Examining LinkedIn's Talent Migration Dataset

Gia Ahn, Michael P. Ramirez, Nishant Yadav, Venk Muriki.

Abstract

In a collaborative effort, the World Bank Group (WWG) and LinkedIn created the Digital Data for Development (DDD). The main objective is to provide new data and insight to guide policymakers to assist underdeveloped countries. The raw LinkedIn membership data is filtered and transformed. Lastly, the tidy LinkedIn dataset metric is compared against the WWG government-sourced data.

The migration trends of skilled workers and industry directly impact both the host country and the country of origin. In the past, we could only rely on government-sourced data that provided limited insight. However, LinkedIn provides web-based data that is not captured by government entities. The merging and comparing of both metrics can provide better insight.

We explored nation migration based on income and set out to find the difference between wealth in terms of socioeconomic index that is part of the data. We will also be performing network analysis, especially on the Country Migration data and we will be conducting hoverable and interactive visualizations of the migration actively presented with the graphs below. We will also be looking at the degree of centrality its effect on the exchange of personnel between one country and the other.

After contacting the owners of our dataset, we did not receive a reply. However, we decided to proceed with data visualization, and exploratory analysis to establish findings regarding the type of transfer between the two countries. Overall, we are trying to show the relationship between third world countries who are receiving help from the countries with skill resources.

Background

In an international marketplace, we need better metrics to help aid policymakers, and economists make better-informed decisions.

Our project focuses on studying skilled talent migrations between 2015 and 2019 using data collected in collaboration with LinkedIn and the World Bank Group. The dataset includes industry, skill, and country migration metrics for almost 140 countries, keeping track of 4 major metrics: Industry Employment Shifts, Talent Migration, Industry Skills.

While the data is very detailed and kept up to date, we are particularly interested in the Country Migration data. From this data, we can observe changes and compare labor markets to see which countries are performing well and which countries might need an extra push. Moreover, we can monitor how different social and economic conditions can cause people to immigrate to different countries in hopes of setting their families up for a better life. We hopefully gain a better perception of the “invisible hand of the market.”

Data

The migration trends of skilled workers and industry directly impact both the host country and the country of origin. In the past, we could only rely on government-sourced data that provided limited insight. However, LinkedIn provides web-based data that is not captured by government entities. The merging and comparing of both metrics can provide better insight. The dataset is separated into 3 different datasets: country migration, industry migration, and skill migration.

In all three datasets, the constants variable were a country, world bank region, and the world bank income level. The country migration is gathered from LinkendId's member's location. The industry's and skill observations were obtained from LinkedIn members' C.V and current job titles. Later, the datasets were merged with the World Bank's existing government data.

Data semantics and structure

Units and observations: There are several variables we are interested in. We are primarily interested in the net_per_10K_yyyy variables, where yyyy denotes the year for which it is collected. There are five years, 2015 through 2019. The net_per_10K_yyyy variables are the number of arrivals of LinkedIn members from a base country to a target country, less then number of departures from the target country to the base country, over the number of LinkedIn members of the base country. For example, if in 2015 there were 100 LinkedIn members in the country of Rohan (base country), and 25 of them accepted jobs in the country of Mordor (target country), and simultaneously 10 LinkedIn members left Mordor after accepting jobs in the country of Rohan, then the net_per_10K_2015 for the country of Rohan is (-25+10)/100 = (-15/100) = -0.15. The net_per_10K_yyyy variables have been scaled so that the ratio is describing per 10K LinkedIn members.

Variable descriptions:

Name	Variable description	Type	Units of measurement
base_country_name	Country of Origin	string	Name of Country
base_lat	Base Country Latitude Coordinates	numeric	Latitudinal Coordinates
base_long	Base Country Longitude Coordinates	numeric	Longitudinal Coordinates
base_country_wb_income	World Bank Income Classification of Base Country	string	(High, Upper Middle, Lower Middle, Low)
base_country_wb_region	Region of Origin Country	string	Continental Region
target_country_name	Country of Arrival	string	Name of Country
target_lat	Target Country Latitude Coordinates	numeric	Latitudinal Coordinates
target_long	Target Country Longitude Coordinates	numeric	Longitudinal Coordinates
target_country_wb_income	World Bank Income Classification of Target Country	string	(High, Upper Middle, Lower Middle, Low)
target_country_wb_region	Region of Target Country	string	Continental Region
net_per_10K_2015	net flow of migration for target country in 2015	numeric	ratio of gain or loss per 10,000 people.
net_per_10K_2016	net flow of migration for target country in 2016	numeric	ratio of gain or loss per 10,000 people.
net_per_10K_2017	net flow of migration for target country in 2017	numeric	ratio of gain or loss per 10,000 people.
net_per_10K_2018	net flow of migration for target country in 2018	numeric	ratio of gain or loss per 10,000 people.
net_per_10K_2019	net flow of migration for target country in 2019	numeric	ratio of gain or loss per 10,000 people.

A view of the tidied data:

Out[4]:

	base_country_name	base_lat	base_long	base_country_wb_income	base_country_wb_region	target_country_name	target_lat	target_
0	United Arab Emirates	23.424076	53.847818	High Income	Middle East & North Africa	Afghanistan	33.939110	67.70
1	United Arab Emirates	23.424076	53.847818	High Income	Middle East & North Africa	Algeria	28.033886	1.65
2	United Arab Emirates	23.424076	53.847818	High Income	Middle East & North Africa	Angola	-11.202692	17.87
3	United Arab Emirates	23.424076	53.847818	High Income	Middle East & North Africa	Argentina	-38.416097	-63.61
4	United Arab Emirates	23.424076	53.847818	High Income	Middle East & North Africa	Armenia	40.069099	45.03

In our exploratary data analysis, we discovered there were 140 different countries that appeared on the LinkedIn country-migration dataset, from seven unique regions: South Asia, Middle East & North Africa, Sub-Saharan Africa, Latin America & Caribbean, Europe & Central Asia, East Asia & Pacific, North America. No missing variables appeared in the original dataset.

Aims

We plan to explore some other aspects of the networked data, and potentially combine the other datasets from the LinkedIn talent migration data to further explore the country-to-country migration.

In our initial exploration, we found that there were three datasets: country-migration, industry-migration, and skill-migration. The country migration dataset was the only dataset that specified the target_country of the base_country to the final country relationship. Although the documentation specifies that in the formula for industry and skill migration variables a target_country should be available, there is no correlation. We needed to find a way to combine our uneven dataset based on region, and income level.

The three datasets are also of varying rows, so there is no direct connection that is visible. Since we don't have the target_country specified for the industry-migration and the skill-migration dataset, we could not even conduct a left_join because we needed to join on the base_country and target_country for both of the datasets, so we will mainly work with the country-migration dataset and perform network analysis, as well as try to find some new patterns and explore new ideas starting from the visualization techniques we learned in the class so far. We have already contacted the people responsible for the dataset, but they failed to respond.

Firstly, we wanted to account, for the World Bank's ranking system. There is an uneven distribution of countries assigned to the region, and income level. Hence, the skewed distribution needed to be taken into account.We conducted a network analysis of the flow of talent. Although this visualization is not easily quantifiable and prone to error, it gives us a good visual on understanding the main players of exchanging talent at the very least. The net flows can also be presented in the form of a choropleth map, from country to country, which can also identify the dominant countries on a visual basis. We can also rank the top 10 countries, faceted by year, to see who donated the most talent and to see who received the most talent.

We aggregated over the years and try to run a certain function over the ratios. We can also try to observe the ratio stagnancy over the five years to see which countries maintained their stagnant talent exchange. Since income level is the most interesting covariate in the country-migration dataset, we can try to produce more visual plots that quantify the income level distribution across countries and over the five years. We grouped the countries into different regions since there are so many and tried to observe regional differences to see whether geographical effects are inherently driving factors. The income level distribution can also tell us a lot about the wealth of the populations, which translates to the health of the international trade market and economy. We can also compare the frequency of change over the five years to specifically observe to what degree the international market is dynamic.

Methods and Results

At first, we inspected our datasets and filtered any missing, or duplicated data. We created subsets based on the World Bank's ranking system and merged our datasets by countries, the countries' region, and countries' income levels. We calculated the proportion of a country's connection to other countries based on income level, and we were able to see interesting distribution when we group by specific variables. Afterward, we computed the net flows by aggregating the past five years to see the overall trend.

We are going to now analyze the unique origin countries. We want to analyze the country migration data to see the various exhchanges that will be taking place. We hope to generate information regarding the base countries so that we can understand the base relationship that the target countries will be held accountable for. What we are analyzing is the various ways that a certain home country of a certain population can help with the personnel and industry business growth that other countries wish to recieve similar aid with.

For example, let's consider the country of Mongolia. Our dataset shows that Mongolia networks with four countries. Network analysis can help us understand the connections, especially with a larger dataset with a number of countries with larger-scale attributes such as skill and migration. We are going to be analyzing the income, target countries, and the covariates, which pertain to the understanding of the country in question. Below we show the number of countries, which Mongolia networks with.

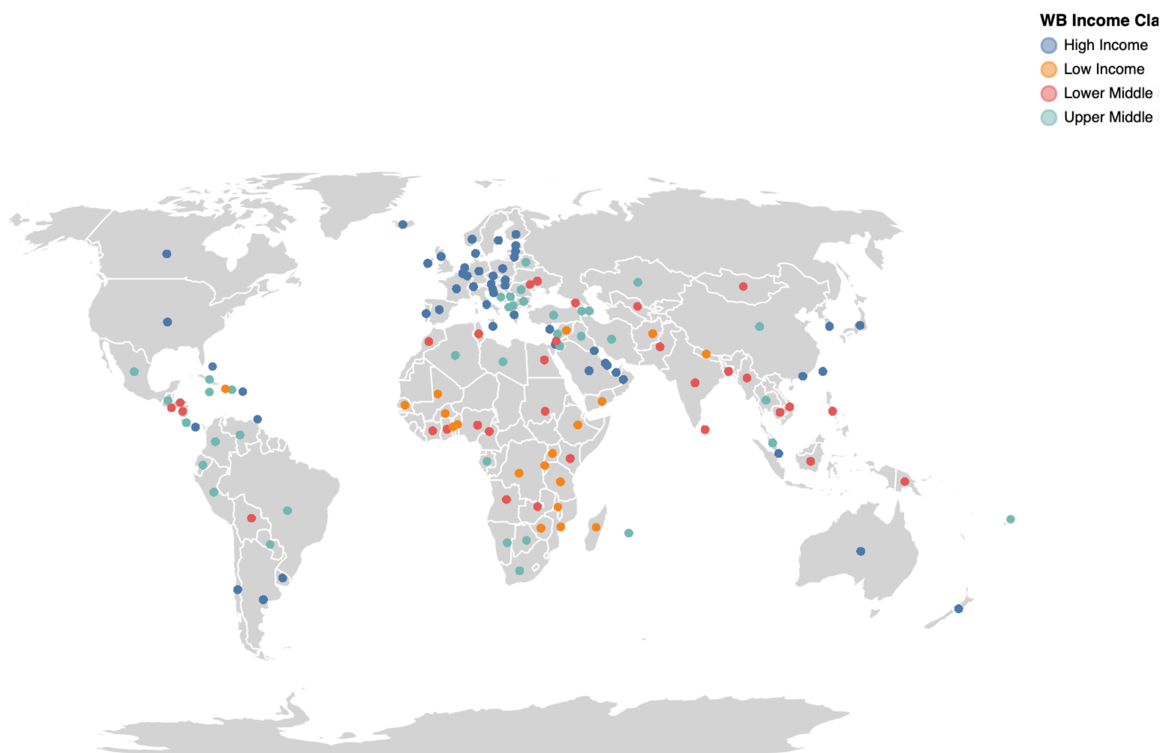
Out[7]:

	base_country_name	base_lat	base_long	base_country_wb_income	base_country_wb_region	target_country_name	target_lat	ta
2584	Mongolia	46.862496	103.846656	Lower Middle Income	East Asia & Pacific	Australia	-25.274398	13
2585	Mongolia	46.862496	103.846656	Lower Middle Income	East Asia & Pacific	China	35.861660	10
2586	Mongolia	46.862496	103.846656	Lower Middle Income	East Asia & Pacific	United Kingdom	55.378051	-
2587	Mongolia	46.862496	103.846656	Lower Middle Income	East Asia & Pacific	United States	37.090240	-9

Exploring the difference in number of net-flow exchanges by country

This visualization shows the number of exchanges one country has with another. Use the mouse to hover over a country to see its net-flow exchanges with another. An interesting thing to note is that if a base country has a net flow to a target country, then that target country has a net flow value to that base country. There is no instance where say Country Narnia sends people to Country Mordor and doesn't receive people back. This indicates that if there is a line from one country to another, they have a relationship.

Out[8]:



Above is a dynamic chart of the country-country migration. It shows the different networks that each country has. Above, we have separated the "lines", which correspond to each network. The networking aspect can result an interactive presentation of the country-country exchange. This is useful in terms of getting the number of exchanges (relationships) for each country. This analysis revolves around the degree centrality regarding the country exchange in question. In python, we are going to generate a list of country exchanges, which pertains to the countries in question.

Examining changes in net-flows for each country over time

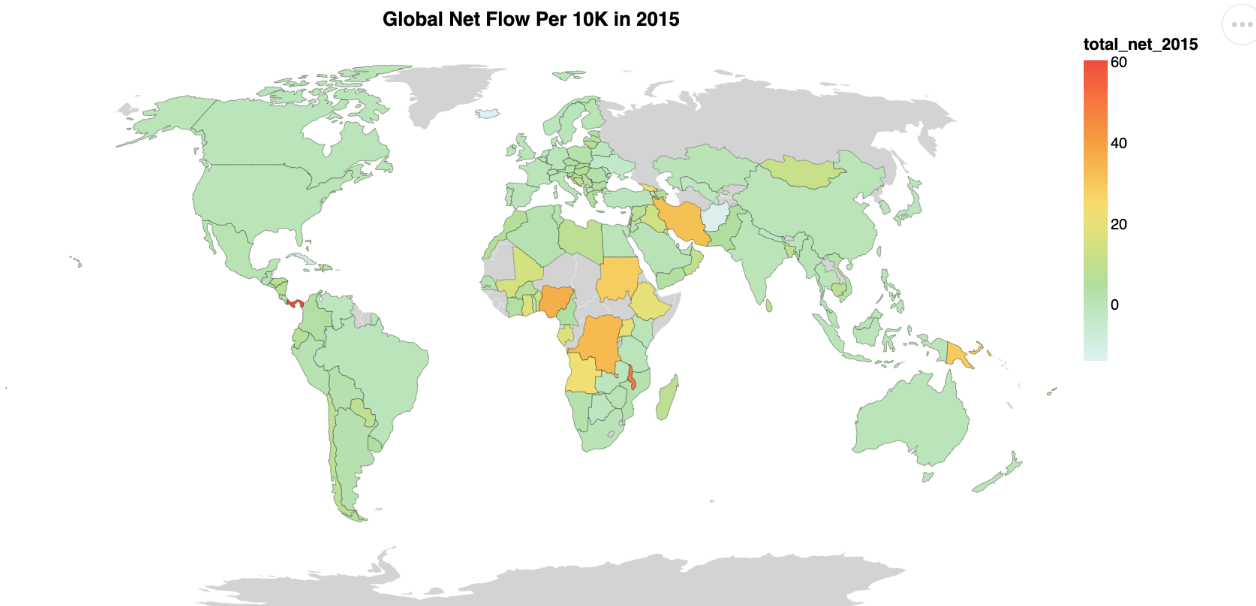
We calculated the minimum, maximum, mean, and standard deviation of net flows for each year:

Net Flow Metrics:

Year	Min	Max	Mean	StdDev
2015	-37.01	150.68	0.4617574734811958	5.006529628076893
2016	-40.89	124.48	0.15024831243973	4.2011179412928366
2017	-43.66	87.0	-0.08027242044358726	3.2030922857484443
2018	-56.22	91.41	-0.04059064609450337	3.5938763075730384
2019	-50.33	87.71	-0.02274349083895855	3.633246602241162

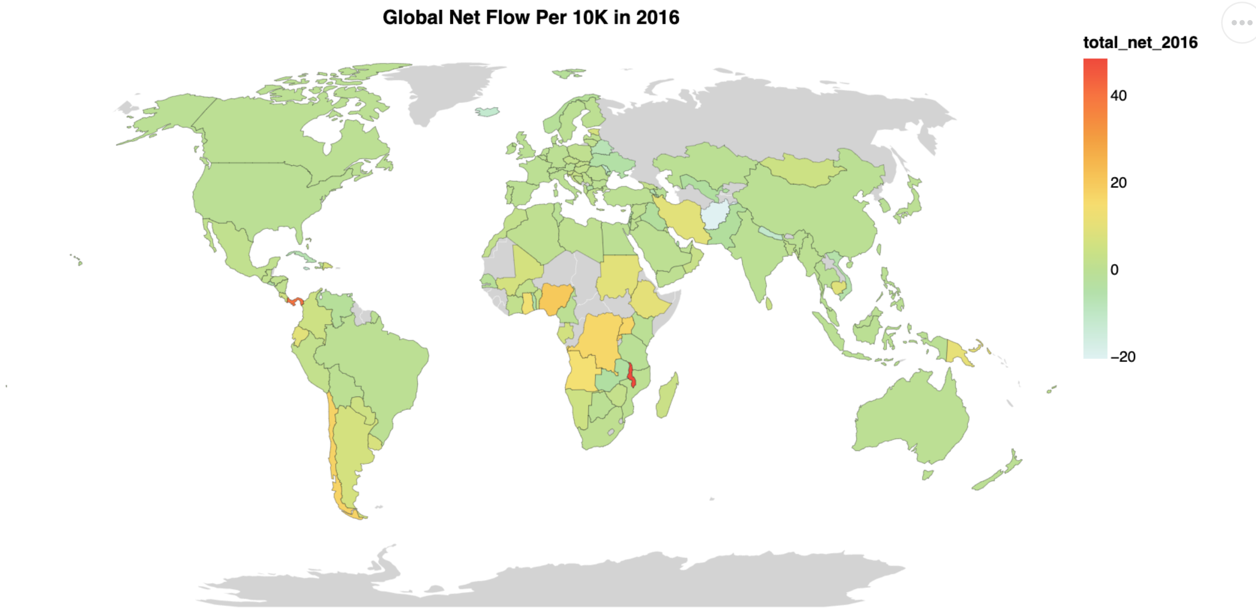
We next explored the total net-flow a country had for each year between 2015 and 2019 to analyze whether any significant event occured over time.

Out[11]:



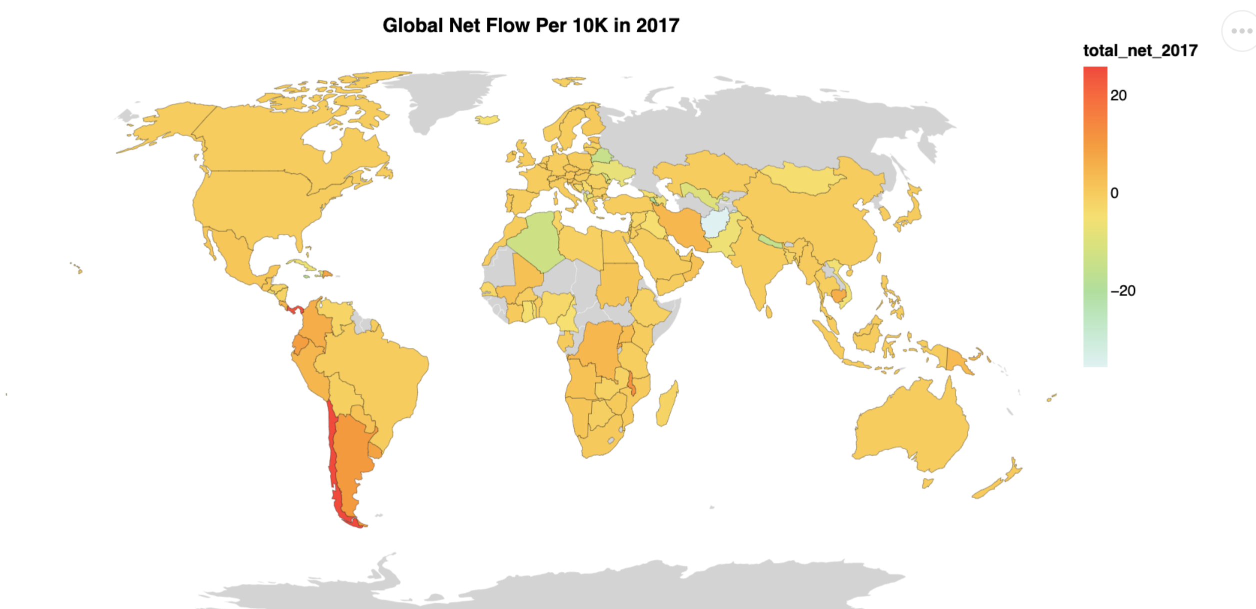
For 2015, the net flow, per the documentation, is per 10,000 people. It is interesting to notice that there is barely any red, except near the strip connecting North and South America and near Madagascar/South Africa. But mostly, we can see that there is high net flow in Africa and some in the middle east and the islands above Australia. But it is intersting to see that North America and most of South America along with some of middle asia and Australia.

Out[12]:



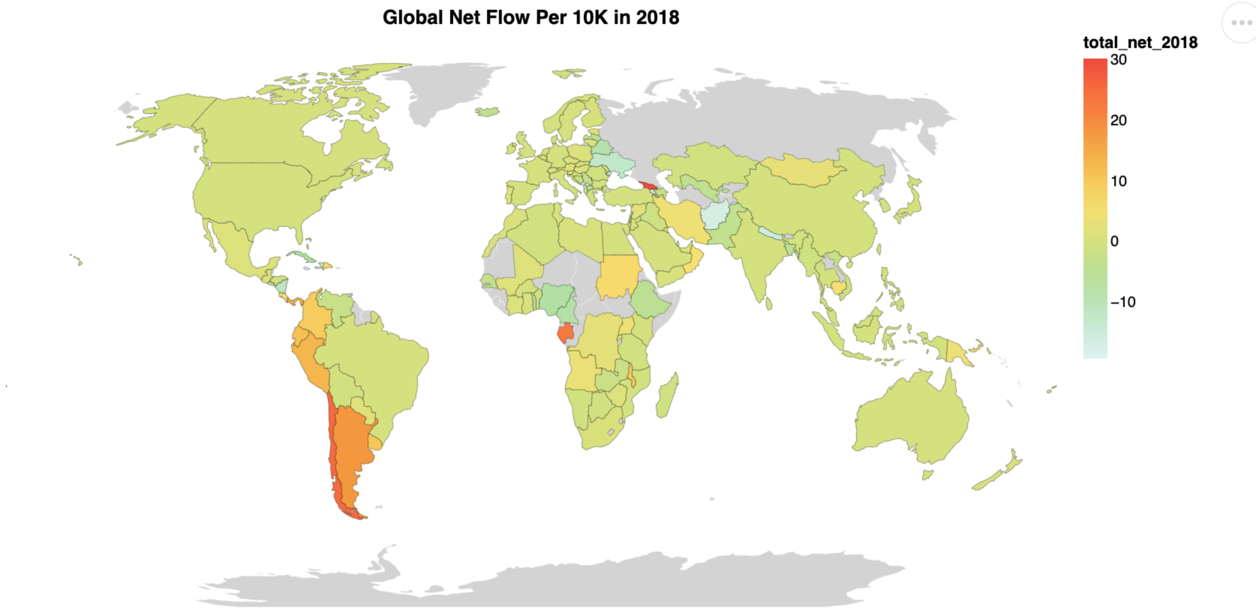
For 2016, we the see that the strip has high net flow, but as high as before and also the same for the region in Africa as noted for 2015. We do see that some negative net flow has entered in this map than before, with the colder regions, some regions in Africa, and in the middle east. The rest are mostly 0 net flow, being green.

Out[13]:



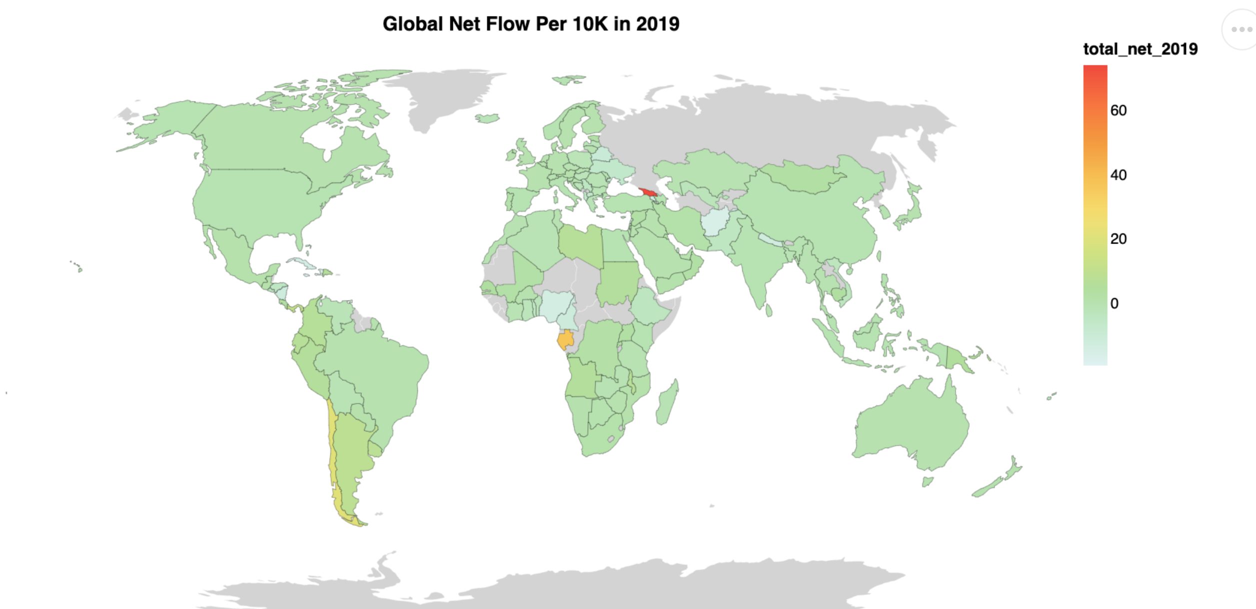
We can see that for the net flow here in 2017, that south west of South America has got high positive net flow and colder regions in the north and south have gotten negative net flow, along with some regions in Africa. So the high netflow has changed from that strip to this south west region.

Out[14]:



We see that other than the south west region of South America, a middle west region of Africa has now high net flow, along with a region in the middle east. The prior pattern of 0 and negative net flow continues in this map.

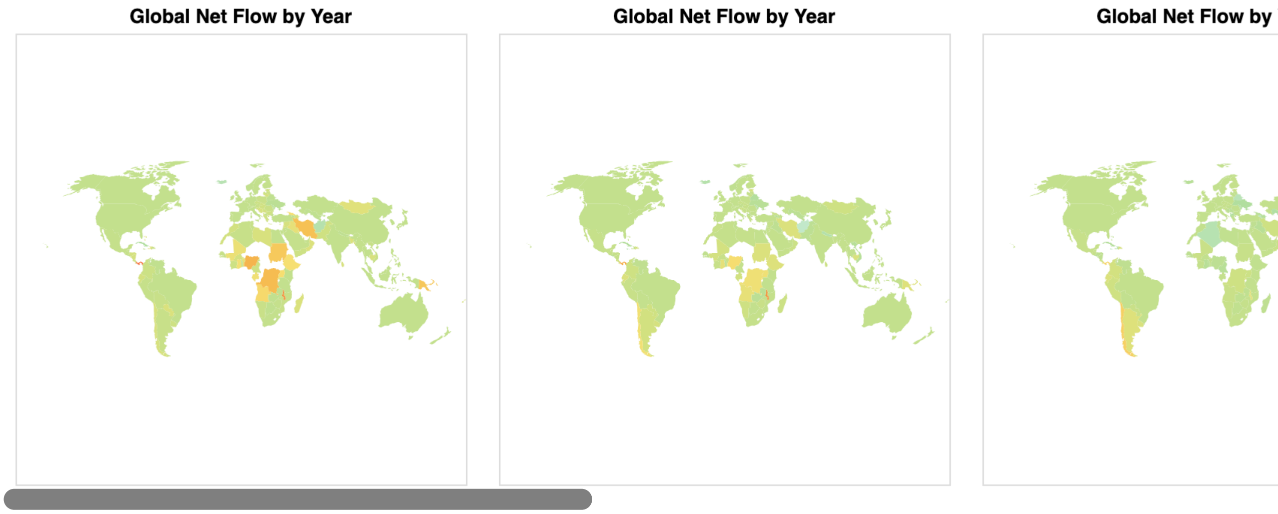
Out[15]:



The middle east region persists to have much higher positive net flow than before and the rest have mostly 0 net flow. The south west region aforementioned has now positive, albeit much lower positive net flow.

Below is a side-by-side comparison of this change over time:

Out[16]:



We can see the regions in South America, the middle east, and Africa have the most fluctuation in terms of net flows. The rest mostly remain the same. These regions have popped up earlier because of our prior analysis regarding time series.

Exploring how income class relates to a country's count of net-flow exchanges

We cacluated each country number of connection to other countries based on wb_income ranking:low-income,low middle income, the upper middle income, high income count. The low_count is the number of low-income countries a country has exchanges with. The lower_middle_count is the number of Lower Middle Income countries that a country has exchanges with. The upper_middle_count is the number of Upper Middle Income countries that a country has exchanges with. The high_count is the number of High Income Countries a country has exchanges with.

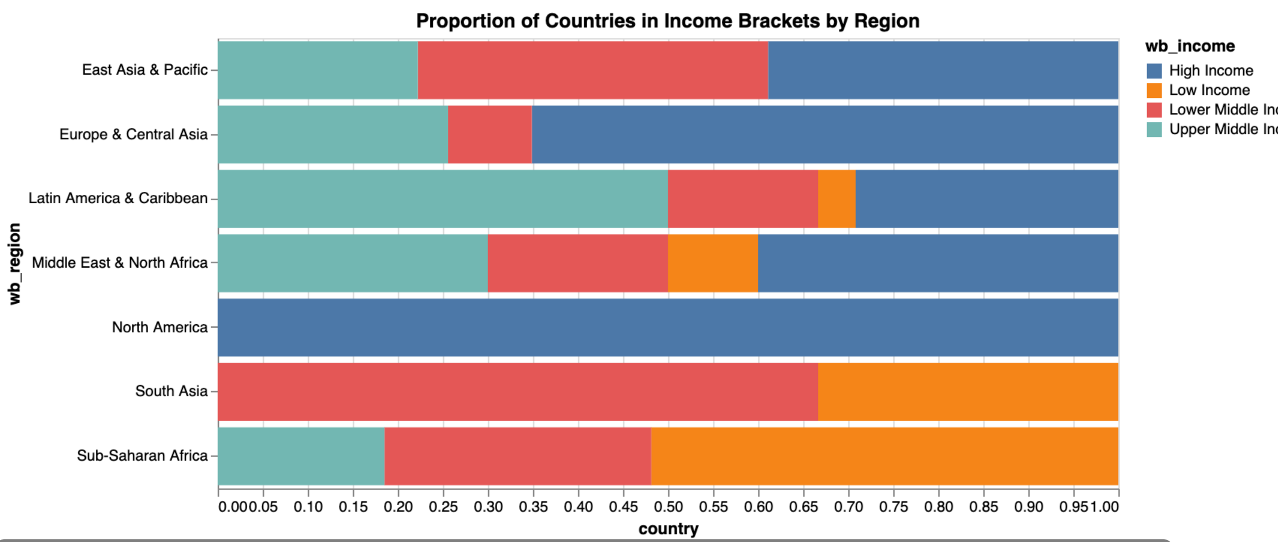
The count of countries per income bracket per region:

Each of the seven regions identified in the data set

Out[21]:

	wb_region	wb_income	country
0	East Asia & Pacific	High Income	7
1	East Asia & Pacific	Lower Middle Income	7
2	East Asia & Pacific	Upper Middle Income	4
3	Europe & Central Asia	High Income	28
4	Europe & Central Asia	Lower Middle Income	4
5	Europe & Central Asia	Upper Middle Income	11
6	Latin America & Caribbean	High Income	7
7	Latin America & Caribbean	Low Income	1
8	Latin America & Caribbean	Lower Middle Income	4
9	Latin America & Caribbean	Upper Middle Income	12
10	Middle East & North Africa	High Income	8
11	Middle East & North Africa	Low Income	2
12	Middle East & North Africa	Lower Middle Income	4
13	Middle East & North Africa	Upper Middle Income	6
14	North America	High Income	2
15	South Asia	Low Income	2
16	South Asia	Lower Middle Income	4
17	Sub-Saharan Africa	Low Income	14
18	Sub-Saharan Africa	Lower Middle Income	8
19	Sub-Saharan Africa	Upper Middle Income	5

Out[22]:

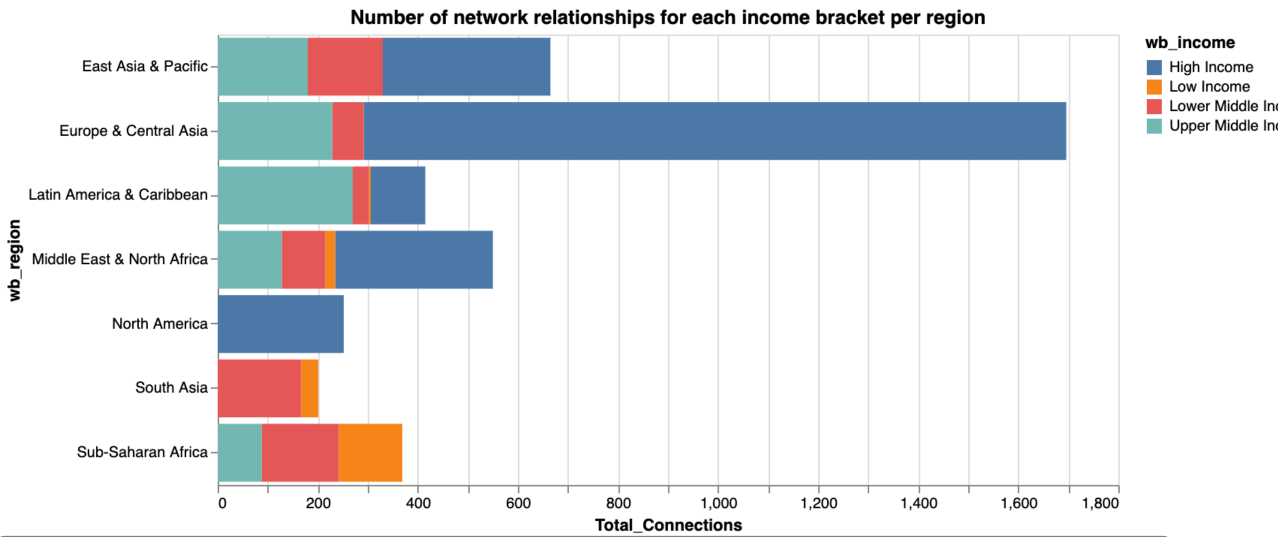


We can see the proportion that is made up by countries in different income brackets per region.We see that Asian, Latin American, and Carribean regions have countries in high income levels but some East Asian and sub-saharan African countries have a higher percentage of lower or middle income levels. We can now check the number unique wb regions, the countries, and group ids, as well as categories and names.

Distribution of network relationships by income for each country

We were interested in examining the proportion of networked relationships each country had by income bracket. For example, in the table below, we see that the United Arab Emirates, which is classified as a High income country and only 8.6% of its network relationships is comprised of low-income countries, where as it has 43% of its network relationships comprised of high-income countries.

Out[23]:



So we see here that North America, which only has High income countries has about 250 total network relationships (canada + us) And South Asia only has low income and lower middle income countries and they show the lower middle has around 180 relationships and the low income country has roughly like 40 relationships.

Skill group analysis

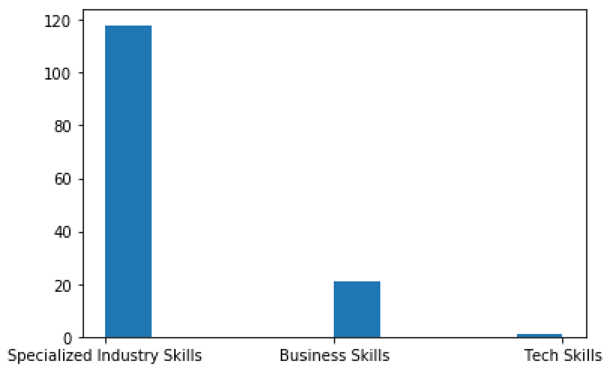
In the Skill Migration dataset, I was curious to see what the most important skill group category was in each country. I sorted the dataset to determine the most frequently applied skill group category for each country and plotted the results. From the plot, we see that the majority of the countries we sampled data from prioritized hiring Specialized Industry Skills, with Business Skills and Tech Skills rounding up the top 3. If I were applying for a job, I could look at the data and develop a toolbox to cater more to an in-demand skill group category to have a higher probability of getting hired. The accompanying table is the sorted dataset I used to create the plot.

Out[25]:

	country_code	country_name	wb_income	wb_region	skill_group_id	skill_group_category	skill_group_name	net_per_10K_2015	net_per
0	af	Afghanistan	Low income	South Asia	2549	Tech Skills	Information Management	-791.59	
1	af	Afghanistan	Low income	South Asia	2608	Business Skills	Operational Efficiency	-1610.25	
2	af	Afghanistan	Low income	South Asia	3806	Specialized Industry Skills	National Security	-1731.45	
3	af	Afghanistan	Low income	South Asia	50321	Tech Skills	Software Testing	-957.50	
4	af	Afghanistan	Low income	South Asia	1606	Specialized Industry Skills	Navy	-1510.71	
5	af	Afghanistan	Low income	South Asia	3139	Disruptive Tech Skills	Materials Science	-1085.03	
6	af	Afghanistan	Low income	South Asia	1315	Specialized Industry Skills	Criminal Law	-687.80	
7	af	Afghanistan	Low income	South Asia	1017	Soft Skills	Problem Solving	-906.42	
8	af	Afghanistan	Low income	South Asia	2130	Tech Skills	Software Development Life Cycle (SDLC)	-1096.96	
9	af	Afghanistan	Low income	South Asia	2265	Specialized Industry Skills	Cybersecurity	-1046.26	

Out[26]:

	country_name	wb_region	skill_group_category
0	Afghanistan	South Asia	Tech Skills
1	Afghanistan	South Asia	Business Skills
2	Afghanistan	South Asia	Specialized Industry Skills
3	Afghanistan	South Asia	Tech Skills
4	Afghanistan	South Asia	Specialized Industry Skills
5	Afghanistan	South Asia	Disruptive Tech Skills
6	Afghanistan	South Asia	Specialized Industry Skills
7	Afghanistan	South Asia	Soft Skills
8	Afghanistan	South Asia	Tech Skills
9	Afghanistan	South Asia	Specialized Industry Skills

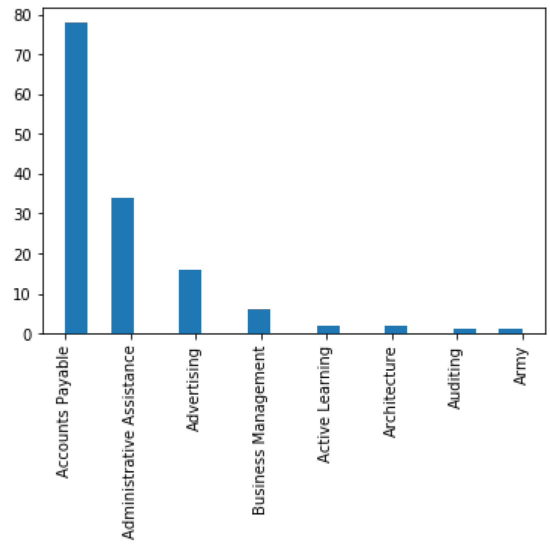


Above are the most popular skill_group_category amongst all 140 countries recorded. Since alot of specialized industry skills are rising, we can see that technological advances are not really the motive as previously hypothesized. We actually service industries, business industries, and other personnel involved with non tech-savvy jobs that are having the most active migrating capacity.

Out[28]:

	country_name	wb_region	skill_group_category	skill_group_name
0	Afghanistan	South Asia	Tech Skills	Information Management
1	Afghanistan	South Asia	Business Skills	Operational Efficiency
2	Afghanistan	South Asia	Specialized Industry Skills	National Security
3	Afghanistan	South Asia	Tech Skills	Software Testing
4	Afghanistan	South Asia	Specialized Industry Skills	Navy
5	Afghanistan	South Asia	Disruptive Tech Skills	Materials Science
6	Afghanistan	South Asia	Specialized Industry Skills	Criminal Law
7	Afghanistan	South Asia	Soft Skills	Problem Solving
8	Afghanistan	South Asia	Tech Skills	Software Development Life Cycle (SDLC)
9	Afghanistan	South Asia	Specialized Industry Skills	Cybersecurity

After creating the previous plot, I was interested in going a little bit deeper to see which skill groups were the most popular by country. I believed that this sort of the dataset would help in identifying the skill groups that are the most in-demand amongst the 140 countries that data was sampled from. From the plot, we see the Accounts Payable was by far the most in-demand skill group with more than half of the 140 countries having that as their most frequent skill group, with Administrative Assistance and Advertising rounding out the top 3. The accompanying table is the sorted dataset I used to create the plot.



Now we can see the histogram for the 8 most popular skill groups amongst the 140 countries. We see that these are financial, administration, and advertising being of the highest demand. This also makes sense since most of the LinkedIn's demographic was outlined to be people looking for on the desk jobs.

Discussion

In summary, we analyzed the relationship between country migration and the flow of industry, and talent. In particular, we note that there was prominent migration in the following regions: South America, Africa, and the Middle East. These regions had a high positive net flow that appeared to be related to lower and middle-income levels. In the last past five years, these regions still had the most activity, for migration.

Most notably, we observed that there positive net flow in high-income levels overall in terms of net flow. After looking at the visualizations, surprisingly found that specialized industries have the highest net flow, along with the financial personnel. We initially thought that due to the technological advancements worldwide, that department have the highest action, but it was actually industrial and business personnel that are experiencing most analysis. Further analysis could have been done on exploring the network relations between skilled staff and industries. We could also explore the income levels changing over the year and try to look into the wealth of the nations in question.