Scilight

Journal of Social Physics

https://www.sciltp.com/journals/jsp

Supplementary Material

In a Society of Strangers, Kin Is Still Key: Identified Family Relations in Large-Scale Mobile Phone Data

Tamás Dávid-Barrett [1,2,3], Sebastian Diaz [2,4], Carlos Rodriguez-Sickert [2], Isabel Behncke [2], Anna Rotkirch [3], Loreto Bravo [4] and János Kertész [5,*]

[1] Trinity College, University of Oxford, Oxford OX1 3BH, UK
[2] CICS, Facultad de Gobierno, Universidad del Desarrollo, Av. Plaza 680, San Carlos de Apoquindo, Las Condes, Santiago de Chile 7610658, Chile
[3] Population Research Institute, Väestöliitto, Kalevankatu 16, 00101 Helsinki, Finland
[4] Data Science Institute, Universidad de Desarrollo, Av. Plaza 680, Las Condes, Santiago de Chile 7610658, Chile
[5] Department of Network and Data Science, Central European University, Quellenstrasse 59, 1100 Vienna, Austria
* Correspondence: janos.kertesz@gmail.com

How To Cite: Dávid-Barrett, T.; Diaz, S.; Rodriguez-Sickert, C.; et al. In a Society of Strangers, Kin Is Still Key: Identified Family Relations in Large-Scale Mobile Phone Data. *Journal of Social Physics* **2026**, *1*(1), 5.

S1. Demographic Data from Chile, 2015

Table S1 summarizes the demographic structure data from Chile in the year of 2015 [62].

Table S1. Population structure in Chile, in 2015.

Estimated Population as of June 30, Based on Demographic Projection			
Age Group (Years)	**Total**	**Men**	**Women**
TOTAL	18,006,407	8,911,940	9,094,467
0–4	1,236,555	629,883	606,672
5–9	1,222,682	623,590	599,092
10–14	1,207,255	615,595	591,660
15–19	1,323,480	676,381	647,099
20–24	1,460,830	743,660	717,170
25–29	1,498,935	757,921	741,014
30–34	1,357,954	683,722	674,232
35–39	1,245,192	623,740	621,452
40–44	1,243,826	619,735	624,091
45–49	1,258,457	623,852	634,605
50–54	1,220,976	601,533	619,443
55–59	1,050,355	513,547	536,808
60–64	822,496	396,985	425,511
65–69	642,018	301,766	340,252
70–74	487,665	220,125	267,540
75–79	342,975	145,320	197,655
80 and more	384,756	134,585	250,171

The mean age at first birth of women in 2015 was 24.72 with a standard deviation of 6.09 [63]. For completeness, we give the similar data from the years 2008–2015 (no earlier data were found).

Table S2. Age of women at first birth in Chile, 2008–2015.

Year	Mean Age at First Birth	Standard Deviation
2008	23.12	5.81
2009	23.17	5.82
2010	23.38	5.9
2011	23.52	5.94
2012	23.72	5.98
2013	24.04	6.06
2014	24.33	6.07
2015	24.72	6.09

The fertility rate in the period 2008–2015 is summarized in Table S3. There is a clear downward trend reaching by 2023 the value of 1.17, according to https://ourworldindata.org/fertility-rate (accessed on 27 May 2026)

Table S3. Fertility rate in Chile 2008–2015.

Year	Fertility Rate
2008	1.963
2009	1.988
2010	1.956
2011	1.924
2012	1.875
2013	1.858
2014	1.919
2015	1.859

S2. Comparison with Previous Methodology

Applying the surname filters to the data only partly confirmed the validity of identifying relationship types between people based on age, sex, and frequency of phone calls introduced by David-Barrett et al. [21].

Figure S1. shows the percentage of kin identification, based on age, sex, and frequency of phone calls introduced by David-Barrett et al. [21], which ranges from about 5% to 35%. This was expected, and a logical implication of working with censored data, given that the company's market share is only about 30%.

Children: only a minority of children identified using the frequency–age criterion were confirmed as the children of the egos based on surname identification. For female egos (i.e., assumed mothers of a child) 5–35% of assumed children were identified as filtered sons or daughters. For male egos (i.e., assumed fathers) the ratio is lower, the ratio is lower—especially for younger children—ranging between 5% and 25%.

Parents: the majority of alters that were identified using the frequency-age criteria as parents were not confirmed as the parents based on surname data. Mothers were confirmed in 3–30% of cases, while fathers were confirmed 1–20% of the time.

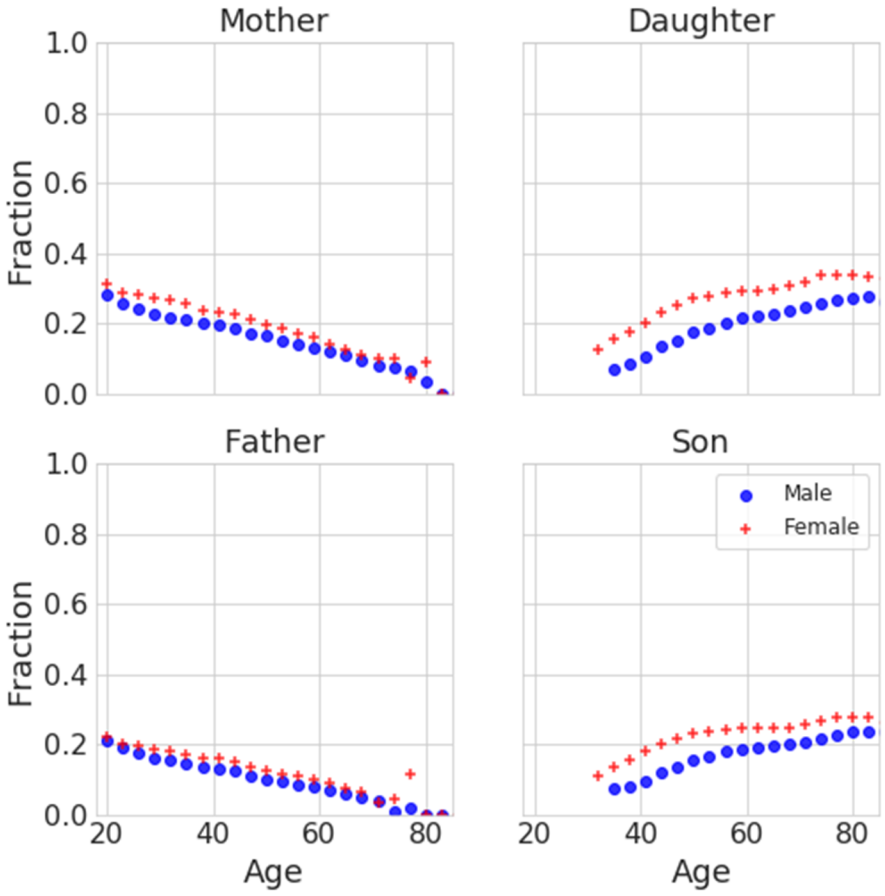

Figure S1. Ratios of relationship types filtered using the last-names technique from the frequency-age-sex method.

S3. Statistical Analysis

Table S4 contains the results of the statistical analysis.

Table S4. Life-course-dependent differences in distribution, mean, and median of kin and non-kin, for every category. (Values smaller than 10^{-6} are set to 0). For this comparison to be valid, every non-kin group was downsampled considering the size of every Age-Sex-Relationship cohort. For example, If 23 year old mothers are 5230 and non-mothers 21,000, Non mothers were randomly sampled to 5230. This procedure was done, because when age increases, kin detection decreases and non-kin group increases by default and this produces a non-balanced sample.

Variable	Alter Type	Ego Sex	Kol-Mogo-Rov-Smir-Nov	*p*-Val.	Kin (Mean)	Non-Kin (Mean)	Kin (Median)	Non-Kin (Median)	*t*-Test	Wilcoxon-Mann-Whitney
Frequency	Mother	female	0.37	0	275.4	76.1	102	14	0	0
		male	0.35	0	199.5	80.8	81	14	0	0
	Father	female	0.34	0	190.9	80.2	72	12	0	0
		male	0.32	0	183.3	73.2	71	14	0	0
	Daughter	female	0.37	0	321.5	95.6	138	20	0	0
		male	0.33	0	226.8	106.6	98	19	0	0
	Son	female	0.33	0	229.7	101.8	103	18	0	0
		male	0.31	0	217.2	96.1	96	22	0	0
Fraction of time	Mother	female	0.38	0	0.13	0.05	0.07	0.01	0	0
		male	0.35	0	0.11	0.05	0.05	0.01	0	0
	Father	female	0.33	0	0.09	0.04	0.04	0.01	0	0
		male	0.33	0	0.09	0.04	0.04	0.01	0	0
	Daughter	female	0.38	0	0.14	0.05	0.08	0.01	0	0
		male	0.33	0	0.10	0.05	0.05	0.01	0	0
	Son	female	0.33	0	0.10	0.05	0.05	0.01	0	0
		male	0.32	0	0.09	0.05	0.05	0.01	0	0
Out-Call Fraction	Mother	female	0.15	0	0.49	0.46	0.49	0.43	0	0
		male	0.14	0	0.49	0.48	0.50	0.48	0	0
	Father	female	0.15	0	0.44	0.45	0.42	0.40	4×10^{-6}	0
		male	0.12	0	0.45	0.46	0.43	0.45	0	6.9×10^{-5}
	Daughter	female	0.12	0	0.50	0.52	0.50	0.52	0	0
		male	0.12	0	0.56	0.54	0.57	0.57	0	0.051
	Son	female	0.11	0	0.50	0.50	0.50	0.50	0.116	0.134
		male	0.10	0	0.55	0.52	0.56	0.51	0	0
Call length	Mother	female	0.10	0	105.7	95.3	69.0	58.1	0	0
		male	0.08	0	94.2	88.1	61.1	54.0	0	0
	Father	female	0.08	0	90.5	86.6	60.5	53.5	0	0
		male	0.07	0	86.1	80.0	57.2	51.3	0	0
	Daughter	female	0.08	0	106.0	101.7	70.6	63.3	0	0
		male	0.06	0	89.2	91.8	60.8	58.3	0	0
	Son	female	0.06	0	95.3	94.3	62.6	58.7	0.0049	0
		male	0.05	0	85.8	85.0	58.0	55.6	0.0290	0

Table S5. Standard deviation of means across time, and total standard deviation of groups. (Values smaller than 10^{-6} are set to 0).

Variable	Alter Type	Ego Sex	Standard Deviation of Means Across Time		*p*-Value	Total Standard Deviation	
			Kin	Non-Kin		Kin	Non-Kin
Frequency	Mother	female	69.15	23.21	0	437.5	243.9
		male	68.40	23.67	1.25×10^{-5}	323.4	263.4
	Father	female	53.74	21.55	0	328.2	287.6
		male	54.22	22.68	0	319.7	227.7
	Daughter	female	43.98	17.49	0	475.4	273.3
		male	30.97	22.82	2.41×10^{-2}	359	328.2
	Son	female	25.89	18.11	1.06×10^{-2}	350.1	294.1
		male	30.65	18.97	1.07×10^{-3}	350.7	264.7
Fraction of time	Mother	female	0.034	0.018	1.05×10^{-5}	0.17	0.11
		male	0.031	0.018	6.93×10^{-5}	0.15	0.11
	Father	female	0.028	0.016	7.65×10^{-5}	0.13	0.11
		male	0.030	0.016	4.89×10^{-6}	0.13	0.1
	Daughter	female	0.033	0.016	1.60×10^{-6}	0.17	0.11
		male	0.034	0.018	4.38×10^{-5}	0.14	0.12
	Son	female	0.021	0.013	6.27×10^{-4}	0.13	0.11
		male	0.027	0.015	5.98×10^{-5}	0.12	0.1
Out-Call Fraction	Mother	female	0.038	0.015	1.75×10^{-4}	0.3	0.37
		male	0.047	0.016	0	0.31	0.37
	Father	female	0.054	0.022	0	0.3	0.38
		male	0.058	0.021	0	0.3	0.37
	Daughter	female	0.027	0.008	0	0.29	0.36
		male	0.043	0.007	0	0.29	0.36

Table S5. *Cont.*

Variable	Alter Type	Ego Sex	Standard Deviation of Means Across Time		*p*-Value	Total Standard Deviation	
			Kin	Non-Kin		Kin	Non-Kin
Out-Call Fraction	Son	female	0.036	0.010	0	0.3	0.36
		male	0.052	0.010	0	0.29	0.35
Call length	Mother	female	28.22	16.74	0	118.4	130.3
		male	21.24	8.82	1.75×10^{-4}	106.6	123.5
	Father	female	29.05	12.71	0	99.7	121
		male	23.68	11.33	0	95.7	107.8
	Daughter	female	19.40	12.19	1.45×10^{-3}	114.1	133
		male	14.43	8.43	3.09×10^{-4}	94.3	118.4
	Son	female	15.19	8.39	8.38×10^{-5}	104.8	126.5
		male	13.79	7.33	3.29×10^{-5}	92.6	105.6

S4. Data Anonymization and Metadata

The data is aggregated anonymized Call Detail Records (CDRs) from a Chilean Mobile Call Company accounting for ∼40% of the market share for the respective period. CDRs are generated automatically by the telephone company every time that a call is made or received by a person inside the network. This data is collected automatically for billing purposes.

Each record stores, among other things, the origin and destination number and antenna, a time stamp (day, hour, minute, second) and the duration of the call in seconds. The data used in this study was anonymized using several techniques (see the supplementary material for details). First, we used hash functions to convert phone numbers into a different string while still being able to connect all the phone calls that were made by that number. We also excluded attributes from the CDR that are not required or cannot be used for ethical reasons (antennas, exact time of phone calls). The information extracted from CDRs includes phone of origin (Origin Phone), destination phone (Destination Phone) and duration (Duration).

This information was aggregated for every pair of phone numbers that appeared at least once in the CDRs. The final database used was an aggregation of all of the phone calls made during 2015, for which at least one phone was from Movistar as seen in Table 1.

Table S6. Aggregated CDRs.

Phone_A	Phone_B	OutCalls	InCalls	Sec
71e61e625c967f98da69	bbb818a312f0fdb0771d	3	1	512
da1f483278cf32d22aa5	562a74c3d213871edf6b	1	1	333
71e61e625c967f98da69	562a74c3d213871edf6b	1	0	957

Besides the CDR data, we have an anonymized version of Movistar's Clients Registry with metadata about the owners of the phone lines. This metadata is collected by the company and includes, among other things, date of birth, gender, names, last names and also information about the type of contract (individual or family contract).

For our project the data was restricted to attributes: age (on 1 January 2015), gender, first and second last name, and type of contract. For data protection we replaced the date of birth by age on 1 January 2015. For family contracts we used the detailed records of phone owners given by the company and we only left the oldest phone number with the metadata. The rest of the phones of the plan were left as null values.

Table 2 is an example of the client database with anonymized cell phone numbers (Phone), paternal last name (LNp), maternal last name (LNm), anonymized owner ID (to identify family and individual plans).

In this example, both phones 71e61e625c967f98da69 and da1f483278cf73d22aa5 have the same owner adfr54kjhy5687lootek. The metadata was always associated to phone that had the oldest contract. In this case 71e61e625c967f98da69.

Table S7. Client' Registry Database

Phone	LNp	LNm	Sex	Age	OwnerID
71e61e625c967f98da69	X1	X10	Male	42	adfr54kjhy5687lootek
da1f483278cf32d22aa5	NULL	NULL	NULL	NULL	adfr54kjhy5687lootek
562a74c3d213871edf6b	X24	X5	Female	20	oujh65dfhk87rfjih677
bbb818a312f0fdb0771d	X24	X8	Male	47	gp984asw12rcyy998rjh

The data was collected for the 12 consecutive months of the year 2015, totalling 3,994,595,128 calls. Using all this data we created a Mobile Call Graph (Gmc). This directed graph Gmc = (Vmc, Emc) there is a node A ∈ Vmc

for every phone number in the dataset, and there is an edge (A, B) ∈ Emc if a, b ∈ Vmc, there is at least one phone call from phone A to phone B. Each node has labels or attributes that correspond to the metadata of the phone. If the phone is not the main one of a Movistar plan or the node does not belong to the Movistar network, all the labels associated to it are null values. The edges are labelled with the number of Out-calls, In-Calls and TotalSec. The final graph created consisted in 8,907,140 vertices and 112,744,511 edges.

One of the aspects that we considered together with Telefonica in the definition of the data to be delivered is that they comply with the handling of personal data required by the GDPR of the European Union. The process described above effectively meets the requirements since the data cannot be attributed to a particular subject without additional information.

Note on the Metadata

The metadata contain also information about the type of telephone contract (individual or family contract). We only used nodes where metadata was available for both egos and at least one of the alters. For family contracts we used the detailed records of phone owners given by the company. If an individual moved from a private phone contract to a family contract, it was assumed that the person kept using the same number instead of switching the phone with another member of the family. Thus, for each family contract we only left one node, the node where ownership of the phone can be traced to the person signing the family contract. The rest of the phone numbers included in the family contract were not included in the analysis left as grey nodes f (no metadata).