Online "Brain Gain": Do Migrants Return Knowledge Home?[*]

Olga Slivko[**]

April 17, 2019

Abstract

While migrants often deliver economic benefits to the receiving countries, are there any positive side effects from emigration for the sending countries? In this paper, I assess the role of migration in knowledge diffusion focusing on the migrants' online content contributions in their native languages from (about) destination countries on the world's most popular online knowledge platform Wikipedia. I find more content editing from the destination countries in the languages of sending countries due to immigration, especially, due to the inflow of educated migrants. However, the size of content increases only marginally, which means edits can be considered quality-improving. The mechanism driving content contributions is the inflow of new content editors. I address the endogeneity concern by constructing an instrument for the migrant flows based on the historical migration of female migrants. The findings of this paper highlight the potential of knowledge diffused by migrants to mitigate the digital divide between countries.

Keywords: Immigration; Knowledge Diffusion; User-generated Content; Wikipedia.

JEL Classification Numbers: O15, O33, H41, L86.

[*] I benefitted a lot from discussions with Irene Bertschek, Sabine Gralka, Shane Greenstein, Reinhold Kesler, Michael Kummer, Toomas Hinnosaar, Mary O'Mahony, Thomas Niebel, Marianne Saam, Patrick Schulte, Michele Tuccio, Steffen Viete, Michael Ward, Michael Weber, Xiaoyu Xia and Michael Zhang. I am also indebted to Tobias Werner, Niklas Isaak, Lukas Trotner and Martin Reinhard for their excellent research assistance. I would like to thank the participants of research seminars at ZEW, IWM KMRC Tübingen, Telecom ParisTech, Catolica Lisbon Business School and conferences ASSA 2017 (Boston), IIOC 2017 (Chicago), IFO Dresden, IAB Nürnberg, MACCI Mannheim. Financial support from Wissenschaftscampus Tübingen is gratefully acknowledged.

[**]Address: L7, 1 68161 Mannheim, Telephone: +49 (0) 621 1235 358, E-mail: *olga.slivko@zew.de*

1 Introduction

International migration is critical for knowledge diffusion. For the migrants' destination countries, the inflow of high-skilled migrants fosters competitiveness by stimulating creation and diffusion of knowledge, in the form of patents and scientific output (Hunt and Gauthier-Loiselle (2010), Bosetti et al. (2015), Kerr and Lincoln (2010), Moser et al. (2014) and Ganguli (2015)). Accordingly, many countries launch policies aimed at attracting high-skilled migrants.

In the sending countries, on the contrary, the outflow of skilled human capital has been concerning to policy makers as it causes long-run decline in innovation (Waldinger (2010), Waldinger (2016)) as well as in the quality of institutions (Anelli and Peri (2017), Docquier et al. (2016)). Apart from widely acknowledged negative effects of emigration, a handful of studies focused on knowledge externalities to the sending countries due to diaspora located abroad (Kapur (2001), Douglas (2015), Bahar and Rapoport (2016), Fackler et al. (2018)). Bahar and Rapoport (2016) suggest that migration transforms the structure of exports in the sending countries increasing the share of products that are intensively exported by the migrants' destination countries. Nanda and Khanna (2010) show how migration from developing countries facilitates entrepreneurship there due to network connections with the abroad diaspora. Even the prospects of migration can yield economic benefits fostering investments in education in the developing countries due to the higher potential returns abroad (Beine et al. (2001)). Further understanding of how diasporas can transmit information about technologic opportunities to their home countries remains a very important gap in the literature (Kerr et al. (2017)).

In this paper I assess the impact of immigration on knowledge diffusion, focusing on broader knowledge beyond scientific publications and patents. I analyze whether immigrants to destination countries subsequently engage in contributing online knowledge in their native languages, thereby, transmitting knowledge to their home countries. I construct a dataset combining data on migration flows to OECD countries with the measures of online content contributions to Wikipedia from the destination countries in the native languages of migrants. I determine the locations of foreign contributors using records of IP-addresses of the anonymous contributors to Wikipedia. Then, to identify the effects of interest, I use the variation in the migration flows and assess how it is related to the variation in content generation. Panel regressions with country pair fixed effects allow to account for the unobserved time-invariant heterogeneity between the pairs of countries and control for language-specific trends in content generation. I further address the endogeneity concern regarding the location choices of migrants using an instrumental variables approach drawn from the labour literature (Altonji and Card (1991), Card (2001)). According to this approach, the yearly migration flows between the sending and the destination country are instrumented for with the share of the total yearly outflow of migrants from the origin country that arrive to the destination country due to the past location of the diaspora. While the concern remains that the shares of diaspora from an origin country to a destination country are correlated with knowledge contributions

I alleviate it by using the historical shares of *exclusively* female stock for each country. According to the empirical evidence, the contributions by females represent a minor share of total content contributions to Wikipedia (Lam et al. (2011)). Hinnosaar (2017) explains this by less frequent usage of Wikipedia and lower self-confidence regarding the necessary competencies for knowledge contributions.

The results of the paper suggest that, due to immigration more content contributions take place on Wikipedia from destination countries in the language of the immigrants' origin countries, in both the science-related domain, "Scientists", and the broad-interest domain "Cuisine". While editing activity increases both from and about destination countries, there is only marginal increase in the size of content contributions. The increase in editing is driven by the increase in the number of new contributors on the platform. Following the information systems literature on Wikipedia, immigration seems to channel occasional contributors, who may have more expertise or be more interested in a specific topic as compared to loyal (registered) Wikipedia community members (Anthony et al. (2009)). With the instrumental variables approach, the estimates become stronger in magnitude indicating that the panel data fixed effects regressions may underestimate the causal effects of interest.

The research question tackled in this paper gained importance with the alarming growth in the "digital divide". The skyrocketing development of ICT in the past decades made the ability to use computers and the Internet one of the key factors for individuals to completely immerse oneself in the economic, political, and social aspects all over the world. However, despite the increasing access to computers and Internet, the gap between individuals in terms of access to the benefits provided by ICT also increases. According to Pew Research Center, in 2018 there are still substantial differences across countries as well as large age, education and income gaps in Internet usage and smartphone ownership within the population of American and European countries.[1] In addition to the differences in Internet infrastructure and hardware, "digital divide" manifests itself in gaps in software and web-content availability. Figure A1 illustrates this by displaying the differences in knowledge coverage in the largest Wikipedia language editions measured by the total number of articles in each language. Even within the top-20 language editions, the coverage of knowledge in the third most represented language, Swedish, and the fifteenth, Portuguese, differs almost by the factor of 4. As immigration is skewed towards developed countries with the US leading (OECD, 2013), if immigrants were to transmit the acquired knowledge in their native languages, this could be a factor potentially mitigating the "digital divide". Moreover, the rapidly gaining importance platforms, which rely on user-generated content, facilitate such knowledge diffusion allowing the consumers of knowledge to assume the roles of knowledge contributors.

Knowledge diffusion via Wikipedia can have important economic implications. Studies conducted randomized field experiemnts and showed that online content on Wikipedia matters for the spread of scientific knowledge (Thompson and Hanley (2017)) and for consumer decisions (Hinnosaar et al. (2017)).

[1] http://www.pewresearch.org/fact-tank/2017/04/21/smartphones-are-common-in-advanced-economies-but-digital-divides-remai

Therefore, policy makers in the origin countries could take a different approach to emigration. Contrary to the negative perception of the outflow of skilled migrants from the origin countries, migrants could be considered as a source of knowledge about foreign advances in science and culture, and could be given incentives to contribute knowledge in their native languages.

The remainder of this paper is structured as follows. Section 2 presents the literature related to the study. Section 3 describes the data sources and the resulting sample construction. Section 4 presents the baseline empirical analysis and Section 5 reports its results. Section 6 addresses the endogeneity concern proposing the instrumental variable approach. Finally, Section 7 discusses the results and concludes.

2 Relation to the Literature

This paper bridges two strands of literature. The first strand aims at assessing the impact of migration on knowledge diffusion, both in the destination and the origin countries of immigrants. The studies focus on technological knowledge measured by patents (Kerr (2008), Agrawal et al. (2011), Breschi et al. (2017), Miguelez and Temgoua (n.d.), Moser et al. (2014) and Hunt and Gauthier-Loiselle (2010)) and on scientific knowledge measured by scientific articles (Waldinger (2010), Waldinger (2016), Ganguli (2015) and Borjas and Doran (2015)). Diaspora is shown to generate knowledge spillovers affecting the levels of manufacturing in the migrants' home countries (Kerr, 2008). Agrawal et al. (2011) show that while the net effect of high-skilled emigration on domestic knowledge production is negative, a better access to knowledge via diaspora links is valuable for the most prominent inventions. Breschi et al. (2017) find no evidence of "brain gain" effect measured as the frequency of citations of migrant inventors in their home countries. Overall, the studies highlight the importance of geographic proximity and interpersonal communication between researchers or inventors for exchanging relevant scientific knowledge and know-how (Ganguli (2015), Agrawal et al. (2006), Breschi and Lissoni (2009), Kerr (2008), Agrawal et al. (2011), Bahar and Rapoport (2016), Kapur (2001), Foley and Kerr (2013)).

Remarkably, the power of diaspora networks persists even in the era of online communications and on unified online labour markets. Ghani et al. (2014) show that diaspora connections determine the choice of employees for outsourcing tasks: ethnic Indians are more likely to choose a worker in India when hiring on the online platform oDesk, despite the efforts of the platform to minimize frictions and provide full information about workers. Similarly, Agrawal et al. (2012) suggest that while workers in developing countries may face initial disadvantages on oDesk, these diaspora-based links could provide an opportunity to overcome initial uncertainty about workers. Such diaspora effects can be important for the economic integration of developing countries, their economic transition and growth.

Causal identification of the effect of interest has been one of the central issues in the literature on migration. Recent studies increasingly use shocks to migration in the origin countries, such as major economic and political crises. Waldinger (2010), Waldinger (2016) and Moser et al. (2014) use the dismissal of Jewish scientists and inventors in Nazi Germany in the 1930s and the bombing during World War II. Borjas and Doran (2015) and Ganguli (2015) use the collapse of the Soviet Union to estimate the spread of Soviet scientific knowledge to the US and the effect on the subsequent research productivity of US scientists.

The present study adds to this strand of literature by exploring the effect of immigration on the diffusion of broader knowledge, access to which is facilitated by online knowledge repositories, such as the online encyclopedia Wikipedia.

The second strand of literature this paper is related to analyzes the incentives of the web crowds to invest time and effort into contributing online knowledge to open-source knowledge repositories. Since long time ago, the economists were trying to assess the factors of the tremendous growth potential of the open source production model. Following Lerner and Tirole (2003), the open source development captures user-driven innovation taking place in many industries. Among mechanisms driving user participation in open source, scholars highlight social spillovers (Zhang and Zhu (2011), Algan et al. (2013), Jan Piskorski and Gorbatâi (2017), Ren et al. (2015)) and content spillovers (Kummer (2013), Aaltonen and Seiler (2015), and Kane and Ransbotham (2016)). Using the block of Chinese Wikipedia in mainland China in a natural experiment setting, Zhang and Zhu (2011) show that when the recipient group size shrinks, contributors' incentives diminish. Algan et al. (2013) show that reciprocity and social image concerns predict the willingness of individuals to engage in peer-production communities. Gallus (2016) shows that symbolic graphic awards on the user pages have a strong effect on newcomer retention in the Wikipedia community. Similarly, Huang et al. (2018) highlight the role of performance feedback for motivating production of user-generated content. Jan Piskorski and Gorbatâi (2017) provide evidence on how social norm reinforcement works in decentralized online communities, such as Wikipedia. They show that in denser networks violators of norms are more likely to be punished by the community. Ren et al. (2015) relate the disparity in tenure and variety in interests of content contributors with higher productivity in online content generation. Similarly, Ransbotham and Kane (2011) suggest that articles written by a mixture of new and experienced participants are more likely to become high-quality articles.

In addition to social spillovers, content spillovers are shown to be important drivers of content contributions to online communities. Nagaraj (2017) studies information seeding on OpenStreetMaps open-source community in a natural experiment setting and suggests that information seeding can crowd out individual incentives to contribute content. Opposingly, Aaltonen and Seiler (2015) suggest the presence of a cumulative effect in online content growth based on observational data from the Wikipedia category "Roman Empire". Kane and Ransbotham (2016) show that in open collaboration communities consumption and contribution mutually reinforce each other at the earlier stages of the content lifecycle. Zhu et al. (2018) leverage a large-scale natural experiment and suggest that exogenously added content increases both content consumption and subsequent contributions, as well as attention to downstream hyperlinked articles. Hinnosaar et al. (2019) show no sizable externalities in content production in a field experiment setting. Kummer (2014) highlights the role of content structure, namely, hyperlinks connecting articles, in channeling attention spillovers from focal articles to neighboring ones. The present study contributes by pointing out an additional channel that can play a role in online knowledge diffusion.

Literature on user-generated content has been concerned with information biases (Greenstein and Zhu (2012), Greenstein and Zhu (2018), and Hinnosaar (2017)). Greenstein and Zhu (2012) show that article slant on Wikipedia reduces over time driven by the entry of new articles representing the opposing points of view to the older ones. Greenstein and Zhu (2018) show that political bias of articles biases reduces over time due to editing activity, but this effect is strong only for the most viewed articles, which receive enough edits to achieve no difference in terms of slant and bias with Encyclopedia Britannica, while the rest of articles show considerable difference in slant and bias from their Britannia counterparts. While Greenstein and Zhu (2018) focus on contested knowledge, concretely, political bias of articles, Hinnosaar (2017) studies gender the presence and the causes of the gender bias. All studies highlight the profound long-term effects of biases and the importance of exposing diverse opinions and views online. The present paper suggests that migration mitigates inequalities in knowledge representation and, therefore, decreases biases in topic coverage between languages. Further, some information system studies also tackle the patterns of user-generated content contributions, specifically, geographic and gender representation (Graham et al., 2014; Lam et al., 2011; Sen et al., 2015). (Graham et al., 2014) and Sen et al. (2015) show that the geographical inequalities in knowledge production increase the coverage of areas with higher socioeconomic status on Wikipedia, while developing countries might fail to reach a critical mass of editors.

Furthermore, this paper relates to other studies that draw attention to the implications of migration for the emigrants' origin countries due to money remittances (Asatryan et al. (2017) allowing people to pay for better education and health or to start businesses, Di Giovanni et al. (2015)), Foreign Direct Investment (Javorcik et al. (2011), Kugler and Rapoport (2007), Foley and Kerr (2013)), international trade (Gould (1994)), international R&D collaborations (Miguelez (2016)) and diffusion of ideas (Kerr (2008)), entrepreneurial connections (Nanda and Khanna (2010), Foley and Kerr (2013)) and democratization (Docquier et al. (2016), Barsbai et al. (2017)).

3 Data

3.1 Migration

For the analysis of the effect of migration on online knowledge contributions I combine data from several sources. The migration flows between country pairs come from the OECD International Migration

Database. The data include figures of inflows, outflows and stocks of migrants for each country of destination belonging to OECD distinguishing the migrants' countries of origin. The observations cover the years 2000 to 2016.

The cross-country immigration flows in the sample are highly skewed with a long right tail. Figure 1 presents the median flow of immigrants between origin and destination countries for the country pairs with the highest migration flows in the final sample. The high levels of migration are observed between the neighbouring countries of Germany and Poland. Germany along with the US are in general among the most popular destination countries. According to the bar histogram, the highest migration takes place within Europe as well as from the Asian countries to the US.

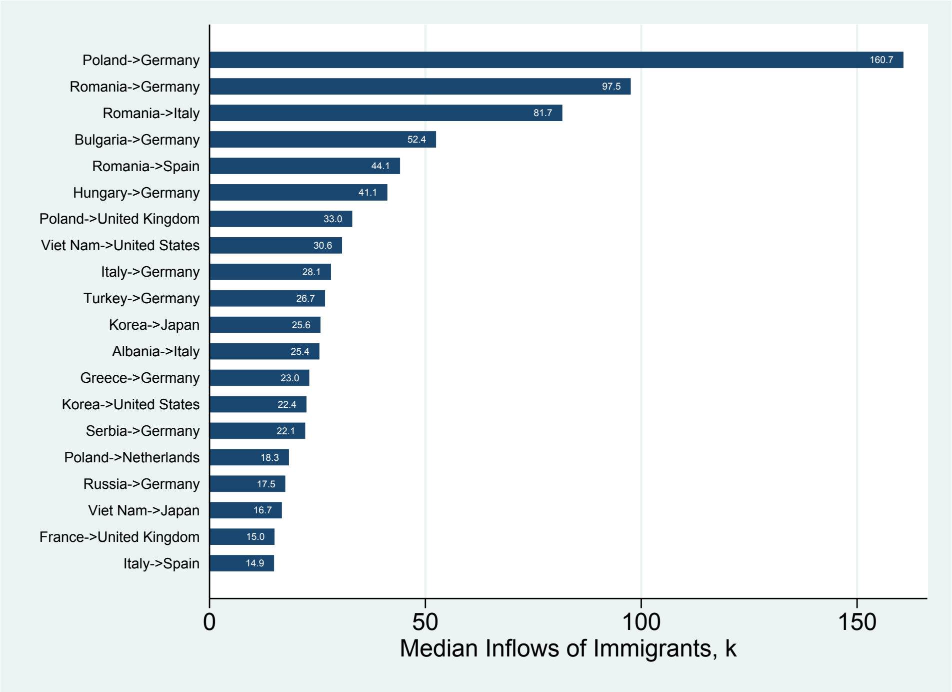
OECD yearly data on migration does not distinguish the education levels of migrants. Therefore, I add the data from the Database on Immigrants in OECD Countries (DIOC), which contains data from a survey conducted only in the years 2000/01, 2005/06 and 2010/11. The questions in the survey cover the age, nationality, duration of stay, education and labour force status of respondents. The data distinguish three levels of education: (1) primary and lower secondary, (2) (upper-) secondary or post-secondary non-tertiary education, and (3) the first (Bachelor or Master) or the second stage of tertiary education (PhD). For each pair of origin-destination countries I construct the share of foreign-born individuals[2] from the origin country residing in the destination country with tertiary education (3). Further, I apply these shares to OECD yearly immigration flows to approximate the flows of educated migrants. However, since the data on the shares of education levels are only available for few years and only a limited subset of countries, using this data implies fewer observations in the analysis. Therefore, I estimate the effect of educated migration on online content contributions only in the robustness checks.

The shares of immigrants with tertiary education for the top 20 country pairs from Figure 1 are presented in Figure 2. The share equal to 1 for the origin country Bulgaria and the destination country Germany are due to the fact that the DIOC data do not contain information about other education groups for the immigrants from Bulgaria to Germany.

To conduct the analysis of the effect of immigration on online knowledge diffusion, I further merge data on immigration flows with the measures of online knowledge contributions extracted from Wikipedia.

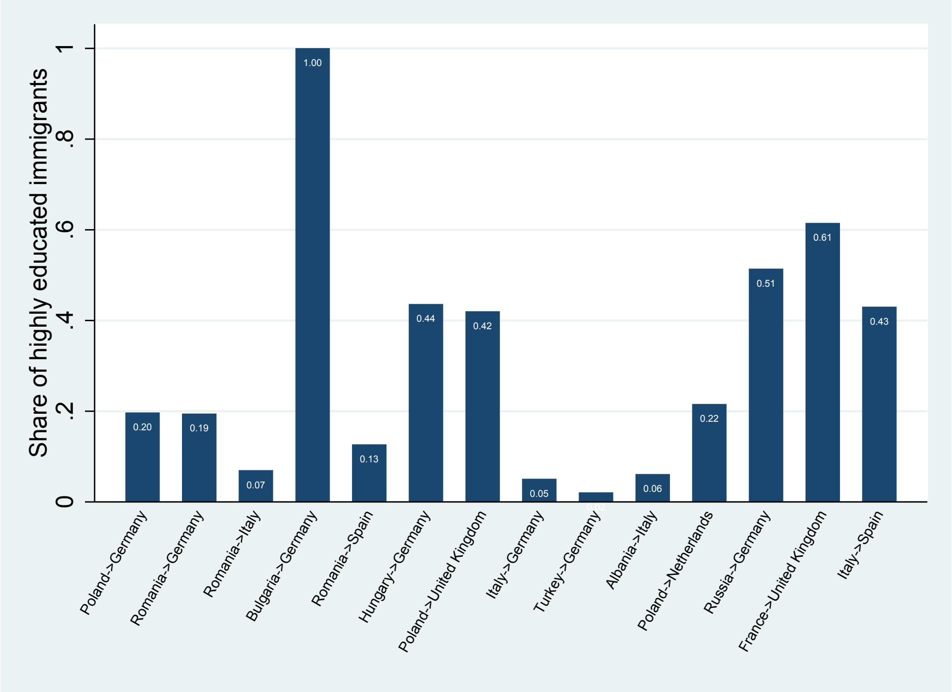
[2] I use educational levels of foreign-born individuals, who at the time of the survey had not yet been granted citizenship in their destination country.

Figure 1: Top-20 Immigration Flows to OECD Countries in 2006-2016.



NOTE: This figure illustrates the pairs of origin and destination countries with the highest immigrant flows averaged over the years in the sample.

Figure 2: The Shares of Immigrants with Tertiary Education for Top-20 Immigration Flows to OECD Countries in 2006-2016.



NOTE: This figure illustrates for each pair of countries from Figure 1 the share of immigrants with tertiary education based on DIOC OECD data.

3.2 Background on Wikipedia

Wikipedia was created in 2001 as a side project of Nupedia, a community where knowledge was produced by experts and and licensed as free content. Wikipedia enabled collaboration on articles prior to entering the peer-review process. Within few months, international language editions of Wikipedia emerged. In 2003, English edition of Wikipedia first passed 100,000 articles and the next largest edition, the German Wikipedia, passed 10,000 articles. By 2019, Wikipedia is the world's largest online knowledge repository and 5th most visited web-site in the world, behind Google, Youtube, Facebook, and Baidu with texts available in 301 languages. More than 35,6 million registered users contribute content to more than 47 million articles.[3]

3.3 Knowledge domains on Wikipedia

As the largest and the most viewed free-access online encyclopedia, Wikipedia represents an ideal setting for measuring the dissemination of online knowledge.[4] While Wikipedia articles cover a wide range of topics, I focus on online knowledge in two domains. For these domains, I collect the revision history of all articles dating back to 2006, because this was the year that Wikipedia became the most popular reference website on the Internet according to traffic monitoring company Hitwise.[5]

The Choice of Knowledge Domains. All contributions to Wikipedia are recorded and available to the public. Although the full data on Wikipedia articles edit history is publicly available, in order to have a better understanding of the composition of the data as well as the data generating process, studies focusing on Wikipedia typically analyze specific knowledge domains, for example, "Economics" (Kummer et al., 2016), "Chemistry" (Thompson and Hanley, 2017) or "Medicine" (Kane and Ransbotham, 2016). In this study I choose a large domain, "Scientists", describing scientists' biographies and scientific work as a proxy for the dissemination of knowledge relevant for innovation and technological progress. This domain is large enough to be representative for knowledge on Wikipedia. Knowledge in this domain can have important implications for individual choices related to education and career. Also, contributing to this domain would require a certain minimum level of education from the volunteering contributors. I contrast the findings for this domain with another domain that is accessed and contributed to by a potentially broader public, "Cuisine".

These two domains have an additional feature used in the analysis. Within them, there is a division of the domain by country. These country-specific categories are created and attributed by the Wikipedia community members. This allows me to test whether knowledge contributed by the migrants from the destination countries describes destination countries.

[3] The data refer to March, 2019.

[4] Wikipedia is the 5th most visited site in World Wide Web, after Google, Youtube, Facebook and Baidu according to alexa.com (March, 2019).

[5] `https://en.wikipedia.org/wiki/History_of_Wikipedia`

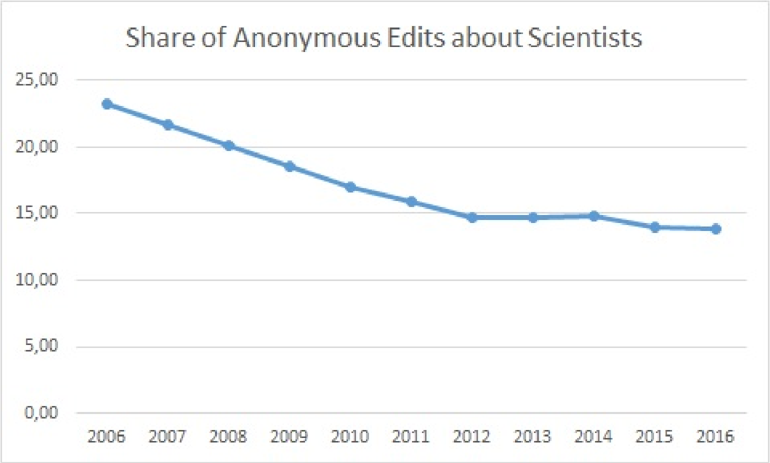
Data Extraction. For extracting information from the domains of interest I use the Wikipedia API tool. I determine the set of root categories describing country-specific topics in English, for example, “French scientists” (Figure A2 in Appendix), “German scientists”, “Italian scientists”, etc. For each of these root categories I download language links, as seen in the lower left corner in Figure A2. These language links imply in which languages I can obtain information about each root category. Then, for each language listed in the language links I can extract subcategories of the root category and articles that are directly related to the category (see panel (a) in Figure A2). Subsequently, for each subcategory I can go one level deeper in the category tree. For example, panel (b) in Figure A2 shows the outcome of going one level deeper from the subcategory “French scientists”. For each subcategory, I collect subsubcategories and articles belonging to it (this is one iteration). My goal is to automatically collect as many articles belonging to each domain as possible, while avoiding the inclusion of articles that do not directly correspond to the topic of the domain. At each iteration, I manually compare the fit of the articles I gain by going one step further with the amount of articles that do not fit and appear in the sample. Overall, for the domain “Scientists” I collected all articles for the four levels of subcategories of each country-specific category, in other words, the overall data collection consisted of four iterations. For the domain “Cuisine” in a country, it took five iterations to collect articles that fit the domain well.

After collecting the titles (and identifiers) of all articles describing knowledge domains “Scientists” and “Cuisine” about OECD destination countries in various languages, I extracted the full revision history of each article. It records all edits (in the Wikipedia community slang, “revisions”) to each article made over time, and for each edit the information available includes the identifier of the edit, the time stamp, the name of the user who made the edit, and the size of the article in bytes after the edit. Article revision history recorded in this way allows the creation of various measures of online content contributions for the purposes of my analysis. For each article, I can compute the total number of edits and the amount of added and deleted content (in bytes).

An important feature of Wikipedia is that the encyclopedia allows Internet users to edit articles in the two regimes. First, every online user can go to the “Edit” page of each article and introduce and save the modification directly, without logging-in. In this case the modification will be recorded in the edit history of the article with an IP-address of the user. Second, individuals who already have accounts on the Wikipedia platform can introduce their credentials and turn to editing. In this case, their modifications will be recorded in the edit history of articles with their user names, and their IP-addresses are not available to the public.

Since I can determine location only for the anonymous users, the analysis in this paper focuses on content contributed anonymously. Over time the share of anonymous contributions varies from 23% to 15% (Figure 3). According to the researchers, contributions from anonymous users are as reliable in terms of quality as those of registered users, and have a better quality compared to registered users with few contributions. Anthony et al. (2009) suggest that while “registered participants ... make many contributions with high reliability”, “the highest reliability comes from the vast numbers of anonymous ‘Good Samaritans’ who contribute only once...”. Moreover, anonymous editors often subsequently register as the members of the community.[6] This is why contributions of anonymous users enhanced by migration are important.

Figure 3: The Share of Anonymous Contributions in Total Contributions to Wikipedia in the Sample



Note: This figure displays the share of contributions in the sample performed by anonymous contributors.

Using the IP-addresses of anonymous contributors I match their contributions to Wikipedia with their geolocations. The R library “rgeolocate” returns the country and the city of the IP-address.[7] Once each anonymous content contribution is matched with its geolocation,I aggregate the measures of content contributions at the level of languages of contributions (spoken in the origin countries) and geolocations in destination countries.

While computing the aggregate indicators of content contribution I distinguish between content generation performed by humans and automated activity. To ensure that automated activity is not affecting my main results, I exclude from the analysis all the activity performed by automated algorithms, i.e. bots. Further, I also separately account for unproductive activity by humans, which occurs on Wikipedia due to content reversions. By design, every user on Wikipedia can make edits of the content. Moreover, every content modification can be reverted restoring the previous version of the article. If an edit represents a clear act of vandalism, or simply if the other community members do not agree with the modification, the edit can be reverted. In this case, the subsequent revisions add or delete the same modifications, and this can not be accounted for as productive (valuable) content contributions. Therefore, I separately analyze content contributions that were not accepted by the community and later reverted.

[6] As Wikimedia blog suggests, 10% of new registrations in English are editing anonymously before registration, while in other major languages this proportion can be even higher : 18% in German Wikipedia, 19% in Spanish Wikipedia, and 21% in Japanese Wikipedia (`https://blog.wikimedia.org/2014/05/16/anonymous-editor-acquisition/`).

[7] I further verify that the precision of the match of IP-addresses to countries remains robust to the use of a specific geolocation library in R.

An increase in reverts can in itself be an interesting indicator for the activity occurring on the platform. It can signal that due to the arrival of new and unexperienced users the quality of content contributions has fallen and the community has to take measures to filter the content arrival. Finally, I exclude from the analysis all content modifications resulted from the act of edit reversion as they do not add anything new to the content on Wikipedia.

Merging data on migration flows with indicators of online content generation on Wikipedia yields the final data set with observations over the years 2006-2016. In order to properly merge migrants' contributions to Wikipedia from destination countries in the languages of origin countries with migration flows, I have to exclude the origin countries whose main languages are spoken all over the world, such as English, Spanish, Portuguese, Persian or Arabic, because it be would difficult to attribute knowledge about a destination country contributed in these languages, for example, in English, to any particular country of migrants' origin. Also, excluded are contributions in German from all German speaking countries (Austria, Switzerland and Germany). Similarly content contributions in French and Dutch from, correspondingly, French and Dutch speaking countries are excluded. Importantly, I also exclude the Chinese Wikipedia from my dataset, because it has been banned in mainland China since 2008. As Zhang and Zhu (2011) show that after the ban of Wikipedia in mainland China contributions in Mandarin from other neighbouring Mandarin-speaking countries, such as Singapore or Taiwan, also decreased due to the drastic reduction of the recipient group. Overall this ban substantially reduced the quality and the relevance of Chinese Wikipedia within as well as outside mainland China. The resulting set of countries of immigrant origin is displayed in Figure 4.[8]

The countries of immigrants' origin include European countries, former Soviet Union countries and Asian countries (see the map on Figure 4 and Table A2 in Appendix). The highest outflows of migrants could be observed from Poland and Romania.

3.4 Descriptive statistics

After merging the OECD immigration data with Wikipedia content generation indicators, I obtain the sample composed of pairs of origin and destination countries of migrants. There are in total 38 origin and 31 destination countries.

The destination countries include all OECD countries (Table 1). The table presents some descriptive evidence that strong immigration inflows into countries like the US, the UK, Germany, Italy or Spain are associated with large amounts of content added to the research-related domain, "Scientists", as well as the general-interest domain, "Cuisine".

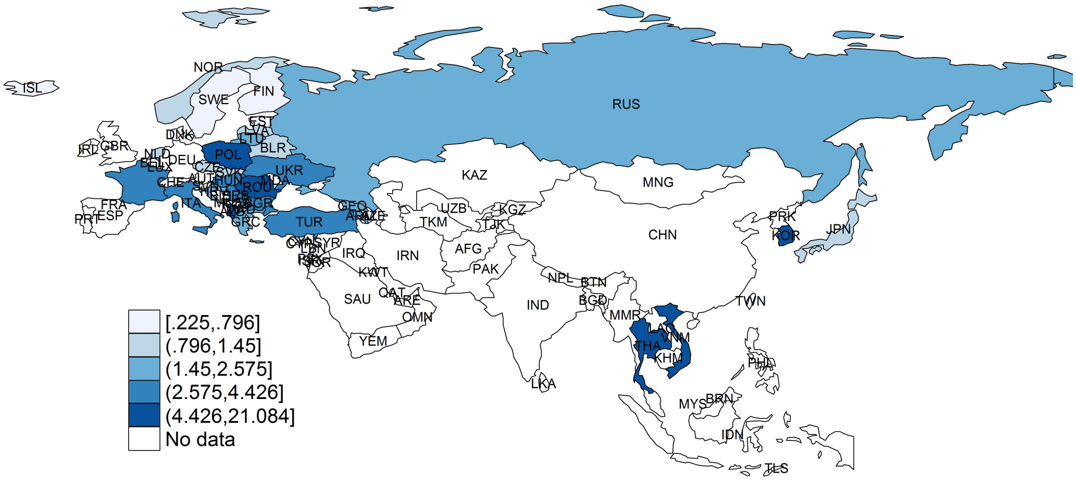
[8] In the baseline specifications, where I cannot trace the locations of content contributed, I exclude the two countries, which are technological leaders, Germany and the United States. The Wikipedia language editions for these languages are among the largest in the world and, therefore, they attract contributions about these countries in a variety of languages spoken in the world. In the main specifications (my preferred results), I do not need to exclude them as I can perfectly map contributions made in foreign language from the territory of Germany or the United States.

Table 1: Total immigration inflows into destination countries and total content contributions in the domains Scientists and Cuisine about each host country over the years 2006-2016.

	Total Immigrant Inflow (k)	Anonymous Content (kbytes), ***Scientists***	Anonymous Content (kbytes) ***Cuisine***
Australia	175.63	156.11	490.67
Austria	652.35	97.68	14.11
Belgium	359.34	219.40	33.00
Canada	213.13	220.27	14.48
Chile	9.76	28.25	1.69
Denmark	111.67	50.82	6.30
Estonia	13.88	121.08	7.86
Finland	101.02	237.35	12.93
France	295.53	176.89	40.91
Germany	6594.29	2111.48	463.03
Greece	7.40	32.42	0.30
Hungary	98.85	23.90	5.33
Iceland	11.14	2.80	0.09
Israel	37.37	23.15	9.79
Italy	1650.95	344.09	105.15
Japan	888.95	124.34	38.01
Latvia	11.03	152.27	8.30
Luxembourg	16.40	15.53	0.39
Mexico	21.42	82.04	5.32
Netherlands	469.01	129.19	15.99
New Zealand	50.73	27.21	1.72
Norway	269.06	44.38	11.00
Poland	325.77	204.51	23.66
Portugal	73.06	54.65	11.05
Slovak Republic	31.76	182.17	7.14
Slovenia	54.47	8.62	0.87
Spain	1415.14	306.92	420.19
Sweden	270.08	131.23	27.27
Switzerland	330.18	89.46	54.55
United Kingdom	1309.00	294.43	81.73
United States	1431.20	2731.38	593.89

NOTE: For each destination country columns (1)-(3) display the total number of immigrants and the number of high-skilled immigrants coming to each country of destination (in rows) together with the total number of edits made in the destination country in each knowledge domain aggregated across all countries of immigrants' origin.

Figure 4: Origin Countries of Migrants' to OECD Countries matched to Wikipedia Language Editions



Note: This figure displays the final set of migrants' origin countries which are matched with the corresponding language editions of Wikipedia for the analysis. The intensity of the blue colour corresponds to the average migration flow from an origin country to any of the destination countries of immigrants.

Table 2 displays descriptive statistics for the explanatory variables of the analysis. In the sample, on average 3778 immigrants per year leave their country of origin and move to a destination country. Among them, the share of immigrants with tertiary education is on average 0.4, which yields on average 1,291 immigrants with tertiary education. Also, the overall statistics for the Wikipedia editions in languages of immigrants' origin countries suggest that, in the sample, the average language edition sees a yearly increase in the number of registered Wikipedia users by 2,759 persons, and receives 204 new articles per day and 2,250,000 edits each year.

Table 2: Descriptive statistics of immigration flows

	Mean	Std. dev.	Min	Max	# Obs.
Immigrants	3781	13757	1	271443	4576
Male Immigrants	2126	8872	0	145860	3673
Female Immigrants	1830	6255	0	146190	3673
Share of Educated Immigrants	.368	.178	.00195	1	2443
Educated Immigrants	1291	4951	.481	86274	2443
New Wikipedians	2757	3066	24	12106	4530
Total Articles in Language, k	458	519	3.3	3800	4530
New Articles per Day	204	353	1.75	3781	4530
All Edits in Language, k	2250	2322	32.2	9595	4530

Table 3 shows summary statistics for the measures of content contributions in both knowledge domains of interest, which are used as dependent variables throughout this study. Each indicator refers to the total activity from the destination country in the language of an immigrants' origin country. The content contributions are measured by the number of edits made from the destination countries in the languages of sending countries (including the number of edits that add or delete content, that change short or substantial amount of text), the number of new users (distinct IP-addresses that edit articles from destination countries in the foreign languages) and the edit distance measured as the number of symbols modified (including the total modified symbols, added or deleted symbols) due to content edits. The domain "Cuisine" receives 2 yearly anonymous edits on average, which is lower than for the domain "Scientists" (10 edits). In both domains, more content is being added than deleted, and more edits make longer content contributions, which means they modify more than a word in the text in of the article. A small share of edits introduced gets reverted by the community.

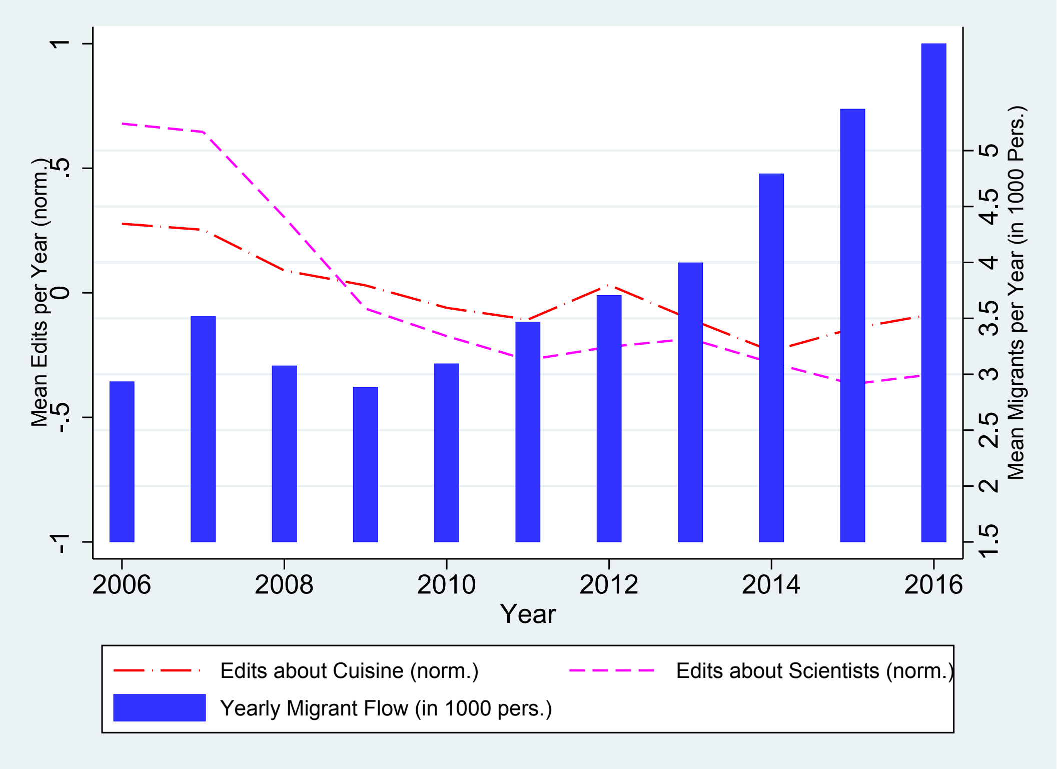
Table 3: Descriptive Statistics on Wikipedia Content Contributions

	Mean	SD	Min	P25	P50	P75	Max	Count
Panel A: Edits in knowledge domain *"Scientists"*								
Edits								
Total Anonymous Edits	9.7	32.9	0	0	2	6	776	4576
Edits Adding Content	7.4	25.6	0	0	1	5	611	4576
Edits Deleting Content	1.8	7.2	0	0	0	1	247	4576
Edits Shorter than a Word	2.8	9.1	0	0	0	2	192	4576
Edits Longer than a Word	6.9	24.6	0	0	1	4	588	4576
Edits Subsequently Reverted	1.0	3.1	0	0	0	1	53	4576
Total Edits in English	1868.7	5395.9	0	78	147	582	28924	4573
Users								
Total Number of Users	5.7	16.8	0	0	1	5	445	4342
Content								
Total Content Modified (B)	1885.1	14807.1	0	0	51	430	775536	4576
Total Content Added (B)	1346.9	8916.7	0	0	42	329	387318	4576
Total Content Deleted (B)	538.2	7643.7	0	0	0	15	388218	4576
Content Subsequently Reverted (B)	1386.0	21597.7	0	0	0	18	1370590	4576
Panel B: Edits in knowledge domain *"Cuisine"*								
Edits								
Total Anonymous Edits	2.0	6.3	0	0	0	1	107	4576
Edits Adding Content	1.4	4.7	0	0	0	1	83	4576
Edits Deleting Content	0.5	1.6	0	0	0	0	41	4576
Edits Shorter than a Word	0.7	2.0	0	0	0	0	28	4576
Edits Longer than a Word	1.3	4.5	0	0	0	1	80	4576
Edits Subsequently Reverted	0.3	0.9	0	0	0	0	13	4576
Total Edits in English	1568.6	4919.4	0	42	92	387	32184	4563
Users								
Total Number of Users	2.4	5.4	0	0	1	2	91	2360
Content								
Total Content Modified (B)	560.9	10439.3	0	0	0	24	495089	4576
Total Content Added (B)	488.0	10312.8	0	0	0	16	495089	4576
Total Content Deleted (B)	73.0	781.5	0	0	0	0	30851	4576
Content Subsequently Reverted (B)	256.6	6242.7	0	0	0	0	378338	4576

Note: The unit of observation is Wikipedia content contributed from a destination country in the native language of an origin country over the year.

To describe the variation over time for the independent and dependent variables in the resulting data set, I plot the yearly median migration flows between pairs of origin and destination countries and content contributions in the two studied domains (normalized with respect to mean and standard deviation of contributions in each domain) in Table 5. Contributions of online content overall seem to follow a downward trend, which is in line with the perceived general decline of Wikipedia in the past five years. At the same time, content contributions increased in the last two years. Immigration flows tend to decrease until 2009-2010, and then grow again.

Figure 5: Median Migration Flows and Anonymous Edits in The Studied Domains: Temporal Variation in the years 2006-2016.



NOTE: This figure illustrates the temporal variation in migration flows to destination countries and the number of edits being made from the destination countries in the languages spoken in immigrants' origin countries in the analyzed domains. The numbers of edits are normalized with respect to mean and standard deviation.

4 Immigration Flows and Wikipedia: Baseline Analysis

To estimate the impact of migration on knowledge diffusion, I exploit the variation between the migrant flows from the set of origin countries to the destination countries and the amount of knowledge generated from the destination countries in the immigrants' native languages. For that, I estimate the following equation:

$$Content_{dot} = \alpha_{do} + \tau_t + \beta \ ImmigrationFlow_{dot} + X_{ot} \ \gamma + X_{dt} \ \delta + \epsilon_{dot}, \tag{1}$$

where d stands for the destination country (for immigration flows) and the location of content contributions on Wikipedia, o indicates the country of immigrants' origin or the language in which content is contributed, t is the year of observation. $Content_{dot}$ is the amount of content about destination countries in immigrants' native languages, i.e. the languages commonly used in their origin countries. The explanatory variable of primary interest is $ImmigrationFlow_{dot}$ measured by the log number of migrants from origin country o arriving to destination country d in year t. I include a vector of year fixed effects, τ_t, to control for unobservable variation in online content generation over time that is common across country pairs. A vector of country pair fixed effects, α_{do}, is included to control for the time-invariant unobserved heterogeneity across country pairs, which might affect the presence of online content about a country in a language. For example, the variation in Wikipedia content available about a country in a certain language might come from the fact that there are strong cultural ties and historical immigration between the two countries. The main coefficient of interest is β. Since the dependent variables are transformed into natural logarithms due to variable skewness, β measures the elasticity of content generation to migration. Following the previous evidence on the positive impact of immigration on knowledge spillovers (Kerr and Lincoln (2010), Bahar and Rapoport (2016), Fackler et al. (2018)), I expect β to be positive.

I use additional control variables to account for trends in content contributions in general in the native languages of migrants and from the destination countries. First, I include a measure of the development of content for each entire Wikipedia language edition, X_{ot}. I collecte yearly statistics at the level of Wikipedia languages: total edits per year, new articles per day, new Wikipedians per year (see Table 2 for descriptive statistics). The inclusion of these variables allows to account that some language editions of Wikipedia may grow faster than others, which could affect the increase in the amount of information in a particular domain in the origin language of immigrants just because of more general growth of content in this language. Similarly, to account for growth in the domains describing destination countries, I add as a control variable, X_{dt}, the amount of content generated about the destination country, d, in English. This helps to tease out the variation in the content from the destination country, which is common for all languages of contributions. For example, Wikipedia may be a standard source of information in certain destination countries. Including total editing activity in English from the destination country would take this variation into account.

In the baseline specifications, migration is measured by the contemporaneous migration flows, meaning that due to the change in immigration the changes in content contributions are expected to be observed within the same year. Some studies on the impact of immigration on patenting (for example, Hunt and Gauthier-Loiselle (2010)) lead the dependent variable by one year in order to allow for research time between the immigrant arrival and the patent application. However, compared to patenting, Wikipedia has reasonably low barriers to contribution: an individual should have access to Internet, the interface is very simple and intuitive and knowledge needs to be very general, as a contribution can be a simple typo correction. Moreover, since my analysis focuses on contributions in the immigrants' native languages, I expect contributions to take place in a relatively short time after the arrival to destination countries.

Identifying Assumption. This analysis relies on the the identifying assumptions that (i) for the pairs of origin and destination countries the pace of content contributions changes due to the change in migration flows conditionally on the other observables, and that (ii) no other unobservable confounders drive both emigration decisions and the pace of content generation on Wikipedia.

The baseline specification addresses the identification challenge in several ways. First, the regressions include country pair fixed effects thus eliminating all time-invariant factors that could be correlated with stronger links between pairs of countries in terms of both immigration flows and knowledge flows, for example, distances between countries or cultural / language proximity. The stock of migrants in the destination countries as well as countries' population change slowly over the years, therefore, the fixed effects capture these influences to a great extent. Second, I include country-specific time trends in content generation for origin and destination countries. This is useful, for example, if content in a language of an origin country was growing because of internal factors in the origin country that make the local Wikipedia language edition a more important information source for the population. This measure provides a better control than the origin country-specific linear time trends as it better captures the variation in the popularity of Wikipedia language editions and does not blow up the degrees of freedom in the estimation. In addition, the content produced within destination countries may change due to destination country-specific trends in content consumption and production. Therefore, I include a measure of content production from destination country in English language as a proxy for popularity of Wikipedia in the destination country.

Another challenge to identification is a potential reverse causality, which may arise if content in specific domains on Wikipedia attracts immigrants to the destination countries described in those domains. I deal with this challenge by approaching this analysis as a natural experiment setting. Recent literature on migration increasingly uses shocks to immigration to identify causal effects of interest (Borjas and Doran (2015), Ganguli (2015), Anelli and Peri (2017), and Barsbai et al. (2017)). For example, Borjas and Doran (2015) and Ganguli (2015) exploit the collapse of Soviet Union in 1991 as a shock in a natural experiment to estimate how the sudden influx of high-skilled Soviet immigrants affected scientific publishing in the US. Anelli and Peri (2017) suggest that similar shocks to immigration can occur due to macroeconomic conditions: following the economic recession in 2008-2009, emigration from Italy experienced a strong shock in the beginning of 2010. Barsbai et al. (2017) show that after the economic crisis in Russia in 1998, Moldova was affected to such magnitude that emigration rose dramatically both to the East and to the West. I follow these studies and also rely on the finding that economic and political crises trigger emigration from the countries.

To estimate the causal effect of interest, the impact of immigration on knowledge diffusion one could compare the country pairs that are affected by shocks (the recent economic and political crises) with those less affected, and estimate the following equation:

$$Content_{dot} = \alpha_{do} + \tau_t + \beta_1 \; T_{do} + \beta_2 \; T_{do} \times After_{ot} + X_{ot} \; \gamma + X_{dt} \; \delta + \epsilon_{dot}, \tag{2}$$

where T_{do} indicates that a pair of countries o and d is affected by the shock, and $T_{do} \times After_{ot}$ is the interaction term between T_{do} and $After_{ot}$, indicating the observation period after the shock. The parameter of interest, β_2 could then be estimated as triple difference where differences would be taken over time, across origin countries and across destination countries. For estimating Equation 2, I would need to define country pairs affected by the shocks to immigration in an arbitrary way. Further, using dummy variables instead of the measure of migration flows would yield loss in precision.

In fact, because the observed period in my data set contains shocks to immigration flows, including economic crises in the European countries and political crises (mass protests with subsequent mass immigration in the countries of the former Soviet Union), following Angrist and Pischke (2014), the fixed effects estimation could be interpreted causally as difference-in-differences under an assumption of randomization into treatment.[9] However, provided that in the context of migration achieving randomization is not possible, the endogeneity concern remains. The immigrants pushed from origin countries due to the crises may choose their destination countries endogenously, under the impact of the same factors that may be driving online content contribution. Therefore, I apply an instrumental variable approach in Section 6.

5 Results

The estimation results for Equation 1 are presented in Table 4. The independent variable of interest is the number of migrants moving from an origin country to a destination country. For both knowledge domains, the estimates show the impact of immigration on the following measures of online content contributions by anonymous Wikipedia users: the total number of edits (Col. 1), the number of edits that add (Col. 2) and delete (Col. 3) content, the number of short edits (Col. 4) that modify less than a word, the number of longer edits (Col. 5), and, finally, the number of unproductive edits that were subsequently reverted by the community members (Col. 6). Since all dependent and independent

[9] The fixed effects model estimated in Equation (1) could be seen as a triple-differences model, where time-varying characteristics of content in the languages of origin countries and in English for destination countries are included, time-invariant country-pair characteristics (for example, distance and similar culture) are included in fixed effects, and the independent variable of interest, $ImmigrationFlow$, contains shocks to immigration arising due to economic and political crises (for example, see Figure A3 in Appendix).

variables are skewed, they are transformed into natural logarithms.[10]

Panels A and B of Table 4 present the results for both studied domains, “Scientists” and “Cuisine”. The results suggest that in both domains, immigration increases content contribution measured by total edits in the native languages of migrants from the IP-addresses within the destination countries. Col. (2) - (5) suggest that the increase in edits is driven by content additions rather than by content deletions and by longer rather than by shorter edits. The effects are robust to the inclusion of control variables.

The magnitude of the effect of immigration on knowledge diffusion is not very high: an increase in a migration frow from the origin country to the destination country by 1% yields 0.09% more total edits about scientists and 0.05% edits about cuisine. This effect can be interpreted as follows: a standard deviation increase in migration can bring 5 edits about scientists and about 1 edit about cuisine contributed by migrants from the destination country in the language spoken in their origin country. While this magnitude is not high, it is generally consistent with the observed share of individuals who contribute content among all consumers of content on Wikipedia.

The evidence that the content modifications made by migrants are adding longer pieces of content is consistent with the perception in the Wikipedia community that contributors who skip registration and make modifications anonymously are less likely to care about their reputation in the Wikipedia community. Anthony et al. (2009) describe the two types of anonymous contributors to Wikipedia. The first type are users who see small mistakes in the text and fix them. The second type are experts in their particular fields who come across articles related to their field of expertise and contribute content. As they do not focus on building their reputation within the Wikipedia community, they skip the registration procedure and contribute directly. They are also unlikely to make many contributions to the community. The results point that migration channels this second type of contributors to the online community.

In addition to the “productive”, or “valuable”, content contributions, the “unproductive” activity increases as well. The coefficient in Col. (6) is positive and statistically significant only for the more specialized domain, “Scientists”. However, the magnitude of this increase is lower than for the total editing activity. The evidence on an increase in reverted edits is consistent with the anecdotal observation that new contributors on Wikipedia are often unfamiliar with the code of conduct and, thus, their contributions are more likely to be reverted by more experienced members of the community.

Once we learnt that immigration yields knowledge production, the subsequent question would be whether the effort of contributing yields a change in the growth rate of knowledge. The effect of immigration on size of content contributed is estimated with panel data fixed effects regressions, where the dependent variables are the edit distances of the content added, deleted and reverted. As Table 5 reveals, while immigration brings more edits to Wikipedia content, the increase in the rate of content generation is small and marginally significant. However, the few marginally significant effects seem to be driven by

[10] In order to preserve in my sample observations with zero contributions I computing the natural logarithm of the value adding one, which is the minimum non-zero value for every variable.

the outliers. If the dependent variables are censored at the 90th percentile, the significance vanishes.

Table 4: Immigration and Anonymous Content Contributions: Immigrants Located in Destination Countries Contribute Content to Wikipedia in Their Native Languages

	Total Edits (1)	Edits (Add.) (2)	Edits (Del.) (3)	Edits (Short) (4)	Edits (Long) (5)	Edits (Rev.) (6)
			Panel A: Knowledge Domain "Scientists"			
Immigrants	0.096***	0.092***	0.035**	0.048***	0.084***	0.045***
	(0.029)	(0.027)	(0.017)	(0.018)	(0.029)	(0.015)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	7.585	5.547	1.547	2.475	5.110	1.164
Observations	4576	4576	4576	4576	4576	4576
Country Pairs	609	609	609	609	609	609
R^2 within	0.187	0.210	0.019	0.034	0.217	0.020
			With Controls			
Immigrants	0.089***	0.083***	0.038**	0.049***	0.077***	0.044***
	(0.028)	(0.026)	(0.016)	(0.018)	(0.028)	(0.015)
Edits in Native	0.230***	0.206***	0.096***	0.114***	0.209***	0.087***
Language	(0.041)	(0.041)	(0.029)	(0.030)	(0.041)	(0.023)
Edits in English	-0.060	-0.074**	0.038*	0.039	-0.082**	0.015
	(0.037)	(0.036)	(0.022)	(0.029)	(0.035)	(0.021)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	7.585	5.547	1.547	2.475	5.110	1.164
Observations	4527	4527	4527	4527	4527	4527
Country Pairs	609	609	609	609	609	609
R^2 within	0.194	0.215	0.024	0.039	0.223	0.025
			Panel B: Knowledge Domain "Cuisine"			
Immigrants	0.049***	0.039**	0.028***	0.033***	0.040***	0.007
	(0.018)	(0.017)	(0.010)	(0.012)	(0.015)	(0.008)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	2.606	1.780	0.651	0.936	1.670	0.621
Observations	4576	4576	4576	4576	4576	4576
Country Pairs	609	609	609	609	609	609
R^2 within	0.029	0.042	0.007	0.008	0.045	0.006
			With Controls			
Immigrants	0.049***	0.041**	0.026***	0.032***	0.042***	0.007
	(0.018)	(0.017)	(0.010)	(0.012)	(0.015)	(0.008)
Edits in Native	0.143***	0.121***	0.060***	0.071***	0.132***	0.055***
Language	(0.037)	(0.034)	(0.018)	(0.024)	(0.034)	(0.017)
Edits in English	0.088***	0.084***	0.032*	0.041**	0.079***	0.023*
	(0.024)	(0.023)	(0.017)	(0.018)	(0.023)	(0.013)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	2.606	1.780	0.651	0.936	1.670	0.621
Observations	4517	4517	4517	4517	4517	4517
Country Pairs	609	609	609	609	609	609
R^2 within	0.041	0.053	0.011	0.013	0.057	0.010

Notes: Panel A presents the results for the knowledge domain "Scientists", and panel B for the knowledge domain "Cuisine". Each column contains linear panel data estimates for the dependent variables: (1) total edits, (2) edits adding content, (3) edits deleting content, (4) edits making short changes up to one word, (5) edits making changes longer than one word, and (6) edits that were contributed and subsequently removed by the community. All dependent and independent variables are transformed in logarithms, therefore, the coefficients can be interpreted as elasticities. Robust standard errors (clustered at the origin - destination country pair level) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

Table 5: Immigration and Anonymous Content Contributions: The Amount of Content

	Tot. Net Distance	Tot. Distance (Add.)	Tot. Distance (Del.)	Tot. Distance (Rev.)
	(1)	(2)	(3)	(4)
	Panel A: Knowledge Domain "Scientists" with Controls			
Immigrants	0.112	0.124*	0.068	0.164**
	(0.075)	(0.074)	(0.057)	(0.073)
Edits in Native	0.446***	0.423***	0.253***	0.292***
Language	(0.101)	(0.100)	(0.077)	(0.099)
Edits in English	-0.194*	-0.216**	0.079	0.099
	(0.101)	(0.100)	(0.077)	(0.099)
Country Pair FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Mean dep. v.	1819.898	1225.258	594.640	1722.375
Observations	4527	4527	4527	4527
Country Pairs	609	609	609	609
R^2 within	0.105	0.117	0.018	0.019
	Panel B: Knowledge Domain "Cuisine" with Controls			
Immigrants	0.076	0.061	0.074*	0.041
	(0.059)	(0.057)	(0.040)	(0.049)
Edits in Native	0.328***	0.311***	0.147***	0.190***
Language	(0.081)	(0.077)	(0.055)	(0.066)
Edits in English	0.295***	0.294***	0.092	0.101
	(0.083)	(0.080)	(0.057)	(0.068)
Country Pair FE	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes
Mean dep. v.	507.021	356.926	150.096	364.565
Observations	4517	4517	4517	4517
Country Pairs	609	609	609	609
R^2 within	0.032	0.038	0.008	0.007

Notes: Panels A and B present results for the knowledge domains "Scientists" and "Cuisine", correspondingly. Each column presents linear panel data estimates for the dependent variables: (1) total absolute edit distance, (2) total absolute edit distance of added content, (3) total absolute edit distance of deleted content, and (4) total absolute edit distance of subsequently reverted content. All dependent and independent variables are transformed in logarithms, therefore, the coefficients represent the elasticities. Robust standard errors (clustered at the origin - destination country pair level) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

5.1 Robustness Check

This Section presents the robustness test of the results with the alternative econometric specifications as well as with the alternative explanatory and dependent variables. The first robustness check estimates the main specifications in the count data model. The distribution of the dependent variables measuring the number of edits is highly skewed, and edits per year arriving to Wikipedia articles can be considered as the outcome of an underlying count process. Therefore, I estimate the baseline model in Equation 1 using a negative binomial specification for panel data with the origin-destination country-pair fixed effects. This model addresses the overdispersion of the dependent variables.[11]

The results in Table **??** provide support to the baseline findings. The independent variable of interest, the yearly size of the migrant flow between each pair of countries, is measured in thousands of persons. The resulting effects are comparable to the panel data linear estimates. The coefficient on the number of total edits contributed by anonymous users (Col. 1) suggests that an increase in the number of migrants by one thousand persons is related to a 0.6% increase in mean total edits about scientists and 0.9% about cuisine. This is equivalent to an increase by about 1 edit in each domain with an increase in the explanatory variable of interest by one standard deviation. The coefficients also suggest that online users make shorter edits to the popular domain "Cuisine", while both shorter and longer edits to the more specialized domain "Scientists".

The empirical results strongly suggest an increase in content contributed by migrants from the destination countries. This calls for the related question regarding the content immigrants contribute: are contributions related mainly to knowledge about destination countries? One could expect that migrants gain new location-specific knowledge as they arrive to destination countries. Table 6 presents the results for contributions of immigrants located in destination countries *about destination countries.* The results suggest that, especially, with migration the number of edits deleting content about destination countries grows. Though, these are "productive" content deletions, as there is no increase in subsequently reverted edits observed. Overall, these results suggest that immigrants' knowledge contributions represent an important channel of general knowledge diffusion between the countries.[12]

Finally, we can shed light on the mechanism by which migration channels online content contributions. Table 7 presents the heteregeneity analysis for the effect of the total migration flows and flows of migrants with higher education on online content contributions. Consistent with the expectation, content contributions increase stronger with the increase in the educated migrant flows is stronger in magnitude: a 7% increase in the number of anonymous users writing to Wikipedia about scientists and a 9% increase in the number of users writing about cuisine. Overall, Table 7 confirms that the effect of immigration on knowledge diffusion is driven by educated immigrants. As mentioned in Data Section, the data on

[11] Following the test based on the auxiliary regression, suggested by Cameron and Trivedi (2013).

[12] Table A3 (in the Appendix) similarly studies on contributions of migrants about their native countries and shows no evidence that migrants contribute about their native countries, which could suggest home sickness motives.

the education level of immigrants are available only for two years of my sample, 2006 and 2011, and for fewer country pairs, therefore, the regressions in Col. (2) have substantially fewer observations and do not include the country pair fixed effects.

As expected, the effect on knowledge contributions in the domain that requires stronger educational background from the contributor is higher when we consider migrants with tertiary education, as it is true for the other domain too. It is important to note, that each IP address can map to multiple contributors, therefore, the shown effect of immigration on new content contributors represents the lower bound of the actual increase in the number of new users.

Table 6: Immigration and Anonymous Content Contributions: Content **about** Destination Countries

	Total Edits	Edits (Add.)	Edits (Del.)	Edits (Short)	Edits (Long)	Edits (Rev.)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Knowledge Domain ”Scientists”						
Immigrants, k	1.004*	1.004*	1.005**	1.004	1.004**	1.003
	(0.002)	(0.002)	(0.002)	(0.003)	(0.001)	(0.007)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	6.650	4.926	1.329	2.083	4.567	0.559
Observations	2234	2213	1788	1972	2195	1502
Country Pairs	302	298	217	246	295	174
With Controls						
Immigrants, k	1.004*	1.004***	1.006***	1.005*	1.004***	1.003
	(0.002)	(0.001)	(0.002)	(0.003)	(0.002)	(0.006)
Edits in Native	1.083***	1.082**	1.138***	1.102***	1.099***	1.215***
Language, mln.	(0.024)	(0.036)	(0.041)	(0.031)	(0.027)	(0.048)
Edits in English, k	1.011	1.013	0.996	1.000	1.018**	1.005
	(0.008)	(0.011)	(0.012)	(0.011)	(0.008)	(0.011)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	6.650	4.926	1.329	2.083	4.567	0.559
Observations	2218	2197	1762	1955	2170	1482
Country Pairs	302	298	215	245	294	173
Panel B: Knowledge Domain ”Cuisine”						
Immigrants, k	1.007	1.005	1.034**	1.019**	1.005	1.010
	(0.006)	(0.006)	(0.014)	(0.008)	(0.006)	(0.011)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	3.010	2.117	0.714	1.065	1.946	0.622
Observations	1685	1656	1345	1513	1615	1172
Country Pairs	210	204	156	178	199	133
With Controls						
Immigrants, k	1.007	1.006	1.038***	1.019*	1.005	1.012
	(0.006)	(0.005)	(0.012)	(0.011)	(0.007)	(0.013)
Edits in Native	1.108**	1.102**	1.175***	1.140***	1.138***	1.269***
Language, mln.	(0.048)	(0.045)	(0.065)	(0.040)	(0.042)	(0.100)
Edits in English, k	0.995	0.986	1.023	1.014	0.991	1.042*
	(0.020)	(0.028)	(0.023)	(0.019)	(0.022)	(0.025)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	3.010	2.117	0.714	1.065	1.946	0.622
Observations	1499	1470	1175	1341	1438	996
Country Pairs	190	184	138	160	180	114

Notes: Panels A and B present results for the knowledge domains “Scientists” and “Cuisine”, correspondingly. Each column presents negative binomial estimates for the dependent variables: (1) total edits, (2) edits adding content, (3) edits deleting content, (4) edits making short changes up to one word, (5) edits making changes longer than one word, and (6) edits that were subsequently removed. All regression coefficients are reported as incidence rate ratios. Robust standard errors (bootstrapped) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

Table 7: Immigration and New Contributors on Wikipedia: Education Level of Immigrants

	Total # Users (1)	Total # Users (2)
	Panel A: Knowledge Domain "Scientists"	
Immigrants, k	1.037*	
	(0.021)	
Educated Immigrants, k		1.139***
		(0.053)
Edits in Native Language, mln.	1.513*** (0.051)	1.492*** (0.060)
Edits in English, k	1.100***	1.096***
	(0.012)	(0.011)
alpha	1.948***	1.924***
	(0.215)	(0.227)
Year FE	Yes	Yes
Mean dep. v.	6.579	6.579
Observations	450	450
	Panel B: Knowledge Domain "Cuisine"	
Immigrants, k	1.015	
	(0.010)	
Educated Immigrants, k		1.047**
		(0.023)
Edits in Native Language, mln.	1.282*** (0.033)	1.276*** (0.028)
Edits in English, k	1.104***	1.100***
	(0.014)	(0.013)
alpha	1.091	1.092
	(0.168)	(0.175)
Year FE	Yes	Yes
Mean dep. v.	3.379	3.379
Observations	284	284

Notes: Panels A and B present results for the knowledge domains "Scientists" and "Cuisine", correspondingly. Each column presents negative binomial estimates for the dependent variable - total number of new users (unique IP-addresses, from which users contributed from the destination countries in the corresponding foreign languages) in the domain. The independent variables of interest are the total inflow of immigrants (Col. (1) and (2)) and the inflow of educated immigrants (Col. (3)). All regression coefficients are reported as incidence rate ratios. Robust standard errors (clustered at the origin - destination country pair level) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

6 Instrumental Variables Approach

Baseline fixed effects estimates are biased if some unobserved time-variant confounders drive both migrant choices of the destination countries and content generation in the foreign languages from the destination countries. One example of such confounding factor can be the popularity growth of a certain destination country in the origin country, this may cause an upward bias to the estimates. Alternatively, if a certain destination country becomes more prominent technologically and, at the same time, rises the migration barriers, this would drive the estimates down.

To address endogeneity, I follow an instrumental variable approach widely employed in the labour literature (Altonji and Card (1991), Card (2001)). According to it, the current location choices of migrants are instrumented for with the past location choices of the diaspora. It is based on the notion that the newly arriving migrants tend to settle in the enclaves established by earlier migrants from the same origin countries (Beine et al. (2009)). This means that if there is an exogenous shock to migration in the source countries, some immigrants would choose the destination country based on the location of the diaspora, and this would constitute an arguably exogenous measure of immigration with respect to the outcome of interest. This instrument is an imputed share of migrants which nets out the component of the migration flows that are attributed to economic opportunities. The advantage of imputed shares is that they are determined only by the initial migration mix by origin and by variation in flows across origin groups. Given the importance of ethnic networks, migrants tend to settle in established communities of similar origin.

Following this approach, I construct a measure of plausibly exogenous shocks to the "supply" of immigrants in the destination countries, known in the labour literature as labour "supply-push" factors. The idea behind this instrument is that potential immigrants, when making a decision regarding the destination countries, tend to choose countries that already became home to their former compatriots with a greater probability. The underlying mechanism could be, for example, the past migration from the country, due to which the opportunities for the next generation of migrants improve or information about career opportunities and infrastructure for assimilation could be disseminated. Therefore, when origin countries experience economic and political shocks that lead to stronger immigration outflows, immigrants could be more likely to choose their new destination, to some extent, based on the past geographical distribution of the diaspora. This is the exclusion restriction of the supply-push approach.[13]

I compute the share of total emigration flow from each origin country arriving to the destination countries based on the *past* share of the overall diaspora from this origin country residing in each destination country in years preceding the growing popularity of Wikipedia and the beginning of my observation sample.

[13] A similar approach has been widely applied in the literature on local labor market conditions, where the labour supply is instrumented by local structure and global shocks (Bartik (1991)).

The remaining concern is that the shares of diaspora from an origin country in a destination country are correlated with knowledge contributions. I alleviate it by using only historical shares of female stock for each country, instead of shares of the total stock of earlier migrants. This is because, similarly, to other traditional and new media, Wikipedia suffers from the gender gap in contributions. Numerous studies and reports suggest that only a small fraction of Wikipedia's contributor base are female. For example, Lam et al. (2011) found that females comprised only 16.1% of the 38,497 editors who started editing Wikipedia during 2009, and they only accounted for 9.0% of edits made by the cohort. Moreover, only 6% of editors who made more than 500 edits were female, with the average male editor having twice as many edits. According to New York Times, in 2011 only 15% of editors are female.[14] As Wikimedia Foundation reveals, in 2018 the share of female editors remains low.[15] Hinnosaar (2017) uses data from a survey and a randomized survey experiment and finds that men and women contribute to different articles, which implies that the gender gap leads to unequal coverage of topics on Wikipedia. This gender gap is caused by the differences in the frequency of Wikipedia use and in beliefs about one's competence and results in profound inequalities in the male and female topic coverage. Therefore, for the construction of the instrumental variable I use the data on the stock of female migrants in 2002, which is the year before Wikipedia became a little known.

The measure of the supply-push component of yearly immigration flows from origin country o to destination country d in year t is constructed as follows:

$$ImmigrationPush_{dot} = ShareFemaleStock_{do,2002}\ ImmigrationFlow_{ot}, \tag{3}$$

where $ShareFemaleStock_{do,2002}$ is the share of the stock of female migrants from a country of origin o in a country of destination d over the total stock of migrants from a country of origin o in the year 2002. Using this measure, at the first stage I estimate the equation:

$$ImmigrationFlow_{dot} = \alpha_o + \alpha_d + \tau_t + \nu\ ImmigrationPush_{dot} + X_{ot}\ \gamma + X_{dt}\ \delta + \epsilon_{dot}. \tag{4}$$

The strongly significant coefficient of 0.204 and the Kleibergen-Paap F-statistic (36) on the excluded instrument (Table 8) confirm that the supply-push component of the current immigrant flow is a strong predictor for the origin-destination immigrant flows.

[14] https://www.nytimes.com/2011/01/31/business/media/31link.html?_r=1

[15] https://commons.wikimedia.org/wiki/File:LE15_Gender_overall_in_2018.png

Table 8: Immigration and Anonymous Content Contributions: Instrumental Variables Estimation

	First Stage	Reduced Form					
	Immigrants	Total Edits	Edits (Add.)	Edits (Del.)	Edits (Rev.)	Tot. Dist. (Add.)	Tot. Dist. (Del.)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: Knowledge Domain "Scientists"							
ImmigrationPush	0.203***	0.068***	0.061***	0.036***	0.031***	0.128***	0.101***
	(0.034)	(0.018)	(0.017)	(0.012)	(0.010)	(0.036)	(0.031)
Edits in Native Language	0.018	0.234***	0.207***	0.099***	0.084***	0.464***	0.276**
	(0.054)	(0.044)	(0.042)	(0.029)	(0.023)	(0.121)	(0.116)
Edits in English	-0.070**	-0.056	-0.072*	0.042*	0.018	-0.202*	0.109
	(0.034)	(0.039)	(0.037)	(0.022)	(0.023)	(0.114)	(0.080)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4485	4485	4485	4485	4485	4485	4485
R^2	0.089	0.113	0.128	0.006	0.005	0.080	0.001
Panel B: Knowledge Domain "Cuisine"							
ImmigrationPush	0.202***	0.033***	0.028***	0.016**	0.015**	0.084***	0.060***
	(0.034)	(0.012)	(0.010)	(0.006)	(0.006)	(0.028)	(0.021)
Edits in Native Language	0.019	0.140***	0.120***	0.059***	0.055***	0.352***	0.177***
	(0.054)	(0.036)	(0.032)	(0.018)	(0.016)	(0.098)	(0.063)
Edits in English	-0.015	0.082***	0.078***	0.029*	0.022*	0.275***	0.106
	(0.053)	(0.026)	(0.023)	(0.017)	(0.013)	(0.074)	(0.077)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4475	4475	4475	4475	4475	4475	4475
R^2	0.088	0.018	0.026	-0.006	-0.005	0.019	-0.008

Notes: Panels A and B present the results for the knowledge domains "Scientists" and "Cuisine" using the supply-push instrument based on the shares of stock for female migrants. Col. (1) presents the first stage estimates for the inflow of migrants, and the remaining columns present the reduced form estimates for the dependent variables of interest: (2) total edits, (3) edits adding content, (4) edits deleting content, (5) edits making short changes up to one word, (6) edits making changes longer than one word, and (7) edits that were contributed and subsequently removed by the community. Robust standard errors (clustered at the origin - destination country pair level) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

At the second stage, I estimate the model analogous to 1, substituting the value of the immigration flow with the fitted values from the first stage. Table 9 presents the results of the second stage. The local average treatment effect of IV regressions estimates the increase in content generated about destination countries in the native languages of immigrants as 0.2-0.4% more content due to a 1% increase in the arrival of migrants from country o to country d.

Table 9: Immigration and Anonymous Content Contributions: Instrumental Variables Estimation

					Tot. Dist.	Tot. Dist.
	Total Edits	Edits (Add.)	Edits (Del.)	Edits (Rev.)	(Add.)	(Del.)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Knowledge Domain ”Scientists”						
Immigrants	0.336***	0.302***	0.179***	0.155***	0.632***	0.499***
	(0.089)	(0.084)	(0.062)	(0.047)	(0.178)	(0.152)
Edits in Native	0.228***	0.201***	0.096***	0.081***	0.453***	0.267**
Language	(0.042)	(0.041)	(0.029)	(0.023)	(0.116)	(0.115)
Edits in English	-0.033	-0.051	0.055**	0.028	-0.158	0.144*
	(0.037)	(0.036)	(0.022)	(0.022)	(0.112)	(0.079)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	7.585	5.547	1.547	1.164	1225.258	594.640
Observations	4485	4485	4485	4485	4485	4485
R^2	0.164	0.169	0.055	0.054	0.116	0.048
F-statistic	35.379	35.379	35.379	35.379	35.379	35.379
Anderson-Rubin P-val.	0.000	0.000	0.003	0.002	0.000	0.001
Panel B: Knowledge Domain ”Cuisine”						
Immigrants	0.165***	0.137***	0.079**	0.072**	0.414***	0.296***
	(0.056)	(0.051)	(0.032)	(0.029)	(0.138)	(0.106)
Edits in Native	0.137***	0.117***	0.058***	0.054***	0.344***	0.171***
Language	(0.035)	(0.032)	(0.018)	(0.016)	(0.097)	(0.063)
Edits in English	0.085***	0.080***	0.030*	0.023*	0.281***	0.111
	(0.025)	(0.023)	(0.017)	(0.012)	(0.073)	(0.076)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	2.606	1.780	0.651	0.621	356.926	150.096
Observations	4475	4475	4475	4475	4475	4475
R^2	0.051	0.051	0.012	0.006	0.041	0.006
F-statistic	35.262	35.262	35.262	35.262	35.262	35.262
Anderson-Rubin P-val.	0.003	0.007	0.012	0.012	0.003	0.004

Notes: Panels A and B present the results for the knowledge domains “Scientists” and “Cuisine” using the supply-push instrument with total migrant stock shares. Panels C and D employ the alternative supply-push instrument based on shares of stock for female immigrants. Each column presents the two-stage least squares estimates for the dependent variables: (1) total edits, (2) edits adding content, (3) edits deleting content, (4) edits making short changes up to one word, (5) edits making changes longer than one word, and (6) edits that were contributed and subsequently removed by the community. All dependent and independent variables are transformed in logarithms, therefore, the coefficients represent the elasticities. Robust standard errors (clustered at the origin - destination country pair level) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

7 Conclusion

International migration is on the rise in the past decade, and, according to the United Nations 2017 report, migration has grown faster than the world's population (UN, 2017). Due to an increasing impact of migration and in the in light of the recent policy debates around it, deeper understanding of the implications of migration for the diffusion of knowledge can inform policy makers about potentially beneficial policies.

This paper measures the impact of migration on the diffusion of broader knowledge online using data on migration flows to OECD countries combined with content contributions to Wikipedia. The findings suggest that due to an increase in migration more online content contributions take place from and about the destination countries in the languages spoken in the migrants' origin countries. However, there is only marginal increase in the size of content contributions. The increase in editing is driven by the increase in the number of new contributors on the platform, especially, educated migrants. These findings are confirmed using the instrumental variables approach, which yields higher estimates of the effect.

As migration is unlikely to slow down in the coming years,[16] policy makers in the migrants' origin countries could take advantage of the potential knowledge inputs by their former compatriots in order to increase the knowledge stock available to the countries' population. Currently, online content contributions tend to be centered around the areas with higher socioeconomic status. Knowledge flows across countries mitigate these inequalities, i.e. the digital divide (Chinn and Fairlie, 2007; Graham et al., 2014; Sen et al., 2015). In the light of these findings, migration could be seen as a phenomenon reducing these information asymmetries in the online knowledge representation. This has numerous implications for research cooperation, technology adoption and individual knowledge-related choices and can, therefore, shape the future economic development in the sending countries. For the knowledge domains studied in this paper, one potential channel how the coverage of the topic can impact individual choices is via creating the role models. The stronger is the coverage of biographies of different scientists, the more likely the readers are to get interested in science-related topics and to choose the education path accordingly. The literature confirms that role models do play an important role for the individual choices. (Beaman et al., 2012) show that female leadership influences adolescent girls' career aspirations and educational attainment using the evidence from a randomized natural experiment in India. (Fairlie et al., 2014) find the positive effect of the role model on the college performance of minority students with minority instructors.

While this analysis sheds light on the process of knowledge diffusion due to international migration, some limitations are acknowledged. The first concern is the granularity of the data. It would be highly desirable to analyze data on migration and online knowledge-related activities at the individual level.

[16] https://www.nytimes.com/interactive/2018/06/20/business/economy/immigration-economic-impact.html

However, individual data on migration combined with online activities can hardly be available due to the privacy concern. Furthermore, conducting a similar analysis in the field-experimental framework is very hard to implement, therefore, this study, as well as the most of the literature on migration, relies on observational data, with all the potential drawbacks this choice yields.

Overall, the digital economy opens up massive amounts of relevant data, in particular from platforms relying on user-generated content, for advancing economic research and creating more and better measures of online knowledge-related activities (Glaeser et al., 2018). In the future research, these data could help in addressing further important questions related to immigration, for example, what kind of knowledge do immigrants share online? Sharing of specific kinds of knowledge online could indicate to which extent immigrants assimilate in destination countries, adopting foreign knowledge and culture.

References

Aaltonen, Aleksi and Stephan Seiler, "Cumulative growth in user-generated content production: evidence from Wikipedia," *Management Science*, 2015, *62* (7), 2054–2069.

Agrawal, Ajay, Devesh Kapur, John McHale, and Alexander Oettl, "Brain drain or brain bank? The impact of skilled emigration on poor-country innovation," *Journal of Urban Economics*, 2011, *69* (1), 43–55.

_ , Iain Cockburn, and John McHale, "Gone but not forgotten: knowledge flows, labor mobility, and enduring social relationships," *Journal of Economic Geography*, 2006, *6* (5), 571–591.

_ , Nicola Lacetera, and Elizabeth Lyons, "How Do Online Platforms Flatten Markets for Contract Labor?," Technical Report, Working Paper 2012.

Algan, Yann, Yochai Benkler, Mayo Fuster Morell, and Jérôme Hergueux, "Cooperation in a Peer Production Economy Experimental Evidence from Wikipedia," in "Workshop on Information Systems and Economics, Milan, Italy" 2013, pp. 1–31.

Altonji, Joseph G and David Card, "The effects of immigration on the labor market outcomes of less-skilled natives," in "Immigration, trade, and the labor market," University of Chicago Press, 1991, pp. 201–234.

Anelli, Massimo and Giovanni Peri, "Does emigration delay political change? Evidence from Italy during the great recession," *Economic Policy*, 2017, *32* (91), 551–596.

Angrist, Joshua D and Jörn-Steffen Pischke, *Mastering'metrics: The path from cause to effect*, Princeton University Press, 2014.

Anthony, Denise, Sean W Smith, and Timothy Williamson, "Reputation and reliability in collective goods: The case of the online encyclopedia Wikipedia," *Rationality and Society*, 2009, *21* (3), 283–306.

Asatryan, Zareh, Benjamin Bittschi, and Philipp Doerrenberg, "Remittances and public finances: Evidence from oil-price shocks," *Journal of Public Economics*, 2017, *155*, 122–137.

Bahar, Dany and Hillel Rapoport, "Migration, knowledge diffusion and the comparative advantage of nations," *Economic Journal*, 2016.

Barsbai, Toman, Hillel Rapoport, Andreas Steinmayr, and Christoph Trebesch, "The effect of labor migration on the diffusion of democracy: evidence from a former Soviet Republic," *American Economic Journal: Applied Economics*, 2017, *9* (3), 36–69.

Bartik, Timothy J, "Who benefits from state and local economic development policies?," 1991.

Beaman, Lori, Esther Duflo, Rohini Pande, and Petia Topalova, "Female leadership raises aspirations and educational attainment for girls: A policy experiment in India," *science*, 2012, *335* (6068), 582–586.

Beine, Michel, Frédéric Docquier, and Caglar Ozden, *Diasporas*, The World Bank, 2009.

_ , _ , and Hillel Rapoport, "Brain drain and economic growth: theory and evidence," *Journal of development economics*, 2001, *64* (1), 275–289.

Borjas, George J and Kirk B Doran, "Cognitive mobility: Labor market responses to supply shocks in the space of ideas," *Journal of Labor Economics*, 2015, *33* (S1), S109–S145.

Bosetti, Valentina, Cristina Cattaneo, and Elena Verdolini, "Migration of skilled workers and innovation: A European Perspective," *Journal of International Economics*, 2015, *96* (2), 311–322.

Breschi, Stefano and Francesco Lissoni, "Mobility of skilled workers and co-invention networks: an anatomy of localized knowledge flows," *Journal of Economic Geography*, 2009, *9* (4), 439–468.

_ , _ , and Ernest Miguelez, "Foreign-origin inventors in the USA: testing for diaspora and brain gain effects," *Journal of Economic Geography*, 2017, *17* (5), 1009–1038.

Cameron, A Colin and Pravin K Trivedi, *Regression analysis of count data*, Vol. 53, Cambridge university press, 2013.

Card, David, "Immigrant inflows, native outflows, and the local labor market impacts of higher immigration," *Journal of Labor Economics*, 2001, *19* (1), 22–64.

Chinn, Menzie D and Robert W Fairlie, "The determinants of the global digital divide: a cross-country analysis of computer and internet penetration," *Oxford Economic Papers*, 2007, *59* (1), 16–44.

Docquier, Frédéric, Elisabetta Lodigiani, Hillel Rapoport, and Maurice Schiff, "Emigration and democracy," *Journal of Development Economics*, 2016, *120*, 209–223.

Douglas, Kacey N, "International knowledge flows and technological advance: the role of migration," *IZA Journal of Migration*, 2015, *4* (1), 1.

Fackler, Thomas, Yvonne Giesing, and Nadzeya Laurentsyeva, "Knowledge Remittances: Does Emigration Foster Innovation?," 2018.

Fairlie, Robert W, Florian Hoffmann, and Philip Oreopoulos, "A community college instructor like me: Race and ethnicity interactions in the classroom," *American Economic Review*, 2014, *104* (8), 2567–91.

Foley, C Fritz and William R Kerr, "Ethnic innovation and US multinational firm activity," *Management Science*, 2013, *59* (7), 1529–1544.

Gallus, Jana, "Fostering public good contributions with symbolic awards: A large-scale natural field experiment at wikipedia," *Management Science*, 2016, *63* (12), 3999–4015.

Ganguli, Ina, "Immigration and Ideas: What Did Russian Scientists "Bring" to the United States?," *Journal of Labor Economics*, 2015, *33* (S1 Part 2), S257–S288.

Ghani, Ejaz, William R Kerr, and Christopher Stanton, "Diasporas and outsourcing: evidence from oDesk and India," *Management Science*, 2014, *60* (7), 1677–1697.

Giovanni, Julian Di, Andrei A Levchenko, and Francesc Ortega, "A global view of cross-border migration," *Journal of the European Economic Association*, 2015, *13* (1), 168–202.

Glaeser, Edward L, Scott Duke Kominers, Michael Luca, and Nikhil Naik, "Big data and big cities: The promises and limitations of improved measures of urban life," *Economic Inquiry*, 2018, *56* (1), 114–137.

Gould, David M, "Immigrant links to the home country: empirical implications for US bilateral trade flows," *The Review of Economics and Statistics*, 1994, pp. 302–316.

Graham, Mark, Bernie Hogan, Ralph K Straumann, and Ahmed Medhat, "Uneven geographies of user-generated information: patterns of increasing informational poverty," *Annals of the Association of American Geographers*, 2014, *104* (4), 746–764.

Greenstein, Shane and Feng Zhu, "Is Wikipedia Biased?," *American Economic Review*, 2012, *102* (3), 343–48.

_ and _ , "Do Experts or Crowd-Based Models Produce More Bias? Evidence from Encyclopædia Britannica and Wikipedia," *MIS Quarterly*, 2018, *42* (3), 945–959.

Hinnosaar, Marit, "Gender inequality in new media: Evidence from Wikipedia," *Available at SSRN 2617021*, 2017.

_ , Toomas Hinnosaar, Michael E Kummer, and Olga Slivko, "Wikipedia Matters," 2017.

_ , _ , Michael Kummer, and Olga Slivko, "Externalities in Knowledge Production: Evidence from a Randomized Field Experiment," *arXiv preprint arXiv:1903.01861*, 2019.

Huang, Ni, Gordon Burtch, Bin Gu, Yili Hong, Chen Liang, Kanliang Wang, Dongpu Fu, and Bo Yang, "Motivating user-generated content with performance feedback: Evidence from randomized field experiments," *Management Science*, 2018, *65* (1), 327–345.

Hunt, Jennifer and Marjolaine Gauthier-Loiselle, "How much does immigration boost innovation?," *American Economic Journal: Macroeconomics*, 2010, *2* (2), 31–56.

Javorcik, Beata S, Çağlar Özden, Mariana Spatareanu, and Cristina Neagu, "Migrant networks and foreign direct investment," *Journal of Development Economics*, 2011, *94* (2), 231–241.

Kane, Gerald C and Sam Ransbotham, "Content as community regulator: The recursive relationship between consumption and contribution in open collaboration communities," *Organization Science*, 2016, *27* (5), 1258–1274.

Kapur, Devesh, "Diasporas and technology transfer," *Journal of Human Development*, 2001, *2* (2), 265–286.

Kerr, Sari Pekkala, William Kerr, Çağlar Özden, and Christopher Parsons, "High-skilled migration and agglomeration," *Annual Review of Economics*, 2017, *9*, 201–234.

Kerr, William R, "Ethnic scientific communities and international technology diffusion," *The Review of Economics and Statistics*, 2008, *90* (3), 518–537.

_ **and William F Lincoln**, "The supply side of innovation: H-1B visa reforms and US ethnic invention," *Journal of Labor Economics*, 2010, *28* (3), 473–508.

Kugler, Maurice and Hillel Rapoport, "International labor and capital flows: Complements or substitutes?," *Economics Letters*, 2007, *94* (2), 155–162.

Kummer, Michael, "Spillovers in networks of user generated content: Pseudo-experimental evidence on Wikipedia," 2014.

Kummer, Michael E, "Spillovers in Networks of User Generated Content," *Available at SSRN*, 2013.

_ **, Marianne Saam, Iassen Halatchliyski, and George Giorgidze**, "Centrality and content creation in networks-The case of economic topics on German wikipedia," *Information Economics and Policy*, 2016, *36*, 36–52.

Lam, Shyong Tony K, Anuradha Uduwage, Zhenhua Dong, Shilad Sen, David R Musicant, Loren Terveen, and John Riedl, "WP: clubhouse?: an exploration of Wikipedia's gender imbalance," in "Proceedings of the 7th international symposium on Wikis and open collaboration" ACM 2011, pp. 1–10.

Lerner, Josh and Jean Tirole, "Some simple economics of open source," *The journal of industrial economics*, 2003, *50* (2), 197–234.

Miguelez, Ernest, "Inventor diasporas and the internationalization of technology," *The World Bank Economic Review*, 2016, p. lhw013.

_ **and Claudia Noumedem Temgoua**, “Highly Skilled Migration and Knowledge Diffusion: A Gravity Model Approach.”

Moser, Petra, Alessandra Voena, and Fabian Waldinger, “German Jewish émigrés and US invention,” *American Economic Review*, 2014, *104* (10), 3222–55.

Nagaraj, Abhishek, “Information Seeding and Knowledge Production in Online Communities: Evidence from OpenStreetMap,” 2017.

Nanda, Ramana and Tarun Khanna, “Diasporas and domestic entrepreneurs: Evidence from the Indian software industry,” *Journal of Economics & Management Strategy*, 2010, *19* (4), 991–1012.

Piskorski, Mikołaj Jan and Andreea Gorbatâi, “Testing Coleman’s soc¡ial-norm enforcement mechanism: Evidence from Wikipedia,” *American Journal of Sociology*, 2017, *122* (4), 1183–1222.

Ransbotham, Sam and Gerald C Kane, “Membership turnover and collaboration success in online communities: Explaining rises and falls from grace in Wikipedia,” *Mis Quarterly*, 2011, pp. 613–627.

Ren, Yuqing, Jilin Chen, and John Riedl, “The impact and evolution of group diversity in online open collaboration,” *Management Science*, 2015, *62* (6), 1668–1686.

Sen, Shilad W, Heather Ford, David R Musicant, Mark Graham, OS Keyes, and Brent Hecht, “Barriers to the localness of volunteered geographic information,” in “Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems” ACM 2015, pp. 197–206.

Thompson, Neil and Douglas Hanley, “Science Is Shaped by Wikipedia: Evidence from a Randomized Control Trial,” 2017.

UN, “International migration report 2017,” 2017.

Waldinger, Fabian, “Quality matters: The expulsion of professors and the consequences for PhD student outcomes in Nazi Germany,” *Journal of Political Economy*, 2010, *118* (4), 787–831.

_ , “Bombs, brains, and science: The role of human and physical capital for the creation of scientific knowledge,” *Review of Economics and Statistics*, 2016, *98* (5), 811–831.

Zhang, Xiaoquan Michael and Feng Zhu, “Group size and incentives to contribute: A natural experiment at Chinese Wikipedia,” *American Economic Review*, 2011, *101* (4), 1601–1615.

Zhu, Kai, Dylan Walker, and Lev Muchnik, “Content Growth and Attention Contagion in Information Networks: A Natural Experiment on Wikipedia,” 2018.

A Appendix

(included for convenience of the referees)

A.1 Descriptive Figures and Tables

Figure A1: Top-20 Wikipedia Language Editions Ranked by the Total Number of Articles.

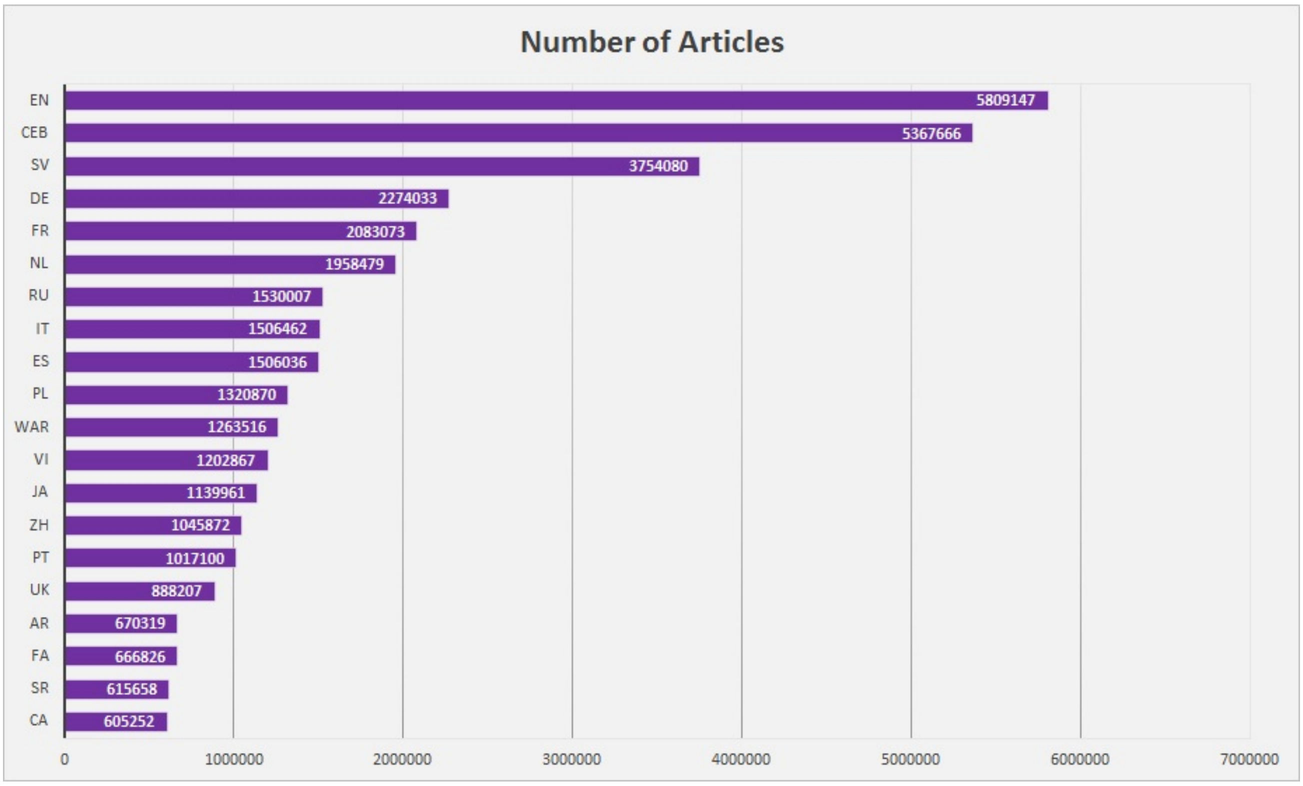
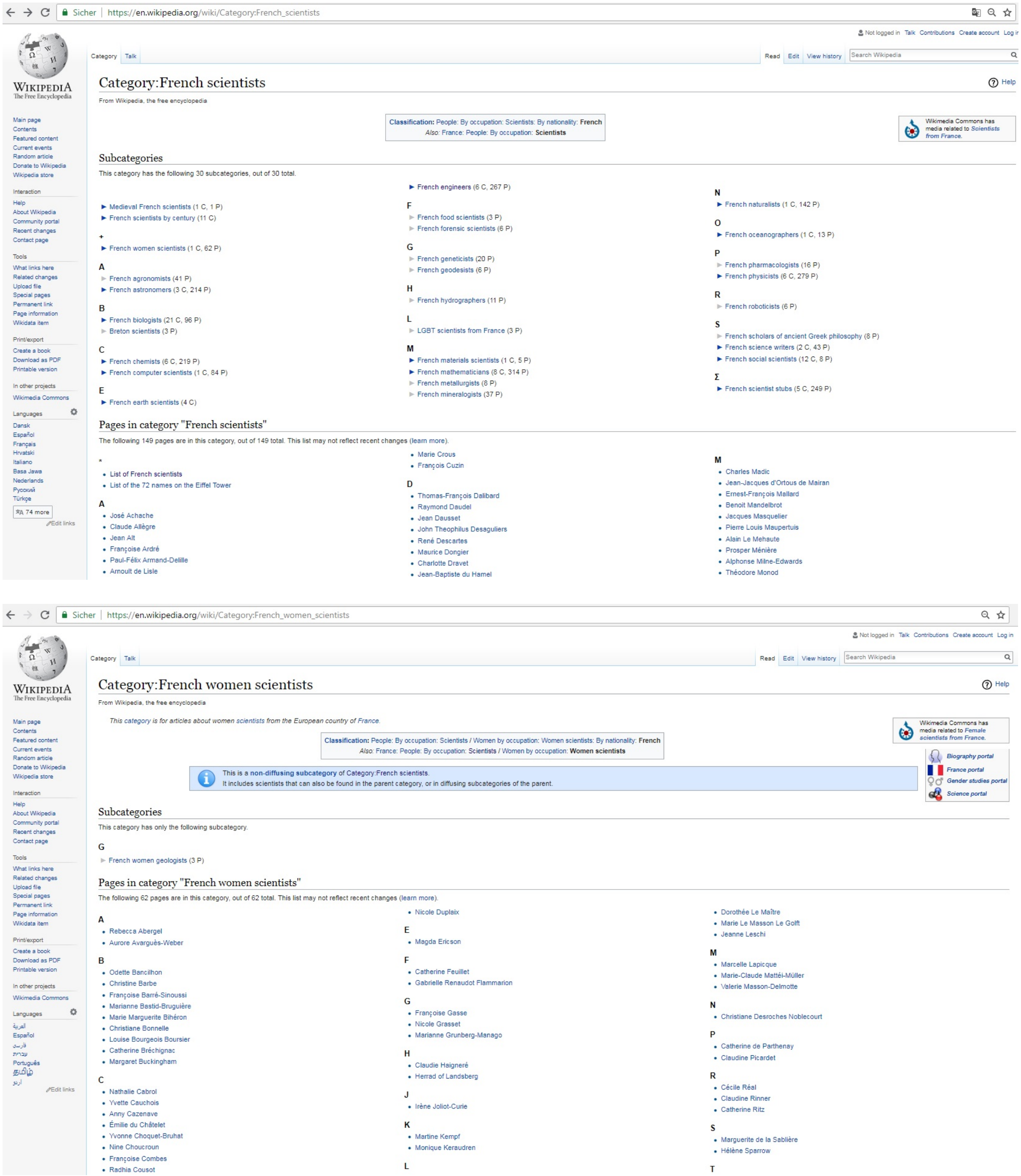


Figure A2: The main category and its subcategory on Wikipedia: “French scientists” in English language

(a) The main category “French scientists” in English (with its subcategories and articles)



Category:French scientists

From Wikipedia, the free encyclopedia

Classification: People: By occupation: Scientists: By nationality: **French**
Also: France: People: By occupation: **Scientists**

Wikimedia Commons has media related to *Scientists from France*.

Subcategories

This category has the following 30 subcategories, out of 30 total.

- ► Medieval French scientists (1 C, 1 P)
- ► French scientists by century (11 C)

+

- ► French women scientists (1 C, 62 P)

A

- ► French agronomists (41 P)
- ► French astronomers (3 C, 214 P)

B

- ► French biologists (21 C, 96 P)
- ► Breton scientists (3 P)

C

- ► French chemists (6 C, 219 P)
- ► French computer scientists (1 C, 84 P)

E

- ► French earth scientists (4 C)
- ► French engineers (6 C, 267 P)

F

- ► French food scientists (3 P)
- ► French forensic scientists (6 P)

G

- ► French geneticists (20 P)
- ► French geodesists (6 P)

H

- ► French hydrographers (11 P)

L

- ► LGBT scientists from France (3 P)

M

- ► French materials scientists (1 C, 5 P)
- ► French mathematicians (8 C, 314 P)
- ► French metallurgists (8 P)
- ► French mineralogists (37 P)

N

- ► French naturalists (1 C, 142 P)

O

- ► French oceanographers (1 C, 13 P)

P

- ► French pharmacologists (16 P)
- ► French physicists (6 C, 279 P)

R

- ► French roboticists (6 P)

S

- ► French scholars of ancient Greek philosophy (8 P)
- ► French science writers (2 C, 43 P)
- ► French social scientists (12 C, 8 P)

Σ

- ► French scientist stubs (5 C, 249 P)

Pages in category "French scientists"

The following 149 pages are in this category, out of 149 total. This list may not reflect recent changes (learn more).

*

- List of French scientists
- List of the 72 names on the Eiffel Tower

A

- José Achache
- Claude Allègre
- Jean Alt
- Françoise Ardré
- Paul-Félix Armand-Delille
- Arnoult de Lisle
- Marie Crous
- François Cuzin

D

- Thomas-François Dalibard
- Raymond Daudel
- Jean Dausset
- John Theophilus Desaguliers
- René Descartes
- Maurice Dongier
- Charlotte Dravet
- Jean-Baptiste du Hamel

M

- Charles Madic
- Jean-Jacques d'Ortous de Mairan
- Ernest-François Mallard
- Benoit Mandelbrot
- Jacques Masquelier
- Pierre Louis Maupertuis
- Alain Le Mehaute
- Prosper Ménière
- Alphonse Milne-Edwards
- Théodore Monod

Category:French women scientists

From Wikipedia, the free encyclopedia

This category is for articles about women scientists from the European country of France.

Classification: People: By occupation: Scientists / Women by occupation: Women scientists: By nationality: **French**
Also: France: People: By occupation: **Scientists** / Women by occupation: **Women scientists**

Wikimedia Commons has media related to *Female scientists from France.*

This is a **non-diffusing subcategory** of Category:French scientists.
It includes scientists that can also be found in the parent category, or in diffusing subcategories of the parent.

Subcategories

This category has only the following subcategory.

G

- ► French women geologists (3 P)

Pages in category "French women scientists"

The following 62 pages are in this category, out of 62 total. This list may not reflect recent changes (learn more).

A

- Rebecca Abergel
- Aurore Avarguès-Weber

B

- Odette Bancilhon
- Christine Barbe
- Françoise Barré-Sinoussi
- Marianne Bastid-Bruguière
- Marie Marguerite Bihéron
- Christiane Bonnelle
- Louise Bourgeois Boursier
- Catherine Bréchignac
- Margaret Buckingham

C

- Nathalie Cabrol
- Yvette Cauchois
- Anny Cazenave
- Émilie du Châtelet
- Yvonne Choquet-Bruhat
- Nine Choucroun
- Françoise Combes
- Radhia Cousot
- Nicole Duplaix

E

- Magda Ericson

F

- Catherine Feuillet
- Gabrielle Renaudot Flammarion

G

- Françoise Gasse
- Nicole Grasset
- Marianne Grunberg-Manago

H

- Claudie Haigneré
- Herrad of Landsberg

J

- Irène Joliot-Curie

K

- Martine Kempf
- Monique Keraudren

L

- Dorothée Le Maître
- Marie Le Masson Le Golft
- Jeanne Leschi

M

- Marcelle Lapicque
- Marie-Claude Mattéi-Müller
- Valerie Masson-Delmotte

N

- Christiane Desroches Noblecourt

P

- Catherine de Parthenay
- Claudine Picardet

R

- Cécile Réal
- Claudine Rinner
- Catherine Ritz

S

- Marguerite de la Sablière
- Hélène Sparrow

T

(b) The subcategory of the main category “French women scientists” (with its subcategories and articles)

Table A1: Countries' most spoken languages by the share of population.

Origin Country	Main Language	Share of Population, %
Albania	Albanian	98.8
Armenia	Armenian	97.9
Azerbaijan	Azerbaijani	92.5
Belarus	Russian/Belarusian	70.2/23.4
Bulgaria	Bulgarian	76.8
Croatia	Croatian	92.6
Czech Republic	Czech	95.4
Denmark	Danish	86.9
Estonia	Estonian	65.5
Finland	Finnish/Swedish	87.9/5.2
France	French	100.0
Georgia	Georgian	87.6
Greece	Greek	99.9
Hungary	Hungarian	99.6
Iceland	Icelandic	93.2
Ireland	English/Irish	99.9/39.8
Israel	Hebrew	60
Italy	Italian	91.3
Japan	Japanese	98.5
Kazakhstan	Kazakh/Russian	74.0/94.4
Korea	Korean	100.0
Latvia	Latvian	56.3
Lithuania	Lithuanian	82.0
Macedonia	Macedonian	66.5
Mongolia	Mongolian	90.0
Netherlands	Dutch	80.9
Norway	Bokmal Norwegian/Nynorsk Norwegian	86
Poland	Polish	98.2
Romania	Romanian	85.4
Russia	Russian	85.7
Serbia	Serbian	88.1
Slovak Republic	Slovak	78.6
Slovenia	Slovenian	91.1
Sweden	Swedish	85.7
Thailand	Thai	90.7
Turkey	Turkish	84.54
Ukraine	Ukrainian	67.5
Viet Nam	Vietnamese	85.7

NOTE: This table contains information on the language spoken in the country by the majority of population including the corresponding shares of population. The main source: https://www.cia.gov/library/publications/the-world-factbook/fields/2098.html. The shares of population not available in the main source are added from Wikipedia.

Table A2: Immigration inflows from the source countries over the years 2006-2016.

	Total Emigrants(K)	Av. Yearly Emigrants(K)	Std. dev.	Maximum
Albania	153.91	6.41	9.92	35.72
Armenia	24.36	2.21	1.53	4.35
Azerbaijan	32.31	0.67	0.84	3.98
Belarus	39.14	1.45	0.99	3.15
Bulgaria	964.85	4.45	12.41	86.27
Croatia	275.55	2.10	7.24	60.98
Czech Republic	188.78	1.03	1.99	10.97
Denmark	118.95	0.78	1.05	5.14
Estonia	62.81	0.61	1.22	6.04
Finland	109.24	0.59	0.71	3.05
France	686.18	2.64	4.24	25.00
Georgia	24.70	1.90	1.15	5.60
Greece	332.61	1.96	5.50	32.66
Hungary	695.78	4.02	10.62	59.99
Iceland	0.07	0.03	0.00	0.04
Italy	1143.49	4.45	8.68	57.19
Japan	295.82	1.43	1.98	8.27
Korea	756.46	4.79	8.06	30.04
Latvia	74.34	2.01	2.82	10.03
Lithuania	211.40	2.55	3.82	17.00
Macedonia	84.46	2.01	3.63	15.63
Netherlands	298.93	1.23	2.22	11.20
Norway	37.25	0.81	0.80	2.49
Poland	3054.08	11.98	32.73	192.17
Romania	3626.40	21.08	45.37	271.44
Russia	741.68	2.58	4.10	31.37
Serbia	331.73	4.05	7.68	39.72
Slovak Republic	233.42	1.70	3.20	15.52
Slovenia	52.11	0.45	0.94	4.75
Sweden	205.35	0.85	1.23	8.20
Thailand	230.45	6.23	5.19	15.45
Turkey	633.25	3.40	6.07	29.59
Ukraine	761.38	4.38	7.59	63.84
Viet Nam	818.34	7.24	13.31	77.54

NOTE: In columns (1)-(4), this table shows main statistic measures of aggregate immigration flows from each source country: the total emigration in the period of observation, average yearly outflow from the origin country per country of emigrants' destination, standard deviation and maximum values. All values are computed only for the sample used in the estimations.

A.2 Robustness check: Fixed Effects Estimation

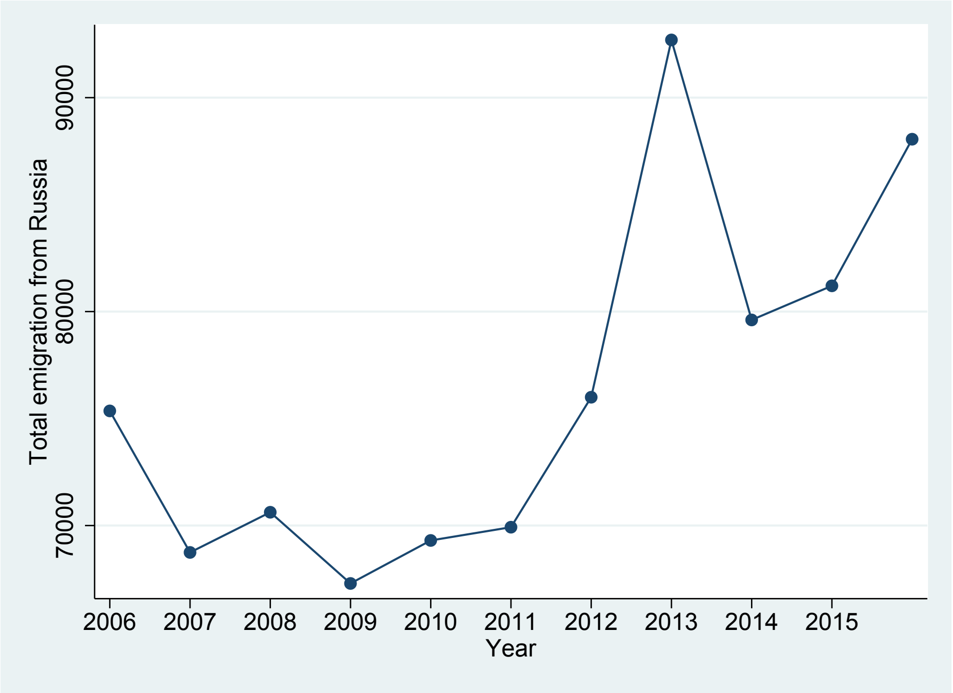
Table A3: Immigration and Anonymous Content Contributions: Immigrants Located in Destination Countries and Contributing Content to Wikipedia in Their Native Languages about their Origin Countries

	Total Edits	Edits (Add.)	Edits (Del.)	Edits (Short)	Edits (Long)	Edits (Rev.)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Knowledge Domain ”Scientists”						
Immigrants	0.070*	0.056	0.040	0.057*	0.052	0.011
	(0.040)	(0.037)	(0.028)	(0.029)	(0.038)	(0.022)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	9.314	6.888	1.896	3.021	6.293	1.412
Observations	2960	2960	2960	2960	2960	2960
Country Pairs	345	345	345	345	345	345
R^2 within	0.072	0.092	0.008	0.014	0.093	0.006
With Controls						
Immigrants	0.072*	0.058	0.040	0.057*	0.055	0.013
	(0.040)	(0.037)	(0.027)	(0.029)	(0.038)	(0.022)
Edits in Native	0.026	-0.017	0.114**	0.031	0.011	-0.013
Language	(0.067)	(0.065)	(0.049)	(0.054)	(0.068)	(0.036)
Edits in English	0.028	0.019	0.033	0.048	0.026	0.011
	(0.039)	(0.038)	(0.029)	(0.033)	(0.036)	(0.022)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	9.314	6.888	1.896	3.021	6.293	1.412
Observations	2949	2949	2949	2949	2949	2949
Country Pairs	345	345	345	345	345	345
R^2 within	0.072	0.091	0.012	0.015	0.092	0.007
Panel B: Knowledge Domain ”Cuisine”						
Immigrants	0.055	0.057	0.052	-0.004	0.063	-0.000
	(0.063)	(0.060)	(0.038)	(0.045)	(0.057)	(0.030)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	4.516	3.168	1.065	1.528	2.988	0.795
Observations	1597	1597	1597	1597	1597	1597
Country Pairs	236	236	236	236	236	236
R^2 within	0.105	0.129	0.019	0.031	0.118	0.013
With Controls						
Immigrants	0.045	0.049	0.043	-0.009	0.054	-0.008
	(0.063)	(0.060)	(0.037)	(0.046)	(0.057)	(0.030)
Edits in Native	0.126	0.049	0.146**	0.077	0.078	0.156***
Language	(0.088)	(0.087)	(0.065)	(0.069)	(0.084)	(0.057)
Edits in English	0.062	0.100**	0.009	0.029	0.081*	-0.017
	(0.052)	(0.050)	(0.033)	(0.040)	(0.045)	(0.037)
Country Pair FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Mean dep. v.	4.516	3.168	1.065	1.528	2.988	0.795
Observations	1592	1592	1592	1592	1592	1592
Country Pairs	236	236	236	236	236	236
R^2 within	0.107	0.131	0.023	0.032	0.119	0.021

Notes: Panels A and B present results for the knowledge contributions in the domains “Scientists” and “Cuisine” where knowledge contributions are made from the destination countries in the languages of migrants’ origin countries and cover knowledge migrants’ origin countries. These results suggest the absence of “home sickness” motives in migrants’ contributions. Each column presents linear panel data estimates for the dependent variables: (1) total edits, (2) edits adding content, (3) edits deleting content, (4) edits making short changes up to one word, (5) edits making changes longer than one word, and (6) edits that were contributed and subsequently removed by the community. All dependent and independent variables are transformed in logarithms, therefore, the coefficients represent the elasticities. Robust standard errors (clustered at the origin - destination country pair level) are reported in parentheses: *** indicates significance at the 1 percent level, ** at 5 percent level, * at 10 percent level.

A.3 Identification

Figure A3: Emigration from Russia after the political crisis in 2012.



Note: This figure illustrates how the flow of total emigration from Russia to OECD countries changed after the political crisis in 2012.