Kinship networks and emigration: A case study of 19th- and early 20th-century Guangdong, China

Tianning Zhu

January 22, 2025

1 Introduction

Historians have long used chain migration to explain emigration flow that was self-reinforcing due to migrant network effects (e.g., Baines 1994). A substantial body of empirical evidence from European transatlantic mass migration in the nineteenth century supports this idea, showing that networks could influence (1) potential migrants' decision to move (e.g., Connor 2019) and (2) their choices of where to move (e.g., Pérez 2021). While this recent literature integrates migrant surname information into traditional measures of network strength, such as the total number or share of emigrants, to capture the readily observable family connections between past and present migrants, the data sources do not allow for a nuanced measure for such family or kin-based connections. This limits the explanatory power of these network measures, as they still reflect persisting local conditions that led to previous migration and cannot explain why not everyone with the same surname moved.

Descending genealogies are one source of data that offers a unique opportunity to measure kinship ties with greater granularity and to improve existing network measures. They keep track of all the descendants of a known ancestor and provide detailed information on the extended kinship networks that are not observable in other sources. Chinese have recorded their family history with this type of genealogy and shared a similar mass migration experience during the 19th- and early 20th centuries. Around 20 million Chinese are estimated to have migrated overseas from 1840 to 1940, with 90 per cent of them moving to Southeast Asia. The origins of these emigrants

barely changed during this period. Less than 4 per cent of them moved under long-term labour contracts with Europeans as coolie, and many were likely to migrate indentured to other Chinese or collectively operated mines or farms under some form of debt or labour obligation, the same as European free migrants (McKeown 2010). Though it has received relatively limited attention in the literature compared to the European mass migration, with its high-quality genealogical data available, historical Chinese mass migration provides a unique setting to study migrant network effects through a comparative lens.

This paper uses Chinese genealogical data to identify and quantify kinship networks more accurately than the existing literature can do and study the chain migration inside a kinship group that existing research has not considered. Specifically, it tests whether kinship networks affect people's propensity to move and migrants' destination choices. Historical Chinese are argued to have relied predominantly on kin-based organization for their social organization (Greif and Tabellini 2017). The kinship networks, revealed in Chinese genealogies, thus should have constituted a critical part of migrant networks in the context of Chinese migration during the 19th and early 20th centuries.

2 Data

Chinese genealogies were compiled by members of lineages, following all male descendants of an ancestor and providing demographic and socioeconomic information, including their marriages, of these descendants. The information available for each male descendant varies in coverage and completeness. Women are recorded as wives or daughters of these men and are underreported. The kinship networks that are traceable in Chinese genealogies are thus patrilineal.

A comprehensive entry would include the member's

1. name (and Zi/Hao, courtesy name; sometimes Shi/Hui, name taboos);
2. vital information, namely birth date, death date and age at death;
3. wife's information, which may include her place of birth, her father's name, her birth order and her vital information ;

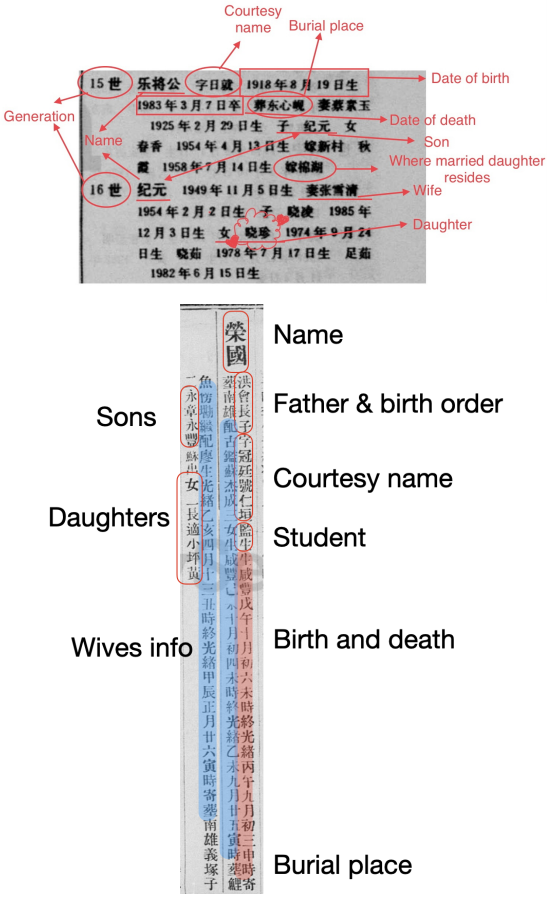


Figure 1: Entries from two Chinese genealogies

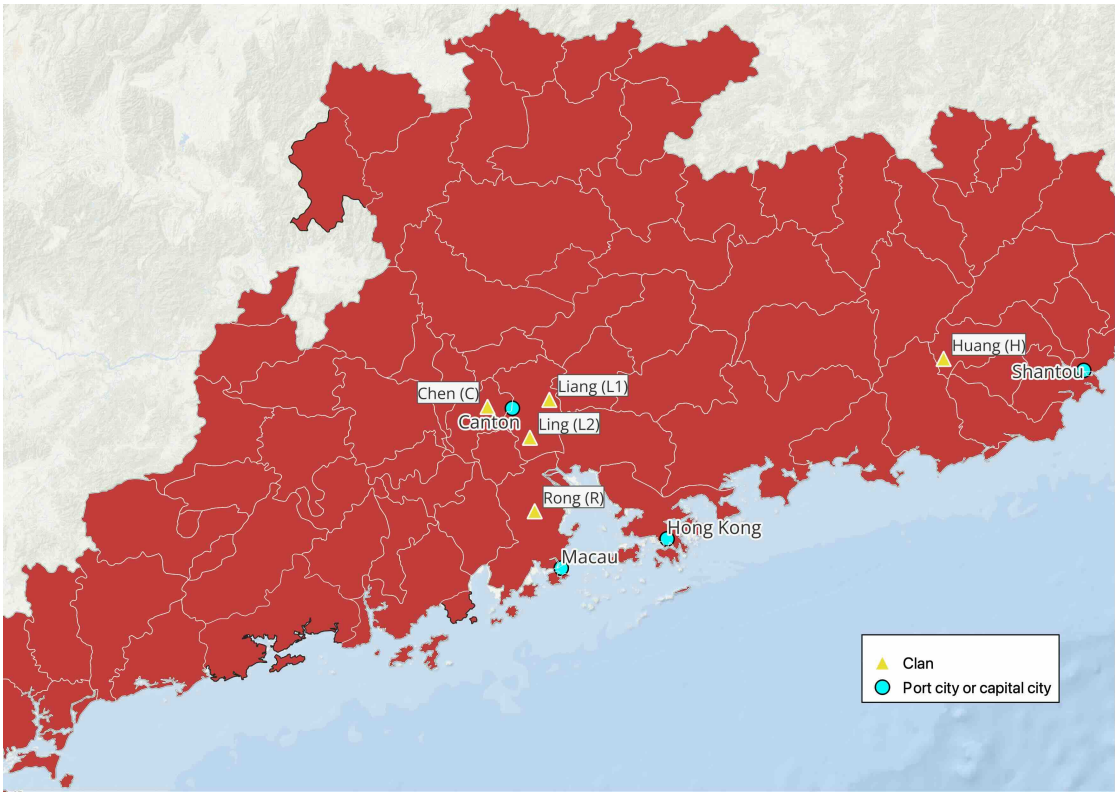
4. son's information, including the number of sons and their birth orders (which might not be explicitly stated but can be inferred as the names are listed in birth order);

5. daughter's information, including the number of daughters and each daughter's marriage (her in-laws' residence and her husband's information);

6. occupation (as a government official or military official) or achievement in the civil service examination system.

Figure 1 shows snippets from two different genealogies with information for their male members. As privately compiled documents, the format of these different genealogies and their entries are different, but the contents are the same, albeit with varying levels of detail. Chinese genealogies are argued to have been written for cultural reasons, such as ancestral worship, and also practical

Figure 2: Clan location



reasons, such as specifying lineage membership for entitlement to lineage resources (see Shiue 2016). The accuracy of the information recorded would, therefore, be important to the compilers and data collected from Chinese genealogies are broadly credible, although not necessarily without errors.

This paper uses 15,653 male members' information collected from five genealogies that were compiled by five lineages living in Guangdong province, a province in China that saw many residents emigrating internationally in the 19th and early 20th centuries. These males were from the most recent 10 to 12 generations recorded in these genealogies and were mostly born between the 17th and early 20th centuries. Figure 2 shows where the compilers of these genealogies were from. Lineage of Huang (H) lived in the eastern part of the province while the other four lineages (Chen (C), Liang (L_1), Ling (L_2) and Rong (R)) lived in the central part and close to the provincial capital (Canton). H was close to the port city of Shantou while the other four lineages were close to the port city of Hong Kong and Macau.

The migration status can be identified in these records. In total, I have found 1,236 migrants. The majority of the migrants are identified if they were reported as (1) moving to or living in a named location, (2) moving away without a known destination, or (3) having died on a trip or been buried in a named location that is outside the local county. Occasionally, some migrants were found as (1) returned migrants or were born elsewhere, (2) having moved away with their remarried mothers or fathers or (3) raised by relatives living elsewhere or adopted by people with different surnames. Table 2.1 shows the breakdown of the migrants across the genealogies. The distribution of the migrants is uneven although the lineages were from the same emigrant-sending province. This is in accord with chain migration literature.

Given that genealogies record life events, most of the migration found in Chinese genealogies is permanent. The migration timing is generally not reported but can be proxied with birth year-based functions. I impute the missing birth year from the known birth year of his father or son based on the average father-son age gap by birth order in the sample, which I can obtain from linear regression. This imputation process is iterated to fill in all missing birth year information of males on the same line of descent that has at least one known birth year.

Table 2.1: Movement and Identification by Clan

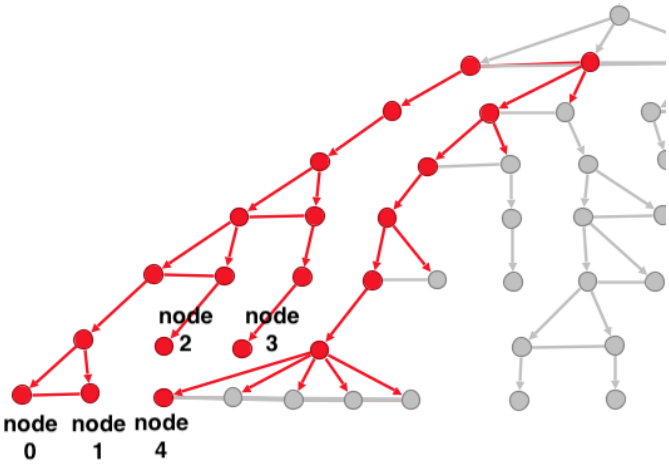
Clan	Moved (location known)	Moved (location unknown)	Death/Burial	Family related reasons	Return	**Total**	**Total obs.**
Chen	3	10	6	0	0	19	1,376
Huang	559	0	1	0	6	566	2,076
Liang	13	15	1	0	0	29	1,382
Ling	174	73	113	33	1	394	4,888
Rong	70	8	143	7	0	228	5,931
Total	819	106	264	40	7	**1,236**	**15,653**

3 Method

To quantify the strength of the kinship network, I first reconstruct the family trees from collected genealogical records by linking the father-son and brother pairs. These reconstructed trees are branches from the original pedigree that consists of every male member in the genealogy. I have reconstructed 682 branches for these 15,653 observations. While an average branch has 23 people and lasts for 4 generations, there are considerable variations in size and duration among these branches.

Then I construct a measure called connectedness to measure how well one is connected with the migrants in his patrilineal kinship network (i.e., the reconstructed branch) by counting the number of ties he has with previous migrants. I assign each tie a weight to distinguish the strong and weak kinship ties. The closeness between two nodes on a graph can be visualised with the shortest path between them. The length of the shortest path measures the minimum steps from one node to the other. When the graph represents a family tree, the longer the shortest path is, the more distant the kinship relation is. I transform the shortest path with the negative exponential function $y = e^{-x}$ to capture this inverse relation and use the transformed shortest path as weight.

Figure 3: Example kinship network



Note: The arrowed edges represent father-son links. The non-arrowed edges represent brother-brother links.

Table 3.1 is an example of measuring the connectedness of node 0 with the other four randomly picked nodes in the network depicted in figure 3. I first calculate the length of the shortest path from node 1-4 to node 0 in column (1). Then I calculate the connection strength from the length with the exponential function for these four nodes in column (2). Despite being from the same family tree, the connections between node 0 and the other four different nodes are hardly the same. Finally, I sum up these calculated connection strengths to have the connectedness score.

As mentioned in the previous section, migration timing can be proxied with birth year-based functions. Different functions can be used for this purpose, allowing flexibility in dealing with

Table 3.1: Shortest path and connection strength

	(1) Shortest path to node 0	(2) Connection strength
node 1	1	e^{-1}
node 2	4	e^{-4}
node 3	6	e^{-6}
node 4	13	e^{-13}
Connectedness =		0.389 $(=e^{-1}+e^{-4}+e^{-6}+e^{-13})$

the migration sequence of migrants from the same birth cohort. For simplicity, my baseline connectedness is calculated using birth year for sequencing migration, assuming that the permanent migration decision is made at the same age for every migrant. For robustness check, I also construct two connectedness measures with stricter assumptions that (1) previous migrants should be at least 25 years older or (2) previous migrants should be in a previous 25-year birth cohort, which allows for contemporaneous migration of brothers or cousins close in age.

To relax the age-invariant assumption about migration timing, I also use a random number drawn from a uniform distribution with an interval of 20 to 50 to simulate each individual's migration age. I then add this age to their birth year to calculate the year when they decided to move away permanently. Namely,

$$Migration\ year = Birth\ year + Migration\ age$$

$$Migration\ age \sim U(20, 50)$$

Finally, to account for the mechanical difference in the connectedness score which arises from different sizes of the networks, I also normalize the baseline connectedness. To do so, I divide the baseline connectedness by the total number of all previous connections which are also weighted in the same way.

Table 3.2 shows the correlation table of 5 different connectedness measures. They are highly correlated with each other, suggesting that (1) contemporaneous migration was not a major contributor to the baseline network measures and (2) the sizes of the networks matter in baseline network measures but only to a limited extent.

Connectedness	Baseline	25 years	25-year cohort	Randomized	Normalized
Baseline	1.00				
25 years	0.78	1.00			
25-year cohort	0.89	0.85	1.00		
Randomized	0.91	0.76	0.84	1.00	
Normalized	0.89	0.81	0.84	0.83	1.00

Table 3.2: correlation

4 Results

Using this connectedness measure, I examine the effect of having migrant networks on people's propensity to move. I use logistic regression for the following model

$$Migration_i = \beta_0 + \beta_1 Connectedness_i + \gamma\prime Z_i + \delta_{B_i} + \alpha_{C_i} + \epsilon_i$$

where the coefficient β_1 measures the effect of having migrant networks on people's migration propensity. $Migration_i$ is a binary outcome variable with 1 meaning person i is a migrant and 0 meaning otherwise. Matrix Z_i represents individual-level control variables, including 1) whether one's courtesy name (Zi/Hao) is reported in the records, (2) whether one was an office or a degree holder, (3) one's birth year estimates, (4) the number of brothers, measured by the number of sons one's father had, (5) one's birth order, (6) whether one was adopted and (7) whether one was the only son. δ_{B_i} represents the branch controls and α_{C_i} represents the clan (genealogy) controls.

Table 4.1: Migration Propensity

$Log(odds)$	(1) Migration	(2) Migration	(3) Migration	(4) Migration
Connectedness (baseline)	48.69***	19.11***	34.81***	79.96***
	(24.70)	(18.17)	(23.49)	(3.76)
Individual-level controls	Y	Y	Y	Y
Clan controls	Y	Y	Y	Y (interacted)
Branch controls	N	Y	N	N
Observations	14935	11595	11595	14935
Pseudo R^2	0.234	0.224	0.192	0.236

Return migrants are excluded in the regressions
Exponentiated coefficients; t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Column (1) of table 4.1 shows the result from baseline regression without branch controls.

The coefficient suggests that connectedness with previous migrants is positively related to the odds of one also being a migrant. We can use the shortest path to interpret the odds ratio. Other things being equal, compared to people who had no ties to a previous migrant, the odds ratio of one also being a migrant if he only had a migrant father or a migrant brother (1 step away from the ego) is 4.18 times higher.

Namely,

$$OR = \frac{odds(Connectedness = 1)}{odds(Connectedness = 0)} = e^{\beta_1} = 48.69 \tag{1}$$

$$\beta_1 = 3.89$$

when

$$Shortestpath = 1$$

$$Connectedness = e^{-shortestpath} = e^{-1}$$

thus,

$$OR = \frac{odds(Connectedness = e^{-1})}{odds(Connectedness = 0)} = e^{\beta_1 e^{-1}} = 4.18 \tag{2}$$

As shown in table 4.2, after 6 steps away, the kinship ties would have little influence on one's odds of being a migrant individually, as the odds ratio hardly differs from 1, unless considered aggregatedly. The migration in the immediate family (father and brother) seems to improve one's odds of moving by 4 times, while the migration in the extended family (grandfather and uncle) has a much smaller impact of 1.7 times. The migration of more distant relatives can make a difference, but the scale is way much smaller.

Table 4.2: Different ties and their influence

kinship ties	father	grandfather	great grandfather	distant relatives			
	brother	uncle	father's uncle				
shortest path	1	2	3	4	5	6	7
odds ratio	4.18	1.69	1.21	1.07	1.03	1.01	1.00

Column (2) of table 4.1 shows the result from baseline regression with branch controls. The coefficient remains significant with a reduced scale, suggesting that people on different parts of

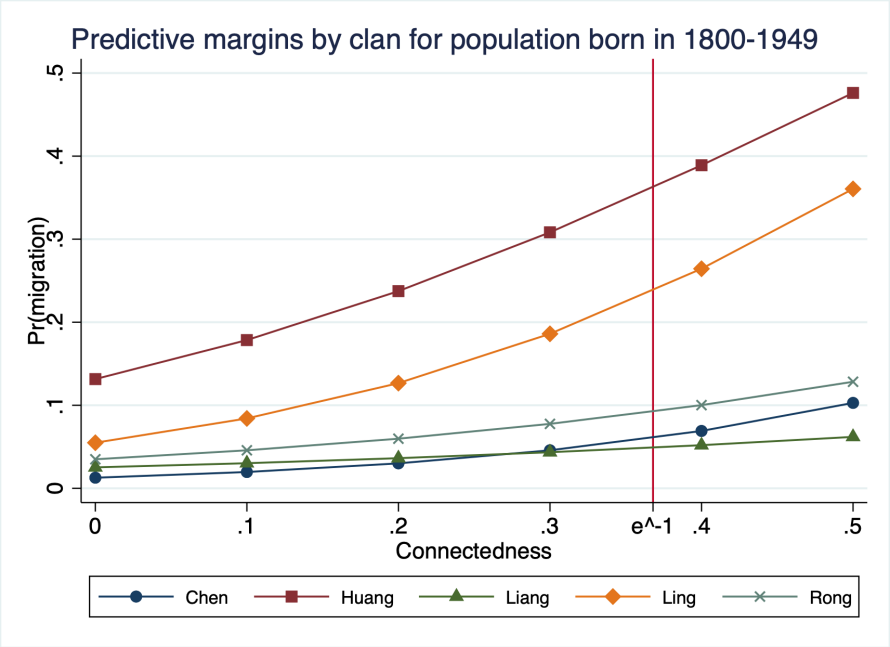


Figure 4

the same family tree would have a different propensity to move. Logistic regression would drop the observations from branches in which its members either all moved or stayed. To check for the robustness of the results, column (3) of table 4.1 shows the result from the baseline regression without branch controls with the same sample of observation as the regression of column (2) uses. The coefficient is still significant but the scale is increased, suggesting that members from different branches would have different baseline migration propensity.

Column (4) shows the result of interacting the clan controls with the connectedness measure while dropping the branch controls in the baseline regression. The independent effect of the network is both large and significant. I plot the marginal effects of clan dummies on people's migration propensity in figure 4. Members from different clans seem to respond to the networks differently. The scale of the odds ratio is probably driven by the small probability of moving when people have little connection.

5 Migrants' destination

Figure 5 shows the destination choices of migrants from these five genealogies. I break down the connectedness measure by these destination choices and test (1) whether the increase in migration propensity was related to the information and (2) whether the destination of previous migrants would affect later migrants' destination choices.

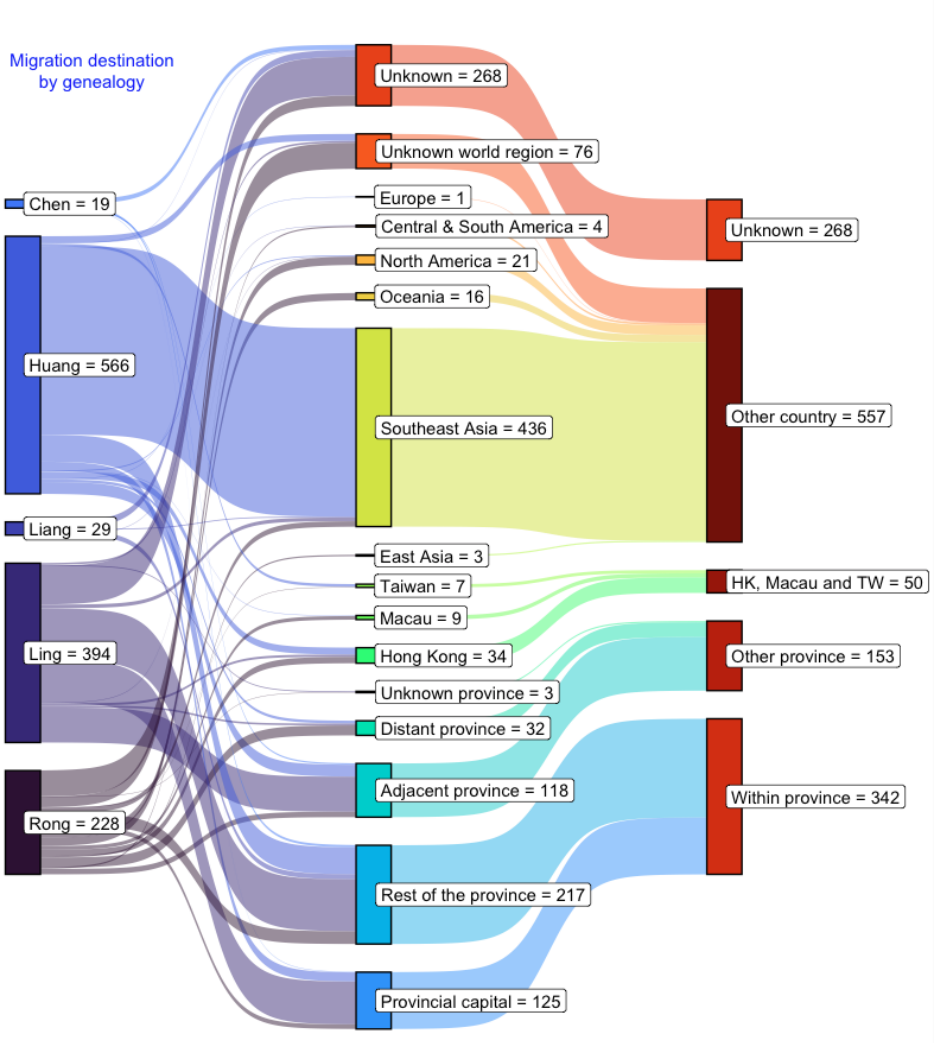


Figure 5: Migrants' destination by clan

5.1 Information

The literature suggests that the spread of information led to chain migration (e.g., Baines 1994). To test this idea, I decompose the connectedness according to whether the migrants' destination information is recorded in the genealogy. Some migrants are not reported with their destinations, suggesting their probable alienation from their lineages. These migrants comprise a less informative part of one's migrant network than the part consisting of migrants with known destinations.

Then I use the decomposed network measure in the baseline regression. As is shown in table 5.1, the uninformative part of the network has little impact on people's migration propensity, which supports the information argument for chain migration.

Table 5.1: Information channel

$Log(odds)$	(1) Migration	(2) Migration	(3) Migration
Connectedness (informative)	19.90***	19.75***	
	(17.97)	(17.92)	
Connectedness (uninformative)	6.512*		4.508
	(2.03)		(1.61)
Individual-level controls	Y	Y	Y
Clan controls	Y	Y	Y
Branch controls	Y	Y	Y
Observations	11595	11595	11595
Pseudo R^2	0.225	0.224	0.176

Return migrants are excluded in the regressions
Exponentiated coefficients; t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

5.2 Malaysian network

As is shown in figure 5, there is a substantial flow to Southeast Asia from clan Huang. A majority of the Southeast Asian migrants from this clan moved to Malaysia. I construct a Malaysian connectedness for migrants from this clan and use logistic regression to examine their destination choices. Table 5.2 shows that Malaysian networks indeed affect migrants' decision to go to Malaysia positively.

Table 5.2: Malaysian network

$Log(odds)$	(1) Malaysia
Connectedness (Malaysian)	18.09*** (6.94)
Individual-level controls	Y
Branch controls	Y
Clan	Huang
Observations	528
Pseudo R^2	0.140

Return migrants are excluded in the regressions
Exponentiated coefficients; t statistics in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

6 Conclusion

This paper shows that migrant networks affect people's migration propensity positively in the historical Chinese mass migration setting. The impact mostly comes from the informative networks. It also finds that destination networks would attract more migrants to the same location, causing a reinforcing migration flow to a certain destination.

This paper thus contributes to the existing literature with empirical evidence on migrant network effects in historical China, which also adds to the debate around the distinctness and "Chineseness" of historical Chinese mass migration. Additionally, the method developed in this paper for analysing migrant networks with genealogical data can be adapted for European mass migration, which enables further comparative studies for better generalizations of mass migration patterns.

7 References

Baines, Dudley. 1994. 'European Emigration, 1815-1930: Looking at the Emigration Decision Again'. *The Economic History Review* 47 (3): 525–44.

Connor, Dylan Shane. 2019. 'The Cream of the Crop? Geography, Networks, and Irish Migrant Selection in the Age of Mass Migration'. *The Journal of Economic History* 79 (1): 139–75.

Greif, Avner, and Guido Tabellini. 2017. "The Clan and the Corporation: Sustaining Cooperation in China and Europe." *Journal of Comparative Economics* 45 (1): 1–35.

McKeown, Adam. 2010. 'Chinese Emigration in Global Context, 1850–1940'. *Journal of Global History* 5 (1): 95–124.

Pérez, Santiago. 2021. 'Southern (American) Hospitality: Italians in Argentina and the United States During the Age of Mass Migration'. *The Economic Journal* 131 (638): 2613–28.

Shiue, Carol H. "A culture of kinship: Chinese genealogies as a source for research in demographic economics." *Journal of Demographic Economics* 82, no. 4 (2016): 459-482.