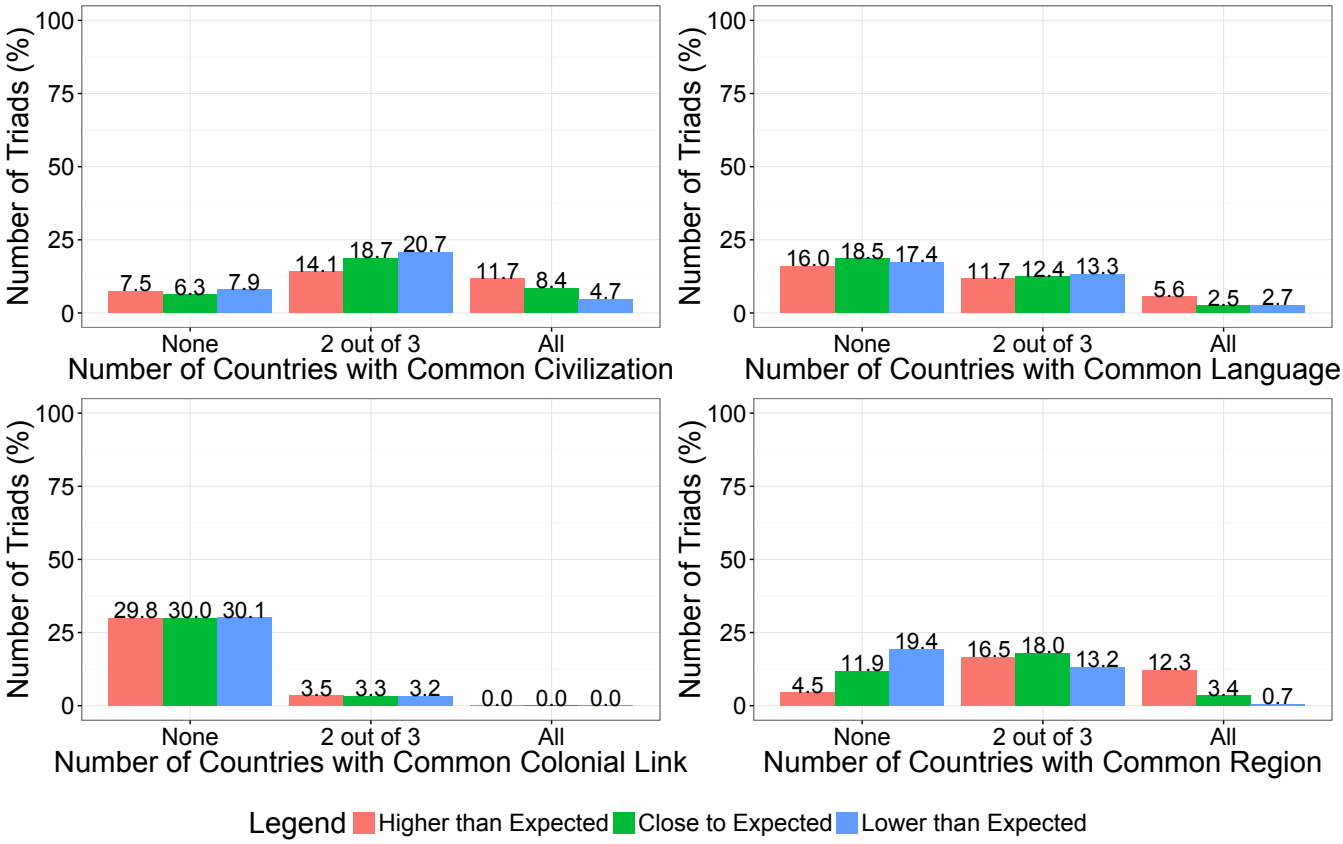

Fig. 4: Distribution of the difference between the ground truth and the expected ranking considering four features that account common factors among countries

Canada, China, and Great Britain offer a similar example of a weaker-than-expected cluster. 6,093 users have lived in Canada and China; 25,696 users have lived in Canada and Great Britain; 8,189 users have lived in China and Great Britain. However, only 623 users have lived in all the three countries. As for the previous example, migration does occur along the corridors but rarely within the whole cluster. For example, a number of Chinese students might go to study to Canada or Great Britain. However, only a relatively small fraction would experience living in both Canada and Great Britain. This example is important because it also highlights one of the limitations of our approach: Google+ is not accessible in China. Thus the values that we observe for this cluster might be skewed, particularly towards Chinese living abroad, or non-Chinese people who have lived in China at some point.

VI. Conclusions and Discussion

We started this paper by saying that new theories and new data move hand-in-hand to advance our understanding of demographic processes. In this article, we showed that new data about 'places lived' can lead to the development of new theories of international migration. We started with the observation that data about 'places lived' for more than two countries (migration histories) are traditionally not available, except for some special subregions within a particular country. This type of information is not equivalent to data about bilateral flows, and is very valuable to identify specific characteristics of high level migration systems. In particular, studies on what leads users to migrate within clusters of countries cannot be performed with data limited to pairwise migration flows.

We believe that this line of research is relevant and timely, and that the increasing availability of information about pseudo-migration histories from online sources opens new and exciting opportunities at the intersection of social network analysis and demography. Here we would like to discuss some of the limitations of our current research and point to some directions for future work.

For this study, we work with a sample of Google+ users that is quite large and that can be collected at low cost. However, Google+ data have several shortcomings. First, as mentioned earlier, we do not know the chronological *order* in which people have lived in the various countries that they list. For our specific application, this is not a problem since we are interested in how people connect countries by living in several of them. However, more elaborate analyses could be performed if we could identify each user's home country and the countries of residence in a chronological order. This

type of information has been used to evaluate bilateral flows of professional migrants on LinkedIn [13]. The same type of dataset could be used to evaluate clusters of countries in terms of professional skills and the direction of flows within a cluster (for example, are people more likely to move from country A to country C via an intermediate step in country B?).

Second, the Google+ dataset that we are using is neither representative of the world population nor of any specific country. Several different types of selection bias mechanisms affect our data. Users in our dataset are, first of all, Internet users. They are more likely to be more highly educated and younger than the average population, especially in the context of developing countries with low Internet penetration rates. As a result our users are most likely more internationally minded and mobile than in the underlying populations. In fact, 9%, 1.96M out of 22.6M users with at least one geo-coded location, are migrants in our dataset. This is substantially higher than the United Nations estimate of the percentage of people who live in a country different from their country of birth, which is between 3% and 4%. In addition, most of the Google+ users are located in North America or in Western Europe. The extent of bias differs from country to country. China is an extreme case, since the country is blocking access to Google and other popular social media services [21]. In our study we did not attempt to calibrate our results in order to remove the bias, as discussed in other venues [10]. Instead, we attempted to control for a number of biases by evaluating the number of people who have lived in three countries conditional on having information about bilateral flows. For example, since Google+ is quite popular in the US, we would expect more people in our data set to have lived in the US and in a second country. Conditional on having lived in these two countries, we considered the fraction of users who have lived in a third one and compared it with the expected value based on the size of bilateral flows. This is an imperfect correction that was appropriate for our specific application, but not necessarily generalizable to other situations. More research to address issues related to selection bias in social media data is certainly needed.

Third, there is a range of data quality issues. These include the free text nature of the "places lived" field, which could lead to ambiguities. In addition, we need to be aware of potential misreporting or intentionally fabricated histories.

In the end, no single dataset is enough to study international migration. In the future, we hope to be able to combine several data sources that include both Web data and traditional demographic sources. We hope that this paper contributes to highlight the potential and weaknesses of Web data for the study of migration processes and that it would stimulate collaborations between researchers in the area of demography and Web science.

Aiming at allowing reproducibility we release our migration dataset to the research community. The dataset is available at http://www.dcc.ufmg.br/~fabricio/migration-dataset/.

ACKNOWLEDGMENT

This work was partially supported by the project FAPEMIG-PRONEX-MASWeb, Models, Algorithms and Systems for the Web, process number APQ-01400-14, and by individual grants from CNPq, CAPES, and Fapemig. We also would like to thank Gabriel Magno for sharing his data collection.

REFERENCES

[1] J. De Beer, J. 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," *European Journal of Population/Revue européenne de Démographie*, vol. 26, no. 4, pp. 459–481, 2010.

[2] F. Laczko, "Factoring migration into the 'development data revolution'," *Journal of International Affairs*, vol. 68, no. 2, 2015.

[3] D. S. Massey, J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino, and J. E. Taylor, "Theories of international migration: A review and appraisal," *Population and Development Review*, vol. 19, no. 3, pp. 431–466, Setembro 1993.

[4] H. Zlotnik, *Empirical identification of international migration systems*. Clarendon Press, Oxford, 1992, pp. 19–40.

[5] J. DeWaard, K. Kim, and J. Raymer, "Migration systems in europe: Evidence from harmonized flow data," *Demography*, vol. 49, no. 4, pp. 1307–1333, 2012.

[6] O. Bakewell, "Relaunching migration systems," *Migration Studies*, p. mnt023, 2013.

[7] G. J. Abel and N. Sander, "Quantifying global international migration flows," *Science*, vol. 343, no. 6178, pp. 1520–1522, 2014.

[8] E. Zagheni and I. Weber, "You are where you e-mail: Using e-mail data to estimate international migration rates," in *WebSci*, 2012, pp. 348–351.

[9] B. State, I. Weber, and E. Zagheni, "Studying inter-national mobility through ip geolocation," in *WSDM*, 2013, pp. 265–274.

[10] E. Zagheni and I. Weber, "Demographic research with non-representative internet data," *International Journal of Manpower*, vol. 36, no. 1, pp. 13–25, 2015.

[11] E. Zagheni, V. R. K. Garimella, I. Weber, and B. State, "Inferring international and internal migration patterns from twitter data," in *WWW*, 2014, pp. 439–444.

[12] B. Hawelka, I. Sitko, E. Beinat, S. Sobolevsky, P. Kazakopoulos, and C. Ratti, "Geo-located twitter as proxy for global mobility patterns," *Cartography and Geographic Information Science*, vol. 41, no. 3, pp. 260–271, 2014.

[13] B. State, M. Rodriguez, D. Helbing, and E. Zagheni, "Migration of professionals to the U.S. - evidence from linkedin data," in *SocInfo*, 2014, pp. 531–543.

[14] R. Kikas, M. Dumas, and A. Saabas, "Explaining international migration in the skype network: The role of social network features," in *1st ACM Workshop on Social Media World Sensors*, 2015, pp. 17–22.

[15] G. Magno and I. Weber, "International gender differences and gaps in online social networks," in *Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11-13, 2014. Proceedings*, 2014, pp. 121–138.

[16] J. E. Cohen, M. Roig, D. C. Reuman, and C. GoGwilt, "International migration beyond gravity: A statistical model for use in population projections," *Proceedings of the National Academy of Sciences*, vol. 105, no. 40, pp. 15 269–15 274, 2008.

[17] H. Abdi, "The kendall rank correlation coefficient," *Encyclopedia of Measurement and Statistics. Sage, Thousand Oaks, CA*, pp. 508–510, 2007.

[18] B. State, P. Park, I. Weber, and M. Macy, "The mesh of civilizations in the global network of digital communication," *PLoS ONE*, vol. 10, no. 5, 2015.

[19] S. P. Huntington, *The clash of civilizations and the remaking of world order*. Penguin Books India, 1997.

[20] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *ICML*, 1997.

[21] D. Bamman and N. A. Smith, "Censorship and deletion practices in chinese social media," *First Monday*, 2012.