Machine Learning for Phone-Based Relationship Estimation: The Need to Consider Population Heterogeneity

TONY LIU, University of Pennsylvania, USA
JENNIFER NICHOLAS, Northwestern University, USA
MAX M. THEILIG, Northwestern University, USA
SHARATH C. GUNTUKU, University of Pennsylvania, USA
KONRAD KORDING, University of Pennsylvania, USA
DAVID C. MOHR, Northwestern University, USA
LYLE UNGAR, University of Pennsylvania, USA

Estimating the category and quality of interpersonal relationships from ubiquitous phone sensor data matters for studying mental well-being and social support. Prior work focused on using communication volume to estimate broad relationship categories, often with small samples. Here we contextualize communications by combining phone logs with demographic and location data to predict interpersonal relationship roles on a varied sample population using automated machine learning methods, producing better performance ($F1 = 0.68$) than using communication features alone ($F1 = 0.62$). We also explore the effect of age variation in the underlying training sample on interpersonal relationship prediction and find that models trained on younger subgroups, which is popular in the field via student participation and recruitment, generalize poorly to the wider population. Our results not only illustrate the value of using data across demographics, communication patterns and semantic locations for relationship prediction, but also underscore the importance of considering population heterogeneity in phone-based personal sensing studies.

CCS Concepts: • **Human-centered computing** → *Ubiquitous and mobile computing systems and tools*; **Empirical studies in ubiquitous and mobile computing**; *Ubiquitous and mobile devices*; • **Information systems** → Data mining; • **Computing methodologies** → *Machine learning approaches.*

Additional Key Words and Phrases: automated machine learning, social relationship prediction, semantic location-based features, population heterogeneity

ACM Reference Format:
Tony Liu, Jennifer Nicholas, Max M. Theilig, Sharath C. Guntuku, Konrad Kording, David C. Mohr, and Lyle Ungar. 2019. Machine Learning for Phone-Based Relationship Estimation: The Need to Consider Population Heterogeneity. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 3, 4, Article 145 (December 2019), 23 pages. https://doi.org/10.1145/3369820

Authors' addresses: Tony Liu, liutony@seas.upenn.edu, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA; Jennifer Nicholas, Center for Behavioral Intervention Technologies, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; Max M. Theilig, max.theilig@northwestern.edu, Center for Behavioral Intervention Technologies, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; Sharath C. Guntuku, sharathg@sas.upenn.edu, Penn Medicine Center for Digital Health, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA; Konrad Kording, koerding@gmail.com, Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA; David C. Mohr, d-mohr@northwestern.edu, Center for Behavioral Intervention Technologies, Feinberg School of Medicine, Northwestern University, Chicago, IL, USA; Lyle Ungar, ungar@cis.upenn.edu, Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA.

2474-9567/2019/12-ART145 $15.00
https://doi.org/10.1145/3369820

1 INTRODUCTION

Being able to understand the nature of interpersonal relationships as revealed by phone-based communication can be useful for studying well-being. Lack of social support is a widespread problem with links to a wide variety of mental and physical health problems such as depression [16], schizophrenia, and substance abuse [36]. An individual's interpersonal context can be an indicator or risk factor: depressed individuals exhibit asymmetric communication patterns where they are less likely to initiate social interactions [18], while individuals with a strong support network are less prone to depressive episodes [15]. Individuals also select who they contact depending on what they are seeking from the interaction – calls to close friends can provide social support of a different nature than communication with family members [31]. Knowing the social role contacts play can be critical in understanding how communication influences well-being.

The proliferation of smart mobile devices that passively collect data on user behavior [19] is leading to an increasing interest in using passive sensing to estimate behavioral markers like social communication [23, 29]. Automated analysis of such passive data can potentially provide avenues for early detection and intervention of mental health [4]. If an automated process could detect elevated risk factors for mental health difficulties or other actionable behaviors, individuals could be targeted for a more thorough assessment (and provided with digital forms of support and treatment), alleviating many of the constraints associated with traditional assessment methods [37].

Previous works using passive sensing to estimate social relationships, tie strength, and friendship networks have used feature modalities such as communication logs, Bluetooth proximity, and geo-location tags to build machine learning models [11, 20, 27, 40]. However, many of these studies focus on broad social categories, such as friend vs. not friend [11], where there can be overlap between roles or contacts that do not fit cleanly into a category. Moreover, we currently know little about the relationship between communication events and demographics and semantic location (location labels such as work, home, shop, place of worship, etc.) data, which can provide additional information about the nature of relationships. Learning more about the fine-grained aspects of relationships and their generalization across demographics will better inform the use of passive sensing data for applications to social support.

Furthermore, many mobile personal sensing studies use small samples from narrow groups such as students to build machine learning models, which may not produce results that generalize to wider populations [29]. This further motivates the need to understand how demographics interact with phone sensor data; as we show below, the underlying communication tendencies of the training sample can skew a model towards features that work well for a particular demographic but transfer poorly to other groups. More generally, as machine learning tools become more ubiquitous, practitioners need to take care in correctly applying methods, especially when working with such small samples. Previous meta-analyses of the phone sensing space have shown that papers use spurious baselines [7] and improper record-wise cross-validation [35] that lead to false optimism in model performance. Our goal is not just to produce high accuracy models, but also to build models that are statistically meaningful and generalize well.

To address these gaps, we take a principled approach to social role prediction by contextualizing communication events with location and demographic data on a relatively large participant population (n = 189) collected from across the United States. We make the following contributions:

- We contextualize participant's communication patterns with semantic location information and demographics, achieving a weighted F1 score of 0.68 on a held-out test set.
- We perform feature analysis and show that there are significant correlations between both the temporality (time of day, day of week) and channel selection (call vs. text) of communication and the age of the participants.

- We use *auto-sklearn*, an automated machine learning technique [13], to build our models in order to reduce the biases introduced by manual model and feature selection.
- We present an experiment where models are trained only on subsets of the population divided into age quartiles to illustrate the impact of population heterogeneity on model performance.

The rest of the paper is organized as follows. We review related work in Section 2 and describe our data collection, feature extraction, and modeling methodology in Section 3. In Section 4, we present our relationship prediction modeling results and analyze feature importance as well as correlations in communication patterns. We introduce our subgroup prediction task and subsequent performance analysis in Section 5. We conclude by discussing implications of our results as well as limitations in Section 6.

2 RELATED WORK

The ubiquity of personal sensing platforms has provided an opportunity for many novel modeling methods of human social behavior. In particular, there have been numerous previous studies for predicting relationship roles or social network analysis through passively collected phone data spanning different sensor modalities, such as Bluetooth and communication logs. However, many of these studies also use data drawn from student or university participants, which are not representative samples of the wider population. We review prior work here to examine the breadth of model features as well as sample populations used in the literature.

Sample Composition. Understanding the composition of the underlying sample population is important for the evaluation of a predictive model's generalizability. An influential work in estimating social systems using phone sensors, Reality Mining [10], gathered Bluetooth proximity, usage, and communication data from an academic population, where the 94 participants in the study were affiliated with MIT. Several groups have used this public dataset to predict friend vs. non-friend relationships with high accuracy through a variety of features such as spatial proximity [11] and communication logs [28] as well as through novel modeling settings such as semi-supervised network inference [42]. Numerous other works have leveraged student samples to produce effective relationship prediction models, albeit at small sample sizes (12 to 25 participants) [9, 20, 33]. Similar personal sensing research with different inference targets such as social support [17], depression [24], and happiness [21] have also been explored in student populations. Though academic populations are accessible for research groups, models trained on these data may not produce results that generalize to samples of differing demographics or even of different academic institutions.

Thus, in order to improve a model's efficacy across a general population, more varied training samples are required. Indeed, some studies have collected data from more heterogeneous populations: Min et al. [27]'s study of predicting *life facets* (family, work, social) from phone data, as well as a follow-up study of tie strength [40], use a sample of 40 participants recruited across the United States, though they are still majority students. Choi et al. [5] mine relationship types from 22 participants within a workplace environment, though their population is homogeneous in other aspects: 20 of the participants were male and all were computer science majors. The data used by the CommSense phone mining framework [1] crowd-source their sample of 106 users from geographically diverse areas in the United States via an online platform. Overall, however, most personal sensing studies draw from small, homogeneous samples that are skewed towards a particular demographic.

Relationship Prediction Feature Modalities. Mobile devices today drive a significant amount of social interaction through digital communication, which is clearly important for assessing the nature of social relationships. Consequently, the primary modeling features used in previous relationship prediction work are communication patterns extracted from text message, call, email, and instant message logs, which are easily accessed through phone sensor data platforms [9, 28]. Co-location information is also often used to infer social network structure [5, 11, 20, 42]. However, co-location information requires "symmetric" data (from both target participants and their contacts) that would is more difficult to collect at scale or when participants are not geographically close. Part of the appeal of personal sensing is that data like communication logs are much easier to collect across a much wider range of populations.

Although phone communication patterns are good indicators for social relationships, differences in how users behave (e.g., due to life stage) or changing context (e.g. communications while traveling) potentially limit the predictive ability of communication features alone [40]. Previous studies address this by leveraging other sensor modalities to improve their models. Min et al. include demographic information (age and gender) as well as survey results that capture additional relationship information such as closeness, finding that the combination of this information with communication features produced the best predictive performance [27]. The CommSense system uses basic semantic location information to contextualize communication events, with labels for calls or texts that occur at home or at work [1]. Overall however, inferring social relationships with other features such as demographics and location needs to be explored further, as these data are easily collected through personal sensing platforms and can better characterize communication patterns.

Building on Previous Work. We use fine-grained demographic and user-categorized location features for relationship prediction on a heterogeneous population. Though our location labels are more detailed than the high-level work vs home vs other categories used in CommSense, previous work by Saeb et al. [34] has demonstrated that semantic location can be accurately inferred from phone sensor data, making these features viable for use in personal sensing studies. The addition of these features should not only improve predictive performance but also give insight into how people interact with contacts of a particular relationship type.

We also make contributions from a modeling perspective by using automated machine learning methods [13]. Lane et al. note that most groups studying personal sensing use hand-coded and hand-tuned models, which lead to questions on how well these models generalize and how they can be shared and standardized [23]. Automated machine learning methods are powerful, but because they limit user intervention when building predictive models, they can facilitate comparison of model performance on different personal sensing tasks.

We also note that the vast majority of these studies use small n samples that are primarily drawn from student populations, which are often homogeneous demographically, particularly in terms of their age distribution. Our age-based subgroup experiment presented in Section 5 explores the impact of training on such homogeneous samples, highlighting the need to collect data from more varied populations to produce personal sensing models that generalize well.

3 METHODS AND DATA

3.1 Participant Recruitment and Enrollment

Participants were recruited between October 28, 2015 and February 12, 2016 across the United States according to procedures approved by our university's Institutional Review Board. We partnered with Focus Pointe Global, a company that works in study recruitment, to distribute our screening survey to potential participants via email. We selected individuals that were at least 18 years old ($\mu = 38.37$ yrs, $\sigma = 10.25$), owned an Android smartphone (OS 4.4 through 5.1), and had access to WiFi. Individuals with psychotic disorders, inability to walk half a mile, or positive screens for alcohol abuse were excluded from our study.

3.2 Data Collection

Participants were enrolled in the study for six weeks, where they installed *Purple Robot* [14], an open source Android application for recording phone sensor data, and *EMA app*, an EMA administration application, on their devices. Participants were compensated using an incentive-based model depending on how long they remained in the study as well as how many EMA questionnaires they answered, with payouts made at the end of every week. Total compensation per participant ranged from $25 up to $270.40.

Data Privacy. Like all projects utilizing personal information, we were concerned with ensuring participant confidentiality. All users were assigned an anonymized hash, with encryption of sensitive information occurring on the mobile device before being transferred to our secure data servers. As part of our IRB, call transcript records and SMS content were not collected as part of the study. Additionally, because we took a broad interdisciplinary approach with our team spanning computer scientists and psychologists across multiple institutions, data distribution issues had to be considered as well. To make the data shareable across institutions we obfuscated the raw GPS coordinates, only preserving the relative relationship between recorded device locations. All sensor data shared across institutions were linked to participants only through their anonymized hash.

Passive sensing. The Purple Robot app logged call and text information, including whether the communication was incoming, outgoing, or missed as well as total call duration. Multiple communication events as well as EMA surveys were linked to participants across the study period using each participant's unique anonymized hash. Raw latitude and longitude location data was also recorded at one minute intervals. Participants otherwise conducted their daily lives with Purple Robot collecting sensor data in the background.

EMA and survey data. In addition to the passive data collection, participants responded to EMA surveys that asked questions about semantic location as well as the contacts they communicate with. After the conclusion of the first logged communication, participants were prompted by the application to answer questions about the contact. Participants were prompted with the question: "What is your relationship to this person?" Contacts were labelled with one of seven possible options:

- Significant Other
- Friend
- Family Member You Live With
- Family Member You Don't Live With
- Colleague/Work-Related
- Task (e.g. Make an Appointment, Reservation, etc.)
- Other

Additionally, participants answered short questionnaires for each contact on a seven point Likert scale:

- How much did you want to communicate with this person today?
- I would talk to this person about important matters
- I would be willing to ask this person for a loan of $100 or more.
- How close are you to this contact?

These questions are adapted from Wiese et al. [40] as measures of relationship strength. We define a single EMA *tie strength* score by summing the survey responses for each contact as a measure of overall social support the contact provides. The range of possible values for tie strength are $0 - 24$, corresponding to the seven-point Likert scale used for the four survey questions.

For our participants we have 8,948 recorded contacts across 52,850 calls and 353,467 SMS messages. The number of contacts recorded per participant varied, with ranges from a single contact to 159 contacts (25^{th} percentile: 24, 50^{th} percentile: 37, 75^{th} percentile: 58).

3.3 Participant Demographics

A total of 208 participants were recruited. Two participants had invalid GPS data, while 17 participants had an insufficient number of labelled contacts. Our final dataset consisted of 189 individuals: 152 female, 33 male, 4 who did not identify as male or female. 152 participants defined their race as white, 23 as black or African American, 6 as Asian, 5 as American Indian, with 3 participants choosing not to indicate their race. Majority of our participants were employed (119 individuals, 63%), with 36 (19%) unemployed participants, 16 (8%) disabled participants, 4

(2%) retired participants, and 14 (7%) participants not indicating their occupation status. We obtained geographic information through mapping zip codes (one participant had an invalid code). One hundred sixty four participants (87%) lived in a metropolitan area as defined by the USDA Economic Research Service [30], with 24 (13%) in a non-metropolitan area. Thirty seven distinct states were represented in our participant sample, with Pennsylvania (17 participants), California (16), Illinois (13), Florida (13), Texas (11) and Michigan (11) the most common.

3.4 Prediction Target: Top Five Contacts

Table 1. Contact type counts for all contacts and top five contacts by communication frequency per participant. When limiting to only top five contacts, classes that we expect to provide more social support are better represented, namely family members, friends, and significant others.

	All contacts	Top 5 contacts
family live separate	1,232	249
family live together	317	104
friend	1,742	314
other	1,939	87
significant other	289	120
task	2,240	66
work	1,189	73

We examined the distribution of contacts types to ensure we have a well-defined prediction target for our relationship modeling task. When considering all contacts across our data, the most frequent contact types that occur are “other” followed by “task” (see “all contacts” column of Table 1) which is likely due to a high volume of one-off communication events. Indeed, we see in Figure 2 that “other” and “task” contacts have much lower communication volume when compared to the other contact types. Consequently, we excluded “other” labels as a category in our prediction task due to the label being a catch-all for contacts that did not fit clearly into one of the predefined categories, which made the classification task more meaningful.

To further focus our relationship modeling, we chose to restrict our data to each participant’s top five contacts by communication volume. Though social support ground truth is difficult to quantify, we believe that communication volume can be used as an approximate measure for social support – more communication would be indicative of a stronger social connection. We see this relationship in the data (Figure 1), as the contacts that communicate most frequently with the participants also tend to have the greatest EMA-measured *tie strength* (Spearman’s $R = -0.29, p < 0.00001$ between communication frequency rank and tie strength). We chose not to filter contacts based on the tie strength score itself because communication frequency is a commonly recorded passive sensing feature, making this relationship modeling extendable to other mobile phone data studies. We also view tie strength as a related but separate prediction target to be explored in future work. By using communication volume as a selection criteria for contacts, we have a measure that can be applied to other personal sensing data while also providing a proxy of the social tie strength between participants and their contacts.

Our decision to use the top five contacts as the specific cutoff was motivated by results from work in social grouping patterns as well as wanting a clean modeling setup. Prior work in social group organization shows that individuals often have three to five others from whom they seek social support [8, 43], making relationship role estimation of the top five contacts particularly important. Indeed, when we restrict to the top five contacts per participant, “friend” and “family live separate” contact categories become the most frequent (“top 5 contacts” column of Table 1), showing this subset of contacts better represent the contact categories we would expect

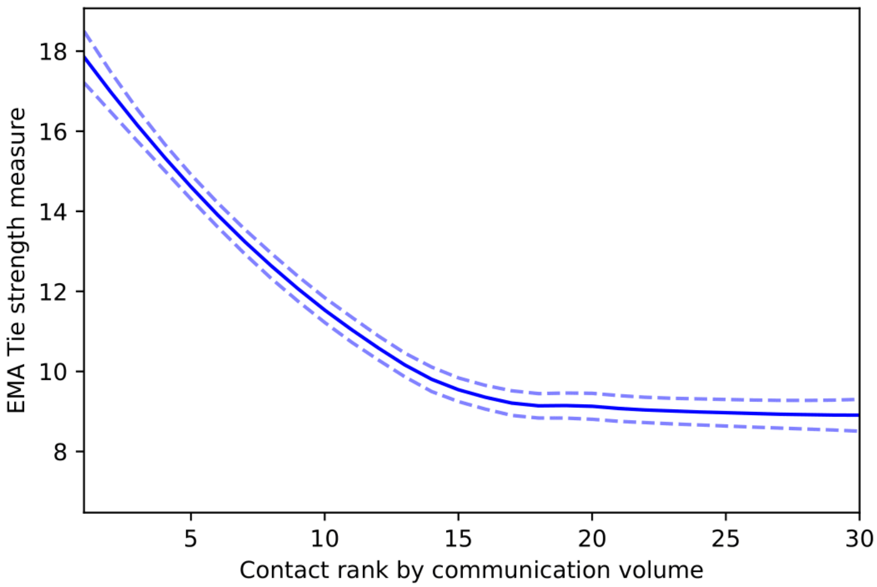


Fig. 1. The relationship between communication volume and tie strength. This plot is generated by fitting a locally weighted regression [6] to all labelled contacts, excluding "other," in our data. The solid line indicates the median fit and the dotted lines indicating 95% confidence intervals. The relative ranking by communication volume for each contact is used to account for variation in overall communication frequency across participants.

to provide more social support. Additionally, removing the less frequent contacts from the data also made the prediction task non-trivial – we see from the class imbalance of "task" labels (Table 1) that a naive model could achieve good performance simply by considering communication volume if all contacts were included as part of the task (Figure 2). Because there is a consistent linear relationship between EMA tie strength and communication volume rank up until the top fifteen contacts (Figure 1), we also evaluated our models on the top fifteen contacts. Though we obtained qualitatively similar results to our top five models (Table A.1), fifteen participants did not have at least fifteen recorded contacts during the study period, so by performing our analysis on the top five contacts we ensure that more participants are evenly represented. With this top five contact design, we have a well-defined prediction task setup that could provide insight into the interactions between participants and their social support group.

3.5 Feature Extraction

To build relationship prediction models, we first needed to derive features from our raw sensor data. We considered a number of distinct feature blocks: communication, demographics, and semantic location.

Communication features. Communication patterns are an important factor in characterizing relationships, so our call and text-based features need to be constructed appropriately. To this end, we produced a comprehensive feature set by deriving 147 communication features based on previous work by Min et al. [27], replicating all of their features except for those relating to SMS length, which is not recorded in our study. We can break these features into four categories corresponding to the particular aspects of communication they capture, as defined in [40] and others (Table 2):

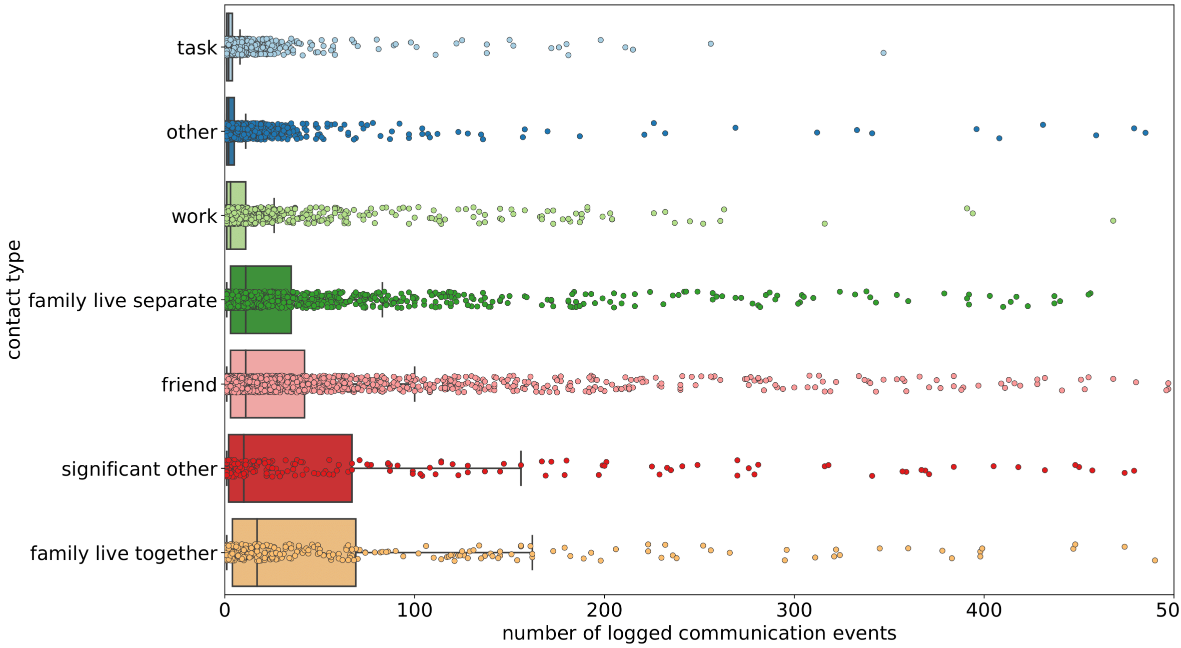


Fig. 2. The relationship between contact type and communication frequency across all contacts. Points overlaid on the box plots are individual contacts in our data. When considering all participants' contacts, we see that the vast majority of "other" and "task" contacts have few logged communication events, illustrating the need for a communication frequency cutoff to model the contacts that truly provide social support for their corresponding participant.

- *intensity and regularity*: features that consider the volume and regularity of communication, such as mean or median weekly calls or number of study days with a logged communication.
- *temporal tendency*: features that capture the time of day or day of week when communication events typically occur.
- *channel selection and avoidance*: features that capture any asymmetry in communication (incoming vs. outgoing) or preference for calls vs. SMS.
- *maintenance cost*: features that consider the amount of effort in maintaining communication relationships.

By considering these aspects of communication, our trained model would be able to capture fine-grained variations in participants' communication tendencies with their contacts.

Demographic features. We also wished to evaluate how demographics impact social relationship inference. Thus, we treated demographic information about each participant as a feature block separate from communication features, which includes age, gender, race, ethnicity, education, marital status, employment status, and whether the participant lived alone or with others (Table 2). All of the categorical responses were encoded in a one-hot representation. Additionally, we hypothesized that the age and gender of a participant is particularly informative of their communication patterns, so we also considered these two features separately.

Semantic location features. To measure the effects of semantic location on our relationship prediction task, we needed to convert the raw GPS data into higher-level features. We used the adaptive k-means clustering

Table 2. Description of extracted features used for relationship prediction. Communication feature categories of *intensity and regularity*, *temporal tendency*, *channel selection*, and *maintenance cost* were derived from Min et al. [27].

Category	Features
demographics	age, gender, education, employment, race, ethnicity, marital status, employment, household size
intensity and regularity	total days of comm., # days of [calls, texts] / days logged, [avg, std] [calls, texts] per day, [avg, min, med, max] call duration
temporal tendency	# [calls, texts] at [time of day, day of week]
channel selection	outgoing comm. / total # comm., # calls / total # comm.
maintenance cost	[calls, texts] within [2, 6] weeks, holiday [calls, texts]
semantic location	[calls, texts] at [home, work, another's home, arts/entertainment, food, nightlife, outdoors/recreation, gym/exercise, professional/medical office, spiritual, travel/transport]
location visit reason	[calls, texts] at [home, work, entertainment, errand, exercise, dining, socialize, travel/traffic]

methodology presented in [34] to accomplish this. First, GPS readings were clustered with a maximum radius of 100 meters. Clusters that had durations of fewer than 10 minutes were discarded to handle transient locations, such as when a participant was stuck in traffic. Participants were then presented a map of the detected location clustered and answered two questions regarding the locations as part of a daily evening survey: “what kind of place is this?" and “why did you visit this place?” Responses were chosen from pre-defined categories, with options such as “home” or “nightlife spot” for the place type and “entertainment” or “errand” for the visit reason (Table 2).

We then cross-referenced the duration of time each participant spent at a location with their communication logs, so that an individual call or text could be labelled as sent or received from a particular location (Figure 3). In total, 175,151 communications (45% of the total communications) were matched to a labelled location. Aggregated counts were normalized over all communication events for each contact, giving us relative proportions of communication events that occurred at a semantic location.

Missing Data Imputation. Due to the passive nature of how phone sensor data are collected, missing data were commonplace and needed to be appropriately handled. Furthermore, it is often the case that data are not missing at random, and the presence or absence of a feature can provide information (e.g., a lack of calls to a contact between 12 am and 4 am). To this end, we not only imputed missing features with the sample population median but also included a new indicator feature for whether the data were missing. This allowed us to capture any potential patterns in how data were missing, which could provide additional information to the predictive models.

3.6 Predictive Modeling Setup and Methodology

Proper cross validation. Because there is shared information between contacts sourced from the same participant, it was critical that we cross-validated over participants as opposed to over contacts to ensure that there is no data leakage. Saeb et al. [35] demonstrated that improper cross validation not only incorrectly inflates performance metrics, but also failed to measure the generalization of a model on true out-of-sample data. In the

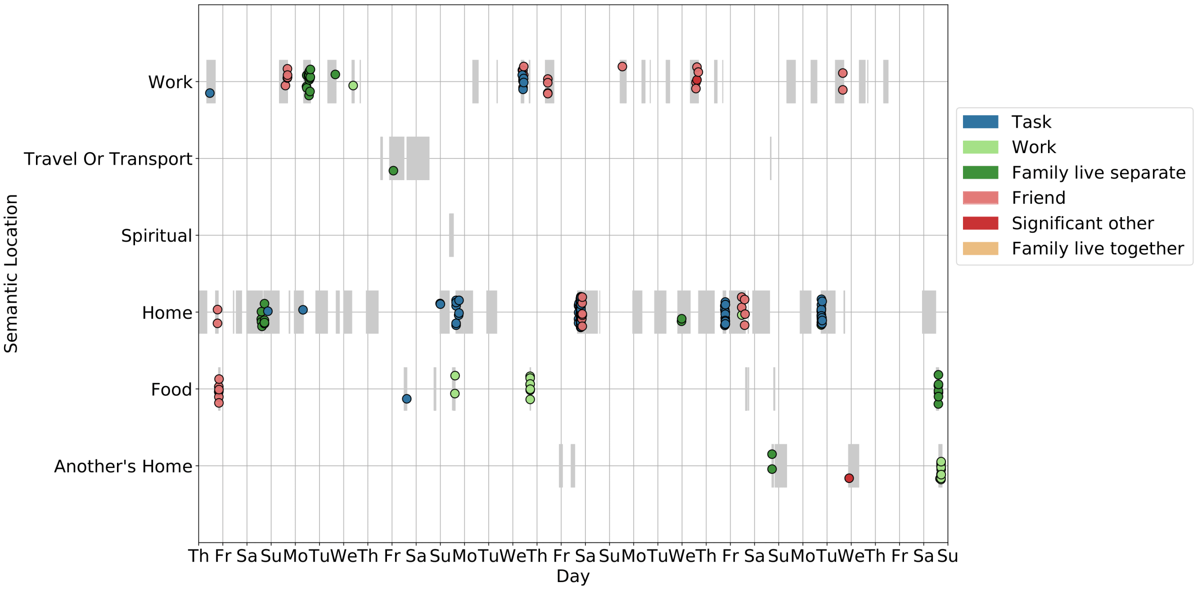


Fig. 3. Semantic location chart over 32 days for a representative participant. Grey regions indicate time spent at a given location while the colored dots represent a communication event. We see regularity in the participant's visits to "home" and "work" locations.

context of relationship prediction, contacts that are associated with a particular participant could have similar communication tendencies based on individual variation as well as identical demographics, which could bias our model's estimation by potentially learning to identify individual participants as opposed to actually learning the characteristics of relationship classes. To this end, we trained our models using *grouped K-fold* cross validation, where all the contacts associated with a participant were included in a single fold and the K folds were defined in terms of groups of participants, with all of a participant's contacts contained within a fold. This cross-validation methodology ensured that our models do not overfit the training data, producing meaningful performance results.

Automated machine learning. In order to avoid overfitting and to apply an objective model construction process to our relationship prediction task, we used *auto-sklearn* [13] for our feature selection, model selection, and hyperparameter tuning. Model building occurred in three stages. First, a meta-learning step selected promising initializations of hyperparameters based on the given data's similarity to existing data. Next, specific algorithms and hyperparameter settings were tuned through Bayesian optimization [3]. Finally, the best models found through this optimization process were used to automatically construct an ensemble model, which is the final model output. Automatic selection of model parameters and ensembling not only tends to give slightly higher performance than manually engineering them, it also prevents user bias from being introduced into the resulting model, such as a preference for one algorithm over another or prior knowledge about the data that can be exploited. Auto-sklearn is feature and algorithm agnostic, which allowed us to essentially treat the model building process as a hands-off task once the feature sets had been constructed.

Feature importance analysis using SHAP. We not only want to produce models with good performance; we also want to interpret the relative importance of the model features. However, because auto-sklearn builds ensembles of various machine learning models, feature weights are not directly accessible. Our solution was to use *SHAP*, which combines numerous additive feature attribution techniques to produce a single “explanation model” [25]. The intuition behind these methods is that they map the original features of a local data point x to a simplified space where the mapped inputs contribute additively to a linear explanation model. These features can then be easily assigned importance based on their contributions to the explanation model, which also describes the actual classifier output for x. By using SHAP, we were able to give more intuitive analyses of our models' performance despite their complexity.

4 RESULTS: RELATIONSHIP PREDICTION ANALYSIS

We used the following feature sets for our relationship prediction:

- age and gender
- communication features
- communication + age/gender features
- communication + demographics + semantic location features

We held out 25% of the participants as an out-of-sample test set, and used accuracy, macro F1 and weighted F1 to evaluate final model performance. Macro F1 is an unweighted average of F1 scores across the predicted classes, while weighted F1 is a weighted average by frequency of those same across-class F1 scores to account for class imbalance. Random forests trained using the auto-sklearn framework served as a baseline comparison to the auto-sklearn ensemble methods. These feature set subdivisions allow us to quantify the effects of different feature modalities on relationship prediction.

4.1 Four-class Relationship Prediction Performance

For a meaningful comparison across relationship roles we must have classes that are sufficiently distinct. We first trained our models for relationship prediction using all six relationship categories: “family live together,” “family live separate,” “friend,” “work,” and “task.” but found that “family live separate” and “friend” contacts are often confused. We hypothesize that instead of predicting the social role, our models were detecting the structure of the relationship (whether the participant lives with the contact or not) rather than the type of the relationship. This could have more impact on when and how participants communicate with contacts than any familial labels: communication patterns between “peer” family members who live separately such as siblings can bear closer resemblance to social relationships than other familial relationships. To this end, we define new labels where the “family live separate” and “friends” labels are collapsed into a single category: “social separate,” and “significant other” and “family live together” labels are collapsed into a single category: “family together.” We note that the six-class prediction class results show similar trends in performance across the blocks of features, which can be found in Supplemental Figure A.2 and Supplemental Table A.2. For the rest of our paper we thus focus on the four-class prediction task.

Our goal is to measure improvement in relationship prediction when adding demographics and location features. Indeed, we find that including these features improves model performance (Table 3). Due to the class imbalance of our data the trivial strategy of always guessing the majority baseline (“social separate”) sets a meaningful lower bound of performance to compare against. A general trend we note is that the auto-sklearn ensemble models perform slightly better than the tuned random forest models, especially when trained with more features. Our best performing model, the auto-sklearn ensemble trained with all of our features, achieves an accuracy of 71% on the held-out test set, a 14 point increase in raw accuracy over the majority baseline prediction. Improvements over baseline are larger in the F1 scores with the ensemble model producing a macro F1 of 0.55 and a weighted F1 of 0.68. Across all three metrics we find that adding demographics and location features considerably improves predictions.

Table 3. Performance metrics for the four-class relationship prediction task. We show the highest scores in bold. Auto-sklearn is used for hyperparameter selection of the "random forest" models, while the "auto-sklearn" models are ensembles of the best performing models found via auto-sklearn's optimization process. Feature blocks correspond to the four feature configurations described above, with "comm" as communication features, "demo" as all demographic features, and "loc" as semantic location features.

model	features	accuracy	macro f1	weighted f1
majority baseline	N/A	0.5714	0.1818	0.4156
random forest	age/gender	0.5714	0.1818	0.4156
random forest	comm	0.6667	0.4795	0.6254
random forest	comm + age/gender	0.6667	0.4750	0.6225
random forest	comm + demo + loc	0.6762	0.4744	0.6326
auto-sklearn	age/gender	0.5714	0.1818	0.4156
auto-sklearn	comm	0.6571	0.4731	0.6195
auto-sklearn	comm + age/gender	0.6905	0.5488	0.6654
auto-sklearn	comm + demo + loc	**0.7095**	**0.5519**	**0.6806**

When evaluating performance, we need to consider which contacts are misclassified. The most common confusion is incorrectly predicting "social separate" (Table 4), which simply happens because this class is so frequent. Despite this, "task" and "family together" are reasonably accurate at about 60%. On the other hand, "work" contacts are usually misclassified as "social separate". We hypothesize that since we only consider the top five contacts of each participants, contacts nominally labelled "work" are often contacts the participant is closer to, akin to a friend. In reality, even the four classes we consider here may be effectively overlapping.

Table 4. Test set confusion matrix for the "auto-sklearn" ensemble model with all features. Contact types are abbreviated work (W), social separate (SS), task (T), and family together (FT), and columns, marked by "p," are predictions. We see reasonable performance on the majority of contact types given the class imbalance present within our data, though most "work" contacts are misclassified.

	pW	pSS	pT	pFT
W	1	15	0	0
SS	0	106	1	13
T	0	7	11	1
FT	0	23	1	31

Breaking our features into four separate blocks also allows us to examine their importance. Across the "auto-sklearn" feature blocks (Table 3), we see the largest increase in performance when adding age and gender to our model with a 3.4% bump in accuracy and 4.5 point increase in weighted F1, while including other demographics and location features on top of age and gender provide much smaller performance improvements. This increase when introducing age and gender features is particularly interesting when we consider the fact that age and gender features alone do not appear to have any predictive value: as shown in Table 3, models trained using only those features do no better than simply guessing the majority class. There seems to be an interaction effect between age/gender and other features.

4.2 Feature Importance

We wish to explore feature interactions with the contact types and understand their fine-grained effects on relationship predictions. We thus use the SHAP framework for evaluating feature importance (see methods). We analyze both individual features and our pre-defined feature blocks. Our analysis will provide a degree of explanation as to why our models made their predictions and also sheds some light on specific communication patterns that characterize particular contact types.

We would like to understand which features contribute to model predictions. All but two of the top fifteen most important features are derived from communication patterns (Figure 4), which is expected given the modest increases in performance when comparing the models trained with only communication features to our models with additional demographic and location features. *Intensity and regularity* as well as *temporal tendency* communication feature modalities have substantial influence on the model output. The most important individual feature, total number of days of communication, is a strong signal for the "family together" and "social separate, " which is an intuitive relationship as this feature encapsulates both the regularity and volume of communication. Through our SHAP analysis, we see that the timing as well as the volume of communication are critical factors in determining the nature of relationships participants have with close contacts.

We want to observe the relative contributions features make to particular relationship classes, even the ones that are sparsely represented such as "work" and "task". The features that contribute most to a "work" classification are texts on Mondays and calls between 8 am and 12 pm (Figure 4). Intuitively, communications during these time periods would be indicative of work-related interactions, demonstrating that communication patterns we would expect to be important for the "work" contact type in fact are the most useful for differentiating it in the data. Regarding "task" contact types, we see that the total number of communications as well as the missed to incoming call ratio are particularly important for the class. Again, these features appeal to our intuition: "task" contacts should communicate in much lower volume with a participant than the other contact types, and the proportion of missed calls from a "task" contact should be much higher perhaps due to the relative lack of importance of the communications or a lack of desire for participants to engage with these contacts. Thus, despite the presence of class imbalance in our underlying data, this method of feature analysis allows us understand which features contribute to the rarer classes.

Certain semantic location features matter when classifying contact types, and we can quantify their effects using SHAP. We see that calls taken at locations where the the visit reason was "errand" and calls taken at shops are important (Figure 4). These two features make the largest contributions to the "family together" class output, a sensible result as communications while running errands or shopping (e.g. for groceries in a common household) will likely be made to contacts that live with the participant. We see that semantic location features contextualize communication events for particular contact types.

4.3 Age Interaction with Communication Patterns

Notably absent among the most important features in our SHAP analysis are demographic features, particularly age. We hypothesize that instead of participant demographics having a direct effect on model output, there is an interaction between age and communication tendencies that contributes to better relationship prediction performance. To investigate this, we examine both the correlation structure between age and communication as well as trends in communication as a function of age across our target relationship types.

Visualizing age and communication patterns. We expect communication patterns to differ across contact types as a function of age. Indeed, we see that the interaction between some of our most important features and age vary depending on the target contact (Figure 5). For "task" contacts, there is a strong negative relationship between age and late night texts while the relationship is much weaker for "social separate" and "family together" contacts. This appeals to our intuition, as communications late at night become less frequent the older the participant

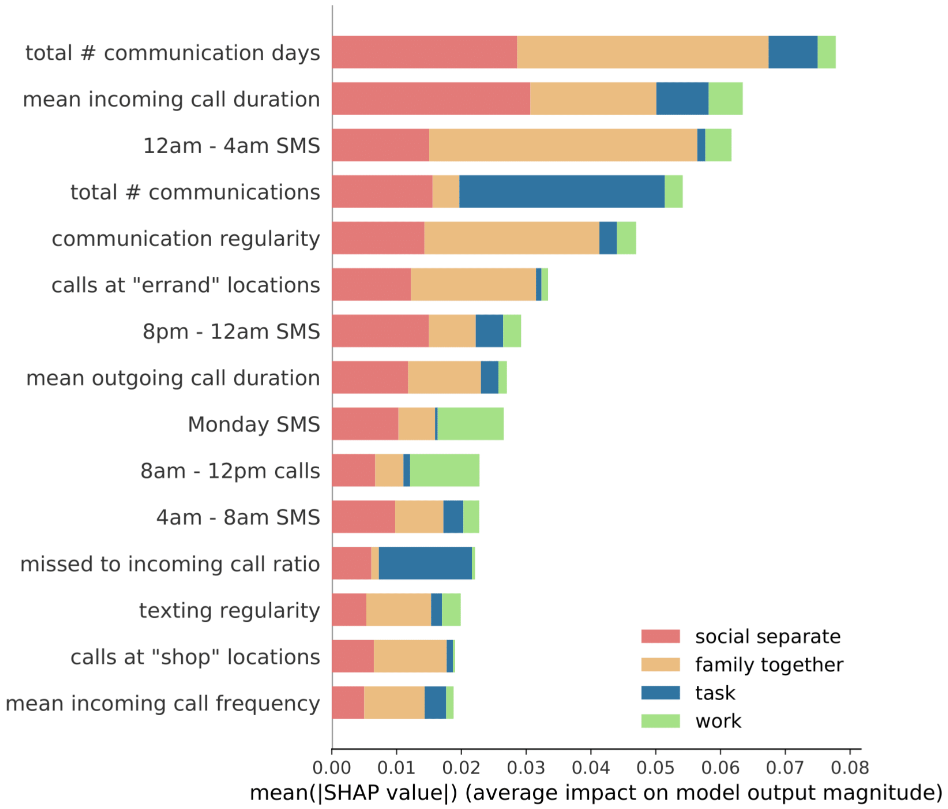
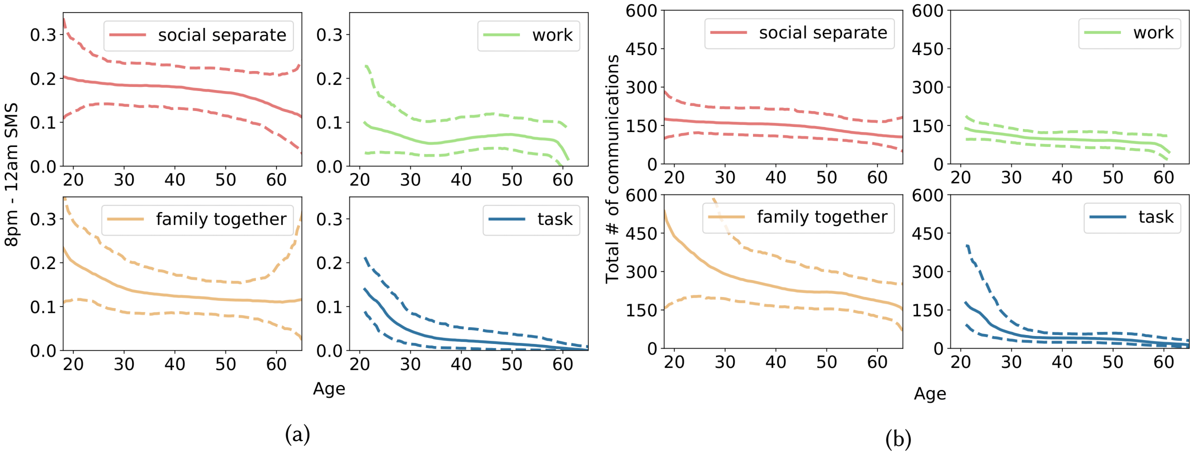


Fig. 4. Top 15 features of the best performing auto-sklearn model ordered by average SHAP value magnitude. We see that features related to communication volume, regularity, and temporality are important. Additionally, we are able to visualize the relative contributions of each feature to a particular contact type output.

is, with the effect being especially strong if the contact is less "close" to the participant, i.e. a task-related communication. We see similar trends in the overall volume of communication as a function of age: there is a sharp decrease in the total number of "task" communications in participants over 30 years old compared to the younger participants, while the volume of communication remains relatively steady across all ages for "social separate" and "work" contacts. Interestingly, there is a higher rate of decrease in communication volume for "family together" contacts, which could be a indication of a shifting preference to call rather than text close family members. For temporal tendency and intensity features, we are able to qualitatively observe substantially different trends in the interaction between age and communication among our different target relationships.

Communication feature correlations. We also attempt to quantify the direct relationships between age and communication patterns. We see several effects of age on *channel selection* and *temporal tendency* within our sample population. As participants get older, they choose to call rather than text, send and receive fewer texts, and tend to communicate less late at night (Table 5). Given both the interaction effects of age and communication among contact types as well as these direct relationships between important features and age, participants across different ages have demonstrably different communication patterns with their contacts in our sample.



(a)

(b)

Fig. 5. Smoothed scatter plots across contact types of age interactions with (a) 8pm - 12 am SMS tendency and (b) total communication volume. We generate these plots by taking 1,000 bootstrap samples from our feature matrix and fitting a locally weighted regression [6]. The solid lines are the median sample, while the dotted lines indicate the 95% confidence interval. We observe different interactions as a function of age across contact types in both features.

5 RESULTS: SUBGROUP GENERALIZATION TASK

Many personal sensing and relationship prediction studies utilize student samples which are skewed towards younger age demographics, raising questions about their generalizability. We have shown in the previous section that participants of different ages have distinct communication patterns, which will likely impact relationship prediction model performance when applied to out-of-sample populations. To quantify this effect, we conduct a prediction task where we split our data into quartiles by age, train models within each quartile, and evaluate performance on the other quartiles. To have consistent comparisons across our relationship prediction tasks, we mirror the best-performing setup on our original relationship prediction tasks presented in Section 4 by utilizing auto-sklearn with all of our derived features (communication, demographics, semantic location) and the same within-participant cross-validation procedure to train models on our four quartile subgroups. This across-group prediction task allows us to explore how age homogeneity may affect generalization.

5.1 Subgroup Data Description and Setup

We need to understand the composition of each of our subgroup populations in order to ensure our prediction task properly evaluates model generalizbility. We initially considered a subdivision where groups were constructed based on whether participants were students, however we lacked the requisite sample size for a meaningful comparison across these groups. We also considered specific age cut-offs (eg participants under the age of 25), but decided that quartile boundaries set according to our data distribution would be the cleanest experimental setup for defining subgroups. The first quartile spans 18 to 31 year-olds, the second quartile spans 32 to 37 year-olds, the third quartile spans 38 to 46 year-olds, and the fourth quartile spans 47 to 66 year-olds (Figure 6). We note that the number of participants and thus contacts are roughly the same in each quartile, so none of the models will have an undue advantage in training set size. Overall, we find that the quartile subgroup splits provide a clean methodology for evaluating relative model performance across different populations.

Understanding the distribution of contact types within each group is also important when examining subgroup model performance. We see that though the proportion of each contact type is approximately the same across Q2,

Table 5. Participant age correlations with communication features. We show significant features with a false discovery rate of <0.01 [2], sorted by correlation magnitude. The most notable trends are that there is a shifting preference from texts to calls and that late night communications become less frequent as age increases.

communication features	corr	p
call tendency	0.181	<0.001
12am - 4am communication	-0.146	<0.001
12am - 4am SMS	-0.145	<0.001
8pm - 12am SMS	-0.141	<0.001
texting regularity	-0.139	<0.001
Sunday SMS	-0.132	<0.001
total # texting days	-0.131	<0.001
communication within last 2 weeks	0.129	0.001
Sunday communication	-0.128	0.001
Max incoming SMS frequency	-0.118	0.002
8pm - 12am communication	-0.116	0.002
standard deviation incoming SMS frequency	-0.111	0.004
call duration within last 2 weeks	0.110	0.005
mean incoming SMS frequency	-0.110	0.005
median incoming SMS frequency	-0.108	0.006
median outgoing SMS frequency	-0.107	0.006
SMS frequency within last 2 weeks	-0.103	0.009
Max outgoing SMS frequency	-0.103	0.009

Table 6. Percentage of relationship labels within each age quartile. Distributions are approximately equal across Q2, Q3, and Q4, while Q1 has roughly 10% more "social separate" contacts.

	family together	social separate	task	work
Q1	20.38%	67.31%	5.00%	7.31%
Q2	25.58%	56.74%	10.70%	6.98%
Q3	24.71%	58.43%	8.63%	8.24%
Q4	21.86%	56.74%	12.09%	9.30%

Q3, and Q4, Q1 has a larger amount of "social separate" labels when compared to the other quartiles (Table 6). To address these relative differences between the proportion of contact types per quartile, labels are resampled so that all classes are equally represented in each cross-validation fold of the training procedure. Additionally, we choose to use macro F1 as our target metric to account for the class imbalance across our quartile samples; by evaluating our models using macro F1, trained models cannot artificially achieve higher out-of-sample performance simply by predicting the majority class. We note that the weighted F1 metrics (Supplemental Table A.3) produces qualitatively the same trends as our subsequent performance analysis that focus on macro F1. These resampling and performance metric choices make the relative performances of each subgroup model more comparable.

Analogous to our question regarding the limits of model generalizability when trained on homogeneous samples, we also hypothesize that models trained on heterogeneous data are able to generalize better across a wider test set population. To this end, we include model runs where we sample twelve participants from each

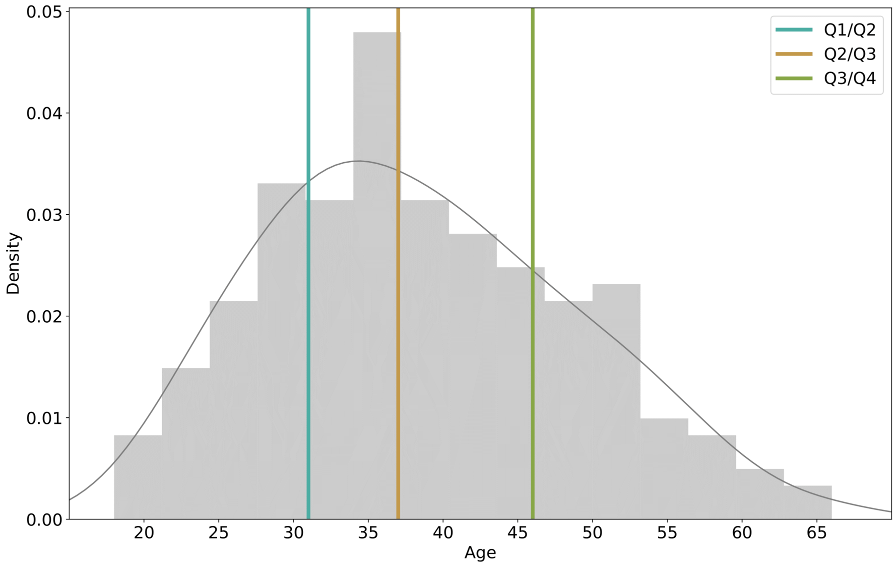


Fig. 6. Histogram of participant ages with quartile cutoffs. As defined by the distribution of participant ages in our sample, the cutoffs are defined at age 31 (Q1 to Q2), 37 (Q2 to Q3), and 46 (Q3 to Q4).

quartile (48 participants total) to produce a heterogeneous training set and evaluate out-of-sample performance on the rest of the data — this "allQ" model result is calculated as the mean performance over ten such samples drawn. The "allQ" model configuration allows us to evaluate the impact of having more varied training data on out-of-sample model performance.

5.2 Subgroup Model Performance

Though we naturally expect lower out-of-sample performance for our subgroup-trained models, we are interested in trends in relative performance across our models that can give insight for the generalizability of particular age quartiles. Indeed, we see that the model trained on the Q1 subgroup generalizes particularly poorly across the other subgroups, producing the worst performance out of all other models for its out-of-sample test subgroups (Table 7). Additionally, the Q2, Q3, and Q4-trained models produce their respective lowest test performance values on the Q1 subgroup. This could be due to the differences in class label distribution of the Q1 subgroup in comparison to the others (Table 6), though our resampling of the training data would mitigate this effect. This could also be due to differences in the distribution of features in the Q1: for example, we have already seen that younger people have higher texting regularity than older people, which could result in the model trained on only Q1 data to over-emphasize texting features, leading to poor out-of-sample performance when those features are not as indicative of relationship labels in the older subgroups. Regardless of the exact source of this discrepancy in performance, what we have seen is that the youngest quartile in our sample is fundamentally different in both feature and relationship label composition than the rest of our sample.

Furthermore, this experiment also illustrates the benefits of having a heterogeneous training sample. Our "allQ"-trained models (last row of Table 7) produce the best average test performance across all four quartiles, and in all subgroups except for Q4. Performance of the overall model is within a standard deviation of the in-sample performance. This result is not entirely surprising, as when given a training set with more variance, models should generalize better to out-of-sample data. Still, the superior performance of the "allQ"-trained models when compared to the others that are trained on more homogeneous data emphasize the standard lessons of machine learning, namely that ML techniques are good at interpolation not extrapolation of data.

Table 7. Performance of models trained on different sample subgroups, optimizing for macro F1. Each row corresponds to a model trained on the indicated age quartile. Within sample results are obtained by 5-fold cross validation, indicated by italics. All-quartile models are trained on an equal number of participants sampled from each quartile, with the sampling procedure conducted ten times to produce the final average performance and standard deviations indicated.

Model	q1 macro f1	q2 macro f1	q3 macro f1	q4 macro f1
majority	0.205	0.174	0.183	0.184
q1-trained model	*0.418*	0.315	0.386	0.349
q2-trained model	0.353	*0.479*	0.426	0.456
q3-trained model	0.312	0.450	*0.483*	0.443
q4-trained model	0.321	0.455	0.397	*0.587*
allq-trained model	0.388 ± 0.038	0.468 ± 0.021	0.425 ± 0.022	0.487 ± 0.041

6 DISCUSSION

We explored the use of demographic and semantic location information in addition to typical communication features on relationship prediction tasks, finding that their inclusion reveals temporal and location tendencies of communication and improves model performance. Additionally, finding significant interactions between demographics, age, and communication patterns, we conducted an age-subgroup model generalization experiment where we highlight both the limits of generalizability when models are trained on homogeneous samples as well as the benefits of training on heterogeneous ones. The use of automatic machine learning techniques throughout these tasks limited manual intervention in the model building and prediction process, a practice that we believe will encourage a more standardized methodology for tackling machine learning problems.

6.1 Contributions

Connections between Well-being and Communication. Our study has built upon previous work by including demographic information and semantic location context when training the relationship prediction models. These additions give intuitive insights on communication patterns across particular social relationships: participants tend to call family members they live with while running an errand, while older participants tend to communicate with non-family members less frequently late at night. These tendencies could inform when digital interventions would be most effective – mobile services designed to promote help-seeking via text message [22] could target particular times of day for interventions or make suggestions of specific contacts to reach out to. For example, previous work by Wilson et al. [41] has shown that intention to seek help from family is associated with seeking mental health care in young adults, so the ability to infer contact types could be used to potentially connect patients to the right contacts for direct intervention, and to improve the quality of care.

Results from this work inform future studies aimed at measuring the amount of communication between people and their closest contacts, which can serve as a marker of mental health [12]. Such communication between contacts is expected to vary as, for example, depression treatments succeed or fail [38]. Considering social network interventions have been found to be effective in reducing social isolation among individuals with severe mental health conditions [32, 39], cell-phone based passive sensing of communication patterns can inform such technology mediated interventions. In addition to being useful for digital interventions, cell-phone based measures of communication to close contacts could eventually become part of evaluating the efficacy of treatment of mental illness, along with standard self-report measures.

Recommendations for Study Design. Our experiences when evaluating model performance has led to lessons learned and recommendations for future personal sensing work. We find that adding demographics and location information increases prediction performance compared to using solely communication-based features, which highlight the promise of these feature modalities in future personal sensing studies as demonstrated by Saeb et al. [34]. Though there is previous work achieving higher accuracy on similar relationship tasks [1, 5, 27], our focus on estimating relationships for participant's closest contacts (to provide insights on the contacts that provide the most social support) necessarily makes our prediction task more difficult. Other studies also consider subsets of contacts in their prediction tasks such as Min et al.'s 70-contact list [27], but the lack of a consistent prediction framework makes it difficult to evaluate results across different works. Our work illustrates that greater granularity of sensor data and standardized evaluation settings are needed to infer meaningful mental health and well-being properties. This calls for the collection and availability of larger, open datasets in the personal sensing domain with well-defined objectives to enable cross-study comparisons that assess their benefits for understanding human behavior.

We highlighted the need to consider population heterogeneity through our age quartile subgroup task. Within our data, training on the youngest population produces particularly poor results, illustrating differences in communication patterns across age demographics that can impact model performance. Furthermore, we demonstrated that training models on more varied samples produces superior performance in comparison to models trained on homogeneous samples. Given the large proportion of studies that rely on younger (e.g. student) populations in the personal sensing space, our subgroup experiments indicate that it is critical for studies that aim to evaluate their model's effectiveness "in the wild" to be mindful of their participant composition.

We also want to emphasize the use of automated machine learning (*auto-sklearn*). Though auto-sklearn is not a silver bullet to all machine learning problems, it does provide a reasonable benchmark of other models as demonstrated both in prior work [13] and in our relationship prediction task by considering a wide gamut of popular machine learning models. It is difficult to evaluate methods across studies when the model choices are different (and unclear). Furthermore, though the introduction of novel modeling techniques can be valuable contributions, baseline comparisons provide little insight if they are not consistent — one study may use a Naive Bayes classifier as a baseline while another may use random forests, drawing different conclusions on model effectiveness. The adoption of automated machine learning as a benchmark would address these concerns, as studies that make data and feature contributions would have a consistent model methodology while studies that make novel method contributions would have an objective performance baseline to measure against. Replacing the usual workflow of ML scientists analyzing data which comes at a risk of overfitting also promises to make us more confident that our data will generalize. We envision that future work in the personal sensing space uses automated machine learning as a baseline modeling choice, enabling meaningful comparisons across studies.

6.2 Limitations and Future Work

Our study has some limitations. Though our dataset is larger than what is typical within the personal sensing community, our sample size is likely still insufficient to significantly demonstrate applicability and robustness of models to a wider population. Furthermore, our data is skewed in terms of gender, as 85% of our participants are female. If there are trends in communication patterns across relationships that vary based on gender, our trained models would be ill-equipped to transfer to populations that have a greater proportion of other genders. However, we are able to quantify the effects of homogeneity within age brackets given our sample's varied age distribution, and we look forward to future work where even larger, more heterogeneous samples are considered for studying transferable relationship estimations.

Additionally, our choice to predict relationships by classifying discrete contact labels may not be the most appropriate task for capturing the nature of social support. To the best of our knowledge, we did not find a consensus in the delineation between relationship classes in the literature, which makes comparisons across studies difficult: for example, we treat significant-others as family as do Min et al. [27], while Bao et al.'s CommSense [1] treat these relationships as distinct classes. We have found that among a participant's closest contacts by communication volume, it is difficult to differentiate "work" contacts from "social" contacts, which could also be an issue of ground truth labelling as a contact could be in multiple relationship categories (our EMAs did not provide the option for multiple labels for a contact). This is an indication that the communication tendencies with these nominally different relationship types are similar — colleagues can be close friends as well. An alternative metric for evaluating social support could be estimating *tie strength* directly as illustrated by Wiese et al. [40], though they note limitations in using solely call and text features: lack of call/message communication does not necessarily indicate weaker tie strength due to the use of alternative channels such as WhatsApp or Facebook as well as demographic differences mediate different communication patterns. Still, with broader data collection across different digital communication channels as well as the contextualization of communication with demographics as we have demonstrated, further work for social support evaluation through tie strength shows promise.

We also see an opportunity in improving model performance where features are derived from disparate data sources, which is common in the personal sensing field. By concatenating all sensor modalities into a single feature vector, we are losing information about data provenance: semantic locations are fundamentally different than communication frequency and should be treated and regularized accordingly, yet they are analyzed in a homogeneous block. Even subcategories such as *maintenance cost* and *temporal tendency* communication features carry different semantic meanings which could be useful for model creation if treated as separate blocks. Future work could see applications of feature grouping techniques such as group lasso [26] given the blocked nature of the sensor features common in phone data.

7 CONCLUSION

Though personal sensing models show promise for interpersonal relationship estimation on smaller samples, the robustness of such models needs to be validated for wider populations. Our study has demonstrated both the benefits of contextualizing communication patterns with additional demographic and location information as well as the effect of age homogeneity and heterogeneity on model performance. More broadly, we have underscored the need for future personal sensing studies to consider data across all demographics as well as objective model evaluation techniques in order to have meaningful, actionable results.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their helpful comments and suggestions. This research was supported by the National Institute of Mental Health, Grant No. R01 MH111610.

REFERENCES

[1] Xuan Bao, Jun Yang, Zhixian Yan, Lu Luo, Yifei Jiang, Emmanuel Munguia Tapia, and Evan Welbourne. Commsense: Identify social relationship with phone contacts via mining communications. In *2015 16th IEEE International Conference on Mobile Data Management*, volume 1, pages 227–234. IEEE, 2015.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300, 1995.

[3] Eric Brochu, Vlad M Cora, and Nando De Freitas. A tutorial on bayesian optimization of expensive cost functions, with application to active user modeling and hierarchical reinforcement learning. *arXiv preprint arXiv:1012.2599*, 2010.

[4] Luca Canzian and Mirco Musolesi. Trajectories of depression: unobtrusive monitoring of depressive states by means of smartphone mobility traces analysis. In *Proceedings of the 2015 ACM international joint conference on pervasive and ubiquitous computing*, pages 1293–1304. ACM, 2015.

[5] Jinhyuk Choi, Seongkook Heo, Jaehyun Han, Geehyuk Lee, and Junehwa Song. Mining social relationship types in an organization using communication patterns. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 295–302. ACM, 2013.

[6] William S Cleveland. Robust locally weighted regression and smoothing scatterplots. *Journal of the American statistical association*, 74(368):829–836, 1979.

[7] Orianna DeMasi, Konrad Kording, and Benjamin Recht. Meaningless comparisons lead to false optimism in medical machine learning. *PloS one*, 12(9):e0184604, 2017.

[8] Robin IM Dunbar and Matt Spoors. Social networks, support cliques, and kinship. *Human nature*, 6(3):273–290, 1995.

[9] Rahul Dwarakanath, Jérôme Charrier, Frank Englert, Ronny Hans, Dominik Stingl, and Ralf Steinmetz. Analyzing the influence of instant messaging on user relationship estimation. In *2016 IEEE International Conference on Mobile Services (MS)*, pages 49–56. IEEE, 2016.

[10] Nathan Eagle and Alex Sandy Pentland. Reality mining: sensing complex social systems. *Personal and ubiquitous computing*, 10(4):255–268, 2006.

[11] Nathan Eagle, Alex Sandy Pentland, and David Lazer. Inferring friendship network structure by using mobile phone data. *Proceedings of the national academy of sciences*, 106(36):15274–15278, 2009.

[12] Frank J Elgar, Wendy Craig, and Stephen J Trites. Family dinners, communication, and mental health in canadian adolescents. *Journal of Adolescent Health*, 52(4):433–438, 2013.

[13] Matthias Feurer, Aaron Klein, Katharina Eggensperger, Jost Springenberg, Manuel Blum, and Frank Hutter. Efficient and robust automated machine learning. In *Advances in Neural Information Processing Systems*, pages 2962–2970, 2015.

[14] Center for Behavioral Intervention Technologies. Purple robot. https://tech.cbits.northwestern.edu/purple-robot/, 2015.

[15] Genevieve Gariepy, Helena Honkaniemi, and Amelie Quesnel-Vallee. Social support and protection from depression: systematic review of current findings in western countries. *The British Journal of Psychiatry*, 209(4):284–293, 2016.

[16] Linda K George, Dan G Blazer, Dana C Hughes, and Nancy Fowler. Social support and the outcome of major depression. *The British Journal of Psychiatry*, 154(4):478–485, 1989.

[17] Isha Ghosh and Vivek K Singh. Modeling social support scores using phone use patterns. *Proceedings of the Association for Information Science and Technology*, 55(1):133–142, 2018.

[18] Jennifer L Hames, Christopher R Hagan, and Thomas E Joiner. Interpersonal processes in depression. *Annual review of clinical psychology*, 9:355–377, 2013.

[19] Gabriella M Harari, Sandrine R Müller, Min SH Aung, and Peter J Rentfrow. Smartphone sensing methods for studying behavior in everyday life. *Current Opinion in Behavioral Sciences*, 18:83–90, 2017.

[20] Hsun-Ping Hsieh and Cheng-Te Li. Inferring social relationships from mobile sensor data. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 293–294. ACM, 2014.

[21] Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. Predicting students' hapiness from physiology, phone, mobility, and behavioral data. In *Affective computing and intelligent interaction (ACII), 2015 international conference on*, pages 222–228. IEEE, 2015.

[22] David Joyce and Stephan Weibelzahl. Student counseling services: Using text messaging to lower barriers to help seeking. *Innovations in education and teaching international*, 48(3):287–299, 2011.

[23] Nicholas D Lane, Emiliano Miluzzo, Hong Lu, Daniel Peebles, Tanzeem Choudhury, and Andrew T Campbell. A survey of mobile phone sensing. *IEEE Communications magazine*, 48(9), 2010.

[24] Jin Lu, Chao Shang, Chaoqun Yue, Reynaldo Morillo, Shweta Ware, Jayesh Kamath, Athanasios Bamis, Alexander Russell, Bing Wang, and Jinbo Bi. Joint modeling of heterogeneous sensing data for depression assessment via multi-task learning. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):21, 2018.

[25] Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, pages 4765–4774, 2017.

[26] Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

[27] Jun-Ki Min, Jason Wiese, Jason I Hong, and John Zimmerman. Mining smartphone data to classify life-facets of social relationships. In *Proceedings of the 2013 conference on Computer supported cooperative work*, pages 285–294. ACM, 2013.

[28] Seyed Hamid Mirisaee, Saman Noorzadeh, Ashkan Sami, and Reza Sameni. Mining friendship from cell-phone switch data. In *2010 3rd International Conference on Human-Centric Computing*, pages 1–5. IEEE, 2010.

[29] David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

[30] United States Department of Agriculture Economic Research Service. Rural-urban continuum codes. https://www.ers.usda.gov/data-products/rural-urban-continuum-codes.aspx. Accessed: 2019-07-25.

[31] Sun Young Park. Social support mosaic: Understanding mental health management practice on college campus. In *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 121–133. ACM, 2018.

[32] Eris F Perese and Marilee Wolf. Combating loneliness among persons with severe mental illness: Social network interventions'characteristics, effectiveness, and applicability. *Issues in mental health nursing*, 26(6):591–609, 2005.

[33] Delphine Reinhardt, Franziska Engelmann, Andrey Moerov, and Matthias Hollick. Show me your phone, i will tell you who your friends are: analyzing smartphone data to identify social relationships. In *Proceedings of the 14th International Conference on Mobile and Ubiquitous Multimedia*, pages 75–83. ACM, 2015.

[34] Sohrab Saeb, Emily G Lattie, Konrad P Kording, and David C Mohr. Mobile phone detection of semantic location and its relationship to depression and anxiety. *JMIR mHealth and uHealth*, 5(8), 2017.

[35] Sohrab Saeb, Luca Lonini, Arun Jayaraman, David C Mohr, and Konrad P Kording. Voodoo machine learning for clinical predictions. *Biorxiv*, page 059774, 2016.

[36] Michelle P Salyers and Kim T Mueser. Social functioning, psychopathology, and medication side effects in relation to substance use and abuse in schizophrenia. *Schizophrenia research*, 48(1):109–123, 2001.

[37] Edward A Suchman. An analysis of" bias" in survey research. *Public Opinion Quarterly*, pages 102–111, 1962.

[38] Lisa A Uebelacker, Emily S Courtnage, and Mark A Whisman. Correlates of depression and marital dissatisfaction: Perceptions of marital communication style. *Journal of Social and Personal Relationships*, 20(6):757–769, 2003.

[39] Martin Webber and Meredith Fendt-Newlin. A review of social participation interventions for people with mental health problems. *Social psychiatry and psychiatric epidemiology*, 52(4):369–380, 2017.

[40] Jason Wiese, Jun-Ki Min, Jason I Hong, and John Zimmerman. You never call, you never write: Call and sms logs do not always indicate tie strength. In *Proceedings of the 18th ACM conference on computer supported cooperative work & social computing*, pages 765–774. ACM, 2015.

[41] Coralie J Wilson, Debra J Rickwood, John A Bushnell, Peter Caputi, and Susan J Thomas. The effects of need for autonomy and preference for seeking help from informal sources on emerging adults' intentions to access mental health services for common mental disorders and suicidal thoughts. *Advances in Mental Health*, 10(1):29–38, 2011.

[42] Chen Yu, Namin Wang, Laurence T Yang, Dezhong Yao, Ching-Hsien Hsu, and Hai Jin. A semi-supervised social relationships inferred model based on mobile phone data. *Future Generation Computer Systems*, 76:458–467, 2017.

[43] W-X Zhou, Didier Sornette, Russell A Hill, and Robin IM Dunbar. Discrete hierarchical organization of social group sizes. *Proceedings of the Royal Society B: Biological Sciences*, 272(1561):439–444, 2005.

A SUPPLEMENTAL TABLES

Table A.1. Test set performance metrics for top fifteen contacts by communication volume. We see that overall performance numbers are comparable to the top five results (Table 3), and that the model trained solely on the top five contacts generalizes well to the top fifteen as it achieves nearly the same performance as the model trained on the top fifteen contacts.

	accuracy	balanced accuracy	macro f1	weighted f1
majority baseline	0.5617	0.2500	0.1798	0.4040
top 5 trained model (all features)	0.6500	0.4766	0.4644	0.6090
top 15 trained model (all features)	0.6567	0.4623	0.4861	0.6152

Table A.2. Performance metrics for the 6 class relationship prediction task with highest scores in bold. Feature blocks correspond to the configurations described above, with “comm” as communication features, “demo” as all demographic features, and “loc” as semantic location features.

model	features	accuracy	balanced accuracy	macro f1	weighted f1
majority baseline	N/A	0.3381	0.1667	0.0842	0.1709
Random forest	comm	0.4000	0.2994	0.2899	0.3559
Random forest	comm + age/gender	0.4476	0.3543	0.3627	0.4123
Random forest	comm + demo + loc	0.4333	0.3733	0.3577	0.4106
auto-sklearn	comm	0.4524	0.3617	0.3534	0.4113
auto-sklearn	comm + age/gender	**0.4619**	0.4150	**0.4294**	**0.4523**
auto-sklearn	comm + demo + loc	**0.4619**	**0.4178**	0.4261	0.4517

Table A.3. Performance of models trained on different sample subgroups, optimized for weighted F1. Each row corresponds to a model trained on the indicated age quartile. Within sample results were obtained by 5-fold cross validation and are indicated by italics. All-quartile models are trained on an equal number of participants sampled from each quartile, with the sampling procedure conducted ten times to produce the final average performances.

Model	q1 weighted f1	q2 weighted f1	q3 weighted f1	q4 weighted f1
majority	0.550	0.411	0.395	0.359
q1-trained model	*0.640*	0.507	0.569	0.519
q2-trained model	0.642	*0.635*	0.567	0.585
q3-trained model	0.658	0.648	*0.634*	0.623
q4-trained model	0.633	0.597	0.610	*0.617*
allq-trained model	0.643 ±0.014	0.610 ±0.042	0.574 ±0.017	0.585 ±0.034