Using Twitter Data to Estimate the Relationships between Short-term Mobility and Long-term Migration

Lee Fiorio
University of Washington
Seattle, USA
fiorio@uw.edu

Guy Abel
Shanghai University
Shanghai, China
guy.abel@shu.edu.cn

Jixuan Cai
Chinese University of Hong Kong
Hong Kong, China
caijixuan@link.cuhk.edu.hk

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

Ingmar Weber
Qatar Computing Research Institute
Doha, Qatar
iweber@hbku.edu.qa

Guillermo Vinué
Austrian Academy of Sciences
Vienna, Austria
guillermo.vinue.visus@oeaw.ac.at

ABSTRACT

Migration estimates are sensitive to definitions of time interval and duration. For example, when does a tourist become a migrant? As a result, harmonizing across different kinds of estimates or data sources can be difficult. Moreover in countries like the United States, that do not have a national registry system, estimates of internal migration typically rely on survey data that can require over a year from data collection to publication. In addition, each survey can ask only a limited set questions about migration (e.g., where did you live a year ago? where did you live five years ago?). We leverage a sample of geo-referenced Twitter tweets for about 62,000 users, spanning the period between 2010 and 2016, to estimate a series of US internal migration flows under varying time intervals and durations. Our findings, expressed in terms of 'migration curves', document, for the first time, the relationships between short-term mobility and long-term migration. The results open new avenues for demographic research. More specifically, future directions include the use of migration curves to produce probabilistic estimates of long-term migration from short-term (and vice versa) and to nowcast mobility rates at different levels of spatial and temporal granularity using a combination of previously published American Community Survey data and up-to-date data from a panel of Twitter users.

CCS CONCEPTS

•Applied computing → Sociology; •Human-centered computing → *HCI theory, concepts and models;*

KEYWORDS

Twitter, Migration, Mobility, Demographic research

WebSci'17, June 25-28, 2017, Troy, NY, USA.

DOI: http://dx.doi.org/10.1145/3091478.3091496

1 INTRODUCTION

Over the last fifty years, migration has played an increasingly important role in population change [16, 25]. Global international migration flows have steadily grown by 7.3 million between 1995-2000 and 2005-2010 [1] and have been at the center of recent political debates across the globe. For example, the Arab and middle East unrest are causing a refugee crisis, with impact on the stability of European societies [10]. Rapid urbanization in China has led to a large number of new immigrants without urban *Hukou* – a record in the Chinese government system of household registration. As a result, there is a large number of immigrants in cities, who hardly enjoy urban welfare [4]. Long-term migration and short-term mobility are also fundamental drivers of the spread of infectious diseases [2, 9].

Despite the growing need for timely and high-quality migration data, measuring geographic mobility and its relationship to long-term migrations remains an elusive goal for demographers. There are two main types of challenges. First, different countries use different definitions of migration. Some countries collect information on immigrants, other on emigrants. Some countries define a migrant as someone who has relocated to a new residence for a period of at least 3 months. Others use temporal frameworks of 6 months or a year. Some statistical offices produce estimate using data from registrations systems. Others rely on surveys. In other words, migration data are often inconsistent across countries. The second main challenge is that data collection and modeling of long-term migration are typically unrelated to the study of short-term mobility, and vice versa. Similarly, researchers who study internal migration typically tie their models to international migration processes only rarely. The lack of interaction between approaches and communities is mostly motivated by the absence of appropriate data that could enable researchers to cross bridges and model migration and mobility at various temporal and spatial scales, within a unified framework.

The primary focus of this paper is to contribute to ongoing work in designing and assessing methodologies for harmonizing migration estimates. In particular, we use geo-referenced Twitter data as a test-bed of migration theory and investigate the link between short-term mobility and long-term migration among a robust sample of social media users followed over several years. By calculating different kinds of migration rates using the same sample of users,

we can test for systematic variation in migration estimates along these dimensions. From these experiments, we can further refine tools to harmonize migration estimates, when they are inconsistent or sparse, and develop tools to produce more timely estimates (i.e. "now-casts") of migration. Broadly speaking, we use estimates of migration and mobility generated from a panel of Twitter users to demonstrate the ways in which migration rates calculated using different transition intervals and durations can be modeled along a single curve and be linked together with quantifiable uncertainty. This has important implications for the study long-term migration from short-term migration (or vice versa) when certain kinds of data are sparse.

2 BACKGROUND

The Balancing Equation, the most fundamental formula in demography, states that population change can be expressed as a function of three basic processes: fertility, mortality and migration. It is the role of the demographer to investigate these processes and their trends, either historical or contemporary, and to make projections [27]. As population processes, fertility, mortality and migration all exist at the nexus of biology and sociology, but it is common for demographers to parse the biological from the sociological by creating models that treat a well-defined biological outcome (e.g. total number of births) as the response variable and socio-economic/socio-cultural characteristics (e.g. GDP, or female secondary education) as predictors [3, 7, 24]. Unlike fertility and mortality, however, it is difficult to reduce migration to a strictly biological population process [21]. Even the most basic contemplation of migration quickly leads to tricky ontological questions – what is a migrant? what is a migration? – that are more overtly sociological and political in nature and that necessitate a theoretical framework for understanding the relationship between people, time and space.

The definitional issues entailed in studying migration mean that there has been comparatively less work refining the methodologies for measuring migration and, as a result, migration data are more difficult to use and to compare across contexts. Estimates of migration are sensitive to the definitions of time and space, and reconciling inconsistent migration rates is a prevalent and enduring problem in the study of migration [20, 22, 23]. While efforts have been made to impose standards in certain settings, such as the attempt by European Union policy makers at regulating international migration rates estimated by member states [5], transformations of migration statistics to meet such requirements are frequently applied post hoc thus highlighting the need for sound harmonization methodologies [20].

In migration research, a significant obstacle has often been a lack of empirical data from which to test for systematic differences between inconsistent migration statistics. Though a migration rate that uses a six month threshold to define residence and a migration rate that uses a one year threshold to define residence are observations of the same underlying process (i.e. the movement of individuals or households in space), it is rare for a single survey to produce multiple estimates of migration for comparison and calibration. At the same time, the increasing availability of geo-referenced user-generated content from social media data is offering new opportunities to study migrations in new ways [8, 17, 28]. With a sufficiently large sample of users and many geo-referenced posts from those users, it becomes possible to convert their activity into streams of individual movement which can then be aggregated into migration rates and migration flows using any number of temporal or spatial criteria of the researchers' choosing [11, 13, 14].

2.1 Conceptualizing Migration

People move around in time and space. Migration literature has historically distinguished between two perspectives from which to observe movement, that of the migration event or movement and that of the migrant or mover [6]. The former consists of event data describing the timing of a move, while the latter consists of stock data describing whether or not an individual has experienced a move (these data are usually counts of the number of foreign born). Other scholars, including Rees ([1977]), have included a third type of migration data, transition data, which describes the case where migration is assessed based on residence at two discrete points in time. Most migration data collected by government agencies and used in demographic projections is transition data. Transition data are sensitive to two kinds of temporal variation. In this paper we set aside issues of geographic scale and focus on temporal dimensions of migration measurement which come in two forms: duration and interval.

2.2 The Concept of Duration

Perhaps the least standardized aspect of migration measurement is duration. Duration can be understood as the minimum length of residence necessary to qualify as a migrant rather than as a visitor or tourist. There is no standard in the European context [26] with a variety of durations used (3 months, 6 months, 1 year).

Methodologically, duration is the most crucial element in converting discrete geo-located data into migration estimates [15]. A standalone geo-located data point can tell us where a particular person was at a specific point in time, but it tells us nothing about how long the person remained at that location. For this reason, it is necessary to group geo-located data points together based on their temporality in order to establish an estimate of where each person spent their time.

Figure 1 illustrates our method for imputing residence and migration using this conceptualization of "duration". First we choose two reference points, t and $t+$. We then select the tweets that occurred within some buffer, d, around each reference point. The residence of a user over duration d is determined as the modal tweet location of the user. To avoid double counting tweets, d is never larger than the time between the two reference points.

The intuition about the relationship of duration on migration rates is as follows: the shorter the duration, the larger the amount of short term migration will be captured, and so the higher the rate of migration. If the duration is two weeks for example, then a good deal of vacations, business trips and other kinds of non-permanent travel will be observed in the estimate, regardless of the distance between the two points in time. If, for example, one wanted to estimate the migration rate between July 1st, 2010 and July 1st, 2011, we expect to estimate a higher rate of migration if we use a two week duration to impute location than if we use a month

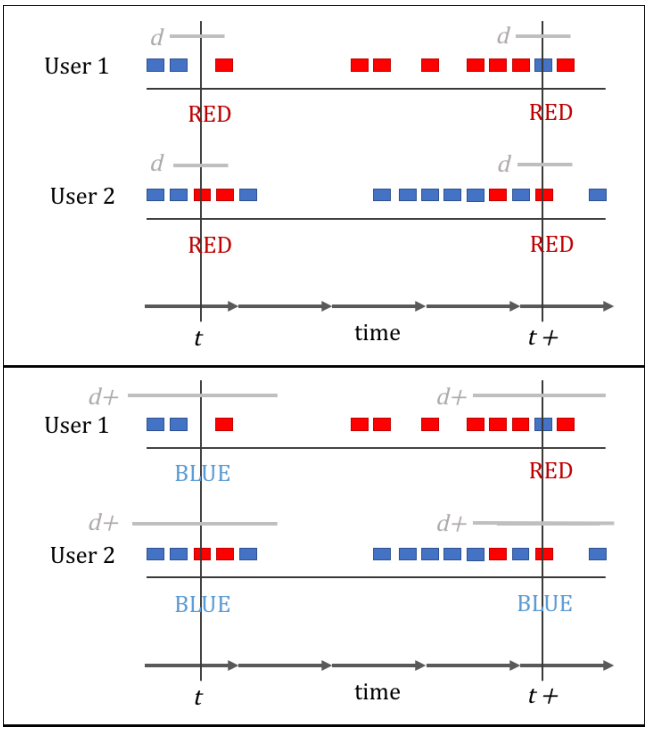

Figure 1: ***Illustrative example of the concept of duration. Between time t and t+, two users tweet from either location blue or red. The location assigned to each user (color-coded text in caps) depends on modal location of tweets occurring within the duration (horizontal gray lines) or buffer around the two reference points, t and t+. The data points are the same in the top and bottom panels, but the duration used to identify the location of residence is shorter in the top panel.***

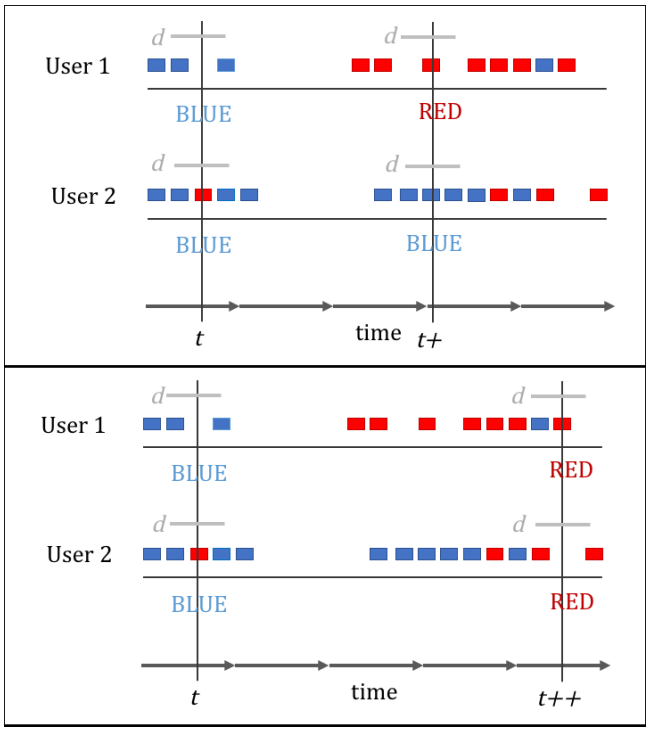

Figure 2: ***Illustrative example of the concept of interval. Between time t and t+, two users tweet from either location blue or red. Holding duration constant, the location assigned to each user (color-coded text in caps) is a function of the interval, the time between t and t+ or t++ (vertical gray lines). The data points are the same in the top and bottom panels, but the interval used to identify the location of residence is shorter in the top panel.***

duration. We also expect that as durations grow very large, the negative effect on migration rates should diminish. That is, the change in migration rate should flatten out at large durations. We expect that this will happen because we hypothesize that the distribution of non-permanent trips approximates a negative exponential: the difference between the number of two-week trips and four-week trips is greater than the difference between the number of 30-week trips and 32-week trips. Therefore, the difference between migration rates estimated using two- and four-week intervals should be greater than the difference between migration rates estimated using 30- and 32-week intervals. In summary, we hypothesize that, holding the interval constant, migration rates should fall as the duration increases, eventually flattening out.

2.3 The Concept of Interval

The concept of interval is more straightforward. In order to determine whether a migration transition took place, it is necessary to select two points in time and compare the estimated location of residence. The interval is the temporal distance between two points of comparison. A common way for migration data to be collected via survey is to ask respondents a question similar to "where did you live five years ago?" Five years in this example would be the interval. In such a question, no data on duration is explicitly collected, though arguably some notion of duration is implied by the concept of "to live"/"residency". That is, even if a respondent was on vacation precisely five years before completing a questionnaire with this question, chances are that the respondent will answer with their place of residence five years ago and not with the location where they traveled to on vacation. This does not mean that duration is the same as the interval in these types of survey questions, just that we can expect the duration to be sufficiently large as to capture some sense of "residency".

Figure 2 illustrates the concept of interval as it relates to our method of estimating migration. First, we fix one reference point at time t and impute a location for this time based on the modal location of tweets that occurred within duration d. Then, we select a second reference point at time t+ and impute a location. To estimate whether a migration occurred for a given person, we compare their imputed locations at the two reference points. As we increase the interval, we keep the left hand reference point, t, fixed while shifting the right hand reference point forward in time, i.e. to t++. While the decision to fix the left hand reference point was predicated

on goal of eventually being able to forecast long term migration from short term migration, it would be just as easy to fix the right hand reference point and work backwards in time. Results from our analysis would be the same.

The intuition about the relationship between interval and migration estimates is as follows: the larger the interval, the longer the period of exposure, and the higher the migration rate. Here it is helpful to think about migration as a risk. Exposure to the risk of migrating accumulates over the interval, so that longer intervals should produce higher rates of migration. That being said, with larger intervals also comes the increased possibility that moves go unobserved. If a person moves from place A to place B to place C over an interval, we would only capture A and C or A and B (depending on the duration and the timing of the moves). Similarly, in the case of return migration, when a person moves from place A to place B to place A, we might capture A and A (depending on the duration and on the timing of the move) and not classify this person as a migrant. Because the risk of return migration also increases with time, we expect the positive relationship between interval and migration rate to be slightly diminishing at large intervals. In summary, we hypothesize that, holding the duration constant, migration rates should rise as the interval increases, but the rate of increase should diminish at larger intervals.

2.4 Linking Duration and Interval

The ultimate goal of this research is to examine the joint effect of duration and interval on migration rates and to assess the potential for estimating long-term migration from short-term mobility and vice versa. When attempting to estimate migration rates from discrete individual-level geo-location data, short-term moves will always be mixed in with long-term moves. It is difficult if not impossible to infer a person's intent to stay in a location based on their (geo-locational) tweeting behavior. A migration rate estimated using a three month interval and a one month duration, for example, will count a relatively large number of individuals as migrants: people who have traveled from place A to place B for a short trip but who will return to A as well as people who have permanently moved from A to B. The question becomes whether one kind of movement is noise to the other's signal or whether short-term mobility and long-term migration can provide information about each other.

In literature on migration there is a theoretical connection between short-term mobility and long-term migration. At the most basic level, greater connectivity and commonality between two places (e.g. capital flows or a shared language) is associated with higher degrees of migration [19]. Similarly, short-term mobility might signal the existence or even presage changing patterns in long-term migration. For example, it is not uncommon that economic migrants who intend to stay abroad for only a short period end up becoming permanent residents of their new country. Alternatively, it is possible that a person who travels to a particular place frequently for business or pleasure simply decides to relocate to that place. Increased short-term mobility between two places results in greater exposure to possible long-term migrants, but the precise relationship has not been explored in the literature.

The growing wealth of social media data provides the means for testing theories about the relationship between short-term mobility and long-term migration in a way that could not be done before. Our analysis is largely exploratory in this sense; however, our basic intuition is as follows: all movements can be plotted along a single curve and that the shape of the curve contains information about the relationship between short-term mobility and long-term migration in whatever context the rates have been estimated.

Our conceptual approach towards understanding the relationship between short-term mobility and long-term migration is to investigate the joint effect of duration and interval on migration estimates. Though conceptually distinct, duration and interval are co-dependent in the sense that duration can never be larger than interval and interval can never be smaller than duration. One method for estimating a curve covering short-term mobility estimates (i.e. estimates with small duration and small interval) and long-term migration (i.e. estimates with large duration and large interval) is to plot a curve using estimates for which duration always equals interval. We hypothesize that such a curve would decrease over short intervals until they reach an inflection point and begin to rise. Demonstrating the possibility of plotting such a curve would be a significant contribution to migration literature in that it would suggest that researchers could then study the nature of these curves (e.g. their shape, the location of their inflection points, and so on) as they relate to different patterns in short- and long-term movement.

3 DATA AND METHODOLOGY

The bulk of data collection occurred in real time from January 2010 to March 2013. In particular we use, as seed, data that were collected for a previous study (see [28]). We then took a subset of the data that includes the geographic coordinates from users who posted geolocated tweets in the U.S. over the period. This resulted in 12.5 million tweets with latitude and longitude from a sample of 62,381 Twitter users. Further data collection was performed in September 2016 by querying individual user timelines from our sample using the Twitter API. This allowed us to collect geo-tagged tweets from the same group of users, but that were posted more recently, since March 2013. We ended up with a total of 15.3 million geolocated tweets. Because the API only provides access to the 3,200 most recent tweets and because geo-tagged tweets are relatively rare, the majority of tweets in the analysis were posted over the original time frame. Duplicates of the original tweets were detected by the unique tweet ID and removed. By using a sample of users who posted over a long period of time and who opened their accounts several years ago, we believe that we are reducing the likelihood of having bots in our sample. In other words, we could have chosen to have a larger sample of users, but we preferred to have a smaller sample of "trusted" users who post on a regular basis. For this analysis, we focused on the U.S., partially because for this first analysis we wanted to avoid the need to model differential usage of Twitter and differences in selection bias across countries.

To make sure that our one-week bins add up to months and years, we made it so each month has exactly four "weeks". We did this by by classifying the first eight days of each month as week one, the next seven days as week two, the next eight days as week three and the remaining days (either eight, seven or five) as week four. While

this means that there is some inconsistency in our unit of time, we argue that any impact on our results is minimal, especially at large intervals and/or durations. From this point on, "week" will refer to a quarter of a month. A year using this schema has 48 weeks.

Tweets are converted into migration flows using the following steps:

(1) Each tweet is geocoded to the US county level by the latitude and longitude and assigned to one of the nine US Census Divisions or to a tenth overseas "division" if the tweet originates from outside the US.
(2) Each tweet is placed in a one week bin based on its time stamp. This allows for more efficient aggregation in later steps. An example of user time lines of state-level movement are visualized in Figure 3 which shows flows of our sampled Twitter users in and out of Arkansas.
(3) The tweets are grouped by user ID to create user-specific time lines.
(4) Twitter users are assigned a location based on the modal location of their tweets during a specified duration around each reference date. For example, if we want to know the location of a given user on March 1st, 2011, and the specified duration is three months, we would determine which of the divisions contained the majority of that user's tweets in the six weeks before and after March 1st, 2011. If there is a tie between two or more locations, we assign the location that occurred first over the duration.
(5) Users are categorized based on whether they changed locations over the interval, i.e. whether the modal location at reference point t is different from the modal location at reference point $t+1$.
(6) Finally, we sum across all eligible users to create counts of movers and non-movers for each division by specified start date, interval and duration. We then use these counts to calculate migration rates. In the analysis that follows, we aggregate the flows between the nine different regions to estimates of the division-level internal migration rate within the US. In future iterations of this project, it is possible to investigate patterns in the rates calculated between individual pairs of divisions.

It should be noted that the tweeting behavior of users is highly irregular with respect to time. There may be users in the data set who only posted a minimal number of geo-tagged tweets and who did so in the span of a month or two in 2011, for example. When estimating rates, any user for which there is insufficient data (i.e. zero tweets during the specified duration for one or both reference points) is excluded from both the numerator and the denominator of the estimated rate. This means that information from a user might appear in the migration rate estimated between reference point t and $t+$ but be excluded in the migration rate estimated between t and $t++$. Similarly, a user might be excluded in the estimate of migration rate t and $t+$ using duration d but be included using duration $d+$. We proceed with our analysis under the assumption that the pattern at which users are included or not included in our rates does not vary systematically with the rates. There may be cases when this assumption may not hold; however, we argue that in this initial exploration of migration curves, the added benefit of including as many users as possible in our analysis is worth the cost.

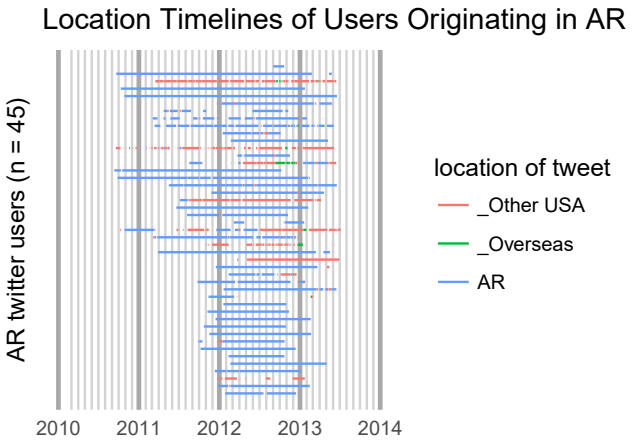

Figure 3: ***Graphical representation of Twitter-estimated migration histories of users originating in Arkansas. Each line represents one user and is color-coded by the estimated location of the user over time (in AR, in a different US state, outside the US).***

4 RESULTS

With data spanning close to six years, there are a large number of migration estimates we can calculate using different combinations of start date, interval and duration. We have calculated migration rates and flows between each of the nine US Census Divisions[1]. For any set of estimates, there are a total of 72 flows, one for each directional pair of divisions (9x8).

4.1 Larger Duration: Smaller Mobility Rate

Our first hypothesis is that there is a negative but diminishing relationship between duration and estimated migration rate. We expect that, as the duration grows, the amount of non-permanent mobility (e.g. holidays, business trips) observed in our estimate will decrease; however, as the duration increases past a certain point, we predict that the changes should flatten out. To test this hypothesis, we estimated a series of migration rates holding the interval constant at 72 weeks (a year and a half) while increasing the duration from two to 72, in two-week increments. To generate multiple estimates for each duration, we shifted the left hand reference date from January 1st 2011 to September 2012 in 6 week increments. This resulted in 15 estimates for all 36 durations for a total of 540 migration estimates.

Results from this analysis are plotted in Figure 4. The blue line shows the estimates of the trend in the data points, from a non-parametric smoother, together with the associate 95% confidence intervals. It is clear from this plot that our hypothesis is largely validated. The relationship between duration and migration rate

[1] https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf

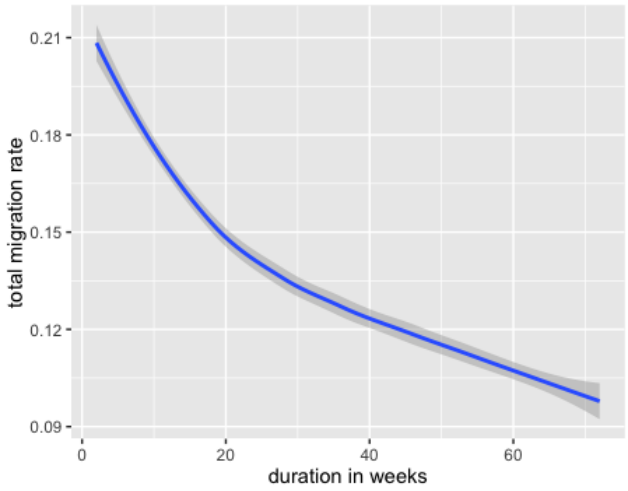

Figure 4: ***Plot of estimated migration rate as a function of duration length, holding interval constant.***

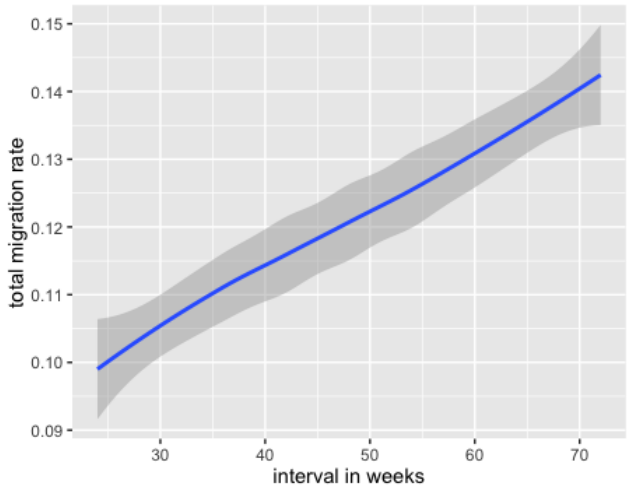

Figure 5: ***Plot of estimated migration rate as a function of interval length, holding duration constant.***

is negative and diminishing. Yet, the plot does not flatten out like we expected. Even at the right hand side where the durations are very large (larger than a year) we still see a fairly pronounced negative relationship. It is unlikely that there are many users who are counted as migrants at a one year duration but who are no longer counted as migrants at one year+one month durations. Perhaps users that tweet infrequently–who are only counted when durations are large–have lower migration propensities than users that tweet frequently. In other words, for very large durations, the data may be more noisy. Future work with larger samples is needed to test the robustness of the trend for large values of duration.

It should be noted here that there appears to be a slight elbow in the trend line between five and six months (20 and 24 weeks). The trend line and information about potential inflection points can be used to evaluate the ideal duration for calculating long-term migration. An unanswered question in the literature is related to the evaluation of the point at which the marginal impact of an increased duration falls below some threshold of acceptable variance. Findings from the duration analysis presented here suggest that the differences between rates estimated with less than six month durations are larger than the differences between rates estimated with more than six month durations. We are not offering a definitive answer, as our results are potentially subject to a number of biases (e.g., Twitter users are not representative of the underlying population, and there may be platform-specific biases in behavior, especially with respect to the use of geolocation). However, we believe that this is a good starting point for further developments using other data sources, and for geographic areas outside of the US.

4.2 Larger Interval: Larger Mobility Rate

Our second hypothesis is that there is a positive relationship between interval and migration rate, but due to return migration, this relationship is diminishing. To test this hypothesis, we estimated a series of migration rates holding the duration constant at 24 weeks (six months) while increasing the interval from 24 to 72, in six-week increments. To generate multiple estimates for each interval, we shifted the left hand reference date from January 1st 2011 to September 2012 in 6-week increments. This resulted in 15 estimates for all nine intervals for a total of 135 estimates.

Figure 5 shows results from this analysis. Here we have evidence that the relationship between interval and migration rate is positive as we expected, but there is no evidence from this plot that the relationship is diminishing. It should be noted that the largest interval in this analysis, 1.5 years, is relatively small. It is not uncommon for migration estimates to be generated from survey questions that ask respondents about their locations two, five, or even ten years before. The amount of return migration that would occur over an 18-month period is likely not very high, and therefore it arguably makes sense that this plot does not show the relationship between interval and migration as diminishing.

Results from this section of the analysis are promising for the prospect of improving data harmonization. The plot clearly suggests that it may be possible to estimate long-term migration from short- to medium-term migration. Certainly two migration rates, one that uses a six-month interval and a second one that uses a one-year interval could be combined to produce an estimate of the migration rate at one and a half years. In future analyses, the interval can be expanded to explore whether the linear relationship between interval and migration rate holds. This would lead to the development of conversion scales that enable researchers to "translate" expected values of migration rates across different definitions of interval.

4.3 Joint Effect of Duration and Interval

Our third hypothesis was that short- and long-term rates of movement could be plotted along a single curve by estimating a series of rates for which duration and interval are always equal. We further hypothesized that this curve would be shaped like a 'U' – declining and then rising as the estimates move from short-term mobility to

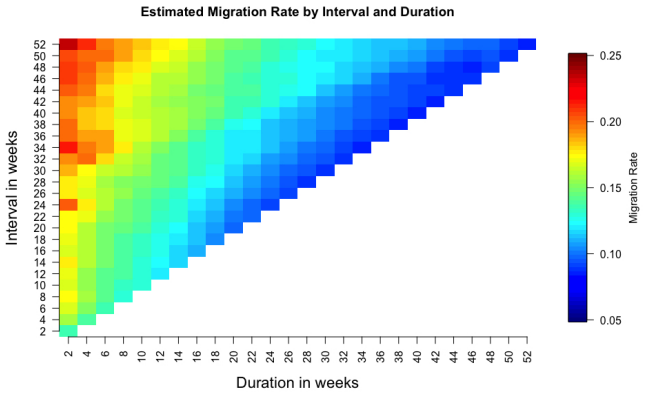

Figure 6: ***Plot of estimated migration rate as a function of interval and duration length. Rates were estimated fixing July 1st 2012 as the starting point.***

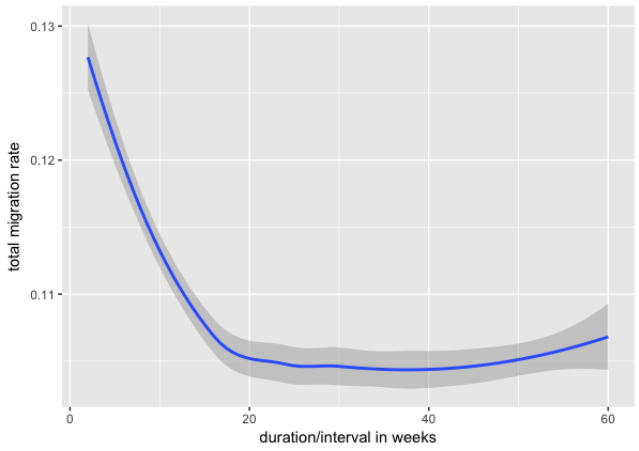

Figure 7: ***Plot of estimated migration rate as a function of duration and interval, where duration = interval.***

long-term migration. To explore the underlying intuition of this hypothesis, we first estimated 351 migration rates with start date July 1st 2011: one for each possible interval-duration pair (2 to 72 by 2) x (2 to 72 by 2) such that interval would always be greater than or equal to duration. The results from this initial analysis are visualized as a contour plot in Figure 6. This plot demonstrates that for nearly all columns (intervals) migration rates increase, and that for nearly all rows (durations) migrations rates decrease, but the two interact in interesting ways.

The next step towards testing our hypothesis involves plotting a single curve using rates for which duration and interval are equal. These rates are equivalent to those plotted along the diagonal of Figure 6. We estimated a large set of such rates using values from two weeks to 60 weeks by increments of two. To ensure that there would be multiple observations of each value, we shifted the left hand start date from January 1st 2010 to January 1st 2011 in one-month intervals. This resulted in 12 estimates of 30 interval-duration lengths, or 360 total estimates.

Results from this analysis are shown in Figure 7, which include the trend line as well as 95% confidence intervals. Once again, our hypothesis regarding the shape of this curve is mostly confirmed. Migration rates decline and then, at some point, begin to increase. What is unexpected, however, is the long trough between 20 and 50 weeks for which the rates stay relatively constant.

5 DISCUSSION AND FUTURE DIRECTIONS

This paper has demonstrated a new use of geo-tagged social media data for exploring unanswered questions related to migration theory. These theoretical questions regarding the relationship between short-term mobility and long-term migration can only be tested empirically using high-volume geo-coded data that come most readily from social media sources. In the analysis above, we have demonstrated that interval and duration behave in ways consistent with our intuitions and migration theory. Importantly, these hypotheses could not have been tested with traditional data sources. Social media data offer new opportunities to validate or confute theories using empirical information. Geo-located social media data suffer from a number of biases and potential issues that may be platform-specific [12, 18]. Nonetheless, data from social media allow for first-approximation type of analyses on which the research community can build. Preliminary findings can potentially lead to the design of new surveys and new data collection strategies.

The analysis that we presented in this paper is motivated by the desire to address an important question related to the theory of migration and mobility. Although the perspectives and approaches that we used mainly come from the toolbox of demographers and geographers, our work resonates well with recent developments in the area of social media analysis and is likely to provide further momentum to a growing field that addresses modeling and understanding human mobility using social media data [13, 14].

Some methods that we used could be refined and the sensitivity of our results to different modeling choices could be tested. For example, we inferred a user's home location based on the modal location of tweets over the course of a defined period of time, which is arguably the most obvious choice. However, one could also use other metrics that include the number of distinct days in a location, the day of the week, and the time of the day [12].

We think of this study as a further step towards a better understanding of human mobility at various temporal and spatial scales. We believe that the research community working with social media data has an unprecedented opportunity to build a unified framework for mobility and migration. This would lead to improved understanding of human movements as well as translate into useful tools for data harmonization that would be very relevant for national statistical offices.

ACKNOWLEDGMENTS

Partial support for this research came from a Eunice Kennedy Shriver National Institute of Child Health and Human Development research infrastructure grant, No.: R24 HD042828, and a Shanahan Endowment Fellowship and a Eunice Kennedy Shriver National Institute of Child Health and Human Development training grant, No.: T32 HD007543, to the Center for Studies in Demography and Ecology at the University of Washington.

REFERENCES

[1] Guy J Abel and Nikola Sander. 2014. Quantifying global international migration flows. *Science* 343, 6178 (2014), 1520–1522.

[2] Duygu Balcan, Vittoria Colizza, Bruno Gonçalves, Hao Hu, José J Ramasco, and Alessandro Vespignani. 2009. Multiscale mobility networks and the spatial spreading of infectious diseases. *Proceedings of the National Academy of Sciences* 106, 51 (2009), 21484–21489.

[3] Stuart Basten, Tomáš Sobotka, Krystof Zeman, M Jalal Abassi-Shavazi, Alicia Adsera, Jan Van Bavel, Caroline Berghammer, Minja Kim Choe, Tomas Frejka, Henri Leridon, et al. 2014. Future fertility in low fertility countries. (2014).

[4] Fang Cai. 2011. Hukou system reform and unification of rural–urban social welfare. *China & World Economy* 19, 3 (2011), 33–48.

[5] European Commission. 2007. Regulation (EC) No 862/2007 of the European Parliament and of the Council of 11 July 2007 on Community statistics on migration and international protection. (2007). http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2007:199:0023:0029:EN:PDF.

[6] Daniel Courgeau. 1973. Migrants et migrations. *Population (french edition)* (1973), 95–129.

[7] A Garbero and E Pamuk. 2014. Future mortality in high mortality countries. (2014).

[8] Michael F Goodchild. 2007. Citizens as sensors: the world of volunteered geography. *GeoJournal* 69, 4 (2007), 211–221.

[9] Brian D Gushulak and Douglas W MacPherson. 2004. Globalization of infectious diseases: the impact of migration. *Clinical Infectious Diseases* 38, 12 (2004), 1742–1748.

[10] Randall Hansen and Shalini Randeria. 2016. Tensions of refugee politics in Europe. *Science* 353, 6303 (2016), 994–995.

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

[12] Brent Hecht and Monica Stephens. 2014. A Tale of Cities: Urban Biases in Volunteered Geographic Information. In *ICWSM*.

[13] Andrea Hess, Karin Anna Hummel, Wilfried N. Gansterer, and Günter Haring. 2015. Data-driven Human Mobility Modeling: A Survey and Engineering Guidance for Mobile Networking. *ACM Comput. Surv.* 48, 3 (Dec. 2015), 38:1–38:39.

[14] Raja Jurdak, Kun Zhao, Jiajun Liu, Maurice AbouJaoude, Mark Cameron, and David Newth. 2015. Understanding human mobility from Twitter. *PloS one* 10, 7 (2015), e0131469.

[15] Kenneth C Land. 1969. Duration of residence and prospective migration: Further evidence. *Demography* 6, 2 (1969), 133–140.

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

[17] Yu Liu, Zhengwei Sui, Chaogui Kang, and Yong Gao. 2014. Uncovering patterns of inter-urban trip and spatial interaction from social media check-in data. *PloS one* 9, 1 (2014), e86026.

[18] Momin Malik, Hemank Lamba, Constantine Nakos, and Jürgen Pfeffer. 2015. Population bias in geotagged tweets. In *ICWSM Workshop on Standards and Practices in Large-Scale Social Media Research*. 18–27.

[19] Douglas S Massey, Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J Edward Taylor. 1993. Theories of international migration: A review and appraisal. *Population and development review* (1993), 431–466.

[20] Beata Nowok and Frans Willekens. 2011. A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place* 17, 5 (2011), 521–533.

[21] Philip H Rees. 1977. The measurement of migration, from census data and other sources. *Environment and Planning A* 9, 3 (1977), 247–272.

[22] Andrei Rogers, James Raymer, and K. Bruce Newbold. 2003. Reconciling and translating migration data collected over time intervals of differing widths. *The Annals of Regional Science* 37, 4 (2003), 581–601.

[23] Peter A. Rogerson. 1990. Migration analysis using data with time intervals of differing widths. *Papers of the Regional Science Association* 68, 1 (1990), 97–106.

[24] Susheela Singh and John Casterline. 1985. The socio-economic determinants of fertility. (1985).

[25] J Edward Taylor and Philip L Martin. 2001. Human capital: Migration and rural population change. *Handbook of agricultural economics* 1 (2001), 457–511.

[26] Xavier Thierry, Anne Herm, Dorota Kupiszewska, Beata Nowok, and Michel Poulain. 2005. How the UN recommendations and the forthcoming EU regulation on international migration statistics are fulfilled in the 25 EU countries. In *XXV International Population Conference, Tours*. 18–23.

[27] Kunniparampil Curien Zachariah, My T Vu, et al. 1988. *World Population Projections: 1987*. JSTOR.

[28] Emilio Zagheni, Venkata Rama Kiran Garimella, Ingmar Weber, et al. 2014. Inferring international and internal migration patterns from twitter data. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 439–444.