Educational Inequality and Mobile Phone Data

MUHAMMAD RAZA KHAN, UNIVERSITY OF CALIFORNIA BERKELEY SCHOOL OF INFORMATION

Introduction

Can mobile phone data serve as a high-resolution monitoring system for educational inequalities between girls and boys? This project uses a mobile phone dataset from Pakistan to predict net primary enrollment rates of children at the district level. Men and women in Pakistan differ considerably in their patterns of phone use and in the structure of their calling networks. These differences reflect mobility, poverty, and other social factors that also drive disparities in education—and thus inequalities in phone data are correlated to inequalities in school enrollment. This method could be applied to other countries where call detail records with gender information are available, greatly reducing the cost and logistical difficulty of gathering high-resolution educational data. Overall, patterns of phone use and the structure of calling networks provides important insight into how women's social freedoms are changing.

Our Approach

We used call detail record (CDR) data from a major operator in Pakistan, containing more than one billion voice and text messages from around six million users. 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. The sex and age of each of the subscribers are also provided by the telecom operator. About 8.4% of the users (~0.57 million) in this dataset were women. The CDR data spans seven days and covers 93 of 128 districts in Pakistan. To validate the link between mobile phone use and education, we used primary enrollment data collected by the gender advocacy NGO Alif Ailaan (Figure 1).

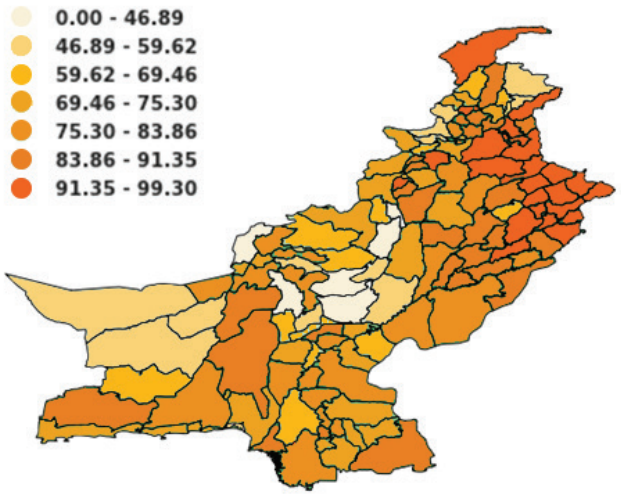
We focused on two key questions in this study:

1) How do the mobile phone-based social networks of women and men differ?

2) Do the differences in these social networks reflect inequalities in primary school enrollment?

The first question is about describing the association between the sex of the phone user and the features of the CDR dataset, including number of calls made, size of network, and various network metrics (number of contacts, friend clustering, etc.). We convert individual network features to district level averages, disaggregated by sex. The second question is about whether these social network features can be used to predict educational gender disparity at the district level. Because of the very large number of possible network features that can be extracted from CDRs, we rank features based on their explanatory power, and then select the top 30 features. We then compare the gender disparity of these features to the educational gender disparity seen in the validation dataset.

Figure 1. Parity of net enrollment in primary school, Pakistan. A score of 100 would indicate gender equality.

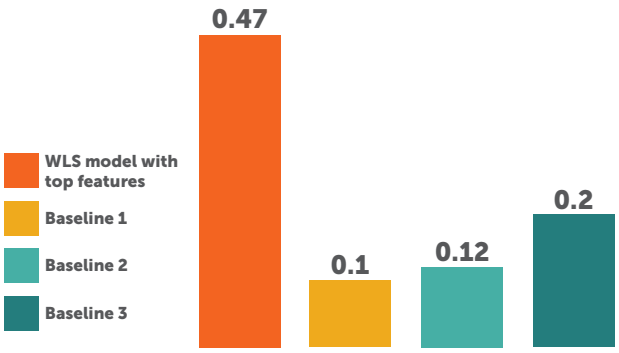


Results

The social networks of women and men differ. Women tend to make more calls within a given period, have a smaller calling network, have more clustered friend groups (their close contacts tend to know each other), function less as "bridges" (key connections between separated groups), and are more limited in mobility. These features draw a portrait of women's social lives in Pakistan. Although women use mobile phones when they have access, the value of the technology appears to be lessened by the social restrictions they face. This type of network analysis is easily performed with any CDR dataset containing gender information.

Three network features perform particularly well in predicting educational disparities: 1) gender diversity of male calling networks; 2) clustering of friend groups ("embeddedness") across all networks; and 3) geographical reach (distance between callers and recipients) across all networks. We find that the overall set of 30 features explains nearly half of district-level educational inequality. This "best features" model easily outperforms simpler approaches (Figure 2).

Figure 2. Comparing R-squared of final models. Baseline 1 is network activity only; Baseline 2 is network size only; Baseline 3 is ratio of female to male users in a district only.



Implications

Assessing gender equality in school enrollment is expensive using traditional methods, especially at local levels and with high frequency over time—important considerations given the rapid pace of economic and cultural change. However, CDR network data can function like a near-real-time surveillance system to track school enrollment, which may be especially important information during conflicts, natural disasters, or other shocks. The data is passively generated and thus very inexpensive. Such a detailed picture can help reallocate educational resources and policy attention towards regions of countries with persistent disparities, as well as focus research attention on the causes of persistent inequality. Governments can greatly ease the usability of CDRs by developing legal and technical protocols by which mobile network operators can safely share anonymized and aggregated data. This is especially critical when analyzing patterns at very localized levels.