Migration and Social Networks: New Insights from Novel Data

by

Guanghua Chi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Information Management and Systems

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Joshua E. Blumenstock, Chair
Professor John Chuang
Professor Marta C. Gonzalez

Spring 2020

Migration and Social Networks: New Insights from Novel Data

Abstract

Migration and Social Networks: New Insights from Novel Data

by

Guanghua Chi

Doctor of Philosophy in Information Management and Systems

University of California, Berkeley

Professor Joshua E. Blumenstock, Chair

Migrants play a central role in the economy and society of most developing countries and are primary drivers of economic mobility among poor and rural households. The decision to migrate is one of the most important economic decisions an individual can make. On the one hand, social networks play a crucial role in influencing people's migration decision. On the other hand, as migrants adapt to a new environment, their social network evolves. My research seeks to shed light on the influence of social networks on migration, as well as the influence of migration on social networks. This dissertation answers three questions on migration and social networks using large-scale social network data: (1) What are the roles of migrants in connecting global social networks? (2) How do social networks affect people's decision to migrate? (3) How do migrants' social networks evolve over the migration process?

In the first chapter, I explore in detail the relationship between international social ties and global migration. Social ties form the bedrock of the global economy and international political order. Yet prior empirical studies have been constrained by a lack of granular data on the interconnections between individuals. In this study, using several billion domestic and international Facebook friendships, I find that long-term migration accounts for roughly 83% of international ties on Facebook. By computing the average shortest path length in a social graph with and without migrants, I find that migrants effectively decrease the length of the average shortest path, and act as conduits for more shortest paths than non-migrants.

The second chapter studies how social networks influence an individual's decision to migrate. Two distinct mechanisms through which social networks provide utility to migrants are disambiguated: first, that networks provide migrants with access to information, for instance about jobs and conditions in the destination; and second, that networks act as a safety net for migrants by providing material or social support. I use a massive 'digital trace' dataset to link the migration decisions of millions of individuals to the topological structure of their social networks. The main analysis indicates that the average migrant derives more utility

from 'interconnected' networks that provide social support than from 'extensive' networks that efficiently transmit information.

In the third chapter, I develop and validate a novel and general approach to detecting migration events in trace data. The most common 'frequency-based' approach to inferring migration events often results in mis-classifications. The novel approach accurately classifies migrations, and also provides more granular insight into migration spells and types than what are captured in standard survey instruments.

The fourth chapter examines how migrants' social networks change over the migration and settlement process based on the migration events and dates that were detected in the third chapter. I characterize changes in network structure before and after migration by observing the evolving social networks of a nation's worth of migrants. I find stark and systematic changes in this structure: within two months of migrating, migrants cease communication with nearly half of their former contacts in their place of origin; these 'lost' relationships are almost exactly offset by the 55% increase in new connections with people in the destination. I also show that friendship persistence and loss is highly predictable: the social ties most likely to persist are those that have frequent communication.

As a whole, the chapters in this dissertation develop methods and theories to understand the interaction between migration and social networks. It lays the groundwork for future researchers answering questions in migration and social networks using population-scale digital trace data.

Dedicated to:

my parents, who give me strength,
Jianxi Chi and Yufeng Zhao;

and

my brother, who guides me along the way,
Guangqing Chi.

Contents

Acknowledgments

I would like first to express my deepest gratitude to my advisor and committee chair Joshua Blumenstock for his unwavering support throughout my graduate study. My dissertation has benefited immensely from his valuable advice and guidance. I am also extremely grateful to my committee members, John Chuang, David Bamman, and Marta Gonzalez, for their insightful comments and constructive advice.

I had great pleasure of working with my collaborators Bogdan State, Lada Adamic, Xu Tan, and Fengyang Lin. Bogdan was my intern mentor at Facebook for two summers. He provided me the opportunity to work with the real-world problems and apply my knowledge to make a difference in the world.

I would also like to acknowledge my colleagues Raza Khan, Niall Keleher, Robert On, and Ott Toomet for the discussions from the DataLab to the Data-Intensive Development Lab.

Finally, my dissertation would not have been possible without the continuous support of my family. I want to thank my wife, Xue, for her encouragement and love throughout this journey. My daughter, Aria, brings me so much joy and keeps me motivated to my new chapter. A special thank you to my parents and my brother for their love and support throughout my life.

It is 9,517 km from my elementary school to graduate school, which took me 25 years to finish. Not fast, but I made it.

Chapter 1

Migrants in Connecting Global Social Networks

1.1 Abstract

Social ties form the bedrock of the global economy and international political order. Understanding the nature of these ties is thus a focus of social science research in fields including economics, sociology, political science, geography, and demography. Yet prior empirical studies have been constrained by a lack of granular data on the interconnections between individuals; most existing work instead uses indirect proxies for international ties such as levels of international trade or air passenger data. In this study, using several billion domestic and international Facebook friendships, we explore in detail the relationship between international social ties and human mobility. Our findings suggest that long-term migration accounts for roughly 83% of international ties on Facebook. Migrants play a critical role in bridging international social networks.[1]

1.2 Introduction

Social connections between individuals in different countries provide a foundation for international trade and commerce, and for global peace and cooperation (Hollis & Smith, 1990; Rauch, 2001). A rich literature documents *how* the world is connected, examining the nature, determinants and consequences of social connections between countries. While early studies relied heavily on customs data, foreign direct investment accounts, and international trade data (Feenstra, 2015), more recent research has integrated data from online sources such as messaging applications and social media sites (Garcia-Gavilanes et al., 2014; Leskovec & Horvitz, 2008; State et al., 2015). Much less is known about *who* connects the world, and

[1]The material in this chapter is based on joint work with Bogdan State, Joshua Blumenstock, and Lada Adamic. Who Ties the World Together? Evidence from a Large Online Social Network. See: Chi et al. (2019)

how micro connections affect macro network structure. Understanding how the world is connected has practical value, as it can provide a starting point for scholars and policy makers who seek to understand international relations from a network perspective (Hafner-Burton et al., 2009), including, for instance, work on the importance of network brokerage (see R. Burt, 2004). More generally, a better understanding of this transition from the individual to the transnational comes to address the *micro-to-macro* problem identified by Coleman (1994) as the fundamental challenge on the path to a science of society.

This study uses Facebook data to provide a disaggregated understanding of the network connections of migrants and non-migrants on one of the world's largest social networks. The Facebook dataset allows for a high-level view of the demographic characteristics and network structures of the world's "international brokers," i.e., the people whose social ties quite literally connect the world. This allows us to ask the central question of our study: who ties the world together?

We present three main results. First, we provide empirical evidence that migrants are a central binding force in the global social network. The act of migration reshapes the network by transforming domestic ties to international ones. The friends they made prior to their move now all know someone who lives in a different country. At the same time, the friends they make in the new country now potentially have a new international tie. These friends now know someone who is *from* another country. With such potential to convert or generate new international ties, it is perhaps unsurprising that over 83% of all international ties involve migrants. These results are consistent with macro-level analyses performed by Perkins and Neumayer (2013), who found migrants to play an important role in international communication networks.

Second, we find that migrants act as a bridging force that shrinks the network distance between other people in the Facebook social graph. This is evident in simple descriptive statistics: migrants have higher betwenness in the Facebook graph, particularly when considering connections across countries. We also run simulations that compare the approximate average shortest path length in two graphs: one containing only ties between non-migrants, and one both locals and migrants. Despite our increasing the number of nodes in the graph, we find that the average shortest path length decreases when migrants are included. Both results emphasize the bridging role of networks in connecting distant sub-networks.

Finally, we expand our analysis to the characteristics of migrants and their *local* social networks, to better understand the role that migrants play in their immediate network neighborhood. We establish that migrants' ego networks have fewer dense cores, and that migrants tend to occupy a less redundant position in their ego network, leading us to the conclusion that migrants are also more likely to act as local network bridges. Taken together, these results emphasize the important role that international migrants play in binding together global communities.

1.3 Related Work

A varied literature has examined social connections between countries. We distinguish between three main areas of research: urban networks, online social networks, and research on international migration.

Traditional international network analysis has focused on understanding urban networks using aggregated datasets such as flight passenger flows, telecommunication volume, and corporate organization (Derudder, 2006; Short et al., 1996). Airline passenger flows have been used to proxy international human flows across urban networks, under the assumption that important cities receive more airline passengers. Common inter-airport passenger flow datasets have been extracted from the International Civil Aviation Organization (ICAO) (Kyoung-Ho & Timberlake, 2000; Short et al., 1996) and Marketing Information Data Transfer (MIDT) (Derudder & Witlox, 2005; Derudder et al., 2007), which have been used to rank key cities in Western Europe and North America (Derudder & Witlox, 2005; D. J. Keeling, 1995; Short et al., 1996), find global hierarchical structures (Smith & Timberlake, 2001; Zook & Brunn, 2005), and detect temporal changes of a city's importance in the global city network (Matsumoto, 2004; Smith & Timberlake, 2001) by adopting network analysis methods. Derudder and Witlox (2008) pointed out several limitations posed by the use of airline passenger flow data, including the lack of origin and destination information because of stopovers, missing inter-state flow, and possible flows to tourist destinations. In spite of these issues, airline passenger flows remain the most commonly used data source to analyze international urban networks.

Internet backbone networks can also reflect the role of cities and the connections between countries, under the assumption that important cities would have more high-speed internet connections and more connections to other cities (Barnett, 2016; Barnett et al., 1996; Malecki, 2002; Townsend, 2001). This assumption is often untenable, however. A small city may act as a gateway between core cities and its centrality in the internet backbone network may exaggerate its importance in the worldwide social system (Rutherford et al., 2004). Another traditional dataset comes from the realm of multinational corporate organization. International business companies create new offices globally to distribute their service for their corporate benefits. The transnational network formed by international offices captures the information flow and products flow (Beaverstock et al., 2000). The use of this dataset comes with its own limitations, given that transnational flows are inferred instead of directly obtained like airline passenger flows (Derudder & Witlox, 2008).

In recent years, the growing availability of large social datasets has enabled a new, fine-grained level for the understanding transnational social networks, thanks to increases in Internet penetration and the development of global social networking platforms, such as Microsoft Messenger instant-messaging system (Leskovec & Horvitz, 2008), Twitter (Garcia-Gavilanes et al., 2014; Leetaru et al., 2013; Takhteyev et al., 2012), Flickr (Cha et al., 2009), and Facebook (Bailey et al., 2018; Ugander et al., 2011). Network structures are analyzed to understand the properties of social networks, including degree distribution, clustering, the small-world effect, and homophily (Backstrom et al., 2012; Onnela et al., 2007; Travers &

Milgram, 1967). For example, Backstrom et al. (2012) found that the degree of separation is 3.74 based on 721 million people at Facebook in 2011. The most recent result is 3.6 degrees of separation in 2016, showing that people have grown more interconnected (Bhagat et al., 2016).

There has been growing interest in combining spatial and social network analyses to understand the relationship between social networks and migration (adams jimi et al., 2012; J. Blumenstock et al., 2019; Cho et al., 2011; Liu et al., 2015). International and internal migration patterns have been explored using different sources of new datasets, such as geo-tagged tweets (Hawelka et al., 2014; State et al., 2015), IP geo-location (State et al., 2014; Zagheni & Weber, 2012), and social network profile fields (Herdağdelen et al., 2016). This research has focused on the factors related to international social networks and migration, including distance and trade, community structure, and interactions across countries. In this line of work, three recent papers are most relevant to this study. Kikas et al. (2015) found that social network features can explain international migration in terms of net migration per country and migration flow between a pair of countries. Herdağdelen et al. (2016) analyzed the social networks of migrants in the United States by leveraging profile self-reports of home countries. Zagheni et al. (2017) showed the viability of conducting demographic research related to international migration through the public Facebook advertising API.

Our research comes to extend the study of international social networks using online data, shifting the focus from the country-to-country to the individuals whose social connections span the boundaries of countries and who quite literally connect the world. We develop a vocabulary to describe social ties in terms of both parties' home and current countries, which we use to provide an examination of both triads and ego networks. Our analysis concludes with a foray into the role of migrants with regard to the connectivity of the global Facebook social graph.

1.4 Data and Methods

Our analysis makes use of de-identified profile and social connection data available on Facebook, presently the world's largest social networking platform, which as of the time of writing numbered more than 2.25 billion monthly active users. These data have several key limitations: the population of Facebook users is not representative, particularly outside of the U.S. and Western Europe; the connections observed on Facebook are a biased sample of actual social connections; and the data are not broadly accessible to the research community (boyd danah & Crawford, 2012; Mellon & Prosser, 2017). Yet the ability to observe the social connections between such a substantial fraction of the world's population also provides unique advantages for social and demographic research.

We use the Facebook data to simultaneously observe social network structure and migration status for the full population of Facebook users (where available through profile self-reports) in 2018. Each active user represents a node in the network; two nodes are

connected by an edge if they have mutually agreed to be 'friends' on the online platform. Example subnetworks are depicted later, in Fig. 1.3.

Separately, we use de-identified Facebook profile information to determine the current and origin country of each user. The country of origin is determined by the self-reported "home town" that users enter on their profile pages. The current country assignment is determined by Facebook for growth accounting purposes, and is based on typical country-level geolocation signals, such as recent IP addresses. There is a considerable amount of measurement error in this approach to inferring migration, as how people report their "home" town is the result of subjective interpretation. While we do not think this measurement error entirely undermines the high-level analysis that we present in this paper, such data may not be well-suited to more disaggregated analysis, or seen as a substittue for official statistics.

By aggregating home and current country of users we were able to generate a migrant stock dataset, showing the current numbers of individuals "from" one country who currently live in another country. We validated the country-to-country dataset we generated against data on international migrant stocks provided by the World Bank (Ratha, 2016). Here we chose those countries with more than 1 million monthly active users, and those country pairs with more than 0.001% of migrants. The magnitude of migrant stocks quantified using Facebook data is highly (though not perfectly) correlated to migrant stock estimates produced by the World Bank (Pearson's ρ: 0.87), which is similar to the findings of Zagheni et al. (2017). Because migration events may be short-lived (e.g. study abroad or volunteer programs) for young adults, we focus our analysis on users aged over 30 at the time of our study.

1.5 Results

Migrants tie the world together

Our first set of results highlight the substantial fraction of international ties on Facebook that are comprised by migrants. Formally, we denote the home and current country of a person i by H_i and C_i, and say that i is a migrant if $H_i \neq C_i$. A social tie exists between i and j if they are friends on Facebook. International ties exist if i and j have different current countries ($C_i \neq C_j$) or different home countries ($H_i \neq H_j$).

A striking result is evident when we look at the fraction of international and domestic ties that involve migrants. While only 17.1% of all ties on Facebook involve a migrant, a staggering 82.91% of interantional ties involve at least one migrant. These results are presented and disaggregated in Table 1.1.

Of the interantional ties we observe, 39.4% exist between migrants and locals in destination countries, and 27.88% of international ties connect migrants with people in the country of origin.

Only 17.09% of all international ties in our sample are between non-migrants – individuals in different countries whose own current countries are the same as their stated home countries.

Table 1.1: Domestic and international ties (univariate statistics)

	International Ties (%)	Domestic Ties (%)	All Ties (%)
Non-migrants	17.09	99.14	82.90
Migrants	82.91	0.86	17.10
... Two migrants	7.66	0.86	2.21
... Migrant to a resident in the destination country	39.40	0	7.79
... Migrant to a resident in the origin country	27.88	0	5.52
... Migrant to a resident in other countries	7.97	0	1.58

This leads to the staggering conclusion that international migration is responsible for over 83% of social ties between countries. Even this statistic may underestimate the percentage of international ties due to migration, given that our analysis does not account for return migration – i.e., the situation in which an individual has returned to their country of origin but maintains ties in their former migrant destination.

Further strengthening the conclusion regarding the crucial role migrants play in providing international ties is Fig. 1.1, which shows that the distribution of the per-individual proportion of international ties is bimodal, comprised of a mixture of migrants, who have a high concentration of international ties (the average migrant's network contains 90.5% international ties), and non-migrants, whose social networks are dominated by domestic ties (only 10% of their ties are international).

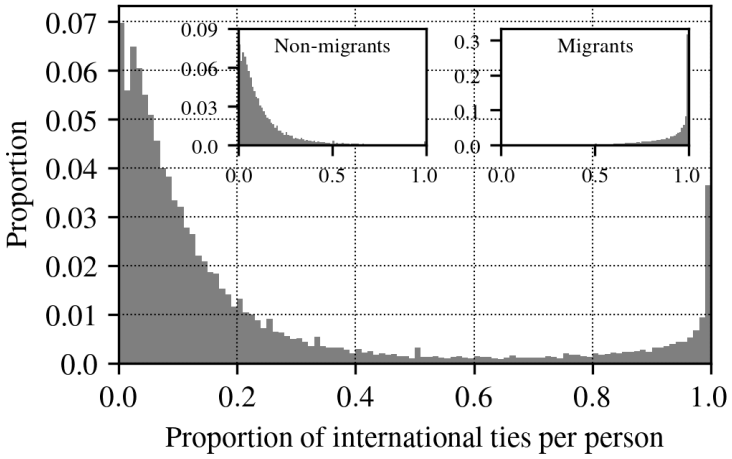
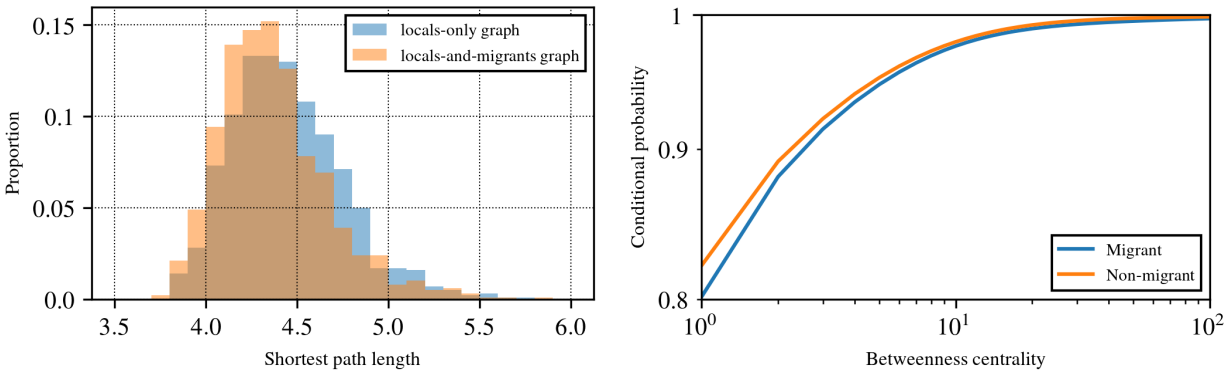


Figure 1.1: Proportion of ties that are international.

Migrants and measures of global cohesiveness

Our second set of results investigate the extent to which migrants play a binding role in the global social network. Here we reproduce the approximation of the average shortest-path computed by Bhagat et al. (2016) and Backstrom et al. (2012), using two graphs as input. The *locals-only* graph, only contains those users for whom the home country is the same as the current country. The *locals-and-migrants* graph results from adding migrants (users with known different home and current countries) to the *locals-only* graph. We sample 1000 seed nodes in each graph to compute the approximate average shortest path using the methodology described in Bhagat et al. (2016). It should be noted that the approximate average shortest path length from these two graphs is not directly comparable to previous results about the entire Facebook social graph, since home-country self-reports are only available for a fraction of Facebook users. We found that the average shortest path length is 4.45 for the *locals-only* graph, and 4.37 for *locals-and-migrants* graph (Fig. 1.2a). In other words, the degree of separation is 3.45 in the *locals-only* graph, and 3.37 in the *locals-and-migrants* graph. A two sample t-test confirms that this difference is statistically significant ($p < 0.001$). Even though there are more nodes in the *locals-and-migrants* graph than the *locals-only* graph, the average shortest path in the *locals-and-migrants* graph is smaller, meaning that the migrants serve as a bridge to bring the world together.



(a) Shortest path length in the *locals-only* graph vs. the *locals-and-migrants* graph.

(b) Betweeenness centrality distribution of migrants vs. non-migrants.

Figure 1.2: Bridging role of migrants in international social networks

In addition to measuring the shrinkage in the global Facebook graph when migrants are added, it is also possible to compute the number of shortest paths which would be routed through migrants and non-migrants when a social search is performed. To this end, we compute weighted approximate betweenness centrality: starting from 24 randomly-selected seeds we compute shortest paths to all nodes in the Facebook social graph (friendships of monthly active users). We then count the number of shortest paths passing through each

vertex in the graph, weighted so that the weights of multiple shortest paths connecting any two vertices all sum to 1. Betweenness statistics for migrants and non-migrants are shown in Table 1.2, suggesting that migrants have higher betweenness despite having lower degree. To better understand what drives this dynamic we plot cumulative distribution function for migrants' and locals' betweenness centrality in Fig. 1.2b. The figure shows that migrants are over-represented among individuals with very high betweenness compared to locals.

Table 1.2: Betweenness centrality statistics for migrants (M) and locals (L).

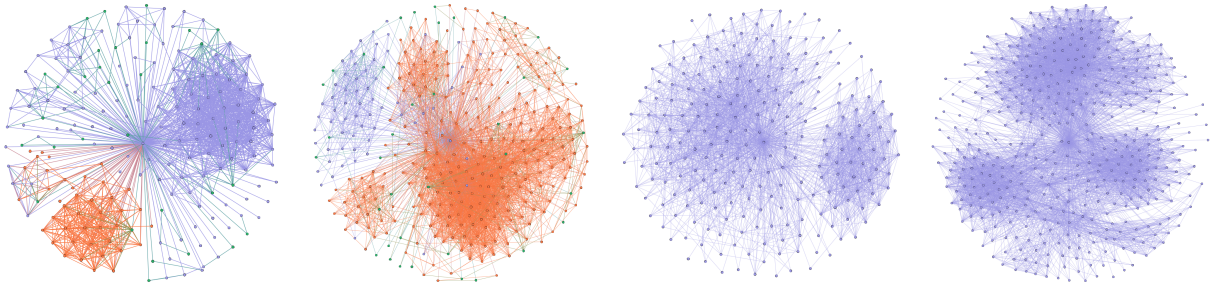
Statistic		Mean	S.D.	Median
Betweenness	M	8.12	25302.26	1.07
	L	7.66	69286.75	1.04
... same	M	45.95	90612.70	1.26
... country	L	79.99	305134.88	1.08
... different	M	6.25	16219.46	1.07
... country	L	3.79	8400.1	1.04
Degree	M	372	513	214
	L	395	544	244

While the majority of both migrants and locals have relatively low betweenness, there are more migrants among those who act as conduits for many of the shortest paths in the Facebook social graph. To better understand the role that migrants play in brokering international ties we can also distinguish between situations where ego and the seed are in the same country or in different countries. When making this distinction we can see in Table 1.2 that, among users in a different country than the seed, migrants help route almost twice as many (6.25) shortest paths as locals (3.79), whereas migrants only route about half as many shortest paths (45.95) as locals (79.99) to a seed in the same current country. This further seems to suggest that migrants have a particularly important role in providing inter-country connectivity: they not only participate in a great number of international ties but their ties are also more likely to function as international network bridges.

Ego-networks

We have seen so far that migrants have more international ties, and that they play an over-size role in improving connectivity in the global social graph. A natural question arises as to whether migrants' *local* networks differ in other structurally meaningful ways from those of non-migrants. The analysis of ego-networks can help establish the extent to which individuals help connect disjoint collections of alters, providing important measures of network brokerage. Fig. 1.3 shows four example ego networks, two of migrants and two of non-migrants, with violet nodes and edges indicating connections in the current country and orange nodes and edges representing connections in the home country. We can see that the

two migrants' home and current country networks are disjoint, with no direct connection between alters in the home and current country. In this case the migrant ego provides a shortest path between each pair of alters in the home and current country, respectively.



(a) Migrant A (b) Migrant B (c) Non-migrant A (d) Non-migrant B

Figure 1.3: Ego networks of two migrants and two non-migrants. *Note*: The center node is the ego. All the other nodes are his or her friends. The node color refers to different countries: orange nodes are living in the ego's home country; violet nodes are living in the ego's current country; green nodes are living in other countries.

To measure the ego-networks of users we measure multiple statistics:

- size of ego network, i.e. a user's number of Facebook friends (alters).
- ego's clustering coefficient, or the proportion of triads ego participates in that are closed.
- k-cores, or the maximal subgraph of the ego graph, in which nodes have degree of at least k. We compute k-cores for all possible k's in the ego-network.

Given the computational requirements of the analysis, running it for all users would be prohibitively expensive. Because we are interested in the structural differences between migrants and non-migrants, we chose to run an analysis on a balanced sample of users. We analyzed a sample of 20,000 users (10,000 migrants and 10,000 non-migrants) drawn at random from among monthly active Facebook users aged between 30 and 80. Ego-network statistics were computed for the entire ego-graph, as well as for two subgraphs: the graph of all users who share their current country (G_C), and the graph of all users who share their home country (G_H). As Table 1.3 reveals, migrants appear to have slightly lower degree than locals. On average, a migrant in our sample had 373 Facebook friends, whereas a local had 388 Facebook friends, this difference being statistically significant at the 0.05 level ($p = 0.04$ using a two-sample t-test).

Migrants were also comparatively less connected to their home and current countries than locals. On average, the home ego-network G_{Hi} of a migrant i – composed of people with

Table 1.3: Ego-network statistics for migrants (M) and locals (L). *Note*: G_H is the graph of all users who share their home country. G_C is the graph of all users who share their current country.

		Whole		G_H		G_C	
Statistic		Mean	S.D.	Mean	S.D.	Mean	S.D.
Degree	M	373	517	129	228	160	312
	L	388	533	255	358	352	491
	p-val.	0.04		< 0.01		< 0.01	
Density	M	0.120	0.134	0.247	0.248	0.209	0.206
	L	0.118	0.119	0.139	0.136	0.126	0.127
	p-val.	0.19		< 0.01		< 0.01	
8-core	M	0.865	0.553	0.462	0.557	0.498	0.548
	L	0.871	0.512	0.732	0.561	0.839	0.515
	p-val.	0.38		< 0.01		< 0.01	
64-core	M	0.070	0.256	0.014	0.116	0.025	0.157
	L	0.077	0.267	0.041	0.198	0.067	0.251
	p-val.	0.07		< 0.01		< 0.01	

the same stated home country as the ego – had 129 nodes, whereas the home ego-network G_{Hj} of a local j had 255 nodes. Similarly, the ego-network in the current country G_{Ci} of a migrant i had a mean of 160 nodes, whereas the ego-network in the current country G_{Cj} of a local j had 352 nodes. Given that their ego networks are split between home and current country, it is not surprising that migrants have fewer alters to draw on in each country. These alters are more likely to be connected to one another however: migrants' home-country ego networks have a density of .247, compared to .139 for locals. The same numbers are reflected when G_{Ci} are considered: .209 for migrants and .126 for locals. This result would seem to suggest that migrants' home and current countries are more cohesive than non-migrants, but one has to consider the fact that degree and clustering coefficient have been found to be inversely correlated (Jacobs et al., 2015; Leskovec et al., 2008; Leskovec & Horvitz, 2008). That is, it is possible that migrants have different network foci split between home and current country, whereas all of a local's foci will be in their current country. For instance, a migrant who leaves after high school to attend university in a different country may have one high school friendship group in the home country and another college friendship group in the current country, whereas a local will have both groups in the same country. Even if the two friendship groups have the same density, the migrants' home and current countries will appear to be denser because they only contain their high school and college friendship groups, respectively.

Table 1.3 also reports the average number of 8- and 64-cores in migrants' and locals' ego-networks. A k-core is defined as a subset of nodes in the ego-network network which have a

degree of at least k when connected to one another. These results reveal that migrants have fewer 8- and 64-cores in their home and current country ego networks, while the difference between the number of k-cores in their overall ego networks is much smaller (.865 for migrants vs. .871 for locals for 8-cores, $p = 0.38$ and .070 for migrants vs. .077 for locals for 64-cores, $p = 0.07$). This suggests that migrants ties' are about as clustered as non-migrants', but the cores in their ego-networks are divided between multiple countries. The k-core structure reinforces the multiple country-foci explanation advanced above.

Triadic closure

Beyond the direct connections between two individuals, larger graph structures can provide insight into the role that migrants play in the broader social network. In particular, network *triads* – which indicate whether two friends of an individual are themselves friends – have long been recognized as fundamental elements of social networks irreducible to their parts (Simmel, 1950).

The triadic view poses a more complex challenge due to the exponential increase in complexity resulting from the various combinations possible between the home and current countries of the three actors who participate in a triad. We therefore downsample the Facebook graph to 10% of all monthly active users for whom both home and current country were available. We counted 15bn triads connecting this subset of users

Fig. 1.4 shows a sample of possible triads. The figure suggests that when two people share a friend in common as well as the same home and current country, they are most likely to be friends themselves. People who share neither home nor current country are unlikely to be friends, even if they share a common friend, while friends-of-friends who share either home or current country are moderately likely to be acquainted themselves. Given that triads – and the extent to which they are closed or not – form the building blocks of social networks, we hope that these closure probabilities can be useful to future research efforts into the topology and dynamics of large-scale social networks.

1.6 Conclusion

Both mundane and essential, social ties underpin the global political and economic system. The connection between social networks and globalization has long elicited a great deal of interest among social scientists. Studies of the global social network have only become possible recently, thanks to increases in Internet penetration and the development of global social networking platforms. Increasingly, we can understand international interactions not just through proxies of international flows such as air passenger data and internet bandwidth between countries, but also through the records of connections between people. In this study, to our knowledge the first of its kind at this scale, we focus on the people who connect the world's social network.

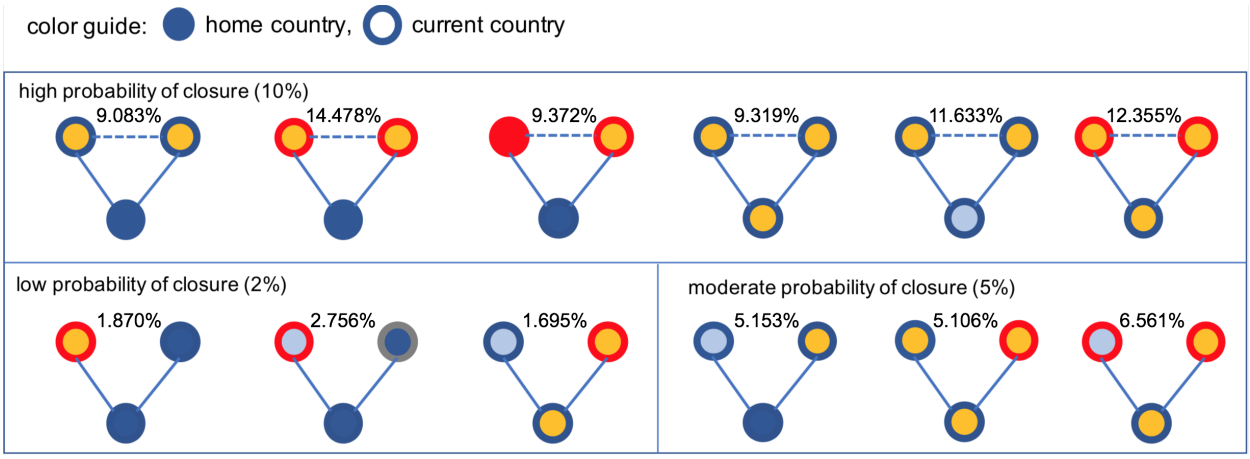


Figure 1.4: Triadic closure probabilities for a sample of triads, illustrating that closure is most likely for migrants sharing home and current country. Each node is an individual, with fill color designating a home country, and the border color designating their current country.

We use an de-identified, aggregated dataset from the Facebook platform to examine the relationship between human mobility and the development of international ties. Our findings suggest that long-term migrations likely account for about 83% of the world's international ties. Our ego network analysis revealed that migrants' networks have higher density, but lower degree, in both home and current countries than non-migrants'.

We also confirmed the "bridging" role of migrants in connecting the world's social network. By computing the average shortest path length in a social graph with and without migrants, we showed that migrants effectively decrease the length of the average shortest path. We also learned that migrants tend to act as conduits for more shortest paths than non-migrants. From these results we can conclude that migrants play an important role in the global economy and society (Lucas, 2015; Todaro, 1980), effectively bringing the world closer together.

We acknowledge the particularly strong tension in network datasets between data privacy and research reproducibility, and hope that both academia and industry will continue working together to find effective ways for sharing large datasets for social science research purposes. To help future researchers with understanding the complex interactions between friendship and international mobility, we have also computed exhaustive triadic closure probabilities between all combinations of migrants and locals. We found that, generally speaking, triads tend to be closed when migrants are present, but only if a current or home country is shared between alters. We hope these aggregations will likewise help advance future social network analysis research, for instance by providing the baseline for simulations.

While this paper has focused on the structure of the network formed by friendship ties between people, there are other types of connections which span the globe. One could ask, for example, what fraction of newspapers' international readership stems from migrants?

For local newspaper readership, do migrants read more international news? Do they share international news with their friends? What role do migrants play in helping artists become globally popular? Since migrants help to make the world just a bit smaller, by stretching their own ties across the globe, it would also be interesting to examine the role of social media in helping to sustain such long-range ties. We leave these and other questions for future work.

Even though much remains to be done until the mechanisms of social networks will be fully understood, the analyses presented in this paper would have been hard to conceive of 50 years ago when Travers and Milgram (1969) performed the first social search experiments. A half century later, it is possible not only to measure the world's connectivity but to ask novel questions of it. We hope that our work will advance scientists' grasp of the social web that envelops the Earth, and of the people who effectively connect the world.

Chapter 2

Migrants and the Value of Social Networks

2.1 Abstract

How do social networks provide utility to migrants? Prior work suggests two distinct mechanisms that have historically been difficult to differentiate: as a conduit of information, and as a source of social and economic support. We use a massive 'digital trace' dataset to link the migration decisions of millions of individuals to the topological structure of their social networks. These data allow us to establish a new set of stylized facts about the relationship between social networks and migration. Our main analysis indicates that the average migrant derives more utility from 'interconnected' networks that provide social support than from 'extensive' networks that efficiently transmit information. We also find evidence of rivalry in information transmission, which suggests that the probability that two people share information is inversely proportional to the (square root of the) size of their social networks.[1]

2.2 Introduction

The decision to migrate is one of the most important economic decisions an individual can make. Many factors influence this decision, from employment prospects and amenity differentials to life-cycle considerations and migration costs. In each of these factors, social networks play a prominent role. It is through social networks that migrants learn about opportunities and conditions in potential destinations; at home, the structure of migrants' social networks shapes their ability and desire to leave.

The central goal of this paper is to better understand exactly *how* social networks influence an individual's decision to migrate, and through the analysis of migration, to provide more general insight into how social networks provide utility. Here, prior work emphasizes

[1]The material in this chapter is based on joint work with Joshua Blumenstock and Xu Tan. Migration and the value of social networks.

two distinct mechanisms: first, that networks provide migrants with access to information, for instance about jobs and conditions in the destination (Borjas, 1992; Dustmann et al., 2016; Munshi, 2003; Topa, 2001); and second, that networks act as a safety net for migrants by providing material or social support (Carrington et al., 1996; Comola & Mendola, 2015; Dolfin & Genicot, 2010; Edin et al., 2003; Munshi, 2014). This distinction between the 'information' and 'social support' value of social networks made in migration literature parallels the contrast between *information capital* and *cooperation capital* made in the theoretical network literature (Jackson, 2018). More broadly, network theory suggests that the utility an individual receives from a social network depends, in part, on the topological structure of the network. Information capital, which reflects the network's ability to efficiently transmit information, is associated with *extensive* subnetworks (e.g., stars and trees) where an individual is linked to many others via short network paths.[2] Cooperation capital is usually motivated by repeated game models of network interaction, where *interconnected* networks (e.g., cliques) best support social reinforcement and sanctioning.[3]

However, there is considerable ambiguity about which types of social capital matter most, and even the nature of each type of social capital in isolation. For instance, the prevailing view in the migration literature is that migrants tend to go to places where they have larger networks, but a handful of studies argue that larger networks may actually deter migration, for instance if migrants compete with one another over opportunities and resources.[4] Similarly, robust risk sharing networks can both facilitate migration by providing informal insurance against negative outcomes (Morten, 2019), and discourage migration if migrants fear those left behind will be sanctioned for their departure (A. V. Banerjee & Newman, 1998; Munshi & Rosenzweig, 2016).

These ambiguities arise because it has historically been difficult to differentiate between distinct sources of social capital in a single empirical setting. In the migration case, linking social network structure to migration decisions is not feasible with traditional data. As Chuang and Schechter (2015) note, "there is little evidence making use of explicit network data on the impact of networks on the initial migration decision... Collecting migration data is quite difficult, and collecting network data is quite difficult; combining the two is even more so" (p.464).[5] Instead, most existing work relies on indirect proxies for a migrant's social

[2]Early models include Kermack and McKendrick (1927) and Jackson and Wolinsky (1996); more recent examples include Calvó-Armengol and Jackson (2004), Jackson and Yariv (2010), and A. Banerjee et al. (2013).

[3]Jackson et al. (2012) and Ali and Miller (2016) provide recent examples. See also Ambrus et al. (2015), Jackson et al. (2012), Ligon and Schechter (2011) and A. G. Chandrasekhar et al. (2018).

[4]Classic papers documenting the 'prevailing' view include M. S. Granovetter (1973), Greenwood (1969), Montgomery (1991), Rees (1966), and Borjas et al. (1992). More recent examples include Bertoli and Ruyssen (2018), Dolfin and Genicot (2010), Fafchamps and Shilpi (2013), Giulietti et al. (2018), Mahajan and Yang (2017), Munshi (2003), Patel and Vella (2012), Winters et al. (2001). Papers that highlight the potential deterrent effect of larger networks include Calvó-Armengol (2004), Calvó-Armengol and Jackson (2004), Wahba and Zenou (2005) and L. A. Beaman (2012).

[5]The difficulty of measuring migration is exacerbated in developing countries, where short-term migration is common (Carletto et al., 2012; Deshingkar & Grimm, 2005; Lucas, 2015; D. J. McKenzie & Sasin, 2007).

network, such as the assumption that individuals from the same hometown, or with similar observable characteristics, are more likely to be connected than two dissimilar individuals.[6] Such proxies provide a reasonable approximation of the size of a migrant's social network, but obscure the higher-order topological network properties that can help disambiguate the mechanism through which social networks provide utility. This higher-order network structure plays a critical role in decisions about employment, education, health, finance, product adoption, and the formation of strategic alliances.[7] Yet, the role of such network structure in migration has not been systematically studied.

We leverage a rich new source of 'digital trace' data to provide a detailed empirical perspective on how social networks influence the decision to migrate. These data capture the entire universe of mobile phone activity in Rwanda over a five-year period. Each of roughly one million individuals is uniquely identified throughout the dataset, and every time they make or receive a phone call, we observe their approximate location, as well as the identity of the person they are talking to. From these data, we can reconstruct each subscriber's 5-year migration trajectory, as well as a detailed picture of their social network before and after migration.[8]

We begin with a reduced form analysis that links each individual's migration decision to the structure of his or her social network several months prior to migration. The purpose of this analysis is to understand whether, ceteris paribus, individuals are more likely to migrate to places where their social networks have particular network topologies (identification is discussed below, and in detail in Section 2.5). A stylized version of our approach is shown in Figure 2.1: we are interested in understanding whether, for instance, individual A is more likely to migrate than individual B, where both A and B know exactly two people in the destination and three people at home, and the only observable difference between A and B is that B's contacts are connected to each other whereas A's contacts are from two disjoint communities.

The reduced form analysis establishes a new set of stylized facts about the relationship between migration and social networks. First, we confirm the longstanding hypothesis that

---

The challenges of measuring social network structure are discussed in Chuang and Schechter (2015) and Breza et al. (2017).

[6]For instance, Munshi (2003) uses rainfall shocks at origin to instrument for network size at destination. L. A. Beaman (2012) exploits exogenous variation in the size of the migrant's social network induced by the quasi-random assignment of political refugees to new communities. Kinnan et al. (2018) take advantage of a resettlement program in China that sent 18 million urban youth to rural areas. Related approaches are used by Card (2001), Hanson and Woodruff (2003) and Dinkelman and Mariotti (2016).

[7]For example: M. S. Granovetter (1973), R. S. Burt (1992), and Karlan et al. (2009) provide examples of how higher-order network structure affects employment prospects. A. Banerjee et al. (2013), L. Beaman et al. (2015), and Ugander et al. (2012) illustrate the importance of higher-order structure in the adoption of microfinance, new plant seeds, and Facebook, respectively. Ambrus et al. (2015) and A. G. Chandrasekhar et al. (2018) relate network structure to contract enforcement and informal insurance. M. J. Keeling and Eames (2005) review how network structure influences the spread of infectious diseases. König et al. (2017) and Jackson and Nei (2015) link political network structure to strategic alliance formation. See Jackson (2010) and Easley and Kleinberg (2010) for an overview.

[8]These data also have several important limitations, which are discussed in detail in Section 2.4.

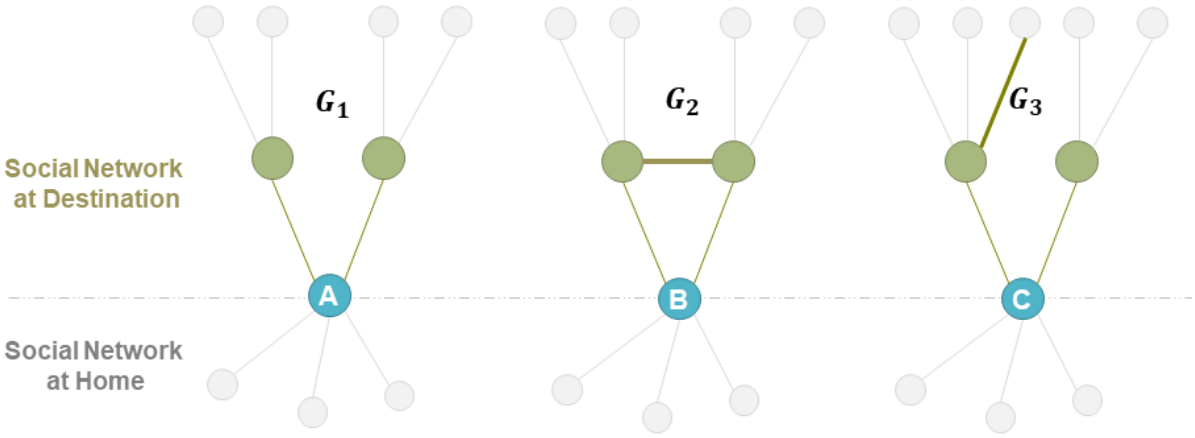


Figure 2.1: Schematic diagrams of the social networks of three migrants. *Notes:* Each of the blue circles (A, B, C) represents a different individual considering migrating from their home to a new destination. Each individual has exactly three contacts in the home district (grey circles below the dashed line) and two contacts in the destination district (green circles above the dashed line). The social network of these three individuals is denoted by G_1, G_2, and G_3.

people move to places where they know more people; conversely, individuals are less likely to leave places where they have larger networks. While these results are expected, an advantage of our setting is that we can observe the nonparametric relationship between migration and network size. We find this relationship to be monotonic and approximately linear with elasticity one, such that the probability of migration roughly doubles as the number of contacts in the destination doubles. Superficially, this result diverges from a series of studies that predict eventual negative externalities from network size, as when members compete for information and opportunities (L. A. Beaman, 2012; Calvó-Armengol, 2004; Calvó-Armengol & Jackson, 2004; Dagnelie et al., 2019). We also find that the probability of leaving home decreases proportional to the size of the home network.

Second, we document, to our knowledge for the first time, the role that higher-order network structure plays in migration decisions. As a proxy for the 'interconnectedness' of the network, we measure the extent to which the individual's local subnetwork is clustered, where a large proportion of neighbors have common friends. As a proxy for the network's 'extensiveness', we measure the size of the individual's distance-2 and distance-3 neighborhood. We find that, conditional on network size, migrants are drawn to locations where their networks are interconnected, but that, on average, they are actually *less* likely to go to places where their networks are extensive — a result that surprised us initially, given the emphasis prior work has placed on the value of connections to socially distant nodes in a network (e.g., M. S. Granovetter, 1973). In other words, of the three potential migrants in

Figure 2.1, B is most likely to migrate and C is least likely, with A somewhere in between.[9]

To better understand this 'surprising' result, we document considerable heterogeneity in the migration response to social network structure. In particular, we find that the negative effect of extensive networks is driven by settings where a migrant's direct contacts have a large number of "strong ties" in the destination (where tie strength is defined by the frequency of communication); when a migrant's destination contacts have many weak ties, migration is not deterred. Such evidence suggests that there may be rivalry in information sharing in networks, which leads migrants to value connections to people for whom there is less competition for attention (as in Dunbar (1998) and A. V. Banerjee et al. (2012)). We also find that while the average migrant is *not* drawn to locations where her friends have more friends (as in G_3), such structure does attract several less common types of migrants. In particular, repeat migrants (who have previously migrated from their home to the destination), long-term migrants, and short-distance migrants — all of whom are presumably better informed about the structure of the destination network — are more likely to migrate to locations where their networks are more extensive.

Building on these reduced-form estimates, our final set of results provide structural insight into the more general question of how people derive value from their social networks. This structure allows us to be more precise about the utility that comes from 'extensive' and 'interconnected' subnetworks, and accounts for more complex network structure than the proxy measures used in the reduced-form analysis. Our model characterizes the migration decision as, ceteris paribus, a tradeoff between the utility an individual receives from the home network and the utility received from a potential destination network, net an idiosyncratic cost of migrating. The focus of the model is on understanding the utility $u_i(G)$ an individual i receives from an arbitrary social network G. We assume that agents derive utility from their networks in two archetypal ways. First, as a source of information capital, where information transmission is modeled as a diffusion process with possible loss of information, as in A. Banerjee et al. (2013). And second, as a source of cooperation capital, where agents engage in repeated cooperation games with their neighbors, as in Jackson et al. (2012) and Ali and Miller (2016).

We estimate this model by maximizing the likelihood of hundreds of thousands of observed migration decisions, and note several results. First, in a departure from benchmark models of diffusion, we find strong support for competition or rivalry in information transmission: a model where information passes from i to j (inversely) proportional to the size of each individual's immediate network fits the data better than standard models where information passes with constant probability. In particular, our results suggest that two people share information with probability roughly inversely proportional to the square root of the (product of the) number of contacts they each have. Our model also allows us to

[9] Our appendices highlight several other empirical regularities between migration and social networks that, to our knowledge, have not been documented — but which are not central to our main analysis and so are only mentioned in passing. For instance, the "pull" of a destination contact is roughly 7 times as strong as the "push" of a home contact; a strong tie is roughly twice as attractive as a weak tie; and recent and co-migrants play an important role in facilitating migration.

decompose the total utility of an agent's network into two components. Consistent with the reduced-form regressions, we find that when information transmission is constrained to be non-rival, most agents receive very little utility from information capital (provided by structures that efficiently diffuse information) relative to cooperation capital (derived from network structures that facilitate repeated cooperation). However, when rivalry is empirically parameterized, information capital and cooperation capital contribute relatively evenly to the migrant's total utility.

Since our approach to studying migration with mobile phone data is new, we devote considerable attention to causal identification, and perform a large number of tests to check the robustness of our results.[10] Perhaps the most important limitation of our approach is that we lack exogenous variation in the structure of an individual's network, so that the social networks we observe are almost certainly endogenous to migration decisions. We address this concern in two principal ways. First, we relate migration decisions in each month to the structure of the social network several months prior in order to minimize the likelihood that the decision to migrate shaped the social network, rather than vice versa.[11] Second, and more important, identification is achieved through an extremely restrictive set of fixed effects that limit the potential for many of the most common sources of endogeneity. Our preferred specification includes fixed effects for each individual migrant (to control for individual heterogeneity, for instance that certain people are both more likely to migrate and to have certain types of networks), fixed effects for each possible origin-destination-month combination (to control for factors that are shared by all people facing the same migration decision, such as wage and amenity differentials), and fixed effects for each possible destination network size (such that comparisons are always between places where the migrant has the exact same number of direct contacts, as in Figure 2.1). Thus, in our preferred specification, the identifying variation comes from within-individual differences in network structure between destinations and over different months in the 5-year window, net the population-average differences that vary by home-destination-month, and net any effects that are common to all people with exactly the same number of friends in the destination. We would observe such variation if, for instance, an individual had been considering a move to a particular destination for several months, but only decided to migrate after his friends in the destination became friends with each other (the G_2 vs. G_1 comparison of Figure 2.1) — and if that tightening of his social network exceeded the average tightening of networks in that destination (as might occur around the holidays, for instance).[12]

To summarize, this paper makes two main contributions. First, it provides a new empir-

[10] Our baseline results assume each individual faces an independent migration decision in each month. She can either stay put, or migrate to one of the 26 other districts in the country of Rwanda. We regress the binary migration decision on (lagged) properties of the migrant's social network, using either a discrete choice (multinomial logit) model or a panel fixed effects specification. Our measurement strategy, these specifications, and the robustness tests are described in detail in Sections 2.4 and 2.5.

[11] One concern is that migrants might begin to strategically reshape their networks long in advance of migrating. We perform several tests to check for such an effect, but find no evidence of anticipatory changes in network structure — see Section 2.5 for an extensive discussion.

[12] In addition to the preferred specification, we perform a series of robustness tests to more precisely isolate

ical perspective on the determinants of migration in developing countries (cf. Lucas, 2015). In this literature, many scholars have noted the important role that social networks play in facilitating migration. Early examples in the economics literature include Rees (1966) and Greenwood (1969); a large number of subsequent studies document the empirical relationship between network size and migration rates.[13] More recently, Munshi and Rosenzweig (2016) document that the fear of losing social network ties may prevent profitable migration, while Morten (2019) shows that the act of migration can change social relationships and risk sharing. Kinnan (2019) theorizes about the two-way inter-connections: migration of one individual can make other network members better off if that individual has a new source of income, but others may be worse if the act of migration improves the outside opportunity for that person or makes it easier to hide income. This paper builds on this line of work by exploiting a new source of data to establish a more nuanced set of stylized facts about networks and migration — highlighting, in particular, the value migrants place on interconnected networks, and substantial heterogeneity in how different types of migrants value networks differently — that have not been documented in prior work.

Second, through the study of migration, we shed light on the more fundamental question of how individuals can derive utility from social networks (cf. A. Banerjee et al., 2013, 2019; Jackson, 2010). Specifically, we use millions of revealed-preference migration decisions to estimate a model of network utility. This allows us to distinguish between the utility provided by network geometries that facilitate the free flow of information from geometries that facilitate repeated cooperation. While the models we test are highly stylized, we hope it can provide a foundation for future work calibrating structural models of network utility with population-scale social network data.

The remainder of the paper is organized as follows. The next section provides a sketch of a model of social capital and migration, which motivates the empirical analysis. Section 2.4 then describes our unique dataset, paying particular attention to how we use it to measure migration and social network structure, and to the limitations of using phone data for such a study. Section 2.5 then discusses our identification strategy, and the assumptions required to make causal inferences about the effect of networks on migration. The reduced form results

the source of identifying variation. In particular, we show the results from regressions that include fixed effects for (a) each individual-month, which isolates the variation between a migrant's potential destinations in a single month; (b) each individual-destination, which isolates variation over time in the structure of an individual's network in a single destination; (c) each individual j in the destination, which removes variation that might be driven by specific destination contacts who are singularly capable of facilitating migration. In these and related cases, the main results are qualitatively unchanged.

[13] Examples include L. A. Beaman (2012), Bertoli et al. (2013), Bertoli and Ruyssen (2018), Borjas et al. (1992), Dolfin and Genicot (2010), D. McKenzie and Rapoport (2010), Montgomery (1991), Munshi (2003), Patel and Vella (2012). Recently, two working papers have used phone data to link spatial mobility and social networks. Büchel et al. (2019) use data from a Swiss cellphone operator to link migration decisions to phone calls, and document patterns similar to the "reduced form" relationship between network size and migration that we note in Section 2.6. Barwick et al. (2019) show that migrant flows in a Chinese city correlate with call volume between regions, and link this information flow to improved labor market outcomes. Both papers focus primarily on how on network size relates to migration, whereas our focus is on the role of network structure, conditional on network size.

and robustness checks are presented in Section 2.6. We use these new stylized facts to put more flesh on a model of network-based social capital, which we develop and estimate in Section 2.7. Section 2.8 concludes.

2.3 A Model of Social Capital and Migration

A central goal of network theory is to understand how the structure of a social network affects the utility that an agent obtains from that network. Our model links social network structure (in both the home and destination) to subsequent migration decisions, to obtain a revealed preference measure of network utility.

Formally, we say that an individual i receives utility $u_i(G)$ from social network G. In deciding whether or not to migrate, the individual weighs the utility of her home network G^h against the utility of the network G^d in the potential destination, and migrates if the difference is greater than an idiosyncratic cost ε_i that can reflect, among other things, wage differentials and i's idiosyncratic costs of migrating.

$$u_i(G^d) > u_i(G^h) + \varepsilon_i. \tag{2.1}$$

How people derive utility from their social networks — and equivalently, how we parameterize $u_i(G)$ — is not known ex ante. The network theory literature links this network-based utility to the topological structure of the underlying network (i.e., to the configuration of connections between nodes in the network). Jackson (2018) summarizes this work, and provides a taxonomy of *social capital* in networks. We focus on two types of social capital that prior studies have emphasized in the decision to migrate: *information capital* and *cooperation capital*.

Information capital. We think of information capital as the potential for the social network to provide access to novel information — about jobs, new opportunities, and the like. Jackson (2018) describes this as the "ability to acquire valuable information and/or spread it to other people through social connections" (p.4). This notion is motivated by a robust theoretical and empirical literature that suggests that the value of a social network stems, at least in part, from its ability to efficiently transmit information (A. Banerjee et al., 2013; Calvó-Armengol & Jackson, 2004; Jackson & Yariv, 2010; Topa, 2001).

The network's ability to transmit information is closely associated with specific network topologies. In particular, efficient information gathering typically requires an *extensive* subnetwork such that one person is linked to many others via short network paths (cf. M. S. Granovetter, 1973). For instance, Jackson and Wolinsky (1996) provide an early measure of information capital as decay centrality, where each agent receives a value $q < 1$ (the probability of information transmission) from each direct friend, a discounted value of q^2 from each friend of friend, and so on. More recently, A. Banerjee et al. (2013) introduce a notion of diffusion centrality, which accounts for the fact that multiple paths could increase the

chance that information makes it from one agent to other. Specifically, agent i's diffusion centrality is the i^{th} element of the vector $DC(G; q, T)$:

$$DC(G; q, T) \equiv \sum_{t=1}^{T} (qG)^t \cdot \mathbf{1}, \tag{2.2}$$

in which the network G is a matrix with $G_{ij} = 1$ if i and j are connected and otherwise $G_{ij} = 0$ (including $G_{ii} = 0$). This measure assumes an information-passing model where, in each period, information is shared with probability q and information is useful if heard within T periods.

In both the decay and diffusion centrality measures, information capital increases with more friends, friends of friends, friends of friends of friends, and so on. Thus, in some of the descriptive analysis that follows, we will initially explore how migration decisions correlate with the size of an individual's second-degree neighborhood (or unique *friends of friends*) and third-degree neighborhood (unique friends of friends of friends). Later, we will develop a structural model of information capital that captures the utility of arbitrarily complex networks.

Cooperation capital. Separately, we consider the cooperation capital of a network to be the network's ability to facilitate interactions that benefit from cooperation and community enforcement, such as risk sharing and social insurance e.g., A. G. Chandrasekhar et al. (2018), Jackson et al. (2012), Ligon and Schechter (2011). This corresponds closely to the notion of favor capital in Jackson (2018), which is described as the network's "ability to exchange favors and transact with others through network position and repeated interaction and reciprocation" (p.4).

Cooperation capital is linked to different network topologies than information capital. In particular, a consistent set of results has shown that such enforcement is strong and cooperation is efficient when local subnetworks are tightly *interconnected*. In particular, Ali and Miller (2016) model a dynamic game of repeated cooperation and find that a clique network (a completely connected network) generates more cooperation and higher average utility than any other networks; Jackson et al. (2012) model a game of repeated favor exchanges and highlight the importance of *supported* relationships, where a link is supported if the two nodes of the link share at least one common neighbor. Related models are cited in footnote 3.

Our initial descriptive analysis thus highlights two related measures of network interconnectedness: *network support*, the probability that a friend has one or more common friends; and *network clustering*, the probability that two friends are connected to each other.

Formally,

$$\text{support}_i(G) \equiv \frac{\#\{j : G_{ij} = 1 \ \& \ (G^2)_{ij} \geq 1\}}{\#\{j : G_{ij} = 1\}} \tag{2.3}$$

$$\text{clustering}_i(G) \equiv \frac{\#\{jk : G_{ij} = G_{ik} = G_{jk} = 1\}}{\#\{jk : G_{ij} = G_{ik} = 1\}} \tag{2.4}$$

Social capital. We make the assumption that the total utility agent i receives from a network G can be expressed as a combination of the information capital u_i^I and cooperation capital u_i^C that i receives from G (we omit G when referring to an arbitrary network):

$$u_i = U(u_i^I, u_i^C). \tag{2.5}$$

We will later develop micro foundations for both u_i^I and u_i^C. That structural analysis is in part motivated by a 'reduced form' analysis that more transparently illustrates how proxy measures of extensiveness (second-degree and third-degree neighborhood size) and interconnectedness (network support and network clustering) correlate with migration decisions. The data and measurement strategy are described in more detail in the following section. Section 2.5 then discusses our identification strategy, and the reduced form results are presented in Section 2.6. The full structural model is developed and estimated in Section 2.7.

Before proceeding, we remark that there are other ways to model information capital and cooperation capital. For instance, in addition to information diffusion, the network literature also examines information aggregation, i.e., agents' ability to form the correct beliefs about the underlying true state, such as whether global warming is true or whether vaccines cause autism. The common prediction regarding network structure for correct learning is that each agent must have a negligible influence on the limit belief (see Golub and Jackson (2010) for myopic learning and Mossel et al. (2015) for Bayesian learning). In the context of migration, we focus on factual information about job openings, housing opportunities and the like, where an information diffusion model seems more natural than an information aggregation model.

Related, in addition to repeated cooperation, the network literature also considers social networks as social collateral for trust and consumption smoothing. Karlan et al. (2009) predict that dense networks generate "bonding social capital" that facilitate valuable cooperation, whereas loose networks generate "bridging social capital" that improves access to information. These two types of social capital are similar to our notion of cooperation capital and information capital. Also related, Ambrus et al. (2015) and Ambrus et al. (2018) study consumption risk sharing with global and local knowledge, respectively. Both find that "expansive" networks (where each subgroup has many links to the rest of the network) and individuals with high eigenvector centrality benefit from consumption risk sharing. We will examine degree centrality as one of the main network measures in our reduced form analysis. But expansiveness and other higher-order measures of centrality depend on global

network structure, which are likely less transparent to the individual than support and clustering, which depend only on the local neighborhood. This is especially true of networks in the potential destination, which are less easily observed by would-be migrants.[14] For these reasons, we focus on the local measures as our main proxies for cooperation capital.

2.4 Data

To study the empirical relationship between networks and migration, we exploit a novel source of data that contains extremely detailed information on the migration histories and evolving social networks of over one million individuals in Rwanda. These data contain the universe of all mobile phone activity that occurred in Rwanda from January 2005 until June 2009. These Call Detail Records (CDR) were obtained from Rwanda's near-monopoly telecommunications company, and contain metadata on every phone call mediated by the mobile phone network. In total, we observe roughly one billion mobile phone calls between roughly one million unique subscribers. For each of these events, we observe a unique identifier for the caller, a unique identifier for the recipient, the date and time of the call, as well as the location of the cellular phone towers through which the call was routed. All personally identifying information is removed from the CDR prior to analysis. In addition, to focus our analysis on individuals rather than businesses, and to remove the potential impact of spammers and call centers, we remove all data involving phone numbers in the 95th percentile or greater of social network size.[15]

This section describes the methods used to observe the structure of each individual's social network over time (Section 2.4), and to extract each individual's complete migration history (Section 2.4). Section 2.4 discusses limitations of these data.

Measuring social network structure with mobile phone data

The mobile phone data allow us to observe all mobile phone calls placed over a 4.5-year period in Rwanda. These pairwise interactions make it possible to reconstruct a detailed measure of the social network structure of each individual in the dataset. As we discuss in greater detail in Section 2.4, this phone-based social network is different from the true underlying social network, but it captures an important dimension of interaction, and the most prominent method for interacting over longer distances (since landlines and email were virtually non-existent in Rwanda at this time). To provide some intuition, the network of a single migrant, in the month before migration, is shown in Figure 2.2. This particular

[14] A. Banerjee et al. (2019) find that within local communities, individuals have good information on (proxies for) their friends' centrality. Several other studies find that people have incomplete information about who their friends are friends with Casciaro (1998), A. Chandrasekhar et al. (2016), Friedkin (1983).

[15] Specifically, we calculate the total degree centrality (i.e., the number of unique contacts) for each phone number in the dataset, for each month. Phone numbers in the 95th percentile of this distribution have roughly 200 unique contacts in a single month. We then remove all incoming and outgoing calls from the dataset that involve those numbers in that month.

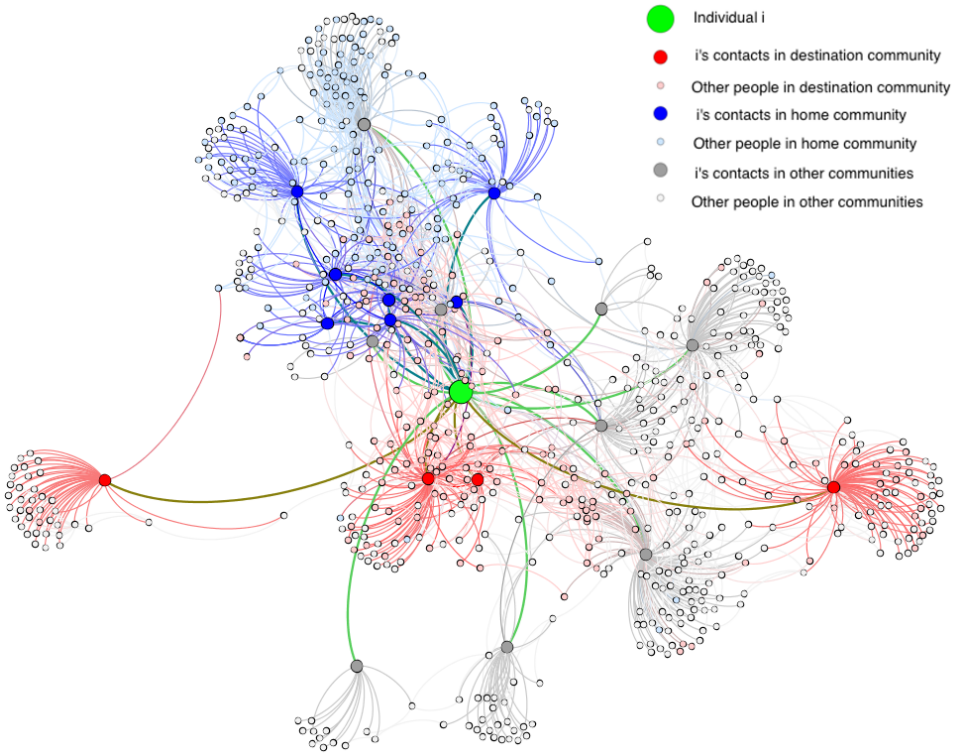


Figure 2.2: The social network of a single migrant. *Notes:* Diagram shows the social network, as inferred from phone records, of a single migrant i. Nodes represent individuals; edges indicate that two individuals communicated in the month prior to i's migration. Direct contacts of i are shown in blue (for people i's home district), red (for people in i's destination district), and solid grey (for people in other districts). Small hollow circles indicate i's "friends of friends," i.e., people who are not direct contacts of i, but who are direct contacts of i's contacts. All individuals within two hops of i are shown. Nodes are spaced using the force-directed algorithm described in Hu (2005).

migrant (the green dot) had 20 unique contacts in the month prior to migration, 7 of whom were in his home district (blue dots), four of whom were in the destination district (red dots), and the remainder were in other districts (grey dots). The large number of friends of friends are also depicted, to provide a sense for the richness of the data.[16]

In the analysis that follows, we relate the network structure of each individual to their subsequent migration decisions. Following the discussion in Section 2.3, we focus on a few statistical properties of networks that prior work suggests are important sources of social capital for migrants. The first is *degree centrality*, which simply counts the number of unique

[16]Throughout, we use the term 'friend' loosely, to refer to the contacts we observe in the mobile phone network. These contacts may be friends, family, business relations, or something else.

individuals with whom each person communicates. This metric most closely reflects the large literature linking migration decisions to the size of an individual's network at the destination (see footnote 4 for classic references). We can separately account for the *strength* of a social tie, which we measure as the number of (undirected) calls between two individuals. In certain analyses we will compare strong and weak ties, where we consider "strong" ties to be those ties in the 90^{th} percentile of the tie strength distribution (equivalent to 5 or more calls per month).[17]

Most importantly, we examine how migration decisions correlate with crude proxies for the information capital u^I and communication capital u^C of a network. We will begin with a reduced form analysis that uses second-degree/third-degree neighborhood size as a measure of network extensiveness (which in turn is a proxy for information capital) and network support/clustering as a measure of interconnectedness (which in turn proxies for cooperation capital). See Section 2.3 for definitions of these metrics. Later, Section 2.7 provides firmer theoretical foundations and a structural approach to measuring u^I and u^C.

Measuring migration with mobile phone data

While fewer than 4% of Rwandan residents are born abroad, internal migration in Rwanda is quite common. According to National Institute of Statistics of Rwanda (2014), 48% of the resident population in urban areas and 14% in rural areas have experienced a lifetime migration. The 1994 genocide and the surrounding conflict were the major cause of internal migration in the 1990s, but since the 2002 census (and including the period that we study from 2005-2009), most migration is driven by economic motives National Institute of Statistics of Rwanda (2014, p.4).

We use mobile phone data to provide rich visibility into the patterns of migration of mobile phone owners in Rwanda. Every time a person uses a mobile phone in Rwanda, the phone company records the time of the event, and the approximate location of the subscriber at the time of the event. We use these logs to reconstruct the migration history of each individual in three steps.

First, we extract the timestamp and cell phone tower identifier corresponding to every phone call and text message made by each individual in the 4.5-year period. This creates a set of tuples {`subscriber_ID, timestamp, tower_ID`} for each subscriber. The tower identifier allows us to approximately resolve the location of the subscriber, to an area of roughly 100 square meters in urban areas and several square kilometers in rural areas. The physical locations of these towers are shown in Figure 2.3. We do not observe the location of subscribers in the time between phone calls and text messages.

Next, we assign each subscriber to a "home" district in each month that she makes one or more transactions. Our goal is to identify the location at which the individual spends the majority of her time, and specifically, the majority of her evening hours.[18] The full details of

[17] By comparison, M. S. Granovetter (1973) defined a weak tie as a tie that was active just once per year.

[18] A simpler approach simply uses the model tower observed for each individual in a given month as the

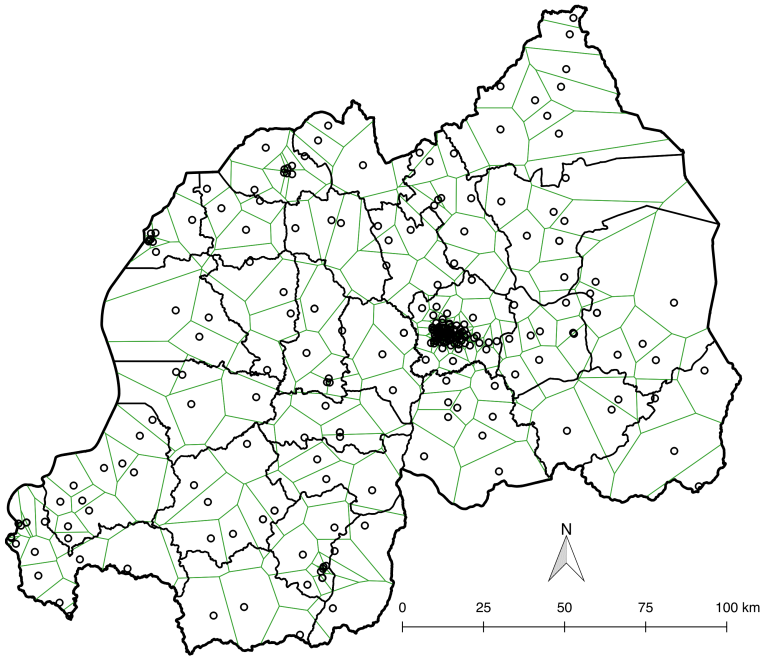


Figure 2.3: Location of all mobile phone towers in Rwanda, circa 2008. *Notes:* Black circles indicate cell tower locations. Black lines represent district borders. Green lines show the voronoi polygons roughly divide the country into the coverage region of each tower.

this assignment procedure are given in Algorithm 1. To summarize, we first assign all towers to a geographic district, of which there are 30 (we treat the three small districts that comprise the capital of Kigali as a single district). Then, for each individual, we compute the most frequently visited district in every hour of the entire dataset (i.e., there will be a maximum of 4.5 years * 365 days * 24 observations for each individual, though in practice most individuals appear in only a fraction of possible hours). We then aggregate these hourly observations, identifying the district where each individual spends the majority of hours of each night (between 6pm and 7am). Finally, we aggregate these daily observations by identifying the district in which the individual spent the majority of nights in each month. The end result is a panel of individual-month districts.[19] After this step, we have an unbalanced panel

"home" location for that person. While our later results do not change if home locations are chosen in this manner, we prefer the algorithm described in the text, as it is less susceptible to biases induced from bursty and irregular communication activities.

[19]At each level of aggregation (first across transactions within an hour, then across hours within a night, then across nights within a month), there may not be a single most frequent district. To resolve such ties, we use the most frequent district at the next highest level of aggregation. For instance, if individual i is observed four times in a particular hour h, twice in district p and twice in q, we assign to i_h whichever of p or q was observed more frequently across all hours in the same night as h. If the tie persists across all hours

indicating the home location of each individual in each month.

Finally, we use the sequence of monthly home locations to determine whether or not each individual i migrated in each month. As in J. E. Blumenstock (2012), we say that a migration occurs in month $t+1$ if three conditions are met: (i) the individual's home location is observed in district d for at least k months prior to (and including) t; (ii) the home location d' in $t+1$ is different from d; and (iii) the individual's new home location is observed in district d' for at least k months after (and including) $t+1$. Individuals whose home location is observed to be in d for at least k months both before and after t are considered residents, or stayers. Individuals who do not meet these conditions are treated as "other" (and are excluded from later analysis).[20] Complete details are given in Algorithm 2.

Using these methods, we are able to characterize very granular patterns of internal migration in Rwanda. Summary statistics are presented in Table 2.1. The first column shows total rates of migration in a single month of the data, using $k = 2$, which defines a migration as an instance where an individual stays in one district for at least 2 months, moves to a new district, and remains in that new district for at least 2 months. The aggregate migration rate in January 2008 is 4.9%; 53.4% of migrants travel from one rural district to another, 23.2% travel from rural to urban districts and 23.4% travel from urban to rural districts.[21]

To validate these methods, Figure A.1 compares the distribution of migration destinations computed from the phone data (red bars) to the distribution of destinations calculated from the 2012 Rwandan census (blue bars), as reported by National Institute of Statistics of Rwanda (2014, p.29). The distributions are not identical, which is expected since the population of phone owners is a non-random sample of Rwandans, but the broad patterns are remarkably consistent across the two approaches to measurement.

While it is reassuring that the aggregate migration rates computed on our data match those reported in traditional surveys, the real advantage of our data is that they can provide a much more granular perspective on internal migration than can be achieved with traditional methods. For instance, the columns of Table A.1 disaggregate migration events into several sub-types that are prominent in the literature on internal migration in developing countries (cf. Lucas, 1997, 2015; Todaro, 1980). We observe a striking number of repeat and circular migrants, with a majority of migrants traveling long distances. The data also make it possible to disaggregate migration rates by length of stay. The rows of Table A.1 show how the implied migration rate decreases as the minimum stay length k is increased. Such comparisons would be difficult with traditional survey data, which typically capture a single definition of migration. In later analysis, we show that certain results depend on this definition. But

on that night, we look at all nights in that month. If a tie persists across all nights, we treat this individual as missing in that particular month.

[20] Individuals are treated as missing in month t if they are not assigned a home location in any of the months $\{t-k, ..., t, t+k\}$, for instance if they do not use their phone in that month or if there is no single modal district for that month. Similarly, individuals are treated as missing in t if the home location changes between $t-k$ and t, or if the home location changes between $t+1$ and $t+k$.

[21] In Table 2.1, we classify the three districts that comprise the capital of Kigali as urban, and the remaining 27 districts as rural.

Table 2.1: Summary statistics of mobile phone metadata

	(1) In a single month (Jan 2008)	(2) Over two years (Jul 2006 - Jun 2008)
Number of unique individuals	432,642	793,791
Number of person-months	432,642	8,121,369
Number of CDR transactions	50,738,365	868,709,410
Number of migrations	21,182	263,208
Number of rural-to-rural migrations	11,316	130,009
Number of rural-to-urban migrations	4,908	66,935
Number of urban-to-rural migrations	4,958	66,264

Notes: Migration statistics calculated from Rwandan mobile phone data. Column (1) based on data from a single month; column (2) includes two years of data, potentially counting each individual more than once. "Migrations" occur when an individual remains in one district for 2 consecutive months and then remains in a different districts for the next 2 consecutive months. We denote as urban the three districts in the capital of Kigali; the remaining districts are considered rural.

unless otherwise noted, our results define migration as a minimum stay length of $k = 2$, as this most closely matches official statistics on internal migration provided by the Rwandan government.[22]

Data limitations

While mobile phone data provide uniquely granular insight into the migration decisions and social networks of a large population, there are several important limitations. First, mobile subscribers are not representative of the larger population; in particular, they are wealthier, older, better educated, and are more likely to be male (J. E. Blumenstock & Eagle, 2012).[23]

While this certainly limits the external validity of our analysis, as we have noted above (and show with Figure A.1 and Table A.1), the patterns of migration inferred from phone

[22] According to the 2012 census: 9% of Rwandans are live in a place other than the place they lived in 5 years prior. According to the 2009 Comprehensive Food Security and Vulnerability Analysis, 12% of Rwanda households have a member who migrated in 3 months prior to survey (Feb-Mar 2009).

[23] There is also a more general concern about extrapolating from Rwanda to migrant networks more generally. Rwanda experienced a devastating genocide roughly 12 years prior to the period we analyze, and ethnic tensions persist. For better or for worse, our dataset does not contain ethnic markers, and the study of ethnic divisions in Rwanda is an extremely delicate subject Hintjens (2008). Indeed, the Rwandan government eliminated the official distinction between Hutu and Tutsi by decree in 2004 Lacey (2004), and has banned subsequent research into ethnic divisions Economist (2019).

data are broadly consistent with existing data on internal migration in Rwanda. While we do not have survey data that make it possible to directly assess whether phone owners are representative of migrants more generally, we do find that the two populations have similar demographic characteristics. In particular, separate survey data indicates that the demographic distribution of migrants and non-migrants (i.e., Figures 11 and 12 in National Institute of Statistics of Rwanda (2014)) are quite similar to the demographic distribution of phone owners and non-owners (i.e., Table 2 in J. E. Blumenstock and Eagle (2012)).[24]

Second, the unique identifiers we observe are for mobile phone numbers, not individuals. As noted above, we attempt to limit the extent to which firms and organizations influence our analysis by removing numbers with very large networks, but this does not fully eliminate potential concerns. When multiple people share the same phone number (which J. E. Blumenstock and Eagle (2012) show was not uncommon during this period), we may overestimate the size of an individual's network. Related, it's possible that a single individual might use multiple phone numbers, which would have the opposite effect (in practice, we believe this was less common, since a monopoly operator existed). In principle, our data make it possible to uniquely identify devices and SIM cards, in addition to phone numbers. However, compared to these alternatives, we believe the phone number (which is portable across devices and SIM cards) bears the closest correspondence to the individual subscriber.

Third, the social network we observe is the network of mobile phone relations, which is a subset of all true social relations in Rwanda. This subset is non-random: it is biased toward the same socio-demographic categories described above; it systematically understates certain types of relationships (such as those that are primarily face-to-face); and may overstate other more transient or functional relationships (such as with a shopkeeper). We address some of these concerns through robustness tests that vary the definition of "social tie," for instance by only counting edges where communication is reciprocated (see Section 2.6). Other concerns are ameliorated by the fact that much of our analysis focuses on long-distance relationships, and during this period in Rwanda the mobile phone was the primary means of communicating over distance. We find it difficult to imagine how the core results we document below could be a byproduct of non-random selection of true social ties into the sample of ties we observe, but this remains a fundamental limitation of using digital trace data to study social networks.[25]

Fourth, the phone data are anonymous and cannot be matched to information about basic economic or demographic information on the individual using each phone. This raises immediate concerns that the network measures we use are simply a proxy for other unobserved confounding variables. However, as we discuss at length in the next section, we use an extremely restrictive set of fixed effects that limits the potential for many of the most worrisome sources of omitted variable bias. However, fixed effects cannot eliminate this potential

[24]We also note that during the period from 2005-2009, there was dramatic adoption of mobile phone technology in Rwanda, and the population of individuals in the sample changes over time. However, as we discuss in Section 2.5, our empirical specification (and in particular the use of time fixed effects) is designed to isolate variation within a relatively short window of time.

[25]See Chuang and Schechter (2015) for a more complete discussion of different approaches to measuring social networks in developing countries.

bias, so in the section below, we carefully articulate the identifying assumption required to interpret our estimates as causal, and provide several robustness tests to explore possible alternative explanations for our results.

2.5 Identification and Estimation

The focus of this paper is on understanding how social networks provide utility that influences the decision to migrate. While a host of other factors also influence that decision — from wage and amenity differentials to physical distance and associated migration costs — we try to understand how, holding all such factors fixed, certain variations in social network structure systematically correlate with migration decisions. In the stylized example of Figure 2.1, we ask whether a person with network G_1 is more likely to migrate than someone with network G_2, whose network is marginally more interconnected and would be expected to provide marginally more cooperation capital. We similarly compare the migration decisions of such individuals to individuals with network G_3, which is slightly more extensive and would be expected to provide slightly more information capital. In practice, of course, the actual network structures are much more complex (as in Figure 2.2). We therefore use statistical models to estimate the effect of marginal changes in complex network structure on subsequent migration decisions.

The central difficulty in identifying the causal effect of social networks on migration is that the social networks we observe are not exogenous: people migrate to places where their networks have certain characteristics, but this does not imply that the network caused them to go there. Here, we describe our estimation strategy, and the identifying assumptions required to interpret our regression estimates.

Simultaneity

An obstacle to understanding the causal effect of networks on migration is that migration decisions may also shape networks. This would be expected if, for instance, migrants strategically formed links to destination communities in anticipation of migration, or simply made a large number of phone calls to their destination before migrating.

We superficially address this concern in two ways. First, we analyze the lagged, rather than contemporaneous, decisions of migrants. Specifically, we relate the migration decision M_{it} made by individual i in month t to the structure of i's social network s months prior. As a concrete example, when t = May 2008 and $s = 2$, we relate the May 2008 migration decision to the structure of the individual's social network in March 2008.[26] Second, rather than focus on the *number* of direct contacts a migrant has at home and in the destination, we focus on the *connections* of those contacts, holding the number of contacts fixed (see Figure 2.1). This is because it seems easier for a migrant to directly control the number of

[26] Our main specifications use $s = 2$, but in robustness tests we also check $s = 3$ and $s = 1$.

contacts she has in the destination and at home than it is for her to alter the higher-order structure of her social network.

These two techniques reduce, but do not eliminate, the potential for simultaneity. In particular, a migrant might plan her migration many months in advance of migration, and in that process could change her higher-order network structure — for instance by asking a friend to make new friends on her behalf, or by encouraging two friends to talk to each other. To gauge the extent to which this might bias our results, we run several empirical tests, and find little evidence of such anticipatory behavior. For instance, Figure 2.4 shows, for a random sample of migrants, how the geographic distribution of migrants' social networks changes over time. Prior to migration, roughly 40% of the average migrant's contacts are in the origin and 25% are in the destination; three months after migration, these proportions have switched, reflecting how the migrant has adapted to her new community. Notably, however, migrants do not appear to strategically form contacts in the destination immediately prior to migrating; if anything, migrants shift their focus to the people in the community they are leaving. These compositional changes do not mask a systematic increase in the *number* of contacts in the destination, or the number of total calls to the destination: Figure A.3 indicates that the total number of contacts increases over time, but there is no sudden spike in the months before migration; Figure A.2b shows analogous results for total call volume. As a sort of 'placebo' test, Figure A.2 shows the corresponding figure for non-migrants, where no changes are observed in the "migration" month, as expected (since no migration takes place for this sample).

What matters most to our identification strategy is that we similarly find no evidence that migrants are systematically altering the *higher-order* structure of their social networks in the months prior to migration. In particular, Figure A.4 indicates that migrants have a relatively constant number of unique friends of friends over time (with no noticeable shift in the months prior to migration). Figure A.5 shows similar results for the level of common support in the network.

Omitted Variables

The second threat to identification is the fact that network structure may be a proxy for other characteristics of the individual (e.g., wealth, ethnicity) and location (e.g., population density, wages) that also influence migration. Our main strategy for dealing with such omitted variables is to include an extremely restrictive set of fixed effects that control for many of the most concerning sources of endogeneity. This strategy is possible because of the sheer volume of data at our disposal, which allow us to condition on factors that would be impossible in regressions using traditional survey-based migration data.

Our preferred specification includes fixed effects for each individual (roughly 800,000 fixed effects), for each origin-destination-month tuple (roughly 18,000 fixed effects), and for the number of direct contacts in the destination. The individual fixed effects absorb all time-invariant individual heterogeneity (such as wealth, gender, ethnicity, personality type, family structure, and so forth), and addresses the fact that some people are inherently more likely to

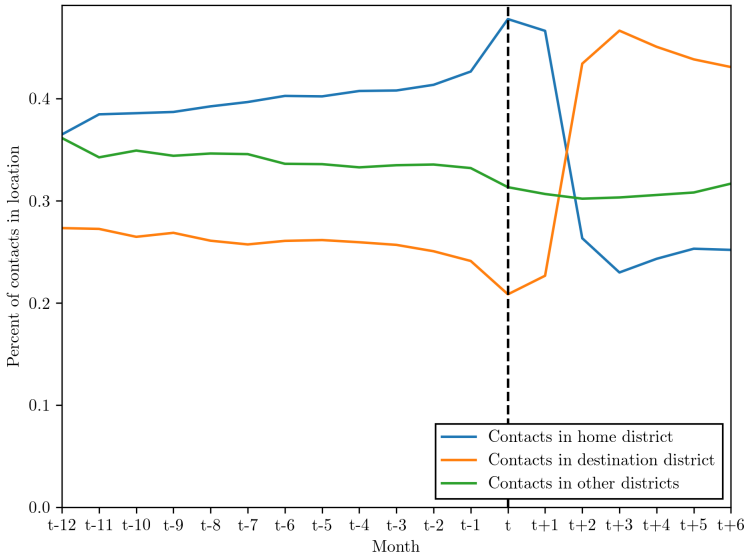


Figure 2.4: Geographic network structure before and after migration – migrants only. *Notes:* Figure shows, for a random sample of 10,000 migrants, the average percentage of the migrant's social network in the home and destination districts, in each of the 12 months before and 6 months after migration. Dashed vertical line indicates the date of migration.

migrate than others (and have inherently different social networks). The origin-destination-month fixed effects control for any factor that similarly affects all individuals considering the same origin-destination migration in the same month. This includes factors such as physical distance, the cost of a bus ticket, location-specific amenities that all migrants value equally, average wage differentials, and many of the other key determinants of migration documented in the literature (including the usual "gravity" effects in a standard trade or migration model).[27] Finally, we include fixed effects for the number of first-degree contacts in the destination in order to isolate the effect of differences in higher-order network structure on migration.

[27]For instance, we know that rates of migration are higher to urban centers, and that social networks in urban centers look different from rural networks. Including a destination fixed effect removes all such variation from the identifying variation used to estimate the effect of networks on migration. The origin-destination-month fixed effects remove destination-specific variation, as well as more complex confounding factors that vary by destination and origin and time, such as the possibility that the seasonal wage differential between two districts correlates with (lagged) fluctuations in social network structure.

Identification

To summarize, the identifying variation we exploit in our main specification is within-individual over time and over potential destinations, net any factors that are shared by all people considering the same origin-destination trip in the same month, and net any effects that are common to all people with exactly the same number of friends in the destination. We would observe such variation over time if, for instance, an individual had been considering a move to a particular destination for several months, but only decided to migrate after his friends in the destination became friends with each other (the G_2 vs. G_1 comparison of Figure 2.1) — and if that tightening of his social network exceeded the average tightening of networks in that destination (as might occur around the holidays, for instance). An example of identifying variation within individual over potential destinations would occur if, in a given month, a single migrant were choosing between two destination districts, had the same number of contacts in each district, and then decided to migrate to the district where his contacts were more interconnected — and if that additional interconnectedness exceeded the extent to which all networks in that destination were more interconnected in that particular month. Prima facie, it may seem unlikely that such small differences would shape the decision to migrate, but our data allow us to ascertain whether, across millions of individual migration decisions, such a general tendency exists.

The fixed effects we include significantly reduce the scope for omitted variables to bias our estimates of the effect of network structure on migration, but they do not eliminate such bias entirely. If, for instance, origin-destination wage differentials are individual-specific, the main fixed effects would not absorb this variation. This might occur if carpenters' networks in a particular district are more interconnected (relative to carpenter networks other districts) than farmers' networks in that district (again relative to farmers' networks in other locations), and if migration rates of carpenters to that district are higher for reasons unrelated to the network. We revisit these concerns, and other possible threats to identification, in Section 2.6, once the main results are established. That discussion acknowledges certain limitations of our identification strategy and performs a series of tests of robustness. For instance, we also test a series of even more restrictive specifications that include fixed effects for the *individual*-destination (this isolates variation within individual-destination over time and would address the carpenter/farmer concern, if we assume that those trends are temporally stable), for the *individual*-month (which isolates variation across potential destinations for a single individual in a single month), and a few other scenarios.

Estimation

Formally, for a migrant i considering moving from home district h to destination district d in month t, we wish to estimate the effect of (s-lagged) network structure $Z_{ihd(t-s)}$ on the migration decision M_{ihdt}, where M_{ihdt} is a binary variable equal to 1 if the migrant chooses to move from h to d at t and 0 otherwise. We estimate this in two ways, using either a linear model or a discrete choice (multinomial logit) model.

In the linear model:

$$M_{ihdt} = \beta Z_{ihd(t-s)} + \pi_{hdt} + \mu_i + \nu_D + \epsilon_{ihdt} \tag{2.6}$$

where π_{hdt} are the (home district * destination district * month) fixed effects; and μ_i are the individual fixed effects. We also condition on i's degree centrality in the destination D using a set of fixed effects ν_D that non-parametrically control for effects that are invariant across all people with the same number of contacts in the destination. The coefficient of interest is β, which indicates the average effect of network property $Z_{ihd(t-s)}$ on the probability of migration. Standard errors are two-way clustered by individual and by home-destination-month.

Specification (2.6) has several attractive properties: it makes it possible to condition on a rich set of fixed effects, and can be estimated relatively quickly even on a very large dataset. The difficulty with estimating equation (2.6) arises in how an observation is defined in the regression. In particular, for non-migrants, it is not clear what should be considered the destination network. We address this by defining an observation at the level of the individual-month-*potential destination*. Thus, in each month, each individual comprises 26 observations, one for each of the 26 potential districts to which that individual could migrate in that month.[28]

Our second approach uses a discrete choice (multinomial logit) model of the migration decision, to address the fact that the 26 observations for each individual in each month are not i.i.d. The multinomial logit is becoming increasingly common in the migration literature (Dahl & Sorenson, 2010; Davies et al., 2001), and has the advantage of providing a sound microeconomic foundation of utility maximization with a random utility model (Mcfadden, 1974; Revelt & Train, 1998). It treats each monthly decision as a single decision with 27 alternatives (one corresponding to staying at home, and 26 migration options).[29] While more natural in this regard, the multinomial logit has several limitations: it is not possible (or at least, quite difficult) to include the same restrictive set of fixed effects as we include in the linear regression, thus increasing the scope for omitted variable bias; it is similarly ill-suited to estimating the impact of individual-specific characteristics (in our case, the attributes of the individual's home network); and the IIA assumption is problematic. Finally, the computational requirements of the multinomial logit are several orders of magnitude greater than that of the corresponding regressions.[30] In practice, the results from the multinomial

[28] An individual is only considered in months where she can be classified as a migrant or a non-migrant in that month. When an individual is classified as "other" (See Section 2.4), she is excluded for that month.

[29] Another possibility is to model the decision to migrate with a nested logit model, where the individual makes two independent decision: the first is whether or not to migrate and the second is, given the decision to move, the choice of destination (Knapp et al., 2001; McFadden, 1984). We believe this approach is less appropriate to our context, as the decision to migrate is closely related to the possible destination choices — Davies et al. (2001) provides a more complete discussion of this point.

[30] Whereas equation 2.6 can be estimated, even with millions of fixed effects and two-way clustered standard errors, in several minutes on our high-performance computing cluster, the panel logit takes several hours, even with minimal fixed effects. This computational constraint is particularly problematic when estimating our effects non-parametrically, as discussed below.

logit are always qualitatively the same as those from linear regression, so our main analysis is based on specification (2.6), with multinomial logit results provided as robustness tests in the appendix (see Table A.5).

Non-parametric estimation

Equation (2.6) and the corresponding multinomial logit indicate the average effect of network characteristic Z on the decision to migrate. We are also interested in disaggregating these effects non-parametrically, to understand how such effects differ for migrants with destination networks of different sizes. We thus present a series of figures that show the coefficients from estimating the model:

$$M_{ihdt} = \sum_{k=1}^{D^{max}} \beta_k Z_{ihd(t-s)} \cdot \mathbb{1}(D = k) + \pi_{hdt} + \mu_i + \nu_D + \epsilon_{ihdt} \tag{2.7}$$

The vector of β_k coefficients from the above model indicates, for migrants with a fixed number of contacts k, the relationship between the migration decision and the higher order network characteristic $Z_{ihd(t-s)}$. As we will see, this analysis helps reveal how the "average" effect of different network structures masks considerable heterogeneity that would not be visible in traditional survey-based data.

2.6 Results

Table 2.2 summarizes the main results from estimating model (2.6). We find that on average, each additional contact in the destination is associated with a 0.37% increase in the likelihood of migration (Panel A, column 1), and each contact at home is associated with a 0.04% decrease in that likelihood (Panel B, column 1). Columns 2-4 indicate the average effect of changes in high-order structure, after controlling for the immediate contacts of the individual (i.e., the "degree centrality" fixed effects). In column 4, for instance, the second row in Panels A and B indicates that migrants are more likely to go to places where their destination networks are more interconnected, and less likely to leave interconnected home networks. The third row indicates that, perhaps surprisingly, people are not more likely to migrate to destinations where their contacts have a large number of contacts, but they are less likely to leave such places.

Where the first column of Table 2.2 separately estimates the "pull" and "push" forces of networks on migration (cf. Hare, 1999), the first two columns of Table A.2 jointly estimate both effects, to allow for a more direct comparison. Comparing the first two coefficients in the first and second rows, we note that in determining migration outcomes, the marginal effect of an additional contact in the destination is roughly 6.5 to 7.5 times as important as an additional contact at home.

In the subsections below, we discuss these "reduced form" results in greater detail, re-estimate each average effect non-parametrically, and discuss heterogeneity in the migration

Table 2.2: Migration and social network structure - base specification

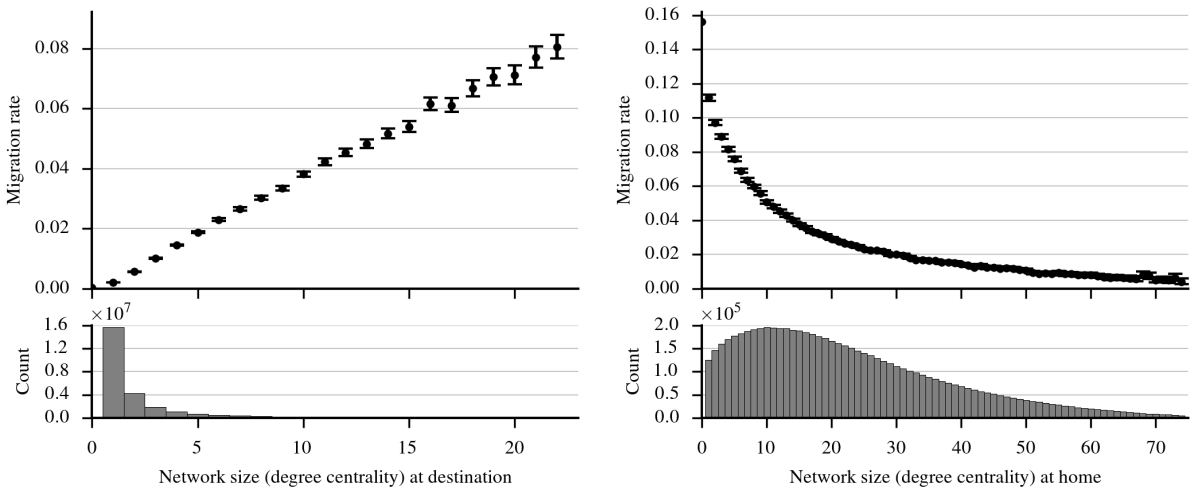
	(1)	(2)	(3)	(4)
Panel A: Destination network characteristics				
Degree (network size)	0.0036547***			
	(0.0000102)			
% Friends with common support		0.0014813***		0.0014808***
		(0.0001146)		(0.0001146)
Unique friends of friends			−0.0000005	−0.0000002
			(0.0000009)	(0.0000009)
Observations	9,889,981	9,889,981	9,889,981	9,889,981
R^2	0.1851423	0.1853017	0.1852869	0.1853017
Panel B: Home network characteristics				
Degree (network size)	−0.0003985***			
	(0.0000049)			
% Friends with common support		−0.0003467		−0.0005710**
		(0.0002422)		(0.0002424)
Unique friends of friends			−0.0000089***	−0.0000089***
			(0.0000004)	(0.0000004)
Observations	9,889,981	9,889,981	9,889,981	9,889,981
R^2	0.1743203	0.1750909	0.1751320	0.1751325
Degree fixed effects	No	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes	Yes
Home*Destination*Month F.E.	Yes	Yes	Yes	Yes

Notes: Each column indicates a separate regression of a binary variable indicating 1 if an individual i migrated from home district h to destination district d in month t. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

response by migrant and location type. The analysis reveals considerable nuance in the relationship between networks and migration, helps explaining the "surprising" result in Table 2.2, and establishes a set of stylized facts that form the basis for structural model of social capital.

The effect of network size, in the destination and at home

Our first result validates a central thesis of prior research on networks and migration, which is that individuals are more likely to migrate to places where they have more connections. The unconditional relationship between degree centrality at destination (i.e., the number of unique contacts of the individual) is shown in Figure 2.5a. A point on this figure can be



(a) Degree Centrality at Destination

(b) Degree Centrality at Home

Figure 2.5: Migration and degree centrality (number of unique contacts in network). *Notes:* In both (a) and (b), the lower histogram shows the unconditional degree distribution, i.e., for each individual in each month, the total number of contacts in the (a) destination network and (b) home network. The upper figure shows, at each level of degree centrality, the average migration rate. Error bars indicate 95% confidence intervals, clustered by individual.

interpreted as the average migration rate (y-axis) across individuals with a fixed number of contacts in the destination (x-axis). For instance, roughly 4% of individuals who have 10 contacts in a potential district d' in month $t-2$ are observed to migrate to d' in month t. The bottom panel of the figure shows the distribution of destination degree centrality, aggregated over individuals, months (24 total), and potential destinations (26 per individual).

This figure also provides intuition for our identification strategy and preferred empirical specification. The average migration rates depicted Figure 2.5a are likely confounded by a variety of omitted variables. For instance, people in rural districts typically know more people in the urban capital of Kigali than in other districts, and rates of migration to Kigali are higher than to other districts. Thus, Figure A.6 re-estimates the migration rates of Figure 2.5a, conditioning on a series of increasingly restrictive fixed effects. In the first panel, Figure A.6a reports the ν_k coefficients and standard errors from estimating:

$$M_{ihdt} = \sum_{d=1}^{D^{max}} \nu_d \mathbb{1}(D = d) + \epsilon_{ihdt} \tag{2.8}$$

Mechanically, these coefficients are identical to unconditional correlations shown in Figure 2.5a, albeit shifted down because of the omitted global intercept. In subsequent panels, Figure A.6b includes destination district fixed effects (which most immediately addresses the Kigali concern described above). Figure A.6c replaces destination fixed effects with more

stringent destination-origin-month fixed effects. Finally, Figure A.6d adds individual fixed effects, resulting in an estimating equation similar to equation 2.7:

$$M_{ihdt} = \sum_{d=1}^{D^{max}} \nu_d \mathbb{1}(D = d) + \pi_{hdt} + \mu_i + \epsilon_{ihdt} \tag{2.9}$$

In all figures, the qualitative relationship is remarkably unchanged. Individuals with more contacts in a destination community are more likely to migrate to that community. We also see that this relationship is positive, monotonic, and approximately linear with elasticity one. In other words, individuals with k times as many contacts in a destination district are k times more likely to migrate to that district.

Just as migrants appear drawn to destinations where they have a large number of contacts, migrants are less likely to leave origins where they have a large number of contacts. Figure 2.5b shows the monotonically decreasing relationship between migration rates and the individual's degree centrality at home.

Higher-order network structure

We next examine how the high-order structure of the individual's network — i.e., the connections of the individual's contacts — relate to subsequent migration decisions. We focus on the proxies for network interconnectedness and extensiveness described in Section 2.4.

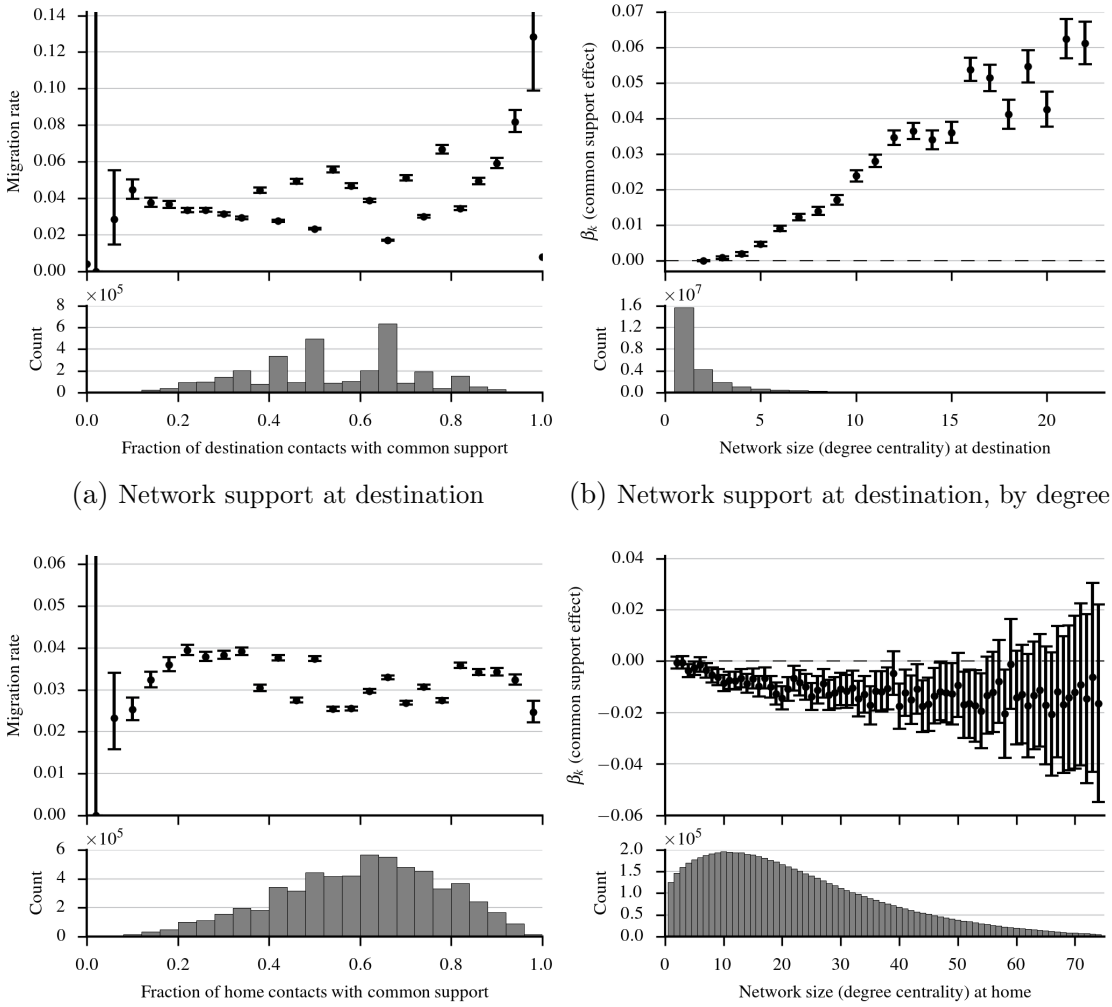
Network 'interconnectedness'

Figure 2.6 documents the relationship between migration decisions and the interconnectedness of the individual's social networks, making the generalized comparison between G_1 and G_2 in Figure 2.1. As described in Section 2.4 and originally proposed in Jackson et al. (2012), we measure this interconnectedness as network "support," or the fraction of i's contacts who have one or more friends in common with i. In later robustness tests, we show that related measures of network interconnectedness, tightness, and clustering, produce qualitatively similar results.[31]

Both at home and in the destination, the unconditional relationship between migration and interconnectedness is ambiguous. Figures 2.6a and Figure 2.6c show how migration varies with network support in the destination and at home, respectively. However, this unconditional relationship is potentially confounded by a large number of omitted variables, including the fact that network support is generally decreasing in degree, since the larger an individual's network, the harder it is to maintain a constant level of support.

Holding degree fixed, a clear pattern emerges: people are systematically drawn to places where their networks are more interconnected. This pattern is evident in Figure 2.6b, which

[31] The distinction between support and clustering is that the former counts the proportion of i's friends with one or more friends in common, the latter counts the proportion of all possible common friendships that exist – see Jackson (2010).



(a) Network support at destination

(b) Network support at destination, by degree

(c) Network support at home

(d) Network support at home, by degree

Figure 2.6: Migration and network "tightness" (friends with common support). *Notes:* Network support indicates the fraction of contacts supported by a common contact (see Section 2.4). In all figures, the lower histogram shows the unconditional distribution of the independent variable. Figures in the left column (a and c) show the average migration rate for different levels of network support. Figures in the right column show the β_k values estimated with model 2.7, i.e., the correlation between migration and support for individuals with different sized networks (network degree) after conditioning on fixed effects. Top row (Figures a and b) characterizes the destination network; bottom row (Figures c and d) characterizes the home network. Error bars indicate 95% confidence intervals, clustered by individual.

plots the β_k coefficients estimated from model (2.7) on the destination social network, all of which are positive. Figure 2.6d show that, holding degree fixed, people are significantly less likely to leave home if their home contacts are more interconnected. Appendix Figure A.7 replicates this analysis using the network clustering, instead of network support, as a measure of interconnectedness. Results are qualitatively unchanged.

The fact that people are more likely to go to places where their networks are interconnected may not be surprising, but in other settings, the opposite result has been documented. For instance, Ugander et al. (2012) show that people are more likely to sign up for Facebook when their pre-existing Facebook friend network is less interconnected.

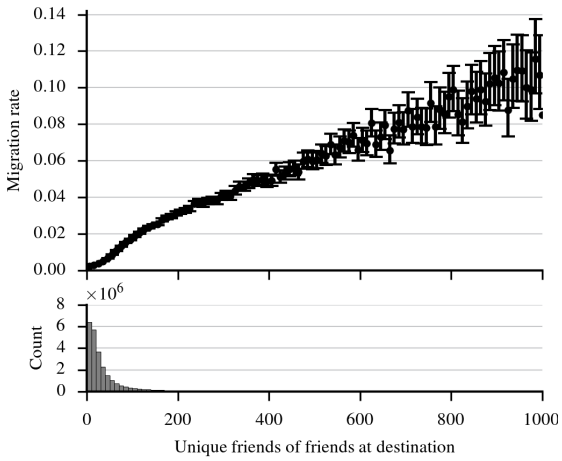
Network 'extensiveness'

The relationship between migration and network extensiveness is more surprising and subtle. Here, we focus on the number of unique friends of friends a person has in a given region, i.e., the generalized comparison between G_1 and G_3 in Figure 2.1. Without controlling for the size of an individual's network, there is a strong positive relationship between migration and extensiveness in the destination (Figure 2.7a), and a strong negative relationship with extensiveness in the origin (Figure 2.7c). The shape of these curves resemble the relationship between migration rate and degree shown earlier in Figure 2.5: the average migration rate increases roughly linearly with the number of friends of friends in the destination, and decreases monotonically but with diminishing returns to friends of friends at home.

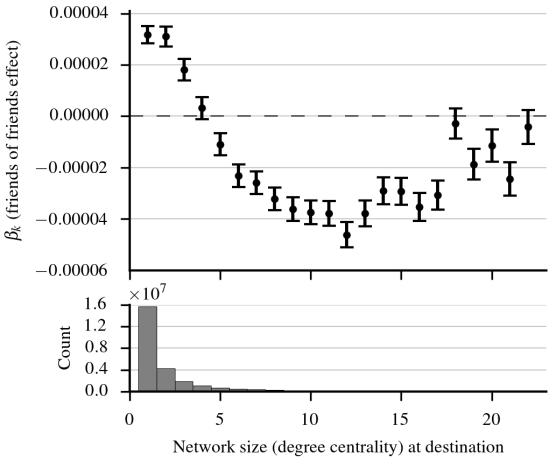
Of course, the number of friends of friends a person has is largely determined by the number of friends that person has. Thus, Figures 2.7b and 2.7d show how the number of friends of friends relates to migration, holding fixed the number of friends (as well as the other fixed effects in model (2.7)). For the home network, Figure 2.7d indicates the expected pattern: the fact that all of the coefficients are negative suggests that given a fixed number of friends at home, people are less likely to leave when those friends have more friends.

The surprising result is Figure 2.7b, which indicates that the likelihood of migrating does not generally increase with the number of friends of friends in the destination, after conditioning on the number of friends. The friend of friend effect is positive for people with $1-3$ destination contacts, but negative for people with >4 destination contacts. Averaged over all migrants, this effect is negative and insignificant (row 3 of Tables 2.2 and A.2). This result is difficult to reconcile with most standard models of information diffusion, such as those proposed in A. Banerjee et al. (2013) and Kempe et al. (2003). Indeed, much of the literature on migration and social networks seems to imply that, all else equal, individuals would be more likely to migrate if they have friends with many friends, as such networks would provide more natural conduits for information about job opportunities and the like.[32]

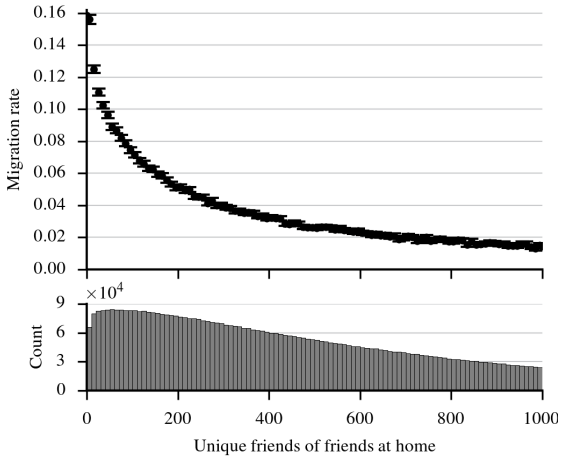
[32] A very similar pattern appears in Figure A.8 when we look at the *home* friends of the friend in the destination. In other words, if migrant i in home district h has a friend j in destination district d, we find that people are less likely to migrate to places where j has more friends located in h. (Where Figure 2.7b analyzes the relationship between migration and the number of j's friends in d, Figure A.8 analyzes the number of j's friends in h).



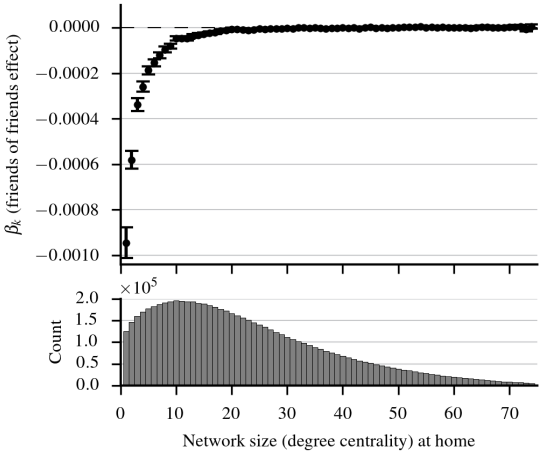
(a) Friends of friends at destination



(b) Friends of friends at destination, by degree



(c) Friends of friends at home



(d) Friends of friends at home, by degree

Figure 2.7: Relationship between migration and "extensiveness" (unique friends of friends). *Notes:* Main figures in the left column (a and c) show the average migration rate for people with different numbers of unique friends of friends. Figures in the right column show the β_k values estimated with model 2.7, i.e., the correlation between migration and unique friends of friends for individuals with different numbers of friends, after conditioning on fixed effects. Top row (Figures a and b) characterizes the destination network; bottom row (Figures c and d) characterizes the home network. Lower histograms show the unconditional distribution of the independent variable. Error bars indicate 95% confidence intervals, clustered by individual.

We run a large number of empirical tests to convince ourselves that this pattern is not an artifact of our estimation or measurement strategy — several of these are described in

Section 2.6. However, the data consistently indicate that the average migrant is no more likely to go to places where she has a large number of friends of friends. This is perhaps most transparent in Figure A.9, which shows the distribution of the count of friends of friends for all migrants and non-migrants with exactly 10 friends in the potential destination. Among this sample of the population, it is apparent that, on average, non-migrants have more friends of friends in the destination networks than migrants.

Heterogeneity and the 'friend of friend' effect

The effect that networks have on the "average migrant" masks considerable heterogeneity in how different types of migrants are influenced by their social networks. In particular, Tables A.6-A.10 disaggregate the results from Table A.2 along several dimensions that are salient in the migration literature: whether the migrant has previously migrated to the destination (Table A.6); whether the migration is between adjacent districts or over longer distances (Table A.7); whether the migrant stays in the destination for a long period of time (Table A.8); and whether the migration is to an urban or rural destination (Tables A.9 and A.10).

Heterogeneity and unawareness of the broader network

Several patterns can be discerned from these tables, but we focus our attention on how the network "extensiveness" effect changes with these different subgroups, as that was the most unintuitive of the above results. Here, we find that for certain types of migration — repeat migrations, short-distance migrations, and long-term migrations — the number of friends of friends is positively correlated with migration rates. Each of these types of migration are significantly less common than the typical migration event (a first-time, long-distance migration), hence the statistically insignificant negative average effect observed in Table 2.2.

This heterogeneity suggests one possible explanation for the unexpected null 'friend of friend' result: the average migrant may simply be unaware of the higher-order structure of their destination network. Such an explanation is supported by several other studies that find that people have incomplete information about the friends of their friends Casciaro (1998), A. Chandrasekhar et al. (2016), Friedkin (1983). This information assymetry is likely to be most severe when the would-be migrant lives far from, or has less experience with, the destination friend's community. And indeed, this is what the heterogeneity suggests: the migrants who are positively influenced by extensive destination networks are the migrants who seem likely to be more familiar with the structure of those networks. When the destination is more familiar, it begins to resemble the home network, where A. Banerjee et al. (2019) have found that people have good information on (proxies for) their friends' centrality.

Strong ties, weak ties, and recent migrants

A different explanation for the 'friend of friend' result is suggested by a closer analysis of the role of strong and weak ties in migration. Here, and consistent with recent work by Giulietti et al. (2018), we find that both strong and weak ties matter in migration: the effect of a strong destination tie is roughly 1.5 times that of a weak destination tie; at home, the effect of a strong tie is roughly twice as large as the effect of a weak tie. These results are shown in Table A.11, which defines a strong tie as one that supports five or more communication events in the reference month (the 90^{th} percentile of communication frequency) — see Section 2.4 for details and justification.

Recent and co-migrants have a similar effect: people are more likely to go to places where they know recent migrants (defined as a contact who previously made the origin-destination migration that the individual is considering). Coefficient estimates in Table A.14 indicate that knowing a recent migrant in the destination increases the likelihood of migration by roughly 3.5X the amount as knowing anyone else in the destination. The effect is slightly larger for recent migrants who arrived in the destination very recently (last month) than for recent migrants who arrived at any point prior. Such evidence is consistent with the fact that households and extended families frequently make joint labor allocation decisions Rosenzweig and Stark (1989).

However, neither strong ties nor recent migrants dominate the migration decision: when controlling for either factor, the main effects reported in Table 2.2 are qualitatively unchanged.

More interesting is the role that *higher order* tie strength plays in modulating the migration decision. In particular, the results in Section 2.6 suggest that a migrant i is drawn to locations where i's contact j has a friend in common k, but that i is indifferent or repelled if k is not a common friend of i. However, this average effect hides a more nuanced pattern: when disaggregating by tie strength, we observe that the negative effect is driven by situations where the i-j tie is weak but the j-k tie is strong — or in other words, when the migrant has a tenuous connection to the destination and that tenuous connection has strong connections to other people in the destination.

These results are presented in Figure 2.8, which summarizes the regression coefficients from Tables A.12 and A.13. The figure indicates the sign of the regression coefficient (using $+/-$ labels) from a regression of i's migration decision on the number of different types of i-j links, where type is determined by the strength of the i-j link (strong ties shown with thick lines, weak ties shown with thin lines) and the existence and strength of the j-k link. The four figures on the left indicate that migrants are generally drawn to places where their contacts have many ties, but that they are deterred when their weak ties have a large number of strong ties. Similarly, the set of triangles on the right, which show all possible configurations of a supported i-j tie, indicate that supported links are positively correlated with migration in all cases except when the i-j tie is weak and the j-k tie is strong.

This heterogeneity is consistent with the notion, proposed by Dunbar (1998) and others, that people might have a capacity constraint in the number of friendships they can effectively

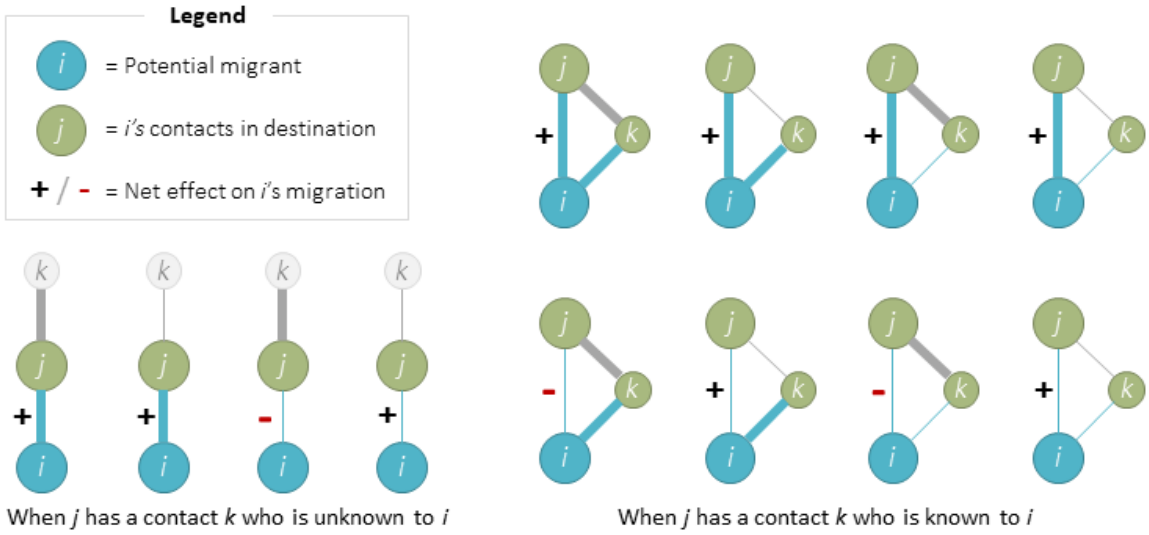


Figure 2.8: The role of (higher order) strong and weak ties in a migrant's network. *Notes:* Thick edges represent "strong" ties and thin edges represent "weak ties" The +/− signs summarize the effect that j has on i's likelihood of migration, based on coefficients in Tables A.12 and A.13.

support, which in turn might induce a degree of rivalry for the attention of a friend.[33] In our context, migrants may be drawn to places where they receive their friends' undivided attention.[34] However, these results — and particularly the results concerning the "friend of friend" effect — are more speculative than conclusive. We take these ambiguities as motivation to develop a more coherent model of how migrants derive utility from networks, which we turn to in Section 2.7.

Robustness and Identification (revisited)

Section 2.5 describes the identifying assumptions behind our regressions. In particular, when estimating models (2.6) and (2.7), we assume $E[\epsilon_{ihdt}|\pi_{hdt}, \mu_i, \nu_D] = 0$. In other words, we assume that the variation in higher-order network structure we observe is exogenous,

[33]Dunbar originally proposed that humans could maintain roughly 150 stable relationships, since "the limit imposed by neocortical processing capacity is simply on the number of individuals with whom a stable inter-personal relationship can be maintained."

[34]Related, L. A. Beaman (2012) and Dagnelie et al. (2019) find evidence that migrants may compete with each other for economic opportunities. See also Wahba and Zenou (2005), who empirically test the tradeoff between information and rivarly in an Egyptian labor market survey. They show that up to a certain (network) size, the network information effect dominates the competition (rivalry) effect so that network is always beneficial for finding a job. However, above a certain size, the second effect dominates the first one so that agents have less chance of finding job when network size increases.

conditional on the identity of the individual making the migration decision, the origin-destination-month choice being made, and the number of direct contacts the individual has in that destination in that month. While we believe these fixed effects address the most concerning sources of bias, it is of course possible to concoct a scenario in which this assumption would be violated (as in the carpenter/farmer example in Section 2.5).

We therefore run a series of robustness checks that further isolate the identifying variation behind the regression results presented above. In particular, Appendix Table A.4 re-estimates the main effect shown in column 4 of Table 2.2 under a variety of increasingly restrictive fixed effect specifications. Column 1 replicates the prior result, including fixed effects for π_{hdt}, μ_i, and ν_D. Column 2 in Table A.4 then includes fixed effects for each *individual-month* pair, so that the identifying variation comes within individual in a given month but across potential destination districts.[35] Column 3, by contrast, includes separate fixed effects for each *individual-destination* pair, so that the β coefficients are identified solely by variation within individual-destination over time.[36] Column 4 includes fixed effects for each *individual-Degree*, exploiting variation between all destinations where a single individual has the exact same number of contacts. Column 5, which includes over 600 million fixed effects, isolates variation within individual-home-destination observations over time. In all instances, the coefficients of interest are quite stable, and in particular, the average effect of additional friends of friends is either negative or insignificant (or both).

In addition to these variations on the core regression specification, we also re-estimate our results using a discrete choice (multinomial logit) model. As noted earlier, this is a more natural specification as it treats each monthly decision as a single decision with 27 alternatives (one corresponding to staying at home, and 26 migration options). Results are shown in Table A.5, and are broadly consistent with the main regression results presented earlier.

Finally, we perform several additional tests to check whether the main results are sensitive to different measurement strategies used to process the mobile phone data. Since these results show a very similar picture and are highly repetitive, we omit them from the paper but can provide them to interested readers upon request:

- **How we define 'migration' (choice of k)**: Our main specifications set $k = 2$, i.e., we say an individual has migrated if she spends 2 or more months in d and then 2 or more months in $d' \neq d$. We observe qualitatively similar results for $k = 1$ and $k = 3$.

[35] Such variation would occur if, for example, in a given month, a single migrant were choosing between two destination districts, had the same number of contacts in each district, and then decided to migrate to the district where his contacts were more interconnected — and if that additional interconnectedness exceeded the extent to which all networks in that destination were more interconnected.

[36] This could reflect a scenario where an individual had been considering a move to a particular destination for several months, but only decided to migrate after his friends in the destination became friends with each other (the G_2 vs. G_1 comparison of Figure 2.1) — and where that tightening of his social network exceeds the average tightening of networks in that destination (as might occur around the holidays, for instance).

- **How we define the 'social network' (reciprocated edges)**: In constructing the social network from the mobile phone data, we normally consider an edge to exist between i and j if we observe one or more phone call or text message between these individuals. As a robustness check, we take a more restrictive definition of social network and only include edges if i initiates a call or sends a text message to j and j initiates a call or sends a text message to i.

- **How we define 'social network' (ignore business hours)**: To address the concern that our estimates may be picking up primarily on business-related contacts, and not the kinship and friendship networks commonly discussed in the literature, we only consider edges that are observed between the hours of 5pm and 9am.

- **Treatment of outliers (removing low- and high-degree individuals)**: We remove from our sample all individuals (and calls made by individuals) with fewer than 3 contacts, or more than 500 contacts. The former is intended to address concerns that the large number of individuals with just one or two friends could bias linear regression estimates; the latter is intended to remove spammers, calling centers, "public" phones, and large businesses.

Stepping back slightly, the relevant question is whether we believe, for instance, that an individual would be more likely to move to a location where his friends happened to become more connected in the months prior to migration. This is what the coefficient 0.00035 in column 3 of Table A.4 indicates: fixing the individual and the destination, rates of migration are higher in the months after friends in the destination become more interconnected. To provide more transparent intuition behind this identifying variation, consider the following: We pull a random sample of 20,000 individuals who have exactly two contacts in a specific district for 4 consecutive months. We then calculate, for each person, whether those two contacts are more likely to become connected or disconnected at the end of the 4-month period (by regressing a dummy for triadic closure on a linear time trend); we then compare the migration rate in month 5 among the population whose two contacts became connected relative to the migration rate in month 5 of the population whose two contacts became disconnected. The migration rate is 2.2% in the former group, and 1.3% in the latter. In other words, when focusing on a sample who consistently have exactly two contacts in the destination, rates of migration are higher when a given individual's two contacts become more connected (over the 4-month period) than when they become more disconnected (over the 4-month period).

This coefficient is of course not perfectly identified. There may be other factors that help drive the observed correlation (for instance, if the migrant induces his friends to connect to each other; or if the interconnections occur because the employment prospects available to that specific migrant improve). But our data clearly indicate that, to continue with the above example, migrants go to places after their networks there become more interconnected – even if it stops short of explaining *why* the network became more interconnected. The presence of this positive correlation is accentuated by the fact that people are *not* more

likely to migrate to places where their friends have recently developed new friends of friends. So there is something unique in the formation of interconnections rather than extensive connections that correlates with subsequent migration decisions. It is this difference between interconnected and extensive networks that we investigate in more detail in the following section.

2.7 Structural Estimation

The reduced form results presented in Section 1.5 highlight *how* social networks influence migration decisions, but offer limited insight into *why* some network structures matter more than others. Since the phone data contain no identifying or socio-demographic information about the individual subscribers, we have limited ability to infer whether, for instance, interconnected networks are influential because they tend to consist of family members, co-ethnics, or some other tightly knit community. The regression specifications are also limited by the fact that different measures of higher-order network structure are highly inter-dependent, so it is difficult to isolate the effect of marginal changes to the network.

For these reasons, we return to the stylized model of Section 2.3, which describes how different subnetwork topologies provide utility to migrants, and use the revealed preference decisions in our data — to migrate or not to migrate — to parameterize a model of network-based social capital and migration. Recall that we say that an individual i receives utility $u_i(G)$ from a social network G. As emphasized in the literature, we assume that $u_i(G)$ is primarily comprised of information capital and cooperation capital. The next two subsections provide micro foundations for these two types of social capital.

Information capital: competition and 'extensiveness'

A robust theoretical and empirical literature suggests that the value of a social network stems, at least in part, from its ability to efficiently transmit information (see footnote 2). We build on recent efforts by A. Banerjee et al. (2013) to model this information capital as an information sharing process with possible loss of information. It is worth noting that A. Banerjee et al. (2013) study a seeding process in which an agent is injected with one unit of information, and this agent's diffusion centrality measures the impact of his information to the network. We study a receiving process in which each agent is initially endowed with one unit of information, and we seek to measure how much information an agent could receive from the network. Using the same information sharing process as A. Banerjee et al. (2013), we will show that the measure we seek turns out to be the diffusion centrality, because the flow of information is symmetric.

In this model, a population of N agents, $N = \{1, \ldots, n\}$, are connected in an undirected network. Let G be the adjacency matrix of the network: $G_{ij} = 1$ if i and j are connected and otherwise $G_{ij} = 0$, including $G_{ii} = 0$. Denote agent i's neighbors as $N_i = \{j : G_{ij} = 1\}$, and agent i's degree as $d_i = |N_i|$, which is the number of his or her neighbors in N_i. Agents

meet with their neighbors repeatedly, and when they meet, they share information with each other with probability $q \in (0, 1)$.

In this benchmark model of information sharing, more extensive networks — where an individual has a large number of short-distance indirect neighbors — provide additional utility. We extend this model by allowing for the possibility that neighbors might compete for the attention of their common neighbor. This is motivated by our earlier observation that more extensive destination networks are not positively correlated with migration, and with the evidence that suggests possible rivalry for attention (see Section 2.6).

We model the source of competition for attention as costly socializing with neighbors, so when an agent has more neighbors, he or she may spend less time with each neighbor. Formally, let cQ^ω be the cost of spending Q amount of time on communicating with neighbors. We assume each agent does not possess additional information about neighbors (such as their degrees), so each agent evenly distributes the total amount of time Q to her d neighbors, that is, she spends $q = Q/d$ amount of time with each neighbor. Her utility from communicating with neighbors is given by $d \cdot v(Q/d)^\beta - cQ^\omega$, in which she receives a value of $v(Q/d)^\beta$ from spending Q/d amount of time with each neighbor, and the total cost of spending time Q is cQ^ω. We assume the cost is convex in time $\omega \geq 1$, the value is concave in time $\beta \leq 1$, and they cannot be linear at the same time $\omega > \beta$. The agent's maximization problem becomes

$$\max_{Q} \; dv(Q/d)^\beta - cQ^\omega. \tag{2.10}$$

To maximize her utility, the agent's optimal time per neighbor is

$$Q/d = \frac{1}{d^\lambda}\left(\frac{\beta v}{\omega c}\right)^{\frac{1}{\omega-\beta}}, \text{ where } \lambda = \frac{\omega - 1}{\omega - \beta} \in [0, 1]. \tag{2.11}$$

Notice that if the cost is linear ($\omega = 1$), then the marginal cost of communicating with one neighbor does not increase when the agent has more other neighbors. Thus, the optimal time per neighbor is independent of her degree: $\lambda = 0$. On the other hand, if the value is linear ($\beta = 1$), time with neighbors are perfect substitutes. Then, the total amount of time Q is independent of her degree, which is then evenly split among neighbors: $\lambda = 1$.

Motivated by this simple exercise, we let the interaction between each pair of linked agents ij depend on their degrees. In particular, let the frequency of their interaction be discounted by $\frac{1}{d_i^\lambda d_j^\lambda}$ due to possible competition for attention. During information sharing, each agent initially has one unit of information. In each period from period 1 up to period T, each agent i shares $\frac{1}{d_i^\lambda d_j^\lambda} q$ fraction of her current information to each neighbor j. Notice that $q < 1$ is the original information sharing discount in A. Banerjee et al. (2013) that is due to loss of information. Then, agent i's information capital is a sum of all the information that she can receive from the network. The vector of agents' *information capital* is the modified diffusion centrality vector, modified to include possible competition for attention. Then,

$$DC(G; q, \lambda, T) \equiv \sum_{t=1}^{T} (q\tilde{G})^t \cdot \mathbf{1}, \quad \text{and} \quad \forall ij, \; \tilde{G}_{ij} = \frac{1}{d_i^\lambda d_j^\lambda} G_{ij}. \tag{2.12}$$

When $\lambda = 0$, this is the original diffusion centrality in (2.2), which assumes that in each period information is shared with probability q and information is useful if heard within T periods. When $\lambda > 0$, there is a tradeoff between the positive discounted utility from indirect neighbors and a negative effect due to competition with them for direct neighbors' attention. We say the *distance* between two agents is 2, if they are not connected but share a common neighbor. To highlight the tradeoff, we compare an agent's information capital with and without a distance-2 neighbor. Let $G \setminus \{k\}$ be the resulting network matrix removing its kth row and kth column.

Proposition 1 *Consider $T = 2$. For any agent i and any of her distance-2 neighbors k, there exists a threshold $\lambda_{ik} \in (0, 1)$ such that when $\lambda < \lambda_{ik}$, agent i's information capital is higher in network G than that in $G \setminus \{k\}$, and when $\lambda > \lambda_{ik}$, the comparison is reverse.*

All proofs are in Appendix A.1. This result shows that when λ is small, having more neighbors of neighbors increases one's information capital, whereas when λ is large (i.e., close to one), having more indirect neighbors decreases one's information capital. Thus, λ allows for extensive networks to be either beneficial or harmful.

Cooperation capital: support and 'interconnectedness'

Social networks also facilitate interactions that benefit from community cooperation and enforcement, such as risk sharing and social insurance. We model this dynamic following the setup of Ali and Miller (2016), which highlights the importance of *supported* relationships, where a link is supported if the two nodes of the link share at least one common neighbor (see also Jackson et al. (2012) and Miller and Tan (2018)).

As before, a population of N players are connected in an undirected network G, with $ij \in G$ and $ji \in G$ if agent i and j are connected (we abuse the notation of G slightly, which differs from the matrix format in the information model). Each pair of connected agents, $ij \in G$, is engaged in a partnership ij that meets at random times generated by a Poisson process of rate $\delta > 0$. When they meet, instead of sharing information, agent i and j now choose their effort levels a_{ij}, a_{ji} in $[0, \infty)$ as their contributions to a joint project.[37] Player i's stage game payoff function when partnership ij meets is $b(a_{ji}) - c(a_{ij})$, where $b(a_{ji})$ is the benefit from her partner j's effort and $c(a_{ij})$ is the cost she incurs from her own effort. We normalize the net value of effort a as $b(a) - c(a) = a$, and assume the cost function c is a smooth function satisfying $c(0) = 0$ and the following assumption.

Assumption 1 *The cost of effort c is strictly increasing and strictly convex, with $c(0) = c'(0) = 0$ and $\lim_{a\to\infty} c'(a) = \infty$. The "relative cost" $c(a)/a$ is strictly increasing.*

Strict convexity with the limit condition guarantees that in equilibrium effort is bounded. Increasing relative cost means a player requires proportionally stronger incentives to exert

[37] The variable-stakes formulation is adopted from Ghosh and Ray (1996) and Kranton (1996).

higher effort. All players share a common discount rate $r > 0$, and the game proceeds over continuous time $t \in [0, \infty)$.

As has been documented in several different real-world contexts, we assume agents have only local knowledge of the network. Specifically, we assume each agent only observes her local neighborhood, including her neighbors, and the links among these neighbors (in additional to her own links). To be precise, it is common knowledge that agent i observes each $j \in g_i \equiv \{i\} \cup N_i$, and all links in $G_i \equiv \{jk : j, k \in g_i\}$. In addition, we consider local monitoring, such that each agent learns about her neighbors' deviation (shirking behavior), and this information travels instantly.[38]

To begin, we seek to minimize contagion of deviation to the rest of the society off the equilibrium path, which follows from Jackson et al. (2012).

Definition 1 *A strategy profile is* ***robust*** *if an agent's deviation only affects partnerships involving herself and between her neighbors.*

Our first result shows that high levels of cooperation can be sustained in a robust manner, with agents needing only local information about the network and other agents' behavior.

Proposition 2 *For any network G, there exists a robust equilibrium of repeated cooperation that maximizes each agent's utility subject to agents' local knowledge of the network.*

Intuitively, each partnership ij uses the maximal level of effort subject to their shared common knowledge of the network. This maximal level of effort depends on the level of efforts i and j can sustain with each of their common neighbors k, which in turn depends on the level of efforts $\{i, j, k\}$ can sustain with their common neighbors l, and so on. Thus, this problem can be solved inductively, starting from the effort level of the largest clique(s) within $g_{ij} = g_i \cap g_j$, which always exists because the population is finite.

However, the optimal equilibrium in Proposition 2 could demand a high cognitive ability and a lot of computational capacity to solve, because one needs to solve (interdependent) effort levels for all subsets of neighbors in her local network. To address this concern, we instead focus on a simple equilibrium strategy profile that maintains the desired properties and sustains high levels of cooperation from the network enforcement.

To do so, we introduce two benchmark cooperation levels. The first one is *bilateral cooperation*, the maximal cooperation attainable between two partners without the aid of community enforcement.

[38]The local monitoring is stronger than the private monitoring in Ali and Miller (2016). It allows us to characterize the optimal equilibrium for any network under only local knowledge of the network, the counterpart of which is unknown with private monitoring (to the best of our knowledge), with the exception that Ali and Miller (2016) find the optimal equilibrium when the network is a triangle.

Bilateral cooperation Consider a strategy profile in which, on the path of play, each agent of the partners exerts effort level a if each has done so in the past; otherwise, each exerts zero effort. The equilibrium path incentive constraints are:

$$b(a) \leq a + \int_0^\infty e^{-rt}\delta a dt. \tag{2.13}$$

The bilateral cooperation level a^B is the effort level that binds the incentive constraint. Since the grim trigger punishment is a minmax punishment and each partner's effort relaxes the other partner's incentive constraint, these are the maximum efforts that can be supported by any stationary equilibrium that does not involve community enforcement.

Triangular cooperation Consider a triangle i, j, k and a strategy profile in which each of them exerts effort level a if each has done so in the past; otherwise, each exerts zero effort.

$$b(a) \leq a + 2\int_0^\infty e^{-rt}\delta a dt. \tag{2.14}$$

The incentive constraint is binding at effort level a^T. Notice that the future value of cooperation is higher in a triangle because there are two ongoing partnerships for each agent, so it can sustain higher level of efforts $a^T > a^B$ and everyone gets a strictly higher utility.

We characterize a particularly simple equilibrium strategy profile that further highlights the value of supported links. Recall that a link ij is *supported* if there exists k such that $ik \in G$ and $jk \in G$; i.e., if i and j have at least one common friend.

Corollary 1 *There exists a robust equilibrium in which any pair of connected agents cooperate on a^T if the link is supported, and on a^B otherwise.*

As the triangular level of effort can be sustained by three fully-connected agents, this strategy profile is robust. For example, consider a triangle ijk plus a link jk'. Even if k' has shirked on j, which reduces the value j gets from the partnership jk', it does not damage j's incentive to cooperate in the triangle ijk because it can sustain a^T by itself.

A benchmark model of migration

We now return to the migration decision. In equation (2.5), we assume that i's utility from a network contains information capital and cooperation capital (u_i^I and u_i^C); here, we further assume that the utility can be expressed as a linear combination of these two capitals. This stylized formulation is not meant to imply that u^I and u^C are orthogonal or that other aspects of the network do not weigh in the decision to migrate. Rather, this linear combination is intended to provide a simple benchmark that contrasts two archetypical properties of network structure, which we can also estimate with our data. Appendix A.2 develops a more general model of network utility, based on a network game approach, which allows for more complex

interactions among agents (for instance that an individual's utility may be affected by her position in the global network as well as her local network structure).[39] Appendix A.3 shows that similar results obtain when we consider a log-linear (Cobb-Douglas) utility function.

As outlined in Section 2.7, we say that agent i's information capital is proportional to their modified diffusion centrality $DC_i(q, \lambda, T)$, which is the i-th element of the vector in (2.12). We derive i's cooperation capital from Corollary 1 in Section 2.7, which implies that supported links are more valuable than unsupported links:

$$u_i^C = u_1 d_i^{NS} + u_2 d_i^S, \tag{2.15}$$

where d_i^{NS} is the number of i's unsupported links, d_i^S is the number of i's supported links, u_1 is the utility of cooperating on an unsupported link, and u_2 is the utility of cooperation on a supported link.

The overall utility is thus

$$u_i = u_0 DC_i(q, \lambda, T) + u_1 d_i^{NS} + u_2 d_i^S. \tag{2.16}$$

We will use this model to contrast the value of information capital against the value of cooperation capital, so we replace the parameters (u_0, u_1, u_2) by (π^I, π^C, α) and rewrite the overall utility:

$$u_i = \pi^I DC_i(q, \lambda, T) + \pi^C \left(d_i + \alpha d_i^S\right). \tag{2.17}$$

Substituting (2.17) into the original migration decision (2.1), we have

$$\begin{aligned} &\pi^{I,d} DC_i(G^d; q, \lambda, T) + \pi^{C,d} \left(d_i(G^d) + \alpha^d d_i^S(G^d)\right) \\ &\quad > \pi^{I,h} DC_i(G^h; q, \lambda, T) + \pi^{C,h} \left(d_i(G^h) + \alpha^h d_i^S(G^h)\right) + \varepsilon_i. \end{aligned} \tag{2.18}$$

Notice that we allow agents to have different weights $(\pi^{I,d}, \pi^{C,d}, \pi^{I,h}, \pi^{C,h})$ for the home and destination networks, because it is possible that the relative value of information and cooperation is different in a home network than in a destination network. For the same reason, we allow α to differ between home and destination networks. However, we assume (q, λ, T) are the same for home and destination networks, because they capture properties of the network that are common across agents and over which the agent has no direct control.[40]

[39] The network game approach follows in the tradition of Ballester et al. (2006), who use a network game to identify the key player, and König et al. (2017), who study strategic alliances and conflict. This approach is formally attractive, but since each agent's utility depends on their position and the entire network structure, it could not be realistically computed on our data. (As a point of comparison, calibration of the far simpler model (2.5) takes several days to complete, even after being parallelized across a compute cluster with 96 cores). See also Guiteras et al. (2019) for a related structural approach to dealing with network inter-dependencies.

[40] In particular, the 'destination' network of a given migrant is actually the 'home' network for all of the contacts that live in that destination.

Model Parameterization

We use the migration decisions made by several hundred thousand migrants over a 4.5-year period to estimate the parameters of model (2.18). The estimation proceeds in two steps. First, we draw a balanced sample of migrants and non-migrants by selecting, for every migrant who moves from h to d in month t, a non-migrant who lived in h in month t, had ≥ 1 contacts in d, but remained in h after t. This provides a total sample of roughly 270,000 migrants and non-migrants.

Second, we use simulation to identify the set of parameters that maximize the likelihood of generating the migration decisions observed in the data. The structural parameters of primary interest are λ, which we interpret as a measure of the competition or rivalry in information transmission; (α^h, α^d), the added value of a supported link, above and beyond the value of an unsupported link at home and in the destination; and the scaling coefficients $(\pi^{I,d}, \pi^{C,d}, \pi^{I,h}, \pi^{C,h})$, which together indicate the relative importance of information capital and cooperation capital at home and in the destination. We normalize $\pi^{C,h} = 1$, and follow A. Banerjee et al. (2013) by setting q equal to the inverse of the first eigenvalue of the adjacency matrix, $\mu_1(G)$, and $T = 3$.[41] Since a very large number of combinations of possible parameters exist, we use an iterative grid-search maximization strategy where we initially specify a large set of values for each parameters, then focus and expand the search around local maxima.[42]

Estimation appears to be well-behaved. For instance, Figure A.12 shows the home and destination utility values for all 270,000 individuals, using the parameterized version of model (2.18). Most of the true migrants (blue dots) have a predicted destination utility that exceeds their home utility; most of the true non-migrants (red dots) have a higher home utility. In aggregate, the calibrated model correctly classifies roughly 70% of the migration events.

To provide more intuition for the model estimation process, Figure 2.9 shows the estimation plots for λ; similar plots for the remaining five parameters are shown in Figure A.11. To produce these figures, we take all possible combinations of 6 parameters, resulting in roughly 50,000 different parameter vectors. We then simulate the migration decisions of the 270,000 migrants and non-migrants using model (2.18), and calculate the percentage of correct classifications. The figures show the the marginal distributions over a single parameter of the accuracy for the top percentile of parameter vectors. In most cases, the likelihood function is concave around the global maximum.

[41] When we treat q as a free parameter and estimate it via MLE, the likelihood-maximizing value of q is very close to $1/\mu_1(G)$. A. Banerjee et al. (2013) show that this approach to measuring diffusion centrality closely approximates a structural property of "communication centrality." However, we cannot directly estimate this latter property on our empirical network, which contains hundreds of thousands of nodes and tens of millions of edges.

[42] Specifically, for each possible set of parameters $< \lambda, \alpha^d, \alpha^h, \pi^{I,d}, \pi^{C,d}, \pi^{I,h} >$, we calculate the utility of the home and destination network for each migrant, and the change in utility after migration. If the change in utility of migration is positive, we predict that individual would migrate. We choose the set of parameters that minimizes the number of incorrect predictions.

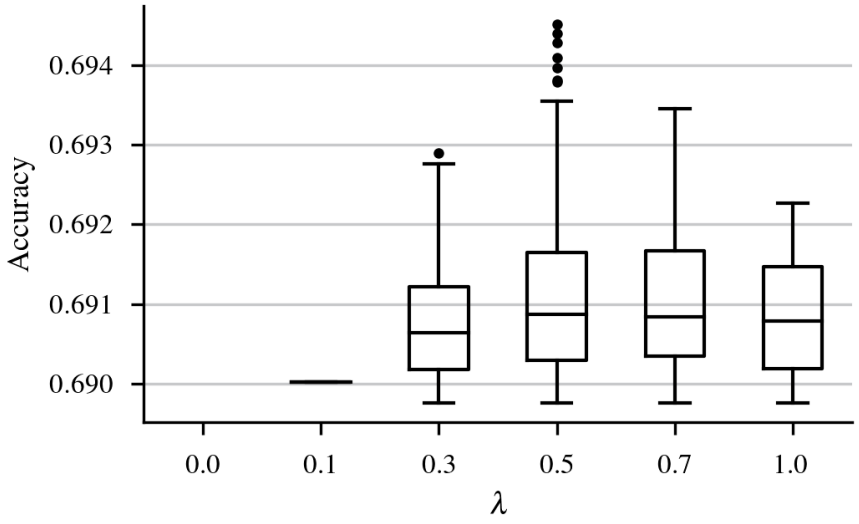


Figure 2.9: Model calibration results for λ. *Notes:* Figure shows the marginal effect of varying λ when calibrating Model (2.18). The full model has 7 parameters $(\lambda, \alpha^d, \alpha^h, \pi^{I,d}, \pi^{C,d}, \pi^{I,h})$; roughly 50,000 different parameter combinations are tested. The top percentile (by accuracy) of these combinations are selected. Each box and whisker plot represents the accuracy distribution within that top percentile, for each value of λ tested.

The structural model is largely being identified by the same variation that drives the reduced-form results. For instance, 97.5% of the variation in the total simulated utility of the destination network can be explained by the three main measures of network structure used in Section 1.5.[43] Moreover, when we take the *simulated* migration decisions $\widehat{M_{ihdt}}$ from the parameterized structural model, and estimate the equivalent of model (2.6) with $\widehat{M_{ihdt}}$ as the dependent variable, the regression results, presented in Table A.15, are qualitatively the same as the regression results using the actual migration decision M_{ihdt} (Table 2.2). The only notable difference is the effect of unique friends of friends in the destination network, which becomes significantly negative in Table A.15 and was insignificant in Table 2.2. This shows that when the rivalry parameter λ is optimally chosen for the structural model, the average effect of one's second-neighborhood becomes negative.

[43]Specifically, we regress the total *simulated* utility in the destination network, using the parameterized structural model, on three 'reduced-form' properties of the individual's social network: the destination degree centrality, the number of unique destination friends of friends, and the destination network support (see Section 2.4 for definitions). In this linear regression (no fixed effects), $R^2 = 0.975$.

Parameterization results

Estimation of the model yields several results. First, we find an optimal value of the rivalry coefficient at $\lambda = 0.5$, as shown in Figure 2.9. This suggests a significant departure from the benchmark information diffusion model of A. Banerjee et al. (2013): having friends who have many friends can actually reduce the utility that the agent receives from the network. The parameterized value of 0.5 implies that the probability of people sharing information with a neighbor is roughly inversely proportional to the (square root of the) size of their social networks. For instance, revisiting individuals A and C from Figure 2.1 (and assuming a two-period transmission model), with the parameterized $\lambda = 0.5$, we expect that A would receive 1.17 times the information capital as C. By contrast, the benchmark model with $\lambda = 0$ would imply that A would receive slightly less (0.99 times) information capital than C.

Second, using the information diffusion measure with the optimally parameterized rivalry coefficient, we find that the total utility from u_i^I (loosely, the 'information capital') and the total utility from u_i^C (loosely, the 'cooperation capital') contribute relatively evenly to the agent's total utility from the network. This can be seen most clearly in Figure 2.10, which shows the distribution of predicted utility from u_i^I and u_i^C for each of the individuals used to estimate the simulation. The bulk of this distribution lies around the 45-degree line, which is where $u_i^I = u_i^C$. This result is perhaps surprising given the reduced-form results presented in Section 1.5, which suggest that friends of friends in the destination have an insignificant (or negative) effect on the migration decision. However, a critical difference between the reduced form and structural results is that the structural results allow for rivalry in information transmission. To further confirm that it is the rivalry parameter drives this difference, we reestimate a version of model (2.18) where the rivalry coefficient is fixed at $\lambda = 0$. In other words, we use the original diffusion centrality (without λ) to measure the information capital and redo the whole simulation to identify the likelihood-maximizing set of parameters. We find that information capital (as the original diffusion centrality) contributes very little to total network utility; as shown in Figure A.13, the bulk of the distribution lies far below the 45-degree line, where $u_i^I < u_i^C$.

Third, and consistent with previous results, we find that supported links are valued more than unsupported links. This can be observed in the calibration plots for α^D and α^H in Figure A.11. In particular, $\alpha^d = 5$ implies that one supported link in the destination is six times as valuable as an unsupported link in the destination, and similarly, $\alpha^h = 1$ implies that one supported link at home is twice as valuable as an unsupported link at home.

Taken together, the structural estimates provide a micro-founded validation of the reduced-form results described earlier. This is an important step, since the reduced form results are based on statistical properties of networks that are correlated in complex ways, which cannot be easily accounted for in a regression specification. The model parameterization also provides independent support for the presence of some degree of rivalry in information diffusion — a possibility that was suggested by the heterogeneity discussed in Section 2.6, but only directly tested through structural estimation.

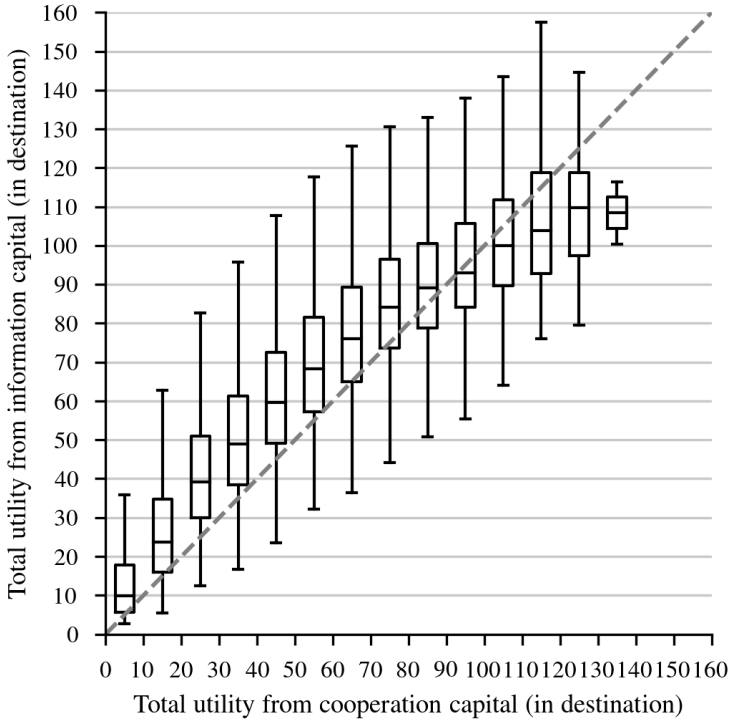


Figure 2.10: Calibration results: 'information' and 'cooperation' utility. *Notes:* Figure shows the distribution of predicted utility from 'information' capital and 'cooperation' capital (i.e., equation 2.5) for 270,000 migrants and non-migrants.

As a final step, Appendix A.3 examines the robustness of the parameterization results. In particular, we allow for the migration decision to include an average migration cost τ, which acts as a linear threshold that is constant across people, in addition to the idiosyncratic error that varies with each individual:

$$u_i(G^d) > u_i(G^h) + \tau + \varepsilon_i. \tag{2.19}$$

Separately, instead of the linear form of (2.18), we consider a Cobb-Douglas utility function which implies a log-linear combination of information capital and cooperation capital. Equation (2.18) becomes

$$\begin{aligned} &\pi^{I,d} \log DC_i(G^d; q, \lambda, T) + \pi^{C,d} \log \left(d_i(G^d) + \alpha^d d_i^S(G^d)\right) \\ &\quad > \pi^{I,h} \log DC_i(G^h; q, \lambda, T) + \pi^{C,h} \log \left(d_i(G^h) + \alpha^h d_i^S(G^h)\right) + \varepsilon_i. \end{aligned} \tag{2.20}$$

Results in Appendix A.3 show that the key qualitative results persist under these alternative specifications of model (2.18).

2.8 Conclusion

Social networks play a critical role in economic decision-making. This paper studies the decision to migrate in order to understand the value of social networks. Relative to prior work on the topic, our data provides uniquely granular visibility into the structure of social networks and the migration events they precipitate.

There are two main sets of findings. The first are specific to migration, and perhaps even to internal migration in Rwanda. These results establish several new stylized facts. Perhaps most surprising, we find that most migrants are *not* drawn to places where their social networks are extensive and diffuse. Our structural results suggest that this aversion may stem from the fact that migrants feel competition for the attention of their well-connected friends. By contrast, migrants respond strongly to the interlinkages of their friend and kinship networks, and are consistently drawn to networks that are interconnected and embedded. Such a finding is consistent with recent evidence that risk sharing and favor exchange play an important role in the migration decision (Morten, 2019; Munshi & Rosenzweig, 2016). But we also find that the notion of the "average migrant" can be a misleading generalization. Our data reveal rich heterogeneity, and we find that different types of migrants — including repeat, long-term, and short-distance migrants — value different properties of social networks differently.

The second set of results speak more generally to the utility that social networks provide to individuals embedded in those networks. In contexts ranging from product adoption (A. Banerjee et al., 2013) and disease transmission (M. J. Keeling & Eames, 2005) to the spread of new ideas and innovations (Kitsak et al., 2010; E. M. Rogers, 1962), simple models of information diffusion have seen remarkable success. Such models imply a prominent (albeit highly stylized) narrative that the primary function of networks is to diffuse information about economic opportunities cf., Ioannides and Datcher Loury (2004), Rees (1966). But the patterns revealed by our data are hard to reconcile with these models, and instead point to a model of network utility where repeated cooperation, and rivalry in information diffusion, play a more prominent role.

More broadly, we are hopeful that this study can illustrate the potential for novel sources of network data to provide deeper insight into how individuals derive utility from their social networks. Such data capture incredibly rich structure that reveal hitherto unobserved correlations between networks and consequential economic decisions. Through a combination of rich descriptives and structural estimation, we see great potential for future work aimed at understanding the value of social networks.

Chapter 3

Novel Approaches to Detecting Migration Events

3.1 Abstract

Empirical research on migration has historically been fraught with measurement challenges. Recently, the increasing ubiquity of digital trace data — from mobile phones, social media, and related sources of 'big data' — has created new opportunities for the quantitative analysis of migration. However, most existing work relies on relatively *ad hoc* methods for inferring migration. Here, we develop and validate a novel and general approach to detecting migration events in trace data. We benchmark this method using two different trace datasets: four years of mobile phone metadata from a single country's monopoly operator, and three years of geo-tagged Twitter data. The novel measures accurately reflect existing knowledge of migration in these contexts, and also provide more granular insight into migration spells and types than what are captured in standard survey instruments.[1]

3.2 Introduction

Migrants play an important role in all aspects of modern society. It is estimated that about 0.6% of the world population migrated internationally from 2005 to 2010 (Abel & Sander, 2014). As many as 750 million people in the developing world are permanent internal migrants (Lucas, 2015). Understanding the causes and effects of migration is a central focus of social science research. For research and policy, it is thus critical to have an accurate quantitative understanding of the scale and scope of migration.

However, empirical research on migration has historically been hindered by the lack of granular migration data. Traditional data on migration are typically derived from population censuses or sample surveys, and are usually based on questions about place of birth and recent

[1]The material in this chapter is based on joint work with Fengyang Lin, Guangqing Chi, and Joshua Blumenstock. A General Approach to Detecting Migration Events in Digital Trace Data.

migrations. But census and surveys are expensive and time-consuming, and are plagued by issues of attrition since migrants, by definition, do not remain in the same place (Bell et al., 2015; Lucas, 2015).

Over the past decade, the mass proliferation of digital devices has created large repositories of 'digital trace' data, which provide new opportunities to measure and model human mobility. The data most commonly used in such studies are collected by mobile phone networks or social media platforms. While the majority of such studies focus on local mobility (Gonzalez et al., 2008; Jurdak et al., 2015; Song, Qu, et al., 2010), several more recent papers have used such data to analyze migration (J. E. Blumenstock, 2012; Hong et al., 2019; Zagheni et al., 2014). In turn, migrant flows have been used to study labor markets (Barwick et al., 2019; J. Blumenstock et al., 2019; Büchel et al., 2019), infectious diseases (Wesolowski et al., 2016; Wesolowski et al., 2012; Wesolowski et al., 2015), disaster response (Bengtsson et al., 2011; Lu et al., 2012), and other social phenomena linked to migration.

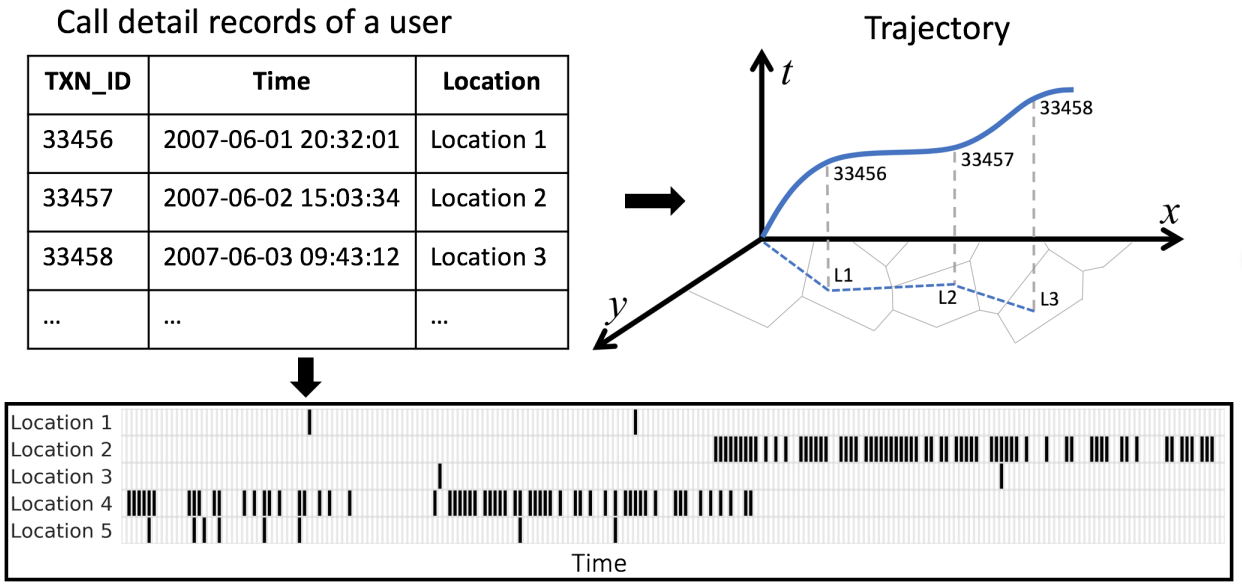


Figure 3.1: Extracting human trajectories from trace data. *Notes:* Raw data (*top left*) contains timestamps and geo-coordinates each time each individual is active on the platform (e.g., making a phone call). From these data, the trajectory of the person through space and time can be reconstructed (*top right*). The bottom figure shows the set of locations (e.g., neighborhoods) in which the individual was observed on each day.

A stylized representation of these data, and how they can be used to reconstruct human trajectories, is shown in Figure 3.1. The top-left table shows the 'raw' data that is logged by, for instance, a mobile phone operator. These transactions can be mapped to physical locations, which indicate the person's trajectory (top right map). We will also use two-dimensional arrays (bottom figure) to visualize location decisions over a long time horizon.

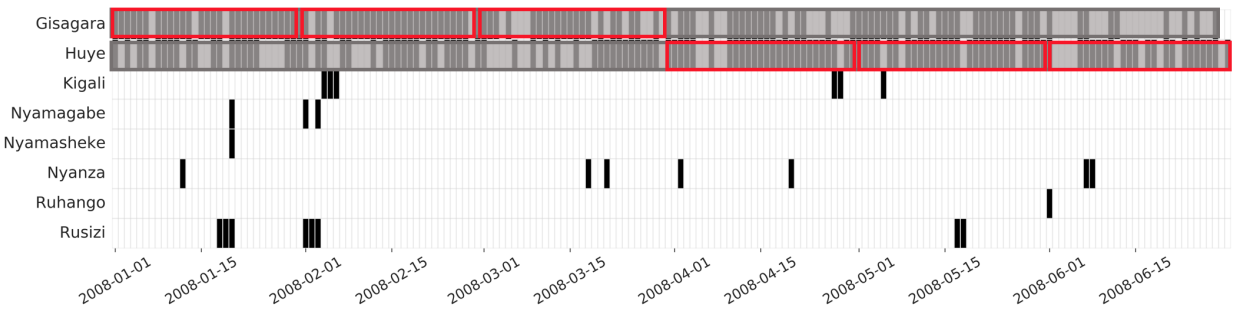


Figure 3.2: One individual's locations over six months.*Notes:* Each row is a different district, each column is a day; cells are black if the individual makes or receives a phone call from that district on that day. The red boxes show the individual's modal district in each month.

Most existing studies of migration based on trace data rely on relatively *ad hoc* methods for measuring migration. Prior work typically breaks the problem down into two distinct steps: First, the data are used to infer the "home location" of an individual at a certain point in time; and second, the sequence of home locations is used to infer migration events. This approach, while simple, has several important limitations. For instance, most studies make the assumption that an individual's 'home' location is the cell tower or the city from which they make the most of the calls or post the most tweets in a defined time period.[2] But such inferences are not robust to the bursty behavior that has been well-documented in phone and media use (Z.-Q. Jiang et al., 2013). More generally, this notion of 'home' is brittle to diurnal patterns (e.g., home vs. work device use) and the measurement technology used (for instance, the fact that phone networks load balance by shifting calls from high-volume cell towers to neighboring low-volume cell towers). An example of how such inferences can go wrong is shown in Figure 3.2, which shows the different locations in which a single mobile phone subscriber was observed over a 6-month period in Rwanda. In the case, the individual's most frequent location for the first three months was different from the most frequent location for the last three months, but in actuality the individual simply lived in the border region between the two locations, and did not actually migrate (the modal cell tower of this mobile phone subscriber in the first three months is roughly 500 meters from the border with Huye).

To fill these gaps in the literature and enable future empirical work on migration, this paper develops a novel and general approach to detecting migration events in large-scale digital trace data. We begin by proposing a new segment-based algorithm for migration detection. This algorithm works by first grouping contiguous segments in time when an

[2]For example, Zagheni et al. (2014) and Fiorio et al. (2017) assign users to the county from which they posted the majority of tweets during a specified period of time. Papers using phone data typically assign home locations based on the cell tower or administrative unit with the densest call activities (J. E. Blumenstock, 2012; Hankaew et al., 2019; Hong et al., 2019; Lu et al., 2016; Phithakkitnukoon et al., 2011). See Appendix Table B.1 for a full inventory of the prior work using trace data to study migration.

individual is likely to be in the same location (subject to some random deviations), and then identifying persistent changes in those segments over time. The algorithm is intuitive, and contains tuning parameters that make it possible to flexibly detect both short-term displacement and long-term migration. It also makes it possible to identify the likely date of migration, and provides a confidence intervals for each inferred migration event.

We then conduct a series of experiments to calibrate and validate this new algorithm using mobile phone and Twitter data. In particular, we hired a team of students to hand-label 1,000 migration diagrams (similar to Figure 3.2), and compared these hand labels to our algorithm's predictions, and to those of alternative methods in the literature. We show that our approach is substantially more accurate than traditional approaches.

By providing a more coherent and robust framework for measuring migration, we hope this study can improve the set of tools available to applied researchers, and in turn advance empirical research on migration. The method we develop is 'data-agnostic' in the sense that it can be applied to any dataset where individuals have spatial and temporal markers. To facilitate adoption by the research and policy community, we have packaged the algorithms into a set of tools that are implemented using an open-source Python-based library.[3]

3.3 Background and Related Work

Empirical analysis of human migration dates back at least to 1885, when Ravenstein analyzed 1881 British census data with the information of birthplace and residence place (Ravenstein, 1885). The research was done at a time of a large number of migrants after the Second Industrial Revolution (Nestorowicz & Anacka, 2019). Ravenstein summarized seven "laws of migration", such as "the great body of our migrants only proceed a short distance, and that there takes place consequently a universal shifting or displacement of the population" (Greenwood & Hunt, 2003; Pisarevskaya et al., 2019).

More recently, the research literature has defined a migration event as "a change in the place of usual residence, which also involves crossing a recognized political/administrative border" (White, 2016). In practice, this usually involves specifying a temporal dimension and a spatial dimension (A. Rogers et al., 2003; Willekens, 2008). The temporal dimension indicates some fixed length of time in which an individual must remain in a location for residency to be established; the spatial dimension typically involves crossing international or internal administrative boundaries (Union, 2016). For instance, the US Census asks the residence of households one year ago and the year when households came to the current residence house.[4] The World Bank's Livings Standards and Measurements Survey, conducted primarily in developing countries, similarly contain a migration module that queries place of birth, year that households moved into the current housing unit, and so on (White, 2016; World Bank, 2009).

[3] See `migration_detector`, available at https://github.com/g-chi/migration_detector

[4] https://www.census.gov/topics/population/migration/surveys-programs.html

Over the past decade, a handful of studies have used novel sources of spatiotemporal 'trace' data to observe human mobility and migration with much greater spatial and temporal granularity.[5] Early research in this area used mobile phone data to characterize patterns of human mobility. For instance, Gonzalez et al. (2008) show that human mobility is highly regular and follows a truncated power-law distribution. Song, Qu, et al. (2010) similarly find that human mobility is highly predictable, and Simini et al., 2012 and others develop statistical models of human mobility.[6]

More recent work has used digital trace data to study human migration. This body of work, and the way the data are used to measure migration, is summarized in Appendix Table B.1. In work most closely related to the current study, J. E. Blumenstock (2012) proposed a rudimentary method for inferring migration from phone data, which defined a migration event as one in which an individual remains within one administrative unit for k consecutive months and then a different administrative unit for k consecutive months. This approach, or slight variations of it, have subsequently been used to study migration using phone data (J. Blumenstock et al., 2019; Hankaew et al., 2019; Phithakkitnukoon et al., 2010) and social media data (Zagheni et al., 2014).[7]

This paper departs from prior work by developing and validating a more robust approach to inferring migration from spatiotemporal trace data. In the next section, we describe this approach and show how it can be used to identify migrants, infer migration dates, and provide measures of confidence for each migration event. Section 3.5 then walks through a series of examples to build intuition for how the algorithm works on data. In Section 3.6 we calibrate and validate the algorithm by comparing the algorithm's predictions to those of human judges and alternative approaches. Section 3.7 concludes.

3.4 Detecting Migration: A 3-Step Algorithm

In this study, we define a migration event as one where an individual's primary residential location remains stable for some minimum amount of time, and then changes to a different location for another minimum amount of time. Following J. E. Blumenstock (2012), we define

[5] We think of 'trace' data as that produced as the result of people's ordinary activities that leave behind a digital footprint, rather than data produced specifically for the purpose of scientific study (Salganik, 2017). *Spatiotemporal* trace data contain spatial and temporal markers.

[6] See also (Chen et al., 2016; Csáji et al., 2013; S. Jiang et al., 2017; Kang et al., 2012; Kang et al., 2013; Lu et al., 2013; Phithakkitnukoon et al., 2010; Schneider et al., 2013; Shi et al., 2015; Song, Koren, et al., 2010; Williams et al., 2015). Barbosa et al. (2018) provide a review.

[7] An important related body of work grapples with the ethical considerations inherent in using digital trace data to study human behavior in general cf. Boyd and Crawford, 2012; Eubanks, 2018, and the privacy concerns that arise in using such data to study human mobility specifically (De Montjoye et al., 2013; Taylor, 2016). The methods we describe in this paper are meant to provide more accurate and robust measurements of human migration, and while our goal is to enable social science research and pro-social applications (e.g., rapid disaster response), we acknowledge the potential for anti-social uses (e.g., discrimination against at-risk populations). The analysis we conduct here uses de-identified data and is governed by strict IRB protocols; we can only urge subsequent applications and extensions to use these methods and data responsibly.

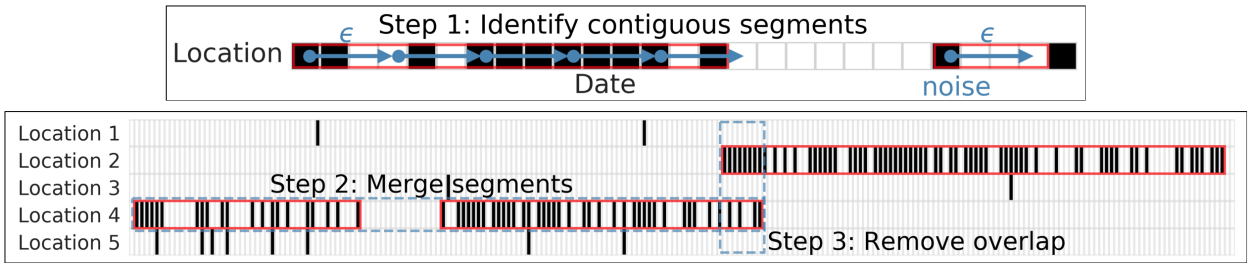


Figure 3.3: Detecting location segments. *Notes:* Step 1 identifies segments where an individual is at a location continuously, with no gaps exceeding ϵ days. Red boxes in the bottom figure are detected segments.

a 'minimum amount of time' flexibly, using a parameter k that can be easily changed to suit the application domain. While most of our empirical examples fix $k = 3$ months, longer-term migrations and shorter-term displacements could be studied analogously by increasing or decreasing k. We define a 'location' as a pre-existing administrative division, such as a district in Rwanda or counties in the United States.

Using this empirical definition of "migration", a key contribution of this paper is to design an algorithm that operationalizes this definition on digital trace data. The algorithm, which is included as Appendix Algorithm 3, operates in three steps: First, it detects contiguous location segments in the raw trace data. Then, it determines which of those segments constitute 'migration' events. Finally, it infers an exact migration data, and assigns a confidence interval to each migration event.

Detecting location segments

The first step in detecting migration requires detecting periods of time when an individual is continuously present in a single location, allowing for some margin of travel from that location (for instance, for an evening on the other side of the city, or for a weekend trip out of town). We accomplish this in 3 substeps. A schematic of this process is depicted in Figure 3.3. The full details of this algorithm are given in Algorithm 3, and summarized below.

Step 1: Identify contiguous segments. The goal of this step is to identify contiguous periods of time during which a person remains in the same location.[8] We use a clustering algorithm, similar to DBSCAN (Ester et al., 1996), that finds periods when an individual remains at a single location continuously, with no gap exceeding ϵ days. To allow for idiosyncratic deviations, from a primary location, we consider all segments where the individual is observed in that location on at least *propDays* percent of

[8]For simplicity, we assume the raw latitude/longitude coordinates can be resolved to locations with preexisting administrative boundaries (such as neighborhoods or municipalities).

days in the segment. Finally, we eliminate segments that are less than $minDays$ days in length.

Step 2: Merge segments. This step merges neighboring segments together if there are no segments in other locations between them.

Step 3: Remove overlap. This step resolves situations when an individual is associated with segments in multiple locations at a single point in time.

Detecting migration

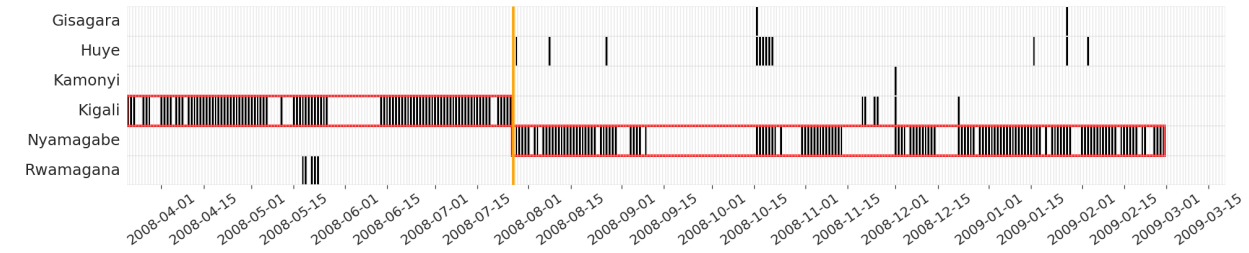
After contiguous location segments are identified, migration events are defined by the existence of two neighboring segments with different locations. This requires specifying a minimum residency k for the individual to be in each location. As there is no universal definition of residency length (Union, 2016), we expect different applications to use different values for k. In the empirical examples below, we use $k = 90$ days.

Inferring the date of migration

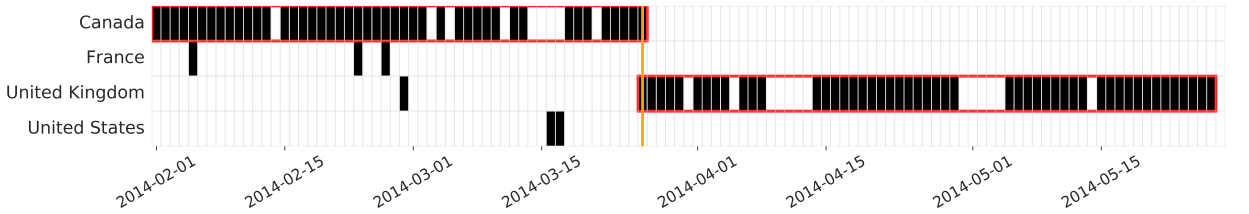
We also design a method to infer the exact date on which a migration occurs, rather than, say, simply recording that a person was in one location in one month and a different location in a subsequent month. In cases where there is a discontinuous break such that an individual appears only at one location until a specific day t, and then only at a different after that day, then we simply say that the person migrated on day t. Often, however, there is some ambiguity, such that an individual is observed in both one location and another, and the exact migration date is not obvious (as in the dotted blue region of Figure 3.3). In such cases, we select the day between the start of the new segment and the end of the old segment that minimizes the number of 'misclassified' days, i.e., the number of days when the migrant appears at destination before the migration date and days when the migrant appears at home after the migration date. In cases where multiple days yield the same number of misclassifications, we select the last day as the migration date.

Measuring the uncertainty of migration dates

For every migration date detected through the above algorithm, we attach a measure of confidence that an actual migration occurred. Confidence measures are useful because, as can be seen in Figure 3.2, the raw data are often quite noisy (and occasionally very sparse), and some migration events are more ambiguous than others. To evaluate the uncertainty of the inferred migration dates, we use the number of gap days between the start of the new segment and the end of the old segment after removing overlap. The larger the gap is, the higher uncertain our estimation is.



(a) Long-term migration in Rwanda (migrated from Kigali to Nyamagabe on 2008-07-27)



(b) International migration (migrated from Canada to United Kingdom on 2014-03-26)

Figure 3.4: Two migration trajectories. *Notes:* Red boxes indicate location segments. Orange line marks the inferred date of migration.

3.5 Empirical Example

Above, we describe a general algorithm for detecting migration events in trace data. To provide some intuition for how this algorithm works in practice, we show the results when applied to two different sequences of location data. Figure 3.4 displays the location history, location segmentation, and inferred migration dates from one mobile phone traces (a) and one Twitter trace (b). Figure 3.4a shows the trajectory of a long-term migrant from Kigali to Nyamagabe. Even though some noise exists in this individual's trace data (i.e., they are seen in multiple locations and there are many days without any data), the method identifies a migration event. Note that in this example, the standard approach of first identifying primary locations in each month and then inferring migrations would fail, since the modal location in October 2008 is Huye rather than Nyamagabe. Figure 3.4b shows the trajectory of an international migrant who moves from Canada to the United Kingdom. Compared to within-country migration using mobile phone trace data, international migration using geo-tagged tweets have less noise and is relatively easier to detect segments.

3.6 Experiments and Validation

The preceding examples provide some intuition for how the algorithm works when applied to data. To provide more rigorous quantitative validation of the method, we run a series of experiments with human judges. This makes it possible to compare the performance of

this algorithm to traditional approaches, highlight the overall robustness of this method, and show how the algorithm's parameters can be tuned.

Experimental design

Digital trace data

To help illustrate the generality of our approach, we perform experiments on two different digital trace datasets.

- **Mobile phone data.** We use a dataset of mobile phone Call Detail Records (CDR), which contains 4.5 years of activity for roughly 1.5 million de-identified individuals in Rwanda. Each time a mobile phone owner makes or receives a call or text message, a new entry is generated in this dataset which contains a unique identifier for the caller and receiver, a timestamp for the event, and the approximate location of the caller and receiver (i.e., the geo-coordinates of the nearest cell phone tower).
- **Twitter data.** The geo-tagged tweets we use include 20,000 randomly selected Twitter users in the U.S. over two years who have at least 1,000 geo-tagged tweets. Each record in this data contains a unique identifier for the individual, the timestamp of the tweet, and the geo-coordinates from which the tweet was posted.

Note that the key commonality between these two datasets — and what is required for our algorithm to work — is that a single transaction record in both datasets contains spatial and temporal information.

Validation data

To validate our approach, we hired a team of five undergraduate students to build a labeled corpus of migration data. Each labeler was randomly assigned a large number of 'samples', where each sample contains a trajectory for a single individual (see Figure B.1 for an example). For each sample, the human labeler was required to indicate (i) whether a migration took place; (ii) how confident they are in that assessment on a scale of 1 to 3. In addition, if a migration was marked, the labeler was asked (iii) the date of migration; (iv) their confidence in that date at an interval of 5 days (0-5, 6-10, 11-15, >15); (v) the first day and last day of home segment and destination segment.

In total, the labelers provided labels for 1,000 different migration trajectories. These 1,000 samples were drawn strategically to compare and contrast our new *segment-based* approach and the 'traditional' *frequency-based* approach used in most prior work, which first assigns individuals to locations in each month (based on the location in which the majority of events occur in that month) and then classifies migrants as individuals whose location changes between subsequent months. Specifically, we drew: (1) 250 samples where both algorithms detect a single migration; (2) 250 samples where neither algorithm detects a

migration; (3) 250 samples where the new method detects a migration but the traditional method does not; and (4) 250 samples where the traditional method detects a migration but the new method does not.

Experimental Results

We compare the performance of the new method of detecting migration to the traditional method in Table 3.1. The accuracy of our method is 81.5%, much higher than 62.7% of the traditional method.

Our method has high accuracy on the estimated migration dates and home and destination segment length. Figure 3.5 shows that most of the migrants have a very small difference between real migration dates and estimated migration dates, implying that our approach has a good performance in estimating migration dates. Figure 3.6a and 3.6b confirm that our approach can also find reasonable residency length.

Table 3.1: Performance of the two approaches at labelers' different levels of confidence

Method	Uncertainty by labelers	Accuracy	Precision	Recall	F1	# Samples
frequency-based	overall	0.627	0.600	0.634	0.617	1000
	not confident at all	0.465	0.351	0.521	0.419	61
	somewhat confident	0.530	0.521	0.603	0.559	264
	very confident	0.695	0.697	0.656	0.676	585
Segment-based	overall	0.815	0.788	0.833	0.810	1000
	not confident at all	0.579	0.457	0.718	0.558	84
	somewhat confident	0.749	0.717	0.811	0.761	355
	very confident	0.877	0.884	0.859	0.871	766

Why does the new segment-based approach out-perform the traditional frequency-based approach? One common (false-positive) error of the traditional method is to falsely identify as migrants people who frequently appear in neighboring locations (as in Figure 3.2). Another common (false negative) error of the traditional method is to erroneously classify transient displacement as a change of primary location (as in Figure 3.4a), so that the person does not remain stable for long enough to be classified as a migrant. In both these situations, the segment-based approach performs better.

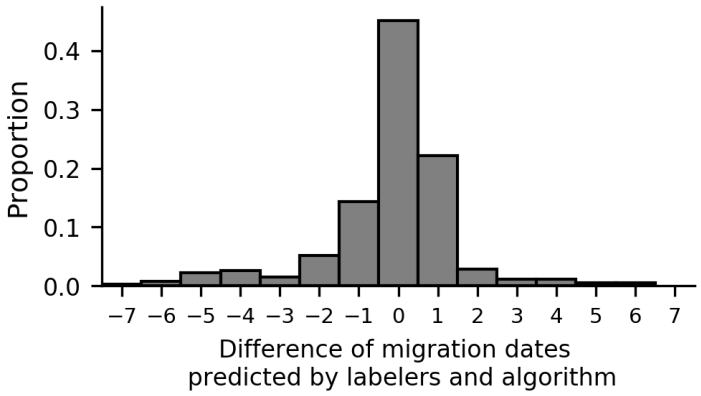
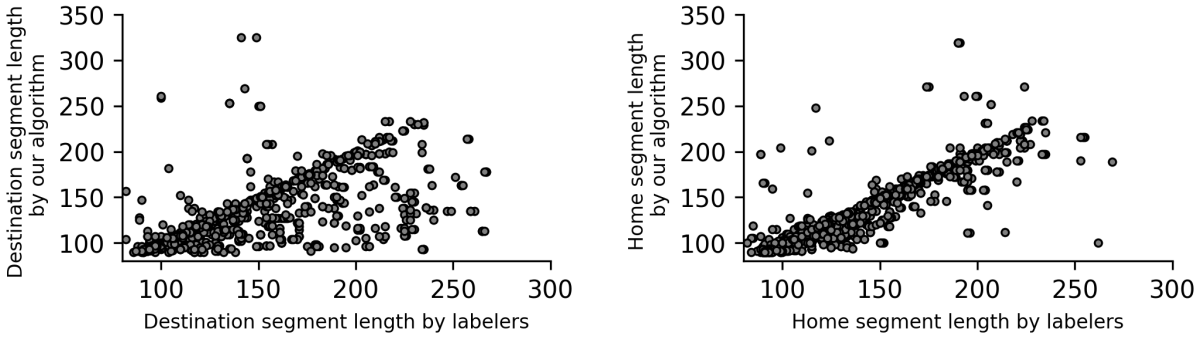


Figure 3.5: Distribution of migration date difference between our approach and labelers.



(a) Distribution of destination duration

(b) Distribution of home duration

Figure 3.6: Accuracy of residency length at home and destination.

Results by sample difficulty

Since many of the samples in the dataset are ambiguous (as in Appendix Figure B.2), we asked labelers to indicate their own level of confidence in classifying the sample as a migration or non-migration. Importantly, we find that the segment-based method performs better than the frequency-based method irrespective of the ambiguity of the underlying sample. This result can be seen in Table 3.1, which disaggregates the performance of each algorithm by the (human-classified) ambiguity of the sample. The new method is at least 10 percentage points more accurate than the traditional method for each of the three types of samples. Note that for those cases where labelers have different opinions on whether a migration takes place in

a sample, we only keep the results agreed by two labelers in this table. This is the reason why the total number of samples in the three categories of uncertainty is different for the two approaches.

Qualitative validation of Twitter samples

While our main experiments focus on the mobile phone data, we also spent some time manually validating the performance of the segment-based method on Twitter data. Specifically, we pulled the Twitter histories of 100 individuals who the segment-based algorithm identified as migrants, and analyzed the contents of their tweets immediately before and after the date of migration inferred by the algorithm. In 91% of these cases there was direct evidence in the contents of the tweets that a migration indeed occurred on the inferred date. To provide one example, Table 3.2 contains the tweets posted by one individual in the time surrounding migration. The algorithm infers a migration date of September 4, which is consistent with the text of the tweets.

Table 3.2: Selected tweets of a detected migrant who moved from Virginia to New York on 2014-09-04 based on our approach.

Date	Tweet
2014-08-28	hmm. second to last day in Charlottesville! #daydreamin
2014-08-29	Moving out of Cville today #ahhhh
2014-09-04	Walking around Union Square with this crazy beautiful weather got me like whoa #nyc
2014-09-07	What else to do in nyc when it's 85 and gorgeous?? Go to the beach! #forttilden

Uncertainty of migration dates

As described earlier, our algorithm associates a measure of confidence with each inferred migration date. We compare this measure of confidence/uncertainty to the level of uncertainty assigned by the human labeler, for the set of samples where the human attempted to assign a migration date to the sample. The strong correlation between the uncertainty of our method and of the human judges is shown in Figure 3.7. As human labelers become less certain about the date of the migration, so too does the algorithm's uncertainty increase.

Tuning algorithm parameters

To understand how the key algorithm parameters impact the resulting migration classifications, we show the impact different parameter values have on the F1 score, based on the

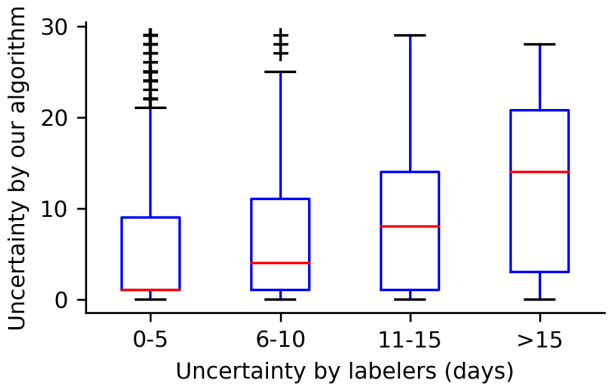


Figure 3.7: Distribution of uncertainty between our approach and labelers.

sample with human labels.[9] Our approach contains three parameters: the maximum gap between consecutive days ϵ, the minimum number of days in a segment $minDays$, and the minimum proportion of days in a segment $propDays$. The relationship between each of these parameters and algorithm performance is shown in Figures 3.8a, 3.8b, and 3.8c.

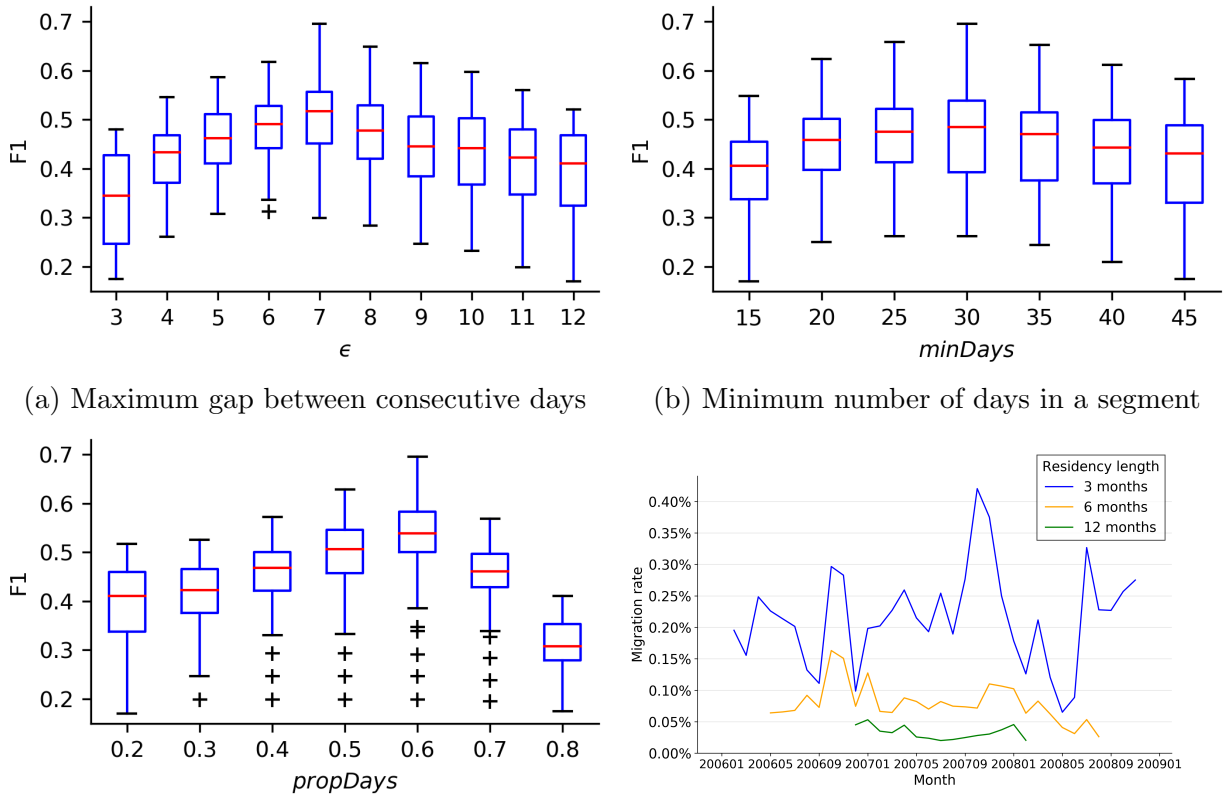
Small values of ϵ overlook cases where a person does not appear in the home location for a few days. But large ϵ generates very long segments, which can leads to substantial overlap with other segments. When $minDays$ is large, it requires a longer sequence of consecutive sightings at the same location, which in turn decreases the number of detected migration events. But if $minDays$ is too small, more segments with a long overlap will be found, decreasing the number of detected migration events. The effect of $propDays$, the proportion of appeared days in segments, is similar to $minDays$. In our sample, based on the performance on the labeled dataset, the optimal tuning of these three parameters is $\epsilon = 7$, $minDays = 30$, and $propDays = 0.6$. Of course, different contents may dictate a different optimal combination.

In settings that are qualitatively different than ours, one would ideally tune these parameters through cross-validation, for instance hand-labeling a sample in order to produce diagnostics similar to those in Figure 3.8. Absent such labels, it is still possible to tune the main parameters by observing the impact of different parameter combination on trajectory maps as those shown in Figure 3.4. As a general rule of thumb, larger values of $minDays$ are useful for detecting long-term migration (to avoid including short-term displacement), whereas small values detect rapid moves. ϵ and $propDays$ will depend on the frequency and volume of trace data for each individual — if individuals appear almost every day, smaller ϵ and larger $propDays$ are appropriate.

Finally, it is worth noting that in addition to the main parameters of the model, perhaps the most important 'hyper-parameter' is k, the minimum time an individual must reside in

[9] F1 is defined as $2 * \frac{precision * recall}{precision + recall}$, and provides a balanced measure of recall and precision.

one location to be considered a resident, as discussed at the beginning of Section 3.4. Different values of k allow for different types of short- and long-term migrations to be counted. As expected, when migration requires a longer residency length, fewer migrants are detected. This effect is evident in Figure 3.8d, which shows the migration rate implied by different thresholds of $k = 3$ months, 6 months, and one year. Also evident in the figure is a seasonal migration pattern in January every year, a time when the migration rates almost double.



(a) Maximum gap between consecutive days

(b) Minimum number of days in a segment

(c) Minimum proportion of days in a segment

(d) Migration rate over time in Rwanda

Figure 3.8: The effect of parameters on performance.

3.7 Conclusion

With the increasing prominence and ubiquity of large-scale digital trace data, new opportunities are emerging to study human migration. While a handful of studies have demonstrated this potential, no prior work has carefully considered or validated the computational methods used to infer human migration from these new sources of 'big' data. The most common 'frequency-based' approach to inferring migration events often results in mis-classifications.

This paper developed a new segment-based approach to measuring migration, and carefully validated the method using a large corpus of migration data labeled by humans. In addition to more accurately classifying migrations, the segment-based approach makes it possible to identify the exact date of migration, and attaches a measure of confidence to each migration event. We have packaged the algorithm in an open-source Python library that is available for public use and modification on GitHub.

While we believe this work represents an important step for researchers using new sources of data to study migration, our hope is that future researchers will continue to adapt our specific algorithm to other data and contexts. Indeed, beyond the specific segment-based algorithm we have proposed, a broader contribution of this paper is to describe a rigorous quantitative framework for evaluating new methods for measuring migration from trace data. In particular, by providing new techniques for visualizing individual location trajectories over time, and by showing how human judges can be used to label and cross-validate algorithmic classifications, we hope to lay the groundwork for future researchers to rigorously document new algorithms that out-perform our own.

Chapter 4

Evolution of Migrants' Social Networks

4.1 Abstract

Personal social networks evolve over time to adapt to major life events, such migration, marriage, and employment. This paper studies how migrants' social networks change over the migration and settlement process. We use a rich mobile phone dataset that allows us to observe the evolving social networks of a nation's worth of migrants, to characterize changes in network structure before and after migration. We document stark and systematic changes in this structure: Within two months of migrating, migrants cease communication with nearly half of their former contacts in their place of origin; these "lost" relationships are almost exactly offset by the 55% increase in new connections with people in the destination. We show that friendship persistence and loss is highly predictable: the social ties most likely to persist are those that have frequent communication, and which have fewer friends.[1]

4.2 Introduction

Personal social networks evolve over time to adapt to migration-related life events, such as marriage and employment (Bidart & Lavenu, 2005; Lubbers et al., 2010). Social relationship plays an important role in migrants' well-being. Migrants might lose social support in origin and feel loneliness after migration (Koelet & de Valk, 2016; Roberts & Dunbar, 2015). Understanding the social network evolution of migrants would help us investigate the strategies migrants adopted to secure their social capital (Jackson et al., 2012; Nisic & Petermann, 2013).

Recent studies have found contradictory results on the network change of migrants. For example, some studies have found that migrants build new connections in the new place

[1]The material in this chapter is based on joint work with Joshua Blumenstock. Understanding social network evolutions of migrants with population-scale network data.

and lose old connections in the original place (Bidart & Lavenu, 2005; Lubbers et al., 2010; Nisic & Petermann, 2013). However, Bolíbar et al. (2015) established that few migrants lost their connections with the original country. One possible reason for the different findings is that traditional analyses on migrant social networks in sociology usually interview migrants for a few waves, which have years of interval and have limited capability to collect dynamic information on social networks, especially during the month of the migration progress (Bidart & Lavenu, 2005; Gill & Bialski, 2011; Lubbers et al., 2010). It is still not clear how migrants' social networks evolve, and what strategies migrants use to integrate into local communities and maintain their friendship in origin.

Different from prior studies using surveys that were collected several years after migrants live in a new location, this research seeks to identify how migrants' social networks evolve months before and after migration by applying the social network analysis approach based on call detail record (CDR) datasets at a very high temporal resolution. The CDR datasets contain both the movement trajectory, which allows us to detect migrants, and dynamic information of the social connections. We focus on the settlement process in the first few weeks pre- and post-migration when migrants need most of the help, in the formats of accommodation, credits, and jobs, which lead to the temporal and spatial dynamics of their social networks (Dolfin & Genicot, 2010).

We examine migrants' network changes in the settlement process from both the network level and the tie level. At the level of personal networks, we find that the percentage of connections with friends in origin decreases by half after migration, while the percentage of connections with destination friends is almost doubled to 55%. Clustering coefficient decreased in destination after migration, suggesting that migrants start making more diverse friends instead of strengthening their existing contacts. At the tie level, we find that the strong ties and densely connected ties are more likely to persist.

Based on the insights from the results of network evolution, we use the Cox proportional hazards model to investigate the factors determining the formation of migrants' social ties and help understand migrants' strategies of forming interpersonal links in the new environment. Our regression shows that migrants who have more friends, fewer calls, and smaller radius of gyration are more likely to lose connections, and ties with more interactions are less likely to dissolve.

To summarize, the main contribution of this work is that we depict the evolution of migrants' social network structures in a fine time-resolution before and after migration. Lubbers et al. (2010) mentioned that "the network dynamics that we observed seem to be comparable with those observed in non-immigrant samples. ... We might find more compositional and structural changes among more recent migrants." With the population-scale digital trace data, we shift the focus to the period when migrants just moved to a new location and examine the question that cannot be answered based on the traditional surveys.

4.3 Related Work

There have been many traditional studies on social network evolution of migrants using cross-sectional datasets rather than longitudinal datasets (Bolíbar et al., 2015; Gill & Bialski, 2011; Nisic & Petermann, 2013). They asked migrants their residency length, based on which, social network changes over time are measured at the aggregated level. For example, Bolíbar et al. (2015) interviewed immigrants from Ecuadorian and Moroccan and asked how long they had resided in Catalonia. The cross-sectional datasets could measure network structures at the aggregated level, but have limited capability to capture individual network changes. One exception is the work by Lubbers et al. (2010), who interviewed 25 Argentineans in Spain twice in a 2-year interval. This interview was conducted after migrants had already settled in the destination for a long time rather than over the migrants' settlement process in the first few weeks pre- and post-migration when migrants are more likely to change network connections to find financial and physical help.

Two recent work has studied network evolution of migrants using CDR which can detect individual network change at a high temporal resolution. Phithakkitnukoon et al. (2011) analyzed the social network change of migrants and found that migrants need seven to eight months to reconstruct a new social network in terms of the average tie distance. However, according to the definition of migrants in this research, a person will be defined as a migrant as long as his or her location changed. This is problematic because people might change location temporarily for vacation. Yang et al. (2018) found that staying migrants have more friends and longer moving distance than temporary migrants. But this study did not focus on the evolution trend of migrants' social networks probably because it is hard to tell the change based on their one-month CDR dataset.

Migrants' social networks change in various formats. Lubbers et al. (2010) proposed a general model of personal network change of migrants, including maintaining kin in the origin, having new ties in the host city, contacts in the origin decreasing, and people from different clusters interconnecting. This framework, focusing on the changes after migration, ignored the network changes before migration, which is crucial to understand migrants' behaviors to prepare for the migration. We will detect the network changes before and after migration according to the four directions proposed by Feld et al. (2007). At the tie level, we will examine what types of ties are more likely to dissolve and form. For those persistent ties, changes of the connection density will be studied. At the level of personal networks, we are concerned with how the networks expand or contract. The change of the network composition will also be analyzed from the perspective of local friends and friends in the origin region.

4.4 Data and Method

The dataset we use in this study covers 4.5 years of mobile phone activity in Rwanda, from January 2005 until June 2009, and contains metadata on roughly 50 billion communication

events. For each record, we know the anonymized caller ID, recipient ID, the time and duration, as well as the location of the mobile phone tower through which the call was routed.

These data make it possible to detect migration events by finding the change of home district for each mobile phone subscriber, where "home" is defined as the district which a person remains in during a contiguous periods of time. We define migrants as those who stay in a district D_1 for K consecutive days, migrate to district D_2, and stay there for at least K consecutive days (e.g. Figure 4.1). Larger values of K imposes a more stringent definition of "migration". Most of our analysis sets $K = 90$, to follow the convention of Rwanda's Comprehensive Food Security and Vulnerability Analysis and Nutrition Survey (CFSVANS), which defines migration as a stay of three or more months. We use the algorithm designed by Chi et al. (2020) and only chose those migrants with $uncertainty = 0$, where uncertainty is the number of days when a migrant appeared at destination before migration date and appeared at home after migration date. This condition of uncertainty can avoid the issue where the change of migrants' network structures near migration dates might be caused by the bias of migration dates.

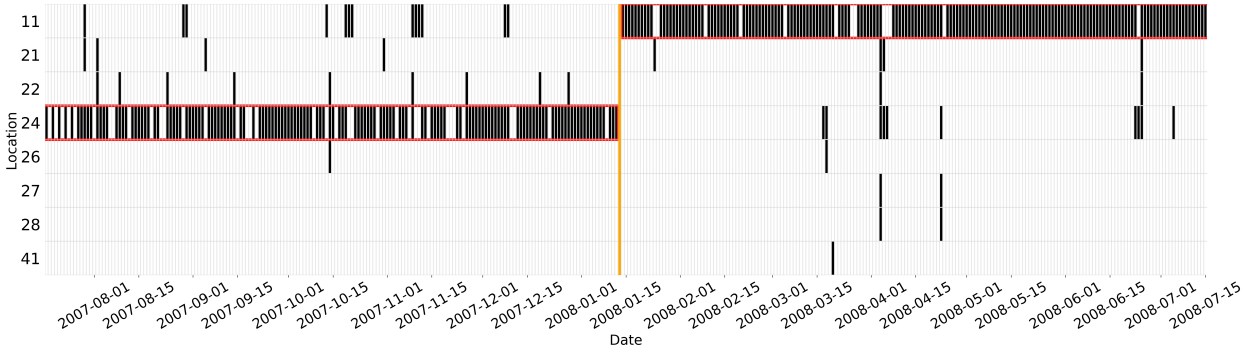


Figure 4.1: Example of a migration event. *Notes:* Each row is one location; each column is one day. A black bar means that this person made a call in that location on that day. The orange line is the detected migration date.

Event history analysis examines the likelihood that an event occurs at time t conditional on that the event did not occur before t. The Cox proportional hazard regression model is

$$\lambda(t|X_i) = \lambda_0(t)\exp(\beta_1 X_{i1} + \cdots + \beta_p X_{ip}) = \lambda_0(t)\exp(X_i \cdot \beta) \tag{4.1}$$

This model contains a baseline hazard function and effect parameters. $\lambda_0(t)$ is the baseline hazard function, describing the change of the event risk over time. This model has no assumption on the shape of the underlying hazard. The second part is the effect parameters, describing how the hazard changes in response to various factors. It is assumed that the relationship between the factors and the log hazard is linear. This model allows us to examine how specific factors influence the rate of a tie forming or dissolving.

Regressions on social networks have the issue of dependency between ties, which violates the assumptions of regressions. To have independent observations, Kossinets and Watts (2006) randomly sampled pairs from a large network. For small longitudinal network data collected based on surveys, dependencies between ties can be resolved by the multilevel design (de Nooy, 2011). For example, de Nooy (2011) incorporated the multilevel design into the discrete-time event history model to understand how ties appear, change, and disappear based on longitudinal network data. Hartl et al. (2015) used discrete-time survival analysis to examine the effect of similarity between friends and individual characteristics on adolescent friendships dissolution and found that the similarity between friends has more contribution to the friendship duration than the personal traits. Dean et al. (2017) developed a friendship duration model, which was built on the multilevel event history analysis by de Nooy (2011), to understand the role of gender and depression in friendship dissolution. In our research, we use the same method to the one of Kossinets and Watts (2009) by randomly sampling ties from the large social network to avoid the issue of dependency between social ties.

4.5 Results

We first analyze the change of social network structures pre- and post-migration, which include degree, the percentage of contacts/calls, clustering coefficient, and the proportion of denominate calls. These structures capture the connection intensity and diversity of migrants. Then we intend to understand which types of ties are more likely to be maintained and lost. Based on these findings, we will answer what factors affect the evolution of migrants' social networks. Note that there might be a large number of one-way connections during the migration period. To make sure the connections represent close relationship instead of random ties, we only keep reciprocal ties.

Descriptive statistics: Network structure change over time

Figure 4.2a shows the trend of contacts proportion 12 weeks before and 12 weeks after migration. Before migration, more than half of a migrant's contacts are from his/her home district, while about 32% of the contacts are living at migrants' destination districts. In total, more than 90% of migrants' contacts are living at either migrants' home or destination districts. For non-migrants, about 63% of their contacts are from their home districts and remain stable over time. After migration, the percentage of contacts to home decreases quickly from 60% to about 35%, while it is opposite for connections to destination friends, the proportion of which is almost doubled to 55%. The change pattern of call percentage at home and destination is similar to that of contact percentage (Figure 4.2b). This confirms the conclusion that migrants build new connections in the new place and lose old connections in the original place (Bidart & Lavenu, 2005; Lubbers et al., 2010; Nisic & Petermann, 2013). It is not surprising that Bolíbar et al. (2015) found few migrants lost their connections with the original country, while our result shows a clear connection shift. This is because their

surveys were collected years after migration when migrants' social networks are stable, while our results focus on the months over the migration process when migrants' social networks change to meet their needs in the form of information, job, accommodation, and so on. In our setting, the contact change remains stable one month after migration.

Figure 4.2c and 4.2d compare the call and contacts of migrants to non-migrants on the base of the first week. The number of calls and contacts of non-migrants keeps increasing slightly over time because more customers used cell phones. The call number and contact number of migrants increase substantially in the week before migration and go back to normal after migration.

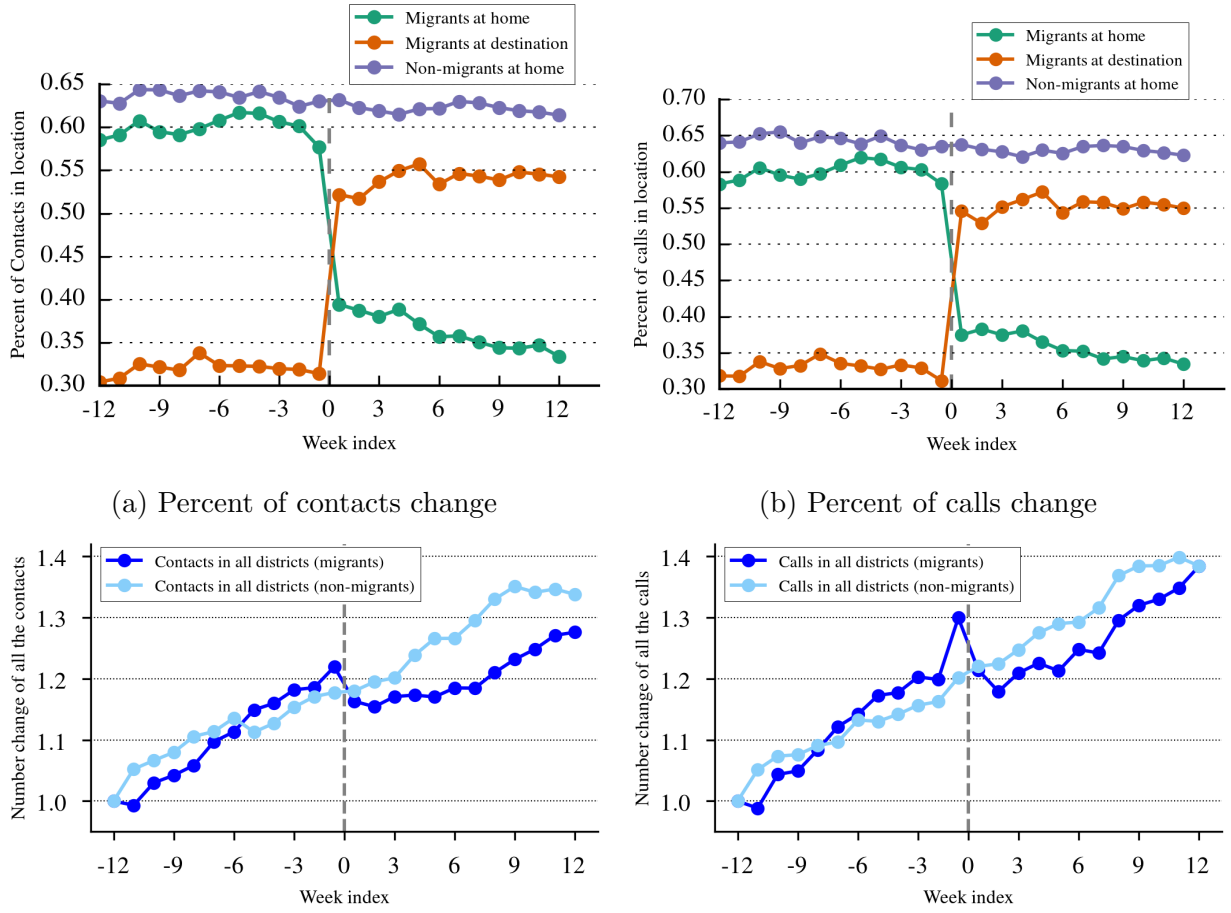
Clustering coefficient measures the extent to which nodes tend to cluster with each other. If all of a person's second-degree friends are also her friends, the clustering coefficient is 1. There is a large number of migrants who only have one or two contacts in destination per week. It would be inappropriate to measure clustering coefficient at the weekly level because most migrants' clustering coefficients per week cannot be calculated, or it is either 0 or 1. Therefore, we decrease our temporal resolution from weekly to monthly to get a better picture of migrants' clustering pattern in each time window. Figure 4.3a shows that migrants only call about four contacts at home and six contacts at destination. Figure 4.3b displays the clustering coefficient change of migrants monthly. As expected, the clustering coefficient at home increases after migration, meaning that the maintained ties of migrants are densely connected, such as family members and close friends. But for destination contacts, clustering coefficient decreased after migration, suggesting that migrants start making more diverse friends instead of strengthening their existing contacts, which confirms the hypothesis of Lubbers et al., 2010 that the heterogeneity of network increases after migration.

Figure 4.3c shows the trend of migrants' dominated calls to one person. On average, about half of an individual's calls are only with one person. Before migration, migrants make 55% of their calls to one person at home. After migration, it increased to 65%. For destinations, migrants' dominated calls keep increasing from 60% to 70% before migration, and keep decreasing after migration until about 55%. We also measure the spatial mobility pattern of migrants (Figure 4.3d). The average distance of all migrants' friends to home district remains stable before migration. After migration, the pattern shifts – the centroid of migrants' friends is closer to destination.

Which types of ties come and go?

Migrants build new social ties in the new location, lose and maintain part of connections in the origin place. In this section, we aim to figure out the different characteristics between maintained ties and lost ties in terms of their network structures and tie types.

We compare the proportion of strong ties and proportion of ties with common friends between maintained ties and lost ties before and after migration. Here the strong ties are defined as those ties with more than four calls, which covers less than 20% of all ties. Figure 4.4a shows that among maintained ties, the proportion of strong ties is much higher than that of lost ties. Overall, among all maintained ties, 69% are strong ties; while among lost



(a) Percent of contacts change

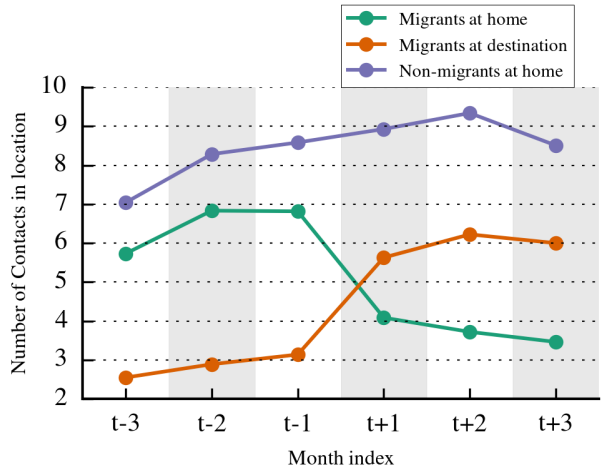
(b) Percent of calls change

(c) Number of contacts change

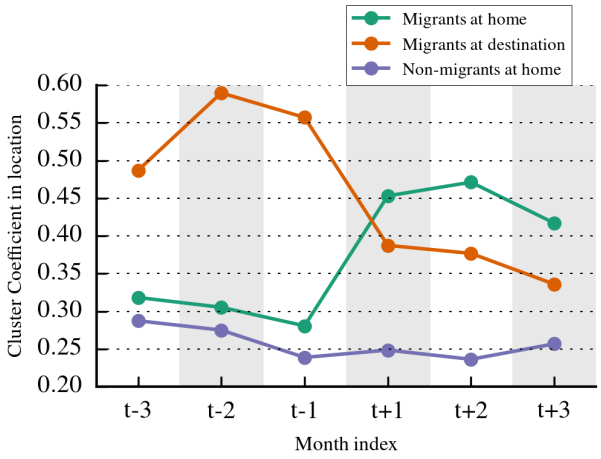
(d) Number of calls change

Figure 4.2: Social network structure changes of migrants (Part 1). *Notes:* There is a clear-cut breaking point in the week when migration events occur (Figures a and b). Migrants' contacts to home and destination shift in the first week after migration. Figures on the bottom row show the contacts/calls change of migrants comparing to non-migrants based on the t-12 week. The contact/call number of migrants increases substantially in the week before migration (Figures c and d).

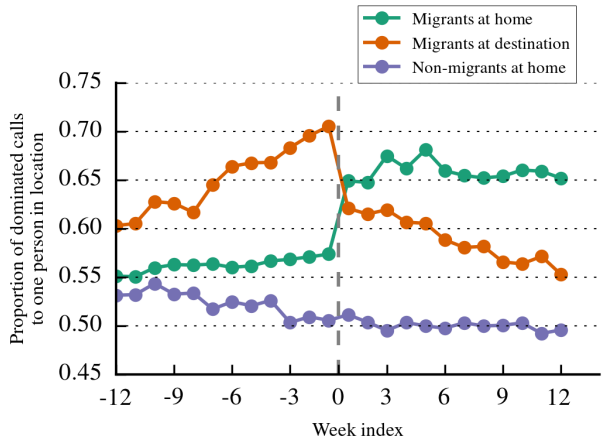
ties, only 41% are strong ties. Figure 4.4b shows the distribution of the maintained or lost friends who are friends of friends. About 38.8% of individuals have higher than 90% of maintained friends that are friends of friends; while it is only 3.9% for lost ties.



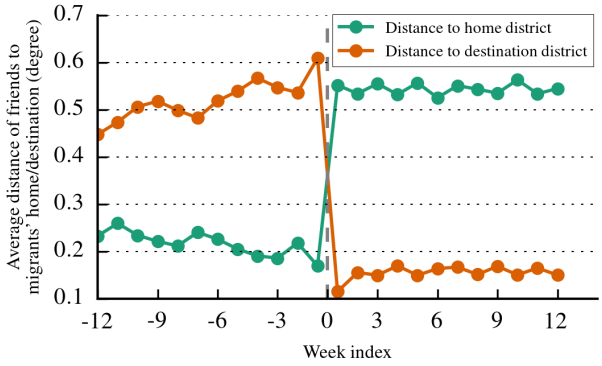
(a) Number of contacts change monthly



(b) Clustering coefficient change



(c) Proportion change of dominated calls to one person

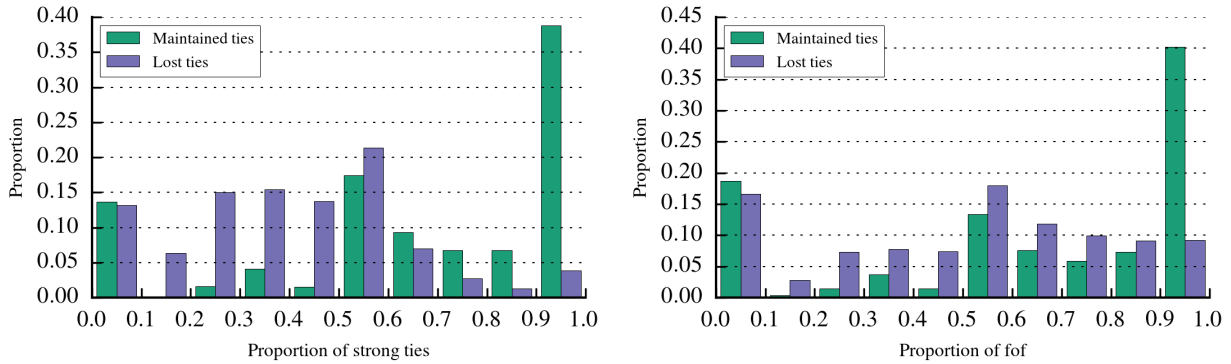


(d) Distance change of friends to home/dest

Figure 4.3: Social network structure changes of migrants (Part 2). *Notes:* Figures on the top row characterized contacts density at a monthly level. Migrants' clustering coefficient at home increases after migration, meaning the maintained ties are densely connected. Figure c shows that migrants make at least half of their calls to one person at home or destination. Figure d displays the spatial distribution change of migrant's friends. The centroid of migrants' friends moves from home to destination.

What factors affect the friends dissolution?

Above, we analyze the change of migrants' network characters before and after migration and what factors are related to tie remaining and losing. To understand how these factors



(a) Strong tie proportion

(b) Proportion of friends of friends (fof) ties

Figure 4.4: Proportion of maintained and lost ties that are strong ties or have common friends.

affect social ties dissolution, we adopt the Cox proportional hazard regression model.

Traditional studies have analyzed two sets of factors that affect network evolution. From the perspective of individual factors, the number of friends and the centrality have been found to have a positive correlation with the tie dissolution. From the perspective of tie factors, homophily, triadic closure, and spatial closeness play an important role in the network change (McPherson et al., 2001; Rapoport, 1953). In this research, we also use individual characteristics and ties factors between a migrant and a potential new friend. The former includes: degree, call number, clustering coefficient, and radius of gyration. The latter includes: difference in degree, whether common friends exist, and the number of calls between a tie. The hazard ratio of variables will illustrate which mechanism migrants prefer in general.

We define the friendship by requiring that there be at least one reciprocal call in each of the month from t-6 to t-4. We assume that a tie was lost if they have not interacted for at least three consecutive months. Specifically, there should be at least a call in each of three months from t-6 to t-4, and no calls within 3 months after migration, which is from t+4 to t+6 (Figure 4.5). The network features are calculated at t-4.

We measure migration rate over time to understand when the risk of losing friends in the migration process is highest (Figure 4.6a). Migrants start losing more connections two weeks before migration, and lose the largest number of connections in the week of migration, indicating that the hazard rate in that week is highest.

The result of Cox proportional hazard regression model is shown in Figure 4.6b and Table 4.1. Migrants who have more friends, fewer calls, and smaller radius of gyration (ROG) are more likely to lose connections. Ties with more interactions are less likely to dissolve, although it is not statistically significant. The hazard ratio represents the probability of losing ties with one unit change of the factor. For example, if a migrant has one more friend, he or she will be 1.37% more likely to lose a tie.

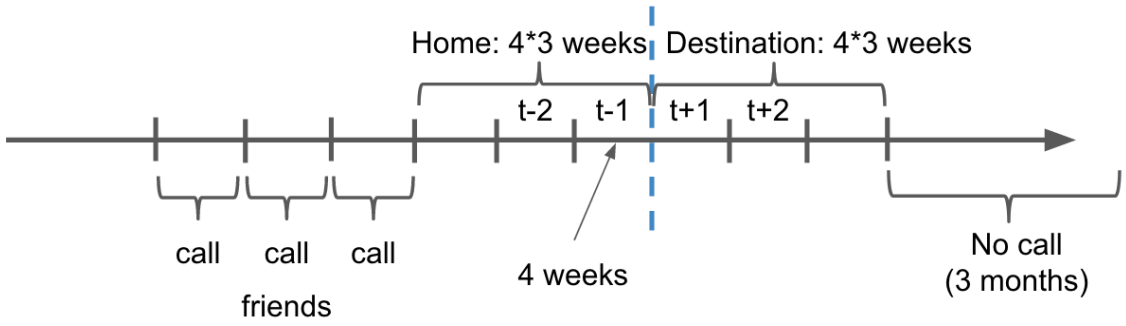
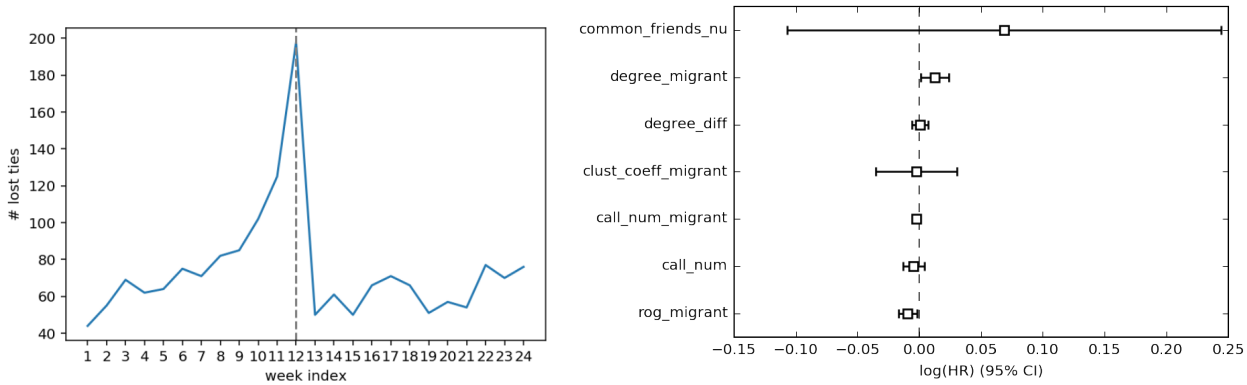


Figure 4.5: Definition of lost ties. *Notes:* Definition of friends: at least one reciprocal call in each of three months from t-6 to t-4; Definition of lost ties: no call for at least three consecutive months.



(a) Trend of lost friends

(b) Results of multivariate survival analysis

Figure 4.6: Results of friends dissolution modeling. *Notes:* Figure a shows that the number of lost ties is highest in the week of migration, but remains stable one month before migration and one week after migration. Figure b shows the hazard ratios and 95% confidence interval.

4.6 Conclusion

M. Granovetter (1983) has called for the attention of dynamic networks since 1983: "The most pressing need for further development of network ideas is a move away from static analyses that observe a system at one point in time and to pursue instead systematic accounts of how such systems develop and change." Thirty years later, it becomes possible using the unprecedented digital trace data not only to measure the network evolution in a fine time resolution but to ask novel questions of it.

How social networks of migrants evolve after migration to a new city is still not clear

Table 4.1: Results of Cox proportional hazard regression models

	(1)	(2)	(3)
Call number of migrant	-0.0028**		-0.0024*
	(0.0013)		(0.0014)
Degree of migrant	0.0137**		0.0124**
	(0.0056)		(0.0059)
Clustering coefficient of migrant	0.0012		-0.0024
	(0.0162)		(0.0167)
ROG of migrant	-0.009**		-0.0093**
	(0.0038)		(0.0038)
Call number between migrant and this friend		-0.0063	-0.0047
		(0.0041)	(0.0045)
Number of common friends between migrant and this friend		0.039	0.0688
		(0.085)	(0.0896)
Degree difference between migrant and this friend		0.0008	0.0005
		(0.0034)	(0.0034)

Note: *p<0.1; **p<0.05; ***p<0.01.

because of the data limitation. In this research, we use call detail records of four and half years to understand the pattern of social network evolution and what strategies migrants adopt to maximize their social utility over the migration process. We find that the connections of migrants to home and destination shift right after migration. This is also the reason why our model focuses on the period right before and after migration because this is the time that migrants establish new friends and lost old ones dramatically. We also find that the heterogeneity of newly formed social networks increases, meaning that migrants intend to have diverse social connections at destination. Migrants maintain strong ties and dense social ties in origin and lost a high proportion of weak ties.

While large-scale digital trace data provide a novel way to understand the relationship between migration and social networks at a granular spatial and temporal resolution, one of the main limitations in this research is that migration reasons are unknown. Migrants might have different strategies on network formation due to different reasons, such as education, family, job, and violence (Bolíbar et al., 2015). The information of migration reasons can provide a natural way to classify migrants into different groups and compare their pattern of network evolution. One future work to achieve a similar goal using CDR is to calculate migrants' network composition over time with the assumption that people with different migration reasons will have different strategies to maintain and form their social ties.

Bibliography

Abel, G. J., & Sander, N. (2014). Quantifying Global International Migration Flows. *Science*, *343*(6178), 1520–1522.

adams jimi, j., Faust, K., & Lovasi, G. S. (2012). Capturing context: Integrating spatial and social network analyses. *Social Networks*, *34*(1), 1–5.

Ali, S. N., & Miller, D. A. (2016). Ostracism and Forgiveness. *American Economic Review*, *106*(8), 2329–2348.

Ambrus, A., Mobius, M., & Szeidl, A. (2015). Consumption risk-sharing in social networks. *American Economic Review*, *104*(1), 149–182.

Ambrus, A., Gao, W. Y., & Milán, P. (2018). Informal risk sharing with local information. *Available at SSRN 3220524*.

Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2012). Four Degrees of Separation, In *Proc. websci'12*, New York, NY, USA, ACM.

Bailey, M., Cao, R., Kuchler, T., Stroebel, J., & Wong, A. (2018). Social Connectedness: Measurement, Determinants, and Effects. *Journal of Economic Perspectives*, *32*(3), 259–280.

Ballester, C., Calvó-Armengol, A., & Zenou, Y. (2006). Who's who in networks. wanted: The key player. *Econometrica*, *74*(5), 1403–1417.

Banerjee, A. V., Breza, E., Chandrasekhar, A., Duflo, E., & Jackson, M. O. (2012). Come play with me: Experimental evidence of information diffusion about rival goods.

Banerjee, A. V., & Newman, A. F. (1998). Information, the Dual Economy, and Development. *The Review of Economic Studies*, *65*(4), 631–653.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2013). The Diffusion of Microfinance. *Science*, *341*(6144), 1236498.

Banerjee, A., Chandrasekhar, A. G., Duflo, E., & Jackson, M. O. (2019). Using Gossips to Spread Information: Theory and Evidence from Two Randomized Controlled Trials. *The Review of Economic Studies*, 2453–2490.

Barbosa, H., Barthelemy, M., Ghoshal, G., James, C. R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J. J., Simini, F., & Tomasini, M. (2018). Human mobility: Models and applications. *Physics Reports*, *734*, 1–74.

Barnett, G. A. (2016). A Longitudinal Analysis of the International Telecommunication Network, 1978-1996. *American Behavioral Scientist*, 1638–1655.

Barnett, G. A., Jacobson, T., Choi, Y., & Sun-Miller, S. (1996). An examination of the international telecommunication network. *Journal of International Communication*, *3*(2), 19–43.

Barwick, P. J., Liu, Y., Patacchini, E., & Wu, Q. (2019). *Information, Mobile Communication, and Referral Effects* (Working Paper No. 25873). National Bureau of Economic Research.

Beaman, L. A. (2012). Social networks and the dynamics of labour market outcomes: Evidence from refugees resettled in the US. *The Review of Economic Studies*, *79*(1), 128–161.

Beaman, L., BenYishay, A., Magruder, J., & Mobarak, A. M. (2015). Can network theory based targeting increase technology adoption. *Unpublished Manuscript*.

Beaverstock, J. V., Smith, R. G., & Taylor, P. J. (2000). World-City Network: A New Metageography? *Annals of the Association of American Geographers*, *90*(1), 123–134.

Bell, M., Charles Edwards, E., Kupiszewska, D., Kupiszewski, M., Stillwell, J., & Zhu, Y. (2015). Internal Migration Data Around the World: Assessing Contemporary Practice. *Population, Space and Place*, *21*(1), 1–17.

Bengtsson, L., Lu, X., Thorson, A., Garfield, R., & Von Schreeb, J. (2011). Improved response to disasters and outbreaks by tracking population movements with mobile phone network data: A post-earthquake geospatial study in haiti. *PLoS medicine*, *8*(8), e1001083.

Bertoli, S., Fernández-Huertas Moraga, J., & Ortega, F. (2013). Crossing the border: Self-selection, earnings and individual migration decisions. *Journal of Development Economics*, *101*, 75–91.

Bertoli, S., & Ruyssen, I. (2018). Networks and migrants' intended destination. *Journal of Economic Geography*, *18*(4), 705–728.

Bhagat, S., Burke, M., Diuk, C., Filiz, I. O., & Edunov, S. (2016). Three and a half degrees of separation. *Facebook Research, https://research.fb.com/three-and-a-half-degrees-of-separation/*.

Bidart, C., & Lavenu, D. (2005). Evolutions of personal networks and life events. *Social Networks*, *27*(4), 359–376.

Blumenstock, J. E. (2012). Inferring Patterns of Internal Migration from Mobile Phone Call Records: Evidence from Rwanda. *Information Technology for Development*, *18*(2), 107–125.

Blumenstock, J. E., & Eagle, N. (2012). Divided We Call: Disparities in Access and Use of Mobile Phones in Rwanda. *Information Technology and International Development*, *8*(2), 1–16.

Blumenstock, J., Chi, G., & Tan, X. (2019). *Migration and the value of social networks* (tech. rep.). CEPR Discussion Papers.

Bolíbar, M., Martí, J., & Verd, J. M. (2015). Just a question of time? The composition and evolution of immigrants' personal networks in Catalonia. *International Sociology*, *30*(6), 579–598.

Borjas, G. J. (1992). Ethnic Capital and Intergenerational Mobility. *The Quarterly Journal of Economics*, *107*(1), 123–150.

Borjas, G. J., Bronars, S. G., & Trejo, S. J. (1992). Self-selection and internal migration in the United States. *Journal of Urban Economics*, *32*(2), 159–185.

boyd danah, d., & Crawford, K. (2012). Critical Questions for Big Data. *Information, Communication & Society*, *15*(5), 662–679.

Boyd, D., & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, communication & society*, *15*(5), 662–679.

Bramoullé, Y., Kranton, R., & D'amours, M. (2014). Strategic interaction and networks. *American Economic Review*, *104*(3), 898–930.

Breza, E., Chandrasekhar, A. G., McCormick, T. H., & Pan, M. (2017). *Using Aggregated Relational Data to Feasibly Identify Network Structure without Network Data* (Working Paper No. 23491). National Bureau of Economic Research.

Büchel, K., Puga, D., Viladecans-Marsal, E., & von Ehrlich, M. (2019). *Calling from the Outside: The Role of Networks in Residential Mobility* (SSRN Scholarly Paper No. ID 3360082). Social Science Research Network. Rochester, NY.

Burt, R. S. (1992). *Structural holes: The social structure of competition*. Harvard university press.

Burt, R. (2004). Structural holes and good ideas. *American Journal of Sociology*, *110*(2), 349–399.

Calvó-Armengol, A. (2004). Job contact networks. *Journal of Economic Theory*, *115*(1), 191–206.

Calvó-Armengol, A., & Jackson, M. O. (2004). The Effects of Social Networks on Employment and Inequality. *The American Economic Review*, *94*(3), 426–454.

Card, D. (2001). Immigrant Inflows, Native Outflows, and the Local Labor Market Impacts of Higher Immigration. *Journal of Labor Economics*, *19*(1), 22–64.

Carletto, C., de Brauw, A., & Banerjee, R. (2012). Measuring migration in multi-topic household surveys. *Handbook of Research Methods in Migration*, 207–228.

Carrington, W. J., Detragiache, E., & Vishwanath, T. (1996). Migration with Endogenous Moving Costs. *The American Economic Review*, *86*(4), 909–930.

Casciaro, T. (1998). Seeing things clearly: Social structure, personality, and accuracy in social network perception. *Social Networks*, *20*(4), 331–351.

Cha, M., Mislove, A., & Gummadi, K. P. (2009). A measurement-driven analysis of information propagation in the flickr social network, In *Proc. WWW '09*, Madrid, Spain, ACM Press.

Chandrasekhar, A. G., Kinnan, C., & Larreguy, H. (2018). Social Networks as Contract Enforcement: Evidence from a Lab Experiment in the Field. *American Economic Journal: Applied Economics*, *10*(4), 43–78.

Chandrasekhar, A., Breza, E., & Tahbaz-Salehi, A. (2016). Seeing the forest for the trees? an investigation of network knowledge. *Working Paper*.

Chen, C., Ma, J., Susilo, Y., Liu, Y., & Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, *68*, 285–299.

Chi, G., Lin, F., Chi, G., & Blumenstock, J. (2020). A general approach to detecting migration events in digital trace data. *Working Paper*.

Chi, G., State, B., Blumenstock, J. E., & Adamic, L. (2019). Who ties the world together? evidence from a large online social network, In *International conference on complex networks and their applications*. Springer.

Cho, E., Myers, S. A., & Leskovec, J. (2011). Friendship and Mobility: User Movement in Location-based Social Networks, In *Proc. kdd'11*, New York, NY, USA, ACM.

Chuang, Y., & Schechter, L. (2015). Social Networks in Developing Countries. *Annual Review of Resource Economics*, *7*(1), 451–472.

Coleman, J. S. (1994). *Foundations of social theory*. Harvard University Press.

Comola, M., & Mendola, M. (2015). The Formation of Migrant Networks. *Scandinavian Journal of Economics*, *117*(2), 592–618.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., & Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications*, *392*(6), 1459–1473.

Dagnelie, O., Mayda, A. M., & Maystadt, J.-F. (2019). The labor market integration of refugees in the United States: Do entrepreneurs in the network help? *European Economic Review*, *111*, 257–272.

Dahl, M. S., & Sorenson, O. (2010). The migration of technical workers. *Journal of Urban Economics*, *67*(1), 33–45.

Davies, P. S., Greenwood, M. J., & Li, H. (2001). A conditional logit approach to US state-to-state migration. *Journal of Regional Science*, *41*(2), 337–360.

De Montjoye, Y.-A., Hidalgo, C. A., Verleysen, M., & Blondel, V. D. (2013). Unique in the crowd: The privacy bounds of human mobility. *Scientific reports*, *3*, 1376.

de Nooy, W. (2011). Networks of action and events over time. A multilevel discrete-time event history model for longitudinal network data. *Social Networks*, *33*(1), 31–40.

Dean, D. O., Bauer, D. J., & Prinstein, M. J. (2017). Friendship Dissolution Within Social Networks Modeled Through Multilevel Event History Analysis. *Multivariate Behavioral Research*, *52*(3), 271–289.

Derudder, B., & Witlox, F. (2005). An Appraisal of the Use of Airline Data in Assessing the World City Network: A Research Note on Data. *Urban Studies*, *42*(13), 2371–2388.

Derudder, B. (2006). On Conceptual Confusion in Empirical Analyses of a Transnational Urban Network. *Urban Studies*, *43*(11), 2027–2046.

Derudder, B., & Witlox, F. (2008). Mapping world city networks through airline flows: Context, relevance, and problems. *Journal of Transport Geography*, *16*(5), 305–312.

Derudder, B., Witlox, F., & Taylor, P. J. (2007). U.S. Cities in the World City Network. *Urban Geography*, *28*(1), 74–91.

Deshingkar, P., & Grimm, S. (2005). *Internal Migration and Development: A Global Perspective*. United Nations Publications.

Dinkelman, T., & Mariotti, M. (2016). *The Long Run Effects of Labor Migration on Human Capital Formation in Communities of Origin* (Working Paper No. 22049). National Bureau of Economic Research.

Dolfin, S., & Genicot, G. (2010). What Do Networks Do? The Role of Networks on Migration and "Coyote" Use. *Review of Development Economics*, *14*(2), 343–359.

Dunbar, R. I. M. (1998). The social brain hypothesis. *Evolutionary Anthropology: Issues, News, and Reviews*, *6*(5), 178–190.

Dustmann, C., Glitz, A., Schönberg, U., & Brücker, H. (2016). Referral-based Job Search Networks. *The Review of Economic Studies*, *83*(2), 514–546.

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Economist. (2019). Rwanda has banned talking about ethnicity. *The Economist*.

Edin, P.-A., Fredriksson, P., & Åslund, O. (2003). Ethnic Enclaves and the Economic Success of Immigrants—Evidence from a Natural Experiment. *The Quarterly Journal of Economics*, *118*(1), 329–357.

Ester, M., Kriegel, H.-P., Sander, J., Xu, X., Et al. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise., In *Kdd*.

Eubanks, V. (2018). *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press.

Fafchamps, M., & Shilpi, F. (2013). Determinants of the Choice of Migration Destination. *Oxford Bulletin of Economics and Statistics*, *75*(3), 388–409.

Feenstra, R. C. (2015). *Advanced international trade: Theory and evidence*. Princeton University Press.

Feld, S. L., Suitor, J. J., & Hoegh, J. G. (2007). Describing Changes in Personal Networks over Time. *Field Methods*, *19*(2), 218–236.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017). Using Twitter Data to Estimate the Relationship between Short-term Mobility and Long-term Migration, In *Proceedings of the 2017 ACM on Web Science Conference - WebSci '17*, Troy, New York, USA, ACM Press.

Friedkin, N. E. (1983). Horizons of Observability and Limits of Informal Control in Organizations. *Social Forces*, *62*(1), 54–77.

Garcia-Gavilanes, R., Mejova, Y., & Quercia, D. (2014). Twitter ain't without frontiers: Economic, social, and cultural boundaries in international communication, In *Proc. icwsm'14*. ACM.

Ghosh, P., & Ray, D. (1996). Cooperation in Community Interaction Without Information Flows. *The Review of Economic Studies*, *63*(3), 491–519.

Gill, N., & Bialski, P. (2011). New friends in new places: Network formation during the migration process among poles in the uk. *Geoforum*, *42*(2), 241–249.

Giulietti, C., Wahba, J., & Zenou, Y. (2018). Strong versus weak ties in migration. *European Economic Review*, *104*, 111–137.

Golub, B., & Jackson, M. O. (2010). Naive learning in social networks and the wisdom of crowds. *American Economic Journal: Microeconomics*, *2*(1), 112–149.

Gonzalez, M. C., Hidalgo, C. A., & Barabasi, A.-L. (2008). Understanding individual human mobility patterns. *Nature*, *453*(7196), 779–782
2008-06-22.

Granovetter, M. (1983). The Strength of Weak Ties: A Network Theory Revisited. *Sociological Theory*, *1*, 201–233.

Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78*(6), 1360–1380.

Greenwood, M. J. (1969). An Analysis of the Determinants of Geographic Labor Mobility in the United States. *The Review of Economics and Statistics*, *51*(2), 189–194.

Greenwood, M. J., & Hunt, G. L. (2003). The Early History Of Migration Research. *International Regional Science Review*, *26*(1), 3–37.

Guiteras, R., Levinsohn, J. A., & Mobarak, A. M. (2019). *Demand Estimation with Strategic Complementarities: Sanitation in Bangladesh* (tech. rep. No. 13498). C.E.P.R. Discussion Papers.

Hafner-Burton, E. M., Kahler, M., & Montgomery, A. H. (2009). Network analysis for international relations. *International Organization*, *63*(3), 559–592.

Hankaew, S., Phithakkitnukoon, S., Demissie, M. G., Kattan, L., Smoreda, Z., & Ratti, C. (2019). Inferring and modeling migration flows using mobile phone network data. *IEEE Access*, *7*, 164746–164758.

Hanson, G. H., & Woodruff, C. (2003). *Emigration and educational attainment in Mexico* (tech. rep.). Mimeo., University of California at San Diego.

Hare, D. (1999).
Push' versus 'pull' factors in migration outflows and returns: Determinants of migration status and spell duration among China's rural population. *The Journal of Development Studies*, *35*(3), 45–72.

Hartl, A. C., Laursen, B., & Cillessen, A. H. N. (2015). A Survival Analysis of Adolescent Friendships: The Downside of Dissimilarity. *Psychological Science*, *26*(8), 1304–1315.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., & Ratti, C. (2014). Geolocated Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, *41*(3), 260–271.

Herdağdelen, A., State, B., Adamic, L., & Mason, W. (2016). The social ties of immigrant communities in the United States, In *Proc. websci'16*, Hannover, Germany, ACM.

Hintjens, H. (2008). Post-genocide identity politics in Rwanda. *Ethnicities*, *8*(1), 5–41.

Hollis, M., & Smith, S. (1990). *Explaining and Understanding International Relations*. Clarendon Press.

Hong, L., Wu, J., Frias-Martinez, E., Villarreal, A., & Frias-Martinez, V. (2019). Characterization of Internal Migrant Behavior in the Immediate Post-migration Period Using Cell Phone Traces, In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development*, New York, NY, USA, ACM.

Hu, Y. (2005). Efficient, high-quality force-directed graph drawing. *Mathematica Journal*, *10*(1), 37–71.

Ioannides, Y. M., & Datcher Loury, L. (2004). Job Information Networks, Neighborhood Effects, and Inequality. *Journal of Economic Literature*, *42*(4), 1056–1093.

Jackson, M. O. (2010). *Social and Economic Networks*. Princeton University Press.

Jackson, M. O. (2018). *A Typology of Social Capital and Associated Network Measures* (SSRN Scholarly Paper No. ID 3073496). Social Science Research Network. Rochester, NY.

Jackson, M. O., & Nei, S. (2015). Networks of military alliances, wars, and international trade. *Proceedings of the National Academy of Sciences*, *112*(50), 15277–15284.

Jackson, M. O., Rodriguez-Barraquer, T., & Tan, X. (2012). Social capital and social quilts: Network patterns of favor exchange. *The American Economic Review*, *102*(5), 1857–1897.

Jackson, M. O., & Wolinsky, A. (1996). A strategic model of social and economic networks. *Journal of economic theory*, *71*(1), 44–74.

Jackson, M. O., & Yariv, L. (2010). Diffusion, strategic interaction, and social structure. *Handbook of Social Economics, edited by J. Benhabib, A. Bisin and M. Jackson*.

Jackson, M. O., & Zenou, Y. (2015). Games on networks, In *Handbook of game theory with economic applications*. Elsevier.

Jacobs, A. Z., Way, S. F., Ugander, J., & Clauset, A. (2015). Assembling thefacebook: Using heterogeneity to understand online social network assembly, In *Proc. websci'15*. ACM.

Jiang, S., Ferreira, J., & Gonzalez, M. C. (2017). Activity-Based Human Mobility Patterns Inferred from Mobile Phone Data: A Case Study of Singapore. *IEEE Transactions on Big Data*, *3*(2), 208–219.

Jiang, Z.-Q., Xie, W.-J., Li, M.-X., Podobnik, B., Zhou, W.-X., & Stanley, H. E. (2013). Calling patterns in human communication dynamics. *Proceedings of the National Academy of Sciences*, *110*(5), 1600–1605.

Jurdak, R., Zhao, K., Liu, J., AbouJaoude, M., Cameron, M., & Newth, D. (2015). Understanding Human Mobility from Twitter. *PLOS ONE*, *10*(7), e0131469.

Kang, C., Ma, X., Tong, D., & Liu, Y. (2012). Intra-urban human mobility patterns: An urban morphology perspective. *Physica A: Statistical Mechanics and its Applications*, *391*(4), 1702–1717.

Kang, C., Sobolevsky, S., Liu, Y., & Ratti, C. (2013). Exploring Human Movements in Singapore: A Comparative Analysis Based on Mobile Phone and Taxicab Usages, In *Proceedings of the 2nd ACM SIGKDD International Workshop on Urban Computing*, New York, NY, USA, ACM.

Karlan, D., Mobius, M., Rosenblat, T., & Szeidl, A. (2009). Trust and Social Collateral. *The Quarterly Journal of Economics*, *124*(3), 1307–1361.

Keeling, D. J. (1995). Transport and the world city paradigm. *World cities in a world-system*, 115–131.

Keeling, M. J., & Eames, K. T. D. (2005). Networks and epidemic models. *Journal of The Royal Society Interface*, *2*(4), 295–307.

Kempe, D., Kleinberg, J., & Tardos, É. (2003). Maximizing the Spread of Influence Through a Social Network, In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, ACM.

Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A*, *115*(772), 700–721.

Kikas, R., Dumas, M., & Saabas, A. (2015). Explaining International Migration in the Skype Network: The Role of Social Network Features, In *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, New York, NY, USA, ACM.

Kinnan, C. (2019). Distinguishing barriers to insurance in Thai villages. *Working paper*.

Kinnan, C., Wang, S.-Y., & Wang, Y. (2018). Access to Migration for Rural Households. *American Economic Journal: Applied Economics*, *10*(4), 79–119.

Kitsak, M., Gallos, L. K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H. E., & Makse, H. A. (2010). Identification of influential spreaders in complex networks. *Nature Physics*, *6*(11), 888–893.

Knapp, T. A., White, N. E., & Clark, D. E. (2001). A Nested Logit Approach to Household Mobility. *Journal of Regional Science*, *41*(1), 1–22.

Koelet, S., & de Valk, H. A. (2016). Social networks and feelings of social loneliness after migration: The case of European migrants with a native partner in Belgium. *Ethnicities*, *16*(4), 610–630.

König, M. D., Rohner, D., Thoenig, M., & Zilibotti, F. (2017). Networks in Conflict: Theory and Evidence From the Great War of Africa. *Econometrica*, *85*(4), 1093–1132.

Kossinets, G., & Watts, D. J. (2006). Empirical Analysis of an Evolving Social Network. *Science*, *311*(5757), 88–90.

Kossinets, G., & Watts, D. J. (2009). Origins of Homophily in an Evolving Social Network. *American Journal of Sociology*, *115*(2), 405–450.

Kranton, R. E. (1996). The formation of cooperative relationships. *The Journal of Law, Economics, and Organization*, *12*(1), 214–233.

Kyoung-Ho, S., & Timberlake, M. (2000). World cities in Asia: Cliques, centrality and connectedness. *Urban Studies; London*, *37*(12), 2257–2285.

Lacey, M. (2004). A Decade After Massacres, Rwanda Outlaws Ethnicity. *The New York Times*.

Leetaru, K., Wang, S., Cao, G., Padmanabhan, A., & Shook, E. (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, *18*(5).

Leskovec, J., Backstrom, L., Kumar, R., & Tomkins, A. (2008). Microscopic evolution of social networks, In *Kdd*. ACM.

Leskovec, J., & Horvitz, E. (2008). Planetary-scale views on a large instant-messaging network, In *Proceeding of the 17th international conference on World Wide Web - WWW '08*, Beijing, China, ACM Press.

Ligon, E. A., & Schechter, L. (2011). Motives for Sharing in Social Networks. *SSRN eLibrary*.

Liu, Y., Liu, X., Gao, S., Gong, L., Kang, C., Zhi, Y., Chi, G., & Shi, L. (2015). Social Sensing: A New Approach to Understanding Our Socioeconomic Environments. *Annals of the Association of American Geographers*, *105*(3), 512–530.

Lu, X., Bengtsson, L., & Holme, P. (2012). Predictability of population displacement after the 2010 haiti earthquake. *Proceedings of the National Academy of Sciences*, *109*(29), 11576–11581.

Lu, X., Wetter, E., Bharti, N., Tatem, A. J., & Bengtsson, L. (2013). Approaching the Limit of Predictability in Human Mobility. *Scientific Reports*, *3*, 2923.

Lu, X., Wrathall, D. J., Sundsøy, P. R., Nadiruzzaman, M., Wetter, E., Iqbal, A., Qureshi, T., Tatem, A., Canright, G., Engø-Monsen, K., & Bengtsson, L. (2016). Unveiling hidden migration and mobility patterns in climate stressed regions: A longitudinal study of six million anonymous mobile phone users in Bangladesh. *Global Environmental Change*, *38*, 1–7.

Lubbers, M. J., Molina, J. L., Lerner, J., Brandes, U., Ávila, J., & McCarty, C. (2010). Longitudinal analysis of personal networks. The case of Argentinean migrants in Spain. *Social Networks*, *32*(1), 91–104.

Lucas, R. E. B. (1997). Internal migration in developing countries, In *Handbook of Population and Family Economics*. Elsevier.

Lucas, R. E. B. (2015). Chapter 26 - African Migration (B. R. Chiswick & P. W. Miller, Eds.). In B. R. Chiswick & P. W. Miller (Eds.), *Handbook of the Economics of International Migration*. North-Holland.

Mahajan, P., & Yang, D. (2017). *Taken by Storm: Hurricanes, Migrant Networks, and U.S. Immigration* (Working Paper No. 23756). National Bureau of Economic Research.

Malecki, E. J. (2002). The Economic Geography of the Internet's Infrastructure. *Economic Geography*, *78*(4), 399–424.

Matsumoto, H. (2004). International urban systems and air passenger and cargo flows: Some calculations. *Journal of Air Transport Management*, *10*(4), 239–247.

Mcfadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior (P. Zarembka, Ed.). In P. Zarembka (Ed.), *Frontiers in econometrics*. Academic Press.

McFadden, D. L. (1984). Econometric analysis of qualitative response models. *Handbook of Econometrics*, *2*, 1395–1457.

McKenzie, D. J., & Sasin, M. J. (2007). *Migration, Remittances, Poverty, and Human Capital: Conceptual and Empirical Challenges* (SSRN Scholarly Paper No. ID 999482). Social Science Research Network. Rochester, NY.

McKenzie, D., & Rapoport, H. (2010). Self-Selection Patterns in Mexico-U.S. Migration: The Role of Migration Networks. *Review of Economics and Statistics*, *92*(4), 811–821.

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology*, *27*(1), 415–444.

Mellon, J., & Prosser, C. (2017). Twitter and Facebook are not representative of the general population: Political attitudes and demographics of British social media users. *Research & Politics*, *4*(3).

Miller, D., & Tan, X. (2018). Seeking relationship support: Strategic network formation and robust cooperation. *working paper*.

Monderer, D., & Shapley, L. S. (1996). Potential games. *Games and economic behavior*, *14*(1), 124–143.

Montgomery, J. D. (1991). Social Networks and Labor-Market Outcomes: Toward an Economic Analysis. *The American Economic Review*, *81*(5), 1408–1418.

Morten, M. (2019). Temporary Migration and Endogenous Risk Sharing in Village India. *Journal of Political Economy*, *127*(1), 1–46.

Mossel, E., Sly, A., & Tamuz, O. (2015). Strategic learning and the topology of social networks. *Econometrica*, *83*(5), 1755–1794.

Munshi, K. (2003). Networks in the Modern Economy: Mexican Migrants in the U. S. Labor Market. *The Quarterly Journal of Economics*, *118*(2), 549–599.

Munshi, K. (2014). Community Networks and the Process of Development. *The Journal of Economic Perspectives*, *28*(4), 49–76.

Munshi, K., & Rosenzweig, M. (2016). Networks and Misallocation: Insurance, Migration, and the Rural-Urban Wage Gap. *American Economic Review*, *106*(1), 46–98.

National Institute of Statistics of Rwanda. (2014). *Migration and Spatial Mobility* (tech. rep.). Kigali, Rwanda.

Nestorowicz, J., & Anacka, M. (2019). Mind the Gap? Quantifying Interlinkages between Two Traditions in Migration Literature. *International Migration Review*, *53*(1), 283–307.

Nisic, N., & Petermann, S. (2013). New City = New Friends? The Restructuring of Social Resources after Relocation. *Comparative Population Studies*, *38*(1).

Onnela, J.-P., Saramäki, J., Hyvönen, J., Szabó, G., Lazer, D., Kaski, K., Kertész, J., & Barabási, A.-L. (2007). Structure and tie strengths in mobile communication networks. *Proceedings of the National Academy of Sciences*, *104*(18), 7332–7336.

Patel, K., & Vella, F. (2012). Immigrant Networks and Their Implications for Occupational Choice and Wages. *The Review of Economics and Statistics*, *95*(4), 1249–1277.

Perkins, R., & Neumayer, E. (2013). The ties that bind: The role of migrants in the uneven geography of international telephone traffic. *Global networks*, *13*(1), 79–100.

Phithakkitnukoon, S., Calabrese, F., Smoreda, Z., & Ratti, C. (2011). Out of sight out of mind–how our mobile social network changes during migration, In *2011 ieee third international conference on privacy, security, risk and trust and 2011 ieee third international conference on social computing*. IEEE.

Phithakkitnukoon, S., Horanont, T., Di Lorenzo, G., Shibasaki, R., & Ratti, C. (2010). Activity-Aware Map: Identifying Human Daily Activity Pattern Using Mobile Phone Data (A. A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli, Eds.). In A. A. Salah, T. Gevers, N. Sebe, & A. Vinciarelli (Eds.), *Human Behavior Understanding*, Berlin, Heidelberg, Springer.

Pisarevskaya, A., Levy, N., Scholten, P., & Jansen, J. (2019). Mapping migration studies: An empirical analysis of the coming of age of a research field. *Migration Studies*.

Rapoport, A. (1953). Spread of information through a population with socio-structural bias: I. Assumption of transitivity. *The bulletin of mathematical biophysics*, *15*(4), 523–533.

Ratha, D. (2016). *Migration and remittances factbook 2016*. The World Bank.

Rauch, J. E. (2001). Business and Social Networks in International Trade. *Journal of Economic Literature*, *39*(4), 1177–1203.

Ravenstein, E. G. (1885). The Laws of Migration. *Journal of the Statistical Society of London*, *48*(2), 167–235.

Rees, A. (1966). Information Networks in Labor Markets. *The American Economic Review*, *56*(1/2), 559–566.

Revelt, D., & Train, K. (1998). Mixed Logit with Repeated Choices: Households' Choices of Appliance Efficiency Level. *The Review of Economics and Statistics*, *80*(4), 647–657.

Roberts, S. B. G., & Dunbar, R. I. M. (2015). Managing Relationship Decay. *Human Nature*, *26*(4), 426–450.

Rogers, A., Raymer, J., & Newbold, K. B. (2003). Reconciling and translating migration data collected over time intervals of differing widths. *The Annals of Regional Science*, *37*(4), 581–601.

Rogers, E. M. (1962). *Diffusion of Innovations*. Simon; Schuster.

Rosenzweig, M. R., & Stark, O. (1989). Consumption Smoothing, Migration, and Marriage: Evidence from Rural India. *Journal of Political Economy*, *97*(4), 905–926.

Rutherford, J., Gillespie, A., & Richardson, R. (2004). The territoriality of Pan-European telecommunications backbone networks. *Journal of Urban Technology*, *11*(3), 1–34.

Salganik, M. (2017). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., & González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, *10*(84), 20130246.

Shi, L., Chi, G., Liu, X., & Liu, Y. (2015). Human mobility patterns in different communities: A mobile phone data-based social network approach. *Annals of GIS*, *21*(1), 15–26.

Short, J. R., Kim, Y., Kuus, M., & Wells, H. (1996). The Dirty Little Secret of World Cities Research: Data Problems in Comparative Analysis. *International Journal of Urban and Regional Research*, *20*(4), 697–717.

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*(7392), 96–100.

Simmel, G. (1950). The sociology of georg simmel (kh wolff, trans.) *Glencoe, IL: The Free Press.(Original work published 1908)*.

Smith, D. A., & Timberlake, M. F. (2001). World city networks and hierarchies, 1977-1997: An empirical analysis of global air travel links. *American Behavioral Scientist*, *44*(10), 1656–1678.

Song, C., Koren, T., Wang, P., & Barabási, A.-L. (2010). Modelling the scaling properties of human mobility. *Nature Physics*, *6*(10), 818–823.

Song, C., Qu, Z., Blumm, N., & Barabási, A.-L. (2010). Limits of predictability in human mobility. *Science*, *327*(5968), 1018–1021.

State, B., Park, P., Weber, I., & Macy, M. (2015). The Mesh of Civilizations in the Global Network of Digital Communication. *PLOS ONE*, *10*(5), e0122543.

State, B., Weber, I., & Zagheni, E. (2014). Studying inter-national mobility through IP geolocation, In *Proc. wsdm'14*, ACM.

Takhteyev, Y., Gruzd, A., & Wellman, B. (2012). Geography of Twitter networks. *Social Networks*, *34*(1), 73–81.

Taylor, L. (2016). No place to hide? the ethics and analytics of tracking mobility using mobile phone data. *Environment and Planning D: Society and Space*, *34*(2), 319–336.

Todaro, M. (1980). Internal migration in developing countries: A survey, In *Population and economic change in developing countries*. University of Chicago Press.

Topa, G. (2001). Social Interactions, Local Spillovers and Unemployment. *The Review of Economic Studies*, *68*(2), 261–295.

Townsend, A. M. (2001). Network Cities and the Global Structure of the Internet. *American Behavioral Scientist*, *44*(10), 1697–1716.

Travers, J., & Milgram, S. (1967). The small world problem. *Phychology Today*, *1*(1), 61–67.

Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. *Sociometry*, 425–443.

Ugander, J., Backstrom, L., Marlow, C., & Kleinberg, J. (2012). Structural diversity in social contagion. *Proceedings of the National Academy of Sciences*, 201116502.

Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The Anatomy of the Facebook Social Graph. *arXiv:1111.4503*.

Union, P. O. o. t. E. (2016). Inferring migrations, traditional methods and new approaches based on mobile phone, social media, and other big data : Feasibility study on inferring (labour) mobility and migration in the European Union from big data and social media data.

Wahba, J., & Zenou, Y. (2005). Density, social networks and job search methods: Theory and application to Egypt. *Journal of Development Economics*, *78*(2), 443–473.

Wesolowski, A., Buckee, C. O., Engø-Monsen, K., & Metcalf, C. J. E. (2016). Connecting mobility to infectious diseases: The promise and limits of mobile phone data. *The Journal of infectious diseases*, *214*(suppl_4), S414–S420.

Wesolowski, A., Eagle, N., Tatem, A. J., Smith, D. L., Noor, A. M., Snow, R. W., & Buckee, C. O. (2012). Quantifying the impact of human mobility on malaria. *Science*, *338*(6104), 267–270.

Wesolowski, A., Metcalf, C. J. E., Eagle, N., Kombich, J., Grenfell, B. T., Bjørnstad, O. N., Lessler, J., Tatem, A. J., & Buckee, C. O. (2015). Quantifying seasonal population fluxes driving rubella transmission dynamics using mobile phone data. *Proceedings of the National Academy of Sciences*, *112*(35), 11114–11119.

White, M. J. (2016). *International handbook of migration and population distribution*. Springer.

Willekens, F. (2008). Models of migration: Observations and judgement. *International migration in Europe: Data, models and estimates*, 117–147.

Williams, N. E., Thomas, T. A., Dunbar, M., Eagle, N., & Dobra, A. (2015). Measures of Human Mobility Using Mobile Phone Records Enhanced with GIS Data. *PLOS ONE*, *10*(7), e0133630.

Winters, P., de Janvry, A., & Sadoulet, E. (2001). Family and Community Networks in Mexico-U.S. Migration. *The Journal of Human Resources*, *36*(1), 159–184.

World Bank. (2009). Ghana socio-economic panel survey household instrument wave one. *https://microdata.worldbank.org/index.php/catalog/2534*.

Yang, Y., Liu, Z., Tan, C., Wu, F., Zhuang, Y., & Li, Y. (2018). To Stay or to Leave: Churn Prediction for Urban Migrants in the Initial Period, In *Proceedings of the 2018 World Wide Web Conference*, Republic, Canton of Geneva, Switzerland, International World Wide Web Conferences Steering Committee.

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014). Inferring international and internal migration patterns from Twitter data, In *Proceedings of the 23rd International Conference on World Wide Web - WWW '14 Companion*, Seoul, Korea, ACM Press.

Zagheni, E., & Weber, I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates, In *Proc. websci'12*. ACM.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's Advertising Platform to Monitor Stocks of Migrants. *Population and Development Review*, *43*(4), 721–734.

Zook, M. A., & Brunn, S. D. (2005). Hierarchies, Regions and Legacies: European Cities and Global Commercial Passenger Air Travel. *Journal of Contemporary European Studies*, *13*(2), 203–220.

Appendices

Appendix A

Chapter 2 Additional Materials

A.1 Proofs

Proof of Proposition 1: Consider any agent i and any of her distance-2 neighbors k, and let $G' = G \setminus \{k\}$. To show the existence of such threshold λ_{ik}, it is sufficient to show the following three parts are true. First, when $\lambda = 0$, agent i's diffusion centrality is higher in network G than that in network G'. This is straight forward, because when there is no competition among neighbors, distance-2 neighbors always increase the diffusion centrality which is a sum of information one gets from her neighbors and distance-2 neighbors. Second, when $\lambda = 1$, agent i's diffusion centrality is lower in network G than that in network $G \setminus \{k\}$. Third, the difference in diffusion centrality for any given q (recall $T = 2$)

$$DC_i(G; \lambda, q) - DC_i(G'; \lambda, q)$$

decreases in λ.

For the second part, let $\lambda = 1$ and let agent j be one of i's neighbors who are connected to agent k. Let d_j be agent j's degree in network G, which is at least two since he or she is connected to both i and k. The information capital agent i gets from agent j in network G is then (recall $\lambda = 1$)

$$DC_{ij}(G; q) = q\frac{1}{d_i d_j} + q^2 \sum_{h \in N_j} \frac{1}{d_i d_j^2 d_h}.$$

The first term is the direct information i gets from j, and the second term is the indirect information i gets from j's neighbors. On the other hand, without agent k, the information capital agent i gets from agent j is

$$DC_{ij}(G'; q) = q\frac{1}{d_i(d_j - 1)} + q^2 \left(\sum_{h \in N_j \setminus g_k} \frac{1}{d_i(d_j - 1)^2 d_h} + \sum_{l \in N_j \cap N_k} \frac{1}{d_i(d_j - 1)^2(d_l - 1)} \right).$$

Without agent k, agent j's degree decreases by one and so does any of j and k's common neighbors l. Also, agent i on longer gets indirect information from k, which is reflected as

$(N_j \setminus g_k) \cup (N_j \cap N_k) = N_j \setminus \{k\}$. We have,

$$\begin{aligned}
&DC_{ij}(G';q) - DC_{ij}(G;q) \\
&\geq q\left(\frac{1}{d_i(d_j-1)} - \frac{1}{d_i d_j}\right) + q^2\left(\sum_{h\in N_j\setminus\{k\}}\left(\frac{1}{d_i(d_j-1)^2 d_h} - \frac{1}{d_i d_j^2 d_h}\right) - \frac{1}{d_i d_j^2 d_k}\right) \\
&\geq q\left(\frac{1}{d_i(d_j-1)} - \frac{1}{d_i d_j}\right) - q^2\frac{1}{d_i d_j^2 d_k} \\
&= q\frac{1}{d_i(d_j-1)d_j} - q^2\frac{1}{d_i d_j^2 d_k} \\
&> 0.
\end{aligned}$$

This is true for all $j \in N_i \cap N_k$. So the second part is true that when $\lambda = 1$, agent i's diffusion centrality in network G' is higher.

Third, we consider the difference in agent i's diffusion centrality from neighbor j:

$$\begin{aligned}
&DC_{ij}(G';\lambda,q) - DC_{ij}(G;\lambda,q) \\
&= q\left(\frac{1}{d_i^\lambda(d_j-1)^\lambda} - \frac{1}{d_i^\lambda d_j^\lambda}\right) - q^2\frac{1}{d_i^\lambda d_j^{2\lambda} d_k^\lambda} + q^2\sum_{h\in N_j\setminus g_k}\left(\frac{1}{d_i^\lambda(d_j-1)^{2\lambda} d_h^\lambda} - \frac{1}{d_i^\lambda(d_j)^{2\lambda} d_h^\lambda}\right) \\
&\quad + q^2\sum_{l\in N_j\cap N_k}\left(\frac{1}{d_i^\lambda(d_j-1)^{2\lambda}(d_l-1)^\lambda} - \frac{1}{d_i^\lambda(d_j)^{2\lambda} d_l^\lambda}\right). \qquad \text{(A.1)}
\end{aligned}$$

Clearly, each of the four terms in (A.1) increases as λ increases. So we prove the third part of the monotonicity of the difference in the two diffusion centrality. ∎

Proof of Proposition 2: We construct the equilibrium as follows. Consider the partnership between i and j; the common knowledge they share about the network includes $g_{ij} = g_i \cap g_j$ and $G_{ij} = G_i \cap G_j$.

First, we identify the maximal effort for each clique with m agents.

$$b(a) \leq a + (m-1)\int_0^\infty e^{-rt}\delta a dt,$$

in which $b(a)$ is the gain from deviation and the right hand side is the payoff of each agent from all m agent cooperating at effort a. The effort $a^{c=m}$ binds this inequality.

Then, we claim there exists a maximal effort for the link ij subject to their shared common knowledge. If $g_{ij} = \{i,j\}$, then this maximal effort is $a^{c=2}$, otherwise it can be found by induction as illustrated below. From now on, we focus on the shared local network (g_{ij}, G_{ij}). We say a subset of agents is *fully-connected* if every agent in the subset is connected to everyone else in the subset. When the largest clique(s) in (g_{ij}, G_{ij}) has $h+2$ agents, then the induction takes h steps:

- In step 1, find the largest clique(s), for example, $g_{ijk_1\dots k_h}$. Then assign the effort $a(k_mk_l|ijk_1\dots k_h) = a^{c=h+2}$ to each link k_mk_l within the clique. That is, it is common knowledge among agents in the clique that each link can sustain effort at least $a^{c=h+2}$.

- In step 2, find all subsets of fully-connected agents containing $h+1$ agents, including i and j (this must always hold for all subsets we discuss, so omitted below). For any of them, say $g_{ijk'_1\dots k'_{h-1}}$, assign $a(k'_mk'_l|ijk'_1\dots k'_{h-1})$ to each link $k'_mk'_l$ to bind the inequality:

$$b(a) \leq a + \int_0^\infty e^{-rt}\delta \left(ha + \sum_{l\in g_{ijk'_1\dots k'_{h-1}}\setminus\{i,j,k'_1,\dots,k'_{h-1}\}} a(il|ijk'_1\dots k'_{h-1}l) \right) dt.$$

  That is, everyone in the clique uses the effort a and for other links that all of them can observe, the effort level is determined in the previous step (step 1).

- ...

- In step η, find all subsets of fully-connected agents containing $(h+3-\eta)$ agents. For any of them, say $g_{ijk''_1\dots k''_{h+1-\eta}}$, assign $a(k''_mk''_l|ijk''_1\dots k''_{h+1-\eta})$ to each link $k''_mk''_l$ to bind the inequality:

$$b(a)\leq a+\int_0^\infty e^{-rt}\delta\left((h+2-\eta)a+\sum_{l\in g_{ijk''_1\dots k''_{h+1-\eta}}\setminus\{i,j,k''_1,\dots,k''_{h+1-\eta}\}} a(il|ijk''_1\dots k''_{h+1-\eta}l)\right)dt.$$

- ...

- In step $h+1$, the only subset containing 2 agents and including i and j is the set $\{i,j\}$. The effort between them (a^*_{ij}) must bind the inequality:

$$b(a) \leq a + \int_0^\infty e^{-rt}\delta \left(a + \sum_{l\in g_{ij}\setminus\{i,j\}} a(il|ijl) \right) dt.$$

By construction, each effort level is the highest effort that is sustainable given the (higher-order) common knowledge of the network. Thus, a^*_{ij} is the maximal effort sustainable between ij subject to their shared knowledge of the network. In addition, as long as no one in g_{ij} has deviated, i and j can sustain a^*_{ij}. Thus, the strategy is robust. ∎

A.2 A Network Game Approach

In the benchmark model, we assume the total utility each agent gets from the network is a linear combination of information capital and cooperation capital as in equation (2.5). To

allow more complex features of network structures to influence the value an agent gets from the social network, one possibility is to consider a network game approach.

Each agent i chooses an action a_i, which could be socializing with friends, cooperating with them or both. Let $\mathbf{a} = (a_1, \ldots, a_n)$ be the strategy profile. We use the matrix format of a network G, such that $G_{ij} = G_{ji} = 1$ when i and j are connected. Let the matrix G^s be the network of links that are supported in the baseline network G, that is $G^s_{ij} = G^s_{ji} = 1$ if and only if ij is supported in G. Agent i derives the following quadratic utility, which has been commonly-used in network games (Jackson & Zenou, 2015):

$$u_i(\mathbf{a}, G) = \pi a_i - \frac{a_i^2}{2} + \phi \sum_{j=1}^{n} G_{ij} a_i a_j + \alpha \sum_{j=1}^{n} G^s_{ij} a_i a_j. \tag{A.2}$$

The first two terms $\pi a_i - \frac{a_i^2}{2}$ represent a linear benefit and a quadratic cost to agent i from choosing a_i. When $\phi > 0$, the third term $\phi \sum_{j=1}^{n} G_{ij} a_i a_j$ reflects the strategic complementarity between neighbors' actions and one's own action.[1] And the last term $\alpha > 0$ reflects the additional complementarity between supported neighbors.

We add two remarks about the utility function. First, the utility differs from a standard network game setup due to the last term, $\alpha \sum_{j=1}^{n} G^s_{ij} a_i a_j$. This is motivated by the theory results in Section 2.7 and the empirical results in Section 1.5 that an agent may derive additional utility from a supported neighbor. Second, if $\alpha = 0$, then the equilibrium action will be in proportion to the diffusion centrality in Section 2.7, $DC(G; q, \lambda, T)$ when $q = \phi$, $\lambda = 0$ and $T \to \infty$. In particular, ϕ can be viewed as the information passing probability q. The equilibrium action of agent i depends on the entire network structure, including her indirect neighbors and her supported links, and thus, this network approach allows for these network structures to jointly determine the equilibrium utility an agent gets from the network.

Let $\mu_1(G)$ be the spectral radius of matrix G, $\mathbf{I}$ be the identity matrix, and $\mathbf{1}$ be the column vector of 1.

Proposition 3 *If $\mu_1(\phi G + \alpha G^s) < 1$, the game with payoffs (A.2) has a unique (and interior) Nash equilibrium in pure strategies given by:*

$$\mathbf{a}^* = \pi(\mathbf{I} - \phi G - \alpha G^s)^{-1}\mathbf{1}. \tag{A.3}$$

Consider the first-order necessary condition for each agent i's action:

$$\frac{\partial u_i(\mathbf{a}, G)}{\partial a_i} = \pi - a_i + \phi \sum_{j=1}^{n} G_{ij} a_j + \alpha \sum_{j=1}^{n} G^s_{ij} a_j = 0.$$

[1] While it is unlikely in our setup, ϕ could be negative in some network games, which then reflects the substitution between neighbors' actions and one's own action.

This leads to

$$a_i^* = \pi + \phi \sum_{j=1}^{n} G_{ij} a_j^* + \alpha \sum_{j=1}^{n} G_{ij}^s a_j^*. \tag{A.4}$$

In the matrix form: $\mathbf{a}^* = \pi \mathbf{1} + \phi G \mathbf{a}^* + \alpha G^s \mathbf{a}^*$, which leads to the solution in (A.3).

A simple way to prove this solution is indeed the unique (and interior) Nash equilibrium, as noted for example by (Bramoullé et al., 2014), is to observe that this game is a potential game (as defined by (Monderer & Shapley, 1996)) with potential function:

$$P(\mathbf{a}, G, \phi) = \sum_{i=1}^{n} u_i(\mathbf{a}, G) - \frac{\phi}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} G_{ij} a_i a_j - \frac{\alpha}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} G_{ij}^s a_i a_j.$$

We omit the details of the analogous proof, which can be found in (Bramoullé et al., 2014) and (Jackson & Zenou, 2015).

In the equilibrium, the utility of agent i is given by

$$\begin{aligned} u_i(\mathbf{a}^*, G) &= \pi a_i^* - \frac{a_i^{*2}}{2} + \phi \sum_{j=1}^{n} G_{ij} a_i^* a_j^* + \alpha \sum_{j=1}^{n} G_{ij}^s a_i^* a_j^* \\ &= a_i^* \left(\pi + \phi \sum_{j=1}^{n} G_{ij} a_j^* + \alpha \sum_{j=1}^{n} G_{ij}^s a_j^* \right) - \frac{a_i^{*2}}{2}. \end{aligned}$$

By equation (A.4), $u_i(\mathbf{a}^*, G) = (a_i^*)^2/2$, which by equation (A.3) depends on (π, ϕ, α, G). So in this way, we can estimate how an agent's utility depends on the interaction with neighbors ϕ, the added value of a supported link α, and his or her position in the network G.

More generally, the network game can be enriched to capture the possibilities of competition with indirect neighbors, as we modeled in Section 2.7. For example, (Ballester et al., 2006) consider a global congestion effect by adding the term $-\lambda a_i \sum_{j=1}^{n} a_j$ to each agent i's utility. Using the corresponding equilibrium utility with this congestion λ, one could also estimate the rivalry or competition with indirect neighbors.

A.3 Robustness of Model Calibration

Our benchmark model assumes that an individual will migrate if the total utility of the destination network exceeds the total utility of the home network (equation 2.1), and assumes that the total utility an agent i receives from an arbitrary network G can be expressed as a linear combination of the information capital and cooperation capital of G (equation 2.5). This highly stylized formulation is intended to contrast, as transparently as possible, what the literature has emphasized are the two main mechanisms through which social networks provide utility. Here, we explore alternative formulations of models (2.1) and (2.5), to test the robustness of the calibration results in Section 2.7.

Fixed migration costs

We first allow for the migration decision (equation 2.1) to include a fixed threshold (cost) τ, in addition to the idiosyncratic error ε_i:

$$u_i(G^d) > u_i(G^h) + \tau + \varepsilon_i. \tag{A.5}$$

Here, τ is meant to capture the possibility that all people might share a common aversion to migrating; accounting for this shared cost might help us identify the main parameters of interest.

When model (A.5) is calibrated with the data, the main observations in Section 2.7 persist. Full calibration plots for all parameters $< \lambda, \alpha^d, \alpha^h, \tau, \pi^{I,d}, \pi^{C,d}, \pi^{I,h} >$ are shown in Figure A.14. Most importantly, the optimal value of the rivalry coefficient remains at $\lambda = 0.5$ (top left). Similar to the results presented in the main text, supported links are more valuable than unsupported links (i.e., α^D and α^H are both greater than 0). In particular, α^D is exactly 5 as in the main model, and α^h decreases slightly from 1 to 0.5.

Second, the total utility from information capital and cooperation capital contribute relatively the same amount to an agent's total utility from the network. This can be seen most clearly in Figure A.15. The bulk of the distribution of u_i^I and u_i^C lies around the 45-degree line, which is where $u_i^I = u_i^C$.

The calibration sensitivity plot for the new parameter, τ, is shown in the middle-right panel of Figure A.14. This calibration is more noisy, with the optimal calibrated threshold at $\tau = -5$. This is perhaps surprising, since a literal interpretation of τ is as an average migration cost, which should be positive. However, the vast majority of agents in our simulation have considerably larger home networks than destination networks (see the bottom panels of Figure 2.5); it is likely that the negative τ is offsetting the fact that in our balanced sample home utility generally exceeds destination utility.

Cobb-Douglas utility

Next, we consider a Cobb-Douglas network utility function, which can be rewritten as the total utility being a log-linear combination of information capital and cooperation capital. Specifically, equation (2.18) becomes

$$\begin{aligned} &\pi^{I,d} \log DC_i(G^d; q, \lambda, T) + \pi^{C,d} \log \left(d_i(G^d) + \alpha^d d_i^S(G^d)\right) \\ &\quad > \pi^{I,h} \log DC_i(G^h; q, \lambda, T) + \pi^{C,h} \log \left(d_i(G^h) + \alpha^h d_i^S(G^h)\right) + \varepsilon_i. \end{aligned} \tag{A.6}$$

We note that the linear utility function and the Cobb-Douglas utility function describe fundamentally different ways that agents value the network. A key difference is that the information capital and cooperation capital are substitutable in the linear utility function, but they are complementary in the Cobb-Douglas utility function. To get a high utility based on the Cobb-Douglas form, an agent needs both a high information capital and a high cooperation capital, while only one is needed based on the linear form. As a result, we want to

confirm the main takeaways are robust, although we do not expect all the parameterizations are exactly the same.

We find that the main observations in section 2.7 persist. The log-linear model correctly predicts 68.6% of the migration decisions, which is close to, though slightly below, the accuracy of the model in the text, which is 69.5%. The parameterization plots for $< \lambda, \alpha^d, \alpha^h, \pi^{I,d}, \pi^{C,d}, \pi^{I,h} >$ are shown in Figure A.16. As before, the optimal value of the rivalry coefficient remains at $\lambda = 0.5$. Similarly, supported links are more valuable than unsupported links, although the particular values differ from the main model: $\alpha^d = 0.5$ and $\alpha^h = 10$.

Figure A.17a shows the extent to which information capital and cooperation capital contribute to the agent's total utility from the network. Cooperation capital contributes roughly twice as much as information capital, which differs from the equal contribution in the main specification. This shows that the fact that both information capital and cooperation capital contribute significantly to the total social capital is a robust result, but the relative weights of the two may depend on their interactions (substitutes or complementary). It's worth to note that it remains the case that when λ is optimally parameterized, the information capital contributes significantly more to total utility than when we remove the possibility for rivalry by setting $\lambda = 0$. This contrast can be seen by comparing the left ($\lambda = 0.5$) and right ($\lambda = 0$) panels of Figure A.17. In other words, regardless of the specific utility functions, the information capital if in the form of the original diffusion centrality does not contribute to the social capital (relative to the cooperation capital), which further supports the finding of rivalry in competing for neighbors' attention.

A.4 Algorithms

Data: $< ID, datetime, location >$ tuples for each mobile phone interaction
Result: $< ID, month, district >$ tuples indicating monthly modal district
Step 1 Find each subscriber's most frequently visited tower;
→ Calculate *overall daily modal districts*;
→ Calculate *overall monthly modal districts*;
Step 2 calculate the *hourly modal districts*;
if *tie districts exit* **then**
  if *overall daily modal districts can resolve* **then**
    return the district with larger occurance number;
    else
      if *overall monthly modal districts can resolve* **then**
        return the district with larger occurance number
      end
    end
  end
end
Step 3 calculate the *daily modal districts*;
if *tie districts exit* **then**
  if *overall daily modal districts can resolve* **then**
    return the district with larger occurance number;
    else
      if *overall monthly modal districts can resolve* **then**
        return the district with larger occurance number
      end
    end
  end
end
Step 4 calculate the *monthly modal districts*;
if *tie districts exit* **then**
  if *overall monthly modal districts can resolve* **then**
    return the district with larger occurance number;
  end
end

Algorithm 1: Home location assignment

Data: Monthly modal district for four consecutive months: D_1, D_2, D_3, D_4
Result: Migration type

if $D_1 == D_2$ *AND* $D_3 == D_4$ **then**
  if $D_2 == D_3$ **then**
    if $D_4 == Kigali$ **then**
      migration type is **urban resident**
    end
    else
      migration type is **rural resident**
    end
  end
  else
    if $D_4 == Kigali$ **then**
      migration type is **rural to urban**
    end
    else
      if $D_1 == Kigali$ **then**
        migration type is urban to rural
      end
      else
        migration type is **rural to rural**
      end
    end
  end
end
else
  migration type is **other**
end

Algorithm 2: Classifying individuals by migrant type for k=2

A.5 Appendix Figures and Tables

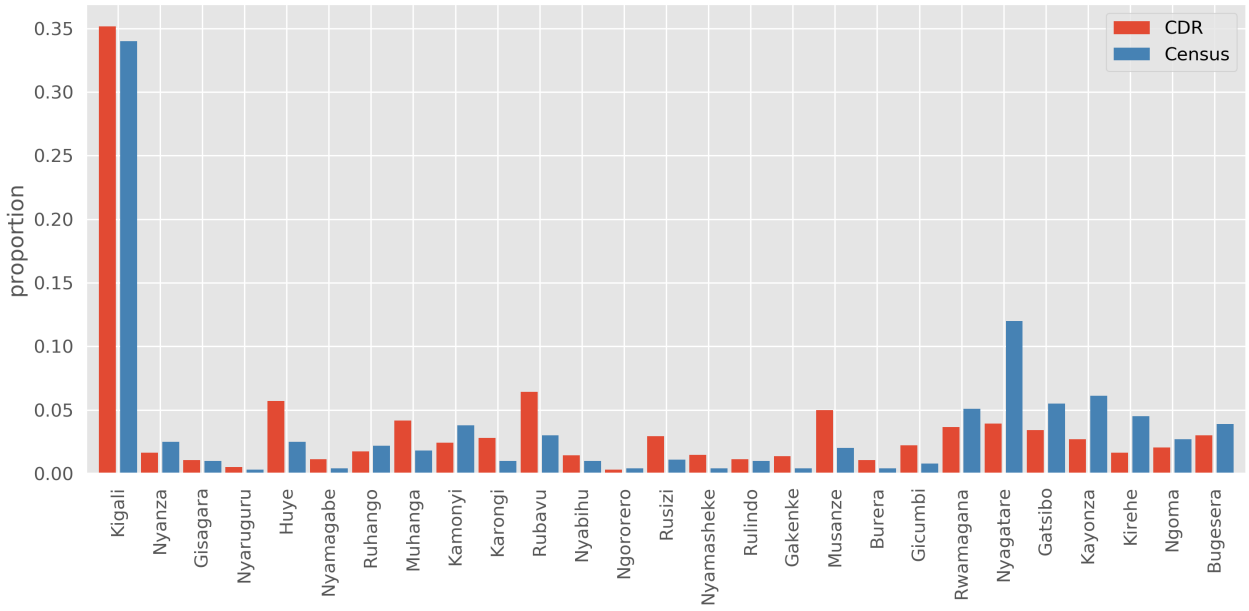
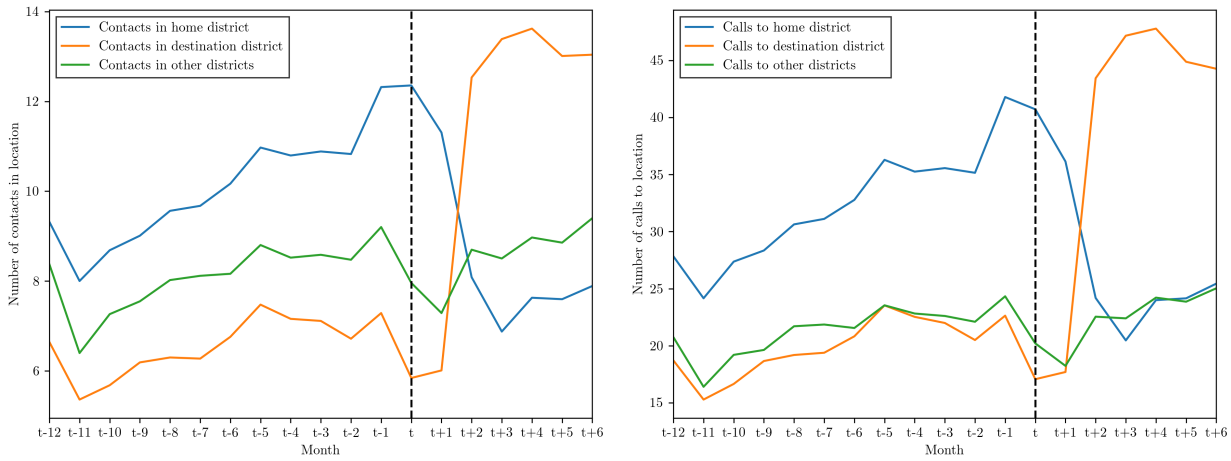


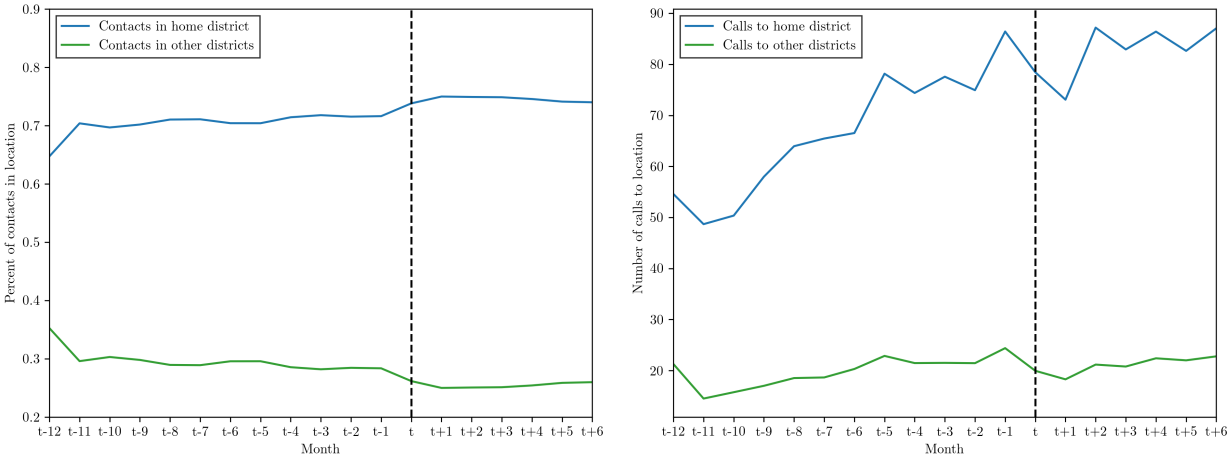
Figure A.1: Validation of Migration Data. *Notes:* Figure shows the proportion of migrants to each district in Rwanda. Red bars indicate the proportion inferred from the mobile phone data; Blue bars indicate the proportion calculated from 2012 Rwandan census data, as reported by National Institute of Statistics of Rwanda (2014).



(a) Number of contacts

(b) Number of calls

Figure A.2: Network structure of non-migrants



(a) Percent of contacts

(b) Percent of calls

Figure A.3: Network structure of migrants. *Notes:* Top figures shows how the network connections of migrants evolves over time, in each of the 12 months before and 6 months after migration. These are similar to Figure 2.4, except that instead of showing the *percent* of calls to each location, Figure A.2a plots the *number* of unique contacts in each location and Figure A.2b indicates the number of *phone calls* to each location. Bottom figures show equivalent figures for *non*-migrants, as a sort of placebo test. For non-migrants, the index month t is sampled from the same distribution of months in which actual migrations occur).

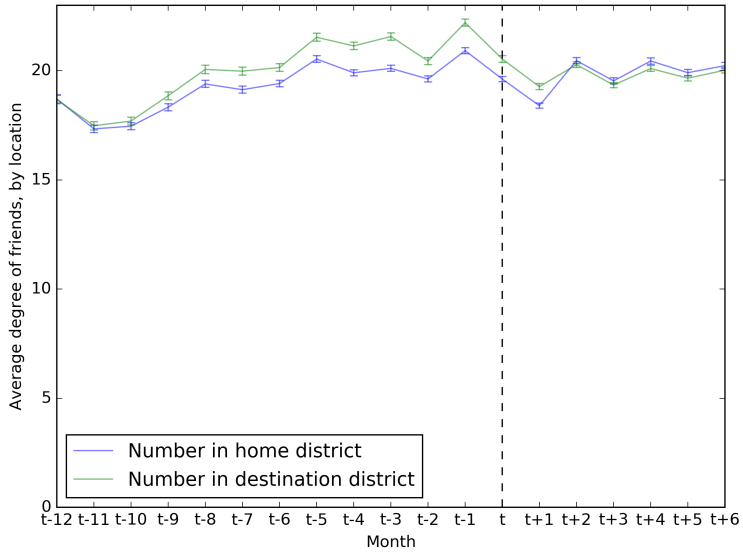


Figure A.4: Number of friends of friends, before and after migration (migrants)

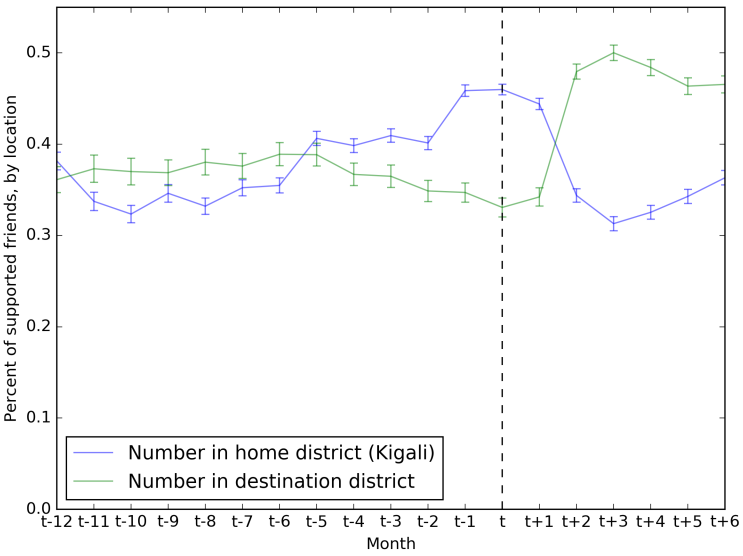
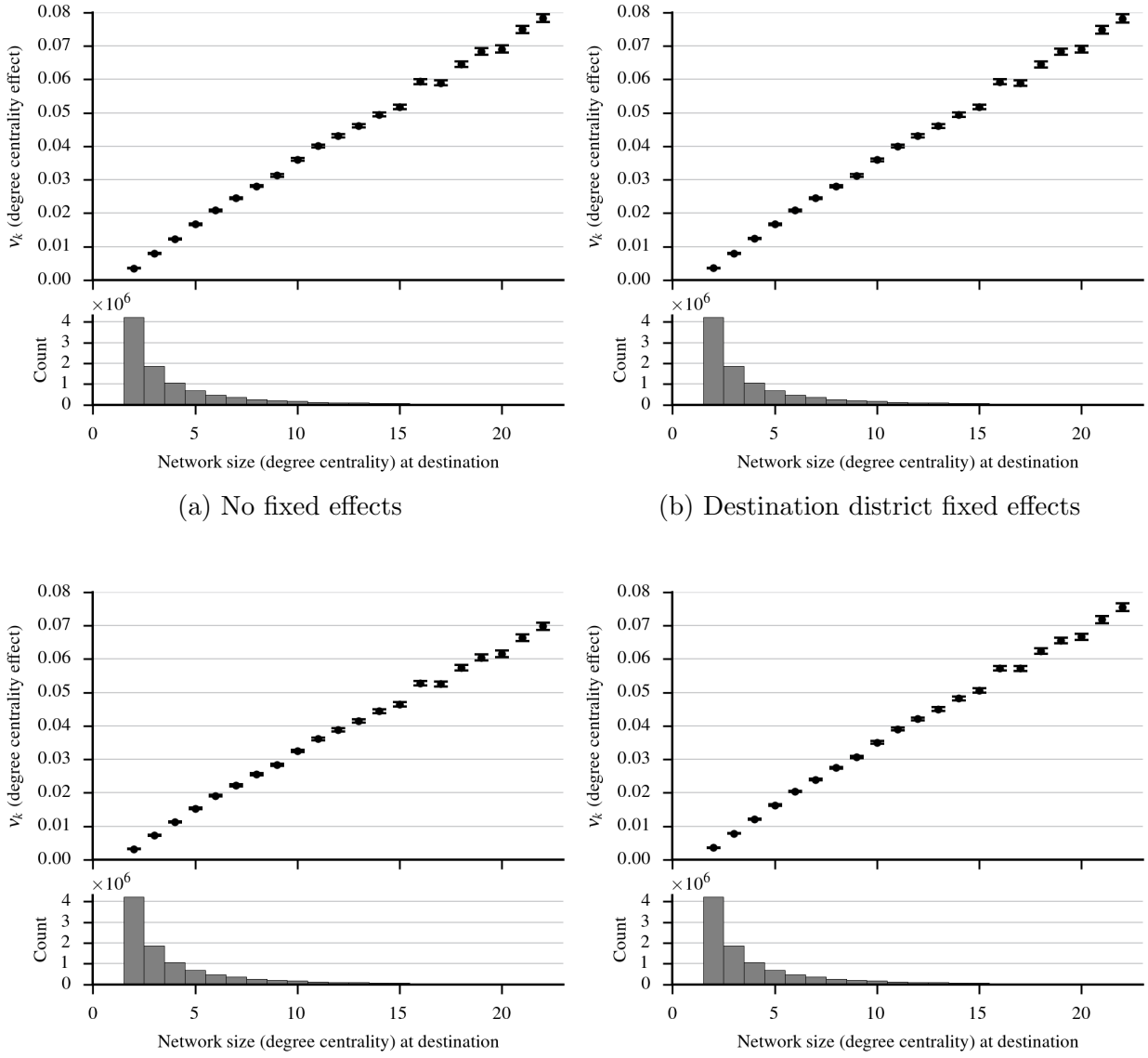


Figure A.5: Percent of friends with common support, before and after migration (migrants). *Notes:* Top figure shows total number of friends of friends migrants have in their home district and their destination district, in each of the 12 months before and 6 months after migration. Bottom figure shows the percent of the mgirants friends who have a common friend.



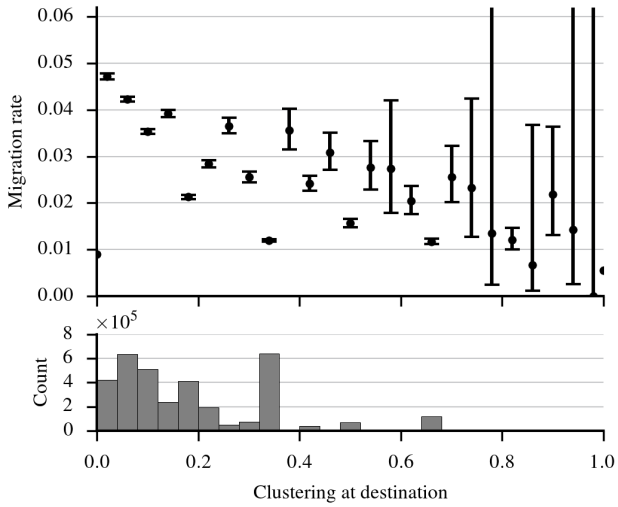
(a) No fixed effects

(b) Destination district fixed effects

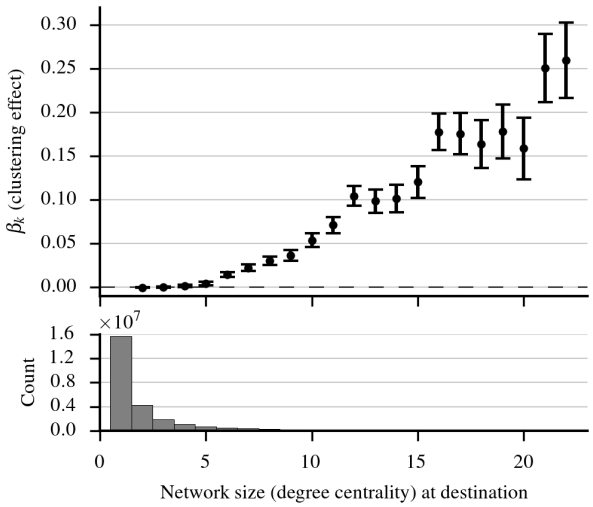
(c) Home-destination-month F.E.'s

(d) Home-dest-month & individual F.E.'s

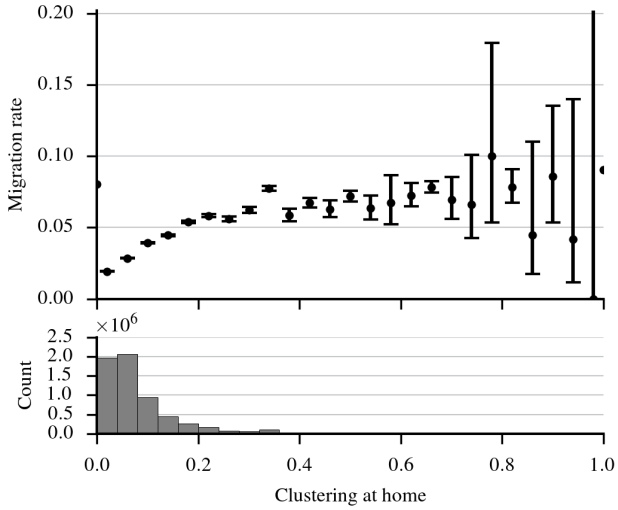
Figure A.6: Migration rate and degree centrality, controlling for different fixed effects. *Notes:* Each figure shows the fixed effect coefficients estimated from a regression of migration on separate fixed effects for each possible destination network size (see Section 2.6). Figure subtitle indicates any other fixed effects included in the specification. Error bars indicate 95% confidence intervals, clustered by individual.



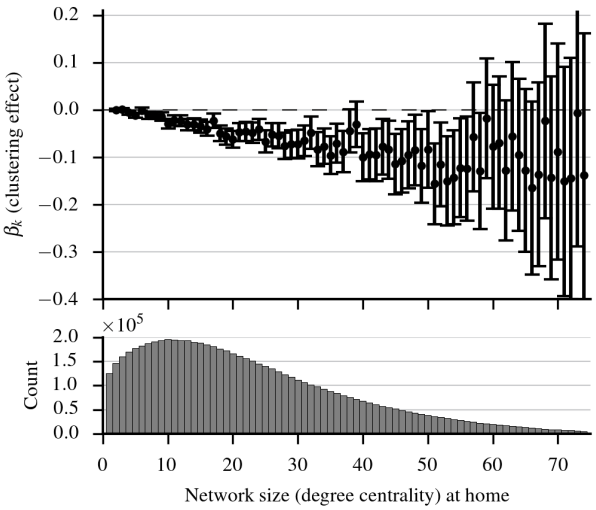
(a) Clustering at Destination



(b) Clustering at Destination, by Degree



(c) Clustering at Home



(d) Clustering at Home, by Degree

Figure A.7: Relationship between migration rate and clustering. *Notes:* "Clustering" denotes the proportion of potential links between i's friends that exist. In all figures, the lower histogram shows the unconditional distribution of the x-variable. Top row (a and b) characterizes the destination network; bottom row (c and d) characterizes the home network. For the left column (a and c), the main figure indicates, at each level of weighted degree, the average migration rate. For the left column (b and d), the main figure indicates the correlation between the migration rate and clustering, holding degree fixed. In other words, each point represents the β_k coefficient estimated from a regression of $Migration_i = \alpha_k + \beta_k Clustering_i$, estimated on the population of i who have degree equal to k. Error bars indicate 95% confidence intervals, clustered by individual.

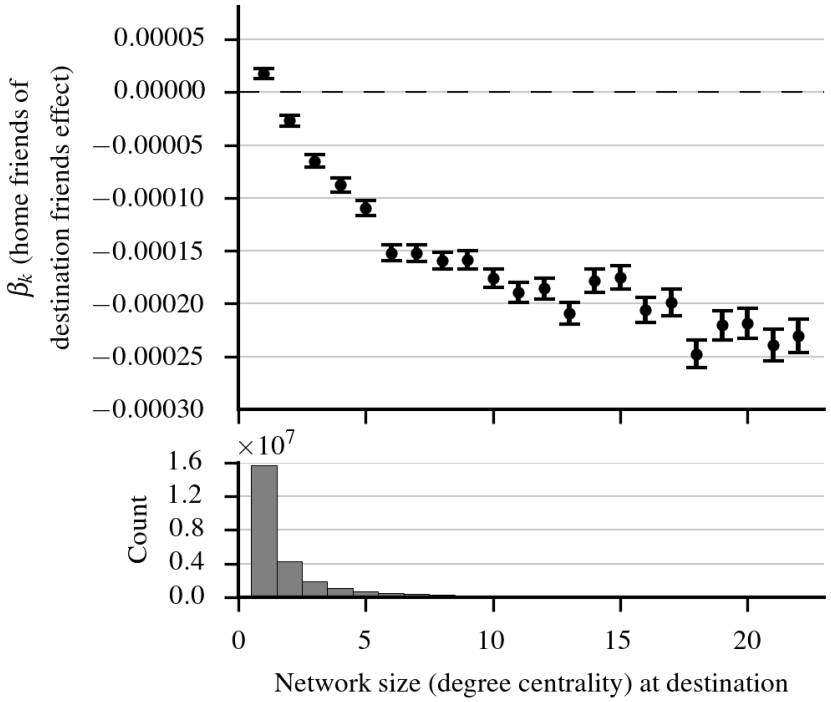


Figure A.8: Migration rate and home friends of friend in destination. *Notes:* Figure shows the β_k values estimated with model 2.7, i.e., the correlation between migration and unique friends (at home) of friends (in the destination) for individuals with different numbers of friends (in the destionation), after conditioning on fixed effects — see Section 2.6. Error bars indicate 95% confidence intervals, clustered by individual.

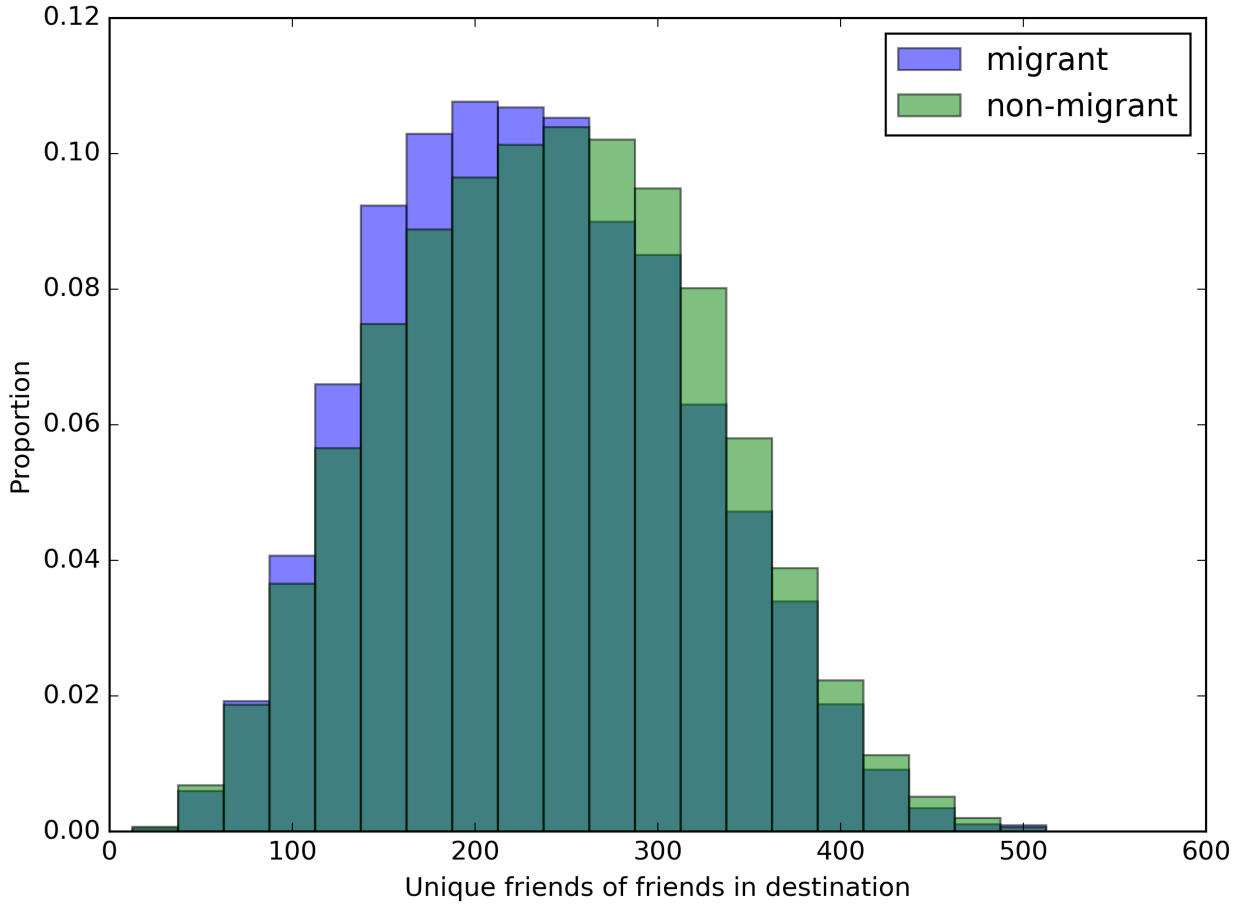


Figure A.9: Migrants have fewer friends of friends than non-migrants. *Notes:* The figure focuses on all individuals who have exactly 10 unique contacts in a potential destination, and shows the distribution of the number of unique "friends of friends" in that destination. Counterintuitively, migrants have fewer unique friends of friends than non-migrants.

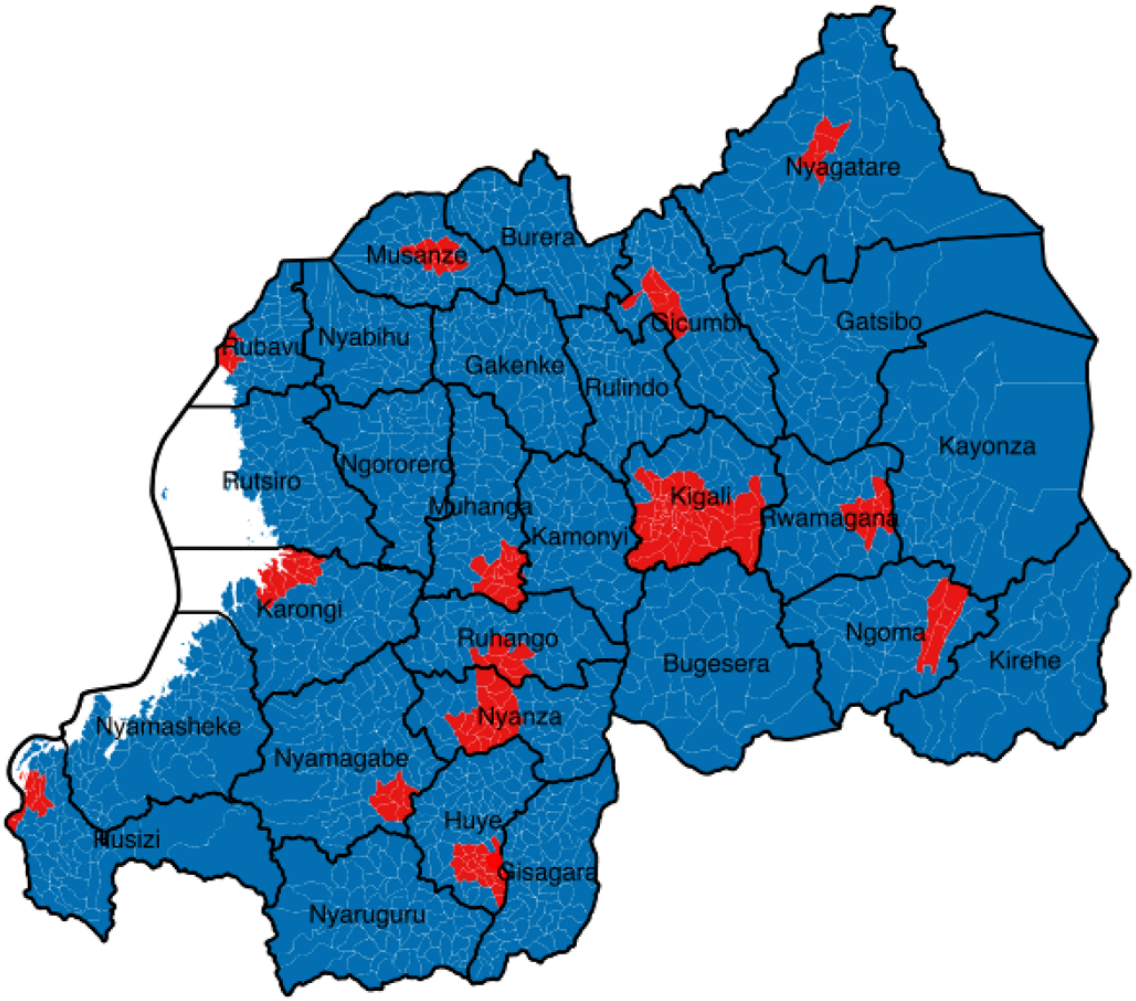


Figure A.10: Urban and rural sectors in Rwanda. *Notes:* Urban zones shown in red; rural zones shown in blue. Urban and rural designations detremined using the sector boundary dataset from the website of National Institute of Statistics Rwanda, available from http://statistics.gov.rw/geodata.

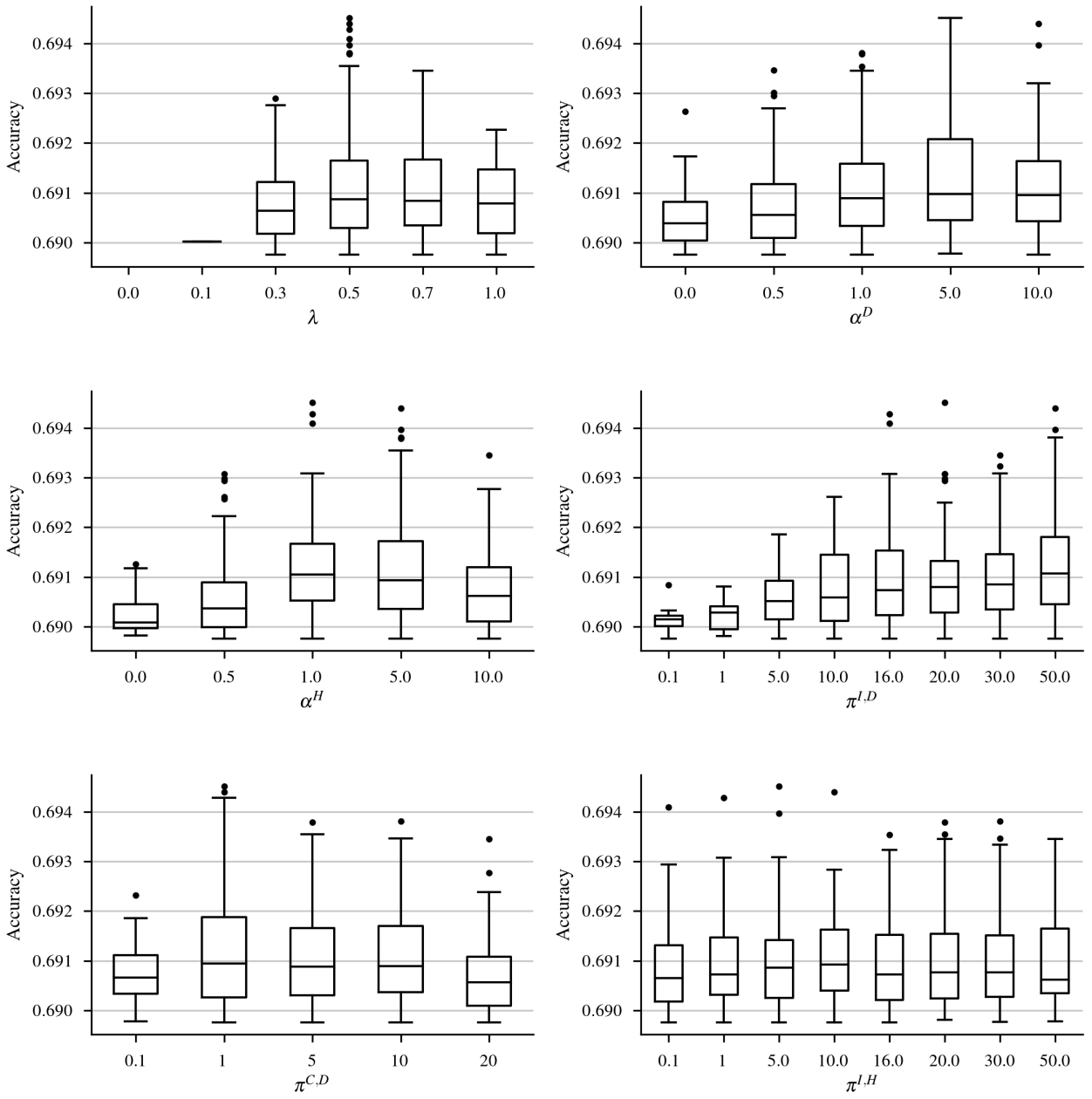


Figure A.11: Calibration results: marginal plots. *Notes:* Figures show the marginal effect of varying λ, α_d, α^h and $(\pi^{I,d}, \pi^{C,d}, \pi^{I,h})$ when calibrating Model 2.18. Each of roughly 50,000 different parameter combinations is tested; the top percentile of simulations are used to generate this marginal plot.

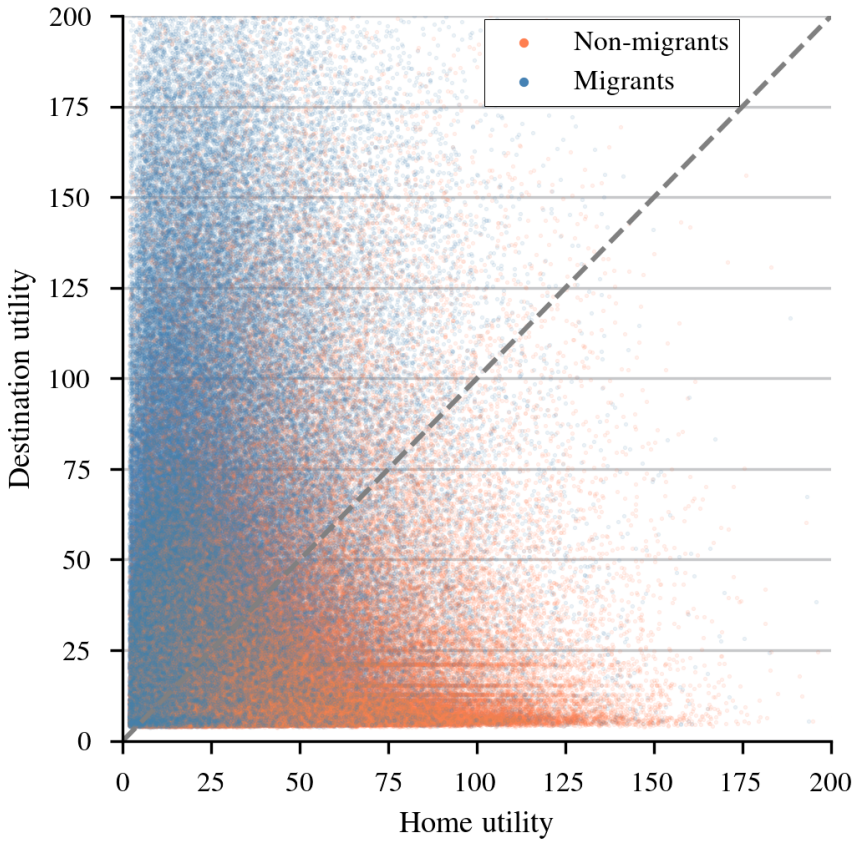


Figure A.12: Simulated balance of home vs. destination utility. *Notes:* After the model is calibrated, the optimal parameters are used to calculate the total utility provided to each individual by the home network and destination network. Each dot represents one individual's combination of predicted home-destination utility. Blue (red) dots above (below) the 45-degree line are correctly classified; blue (red) dots below (above) the 45-degree line are incorrectly classified.

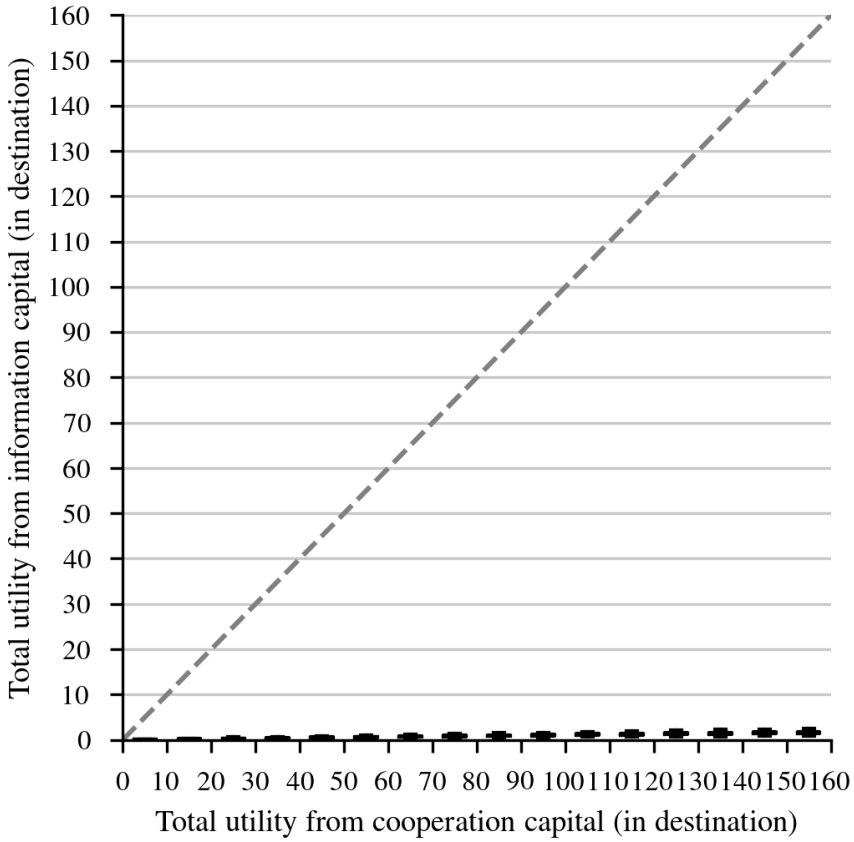


Figure A.13: Calibration results when $\lambda = 0$: 'information' and 'cooperation' utility. *Notes:* Figures show the distribution of predicted utility from 'information' and 'cooperation' (i.e., equation 2.5) for 270,000 migrants and non-migrants. It is calculated using the parameters selected by calibrating Model 2.18 with λ fixed at zero (i.e., no information rivalry).

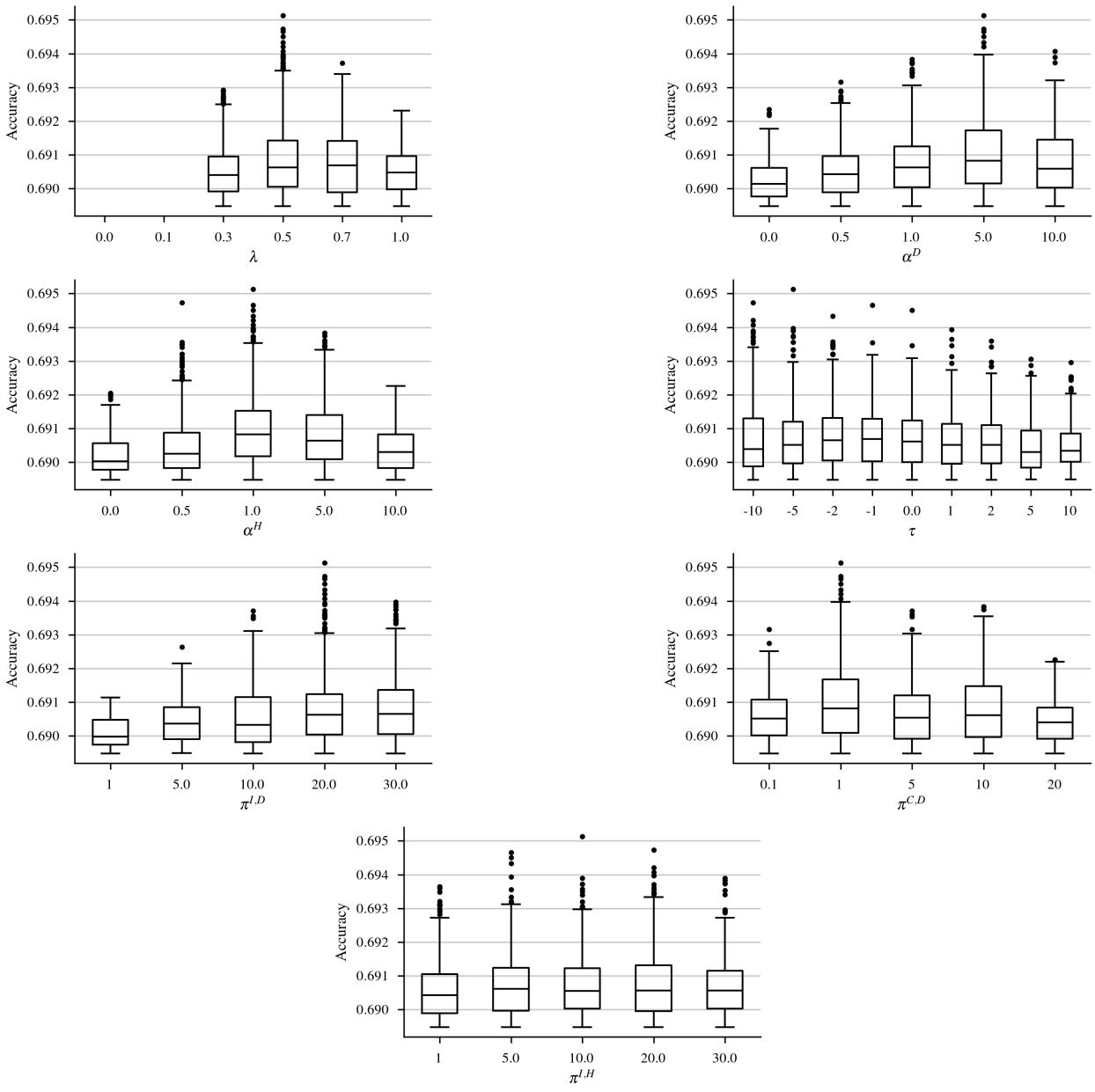


Figure A.14: Calibration results (with τ): marginal plots.

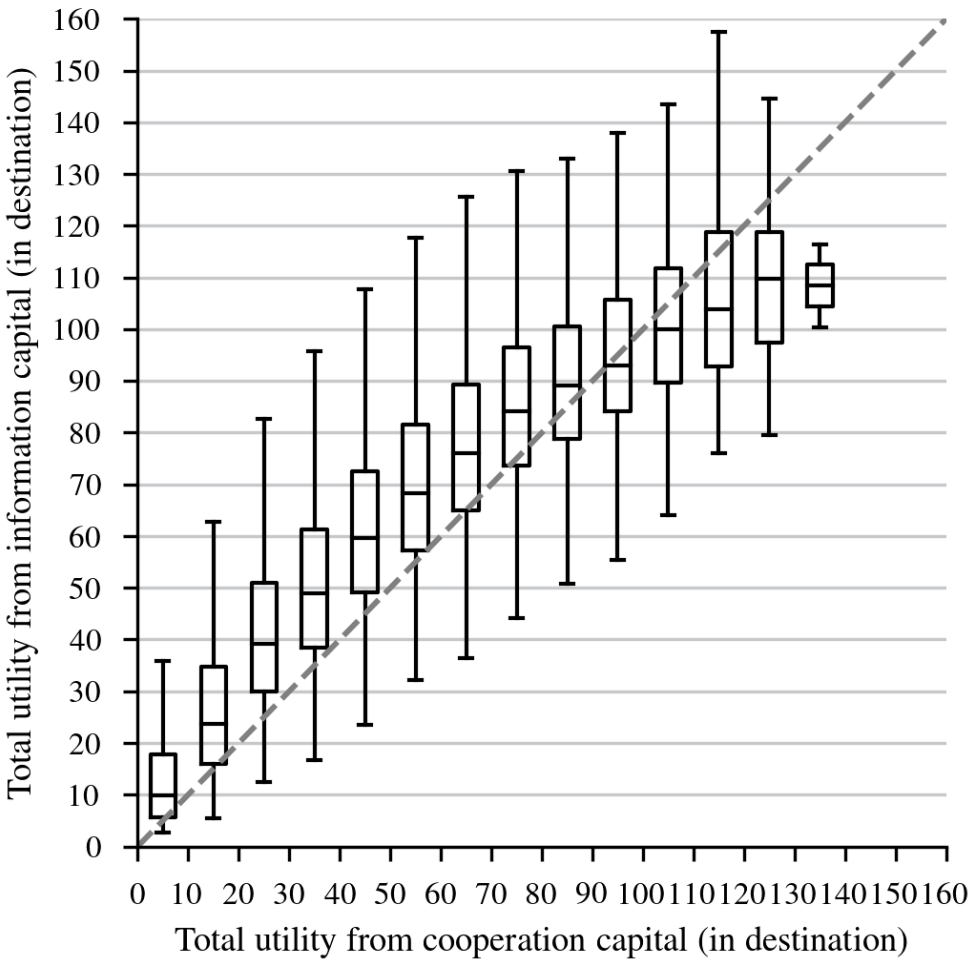


Figure A.15: Calibration results (with τ): 'information' and 'cooperation' utility.*Notes:* Figures show the distribution of predicted utility from 'information' and 'cooperation' (i.e., equation 2.5) for 270,000 migrants and non-migrants.

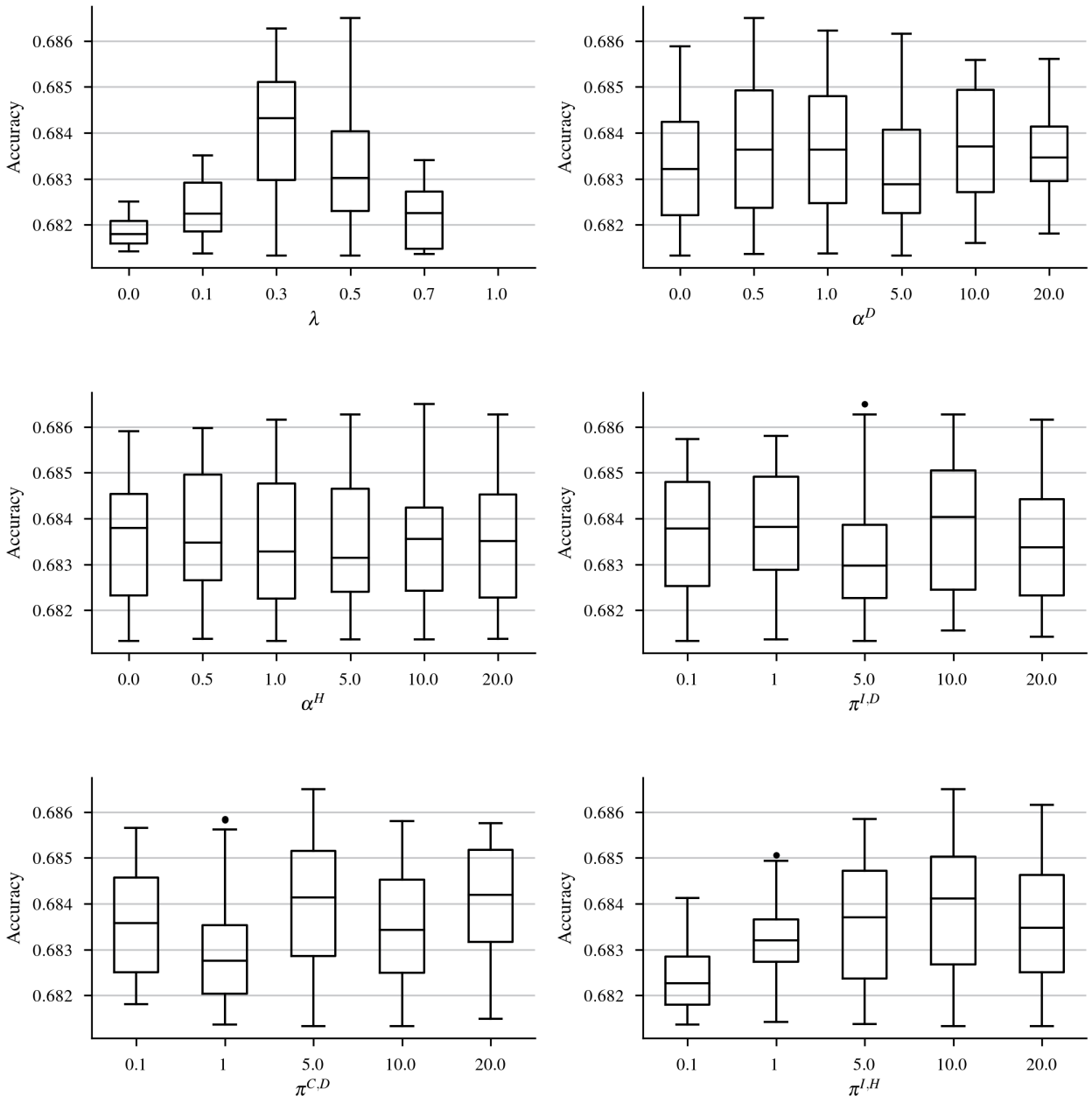
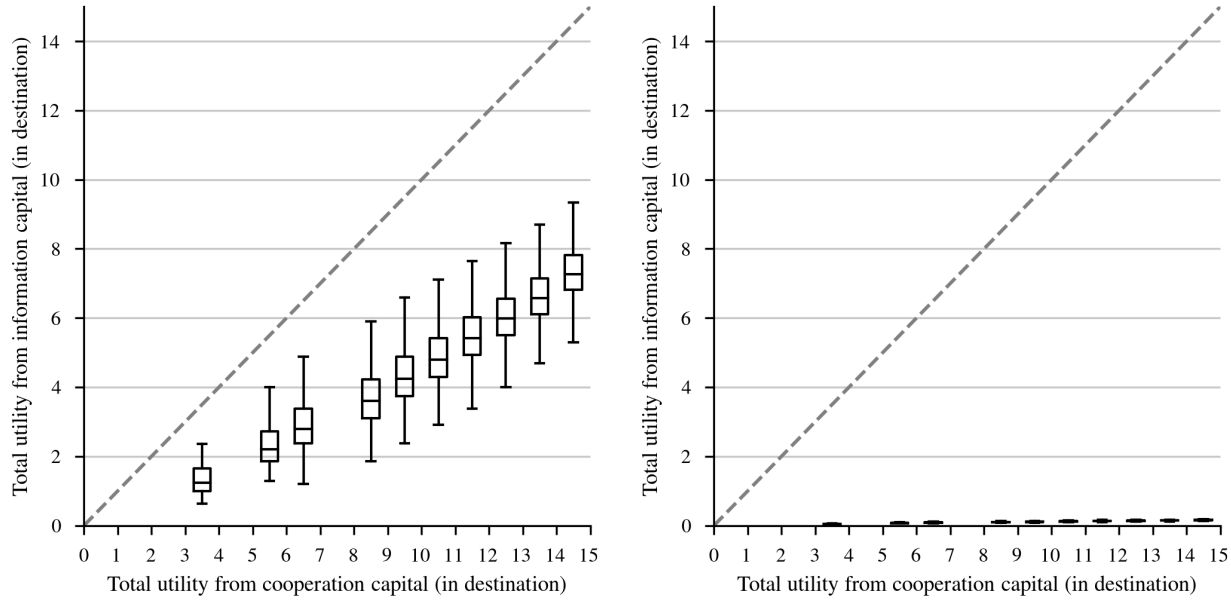


Figure A.16: Calibration results for log linear model: marginal plots. *Notes:* Figures show the marginal effect of varying λ, α_d, α^h and $(\pi^{I,d}, \pi^{C,d}, \pi^{I,h})$ when calibrating Model (A.6). Each of roughly 50,000 different parameter combinations is tested; the top percentile of simulations are used to generate this marginal plot.



(a) Rivalrous information transmission ($\lambda = 0.5$) (b) Non-rival information transmission ($\lambda = 0$)

Figure A.17: Calibration results for log linear model: 'information' and 'cooperation' utility. *Notes:* Figures show the distribution of predicted utility from 'information' and 'cooperation' (i.e., equation 2.5) for 270,000 migrants and non-migrants. The left figure is calculated using the parameters selected by calibrating Model A.6. For the right figure, λ is fixed at zero (i.e., no information rivalry).

Table A.1: Migration events observed in 4.5 years of phone data

Definition of Migrant (k)	Total Individuals (N)	% Ever Migrate	% Repeat migrants (to same district)	% Repeat migrants (to any district)	% Long-distance migrants (non-adjacent districts)	% Circular Migrants
1	935,806	34.565	11.171	21.923	23.181	18.457
2	680,267	21.634	1.933	8.244	13.828	5.934
3	518,156	13.960	0.405	2.893	9.216	2.007
6	263,182	5.294	0.000	0.192	3.547	0.128

Notes: Table counts number of unique individuals meeting different definitions of a "migration event." Each row of the table defines a migration by a different k, such that an individual is considered a migrant if she spends k consecutive months in a district d and then k consecutive months in a different district $d' \neq d$ – see text for details. Repeat migrants are individuals who have migrated one or more times prior to a migration observed in month t. Long-distance migrants are migrants who travel between non-adjacent districts. Circular migrants are migrants who have migrated from d to h prior to being observed to migrated from h to d. The number of individual (N) varies by row, since an individual is only considered eligible as a migrant if she is observed continuously over $2N$ consecutive months.

Table A.2: Jointly estimated effects of home and destination network structure

	(1)	(2)	(3)
Destination Degree (network size)	0.0048033***	0.0037637***	
	(0.0000201)	(0.0000238)	
Home Degree (network size)	−0.0007377***	−0.0005089***	
	(0.0000060)	(0.0000107)	
Destination friends of friends	−0.0000324***	−0.0000059***	−0.0000001
	(0.0000007)	(0.0000009)	(0.0000009)
Home friends of friends	0.0000113***	0.0000059***	−0.0000035***
	(0.0000002)	(0.0000004)	(0.0000004)
Destination % friends with support	0.0037855***	0.0017164***	0.0010618***
	(0.0001088)	(0.0001130)	(0.0001146)
Home % friends with support	0.0081299***	−0.0061902***	0.0002216
	(0.0001336)	(0.0002305)	(0.0002407)
Observations	9,889,981	9,889,981	9,889,981
R^2	0.0213936	0.1858886	0.1868505
Degree fixed effects	No	No	Yes
Home*Destination*Month fixed effects	No	Yes	Yes
Individual fixed effects	No	Yes	Yes

Notes: Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.3: Robustness to alternative fixed effect specifications

	(1)	(2)	(3)	(4)
Panel A: Destination network characteristics				
Degree (network size)	0.0036548***			
	(0.0000183)			
Friends of friends	−0.0000103***	−0.0000160***	−0.00000004	−0.0000002
	(0.0000007)	(0.0000007)	(0.0000008)	(0.0000009)
% Friends with common support	0.0010869***	0.0022076***	0.0028977***	0.0014808***
	(0.0001045)	(0.0001107)	(0.0001112)	(0.0001146)
Observations	9,889,981	9,889,981	9,889,981	9,889,981
Panel B: Home network characteristics				
Degree (network size)	−0.0003957***			
	(0.0000060)			
Friends of friends	0.0000021***	−0.0000109***	−0.0000165***	−0.0000110***
	(0.0000002)	(0.0000001)	(0.0000001)	(0.0000002)
% Friends with common support	0.0325365***	−0.0186718***	−0.0139236***	−0.0087495***
	(0.0001233)	(0.0001673)	(0.0001731)	(0.0002245)
Observations	9,889,981	9,889,981	9,889,981	9,889,981
Degree fixed effects	No	Yes	Yes	Yes
Home*Destination*Month fixed effects	No	No	Yes	Yes
Individual fixed effects	No	No	No	Yes

Notes: Each column indicates a separate regression of a binary variable indicating 1 if an individual i migrated from home district h to destination district d in month t. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.4: Robustness to alternative fixed effect specifications, part 2

	(1)	(2)	(3)	(4)	(5)
Destination friends of friends	−0.0000002	0.0000011	−0.0000064***	−0.0000077***	−0.0000028***
	(0.0000009)	(0.0000011)	(0.0000010)	(0.0000012)	(0.0000010)
% Destination friends with support	0.0014808***	0.0013719***	0.0003458***	0.0006663***	0.0001123
	(0.0001146)	(0.0001491)	(0.0001220)	(0.0000966)	(0.0001204)
Observations	9,889,981	9,889,981	9,889,981	9,889,981	9,889,981
R^2	0.1853017	0.5080845	0.5952072	0.6680641	0.6332967
Fixed effects	$D, h*d*t, i$	$D, h * d * t, i * t$	$D, h * d * t, i * d$	$D, h * d * t, i * D$	$D, h*d*i, t$

Notes: Each column indicates a separate regression of a binary variable indicating 1 if an individual i migrated from home district h to destination district d in month t. All specifications control non-parametrically for the number of unique contacts D that i has in district d. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.5: Conditional logit results

	(1)	(2)	(3)	(4)
Destination Degree (network size)	0.16427***	0.308192***	0.11818***	0.211611***
	(0.00106)	(0.002854)	(0.00114)	(0.003034)
Home Degree (network size)	-0.11931***	-0.261790***	-0.07906	-0.188931***
	(0.00114)	(0.002980)	(0.00128)	(0.003160)
Destination friends of friends		-0.005564***		-0.003503***
		(0.000108)		(0.000108)
Home friends of friends		-0.005442***		0.004055***
		(0.000112)		(0.000110)
Destination % friends with support			2.49114***	2.241620***
			(0.02788)	(0.030131)
Home % friends with support			-1.90396***	-1.57135***
			(0.01924)	(0.042690)
Home choice	6.10215***	6.114159***	6.10313***	6.082535***
	(0.01493)	(0.01514)	(0.01824)	(0.01813)
McFadden R^2	0.88563	0.88709	0.88864	0.88936
N individuals	433,782	433,782	433,782	433,782

Notes: Response variable in conditional logit is a dummy variable indicating whether inidividual i migrates from district h to district d in Jaunary 2008. Each choice represents one of the 27 districts in Rwanda (the three smaller urban districts in Kigali province are treated as a single district). Standard errors in parentheses. *p<0.1; **p<0.05; ***p<0.01.

Table A.6: Heterogeneity by Migration Frequency (Repeat and First-time)

Migration Frequency	(1) Any	(2) Repeat	(3) First-Time
Destination friends of friends	−0.0000001	0.0000171***	−0.0000030***
	(0.0000009)	(0.0000062)	(0.0000008)
Home friends of friends	−0.0000035***	−0.0000511***	0.0000022***
	(0.0000004)	(0.0000043)	(0.0000003)
% Destination support	0.0010618***	−0.0027428*	0.0010934***
	(0.0001146)	(0.0014071)	(0.0000920)
% Home support	0.0002216	0.0037889**	−0.0007294***
	(0.0002407)	(0.0018547)	(0.0001994)
Observations	9,889,981	665,780	9,224,201
R^2	0.1868505	0.4382679	0.1986143
Degree fixed effects	Yes	Yes	Yes
Home*Destination*Month fixed effects	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes

Notes: All specifications include degree fixed effects, (home * destination * month) fixed effects, and individual fixed effects. Repeat migrants are individuals who have migrated one or more times from h to d prior to a $h - d$ migration observed in month t. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.7: Heterogeneity by Distance (Adjacent districts vs. Non-adjacent districts)

Migration Distance	(1) Any	(2) Short Distance (adjacent districts)	(3) Long-Distance (non-adjacent districts)
Destination friends of friends	−0.0000001	0.0000042**	−0.0000159***
	(0.0000009)	(0.0000017)	(0.0000012)
Home friends of friends	−0.0000035***	−0.0000052***	−0.0000028***
	(0.0000004)	(0.0000008)	(0.0000005)
% Destination support	0.0010618***	0.0010032***	0.0010780***
	(0.0001146)	(0.0002282)	(0.0001362)
% Home support	0.0002216	−0.0004295	0.0002990
	(0.0002407)	(0.0004260)	(0.0002933)
Observations	9,889,981	3,337,184	6,552,797
R^2	0.1868505	0.3237450	0.1972246
Degree fixed effects	Yes	Yes	Yes
Home*Destination*Month F.E.	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes

Notes: All specifications include degree fixed effects, (home * destination * month) fixed effects, and individual fixed effects. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.8: Heterogeneity by Migration Duration (Long-term vs. Short-term)

Migration Distance	(1) Any	(2) Long Stay (> 12 months)	(3) Short Stay (< 6 months)
Destination friends of friends	−0.0000001	0.0000156***	−0.0000125***
	(0.0000009)	(0.0000005)	(0.0000007)
Home friends of friends	−0.0000035***	−0.0000068***	0.0000007**
	(0.0000004)	(0.0000002)	(0.0000003)
% Destination "support"	0.0010618***	0.0002180***	0.0008051***
	(0.0001146)	(0.0000626)	(0.0000846)
% Home "support"	0.0002216	0.0000928	0.0001442
	(0.0002407)	(0.0001323)	(0.0001786)
Observations	9,889,981	9,782,384	9,820,778
R^2	0.1868505	0.1445434	0.1857658
Degree fixed effects	Yes	Yes	Yes
Home*Destination*Month F.E.	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes

Notes: All specifications include degree fixed effects, (home * destination * month) fixed effects, and individual fixed effects. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.9: Heterogeneity by destination type (Rural and Urban)

	(1)	(2)	(3)
Destination Type	All	Rural	Urban
Destination friends of friends	−0.0000001	0.0000022	−0.0000019
	(0.0000009)	(0.0000020)	(0.0000012)
Home friends of friends	−0.0000035***	−0.0000037***	−0.0000018***
	(0.0000004)	(0.0000006)	(0.0000006)
% Destination "Support"	0.0010618***	0.0009579***	0.0008771***
	(0.0001146)	(0.0001470)	(0.0001612)
% Home "Support"	0.0002216	−0.0002734	0.0002481
	(0.0002407)	(0.0003254)	(0.0003042)
Observations	9,889,981	4,236,638	5,918,664
R^2	0.1868505	0.3103749	0.2471896
Degree fixed effects	Yes	Yes	Yes
Home*Destination*Month F.E.	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes

Notes: All specifications include degree fixed effects, (home * destination * month) fixed effects, and individual fixed effects. The three districts that comprise th capital of Kigali are denoted as urban and the remaining districts are denoted as rural (see Table A.10 for an alternative definition of urban and rural locations). Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.10: Heterogeneity by destination type (Rural and Urban), using alternative definition of urban and rural areas

	(1)	(2)	(3)
Destination Type	All	Rural	Urban
Destination friends of friends	−0.0000001	0.0000030	−0.0000024**
	(0.0000009)	(0.0000020)	(0.0000012)
Home friends of friends	−0.0000035***	−0.0000034***	−0.0000017***
	(0.0000004)	(0.0000006)	(0.0000006)
% Destination "Support"	0.0010618***	0.0009944***	0.0009398***
	(0.0001146)	(0.0001472)	(0.0001610)
% Home "Support"	0.0002216	−0.0003122	0.0002904
	(0.0002407)	(0.0003260)	(0.0003043)
Observations	9,889,981	4,230,528	5,924,177
R^2	0.1868505	0.3101766	0.2464579
Degree fixed effects	Yes	Yes	Yes
Home*Destination*Month F.E.	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes

Notes: All specifications include degree fixed effects, (home * destination * month) fixed effects, and individual fixed effects. Urban and rural designation detremined using the sector boundary dataset from the website of National Institute of Statistics Rwanda (see Figure A.10). Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.11: The role of strong ties and weak ties

	(1)	(2)	(3)	(4)
Destination "Weak tie"	0.0036077***	0.0037190***	0.0036771***	0.0037849***
	(0.0000123)	(0.0000250)	(0.0000107)	(0.0000240)
Destination "Strong tie"	0.0044319***	0.0045117***	0.0044074***	0.0045034***
	(0.0000495)	(0.0000536)	(0.0001536)	(0.0001549)
Home "Weak tie"	−0.0003855***	−0.0004813***	−0.0004042***	−0.0005021***
	(0.0000050)	(0.0000108)	(0.0000049)	(0.0000107)
Home "Strong tie"	−0.0007742***	−0.0008799***	−0.0014034***	−0.0015449***
	(0.0000152)	(0.0000179)	(0.0000755)	(0.0000761)
Destination friends of friends		−0.0000062***		−0.0000061***
		(0.0000009)		(0.0000009)
Home friends of friends		0.0000058***		0.0000059***
		(0.0000004)		(0.0000004)
% Destination "Support"		0.0018786***		0.0018158***
		(0.0001138)		(0.0001133)
% Home "Support"		−0.0061352***		−0.0061689***
		(0.0002306)		(0.0002305)
Observations	9,889,981	9,889,981	9,889,981	9,889,981
R^2	0.1858262	0.1859473	0.1857898	0.1859106
Degree fixed effects	No	No	No	No
Home*Destination*Month FE's	Yes	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes	Yes
Definition of "Strong"	90th Percentile	90th Percentile	95th Percentile	95th Percentile

Table A.12: Disaggregating the friend of friend effect by the strength of the 2nd-degree tie

	(1)	(2)	(3)	(4)	(5)	(6)
Destination friends of friends (all)	0.0000004					
	(0.0000009)					
Friends of friends (strong-strong)		0.0000175*				−0.0002288***
		(0.0000104)				(0.0000202)
Friends of friends (strong-weak)			0.0000226***			0.0000696***
			(0.0000024)			(0.0000047)
Friends of friends (weak-strong)				−0.0000460***		−0.0001103***
				(0.0000048)		(0.0000072)
Friends of friends (weak-weak)					0.0000016	0.0000224***
					(0.0000011)	(0.0000017)
Observations	10,089,959	10,089,959	10,089,959	10,089,959	10,089,959	10,089,959
R^2	0.1908962	0.1908965	0.1909039	0.1909041	0.1908964	0.1909380

Notes: Each column indicates a separate regression of a binary variable indicating 1 if an individual i migrated from home district h to destination district d in month t. We show the destination "friend of friend" coefficient separately for geometries of different tie strength. "Strong-strong" (column 2) indicates the effect of friends of friends when the potential migrant i is connected to j via a strong tie, and j is connected to k via a strong tie. "Strong-weak" (column 3) indicates the effect when i and j have a strong tie and j and k have a weak tie. Columns 4 and 5 follow this nomenclature. Strong ties are defined as relationships with 5 or more phone calls (the 90th percentile of tie strength) in a given month. *p<0.1; **p<0.05; ***p<0.01

Table A.13: Disaggregating the network support effect by the strength of supported ties

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Support (all)	0.0013***									
	(0.0001)									
Support (sss)		0.0016**								0.0025***
		(0.0006)								(0.0006)
Support (sws)			0.0069***							0.0076***
			(0.0006)							(0.0006)
Support (ssw)				0.0006***						−0.0000
				(0.0001)						(0.0001)
Support (sww)					0.0027***					0.0030***
					(0.0001)					(0.0001)
Support (wss)						−0.0005				−0.0005
						(0.0003)				(0.0003)
Support (wws)							0.0009***			−0.0025***
							(0.0003)			(0.0003)
Support (wsw)								−0.0019***		−0.0019***
								(0.0003)		(0.0003)
Support (www)									0.0015***	0.0012***
									(0.0003)	(0.0003)
Strong tie	0.0013***	0.0013***	0.0013***	0.0013***	0.0013***	0.0014***	0.0013***	0.0014***	0.0014***	0.0012***
	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00005)
Observations	10,089k	10,089k	10,089k	10,089k	10,089k	10,089k	10,089k	10,089k	10,089k	10,089k
R^2	0.1909	0.1909	0.1909	0.1909	0.1910	0.1909	0.1909	0.1909	0.1909	0.1910

Notes: Each column indicates a separate regression of a binary variable indicating 1 if an individual i migrated from home district h to destination district d in month t. We show the Destination network "support" coefficient separately for geometries of different tie strengths. "SSS' (column 2) indicates the effect of network support for triangles where the potential migrant i is connected to j via a strong tie, j is connected to k via a strong tie, and k and i are connected by a strong tie. "SWS" (column 3) indicates the effect when i and j have a strong tie, j and k have a weak tie, and k and i have a strong tie. Columns 4-8 follow a similar nomenclature. Strong ties are defined as relationships with 5 or more phone calls (the 90th percentile of tie strength) in a given month. *p<0.1; **p<0.05; ***p<0.01

Table A.14: The role of recent migrants

	(1)	(2)	(3)
Destination Degree (network size)	0.0037637***	0.0036358***	0.0036513***
	(0.0000238)	(0.0000244)	(0.0000238)
Home Degree (network size)	−0.0005089***	−0.0005171***	−0.0005859***
	(0.0000107)	(0.0000107)	(0.0000107)
Destination friends of friends	−0.0000059***	−0.0000041***	−0.0000060***
	(0.0000009)	(0.0000009)	(0.0000009)
Home friends of friends	0.0000059***	0.0000060***	0.0000075***
	(0.0000004)	(0.0000004)	(0.0000004)
% Destination "Support"	0.0017164***	0.0017326***	0.0017847***
	(0.0001130)	(0.0001130)	(0.0001129)
% Home "Support"	−0.0061902***	−0.0061607***	−0.0063159***
	(0.0002305)	(0.0002305)	(0.0002304)
Recent migrant friends		0.0011090***	0.0126456***
		(0.0000489)	(0.0001135)
Observations	9,889,981	9,889,981	9,889,981
R^2	0.1858886	0.1859340	0.1869832
Degree fixed effects	No	No	No
Home*Destination*Month fixed effects	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes
Definition of "Recent"	NA	Ever	Last month

Notes: Each column indicates a separate regression of a binary variable indicating 1 if an individual i migrated from home district h to destination district d in month t. Column (1) replicates the original result from Table A.2; column (2) controls for the number of migrants that i knows, who ever migrated from h to d prior to t; column (3) controls for the number of recent migrants that i knows, who migrated from h to d in the month prior to t. Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Table A.15: *Predicted* migration (from structural model) and social network structure

	(1)	(2)	(3)	(4)
Panel A: Destination network characteristics				
Degree (network size)	0.0680931***			
	(0.0000450)			
% Friends with common		0.1728557***		0.1707765***
support		(0.0004015)		(0.0004002)
Unique friends of friends			−0.0007402***	−0.0007033***
			(0.0000035)	(0.0000034)
Observations	6,386,523	6,386,523	6,386,523	6,386,523
R^2	0.5967755	0.6359449	0.6271628	0.6386054
Panel B: Home network characteristics				
Degree (network size)	−0.0114922***			
	(0.0000197)			
% Friends with common		−0.1836519***		−0.1846382***
support		(0.0010150)		(0.0010159)
Unique friends of friends			−0.0000240***	−0.0000364***
			(0.0000016)	(0.0000016)
Observations	6,386,523	6,386,523	6,386,523	6,386,523
R^2	0.4676148	0.4948318	0.4919757	0.4948771
Degree fixed effects	No	Yes	Yes	Yes
Individual fixed effects	Yes	Yes	Yes	Yes
Home*Destination*Month F.E.	Yes	Yes	Yes	Yes

Notes: Each column indicates a separate regression of a binary variable $\widehat{M_{ihdt}}$ that takes the value 1 if an individual i was *predicted* to migrate from home district h to destination district d in month t (where this prediction is based on the calibrated structural model, and determined using the actual network properties of i). Standard errors are two-way clustered by individual and by home-destination-month. *p<0.1; **p<0.05; ***p<0.01.

Appendix B

Chapter 3 Additional Materials

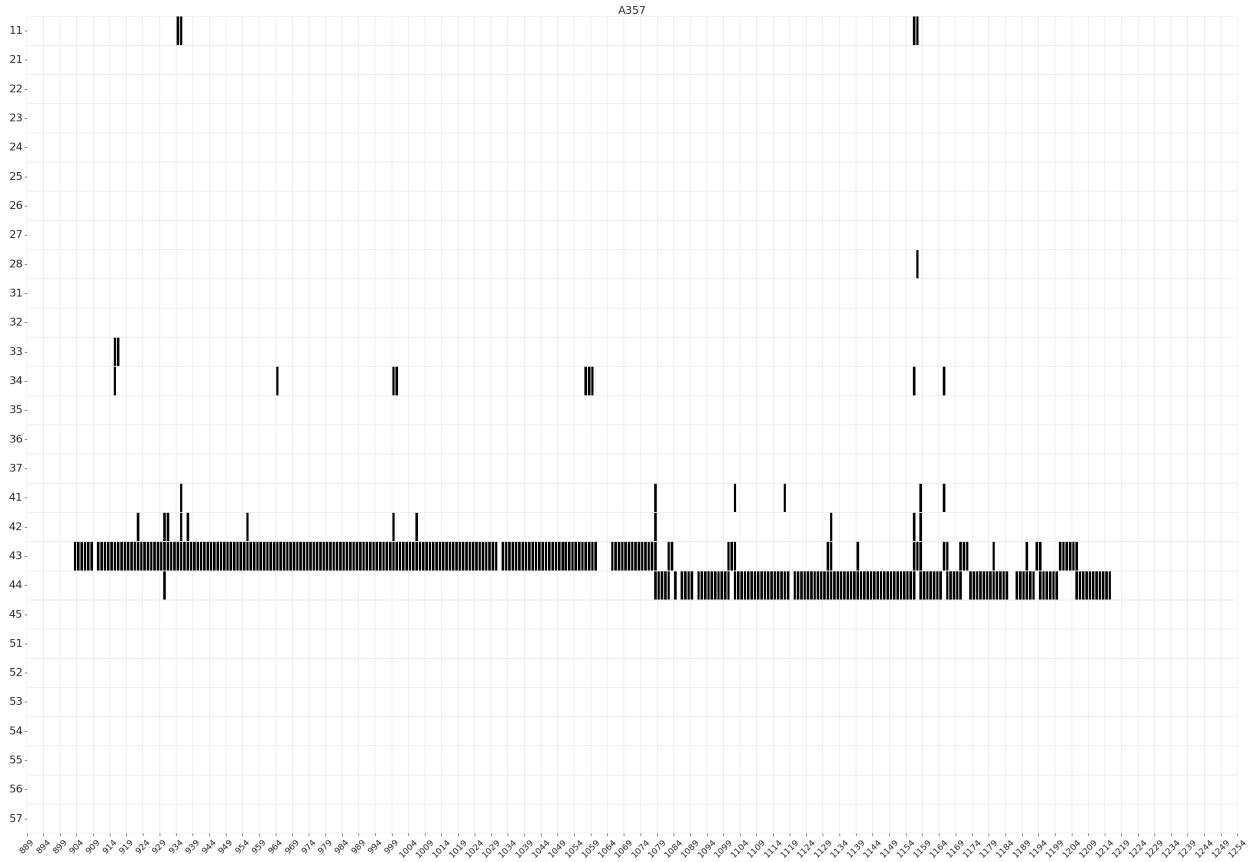


Figure B.1: An example of labeling tasks. *Notes*: Each row is one district in Rwanda. Each column is one day. Labelers are required to answer several questions. For example, whether a migration took place and how confident they are in that assessment on a scale of 1 to 3.

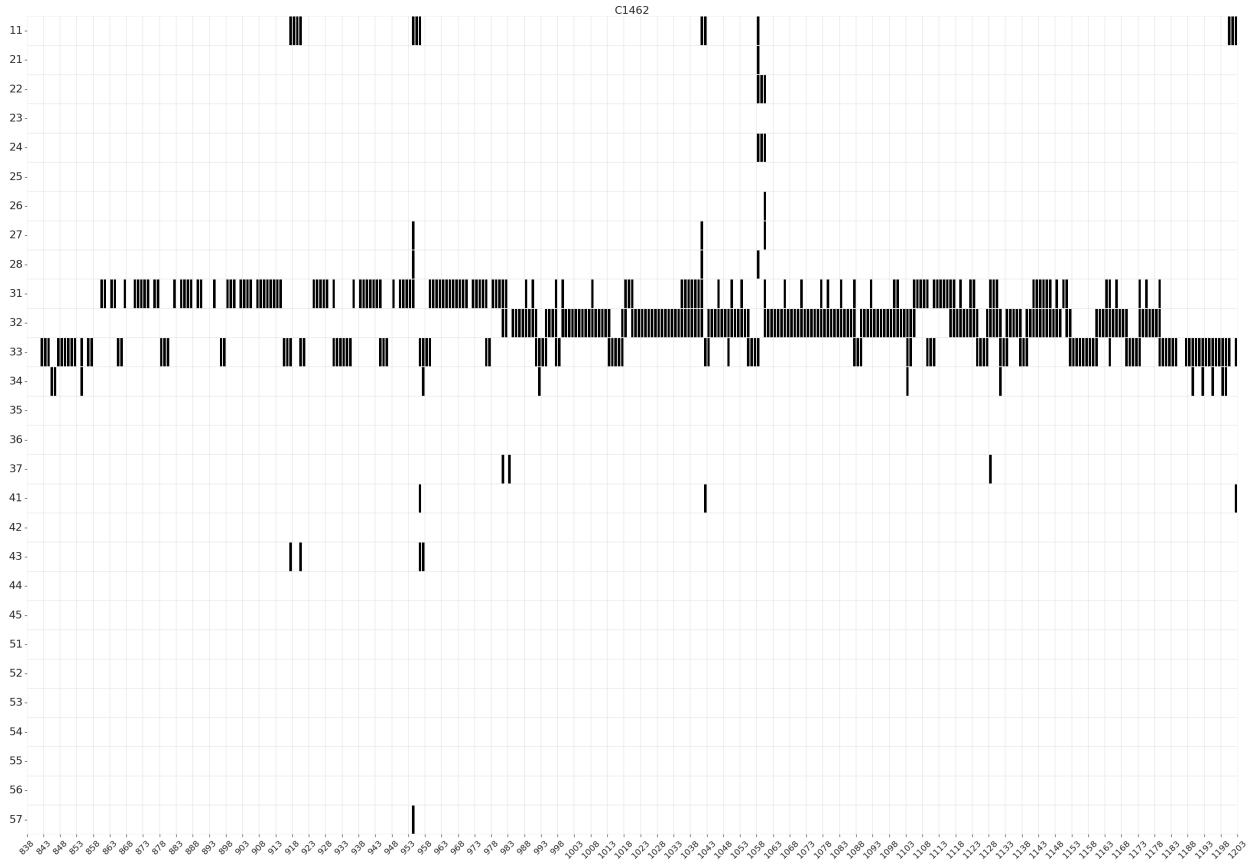


Figure B.2: An ambiguous example. *Notes*: Labelers have different opinions on whether a migration took place in this sample.

Data: $< userID, timestamp, location >$ tuples for each location record. *Note:* For each individual i, his or her location history can be coded into a matrix M_{ld}, where l is the location among all the locations L; d is the date. l can be a city, or a country, which is determined by the definition of location. $M_{ld} = 1$ if this person appears in the location l in the day d. Otherwise, $M_{ld} = 0$.
Result: $< Segments >$ tuples of users' segments

```
for i ∈ U do
    for l ∈ L do
        for d ∈ D do
            if M_ld == 1 & M_l(d+ϵ) == 1 then
                M_l(d+i) ← 1 where i in range(ϵ)
            end
        end
        for d ∈ D do
            if M_ld == 1 for d in range(k) and k ≥ minDays and
             ∑ M^raw_l(d) ≥ k * propDays then
                // save this segment with start date and end date
                Segmt[i][l]+ = [d, d + k]
            end
        end
        for l ∈ L do
            for s ∈ Segmt[i][l] do
                if no other segments exist within (s[t − 1][1], s[t][0]) in other locations
                 then
                    Merge s[t − 1] and s[t]
                end
            end
        end
        for l ∈ L do
            for s ∈ Segmt[i][l] do
                if overlap exist: (s[t − 1][1] > s[t][0]) then
                    Swap s[t − 1][1] and s[t][0]
                end
            end
        end
    end
end
```

Algorithm 3: Detecting location segments

Table B.1: Papers that use trace data to measure migration

Paper	Data	Method to identify home	Method to identify migrants
Phithakkitnukoon et al. (2011)	CDR	Cell tower where the user has the most call activities (10pm to 7pm) each month	Whose home location changed only once with migration distance of more than 50km
J. E. Blumenstock (2012)	CDR	Monthly center of gravity (COG) whose distance to COG of last month is smaller than a proportion of average radius of gyration (ROG) over a certain number of months.	Whose migration distance is greater than a proportion of ROG
Lu et al. (2016)	CDR	Monthly modal location where the user has the most call activities	Whose home location changed after disasters
J. Blumenstock et al. (2019)	CDR	Same monthly modal location in two/three consecutive months, which is calculated based on daily modal location and hourly modal location. Modal location is where the user has the most call activities.	Whose home location changed
Hankaew et al. (2019)	CDR	Same to Phithakkitnukoon et al. (2011)	Whose home location changed only once (One home location in two months and a new home location in another two months)
Yang et al. (2018)	CDR, plus national ID of each user	Birthplace is extracted from national ID; Living city is the location where the phone number was obtained.	Whose birthplace is different from the living city
Büchel et al. (2019)	CDR, plus billing address	Postcode of the billing address	Whose home location changed
Zagheni et al. (2014)	Geo-tagged tweets	Modal country of the user over four months (the modal country should has three times more tweets than the second most frequent country)	Whose home location changed
Fiorio et al. (2017)	Geo-tagged tweets	Modal tweet location (US county) of the user over a specific duration (duration is a threshold)	Whose home location changed over an interval (interval is another threshold)