Project Report
Gender Disparity Signals:
Analyzing gender disparities with mobile phone metadata

Principal Investigator: Muhammad Raza ur Rehman Khan mraza@berkeley.edu
Advisor: Joshua Blumenstock (jblumenstock@berkeley.edu)

SUMMARY

Ending gender discrimination in all its forms has long been a sustainable development goal of the United Nations. Despite initiatives and efforts of different governments across the world, however, gender discrimination still exists. One of the biggest problems has been the lack of accurate information about gender disparity. This project aims at modeling educational gender disparity at the district level in Pakistan using network data extracted from call detail records. Using survey-based educational gender disparity as ground truth data, we first explore the prominent network metrics and patterns that help in better understanding of gender disparities in the population. In the second half of this project, we use these network metrics to predict educational gender disparity at the district level.

Our model uses a dataset of more than 30 million customers, advanced network features and prediction algorithms to model educational gender disparities accurately. Our findings show that call detail records can be effectively used to analyze social networks of men and women in the developing world as an alternative data source to more expensive forms of data collection. Secondly, our analysis shows that the men and women of developing countries like Pakistan manifest significant differences in their social network activities and patterns. Lastly, the predictive model that we have developed enables a relatively accurate and cheaper way of estimating gender disparities.

Our proposed model can be easily applied to other countries, provided gender-annotated call detail records are present.

1. Problem Statement

Getting accurate data about gender disparities can be a challenging task in many countries. There exist countries in the developing world where there has been a gap of more than a decade between consecutive censuses. As a result, getting accurate demographic information about women and gender disparities can be the first hurdle for the researchers working on gender-related issues. Researchers working on gender disparities have primarily relied on the data collected through surveys. Survey data may fit the requirements in many cases, but population-level surveys can be expensive and hard to manage, and thus researchers have been actively using alternative data sources in their research.

Internet-based social networks like Facebook, Twitter and Google+ can go a long way toward solving this problem in the developed world, but in many developing countries the penetration of these social networks remains low. For instance, Magno and Weber(2014) have tried to analyze gender disparity using Twitter and Google+ datasets, but they found that in some countries like Pakistan there can be quite some disparity between the ratio of male and female users on the social network as compared to the ratio of males and females in the actual population [1]. Because of this issue, research based on these sources can result in analytic discrepancies, as women in developing countries having regular access to these social networks may already be more privileged than the random male counterparts of the society (Jackie Robinson effect [1]). Thus, there is a need for extensive research on gender issues at the population level using more comprehensive (conventional or non-conventional) data sources.

Mobile phones have seen good penetration in the developing world. As a matter of fact, in many developing countries, mobile phone networks have higher penetration than financial institutes like banks and Internet-based social networks [2]. High penetration of mobile phones in developing countries makes them an ideal source to passively collect information about the mobility and behavioral patterns of individuals. These patterns have been used by researchers to analyze poverty [3][4], unemployment [5], and migration [6]. However, even with the popularity of the mobile phones in the developing world, the analysis of gender disparities is complicated by two factors: 1) lack of gender information for each of the subscribers; and 2) lack of ground-truth gender disparities data at fine resolution.

In this research, we use gender-annotated call detail records (CDRs) data from a major operator in Pakistan to analyze how gender disparities manifest themselves in social networks. Furthermore, the mobility information present in the CDRs enables us to associate these social networks with district boundaries. We use educational gender parity data, available at the district level, to ground truth the CDR-derived conclusions. The overall goal is to ascertain how much of the educational gender disparity can be accurately inferred from the different type of features learned through the CDR data. This research builds on our exploratory work presented in the ICTD 2015 [7].

2. Research Questions

The main questions that we intend to address in this research include:

1) Research Question1 (RQ1): To what extent are gender disparities reflected in the social networks extracted from the call records?

In other words, do special social network features exist which can tell us about the gender disparities in a district? A lot of research work has been done on the analysis of social networks of less privileged populations. Examples of this work include the analysis of social networks of migrants [8] and the impact of collaboration on the success of individuals [9], but similar analysis has not been done to compare the social networks of women and men. We explore through this question whether prevailing concepts in the literature on social network analysis are valid for gender-annotated networks extracted from the call detail records of developing countries.

2) Research Question2 (RQ2): How accurately can educational gender parity be modeled using the features extracted from the call detail records?

Our second goal in this research is to accurately model gender educational parity through the social network metrics extracted during the analysis for Research Question 1. The first research question is our attempt to better understand the correlations between the social network-related metrics (or features) and gender disparities in society. The second research question aims to develop an accurate predictive model that can help in predicting educational gender disparity in different districts of Pakistan. Both research questions are related in the sense that training the predictive model on high-quality features is a key to the high performance of the model. The predictive quality of the features can be measured in different ways, but simply the features for which males and females have significantly different patterns are expected to have higher predictive power as compared to other features.

3. DATA DESCRIPTION

We used two main data sources for this research:

1) **Communication metadata:** We extract social networks from the Call Detail Records (CDRs) of a major telecom provider in Pakistan. The data consists of more than 1 billion transactions (voice and text messages) around 30 million users and. In addition to the anonymized caller and recipient ids, CDRs also contain the timing of the activity and the location of the cell tower through which the call was made. Furthermore, gender and age of each of the subscribers are also provided by the telecom operator. The CDR data is quite rich in information, as it can be used to collect metrics related to the usage of the network, mobility of individuals and temporal characteristics of the user's network. Summary statistics of the CDR data used in this project are as follows:

CDR Data Characteristics	
Property	**Value**
Male users	~5.51 Million
Female users	~0.57 Million
Number of days	7
Total calls + SMS	~1.07 Billion
Total districts covered	93

Pakistan's Demographic Indicators	
Socio-Economic Indicator	**Value**
Population	185 Million
GDP per capita	$ 4,619
Human Development Index	0.58
Gender Gap Index	0.559
Mobile Subscribers	130 Million

Table 1 Summary statistics of the data

The CDR data that we have is quite rich as, in addition to the information about the caller and the recipient IDs, it also contains information about the cell tower used by the individuals. This information can be used to calculate different mobility related metrics.

Though the CDR data used in this research only spans seven days, it covers 93 out of 128 districts, capturing variation in population density and human development index (Figure 1). The time span of the CDR data does not coincide with any of the major national holidays or any weather-related catastrophe.

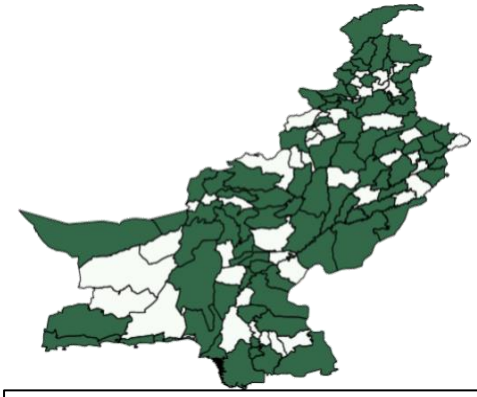

Figure 1 Distribution of districts covered by the CDR Data *(Data was available for the districts colored as green)*

2) **Educational Gender Disparity Data:** To augment the CDR data with the ground truth data about educational gender parity, we use the data collected by the gender advocacy group Alif Ailaan[1]. This dataset contains statistics about the district wise educational gender parity score calculated as a ratio of the net primary enrolment rate of girls to the net primary enrolment rate of boys. Net primary enrolment rate in primary education is the number of pupils of official primary school age who are enrolled in primary education as a percentage of the total children of the official school-age population[2].

Distribution of educational gender parity across different districts of the country is shown in Figure 2 (A). It is clear from Figure 2 (A) that most of the districts with higher gender parity scores are in the northwestern part of the country. The distribution of educational gender parity is positively correlated with the population density of the districts, as shown in Figure 2 (B).

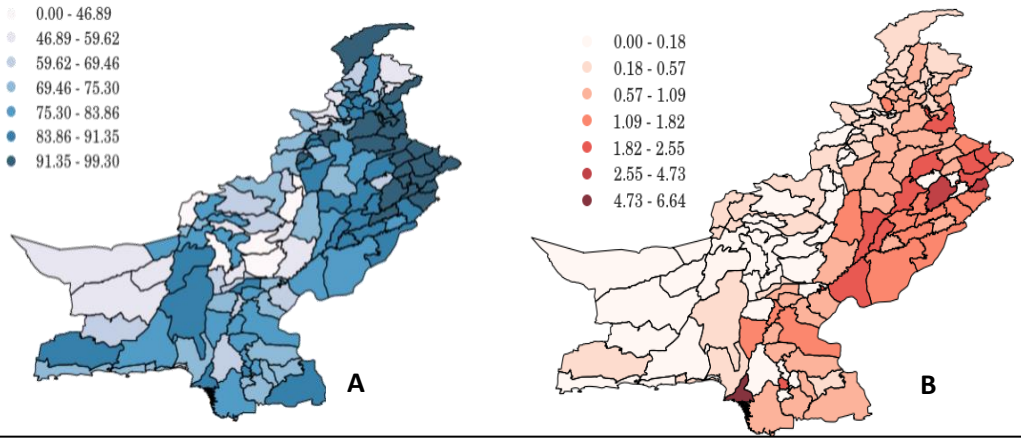

Figure 2 (A) Distribution of educational gender parity. (B) Distribution of district wise population density showing population of the district as percentage of the country's population

[1] http://www.alifailaan.pk

[2] http://uis.unesco.org/node/334718

DATA PREPROCESSING

The CDR dataset contains the details of calls and SMS of millions of customers from a major telecom operator in Pakistan. The general format of the call detail records is as shown in Table 2.

Caller-ID	Recipient-ID	Date	Time	Duration	Caller Cell	Recipient Cell	…
9aes8cd	939b87	2014-01-04	22:00:11	42	1619	1618	

Table 2 General format of CDR data

The networks corresponding to calls and messages can be analyzed separately or jointly; the rest of the analysis in this report corresponds to the joint network, such that the edge between two nodes or subscribers can either represent a call or a message. Furthermore, all the edges are considered undirected. We did explore the call and SMS networks separately with directed edges, but the results were inferior as compared to the network corresponding to the undirected calls and SMS. One possible reason for the better performance of the merged (calls + SMS) undirected network is that such a network is denser as compared to the calls only or the SMS only network.

The location information present in the CDRS (Caller Cell and the Recipient Cell) is useful for calculating different mobility related features as described in the next section.

In addition to the CDR transactions, we also had information about the gender and age of each of the subscriber. This information, coupled with movement patterns, helped us to analyze the social networks along different patterns related to the age, gender and mobility of the ego node (the user whose properties are being analyzed) and the alter (the set of users connected to the ego node).

4. Social Network Analysis

The popularity of social networks like Twitter and Facebook has resulted in an increased research focus on social network analysis, but there is no consensus around ways to generate features, or which features are most important. Many times, the choice of the features used in analysis depends on the personal preferences of the researchers. Instead of relying on a few hand-selected features, we have tried to analyze the social networks of men and women in Pakistan through the lens of as many features as possible, as explained in the following subsections.

Network Activity

Number of Calls and SMS

It is a general perception that underprivileged communities and individuals may be less active in call and SMS networks. However, different researchers have found evidence to support or contradict this assumption. Friaz-Martinez et al. (2010) found that women in the developing country of their study were more active than men on mobile phone networks [10]. Our analysis also confirms this trend, as shown in Figure 3.

In Figure 3, the y-axis shows the number of subscribers (men (green) or women (red)) making a given number of calls or SMS over the 7-day period, as shown by the x-axis. The average number of network transactions made by men and women is also shown through dotted green and red lines respectively.

Figure 3 also shows that the percentage of women making a higher number of calls is greater than the percentage of men.

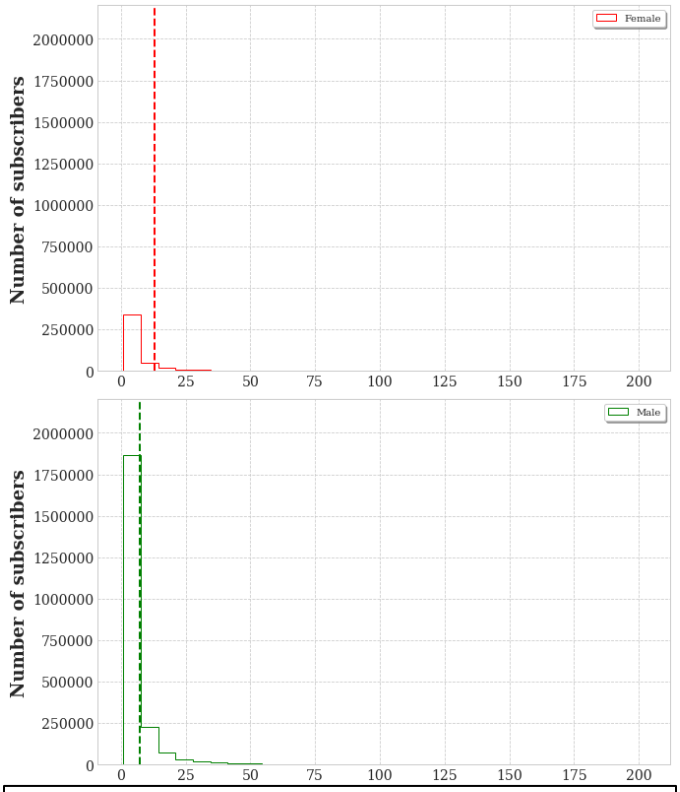

Figure 3 Network activity for females (top panel) and males (bottom panel). *Dotted lines show the averages.*

Network Status

In addition to network activity, the status of the node in the network can be approximated through measures like network size, embeddedness and centrality.

Network Size (Degree Centrality)

The comparison of the network size for women and men is shown in Figure 4. Dotted lines in this figure showing the averages indicate that males in Pakistan have higher degrees on the average as compared to females. This observation is a contrast with the findings of Friaz-Martinez et al. (2010) [1].

This trend shows that women in Pakistan, in general, have a smaller number of contacts, but as shown in Figure 3 the average number of calls and messages for women is higher. Larger network size of males is consistent with social trends in Pakistan, as the number of working men is much higher as compared to working women, and working individuals are expected to have a higher number of contacts.

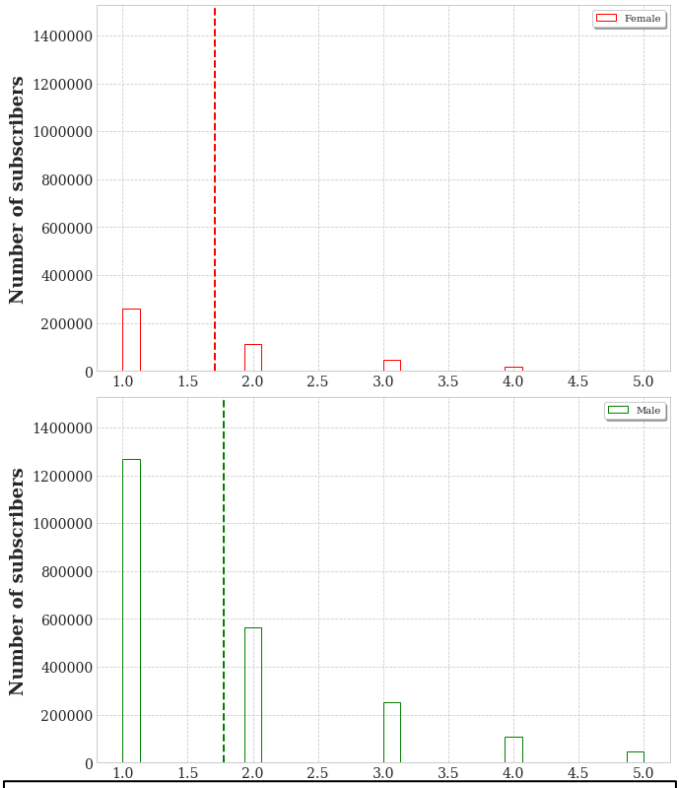

Figure 4 Degree Centrality for females (top panel) and males (bottom panel). *Dotted lines show the averages.*

Network Embeddedness

Embeddedness describes the degree to which the ego nodes are enmeshed in their networks [2]. In other words, embeddedness is the measure of the extent to which the contacts of a node i and the friend of the node i are connected to each other.

Embeddedness of the ego node i can be defined as follows [9]

$$\frac{\sum_{v \in V(i)} \left|\left(V(i) \cap V(v)\right)\right| / \left|\left(V(i) \cup V(v)\right)\right|}{V(i)}$$

Here, $V(i)$ represents the list of the neighboring nodes of the ego node i.

Women in Pakistan have higher average embeddedness as compared to men as shown by the dotted lines in Figure 5. One possible explanation for this trend can be the fact that women have more responsibilities within the house and the family while the men have more responsibilities out of the home. In other words, a woman's network may be largely comprised of

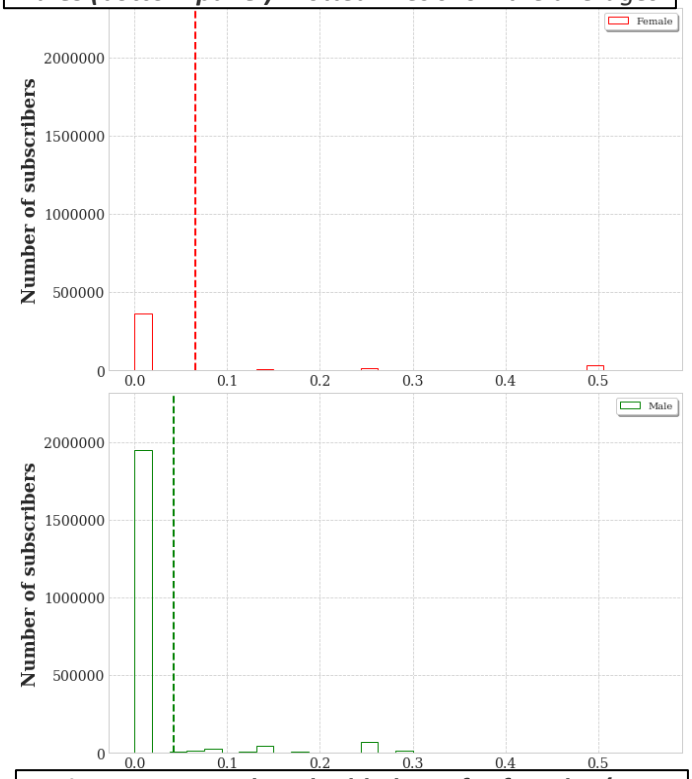

Figure 5 Network Embeddedness for females (top panel) and males (bottom panel). *Dotted lines show the averages.*

relatives, friends or acquaintances, many of whom are connected in ways independent of the woman. Men, however, many have commercial contacts that are not linked in ways independent of the man.

Network Constraints

The concept of structural holes [3] is a popular instrument in social network analysis research and has been used to assess the status of nodes in a network, diffusion of information and many other problems. Network constraints measure the extent to which the network does not span structural holes. If most of the neighbors of a node are connected to each other, then the node has higher constraints and vice versa. Disenfranchised communities like women in the developing world are expected to have higher constraints or lower number of structural holes.

The constraint of a node i denoted as $C(i)$ is defined as follows

$$C(i) = \sum_{j \in V(i)} (p_{ij} + \sum_{q \in V(i)} p_{iq} + p_{qj})^2$$

$V(i)$ represents the list of neighboring nodes of the node i, while the proportional tie strengths p_{ij} is based on the adjacency matrix A and is defined as follows

$$p_{ij} = \frac{a_{ij} + a_{ji}}{\sum_{k \in V(i)} (a_{ik} + a_{ki})}$$

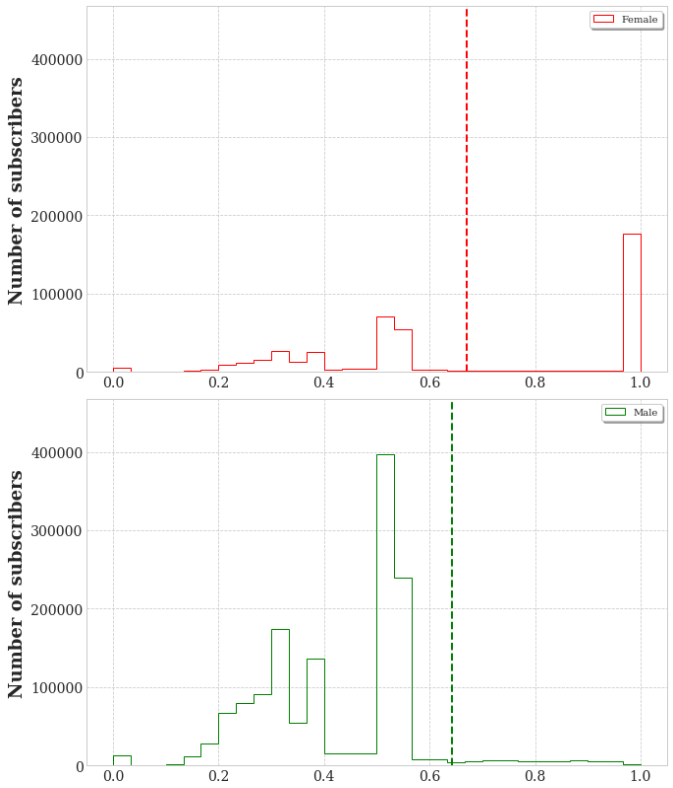

Figure 6 Network Constraints for females (top panel) and males (bottom panel). *Dotted lines show the averages*

Average constraints for women in our network is slightly higher as compared to the constraints for men (0.67 and 0.64 respectively) as shown in Figure 6.

Betweenness Centrality

Importance of a node in a network can also be quantified by calculating the betweenness centrality of the node, i.e., the number of shortest paths that go through the node.

More formally, the betweenness centrality of a node i is given by the following equation

$$g(i) = \sum_{s \neq i \neq t} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Where σ_{st} is the total number of shortest paths from node s to node t, while $\sigma_{st}(i)$ is the number of those paths that pass through i.

Comparison of betweenness centrality of males and females is shown in Figure 7. Figure 7 shows that the betweenness centrality of the men is consistently higher as compared to the betweenness centrality of women in Pakistan, which is a trend that we expected given that the number of men in the network is much higher than the number of women.

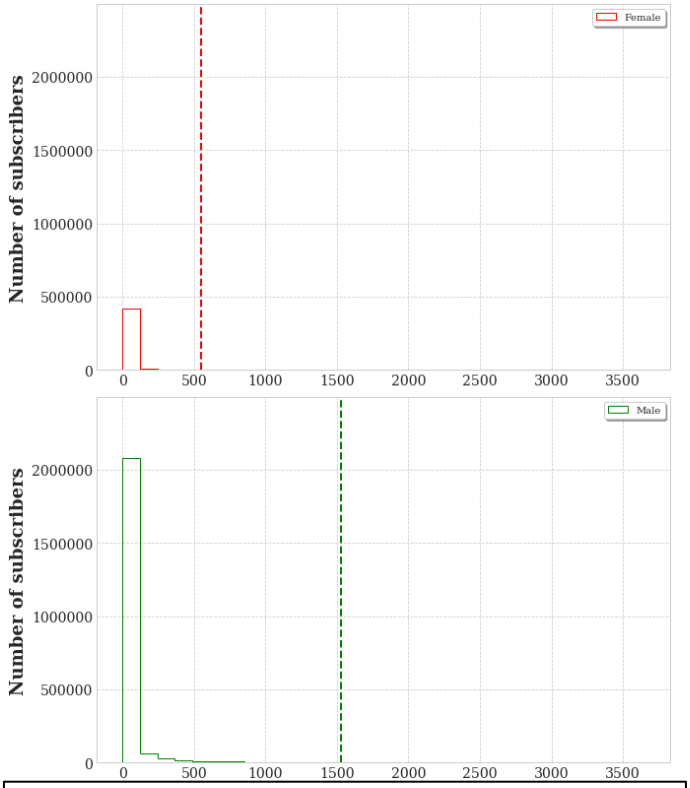

Figure 7 Betweenness centrality for females (top panel) & males (bottom panel). *Dotted lines show the averages*

NETWORK FORMATION

Gender Homophily

People belonging to different categories and communities can have different preferences for network formation, and these preferences can result in patterns of homophily regarding gender, age, and other characteristics. Figure 8 shows that male to male edges in the network are much more frequent as compared to female to female edges.

Homophily in a network tells us about the tendency of a node to form connections with similar nodes. However, plain homophily cannot accommodate the frequency of communication between the edges. Diversity-related measures can be used to measure the extent to which a node interacts with a particular type of node. Network diversity has been shown to be an important feature for predicting the socio-economic status of individuals [4]. Network diversity is defined as a function of Shannon entropy, as shown in the following equation.

$$Diversity = \frac{\sum_{i=1}^{N} -p(i) * \log(p(i))}{logN}$$

Here N indicates the total number of possible groups across which the diversity is to be calculated while p(i) indicates the proportion of calls being

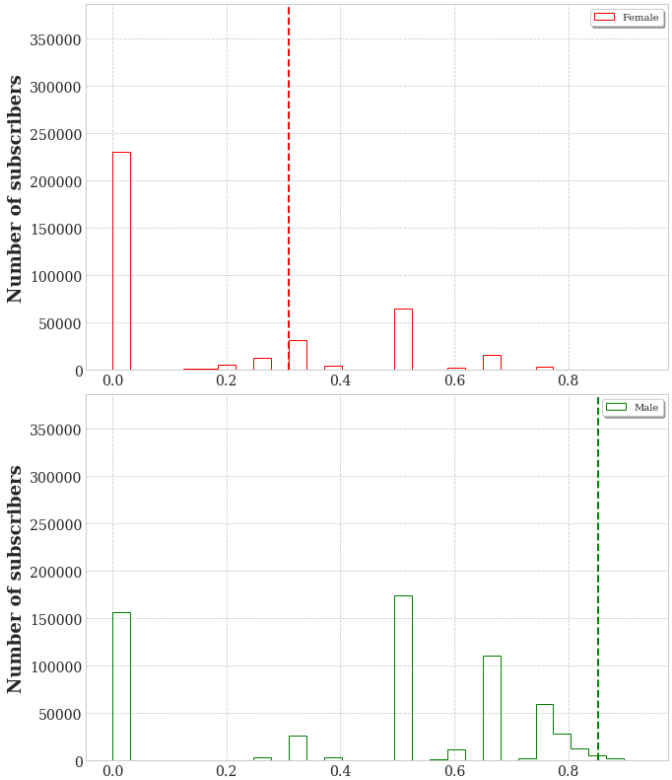

Figure 8 Gender Homophily for females (top panel) and males (bottom panel). *Dotted lines show the averages*

made to the ith group. Based on this equation, we calculate different diversity related metrics, as explained in the following subsections.

Gender Diversity

Gender Diversity calculates the proportion of calls being made by an individual to each gender. Women have higher gender diversity compared to males in our dataset, as shown in Figure 9. This trend seems to be a contradiction of the trend seen in Figure 8 at first, but the concept of gender homophily does not incorporate the volume of calls made to each of the group.

Age Diversity

Just like gender diversity, age diversity calculates what proportion of calls are being made by the individual to each age group. Based on the age distribution of the subscribers in our dataset, we have defined four different age groups as follows: Group 1 (25-24 years), Group 2 (25-39 years), Group 3 (40 – 54 years), Group 4 (55 and beyond).

The comparison of the age diversity of males and females in our dataset is shown in Figure 10. Males have slightly higher age diversity as compared to females on the average (0.057 vs. 0.049). Once again, the larger network size of the males on the average is one possible reason for the higher age diversity in the network.

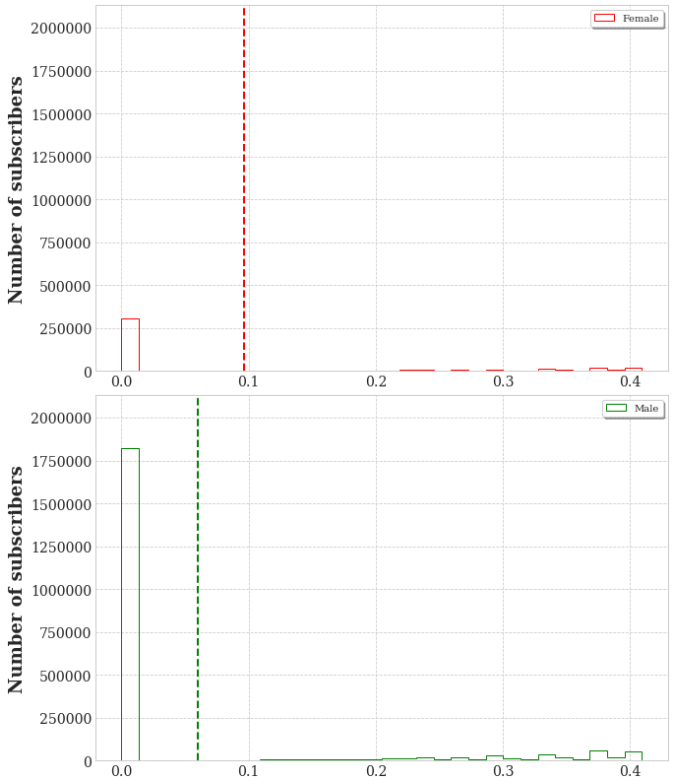

Figure 9 Gender Diversity for females (top panel) and males (bottom panel). *Dotted lines show the averages*

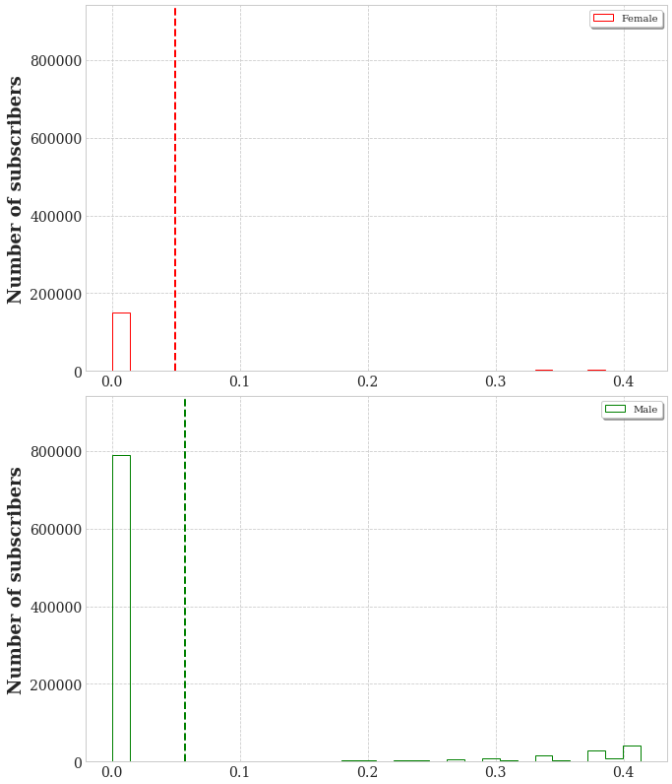

Figure 10 Age Diversity for females (top panel) and males (bottom panel). *Dotted lines show the averages*

Topological Diversity

Topological diversity analyzes the proportion of calls being made to each of the persons in an individual's network. The distribution of topological diversity for males and females in the dataset is shown in Figure 11. The topological diversity of men on the average is slightly higher as compared to the topological diversity of women (0.25 vs. 0.22).

As most of the working class in Pakistan constitutes men, the men are expected to be calling to different contacts in their network, and these contacts may have higher variation in their location as compared to the contacts of the women in general which explains the higher topological diversity of men on the average. But the difference between the average topological diversities between females and males is not very high which may be because of the fact that females are more active in communication with other family members in different locations.

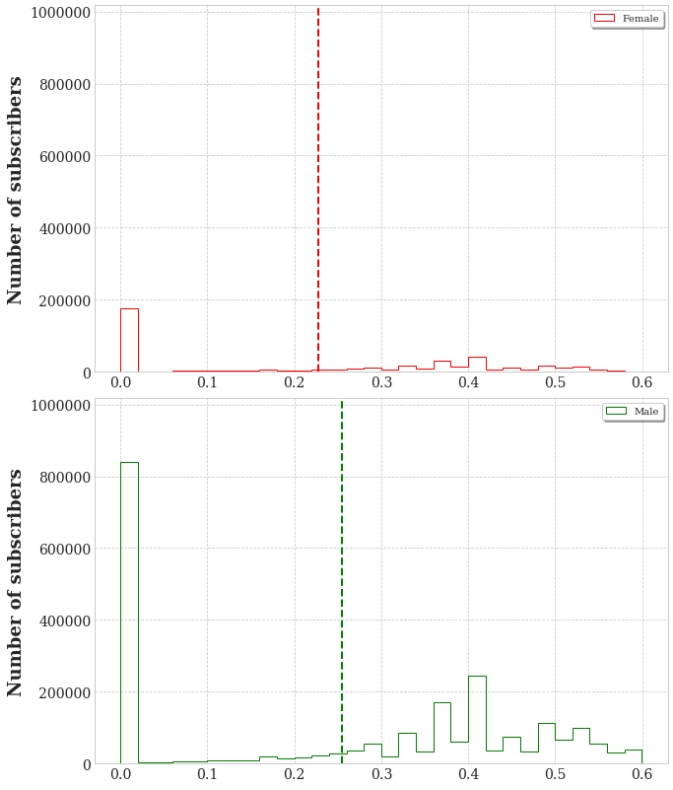

Figure 11 Topological Diversity for females (top panel) and males (bottom panel). *Dotted lines show the averages*

MOBILITY RELATED MEASURES

As shown in Table 2, the location information in the CDR data (Caller Cell Id and Recipient Cell Id) enables us to calculate many location-related or mobility- related features as well. These mobility related features can be really important as it has been shown in different research papers that many of these diversity-related measures may be correlated with the socio-economic status of individuals at the micro level and different regions at the macro level (For example, see [13] and [4]). Amongst these features, the distribution of two, namely location diversity and average geographical reach, of males and females is shown in Figure 12 and 13, respectively.

Location Diversity calculates the proportion of calls being made to each of the locations to which the user has been communicating. Location diversity of males is higher than the location diversity of females on the average, as shown in Figure 12.

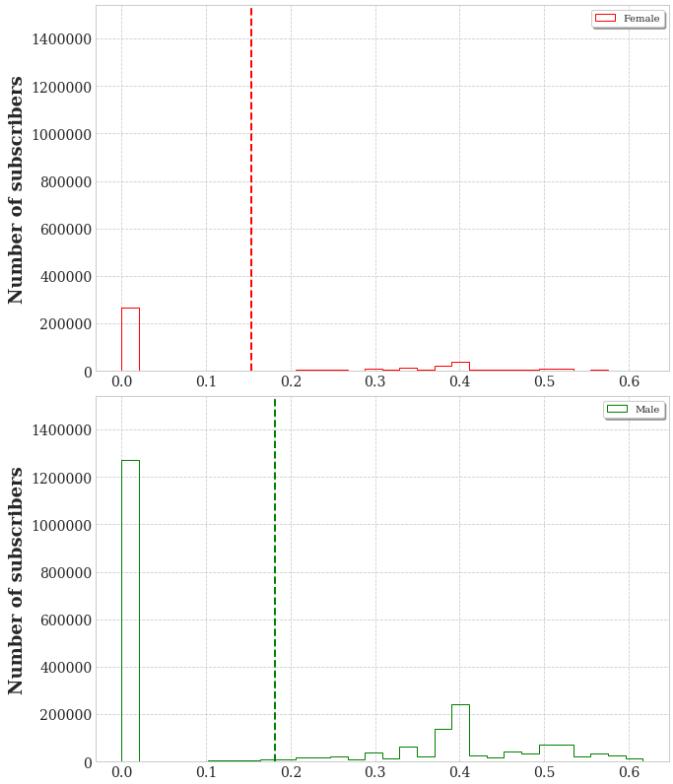

Figure 12 Location Diversity for females (top panel) and males (bottom panel). *Dotted lines show the averages*

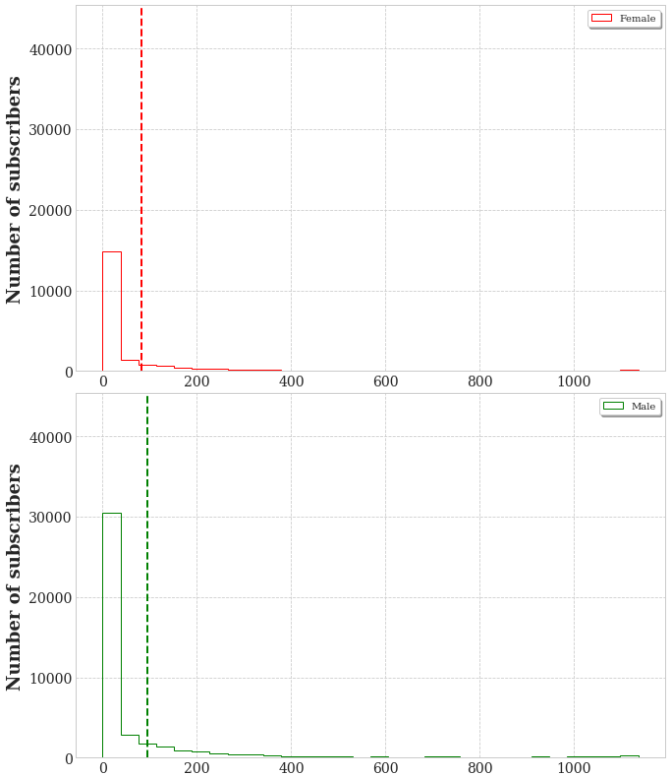

Figure 13 Average Geographical Reach for females (top panel) and males (bottom panel). *Dotted lines show the averages*

Average geographical reach calculates the average geographical distance between the caller and the recipient. The average geographical reach of males is higher as compared to the average geographical reach of women, as shown in Figure 13.

From the distributions of different features highlighted in these sections, we can see three prominent trends.

- Women have a higher average value for the features, for example, network constraints and network embeddedness, which depend on the interconnections of the nodes in the ego's network.
- Men have higher average values for the features which depend on the number of males in the network. Some of the examples of these features include gender homophily and betweenness centrality.
- Lastly, in Pakistan's culture, most of the men work outside of the home whereas women are expected to be the homemakers. This cultural norm implies that the network of the men is bigger on average as compared to the network of women. The higher average network size of men also results in higher topological diversity. The average value of the mobility-related features for the

men is also higher as compared to that of women, as men's networks may contain many geographically scattered professional contacts.

STATISTICAL SIGNIFICANCE OF FEATURES

The statistical significance of different features used in our analysis is shown in Table 3. Avg(M) and Avg(F) indicate the average values of the feature for male and female subscribers in the dataset. The last column shows the difference between the average value of males and females along with the p-value calculated through a t-test. The distribution of each of these features has already been described in the last section; however, this table shows that age diversity and topological diversity are not as significant as compared to the other features.

Feature	Avg(M)	Avg(F)	Diff(F-M)
Number of Calls and SMS	10.54993	18.08716	7.54***
Degree Centrality	2.143208	1.990399	-0.15***
Network Embeddedness	0.051923	0.07631	0.025***
Betweenness Centrality	2035.78	544.2583	-1491.52***
Topological Diversity	0.310245	0.266003	-0.04***
Gender Diversity	0.075737	0.114932	0.04***
Age Diversity	0.066672	0.055562	-0.01*
Topological Diversity	0.310245	0.266003	-0.04*
Gender Homophily	0.843515	0.318313	-0.53***
Average Geographical Reach	92.50838	80.0252	-12.48***
Location Diversity	0.371363	0.358917	-0.012***

Table 3: Statistical Significance of different features*. Avg(M) and Avg(F) indicate the average of the feature values for the males and females respectively. The last column shows the difference between the averages. *** P-Value <=0.001, ** 0.001<P-Value <=0.01, * 0.01<P-Value <=0.1*

5. PREDICTING EDUCATIONAL GENDER DISPARITY AT THE DISTRICT LEVEL

Our second goal in this project was to see how accurately we can predict the educational gender disparity at the district level given the CDR based features described in the last section. Given the individual level features, this is accomplished in three main steps.

1. Converting individual features to district level features

We first convert the individual level features calculated in the last step to district level features. Given the subscriber level features, we apply the mean, median and standard deviation operations to each of the features to get the district-level features. We further calculate the ratio features as the ratio of the district level average feature value of females to the same for males. Furthermore, we also calculate the proportion features as the ratio of the district level average feature value of females to the same value for all subscribers in that district.

2. Top Feature Selection

Given the high number of district-level features that we have, it is important to eliminate redundant or useless features from the final model so that the accuracy of the model can be improved. Elimination of

redundant features also helps in improving the interpretability of the model. We used a cross-validated Random Forest Classifier to rank features based on the value of R-squared, and the most optimal set consisting of 30 features was selected using Recursive Feature Elimination (RFE) [14].

The top 4 features selected through RFE are listed below.

1. Average gender diversity of males in a district
2. Average gender homophily of males in a district
3. Average embeddedness of all the users in a district
4. Average geographical reach of all the users in a district

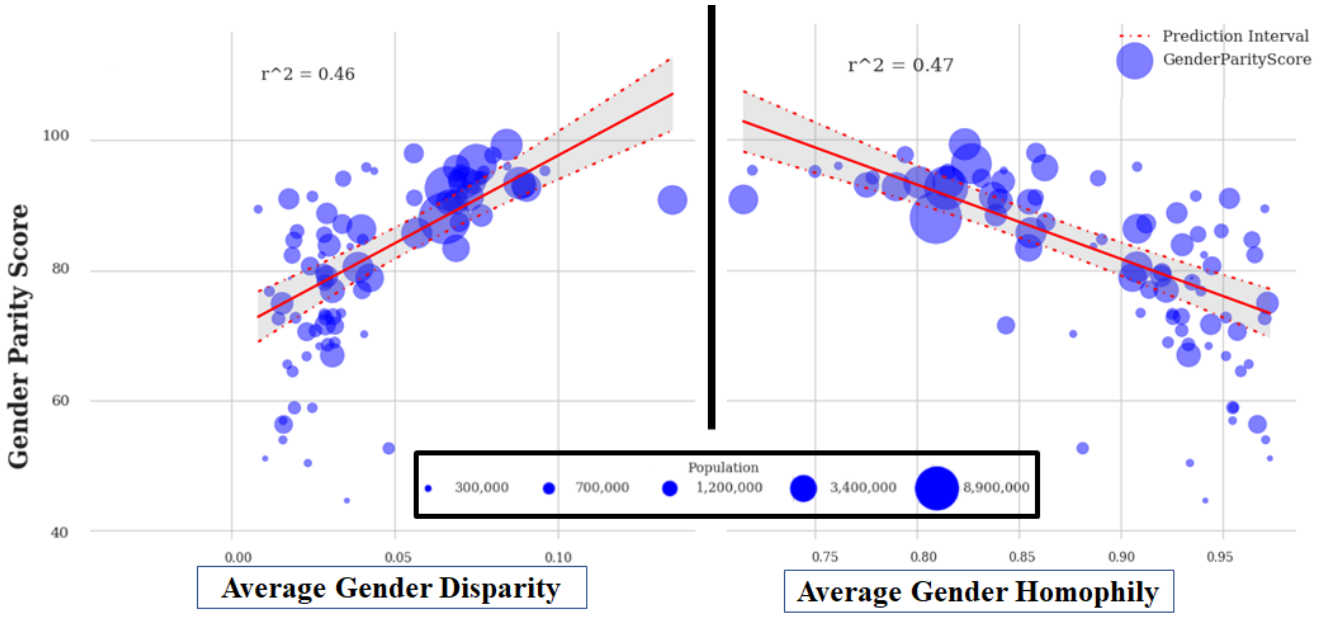

Figure 14 Top features selected through RFE. *Left: Gender parity score vs Average gender diversity of males in a district, Right: Gender parity score vs Average gender homophily of males in a district*

Figure 14 shows the relationship between the top 2 features selected through RFE and the gender parity score. The subfigure on the left shows the predictive performance of a linear least square regression model weighted by the district population and trained on the average gender diversity of males in a district. The subfigure on the right shows the predictive performance of a linear least square regression model weighted by the district population and trained on the average gender homophily of males in a district. The R-squared of these models on the training dataset is 0.46 and 0.47 respectively.

3. Final Prediction Models

In the final prediction models, we use the top 30 features selected through the feature selection process described in the last step and build different machine learning models using these features. Gender

disparity at the district level has not been a widely researched problem, so there is no consensus or existing baselines to serve as comparisons. We thus use network activity (Baseline Model 1), network size of the users (Baseline Model 2) and the ratio of female to male users in a district (Baseline Model 3) as the models against which to compare our models.

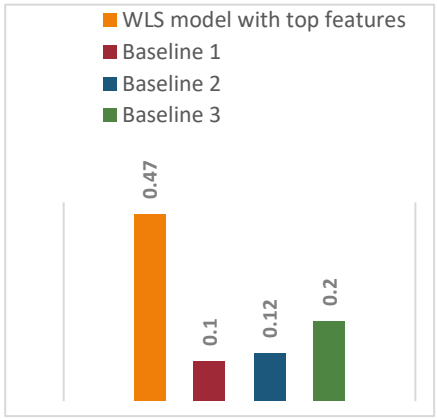

Figure 15 R-Squared for the experiment A

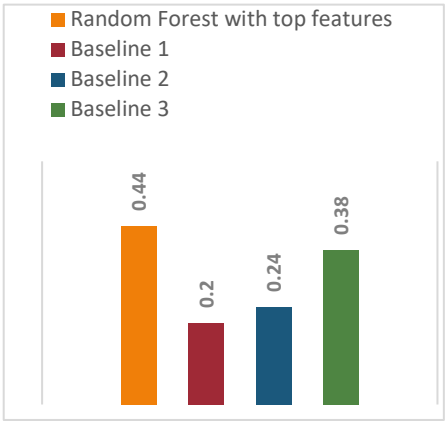

Figure 16 Micro-F1 score for the experiment B

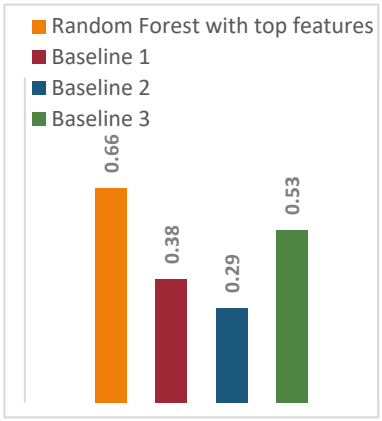

Figure 17 Micro-F1 score for the experiment C

Given the top features selected through the feature selection process, we set up three different type of experiments. Detail of these experiments are as follows:

Experiment A: Prediction of district-level gender parity score

In this experiment, we use different regression models like Weighted Least Squares and Random Forest Regression to predict the gender parity score of each of the districts. In this experiment, the least square regression model weighted by the population of the districts outperformed other regression models. The R-squared for our best performing model in comparison to the baseline models is shown in Figure 15.

Experiment B: Classifying districts into fine-grained categories based on the gender parity score

In this experiment, we have tried to classify the districts based on the gender parity score binned into seven different categories, based on the minimum and maximum value of the gender parity score. The intervals defining these categories are 30-39, 40-49, 50-59, 60-69, 70-79, 80-89 and 90-99. The random forest classifier with top features selected through RFE outperforms other models for this experiment. The performance of random forest classifier and baseline models for this experiment is shown in Figure 16.

Experiment C: Classifying districts as having low, medium or high gender parity score

Some of the categories in experiment B have very few districts. For instance, only two districts have the gender parity score between 30-40. In experiment C, the gender parity score is binned into three different categories: Low, Med and High such that each of these categories has an almost equal number of districts. In this experimental setting, the districts having gender parity score less than 75 are classified

as *low*, the interval 75-89 corresponds to a *medium* gender parity score, and greater than 89 corresponds to the *high* category. Just like the case of experiment C, the random forest classifier with top features outperforms other models for this experiment. The performance of random forest classifier and baseline models for this experiment is shown in Figure 17.

For each of experiments A, B and C we have evaluated different machine learning models using 10-fold cross-validation with a train/test ratio of 80:20.

It is clear from Figure 15 that our approach beats the baseline models by a significant margin. The weighted least squares regression model used in experiment A and trained on the top features selected through RFE beats the weighted least squares model trained on individual top features as well.

Similarly, the Random Forest classifier with top 30 features selected using RFE outperforms baseline models for both experiments B and C (Figure 16 and 17). Performance of the Random Forest classifier with top features is much better for experiment C as compared to experiment B as the class distribution is balanced.

4. DISCUSSION

In this project, we have analyzed the differences in social networks of men and women in a developing country using call detail records, with the aim of developing a predictive model to predict gender disparity at the district level.

Our first contribution in this project is to demonstrate the suitability of CDR-based social networks for research on gender disparities. For large-scale studies like the one discussed here, CDR-based data can provide many advantages. Collecting CDR data does not require a great deal of additional investment; even the poorest of countries have seen a good penetration of mobile phones. Secondly, in comparison to other online social networks like Facebook and Twitter, mobile phones have seen wider adoption across all segments of the society, whereas networks like Facebook and Twitter are relatively more popular among youth, urban residents, and higher income classes.

Many notable differentiating patterns can be seen between the social networks of men and women in Pakistan. The rich information contained in the call detail records enables us to compare the social networks of men and women not only according to simpler features like size of the network and the activity on the network. We also have advanced measures related to network formation and network status, which can provide much more interesting information about the salient differences between the social networks of men and women in the developing world.

As explained in Section 4, the social network of men and women have statistically significant differences in terms of call volume, network size, embeddedness, gender homophily, and average geographical reach.

Interestingly, women in Pakistan use the mobile phone network more actively but consistently have a smaller network size. This pattern indicates that women either prefer relatively stronger ties with fewer

nodes in their network, as compared to men or are constrained to do so. However, as the percentage of women with higher embeddedness is greater than the corresponding percentage of men, this shows that women may have a more central position in the networks. Females show higher gender homophily in accordance with the prevalent social norms of the Pakistani society. Furthermore, the trends of gender and diversity are also in accordance with the prevalent social norms of the Pakistani society, wherein the females have more central roles in the families while the men are the breadwinners. As most of the professional workforce in Pakistani society consists of men, therefore, men, as expected, have higher network size and topological diversity. Furthermore, the network of the men is geographically more spread.

Each of the behavioral features discussed in this report casts a different light on the social networks of men and women in developing countries. On the one hand, these differences highlight how men and women organize their social networks, while on the other hand, these differences highlight how the men and women can be susceptible to diffusion of information and opportunities through their networks. As machine learning models can play a helpful role in the spread of different initiatives (e.g., digital financial services, health services), the knowledge of the factors influencing the diffusion of information can help immensely in the success of these initiatives.

A critical question that requires further exploration is whether the features selected by the feature selection process represent trends in the society or not. We intend to handle this question in our future research as our focus in this project has been on selecting the features which result in the best performance for the machine learning models. Some of the top features selected by the RFE algorithm are easier to interpret in the social context while some others are not easy to interpret. For example, among the top features, gender diversity is positively correlated with gender parity while gender homophily is negatively correlated with the gender parity. Higher gender diversity of males in a district indicates a relatively higher activity of females in the district which can show the better social status of the females in the district. On the contrary, higher gender homophily of males may be an indicator of lower network activity of females in the district. Similarly, the higher geographical reach of the individuals of a district on the average may also be positively correlated with higher HDI of the district. However, the relationship of higher average embeddedness of the users in a district with the gender parity may not be that obvious. Many of the features selected by the RFE algorithm have positive correlations with the network activity, but it is not obvious why these features are more important as compared to the network activity in general. The correlation of these features with the actual social trends is a research topic on its own and needs much more attention, but it was not the focus of this project.

Lastly, the predictive model we have developed beats the performance of other baseline models by quite some margin. Not only we can predict the raw gender parity score, but we are also able to relatively accurately classify the districts as having low, medium or high gender parity scores. From the perspective of the government and social work organizations, we think that this classification model will be useful for these organizations to align their resources and initiatives in the districts with lower gender disparity. In the absence of such classification models, organizations are either dependent on the statistics

collected by the government organizations or the surveys conducted by the social welfare organizations. Government level data collection and analysis can demand a great deal of time and resources, while the surveys collected by the social welfare organizations are usually not comprehensive.

Our study opens some other interesting questions for future exploration as well. For example, do the country-level social network findings correlate with the provincial-level social networks, or do the more progressive provinces have different patterns? Furthermore, to what extent do the patterns found in this society hold for countries with similar cultural and socio-economic background as Pakistan? We intend to handle these questions in future work. Similarly, the application of deep neural networks based models on the population level CDR data is another interesting area for future research.

REFERENCES

[1] G. Magno and I. Weber, "International Gender Differences and Gaps in Online Social Networks," in *International Conference on Social Informatics*, 2014.

[2] M. R. Khan and J. Blumenstock, "Predictors without Borders: Behavioral Modeling of Product Adoption in Three Developing Countries," in *Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining.*, 2016.

[3] J. Blumenstock, G. Cadamuro and R. On, "Predicting poverty and wealth from mobile phone metadata.," *Science,* 2015.

[4] . S.-C. Christopher, A. Mashhadi and L. Capra, "Poverty on the cheap: Estimating poverty maps using aggregated mobile communication networks," in *SIGCHI Conference on Human Factors in Computing Systems*, 2014.

[5] J. L. Toole, Y.-R. Lin, E. Muehlegger, D. Shoag, M. C. González and D. Lazer, "Tracking employment shocks using mobile phone data," *Journal of the Royal Society Interface,* 2015.

[6] J. Blumenstock, "Inferring patterns of internal migration from mobile phone call records: evidence from Rwanda," *Information Technology for Development,* 2012.

[7] P. J. Reed, M. R. Khan and J. Blumenstock, "Observing gender dynamics and disparities with mobile phone metadata," in *Eighth International Conference on Information and Communication Technologies and Development*, 2016.

[8] L. Ryan, R. Sales, M. Tilki and B. Siara, "Social Networks, Social Support and Social Capital: The Experiences of Recent Polish Migrants in London," *Sociology,* vol. 42, no. 4, pp. 672 - 690, 2008.

[9] B. Uzzi and J. Sprio, "Collaboration and creativity: The small world problem," *American journal of sociology,* vol. 111, no. 2, pp. 447-504, 2005.

[10] V. Frias-Martinez, E. Frias-Martinez and N. Oliver, "A Gender-Centric Analysis of Calling Behavior in a Developing Economy Using Call Detail Records.," in *AAAI spring symposium: artificial intelligence for development*, 2010.

[11] Y. Dong, Y. Yang, J. Tang, Y. Yang and N. Chawla, "Inferring User Demographics and Social Strategies in Mobile Social Networks," in *KDD*, 2014.

[12] R. S. Burt, Structural Holes, Cambridge, MA: Harvard University Press, 1992.

[13] N. Eagle, M. Macy and R. Claxton, "Network Diversity and Economic Development," *Science,* 2010.

[14] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning,* vol. 46, no. 1-3, pp. 389-422, 2002.