
This is an electronic reprint of the original article.
This reprint may differ from the original in pagination and typographic detail.

Fudolig, Mikaela Irene D.; Monsivais, Daniel; Bhattacharya, Kunal; Jo, Hang-Hyun; Kaski, Kimmo

Different patterns of social closeness observed in mobile phone communication

Published in:
Journal of Computational Social Science

DOI:
[10.1007/s42001-019-00054-8](https://doi.org/10.1007/s42001-019-00054-8)

Published: 01/01/2020

Document Version
Publisher's PDF, also known as Version of record

Published under the following license:
CC BY

Please cite the original version:
Fudolig, M. I. D., Monsivais, D., Bhattacharya, K., Jo, H.-H., & Kaski, K. (2020). Different patterns of social closeness observed in mobile phone communication. *Journal of Computational Social Science*, 3, 1-17.
<https://doi.org/10.1007/s42001-019-00054-8>



Different patterns of social closeness observed in mobile phone communication

Mikaela Irene D. Fudolig¹ · Daniel Monsivais² · Kunal Bhattacharya² · Hang-Hyun Jo^{1,2,3} · Kimmo Kaski^{2,4}

Received: 1 April 2019 / Accepted: 25 July 2019
© The Author(s) 2019

Abstract

We analyze a large-scale mobile phone call dataset containing information on the age, gender, and billing locality of users to get insight into social closeness in pairs of individuals of similar age. We show that in addition to using the demographic information, the ranking of contacts by their call frequency in egocentric networks is crucial to characterize the different communication patterns. We find that mutually top-ranked opposite-gender pairs show the highest levels of call frequency and daily regularity, which is consistent with the behavior of real-life romantic partners. At somewhat lower level of call frequency and daily regularity come the mutually top-ranked same-gender pairs, while the lowest call frequency and daily regularity are observed for mutually non-top-ranked pairs. We have also observed that older pairs tend to call less frequently and less regularly than younger pairs, while the average call durations exhibit a more complex dependence on age. We expect that a more detailed analysis can help us better characterize the nature of relationships between pairs of individuals and distinguish between various types of relations, such as siblings, friends, and romantic partners.

Keywords Communication patterns · Social closeness · Mobile phone data · Call frequency ranking

Introduction

Traditionally, the studies of human relationships and human social networks have been conducted using questionnaire-based surveys [1, 2]. As these surveys focus on detailed information about social ties between human individuals, they tend to

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s42001-019-00054-8>) contains supplementary material, which is available to authorized users.

✉ Hang-Hyun Jo
hang-hyun.jo@apctp.org

Extended author information available on the last page of the article

be limited by the number of subjects, by the relative uniformity of subjects often recruited from the same social surrounding, and by the memory of the subjects filling the questionnaires. Recent digital technologies such as mobile phones and Radio Frequency Identification (RFID) have enabled researchers to supplement the survey data with much more detailed relational data between subjects [3–5]. Although the data in these studies are still limited in sample size and in the diversity of the subjects, they have paved the way for more accurate and quantitative description of social behavior of human individuals embedded in a social environment or network. An additional benefit of this type of studies is that they allow the cross-validation of the data gathered from different sources [6, 7].

Large-scale mobile phone datasets have also become available due to rapid advance of digital mobile phone technology in the hands of people generating vast amount of data as traces of their behavior. This has facilitated a complementary approach to investigate human sociality even at the population level. The call detail records (CDRs) of mobile phone usage enable us to map the patterns of sociality at diverse scales, i.e., from studying the structure and dynamics of large-scale social networks [8, 9], to the level of communities and groups [10], to immediate social neighborhood of individuals in terms of egocentric networks [11, 12]. In these studies, the strength of ties between individuals has often been quantified in terms of the frequency of contact between them. More recently, such information on the tie strength has been combined with the metadata, such as the age, gender, and billing post code of users, which in turn has enabled us to gain deeper insight into the nature of human sociality [11–14].

However, as the CDRs are anonymized, they do not carry the true nature of relationships between individuals. An approach to circumvent this issue is to utilize the demographic and/or location information of the users and to make plausible assumptions about the nature of relationships between the users [15, 16]. For example, for a given user (an ego), the contacts of the ego (alters) are ranked by the call frequency between the ego and each alter; a few of the top-ranked alters are selected for the study. Then the tie strengths of close relationships are correlated with the age, gender, and location information of the users. The findings in these studies turn out to be indicative and consistent with the well-understood life-course patterns of human sociality [17, 18]. However, since this approach only uses demographic information of the users to infer the relationship between them, it cannot distinguish between pairs with similar demographics but different relationships, e.g., between opposite-gender friends and opposite-gender romantic partners. To address this limitation, we study whether different communication patterns can be found even among pairs with similar demographics. If pairs with similar demographic information exhibit different communication patterns, it may indicate that these pairs correspond to different types of relationships.

In the present study, we extend the above-described approach to analyze the CDRs focusing on getting insight into the nature of relationships between pairs of individuals of similar age. We combine the metadata of users, including age, gender, and billing post code, with the information about the ranks in each other's egocentric networks to find groups of pairs showing different calling patterns in terms of the average daily call frequency, daily regularity, and average call duration. These

groups can be interpreted as corresponding to either “close” or “casual” pairs, with close pairs being characterized by higher call frequency and regularity as well as longer call durations than those of casual pairs. We find that the rank information, which has not been used in any previous work, is crucial to separate these groups. We note that although some socially close pairs do not necessarily call frequently, it has been shown that calling patterns can be reliably used to predict user-validated social closeness [19].

Our paper is organized as follows: in Sect. 2, we present the description of the data used in this study, followed by the methods of data preprocessing and statistical tests. Then in Sect. 3 we focus on distinguishing the pairs of individuals that show different calling patterns in terms of average daily call frequency, daily regularity, and average call duration. In Sect. 4, we discuss how these results can be used to interpret pairs as being either close or casual. We also relate our findings to those from small-scale sociological studies. Finally, we draw conclusions in Sect. 5.

Methods

Data description

We analyze the mobile phone call dataset of a European service provider for the first 7 months of year 2007 (212 days). During this period, which is before the rise of smartphones and social network services, a significant part of the mobile communication was done through voice calls and Short Message Services (SMSs). The service provider had subscribers numbering around 20% of the population of the country [8].

The dataset contains the date and time for all the outgoing and incoming calls between subscribers or users. The duration is included for the calls between the users and for the outgoing calls from the users to those who subscribed to other providers, which we call non-company users. The duration is zero for incoming calls from non-company users to users, but the date and time of such calls are included. We discard the users whose contracts are known to begin or end within the period of interest, i.e., the first seven months in 2007.

For each service contract, some metadata of the users, such as age and gender, are included in the dataset; for most users, the billing post code is also included. There are a number of users, however, with unknown or ambiguous age and gender: these include the non-company users who have no user metadata, as well as users associated with service contracts that have multiple identifiers. Since it is impossible to determine whether the multiple identifiers correspond to different phones owned by the same person or different individuals, only the identifier with the earliest subscription date retained the age and gender information, while other identifiers are considered to have unknown age and gender.

Data preprocessing

For each user with known metadata, which we call an “ego”, we enlist all the other users the ego communicated with, which we call “alters”. The alters, which may include users with unknown age and gender, are ranked in descending order according to the total number of incoming and outgoing calls made between the ego and each alter. By keeping the top five alters for each ego, which likely correspond to the ego’s emotionally closest alters [20, 21], we make the list of ego–alter pairs.

We are interested only in ego–alter pairs who have significant relationships, such as family, friends, and romantic partners. To filter out pairs which do not meet this criterion, we impose regularity by excluding purely transactional calls, which are characterized by lower call frequency and less regularity [22]. Specifically, we exclude pairs who have had calls in less than five out of the seven months. For example, if a pair has one thousand calls but only for the first month, we exclude that pair from the analysis. Further, we also exclude the ego–alter pairs in which the metadata of the alter does not contain the age and gender, including the non-company users. After these filtering steps, we are left with the users with known metadata who make calls regularly to each other. Note that, since the filtering is done after ranking, the ranking preserves the true importance of the alter to the ego, as far as call frequency is concerned.

It is possible that two users appear in each other’s list of the top five alters. In such a case, the total number of incoming and outgoing calls and the total call duration are the same for both users, but this pair appears twice in the list of ego–alter pairs; we keep only one of these two.

After the above-described ranking, filtering, and removing of duplicates, we are left with 322,823 users in 1,236,364 pairs. Of these, we consider the pairs more likely to be in a peer relationship rather than in a parent–child relationship. These two relationships can be distinguished using the age difference of the users; we set the cutoff to be 20 years, based on European census data [23]. Then we study pairs whose age difference is less than 20 years, which are then categorized into nine demographic groups. We first consider three combinations of genders of each pair: (1) opposite gender, denoted by “−”, (2) same-gender female or “+f”, and (3) same-gender male or “+m”. For each gender combination group, we consider three age groups according to the age of the younger user in a pair, being either 18–28 years old (young adulthood or “Y”), 29–45 years old (middle adulthood or “M”), or 46–55 years old (late adulthood or “L”), following the scheme of life stages used in Ref. [15].¹ Consequently, we focus on 768,411 pairs in nine demographically separable groups, denoted by −Y, −M, −L, +fY, +fM, +fL, +mY, +mM, and +mL, respectively. Each of these groups has at least 20,000 ego–alter pairs, as summarized in Table 1. Also, although we consider the maximum age difference to be 20 years, we

¹ Other age categorization schemes were tried, such as considering only pairs where both users belong to the same age group or categorizing pairs based on the average age of the two users. The results are found to be robust even with these changes in the categorization scheme.

Table 1 Numbers of pairs in nine different demographic groups according to the genders and the younger user's age in each pair, with percentages in each group when decomposed by the ranks of users in each other's egocentric networks, i.e., mutual top-rank (1-1), mutual non-top-rank (*n-n*), and non-mutual top-rank (1-*n*)

	Opposite gender (–)			Same-gender female (+f)		Same-gender male (+m)	
Young adulthood (Y)	95,128			47,930		70,945	
	1-1	38.0%		1-1	6.9%	1-1	11.0%
	<i>n-n</i>	45.6%		<i>n-n</i>	75.5%	<i>n-n</i>	69.9%
	1- <i>n</i>	16.5%		1- <i>n</i>	17.7%	1- <i>n</i>	19.2%
Middle adulthood (M)	232,133			98,225		120,256	
	1-1	40.3%		1-1	9.9%	1-1	14.0%
	<i>n-n</i>	38.5%		<i>n-n</i>	68.8%	<i>n-n</i>	65.0%
	1- <i>n</i>	21.2%		1- <i>n</i>	21.3%	1- <i>n</i>	21.0%
Late adulthood (L)	57,085			20,165		26,544	
	1-1	30.1%		1-1	14.7%	1-1	19.9%
	<i>n-n</i>	40.0%		<i>n-n</i>	59.3%	<i>n-n</i>	50.6%
	1- <i>n</i>	29.9%		1- <i>n</i>	26.0%	1- <i>n</i>	29.5%

Note that due to rounding errors, the percentages may not sum to exactly 100%

Table 2 Distributions of the age differences of the pairs

	Age difference (years)	Opposite gender (–)			Same-gender female (+f)			Same-gender male (+m)		
		1-1 (%)	<i>n-n</i> (%)	1- <i>n</i> (%)	1-1 (%)	<i>n-n</i> (%)	1- <i>n</i> (%)	1-1 (%)	<i>n-n</i> (%)	1- <i>n</i> (%)
Y	0–5	73	56	62	68	64	65	71	68	71
	6–10	20	26	23	17	23	20	20	21	19
	11–19	7	18	15	15	13	15	9	11	10
M	0–5	78	58	68	69	64	65	75	67	69
	6–10	16	24	20	16	21	17	17	21	19
	11–19	6	18	13	15	15	17	8	13	12
L	0–5	76	62	72	75	61	71	83	65	76
	6–10	17	23	19	15	23	17	12	21	15
	11–19	6	15	9	10	16	12	6	14	8

For all demographic groups, pairs with an age difference of 0–5 years comprise more than half of each group, followed by those with age differences in the range of 6–10 years. Although the peers are defined to have an age difference less than 20 years, most of the pairs in each group show age differences less than 10 years. Due to rounding errors, the percentages may not sum to exactly 100%

find that most of these pairs show an age difference of only 0–5 years, as shown in Table 2.

Each demographic group can be further divided into three subgroups according to the ranks of users in each other's egocentric networks, as follows: (1) Both users in a pair are the top-rank alters of each other, which can be called mutual top-rank and denoted by “1-1”, (2) both users are not the top-rank alters of each other, i.e.,

mutual non-top-rank or “ n - n ”, and (3) one of the users is the top-rank alter of the other, but it is not mutual, i.e., non-mutual top-rank or “1- n ”. Table 1 shows that for all age groups, mutual top-rank pairs comprise a large portion in the opposite-gender groups, while they are a small minority in the same-gender groups.

In addition to all of the above, we can extract the location information of users with the help of the billing post code, which we assume to correspond to the user’s home address. We will focus on whether the users of each pair have the same post code or different ones.

Statistical test

All the statistical tests are done on the log-transformed variable whenever necessary. To test for statistical significance, we use one-way ANOVA and Tukey’s HSD post hoc test when the variances are found equal by Levene’s test [24]. If heteroscedasticity is obtained, Welch’s ANOVA [25] and the Games–Howell post hoc test [26] are used instead. The tests are implemented using Python’s *scipy* and *statsmodels* as well as using R’s *userfriendlyscience* packages.

Due to the large sample sizes in this study, the power of statistical tests is high [27], and true differences, no matter how small they are, are more likely to be found as significant. For brevity, we only mention the relevant results of the statistical tests where the null hypothesis cannot be rejected. Otherwise, the statistical tests either show a significant difference or are overruled by practical significance. We also report Cohen’s d , an effect size which quantifies the difference in means between two groups [28], whenever useful (see the details in Supplementary Material).

Results

To quantify the calling patterns of ego–alter pairs, we introduce three quantities, i.e., the average daily call frequency, fraction of days active, and average call duration. The distributions of these quantities are then systematically compared across different demographic groups.

Average daily call frequency

We first obtain the number of calls made by each pair, i.e., the call frequency. Dividing this call frequency by the number of days in the observation period, i.e., 212 days, we get the average daily call frequency (DCF) per pair to obtain its distributions. In Fig. 1, we find that the distribution of DCFs for each of nine demographic groups can overall be described by unimodal distributions on a log-scale, except for the opposite-gender young adulthood (–Y) case in Fig. 1a, showing a clear bimodality. This bimodality is resolved by separating the pairs in the –Y group according to the ranks of the users in each other’s egocentric networks. We observe that the mutual top-rank (1-1) pairs and mutual non-top-rank (n - n) pairs successfully account for the right and left peaks of the bimodal distribution, such that the median values

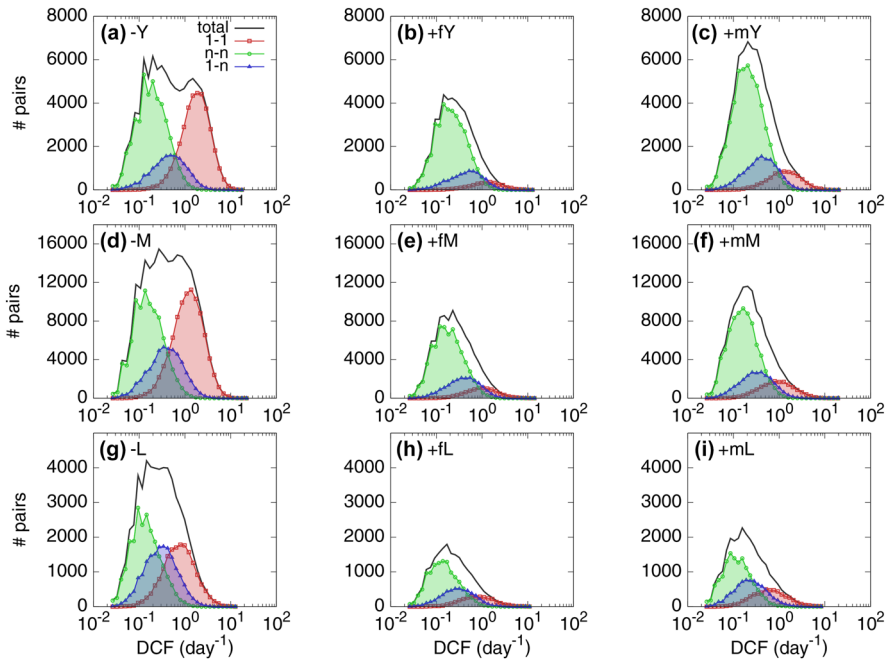


Fig. 1 Distributions of the average daily call frequency (DCF). For each demographic group, the total distribution (black solid line) is decomposed into three subgroups depending on the ranks of users in each other's egocentric networks: mutual top-rank (1-1; red line with squares), mutual non-top-rank (n - n ; green line with circles), and non-mutual top-rank (1- n ; blue line with triangles). The unit of DCF is day^{-1}

for 1-1 and n - n pairs are around 1.75 and 0.18 calls per day, respectively. There is also a non-mutual top-rank (1- n) minority whose distribution shows a peak between those of mutual top-rank and mutual non-top-rank pairs, which will be discussed in Subsect. 4.4. Although the bimodality found in the $-Y$ case is not evident in the rest of the groups, we find that the mutual top-rank pairs show, in general, largely different calling patterns from the mutual non-top-rank pairs in all the other demographic groups, as depicted in Fig. 2a. Note that although mutual top-rank pairs are expected to have more calls than mutual non-top-rank pairs, the successful decomposition of the bimodality using the rank information is not straightforward.

We summarize other relevant findings from the results in Figs. 1 and 2a. For all the gender combinations of pairs, younger pairs tend to call considerably (slightly) more often than older pairs in the mutual top-rank (mutual non-top-rank) case, with the effect sizes most pronounced among opposite-gender peers. For the mutual top-rank case, opposite-gender pairs call more frequently than their same-gender counterparts for both Y and M groups, while for the oldest (L) groups, there is no significant difference at $\alpha = 0.05$ between opposite-gender and same-gender female pairs ($p = 0.18$), but both opposite-gender and same-gender female pairs call more often than the same-gender male pairs. On the other hand, for the mutual non-top-rank case, we find no clear gender dependence of the DCF for each age group.

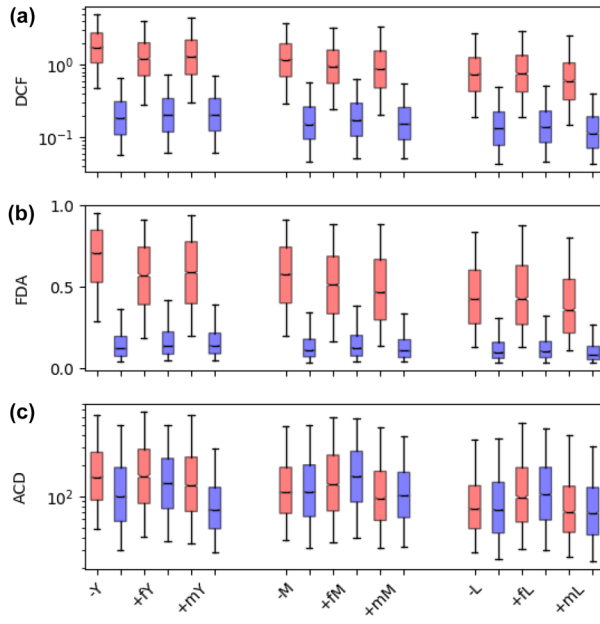


Fig. 2 Summary of distributions of the average daily call frequency (DCF), the fraction of days active (FDA), and the average call duration (ACD). The pink bars correspond to the mutual top-rank (1-1) pairs, while the blue bars are for the mutual non-top-rank ($n-n$) pairs. The black lines in the middle indicate the medians, the bars include data points from the 25th to the 75th percentile, and the whiskers show the 5th and 95th percentiles. The notches show confidence intervals generated by bootstrapping with 10000 resamples. The units of DCF and ACD are day^{-1} and s, respectively

Daily regularity

To quantify the temporal regularity of the calling patterns on a daily basis, we define the fraction of days active (FDA) as the fraction of days in the observation period in which at least one call was made between the users of each pair. By the FDA, one can distinguish, e.g., the case of 10 days with 10 calls per day from the case of 100 days with one call per day, which cannot be distinguished by the average daily call frequency (DCF).

We find that the FDA is highly correlated with the DCF ($r = 0.681$). It should be noted that the number of days active cannot be greater than the call frequency for each pair, which possibly enables the strongly positive correlation between FDA and DCF. However, how the calls are distributed over the observation period is yet an interesting question, in particular, for pairs with high DCF: the pairs with a high DCF tend to have a high FDA, implying that the calls are made rather regularly instead of being lumped into a few days. The distributions of FDAs are presented in Fig. 3 and summarized in Fig. 2b. Overall, we find similar behavior to that observed in the case of DCF, except that the shapes of the distributions are highly skewed either to the left or to the right, probably due to the intrinsic range of the quantity, i.e., $\text{FDA} \in [0, 1]$. The most pronounced difference between the

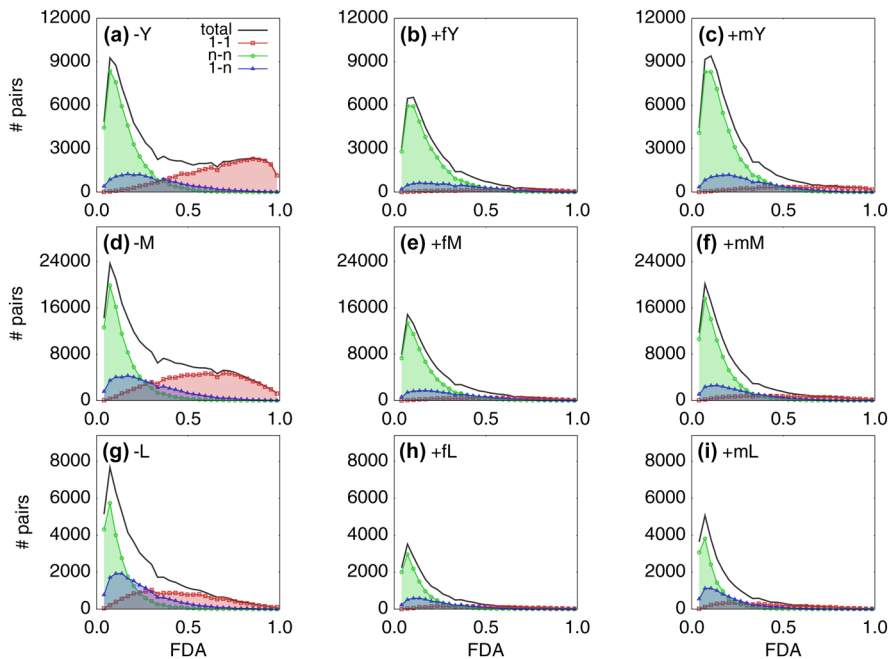


Fig. 3 Distributions of the fraction of days active (FDA). All notations are the same as those in Fig. 1

mutual top-rank and the mutual non-top-rank pairs is observed again in the $-Y$ group as their median values are 0.71 and 0.13, respectively.

Average call duration

To study the calling patterns in more detail, we calculate the average duration per call or the average call duration (ACD) for each pair by dividing the total call duration (in seconds) by the number of calls. The ACD turns out to be positively but weakly correlated with the DCF ($r = 0.087$) as well as with the FDA ($r = 0.075$). As shown in Figs. 2c and 4, unlike the DCF and FDA, there seems to be no clear demographic dependence of the ACD across the different age and gender groups. However, with the medians of the distributions we observe that for all the gender groups, younger pairs tend to have longer calls than older pairs only in the mutual top-rank case. Interestingly, the same-gender female pairs make longer calls than their opposite-gender and same-gender male counterparts for all the age groups, regardless of the ranks, except for one case; there is no significant difference between the mutual top-rank pairs in $-Y$ and $+fY$ groups ($p = 0.79$). We also find that in the young adulthood (Y) case, the mutual top-rank pairs have longer calls than the mutual non-top-rank pairs for all gender combinations, while the opposite tendency is significantly observed for the $+fM$ group.

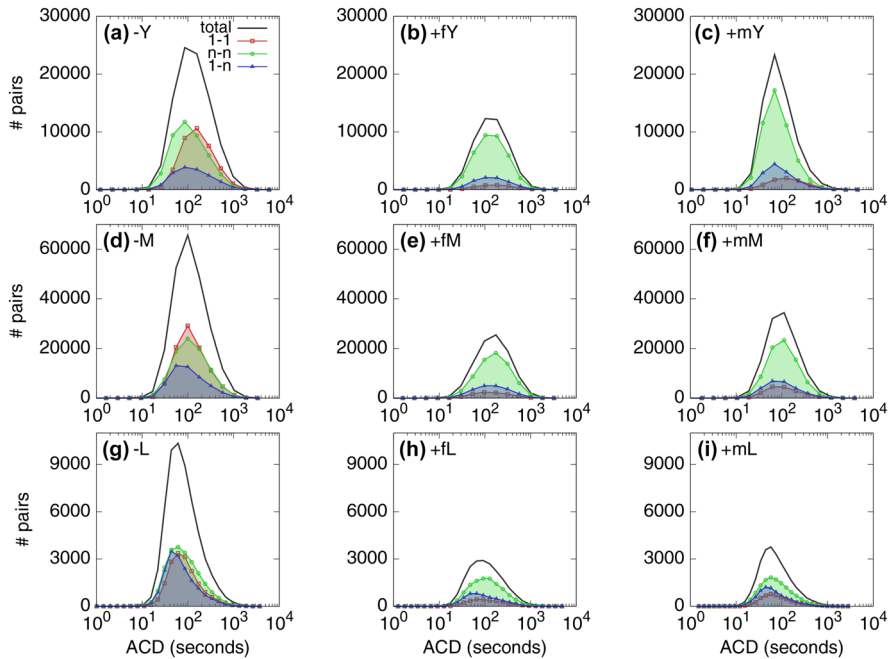


Fig. 4 Distributions of the average call duration (ACD). All notations are the same as those in Fig. 1, except that the unit of ACD is seconds

Location analysis

Next we analyze the location information of the pairs in close relationships by assuming that the billing post codes correspond to the home address of the users. It is known that the frequency of the face-to-face interaction, constrained by the location, is positively correlated with the frequency of contact by telephone and other media [29–31]. This enables us to study how the locations of users in close relationships are related to each other. In Fig. 5, we find that although the fraction of pairs living in the same post code increases with age of the group for all mutual top-rank pairs, the difference is most pronounced among opposite-gender pairs.

Discussion

Based on the above empirical observations, we hypothesize that across all demographic groups, mutual top-rank pairs and mutual non-top-rank pairs have essentially different calling patterns, implying different types of relationships that underlie the communication patterns. Mutual top-rank (1-1) pairs are characterized by a high number of calls and high daily regularity, and hence these relationships can be interpreted as close. On the other hand, mutual non-top-rank (n - n) pairs have fewer calls and very low regularity, leading us to interpret these relationships as casual. As

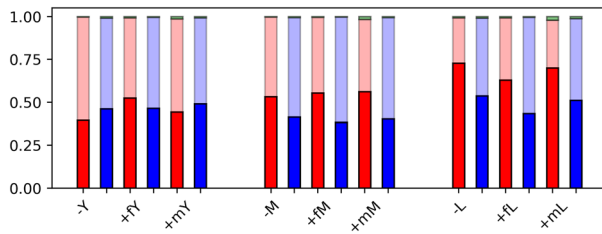


Fig. 5 Decomposition of groups by the billing post code. The red bars correspond to mutual top-rank (1-1) pairs, while the blue bars correspond to mutual non-top-rank (n - n) pairs for each group. The darker bars denote the fraction of pairs with the same post codes, while the lighter bars denote the fraction of those with different post codes. The fractions of pairs in which at least one post code is not available are denoted by the green bars

the calling patterns of mutual top-rank pairs are also differentiated by their genders, in the following we discuss three types of relationships: close relationships of opposite-gender pairs, close relationships of same-gender pairs, and casual relationships.

Close relationships of opposite-gender pairs

The opposite-gender, mutual top-rank pairs show the highest level of call frequency and regularity compared to all other gender and rank cases. This indicates a high level of closeness and suggests that these relationships are possibly romantic in nature. Although this interpretation cannot be confirmed by ground truth, our findings are consistent with those of small-scale sociological studies involving college students (corresponding to the age group of Y) in confirmed romantic relationships [32, 33]. These studies find that pairs with greater frequency or duration of phone calls have less relational uncertainty and higher intimacy. Moreover, those in romantic relationships are, on average, found to call each other more regularly [33]. In addition, we find that the mutual top-rank pairs form a significant chunk in the opposite-gender groups, but only a small minority in the same-gender groups. Since this reflects the profile of marriages and registered partnerships in a number of countries in Europe, where same-sex unions account for less than 5% of all registered unions [34–36], romantic partnerships seem to be the most plausible characterization of these opposite-gender pairs.

We also observe that for this kind of relationship, younger pairs tend to have more frequent, more regular, and longer calls than older pairs, as depicted in Fig. 2. To study whether this tendency is due to the lower usage of mobile phones among the older generation [37] or due to the actual communication patterns of older users, more work is called for.

Recall from Fig. 5 that the fraction of pairs living in the same post code increases with age of the group for all mutual top-rank pairs, especially among opposite-gender pairs. If these opposite-gender pairs are romantic, this may be explained by the higher likelihood of cohabitation and/or marriage as people get older, which is consistent with the previous empirical findings using the same dataset [12].

Close relationships of same-gender pairs

The same-gender, mutual top-rank pairs can also be considered as being close across all age groups as their calling patterns are clearly more active and regular than their mutual non-top-rank counterparts. Yet, they are less active and less regular than their opposite-gender counterparts, which implies that the same-gender, mutual top-rank pairs have a different type of relationship compared to close opposite-gender pairs. However, such differences turn out to get smaller for older age groups. In addition, in terms of the median of the distribution of average call duration, the same-gender female pairs tend to have longer calls than the opposite-gender and same-gender male pairs, which is consistent with the findings in Refs. [13, 15].

To characterize some of the same-gender close relationships as romantic, we need more supporting evidence for the communication patterns of same-gender romantic relationships. The most one can say is that these pairs may be a mixture of romantic, familial, and other relationships.

Casual relationships

The mutual non-top-rank pairs in all demographic groups are here considered as casual relationships, as they are characterized by the lowest level of call frequency and daily regularity compared to their mutual top-rank counterparts. As for the average call duration, in terms of the median of its distribution, the mutual non-top-rank casual pairs tend to have shorter or similar call durations than the mutual top-rank close pairs in most cases, except for the +fM case, where the average call duration of casual pairs (around 157 s) is significantly larger than that of the close pairs (around 134 s). Moreover, among the mutual non-top-rank pairs, the +fM group shows the longest average call duration, which could be due to requirements of child rearing, job demands in the mid- to high-level careers, or other life events.

Similarly to the opposite-gender and same-gender close relationships, the average daily call frequency and the fraction of days active in casual relationships are decreasing with their age. However, the average call duration is the highest for the middle adulthood (M) group irrespective of gender.

We also note that unlike for the close pairs, the fraction of casual pairs with the same billing post code does not show a clear dependence on age group as is evident in Fig. 5.

Other relevant issues

So far we have focused on the ego–alter pairs as if they are separated from the rest of the social network. By incorporating the network structure surrounding those pairs, one can tackle some unresolved issues. For example, friends, family, and romantic relationships may be differentiated using the information about their common contacts, while the non-mutual top-rank (1-*n*) pairs may be studied in the context of directed relationships [38, 39]. Since the peak of the DCF distribution in the 1-*n*

case lies between the mutual top-rank (1-1) and mutual non-top-rank (n - n) peaks, we can hypothesize that they may exhibit different behaviors from both. They may also be composed of two subgroups, one resembling the mutual top-rank pairs, the other resembling the mutual non-top-rank pairs.

Conclusion

We have analyzed the large-scale call detail records (CDRs) with the metadata, such as the age, gender, and billing post code of mobile phone users, by focusing on around 770,000 pairs with an age difference of less than 20 years. We show that in addition to the metadata, the ranks of users in each other's egocentric networks, determined by the call frequency between them, reveal different communication patterns even among pairs of individuals with similar demographic information. Whereas previous studies have used only demographic information to determine the most likely relationship between users [15, 16], our findings show that adding rank information may give us insights into how the nature of relationships can vary even within the same demographic group. In particular, mutual top-rank pairs have markedly different calling patterns from mutual non-top-rank pairs, not only in terms of call frequency but also in terms of daily regularity. These differences could enable us to interpret mutual top-rank pairs as being close relationships and mutual non-top-rank pairs as being casual relationships, respectively. Although there is no information on the actual nature of the relationships among the communicating individuals, our interpretation is consistent with the findings from small-scale sociological studies with verified ground truth.

We have found that mutual top-rank pairs are much more common among opposite-gender pairs. This, as well as the consistency of their calling patterns with those observed for romantic couples, makes it plausible that mutually top-ranked opposite-gender pairs reflect romantic relationships. On the other hand, although mutually top-ranked same-gender pairs also have relatively high call frequency and regularity, they have different calling patterns compared to their mutually top-ranked opposite-gender counterparts. This may be because the same-gender group is not solely composed of romantic partnerships, but may also include platonic or familial ties as well. In contrast to the mutual top-rank pairs, the mutual non-top-rank pairs exhibit the lowest levels of daily regularity and call frequency. We suppose that these pairs are very unlikely to be romantic partners, but instead they are more likely to be platonic or familial pairs.

The calling patterns between pairs of individuals have also been found to vary with the age of the users. We find that older pairs tend to call less frequently and less regularly than younger pairs. The mutual top-rank pairs tend to also have longer average call durations than the mutual non-top-rank pairs for younger pairs of 18–28 years old. For the older pairs, the difference between the mutual top-rank and the mutual non-top-rank pairs is smaller. Interestingly, we find that in the case of the female pairs in the age range of 29–45 years, the mutual non-top-rank pairs make significantly longer calls than the mutual top-rank pairs. This age range corresponds to the period when most couples begin families; hence, such calling patterns may be

due to the demands of family and work on women of that age range. Our findings can be related to the shift in social focus: while both men and women in their young adulthood are likely to maintain stronger social focus on their partners [40], the attention of individuals in middle adulthood gets distributed to alters other than their partners due to time constraints and the increase in the number of familial ties [13].

We also comment on the limitations of our work. First, although non-company users were included in the ranking of the alters by call frequency, they could not be analyzed in more detail as their age and gender were unknown. Thus, our work is biased towards pairs that belong to the same service provider. Second, our dataset was extracted in the year 2007 before the rise of internet-based messaging applications; thus, we expect our dataset to cover most mobile communication at that time, which is via calls and texts. However, advances in technology may have changed the way people communicate, and it would be interesting to apply our methods to more current datasets if they are made available by the application developers.

Finally, we discuss possible future studies. While we have focused exclusively on peers, we can also investigate the parent–child relationships. In addition, as our analysis has focused on the ego–alter pairs, network analysis may help us to uncover more about the users' relationships, and even to distinguish between other types of relationships, such as platonic and familial relationships. These are all interesting issues for future work.

Acknowledgements Open access funding provided by Aalto University. We acknowledge Takayuki Hirakawa for helpful discussions and we also acknowledge the computational resources provided by the Aalto Science-IT project. D.M. acknowledges CONACYT, Mexico, for supporting Grant 383907. D.M., K.B., and K.K. acknowledge support from EU HORIZON 2020 FET Open RIA project (IBSEN) No. 662725, and EU HORIZON 2020 INFRAIA-1-2014-2015 program project (SoBigData) no. 654024. K.K. also acknowledges the Rutherford Foundation Visiting Fellowship at The Alan Turing Institute, UK. H.-H.J. acknowledges financial support by Basic Science Research Program through the National Research Foundation of Korea (NRF) Grant funded by the Ministry of Education (NRF-2018R1D1A1A09081919).

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Zachary, W. W. (1977). An information flow model for conflict and fission in small groups. *Journal of Anthropological Research*, 33(4), 452. <https://doi.org/10.1086/jar.33.4.3629752>.
2. Wasserman, S., & Faust, K. (1994). *Social Network Analysis* (Cambridge University Press, Cambridge). <https://doi.org/10.1017/CBO9780511815478>. <http://ebooks.cambridge.org/ref/id/CBO9780511815478>.
3. Eagle, N., & Pentland, A. S. (2006). Reality mining: sensing complex social systems. *Personal and Ubiquitous Computing*, 10(4), 255. <https://doi.org/10.1007/s00779-005-0046-3>.

4. Jo, H. H., Karsai, M., Karikoski, J., & Kaski, K. (2012). Spatiotemporal correlations of handset-based service usages. *EPJ Data Science*, 1(1), 10. <https://doi.org/10.1140/epjds10>.
5. Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J. F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PLoS ONE*, 5(7), e11596. <https://doi.org/10.1371/journal.pone.0011596>.
6. Stopczynski, A., Sekara, V., Sapiezynski, P., Cuttone, A., Madsen, M. M., Larsen, J. E., et al. (2014). Measuring large-scale social networks with high resolution. *PLoS ONE*, 9(4), e95978. <https://doi.org/10.1371/journal.pone.0095978>.
7. Wang, R., Chen, F., Chen, Z., Li, T., Harari, G., Tignor, S., Zhou, X., Ben-Zeev, D., & Campbell, A. T. (2014). StudentLife: assessing mental health, academic performance and behavioral trends of college students using smartphones. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct* (ACM Press, New York, New York, USA), (pp. 3–14). <https://doi.org/10.1145/2632048.2632054>.
8. 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 of the United States of America* 104(18), 7332. <https://doi.org/10.1073/pnas.0610245104>.
9. Karsai, M., Kiveliä, M., Pan, R. K., Kaski, K., Kertész, J., Barabási, A. L., & Saramäki, J. (2011). Small but slow world: how network topology and burstiness slow down spreading. *Physical Review E*, 83(2), 025102. <https://doi.org/10.1103/PhysRevE.83.025102>.
10. Blondel, V. D., Guillaume, J. L., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, P10008. <https://doi.org/10.1088/1742-5468/2008/10/P10008>.
11. Palchykov, V., Kaski, K., Kertész, J., Barabási, A. L., & Dunbar, R. I. M. (2012). Sex differences in intimate relationships. *Scientific Reports*, 2(1), 370. <https://doi.org/10.1038/srep00370>.
12. Jo, H. H., Saramäki, J., Dunbar, R. I. M., & Kaski, K. (2014). Spatial patterns of close relationships across the lifespan. *Scientific Reports*, 4(1), 6988. <https://doi.org/10.1038/srep06988>.
13. Bhattacharya, K., Ghosh, A., Monsivais, D., Dunbar, R. I. M., & Kaski, K. (2016). Sex differences in social focus across the life cycle in humans. *Royal Society Open Science*, 3(4), 160097. <https://doi.org/10.1098/rsos.160097>.
14. Dong, Y., Yang, Y., Tang, J., Yang, Y., & Chawla, N. V. (2014). Inferring user demographics and social strategies in mobile social networks. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '14* (ACM Press, New York, New York, USA), (pp. 15–24). <https://doi.org/10.1145/2623330.2623703>.
15. David-Barrett, T., Kertész, J., Rotkirch, A., Ghosh, A., Bhattacharya, K., Monsivais, D., et al. (2016). Communication with family and friends across the life course. *PLOS ONE*, 11(11), e0165687. <https://doi.org/10.1371/journal.pone.0165687>.
16. David-Barrett, T., Rotkirch, A., Ghosh, A., Bhattacharya, K., Monsivais, D., Behncke, I., Kertész, J., & Kaski, K. (2017). Peer relations with mobile phone data: best friends and family formation. *arxiv* :1708.07759.
17. Hawkes, K. (2004). The grandmother effect. *Nature*, 428(6979), 128. <https://doi.org/10.1038/428128a>.
18. Kalmijn, M. (2012). Longitudinal analyses of the effects of age, marriage, and parenthood on social contacts and support. *Advances in Life Course Research*, 17(4), 177. <https://doi.org/10.1016/J.ALCR.2012.08.002>.
19. Phithakkitnukoon, S., & Dantu, R. (2010). Mobile social closeness and communication patterns. In *2010 7th IEEE Consumer Communications and Networking Conference, CCNC 2010* (IEEE), (pp. 1–5). <https://doi.org/10.1109/CCNC.2010.5421787>.
20. Dunbar, R. I. M., & Spoors, M. (1995). Social networks, support cliques, and kinship. *Human Nature*, 6(3), 273. <https://doi.org/10.1007/BF02734142>.
21. Zhou, W. X., Sornette, D., Hill, R. A., & Dunbar, R. I. M. (2005). Discrete hierarchical organization of social group sizes. *Proc. R. Soc. B*, 272, 439. <https://doi.org/10.1098/rspb.2004.2970>.
22. Granovetter, M. S. (1973). The strength of weak ties. *American Journal of Sociology*, 78(6), 1360. <https://doi.org/10.1086/225469>.
23. Eurostat. Eurostat data on population (demography, migration and projections). <http://ec.europa.eu/eurostat/data/database>.
24. NIST/SEMATECH e-Handbook of Statistical Methods (2013). <https://www.itl.nist.gov/div898/handbook/index.htm>.

25. Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. *Biometrika*, 34(1–2), 28. <https://doi.org/10.1093/biomet/34.1-2.28>.
26. Games, P. A., & Howell, J. F. (1976). Pairwise multiple comparison procedures with unequal N's and/or variances: a Monte Carlo study. *Journal of Educational Statistics*, 1(2), 113. <https://doi.org/10.2307/1164979>.
27. Columb, M., & Atkinson, M. (2016). Statistical analysis: sample size and power estimations. *BJA Education*, 16(5), 159. <https://doi.org/10.1093/bjaed/mkv034>.
28. Sawilowsky, S. S. (2009). New effect size rules of thumb. *Journal of Modern Applied Statistical Methods*, 8(2), 597. <https://doi.org/10.22237/jmasm/1257035100>.
29. Roberts, S. G., Dunbar, R. I., Pollet, T. V., & Kuppens, T. (2009). Exploring variation in active network size: constraints and ego characteristics. *Social Networks*, 31(2), 138. <https://doi.org/10.1016/J.SOCNET.2008.12.002>.
30. Roberts, S. G. B., & Dunbar, R. I. M. (2011). Communication in social networks: effects of kinship, network size, and emotional closeness. *Personal Relationships*, 18(3), 439. <https://doi.org/10.1111/j.1475-6811.2010.01310.x>.
31. Saramäki, J., Leicht, E. A., López, E., Roberts, S. G. B., Reed-Tsochas, F., & Dunbar, R. I. M. (2014). Persistence of social signatures in human communication. *Proceedings of the National Academy of Sciences of the United States of America*, 111(3), 942. <https://doi.org/10.1073/pnas.1308540110>.
32. Jin, B., & Peña, J. F. (2010). Mobile communication in romantic relationships: mobile phone use, relational uncertainty, love, commitment, and attachment styles. *Communication Reports*, 23(1), 39. <https://doi.org/10.1080/08934211003598742>.
33. Morey, J. N., Gentzler, A. L., Creasy, B., Oberhauser, A. M., & Westerman, D. (2013). Young adults' use of communication technology within their romantic relationships and associations with attachment style. *Computers in Human Behavior*, 29(4), 1771. <https://doi.org/10.1016/J.CHB.2013.02.019>.
34. Statistisches Bundesamt (Federal Statistical Office of Germany). https://www.destatis.de/EN/Home/_node.html.
35. Instituto Nacional de Estadística (Spanish Statistical Office). <https://www.ine.es/en/welcome.shtml>.
36. Office for National Statistics. <https://www.ons.gov.uk/>.
37. Kurniawan, S. (2008). Older people and mobile phones: a multi-method investigation. *International Journal of Human-Computer Studies*, 66(12), 889. <https://doi.org/10.1016/J.IJHCS.2008.03.002>.
38. Ball, B., & Newman, M. (2013). Friendship networks and social status. *Network Science*, 1(01), 16. <https://doi.org/10.1017/nws.2012.4>.
39. Almaatouq, A., Radaelli, L., Pentland, A., & Shmueli, E. (2016). Are you your friends' friend? Poor perception of friendship ties limits the ability to promote behavioral change. *PLOS ONE*, 11(3), e0151588. <https://doi.org/10.1371/journal.pone.0151588>.
40. Burton-Chellew, M. N., & Dunbar, R. I. M. (2015). Romance and reproduction are socially costly. *Evolutionary Behavioral Sciences*, 9(4), 229. <https://doi.org/10.1037/ebs0000046>.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Mikaela Irene D. Fudolig¹ · Daniel Monsivais² · Kunal Bhattacharya² · Hang-Hyun Jo^{1,2,3}  · Kimmo Kaski^{2,4}

Mikaela Irene D. Fudolig
mikaela.fudolig@apctp.org

Daniel Monsivais
monsivais.daniel@gmail.com

Kunal Bhattacharya
kunal.bhattacharya@aalto.fi

Kimmo Kaski
kimmo.kaski@aalto.fi

¹ Asia Pacific Center for Theoretical Physics, Pohang, Republic of Korea

² Department of Computer Science, Aalto University, Espoo, Finland

³ Department of Physics, Pohang University of Science and Technology, Pohang, Republic of Korea

⁴ The Alan Turing Institute, London, UK