The Thirty-First International Florida
Artificial Intelligence Research Society Conference (FLAIRS-31)

Refugee and Immigration: Twitter as a Proxy for Reality

Firas Aswad, Ronaldo Menezes
BioComplex Laboratory, School of Computing
Florida Institute of Technology, Melbourne, USA
faswad2013@my.fit.edu, rmenezes@cs.fit.edu

Abstract

Human migration research is quite multidisciplinary and has yielded works in social sciences, physics, and even the new field of city science. International immigration has societal impacts on both the source and the destination countries at several levels such as economy, city planning, politics, and law enforcement; it leads to changes in the demographics landscape. Existing studies concentrate on modeling the phenomena or on explaining the causal reasons for immigration. In this work, we investigate the role of social media and postulate whether it can be considered a proxy for the reality of international immigration (which includes refugee placement). Our data analysis supports the argument that Twitter may be considered a reasonable source for information about immigration and refugee placement with the benefit that it has a real-time dimension to the information being tracking.

Introduction

Throughout history, wars and natural disasters have left an indelible mark on humans; a common consequence is the displacement of people across countries. Take for instance the *Arab Spring* of 2012 which left thousands of people without shelter, leading to mass migration to Syria and Libya—the instability and violence in the Middle East region have led refugees to leave their homes (Fargues and Fandrich 2012). In fact, this event has played a major role on the increase of refugees around the globe (Fawcett 2016; Carrera, Den Hertog, and Parkin 2012). Recently, we have witnessed a refugee crisis culminating in boats crossing the Mediterranean Sea towards Europe, and thousands of illegal immigrants crossing the Mexico-United States border (Martinez and Slack 2013; Düvell 2008; Fargues and Bonfanti 2014).

The flow of people has positive and negative impacts on both origin and destination countries. On the one hand, the immigration may negatively affect the economy in host countries. For example, when employers hire refugees from parent countries instead of local workers, they save money but lose the support of some of the locals. Also, it is common for refugees to get into debt due to the low wages (Del Carpio, Wagner, and others 2015). On the other hand, migrants may allow for high-wage formal jobs to be offered to locals, allowing an occupational upgrading of local workers (Del Carpio, Wagner, and others 2015; Kapur 2014). In this case, average wage increases primarily because of changes in the composition of the employees. Furthermore, immigration tends to promote diversity in societies, creating significant social benefits, but with the drawbacks of affecting the political stability, security of host countries, and the demographical balance (Kapur 2014).

Indeed, immigration has direct and indirect impacts on our lives making us argue that we need a better understanding of such social phenomenon. Until recently, the lack of data drastically limited our ability to present a comprehensive analysis of the real situation of refugees. In our work, we attempt to address the data limitation using information from social network sites, such as Twitter. Twitter is a great data source because of its open API; something not available on other social networks such as Facebook and Google+.

Here, we investigate whether people's opinions on social network reflect the reality of immigration but more specific related to refugees. Our paper attempts to verify whether social network sites give a valid perception of refugee phenomenon because that could lead to a real-time framework to gauge immigration. In this paper we build a network from Twitter data and another using the United Nations High Commissioner for Refugees data (UNHCR). The data from Twitter was collected from September 2016 to January 2017 using refugee-related keywords in tweets in 8 different languages. From each network, we extracted their structural characteristics and analyze the relationship between the two networks using these characteristics. We found a moderate correlation between the two networks, implying that we can gain information from the social network about the refugees' cases. However, we also find that the level of immigration (refugee) is not fully captured in social media; the structure of the networks is similar but the weights representing the number of immigrants/refugees do not match very well.

Related Work

The availability of data and computational techniques for dealing with such data is the main driving force in Data Science. In the social media world, Twitter is probably the best source of data given its openness. Data from Twitter has been used in many disciplines, including biology, psy-

chology, sociology, and linguistics (Holmberg and Thelwall 2014; Bodnar and Salathé 2013; Murthy 2012; Page 2012). Many of these works attempt to look at society from the optics of the social media platform. For example, in health, users tweeting about their health conditions can eventually lead to an increase the knowledge and awareness about human health (Al-Rubaye and Menezes 2016). In sports and sociology, Pacheco et al. use Data Science techniques to characterize football supporters by observing their activity on Twitter (Pacheco et al. 2016). In linguistics, Saha and Menezes investigated the language of users as an aspect in the spreading knowledge in a social network (Saha and Menezes 2016).

Immigration of refugees is not a new issue and it is generally caused by hardship and wars but also by economic issues and natural disasters. The refugee phenomenon has been the focus of many research studies. However, most of them have been conducted in terms of a global perspective of migration, or using theories to understand the patterns of immigrants at cities and countries levels (Lamanna et al. 2016; Hawelka et al. 2014; Messias et al. 2016).

Hawelka et al. studied human mobility to uncover the patterns of mobility by looking at the mobility rate, the radius of gyration, and the balance of the inflows and outflows of people (Hawelka et al. 2014).

Messias et al. studied human migration using another classic aspect in Network Science, the clustering coefficient. Most studies have focused on the flow between pairs of countries; however, Messias et al. used the concept of triads of countries to cluster instead (Messias et al. 2016). They showed that having clusters using triads reached a better explanation of the phenomena than bilateral flows.

Lamanna et al. shed light on community integration of immigrants using the concept of global cities for 53 cities across the globe. They used a Twitter dataset to detect the spoken languages between users in order to quantify the relationships of cultures between host country and parent country (Lamanna et al. 2016).

Hadgu et al. have examined the discussion about refugees and how it has changed from time to time in countries that accept refugees (Hadgu, Naini, and Niederée 2016). They used a data from Twitter related to refugee situations using keywords and hashtags. They demonstrate that news media plays an important role as a mediator between the actual situation and the perceived refugee situation (Hadgu, Naini, and Niederée 2016). Still, their work was limited to Europe and does not provide a perception on the number of refugees, or how often they are migrating, and which host counties are preferable by them. The importance of our work is in analysis at a country level but considering the entire world.

Datasets and Methods

In this work we use two datasets. First, a dataset collected from Twitter from September 2016 to January 2017. This dataset is built by looking for words broadly related to refugees in 8 languages. Second, the data from the UNHCR official website (United Nations High Commissioner for Refugees 2017).

Twitter Network

We collected a Twitter dataset by gathering tweets containing refugee-related keywords such as refugee, asylum, emigrant, migration, etc. in the 8 most common languages in Twitter: English, Japanese, Spanish, Malay, Portuguese, Arabic, French, and Turkish (MIT Technology Review 2013). The total number of tweets in the period is 12,091,393. After the collection was completed we looked for country names also in the 8 languages. We used the 204 countries listed in the official website of the U.S. Department of State (United States State Department 2017).

Recall that we want to build a network of countries. Hence, the link between countries is done from co-occurrences of country names in a single tweet; if a tweet mentions at least two names of countries, they are linked in the network. For instance, if we assume a user tweet was, "Germany has welcomed more than a million refugees and asylum seekers from Syria", the link Germany-Syria is added to the network. If there are three or more countries mentioned in a single tweet, then we link them with one another as a clique. We assume that the order of country' names in a tweet is not fundamental; thus we represent the links using undirected edges. If a link already exists between to countries, the link weight is increased by one; we have a weighted network. Figure 1(left) depicts how the Twitter network looks like after all the tweets are processed.

When looking for country names we took the precaution to include alternative spellings for several countries. For instance, we used "USA", "U.S.A", "United States", etc. when looking for mentions for the "United States of America".

UNHCR Network

Refugees information is available on the website of the United Nations High Commissioner for Refugees (UNHCR). The UNHCR provides statistics about the number of refugees, the origin country and the host country; we use the data for 2015 (United Nations High Commissioner for Refugees 2017).

We used the UNHCR dataset to extract information about where the refugees are from and where they go to. Similarly to the case with the Twitter data, we created a network but unlike Twitter dataset the network is directed because the source and the destination countries are known. Countries are the nodes and the links are the number of refugees migrating from origin to host country; then we convert the network from directed to undirected in order to have an equitable comparison with the undirected Twitter network.

Three conversion methods were considered, namely: addition, subtraction, and maximum. The *addition* method adds the number of refugees in both directions; for example, if 100 refugees migrate from Russia to the United States and 13 migrate in the opposite direction, the resulting undirected edge between the United States and Russia will have the 113 as its weight. In the *subtraction*, instead of summing links, we subtract them. In our work, subtraction means that some of the refugees may have returned to their origin country and captures the "net" gain/loss of people as part of the migrations. For the example above, the subtraction method will

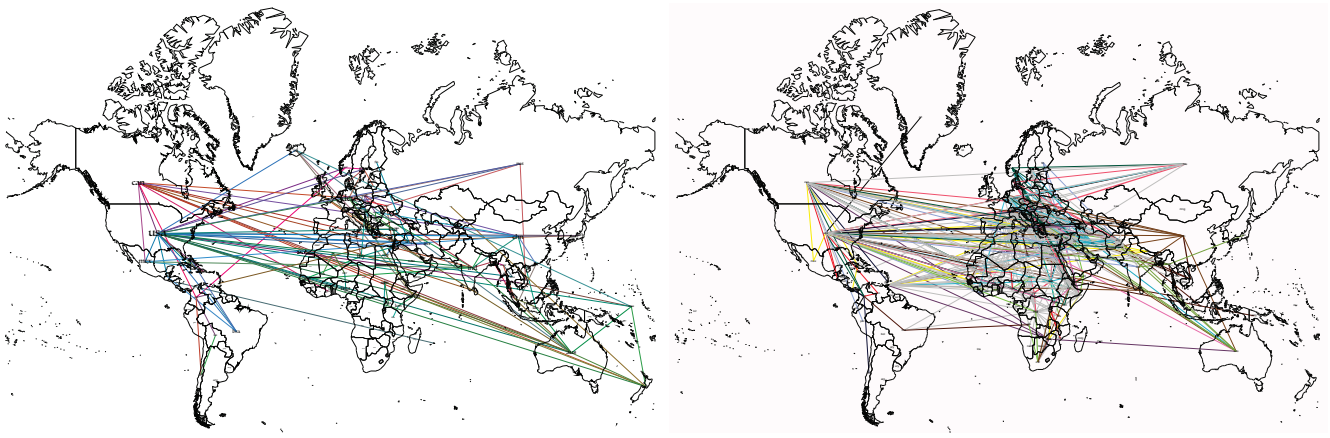

Figure 1: Undirected Twitter network (left). Nodes represent countries and links are the names of countries mentioned in a single tweet. Undirected UNHCR network (right). Nodes represent countries and the links are the refugees flow among countries according to the UNHCR data. For clarity, we are showing edges with weight above 1,000 for UNHCR and above 10 for Twitter.

yield a link with weight 87 between Russia and the United States. The last method is to take the maximum number of refugees from both directions; by applying this concept in the same example above we will have 100 as the weight of the edge between Russia and the United States.

To decide which of the three methods we should use in order to convert the UNHCR directed network to an undirected network we did a Pearson's correlation coefficient analysis. Our test shows high correlation coefficient ranging from 0.98 to 0.99 for all three pairs (see Figure 2). This means that we can use any of the methods to build the network without affecting the results. Thus, we decided to use the addition method to convert the network. Henceforth, we will refer to the UNHCR network as an undirected network which uses the addition method. Figure 1(right) depicts such network.

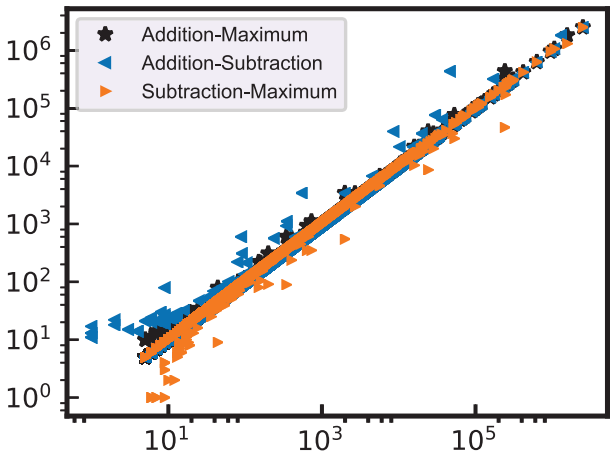

Figure 2: Pearson's correlation coefficient among the three approaches (pairwise comparison) for generating a UNHCR network: Addition-Maximum, Addition-Subtraction, and Subtraction-Maximum. In all instances the correlation is very high making the methods statistically indistinguishable.

Results and Discussion

We have extracted basic properties of networks such as node degree and the weighted node degree distributions. Node degree is one of the most common properties in networks. It is important for our work because it could give us a perception of the popularity of country, that is, the number of different relations a country has. Hence, the higher the node degree, the more diversity of cultures, and the more economic benefit that country may have. Studies on the economy in countries that have cultural diversity suggest that locals have experienced a significant increase in their wages in the rental price of their houses and the benefits appear to outweigh the drawbacks. (Ottaviano and Peri 2006; Bellini et al. 2013; Woodward, Skrbis, and Bean 2008; Gören 2014).

We extracted node degree for both networks described earlier. We then performed a statistical analysis to investigate whether the node degree for the Twitter and the UNHCR are correlated with each other. We used the Spearman's rank correlation coefficient because we are interested in the ranks of the nodes by degree instead of the actual value of the degrees; note that the scale of these degrees in both networks vary a lot given the different domains they come from. We calculated the correlation between the two networks and found a 0.63 correlation, which shows that we have a positive moderate correlation coefficient. Furthermore, we performed a significance test and found a p-value < 0.05 that confirms the significance of the results.

To examine whether our results can be applied to the refugee phenomenon across the globe, we analyzed the confidence interval of the regression. The most convenient

approach involves calculating the confidence interval with 95% confidence around the mean (Kleinbaum et al. 2013). Moreover, it is also useful to estimate the interval in which future refugees will fall, with the probability that given confidence interval already been calculated; thus, we used a prediction interval with 95% (see Figure 3).

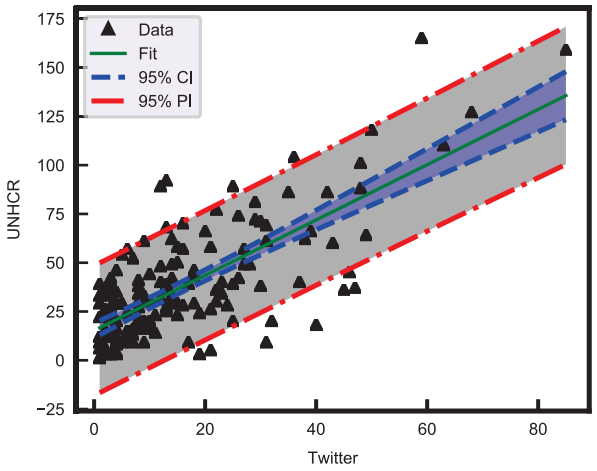

Figure 3: Spearman's correlation coefficient. Twitter node degree vs. UNHCR node degree. Light blue color area represent the 95% of Confidence Interval (CI), and light gray area the represent 95% of Prediction Interval (PI).

From a general perspective, the node degree provides us with information about the diversity of receiving people from other countries. Hence, we analyzed the node degree distribution of the Twitter dataset in order to check the distribution that may fit the data; we performed a log-likelihood ratio test. In our experiments, we found that the distribution tends to follow a truncated power-law. The highest node degree was 96 for the United Kingdom, followed by 85 for the United States, and 67 for Germany, which indicates that United Kingdom is the location with the most diverse set of immigrants/refugees . Recall that this is the perception from social media (see Figure 4).

Similarly, we analyzed the node degree distribution for the UNHCR dataset and our finding also shows that it also follows a truncated power-law distribution. The highest node degree was 165 for Canada, followed by 159 for the United-States, and 127 for Germany, which indicates that Canada is the country receiving refugees from the most diverse set of nations according to the UN (see Figure 5). It is interesting then to see that despite having similar distributions the highest degree node differ from each other. While people believe the UK receives the most diverse set of immigrants the data from the UN appears to show that to be Canada. Regardless of these differences, the Spearman's correlation shown in Figure 3 shows that the general perception reflects well the reality.

Degree distribution captures diversity but not the amount of people. To look at that, we worked with the weighted degree distribution. Here, we want to find out whether the

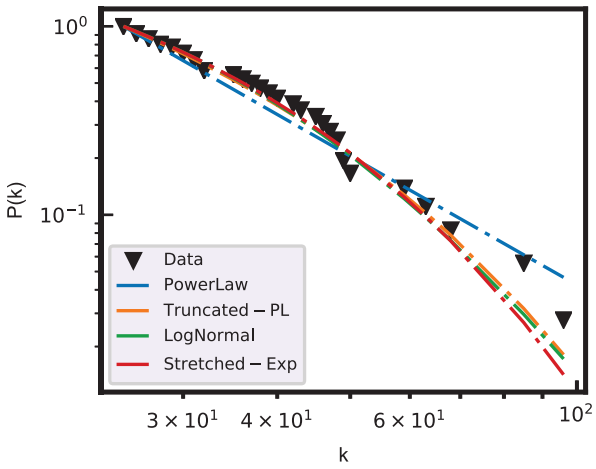

Figure 4: Twitter node degree distribution. We used log-likelihood ratio to test several functions that could fit the distribution. The truncated power law appears to have the best fit.

number of refugees in Twitter and UNHCR dataset is correlated or not. Likewise, in node-degree correlation, we calculated the Spearman's rank correlation coefficient and we obtained positive and moderate and it was 0.453 with significant p-value < 0.05. Similarly, we calculated the confidence and prediction interval with 95% (see Figure 6). Note that the correlation related to the amount of refugees/immigrants is a weaker than for the degree distribution but the ranks are still somewhat preserved although here the public (Twitter) perception of the reality is not so accurate.

The weighted degree, in the context of refugee network, tells us about the number of refugees that country receives regardless of origin. For instance a country could have a high weighted degree but have most of the refugees arriving from few countries. Hence, we analyzed the weighted degree distribution for the Twitter dataset and our findings show that the dataset tends to follow a truncated power-law distribution. The highest weighted degree was 7273 for the United States, followed by 4417 for the United Kingdom, and 4112 for Austria (see Figure 7).

The weighted degree distribution for the UNHCR dataset followed the truncated power-law distribution. The highest weighted degree was 4.8M for Syria, followed by 2.5M for Turkey, and 2.4M for Afghanistan (see Figure 8). Note that these 3 countries do not match the highest in the Twitter dataset. The reason is that these countries recently had a lot of refugees settled in other countries but the population (from Twitter) appears to concentrate on the host nations names more than the origin of the refugees.

Conclusion

The movement of refugees/immigrants and their impact on communities have become the subject of interest and concern to society. The availability of a large datasets has enabled a better understanding of the problem because they

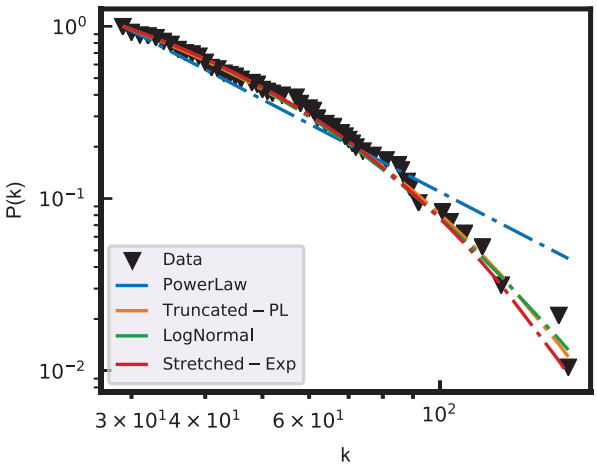

Figure 5: UNHCR node degree distribution. Log-likelihood ratio test was used to fit the distributions. The truncated power law appears to have the best fit.

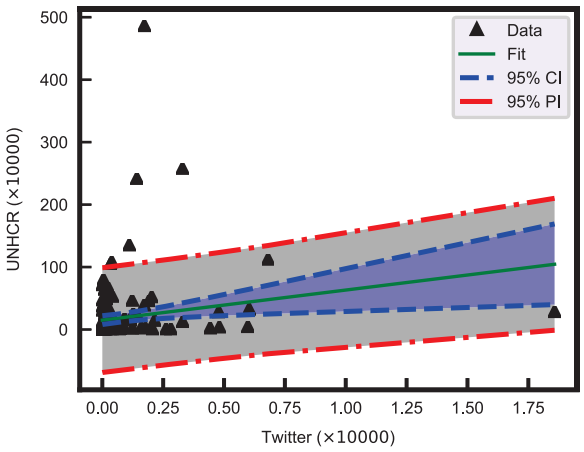

Figure 6: Twitter vs. UNHCR weighted degrees. Light blue color area represent the 95% of Confidence Interval (CI), and light gray area the represent 95% of Prediction Interval (PI).

provide the means to investigate the phenomenon from a data-centered angle; yet, we need to analyze how much we can tell about the phenomenon using the data from social networks. Moreover, the analysis between social media and reality gives us a good indication as to whether society see the issue as it actually is.

In our work, we built two networks, one from Twitter and another from official UNHCR data on refugees. We then looked at the degree and weighted degree distributions in both networks to gauge the perception of people against reality. Degree gives the diversity of immigration whereas weighted degree indicates the amount of people who immigrated.

We used Spearman's rank correlation coefficient to see if the networks have similar ranks. The distributions of both networks show that we have very few countries concentrating most of the refugee/immigrant population. Yet, we showed that the public perceives the rank of diversity a lot better than the amounts each country is hosting. The countries with most diversity coincide more in both networks than the amounts.

Our work is an important first step towards the use of social as a gauge of reality when it comes to understanding immigration. The difference in the results between the two datasets showed us that the public (at least the demographics represented on Twitter) has a fairly good view of the reality but slightly distorted by the amounts of refugee/immigration.

In this paper, we treated immigration and refugee in a combined way. However, these are separate issues and in the future we will study them separately; for that we are working on collecting more data from Twitter. Yet, the results stand, meaning that they help us understand better how the public sees the immigration phenomena.

Acknowledgments

Firas M. Aswad would like to thank the Ministry of Higher Education and Scientific Research (MoHESR, Iraq) and the University of Mosul for financial support under grants 16052 in 19/05/2014 followed by 1600 in 18/05/2014.

References

Al-Rubaye, A., and Menezes, R. 2016. Extracting social structures from conversations in twitter: A case study on health-related posts. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*, 5–13. ACM.

Bellini, E.; Ottaviano, G. I.; Pinelli, D.; and Prarolo, G. 2013. Cultural diversity and economic performance: evidence from european regions. In *Geography, institutions and regional economic performance*. Springer. 121–141.

Bodnar, T., and Salathé, M. 2013. Validating models for disease detection using twitter. In *Proceedings of the 22nd International Conference on World Wide Web*, 699–702. ACM.

Carrera, S.; Den Hertog, L.; and Parkin, J. 2012. Eu migration policy in the wake of the arab spring: What prospects for eu-southern mediterranean relations?

Del Carpio, X. V.; Wagner, M.; et al. 2015. *The impact of Syrians refugees on the Turkish labor market*. World Bank.

Düvell, F. 2008. Clandestine migration in europe. *Social Science Information* 47(4):479–497.

Fargues, P., and Bonfanti, S. 2014. When the best option is a leaky boat: why migrants risk their lives crossing the mediterranean and what europe is doing about it.

Fargues, P., and Fandrich, C. 2012. Migration after the arab spring. Technical report.

Fawcett, L. 2016. *International relations of the Middle East*. Oxford University Press.

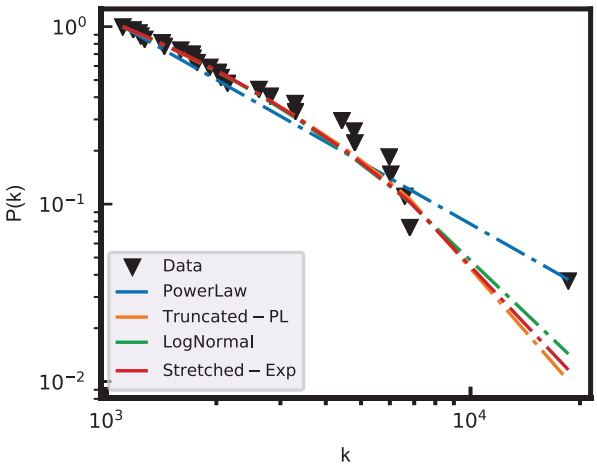

Figure 7: Twitter weighted degree distribution. Log-likelihood ratio test using the Power-law (PL) package among different distribution.

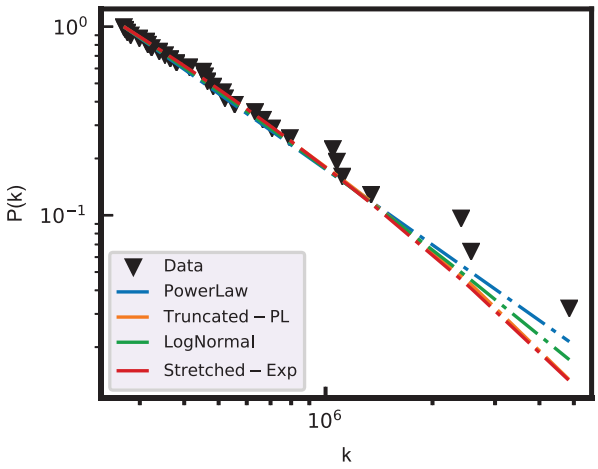

Figure 8: UNHCR weighted degree distribution. Log-likelihood ratio test using the Power-law (PL) package among different distribution.

Gören, E. 2014. How ethnic diversity affects economic growth. *World Development* 59:275–297.

Hadgu, A. T.; Naini, K. D.; and Niederée, C. 2016. Welcome or not-welcome: Reactions to refugee situation on social media. *arXiv preprint arXiv:1610.02358*.

Hawelka, B.; Sitko, I.; Beinat, E.; Sobolevsky, S.; Kazakopoulos, P.; and Ratti, C. 2014. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41(3):260–271.

Holmberg, K., and Thelwall, M. 2014. Disciplinary differences in twitter scholarly communication. *Scientometrics* 101(2):1027–1042.

Kapur, D. 2014. Political effects of international migration. *Annual Review of Political Science* 17:479–502.

Kleinbaum, D.; Kupper, L.; Nizam, A.; and Rosenberg, E. 2013. *Applied regression analysis and other multivariable methods*. Nelson Education.

Lamanna, F.; Lenormand, M.; Salas-Olmedo, M. H.; Romanillos, G.; Gonçalves, B.; and Ramasco, J. J. 2016. Immigrant community integration in world cities. *arXiv preprint arXiv:1611.01056*.

Martinez, D., and Slack, J. 2013. What part of illegal dont you understand? the social consequences of criminalizing unauthorized mexican migrants in the united states. *Social & Legal Studies* 22(4):535–551.

Messias, J.; Benevenuto, F.; Weber, I.; and Zagheni, E. 2016. From migration corridors to clusters: The value of google+ data for migration studies. In *Advances in Social Networks Analysis and Mining (ASONAM), 2016 IEEE/ACM International Conference on*, 421–428. IEEE.

MIT Technology Review. 2013. Most-used languages on twitter. https://www.statista.com/statistics/267129/most-used-languages-on-twitter. Accessed: November 20, 2017.

Murthy, D. 2012. Towards a sociological understanding of social media: Theorizing twitter. *Sociology* 46(6):1059–1073.

Ottaviano, G. I., and Peri, G. 2006. The economic value of cultural diversity: evidence from us cities. *Journal of Economic geography* 6(1):9–44.

Pacheco, D. F.; Pinheiro, D.; de Lima-Neto, F. B.; Ribeiro, E.; and Menezes, R. 2016. Characterization of football supporters from twitter conversations. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, 169–176. IEEE.

Page, R. 2012. The linguistics of self-branding and micro-celebrity in twitter: The role of hashtags. *Discourse & communication* 6(2):181–201.

Saha, P., and Menezes, R. 2016. Exploring the world languages in twitter. In *Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on*, 153–160. IEEE.

United Nations High Commissioner for Refugees. 2017. UNHCR statistical yearbook 2015, 15th edition. http://www.unhcr.org/en-us/statistical-yearbooks.html. Accessed: November 20, 2017.

United States State Department. 2017. A-z list of country. https://www.state.gov/misc/list/.

Woodward, I.; Skrbis, Z.; and Bean, C. 2008. Attitudes towards globalization and cosmopolitanism: cultural diversity, personal consumption and the national economy. *The British journal of sociology* 59(2):207–226.