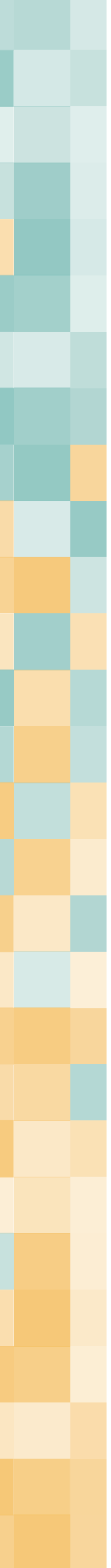


Big Data, Big Impact?

Towards Gender-Sensitive
Data Systems

November 2019

data2x^o

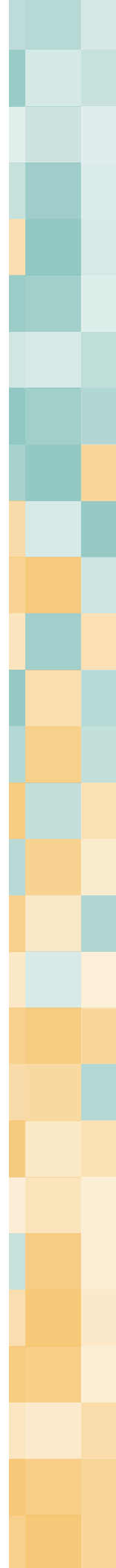


Acknowledgements

This report was compiled by Babu Vaitla, with assistance from grantee researchers. The full list of grantees and research staff (in alphabetical order) are Natalia Adler, Youssef Al Mouatamid, Reham Al Tamime, Linus Bengtsson, Joshua Blumenstock, Girija Borker, Claudio Bosco, Marouane Boujlal, Christopher J. Brooks, Ciro Cattuto, Laura Chioda, Saad Choukry, Daniele de Rigo, Masoomali Fatehkia, Yashmin Fernandes, Leo Ferres, Alina Game, Laetitia Gauvin, Paul Gertler, Kate Glynn-Broderick, Joshua Greenhalgh, Antoine Heuty, Sean Higgins, Abigail Hunt, Ridhi Kashyap, Muhammad Raza Khan, Ian Knowles, Mayank J. Kumar, Emmanuel Letouzé, Youssef Mehdi Oukaja, Mehi Mirpourian, Hajar Mousannif, Kristine Nilsen, Amy Ninneman, Simone Piaggese, Frederic Pivetta, Sarchil Qader, Emma Samman, Kerry Stephens, Michele Tizzoni, Kristýna Tomšů, Anna Tyor, Stefaan Verhulst, Samantha Watson, Ingmar Weber, Richard Wood, Andrew Young, Jihad Zahir, and Rositsa Zaimova.

We are grateful to Emily Courey Pryor, Rebecca Furst-Nichols, Mayra Buvinic, Eleanor Carey, Carlos Mazariegos, Nina Rabinovitch Blecker, Elizabeth Black, Karolina Ramos, David Garrison, and Natalie Cleveland for assistance in creating this report and for helpful comments.

This work was initiated by Data2X, a collaborative technical and advocacy program working through partnerships with United Nations agencies, governments, civil society, academics, and the private sector to improve the quality, availability, and use of gender data to make a practical difference in the lives of women and girls worldwide. Through our research, advocacy, and communications, we build the case and mobilize action for gender data to make it central in global efforts to achieve gender equality. Additionally, we strengthen the production and use of gender data by working with data producers and users to ensure that data collection methods are unbiased, policy-relevant, and gender sensitive. We believe these are necessary steps toward building a more equal world. Data2X is housed at the United Nations Foundation and supported by the Bill & Melinda Gates Foundation and the William and Flora Hewlett Foundation.



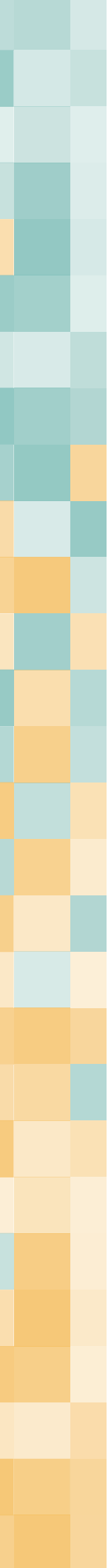
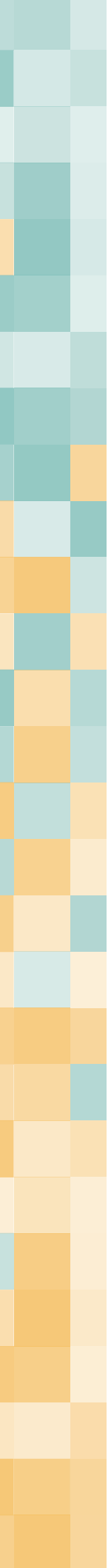


Table of Contents

Executive Summary	1
Case Studies	8
Mobile Phone Data	
Case Study 1: Gender Gaps in Urban Mobility.....	9
Case Study 2: Mobile Money & Gender in Uganda.....	11
Case Study 3: Educational Inequality and Mobile Phone Data.....	13
Geospatial Data	
Case Study 4: Street Harassment and Women's Educational Choices: A Geospatial Analysis.....	15
Case Study 5: Towards High-Resolution Sex-Disaggregated Dynamic Mapping.....	17
Application Data	
Case Study 6: Using Facebook and Google Advertising Data to Measure the Digital Gender Divide.....	20
Case Study 7: Women and the Digital Gig Economy in South Africa.....	23
Case Study 8: Mining the Web for Insights on Violence Against Women in the MENA Region and Arab States.....	26
In-Progress Case Studies	
Case Study 9: Uptake and Usage of Financial Services to Advance Women's Financial Inclusion.....	29
Case Study 10: Gender-Differentiated Credit Scoring Algorithms Using Call Detail Records and Machine Learning.....	29



Executive Summary

The global gender data gap persists and our knowledge of the lives of women and girls remains insufficient to meet the challenge of designing policies to achieve the Sustainable Development Goals (SDGs). Yet we live in an era of big data; massive amounts of information stream from cell phones, laptops, remote sensors, and an ever-growing host of technologies, even in the least developed and most isolated places in the world. Transforming this flood of data into actionable knowledge about the lives of women and girls is one of the great technical and moral tests of the 21st century.

Data2X is proud to be at the forefront of the work at the intersection of big data and gender, investigating the potential of new data sources to improve our understanding of the lives of women and girls. This report summarizes the results from our grants supporting cutting-edge social scientists using big data to fill the global gender data gap. The research projects profiled in this report have succeeded in translating that potential to reality and demonstrate that big data is a powerful source of scientifically rigorous, policy-relevant knowledge.

The featured case study projects span a diverse range of big data types and gender themes. The pages that follow summarize the methods, results, and implications of each project. In this Executive Summary, we offer five cross-cutting messages that emerge from the body of work.

Key Messages

1. Big data offers unique insights on women and girls.
2. Gender-sensitive big data is ready to scale and integrate with traditional data.
3. Identify and correct bias in big datasets.
4. Protect the privacy of women and girls.
5. Women and girls must be central to data governance.

Big Data and Gender Case Studies

Case Study 1: Gender Gaps in Urban Mobility

Data Type: Cell phone call detail records

Location: Santiago, Chile

Grantees & Institutions: The GovLab, UNICEF, Universidad Del Desarrollo, Telefónica R&D, ISI Foundation, University of Bologna

Case Study 2: Mobile Money & Gender in Uganda

Data Type: Cell phone application data and call detail records

Location: Uganda

Grantees & Institutions: Dalberg Data Insights

Case Study 3: Educational Inequality and Mobile Phone Data

Data Type: Cell phone metadata

Location: Pakistan

Grantees & Institutions: Muhammad Raza Khan (University of California Berkeley School of Information)

Case Study 4: Street Harassment and Women's Educational Choices: A Geospatial Analysis

Data Type: Geospatial datasets (Google Maps, safety incident applications), student choice surveys

Location: New Delhi, India

Grantees & Institutions: Girija Borker (World Bank/Brown University)

Case Study 5: Towards High-Resolution Sex-Disaggregated Dynamic Mapping

Data Type: Cell phone call detail records, geospatial datasets, demographic and health data

Location: Nepal

Grantees & Institutions: Claudio Bosco (Flowminder Foundation, WorldPop), Samantha Watson (Flowminder Foundation, WorldPop), Alina Game (Flowminder Foundation, WorldPop), Christopher J. Brooks (Flowminder Foundation, WorldPop), Daniele de Rigo (Maieutike Research Initiative), Sarchil Qader (WorldPop), Joshua Greenhalgh (Flowminder Foundation), Kristine Nilsen (WorldPop), Amy Ninneman (Flowminder Foundation, WorldPop), Richard Wood, (Flowminder Foundation), Linus Bengtsson (Flowminder Foundation, WorldPop)

Case Study 6: Using Facebook and Google Advertising Data to Measure the Gender Digital Divide

Data Type: Facebook, Google advertising API

Location: Global

Grantees & Institutions: Ridhi Kashyap (Oxford University), Ingmar Weber (Qatar Computing Research Institute), Masoomali Fatehkia (Qatar Computing Research Institute), Reham Al Tamime (University of Southampton), Ian Knowles (University of Oxford)

Case Study 7: Women and the Digital Gig Economy in South Africa

Data Type: Digital gig platform data, qualitative interviews

Location: South Africa

Grantees & Institutions: Emma Samman and Abigail Hunt (Overseas Development Institute)

Case Study 8: Mining the Web for Insights on Violence Against Women

Data Type: YouTube, Twitter, Facebook, other social media platforms

Location: Middle East region

Grantees & Institutions: Jihad Zahir, Hajar Mousannif, et al. (Cadi Ayyad University)

Case Study 9: Uptake and Usage of Financial Services to Advance Women's Financial Inclusion

Data Type: Financial accounts data

Location: Nigeria

Grantees & Institutions: Mehi Mirpourian, Kate Glynn-Broderick, Anna Tyor, Kerry Stephens (Women's World Banking)

Case Study 10: Gender-Differentiated Credit Scoring Algorithms Using Call Detail Records and Machine Learning

Data Type: Cell phone call detail records, financial accounts data

Location: Dominican Republic

Grantees & Institutions: Joshua Blumenstock (University of California Berkeley), Laura Chioda (World Bank), Paul Gertler (University of California Berkeley), Sean Higgins (Northwestern University)

Key Message 1

Big data offers unique insights on women and girls.

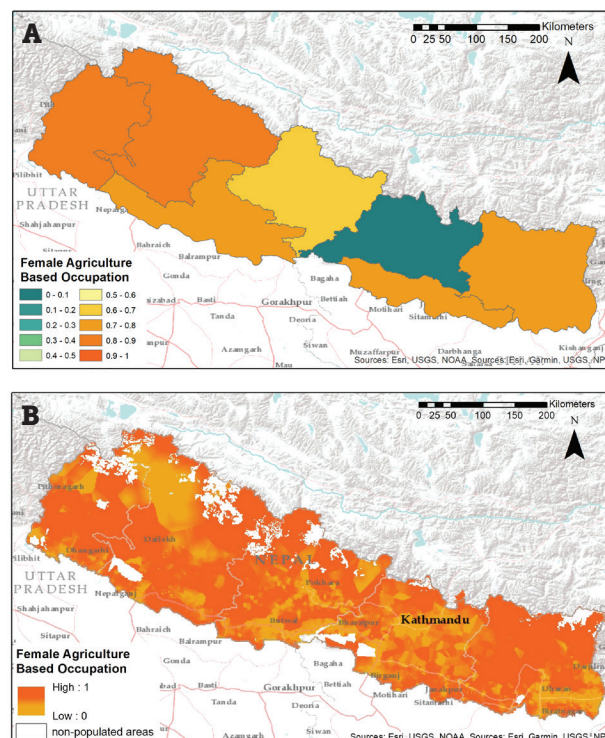
Data2X's work over the past several years has shown that big data can add nuance to our understanding of women's and girls' lives in two important ways: it can provide information that is highly granular in both space and time, and it can offer insights on aspects of life that are often difficult to quantify.

As evidenced in the case studies presented below, big data can help us understand several gendered aspects of economic and social behavior, including mobility (Case Studies 1 & 4), financial flows (Case Study 2), school enrollment (Case Study 3), use of health services (Case Study 4), economic participation (Case Studies 5 & 7), access to technology (Case Study 6), and gender-based violence (Case Study 8).

With investment, big data systems could provide a rich portrait of gendered life in a society. Insights derived from such systems could help policymakers better address the cultural, infrastructural, economic, and political forces that generate gender inequality and by doing so, help achieve the SDGs. For example, Figure 1, taken from Case Study 5, shows how big data can improve our understanding of women's participation in agriculture-based occupations across Nepal, enabling more accurate targeting of interventions. Figure 1 (A) shows the limitations of survey data, which permit only province-level estimates; Figure 1 (B) shows the nuances that emerge from geospatial and cell phone data. Because many forms of geospatial and cell phone data are continuously streaming, these maps update in real time, allowing rapid response to shocks. No traditional data system that currently exists can provide such rich information at scale.

Big data can also provide insight into phenomena

Figure 1. (A) Proportion of women engaged in agriculture-based occupations, province level. This is the highest resolution representative data available from the NDHS. (B) High-resolution landscape of women engaged in agriculture-based occupations, from combination of survey, geospatial, and mobile phone data



that are difficult to capture in standard types of data collection. For example, national socioeconomic surveys typically offer information about the status of the family as a whole, ignoring inequalities within the household. Information gathered from cell phone use, meanwhile, can help us learn more about the well-being of millions of individual women and girls. Similarly, promising sources of gender information like social media and financial mobile applications can yield insights on topics that are otherwise challenging to measure, such as attitudes towards gender-based violence (Case Study 7).

Key Message 2

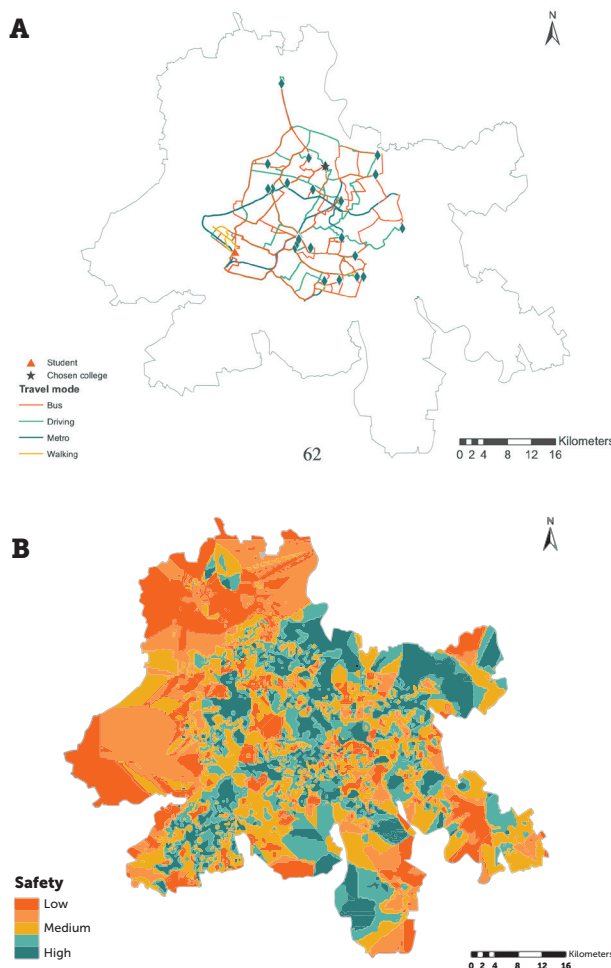
Gender-sensitive big data is ready to scale and integrate with traditional data.

Myriad pilot projects around the world attest that many types of big data methods are ready for scaling up and could be a useful source of information for national statistics offices and other data producers. Creating national-level systems would permit more geographically precise policy and program design and allow public sector agencies to respond more effectively to economic and environmental shocks unfolding in real time.

Additionally, scale up of big data systems would provide the opportunity to build integration with traditional datasets into the governance of future data systems. In fact, nearly every case study in this report relies heavily on a dataset assembled using traditional sources, either for the purposes of context-setting or to ensure accurate interpretation of the results of big data analytics.

The message is clear: big data yields the most powerful gender insights when combined with traditional datasets. For example, Case Study 4 looks at how women's educational choices are affected by the risk of street harassment in New Delhi, India. The research is built on various types of big data, including safety incident reports crowdsourced from mobile phone applications and Google Maps geospatial data but the work is centered on survey responses from students about their educational preferences. Figure 2 (A) below shows these "college choice sets" and the various transport options available to reach them. Figure 2 (B) shows a safety surface of the entire city, created by big data. The key conclusion of the case study — that the safety of transport routes constrains women's educational choices and imposes economic costs — would not be possible without both types of information.

Figure 2. (A) Route options to colleges in student choice sets. (B) Safety surface of Delhi Metropolitan area.



Other case studies also demonstrate the mutual dependence of big and traditional data. For example, the analysis of gendered mobility depends on measuring movement patterns across a large geographical space (Case Study 1) but also requires deep knowledge of the socio-cultural factors that facilitate or constrain the movement of women and girls (Case Study 4). Case Study 7 conducted qualitative interviews with South African women to understand the challenges and opportunities of working in the gig economy. Tracking online activity (Case Study 8) may help in measuring the prevalence of gender-based violence (GBV) as it varies within and across countries but understanding the rapidly transforming technological and economic context (Case Study 6) may reveal the causes of variation in GBV as women negotiate new virtual and real public spaces.

The case studies in the pages that follow are not only exemplars of cutting-edge research with different kinds of big data but also represent an emerging approach to data synergism that social science is increasingly embracing.

Key Message 3

Identify and correct bias in big datasets.

Though big data offers great promise for understanding the lives of women and girls, it is not always representative of the general population. Because of this bias, scaling up big data systems and integrating big and traditional data may not be enough to obtain an accurate picture of the lives of women and girls.

In fact, gendered access to the technology that generates big data may create especially pronounced biases in big datasets. This leads to inaccurate conclusions and could also result in the creation of algorithms that may amplify and further entrench this bias. For this reason, big data's potential will only be realized if complementary investments are made in research methods to identify and correct bias.

For example, Case Study 5 makes the critical point that a better understanding of how families share SIM cards is essential to making inferences about individual well-being from cell phone call detail records. Case Study 8 relies on human judgment to train a machine learning model to recognize and classify online expressions signifying gender-based violence. Validation studies were used by many of our grantees to evaluate the accuracy of individual gender tags within big datasets. This allowed the creation of methods to improve the accuracy of these tags, and thus produce gender-disaggregated results. Alongside efforts to close the gender gap in access to technology, investment in validation research — especially that which allows big data to be gender-disaggregated — would be a “quick win” for the donor community: such studies will rapidly amplify the power of big data.

Key Message 4

Protect the privacy of women and girls.

Big data poses a particular risk to the privacy of women and girls. In many societies, social norms may restrict the ability of women and girls to voice their opinions about privacy standards. In addition, they may have poor access to legal services to protect their consumer rights to privacy and may be excluded from participating in the public debate around issues like ethical private sector use of individual data. More broadly, big data is at a critical juncture: if best practices for data privacy are not codified in international law and/or national legislation in the next few years, public backlash will likely greatly stunt the development of research needed to meet the SDGs. Many of our grantees' original research designs were complicated by the rapidly evolving data privacy conversation. In fact, one of our initial grantees was prevented from carrying out their study after an initial memorandum of understanding around data sharing was rescinded by a telecommunications provider in reaction to concerns about public perceptions about privacy. Other projects were delayed by months for similar reasons, and still others had to agree to restrictive guidelines that hindered the process of answering their original research questions.

But there is hope. The development of sophisticated protocols to enable data access while guaranteeing privacy and security is possible. The overarching message is that with the coordinated effort of private companies, academic researchers, civil society, and the public sector — the latter to both develop legal guidelines for data sharing and take the lead in creating new data systems — the difficult balance between scientific insight and individual security is attainable.

Key Message 5

Women and girls must be central to data governance.

Better big data systems are a means to an end. The ultimate goal is a data landscape that both respects human rights and catalyzes greater gender equality. Attaining this vision depends not only on the refinement of scientific methods and legal guidelines, but also on the participation of women and girls in every stage of data generation and governance. The emerging paradigm of data feminism¹ can provide a guide as it examines the debate around power: who has it and who doesn't, and how those power differentials affect whether data indeed promotes the democratic ideals of justice and equality.

Authentic participation can also help confront problems like algorithmic bias, in which automated procedures reflect the social biases that plague datasets and are held by programmers and their institutions, consciously or unconsciously. Biased input leads to biased output, as we see, for example, in credit scoring algorithms that assign low credit worthiness to women because of their underrepresentation in the financial world. In addition to identifying and correcting bias in big datasets, greater participation by women and girls — not only as data holders but also as leaders in data governance, public policy activists, and data scientists — is the only long-term antidote to algorithmic bias. Such a change in representation will help push the culture of the big data community, including members of all genders, towards inclusivity and mutual respect. We urge all stakeholders in the big data and gender community — researchers, foundations and funding agencies, governments, private companies, and others — to make room for such conversations in their strategic plans and investment portfolios. Big data can make a big difference in the lives of women and girls, but only if it serves as a vehicle for genuine empowerment.

The Future of Gender-Sensitive Data Systems

Data2X exists to improve the lives of women and girls, and gender sensitive big data can help us move more quickly towards that goal. A combination of various types of big data, complemented by traditional data sources, can serve as the basis for socioeconomic information systems that are both highly granular in space and time and broad-scale.

For many big data methods, the time for proofs-of-concept has passed; investment into scaling up will yield strong dividends for the public and private sectors, especially as the volume of big data grows tremendously in the next few years. Difficult questions around individual privacy, data security, and authentically democratic governance persist, but an emerging coalition of communities, nonprofit advocates, companies, and policymakers is ready to meet the challenge. As the case studies in this report demonstrate, the potential benefits for women and girls, indeed for all people, are tremendous.

1. A term coined by Catherine D'Ignazio and Lauren Klein in their book of the same name, <https://bookboook.pubpub.org/data-feminism>.



Case Studies

The eight² research projects profiled in this report can be divided into three overarching categories: mobile phone data, geospatial data, and application data. This categorization is imperfect: as noted earlier, many of the studies utilize more than one type of big data and in some cases use traditional forms of data, especially household surveys, as a complementary source of information. The most innovative and policy-relevant big data studies are increasingly characterized by this data eclecticism. Mobile phone data, for example, is a rich source of information on mobility and broad socio-economic trends. Geospatial data is useful for greatly improving the resolution, in time and space, of already existing low-resolution survey data. As we lead more of our lives online, data from mobile phone and computer applications, as well as from internet activity generally, will shed light on consumer behavior and civic engagement. The most important gender data gaps fall at the intersection of these various uses of big data.

2. Two additional projects focused on sex-disaggregated credit scoring and on uptake of digital financial services are still in progress and due to complete in 2020.

Gender Gaps in Urban Mobility

THE GOVLAB, UNICEF, UNIVERSIDAD DEL DESARROLLO, TELEFÓNICA R&D, ISI FOUNDATION, UNIVERSITY OF BOLOGNA

Introduction

Mobility is gendered. For example, the household division of labor in many societies leads women and girls to take more multi-purpose, multi-stop trips than men. Women-headed households also tend to work more in the informal sector, with limited access to transportation subsidies, and use of public transit is further reduced by the risk of violence in public spaces.

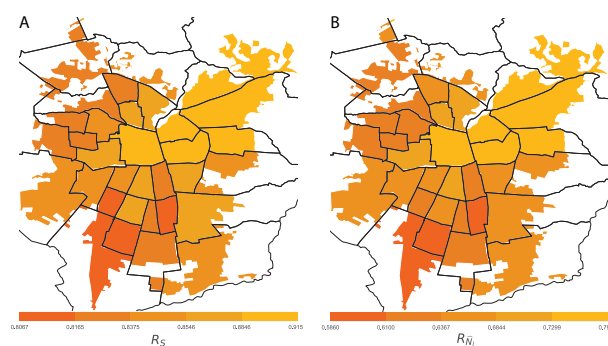
This brief summarizes a recent analysis of gendered urban mobility in 51 (out of 52) neighborhoods of Santiago, Chile, relying on the call detail records (CDRs) of a large sample of mobile phone users over a period of three months. We found that: 1) women move less overall than men; 2) have a smaller radius of movement; and 3) tend to concentrate their time in a smaller set of locations. These mobility gaps are linked to lower average incomes and fewer public and private transportation options. These insights, taken from large volumes of passively generated, inexpensive data streaming in real-time, can help policymakers design more gender-inclusive urban transit systems.

Our Approach

We analyzed the mobility patterns of 418,624 residents of Santiago, using around 2.15 billion CDRs collected between May 1 and July 30, 2016. The sample contains sex and socioeconomic information of individuals, and is strongly representative of the population, sex ratio, and income distribution (Figure 1A) of the Santiago Urban Metropolitan Area Region at the neighborhood (comuna) level. Women made up 51% of the user sample. For each individual, we computed the number of distinct locations visited, the number of distinct locations that make up 80% of a user's calls (Figure 1B), how equally trips were distributed across locations, and the radius

of movement. The study area was divided into 726 cells, approximately one square kilometer in size and regularly spaced based on the positions of cell towers. We also labeled public transit stops in the city area.

Figure 1. (A) Socioeconomic status in Santiago; lighter colors indicate wealthier comunas. (B) Measure indicating the number of distinct locations that make up 80% of calls, by comuna; lighter colors indicate a greater number of distinct locations. Note the relationship between wealthier areas and a greater number of distinct locations.



Results

Over the three month study period, women traveled to, on average, nine fewer locations than men. Women's trips were more localized, with their radius of movement about 1.1 km shorter than men. Women also distribute their trips among a few highly preferred locations, while men distribute their trips among many locations with almost equal probability. These differences are not simply due to gender differences in mobile phone use; when we examine comparable subsamples (for example, of the most active phone users), the gender gaps persist.

We investigated the causes of these patterns, and found that gender inequalities grow wider as socioeconomic status worsens. For example, poorer women tend to be more localized, relative to their male counterparts, than better-off women. However, gender gaps persist even

among the wealthiest classes; income alone does not lead to mobility equality. We also found that mobility inequality is significantly correlated to the gender gap in employment.

Child care duties are also important; a higher fertility rate and larger households (suggesting more dependents) are associated with a greater mobility gap. Education, however, is not a significant predictor of mobility inequality. Public transport options increase the mobility of both women and men, though they do not close the gap entirely—having a stop nearby is associated with 1.39 more locations visited for men, but only 0.76 more for women. The inequality is even more pronounced when considering socioeconomic status; having a stop nearby tends to close the gender gap for the wealthier classes, but the same is not true for poorer individuals (Figure 2).

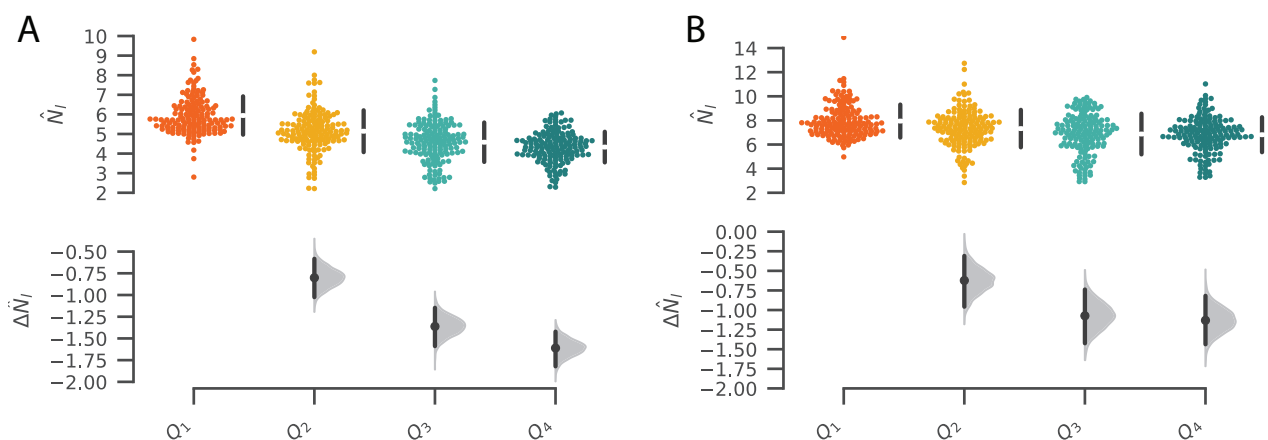
Implications

Women’s and men’s travel behavior in the metropolitan area of Santiago differ, and such differences can be exposed by the analysis of anonymized, sex-disaggregated mobile phone data. These data are sensitive enough to capture the different mobility patterns of men and women, thus providing urban planners the information

needed to design gender-responsive solutions to overcome existing cultural, infrastructure, resource, and safety constraints. Further, our analysis shows that income, employment, and gender mobility equality are all positively correlated, confirming the value of mobile phone-derived mobility metrics as a proxy for human development.

The use of mobile phone data to study gender mobility comes with limitations. First, the user sample might not be representative of the population under study: the sample size and composition will depend on the operator market share. Second, mobility patterns are inferred through user calling activity, which is known to be affected by age and gender, among other individual features. In our work, we controlled for these biases, but we could not address some potential confounding effects, related to users’ age for instance, a variable not available to us. Overall, our work—the result of a data collaborative including academic research centers, international organizations, and private companies—shows the potential of digital data in closing critical gaps in our understanding of the lives of women and girls.

Figure 2. Estimation plots of the difference in the number of locations visited by women (A) and men (B). Each dot is a geographical cell ranked by socioeconomic quartile. Q1 represents the richest quartile and Q4 the poorest.



Mobile Money & Gender in Uganda

DALBERG DATA INSIGHTS

Introduction

Women across the world face social and legal barriers to accessing financial services. Mobile money and other phone-based applications can help, but sex-disaggregated mobile phone data is critical to understand how such applications can effectively address women's financial needs. Such data, however, is rare. Even when available, the information is often unreliable because of sharing of SIM cards, restrictions to women registering accounts in their own name, and other social and cultural factors.

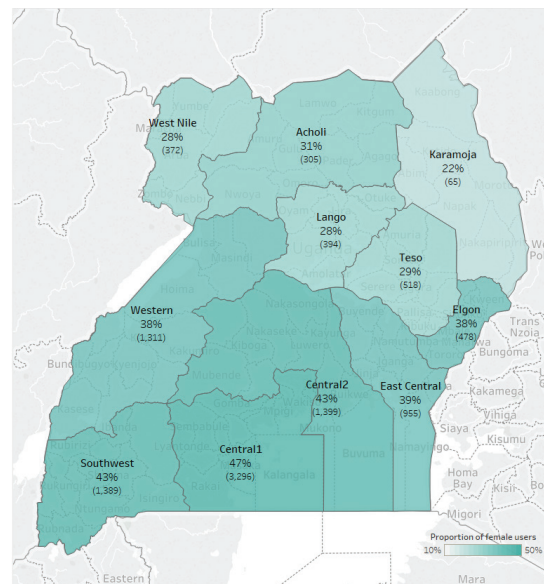
In this study, we outlined the differences in mobile phone and mobile money usage between men and women in Uganda. The research was comprised of three activities. First, we implemented a survey to obtain reliable demographic information about 10,500 subscribers in a major mobile network. Second, we fed this data into a machine learning algorithm, which computes more than 200 phone usage indicators and utilizes machine learning algorithms to predict the sex of subscribers. Finally, we use the algorithm to describe sex-specific patterns in phone and mobile money usage. We found that women tend to be engaged in fewer calls, the majority of their calls are incoming, and their incoming calls have longer average duration. Women also top-up (buy credit for calling and messaging) less frequently and in smaller amounts, have fewer contacts, and travel less.

Our Approach

We obtained information on sex, age, phone sharing, and occupational status from a random sample of 10,500 subscribers of a major telecom operator in Uganda. Nearly 41% of the subscribers in this survey were women, though this varied by region of the country (Figure 1). Validation work

found that about one-third of the respondents did not have a correct sex label in the operator's database. For the main analysis, we characterized usage patterns among this sample from 60 days of anonymized call detail records (CDRs), which contain the logs of all phone activity during the period. We analyzed voice and text CDRs, data CDRs, airtime credit recharges, and mobile money transaction records. Overall, we used 201 indicators, spanning phone usage, mobility patterns, users' social networks, top-up patterns, and mobile money usage.

Figure 1. Representation of women in the ground-truth survey, expressed as percentage of total within each region.



Absolute numbers in parentheses.

Results

The model reached an accuracy of 72% in predicting subscribers' sex. Average top-up value, the number of unique contacts per active day, average call duration during the week, the size of the geographic area in which users moved, and the duration of incoming calls were all important predictors of sex.

With respect to the gender analysis, we found that women have longer calls and a smaller proportion of outgoing phone activity. Men overall connect with more people on a daily basis, and top-up in larger amounts and more frequently. Elderly men travel longer distances in both urban and rural settings than women, though distances are generally longer in rural areas for all groups. However, these differences disappear when considering young women. Young women in both urban and rural settings have a larger network of contacts than their elderly male counterparts, and elderly men typically top up less frequently (though with still higher value) than young and adult women.

Urban residency plays a strong role, with urban women topping up as frequently as their rural male counterparts. In addition, no major differences of data usage between men and women exist—surprising given that gender inequality in mobile internet usage is generally thought to be greater than the gap in phone ownership and usage. Figure 2 summarizes gender differences in a few key indicators: average duration of calls, proportion of calls that are outgoing, unique contacts per day, and the number of total top-ups.

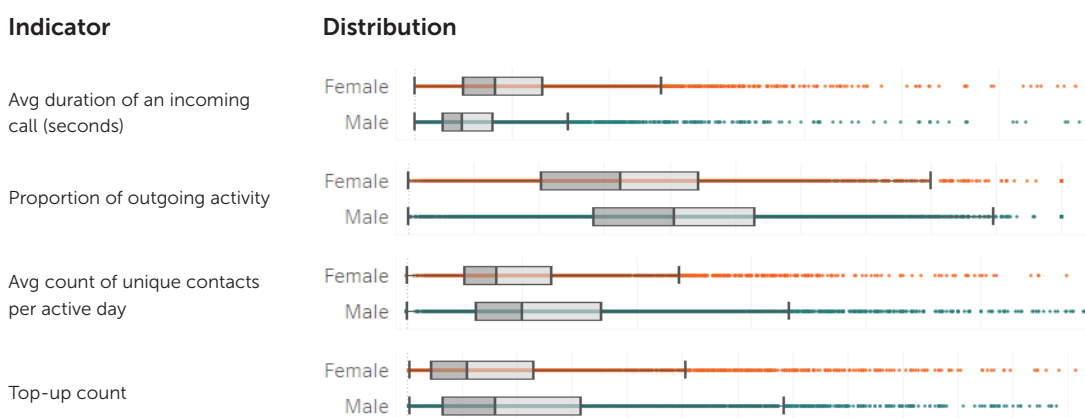
We also found that more women use mobile money than men, but men both deposit and withdraw higher amounts of money, and transact more overall. However, men and women do not differ in their average account balance, the number of accounts they interact with, or the number of

withdrawals they make. Interestingly, women pay on average a higher fee per transaction.

Implications

Increasing financial inclusion among women depends on a better understanding of gendered phone and mobile money usage. The above analysis suggests improving access to digital technologies will not, by itself, fully leverage the power of these technologies to equalize social opportunity. Policymakers and private companies need to consider the barriers that constrain the size and structure of women's social networks, limit their mobility, and determine the differentials in use of voice, text, and data services. At the same time, young women—especially those living in urban areas—do use mobile phone services almost as much as men, and this represents an important opportunity for network operators, governments, and other actors to collaborate for social gender equality. Cell phone data can help describe all these phenomena, and similar analyses in other countries would provide a detailed portrait of women's economic and social lives. Mobile operators that wish to generate accurate estimates of their subscribers' gender can access the Gender Analysis and Identification Toolkit (GAIT), developed in partnership with the GSMA.¹ Complementary research, especially using qualitative interview and focus group methods, would help elucidate the cultural and economic factors that determine women's agency in the uptake and use of technology.

Figure 2. Boxplots of selected phone usage indicators, by sex. The boxes show the range of the middle 50% of users, with the line inside the box indicating the median user.



1. Further details can be found at <https://www.gsma.com/mobilefordevelopment/resources/the-gsmas-gender-analysis-and-identification-toolkit-gait/>

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

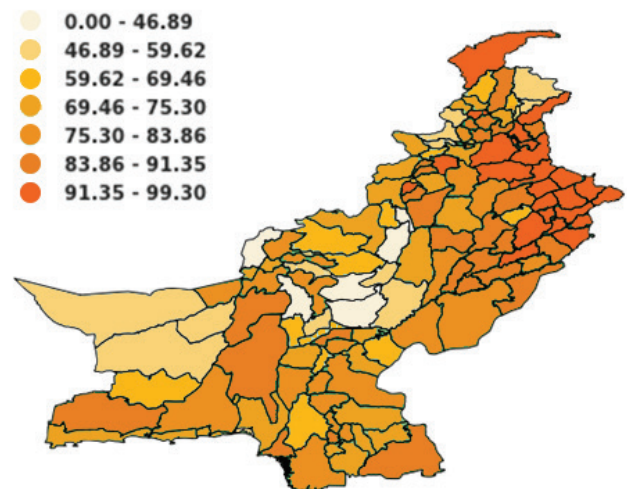
We used call detail record (CDR) data from a major operator in Pakistan, containing more than one billion voice and text messages from around six million users. In addition to the anonymized caller and recipient ids, CDRs also contain the timing of the activity and the location of the cell tower through 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.

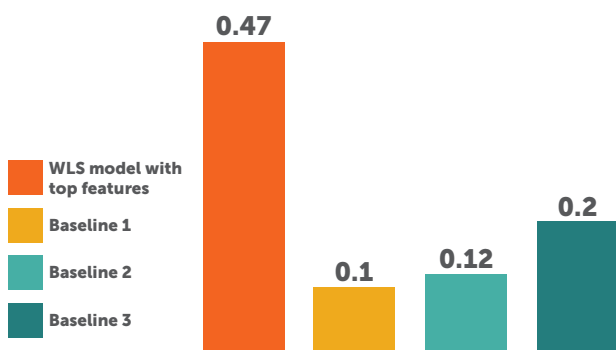


Results

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.

CASE STUDY 4

Street Harassment and Women's Educational Choices: A Geospatial Analysis

GIRIJA BORKER, WORLD BANK/BROWN UNIVERSITY

Introduction

How does the threat of street harassment affect women's lives? Using a combination of student surveys, Google Maps data, and crowdsourced information from mobile applications, this study looks at how safety concerns influence educational choices among women in New Delhi, India. The research finds that women choose poorer quality colleges, spend considerably more on transportation, and accept longer commute times in order to travel by routes that are perceived to be safer. In addition to the risks of harassment and assault, an unsafe public sphere inflicts serious educational and economic consequences on women. These costs are felt both immediately and over the course of a lifetime, in the form of reduced labor force participation and earnings.

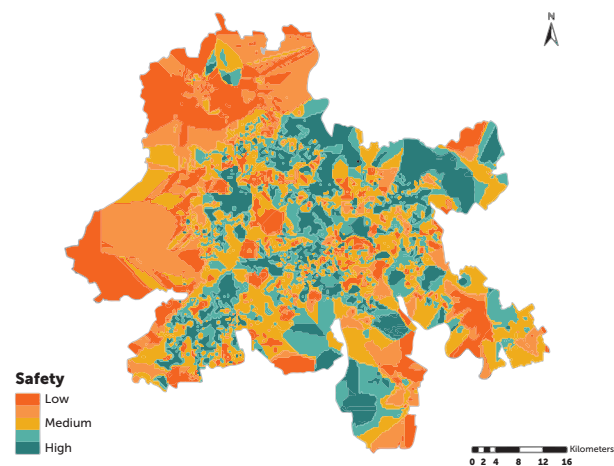
Our Approach

Delhi University (DU) is an umbrella entity comprised of several colleges spread across the city. Each college has its own campus and operates like an independent university, and selectivity in admissions provides a reliable indicator of college quality. A survey of 4,000 DU students identified the "choice set" of colleges available for each student, or the set of colleges that a student is eligible to attend based on their high school exam scores. The colleges captured in the sample span the range of quality in DU as a whole and the students in the sample are representative of the wider student body in the University.

An algorithm developed for this study used Google Maps to map all possible routes available for students to take to each college in their choice

set, where routes are defined as a combination of landmarks and travel modes.

Figure 1. Safety surface of the Delhi metropolitan area.

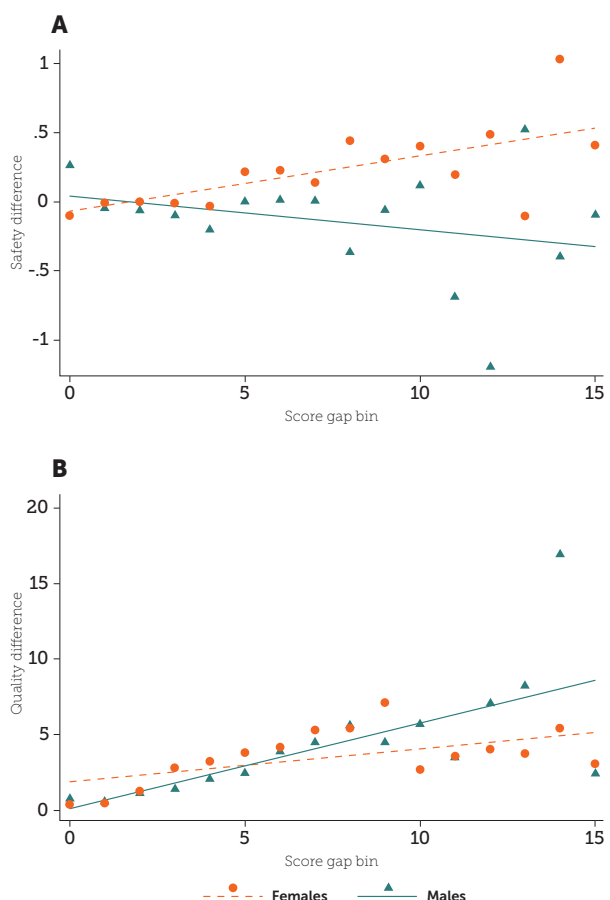


These routes are overlain on a "safety surface" (Figure 1) constructed using a crowd-sourcing mobile application, SafetiPin. Over 26,500 SafetiPin audits from November 2013 to January 2016 provided information about openness of spaces, visibility, presence of security personnel, condition of the walking path, presence of other people (especially women and children), access to public transport, lighting, and the overall feeling of safety. Students' actual and potential routes are assigned a safety score on the basis of the safety surface and safety associated with a travel mode. The latter is estimated using data from Safecity, which has over 5,500 reports of harassment on different modes of transport, including buses, rickshaws, metro, and walking. A comparison of male and female students with similar characteristics (same neighborhood, major, admission year, etc.) allows comparative analysis of the costs students pay — in terms of quality of college chosen, money, and time — for safety.

Results

Despite having better scores on high school exams, women choose worse quality colleges than men — both in absolute terms and within the set of colleges for which they are eligible. Generally, women are willing to trade a college in the top 20% of quality for one in the bottom 50% — on average, nearly nine ranks lower in their choice set — in order to travel by a route that offers about a one standard deviation (SD) increase in safety, which represents around a 3.1% decrease in the risk of rape. Men, on the other hand, choose colleges only about one rank lower in their choice set for the same degree of added safety. Figure 2 illustrates these gendered differences in decision making.

Figure 2. Gendered educational choice.



The horizontal axis represents differences in high school exam scores between students and their neighbors (other students of same gender, from the same residential area, studying the same major with the same admission year).

A higher score gap bin represents higher-achieving students with an expanded choice set of colleges who, absent safety concerns, should choose higher-quality colleges. The vertical axis shows differences in the safety of the chosen travel route (Panel A) and quality of attended colleges (Panel B) between students and neighbors. The graph in Panel A shows that as their choice set expands, female students (red dotted line) choose safer routes to travel by compared to males (solid blue line). Panel B shows the relationship between high school quality and college choice is weak for women, and much stronger for men.

Women also spend INR 20,000 (USD 310) more per year for added safety, nearly 16 times as much as men — a gap that is nearly double the average annual tuition at DU, and about 75% of annual per capita income in Delhi. Women are also willing to travel 40 minutes more daily for a safer route, in comparison to just four minutes more for men.

Implications

In Delhi, 95 percent of women between the ages of 16 and 49 report feeling unsafe in public spaces. In the sample used in this study, nearly nine out of ten female college students have faced some form of harassment, and 40% have been touched, groped, or grabbed. Public areas represent serious daily risks to the security of women and girls.

This study shows that the indirect costs of harassment are also very high. Choosing a worse quality college affects later academic training, peer network, access to jobs, and lifetime earnings — and may affect aggregate economic productivity for a society. In India, labor force participation rates for working-age women have stagnated at 26-28% in urban areas between 1987 and 2011. The drivers of this low participation rate are not clear, but lack of physical security in public areas may be a contributor. This study finds that policies to increase safety of travel routes — improving street lighting, funding self-defense programs, and assuring security on public transit, for example — can have powerful impacts on reducing gender gaps in school quality, costs of transport, and commute time.

Towards High-Resolution Sex-Disaggregated Dynamic Mapping

FLOWMINDER FOUNDATION

Introduction

In this brief, we present results of a mixed-methods study to investigate how novel digital data sources can support gender-equitable development across Nepal. We implemented two bodies of work. First, we combined geolocated survey data, satellite imagery, and mobile phone data to map three key gendered indicators — literacy, agriculture-based occupations, and births in health facilities — at very high spatial resolution. Second, we sought to use de-identified mobile phone data to produce robust, frequently updatable information on gendered mobility and migration patterns within Nepal. This second body of work required us to predict gender among a population of mobile phone subscribers. Our results suggest that SIM sharing is an important complicating factor in predicting gender, and thus inferring individual well-being, from mobile operator data. Overall, we find that combining traditional survey data sources with various forms of digital data holds great promise for a spatially and temporally rich understanding of women's and girls' lives, although more validation work is needed on patterns of SIM use.

Our Approach

In this study, we worked with geo-tagged survey data collected for the 2016 Nepal Demographic and Health Survey (NDHS), focusing on seven indicators: educational attainment, literacy, labor market participation, agriculture-based occupations, attitudes to gender-based violence, births in health facilities, and child stunting. Because of cost and logistical considerations, demographic and health surveys are typically not designed to permit highly local inferences. However, the spatial resolution of survey data

can be improved with the complementary use of geospatial and mobile phone data.

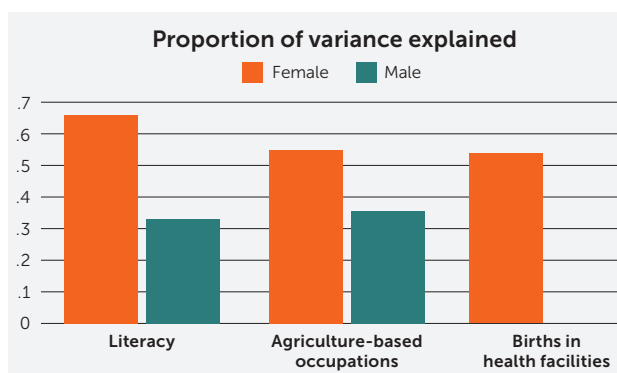
We collated a variety of geospatial information from open-source platforms, including datasets containing physical (topography, climate, land cover, vegetation, biomass, evapotranspiration), social (population, ethnicity), and built environment (urbanization, human settlements) variables. We also partnered with one of Nepal's largest mobile network operators, Ncell, to analyze a subscriber database of 15 million registered SIMs, with information between January-December 2016 on time, duration, location, and parties of each call, as well as daily financial credit 'top-up' totals and counts of recharge type.

While subscriber records may include a gender tag, this information is not validated, and is often entirely missing. We thus built a model to predict gender based on daily and home locations of subscribers, calling patterns, and top-up behavior. To support this model — specifically, to validate existing gender tags — we conducted a survey of a subset of 5,180 subscribers.

Finally, using machine learning and Bayesian geostatistical methods, we built models analyzing the geospatial and mobile phone variables as predictors of the survey indicators of interest. We used these observed relationships to create high-resolution (1km²) gridded datasets of the indicators, with associated uncertainty. The sex-disaggregated maps are updatable, and thus able to track progress on development targets over time, as well as assess short-term fluctuations in well-being associated with economic and environmental shocks.

Results

Figure 1. The proportion of variance explained by the applied models.



The modeling results are generally encouraging: analysis of individual CDR data can enhance our understanding of the spatial variation and temporal dynamics of gender inequality. We found that female literacy, agriculture-based occupations, and births in health facilities were especially amenable to our approach, with the best-performing models explaining around 60% of variance (Figure 1). The models predicting literacy and farm-based livelihoods performed much better for women than men. The reasons for this are not clear but are likely related to the fact that social and economic institutions — especially a strong historical male-child preference in South Asia — play an intervening role in the relationship between geospatial and well-being variables. Educational attainment, child stunting, and gender-based violence were only weakly associated with our chosen set of predictors. However, these indicators may perform better

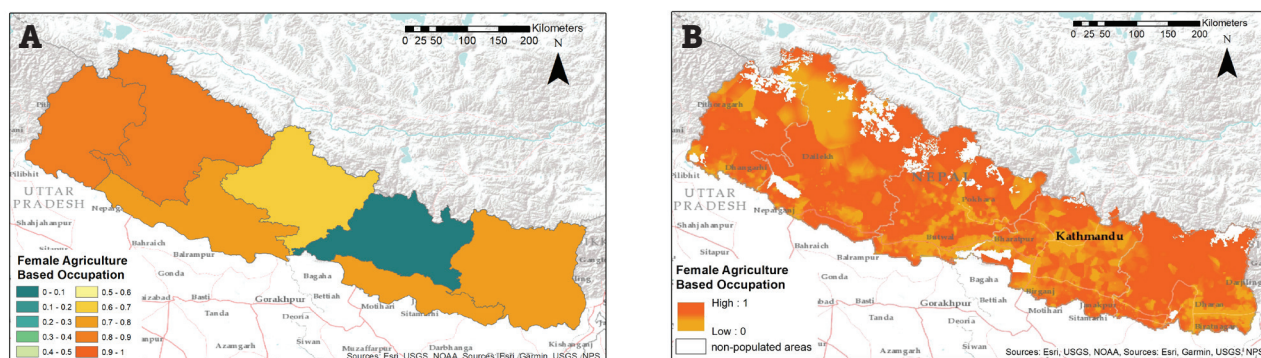
in other contexts, as suggested by our previous work.¹

We also found that SIM sharing within households greatly complicates inferences from mobile operator data, despite the very high overall rates of individual mobile phone ownership in Nepal. Most demographic prediction models have assumed that a single CDR record corresponds to a single individual (the ‘single-SIM/single-subscriber-user’ assumption). Our results suggest that this assumption should be strongly questioned, particularly in Nepal but possibly also in other contexts. Almost one-half of survey respondents indicated shared SIM use, with no difference between men and women in the likelihood of sharing. Most of the sharing (92%) occurred with the family. In addition, nearly one-third of gender tags in the MNO database did not correctly identify the gender of the main user of the SIM.

Implications

The combination of geo-located survey, geospatial, and mobile phone data holds great promise for creating well-being monitoring systems with high resolution in both space and time. Figure 2 shows how the portrait of female participation in agriculture-based occupations is enriched by our modeling approach. Such high-resolution data systems are necessary to monitor progress towards the Sustainable Development Goals, and more generally allocate resources to the places and at the times when they are most needed.

Figure 2. (A) Map of female engagement in farm-based livelihoods at province level (source: authors analysis of weighted NDHS survey data). (B) Map of the population of female engagement in in farm-based livelihoods at 1km2 resolution



However, the use of digital data for spatial modeling must confront several obstacles. First, the underlying survey data is typically limited in sample size, which inhibits the full exploitation of the model architectures. This reinforces the critical point that traditional and new forms of data are complementary, not competing. Second, because of the large number of geospatial covariates and cell phone features available, stronger theory about which combinations of variables are most likely to predict a given indicator would be valuable. Finally, the use of CDR data for improving our understanding of gendered phenomena depends on accurately identifying user gender. Even when tags are available, they may be inaccurate for a variety of reasons, including SIM sharing. Improving our understanding of SIM sharing, which is likely to vary greatly by context — and even in a given context, may change over time as economies evolve — is thus vital. When tags are not available, gender prediction models may help, but again requires careful validation research on calling and sharing behaviors.

These limitations notwithstanding, the potential of these new types of digital data is clear. The data used in this study — geospatial and mobile phone information — is readily available in massive quantity at low cost, and these datasets will continue to increase in size and resolution in the coming years. Building capacity in public sector agencies is an important step to fully realizing this potential. In Nepal, the Flowminder Foundation is working closely with the Central Bureau of Statistics (CBS) to use the modeling approaches outlined in this brief to create a well-being monitoring system embedded within the agency.

Using Facebook and Google Advertising Data to Measure the Gender Digital Divide

RIDHI KASHYAP (UNIVERSITY OF OXFORD), INGMAR WEBER (QATAR COMPUTING RESEARCH INSTITUTE), MASOOMALI FATEHKIA (QATAR COMPUTING RESEARCH INSTITUTE), REHAM AL TAMIME (UNIVERSITY OF SOUTHAMPTON), IAN KNOWLES (UNIVERSITY OF OXFORD)

Introduction

Sustainable Development Goal 5, achieving gender equality, highlights the importance of closing the global gender digital divide — that is, ensuring that women and girls have equal access to the Internet, mobile phones, and other digital technologies. However, the paucity and irregular production of data on these topics, particularly in less developed countries, makes monitoring progress towards this target difficult. In this brief, we show how anonymous, aggregated data from Facebook and Google's online advertising platforms can help fill the gap. We find that Facebook and Google advertising data are strongly correlated with gender gaps in internet access and digital skills. On our website digitalgendergaps.org, we take advantage of the better temporal resolution of the Facebook

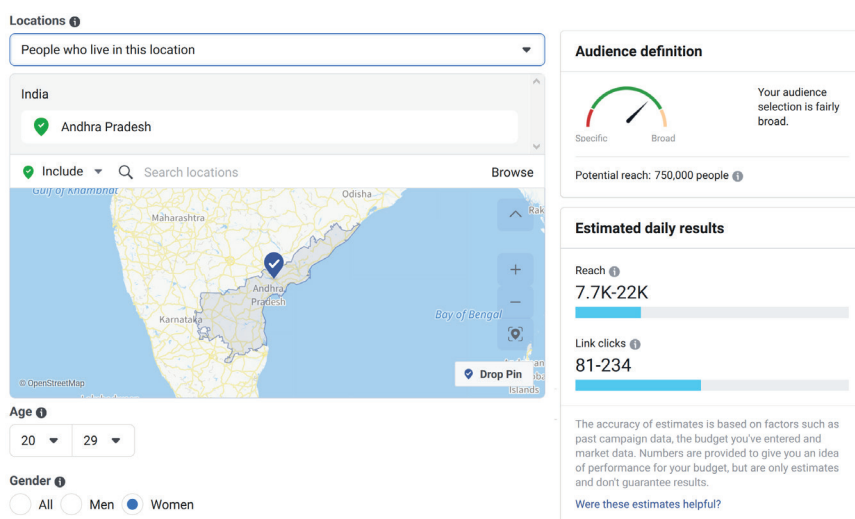
data to provide regularly updated indicators of internet and mobile use gender gaps globally, tracking digital gender inequalities as they evolve over time. Such a tracking system is much less expensive than conventional surveys and can help monitor the effect of policy interventions and economic shocks.

Our Approach

The Facebook and Google advertising platforms are designed to offer potential advertisers data on the group of users that they are interested in targeting. For example, the Facebook advertising platform can provide an answer to the question, "How many female Facebook users between the ages of 20-29 were active in Andhra Pradesh state of India in September 2019?" (Figure 1). When compared with data on men, this real-time information sheds light on the digital gender divide. In addition to age, gender, and location, we can also access information related to the

device types, for example mobile devices, which are used to access Facebook. This is helpful for measuring different forms of digital access. Google's advertising platform, Google AdWords, which has an even broader global reach than Facebook, allows similar insights, though it differs slightly in showing estimates for ad impressions (the number of times an ad is seen by a user) rather than the

Figure 1. Facebook API showing the number of female Facebook users between the ages of 20-29 active in the state of Andhra Pradesh, India, in August 2019.



number of users; more active users create more ad impressions. Similar information is provided by the advertising platforms of Twitter, LinkedIn, Snapchat, and others. The spatial resolution ranges from state level to sub-city postal codes, depending on country and advertising platform. To prevent reidentification of individual users, the platform limits audience estimates to no less than 100 users. This data can be used to construct gender indicators. We used the Facebook data to generate a “Facebook Gender Gap Index” (FB GGI), measuring the ratio of female to male monthly active Facebook users in a given country. For example, in Belgium we observed 3.6M female and 3.5M male monthly active Facebook users, whereas for India there were 65M female and 220M male monthly active Facebook users, as of September 2019. We examined how well this Facebook data, in combination with offline gender and economic indicators (e.g. the UN Human Development Index, gender gaps in education), predicted the latest available survey-based estimates of gender gaps in internet access and digital skills.

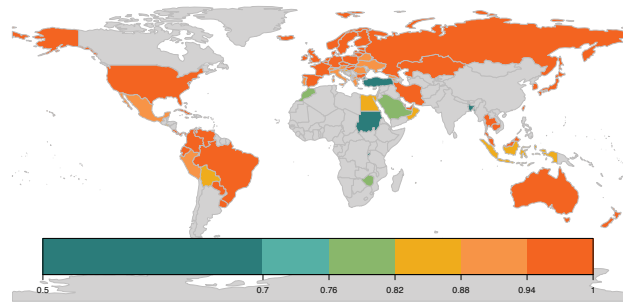
Results

Our results show that both Facebook and Google online indicators are strongly correlated with survey data on internet access gender gaps and low-level digital skills (e.g., using copy and paste tools, transferring files, and sending email). When women are missing on these two online platforms, we can state with a high degree of certainty that they are not online at all, and lack digital skills in these countries. Although models that use Facebook indicators show better predictive performance than Google AdWords, models that combine Facebook and Google online data with offline development indicators perform best in predicting internet access gender gaps. In particular, the combination of Facebook and AdWords data and a country’s Human Development Index explains about 80% of the variation in global internet gender gaps. Figure 2 shows how the global landscape of Internet access differs when viewed by survey data (2a) versus Facebook data (2b). Facebook indicators are better able to predict low-level digital skills

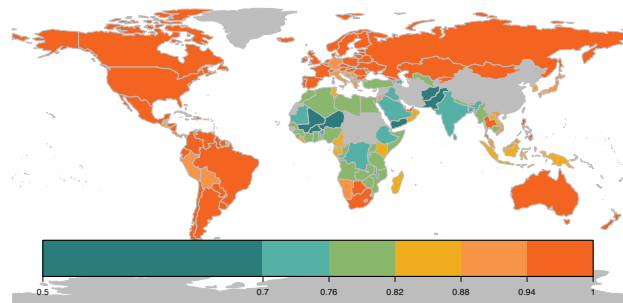
compared with AdWords indicators. Our work highlights how women are disproportionately less online in countries in South Asia and sub-Saharan Africa compared with men, where values of the internet gender gap (the proportion of the female population with internet access divided by the male population with internet access) index lie between 0.7 and 0.8 (see Figure 2).

Figure 2. Two world maps showing the ratio of (percentage of women with Internet access)/(percentage of men with Internet access) on a per-country basis. ITU data from 2015 (top) is compared to model predictions of the online model using Facebook data from 2017 (bottom). The model manages to largely reproduce ITU ground truth data while substantially improving global coverage.

A) Internet access gender gaps according to 2015 ITU data.

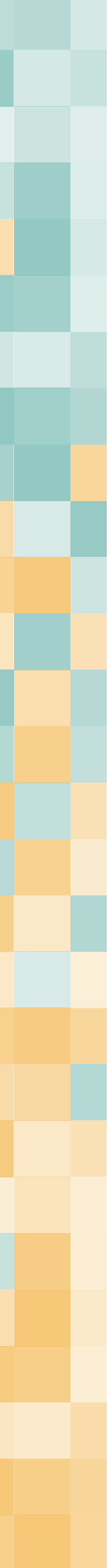


B) Internet access gender gaps according to 2017 Facebook data.



Implications

Our research shows the great value of online advertising audience estimates in complementing existing traditional data sources on the lives of women and girls. All the data sources described here are publicly available free of cost, which enables near-real-time estimates to be made and, more generally, democratizes data access and analysis. When used ethically and responsibly, and in combination with existing data sources, online advertising audience estimates can help to fill gaps on important topics such as digital gender inequalities. These data streams could be used to provide crucial input into policies targeting populations with poor access to technology for infrastructural and educational investment.



Overall, however, this method has limitations. Perhaps most importantly, insights from online data are more likely to represent the behavior of better-off people. However, this type of bias and data distortion is not necessarily problematic for two reasons. First, it is often exactly the missing data that is the signal. For example, in our research on gender gaps, the fact that women are not found in the data at the same rate as men provides a signal on gender inequalities. Second, approaches using supervised machine learning, such as regression models, treat the (biased) data merely as a signal to predict a particular quantity of interest, e.g. internet gender gaps derived from representative survey data. As long as the signal has high predictive power, it has potential value for the task. For such approaches, selection bias is only a challenge when it is non-systematic, e.g. when the reasons for bias differ across countries, and cannot be understood well enough to be corrected. On the other hand, if the reasons for bias are globally consistent, or if the contextual forces driving bias are well-understood, then estimates can be adjusted to account for the bias.

Another important challenge is the fact that the data provided to advertisers comes from proprietary methods. Academics and others cannot easily audit data quality. Whereas some user attributes such as age and gender are most likely derived from self-declared information, more detailed attributes — for example, Facebook’s “lived in [country name]” category, which measures users who have previously lived in a given country but now live in another — are based on a proprietary inference algorithm with unknown accuracy.

Women and the Digital Gig Economy in South Africa

EMMA SAMMAN AND ABIGAIL HUNT, OVERSEAS DEVELOPMENT INSTITUTE¹

Introduction

The gig economy, characterized by digital platforms that bring together workers and the purchasers of their services, is expanding globally. A majority of gig workers engage with these platforms through necessity. These workers often face low and insecure incomes, discrimination, and limited labor protections. Little research to date has focused on gig workers outside of North America and Europe, nor on gendered experiences of gig work. In this study, we analyze data from a two-year project to understand women's involvement in gig work in Kenya and South Africa. We probed workers' time use, how work experiences differ from the employment these workers might otherwise find, and the extent to which gig work offers independent, flexible working patterns that support women workers in managing paid and unpaid work. This briefing focuses on our results for South Africa.²

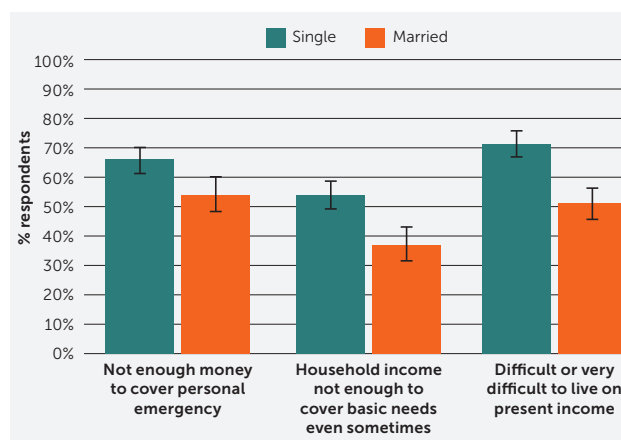
Our Approach

We collected socio-demographic data from 650 domestic gig workers, nearly all of whom were black African women, through a nine-round longitudinal phone-based survey. Our digital platform collaborator provided us with data on their workforce, which we merged with the survey data at an individual level to obtain a fuller picture of worker experiences and company operations. We also analyzed the workforce data to obtain a more comprehensive picture of the services offered and worker availability. We carried out in-depth interviews with gig workers who care for at least one child under 8 years old to understand the trade-offs between gig work and family life.

Results

We find that platforms offered some positive features that workers value, and that improve their working conditions. Platforms offered paid work in an economy with high unemployment, greater hourly pay than that which is available in other jobs, and some flexibility for child care and educational pursuits. Many workers also felt that platforms afforded them the opportunity to improve their skills. The platform companies themselves spoke of the goal of 'professionalizing' the gig workforce through certifying their offerings, providing training in customer relations and other areas, and more broadly dignifying service-based work, which was seen as critical to be able to increase wages eventually.

Figure 1. Economic situation of single and partnered gig workers.



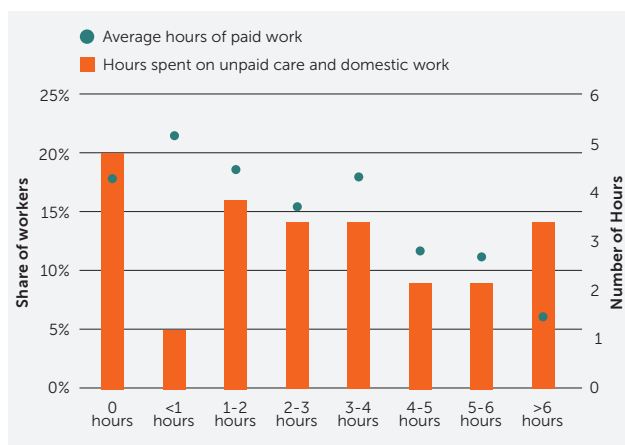
However, workers – particularly the majority who were single (and typically parents) – struggled with the unpredictability of earnings from gig work and felt that their overall earnings were not enough to meet household needs (Figure 1). In turn, the platform company cited an inability

1. For the full report, please see: Hunt, A., Samman, E., Tapfuma, S., Mwaura, G. and Omenya, R. with Kim, K., Stevano, S. and Roumer, A. (Forthcoming, 2019), The gig economy, care and time use in Kenya and South Africa (Working Title), London: ODI. The quantitative data was collected in partnership with Ulula and Data-Pop Alliance.

2. Complications with the study in Kenya restricted the quantitative data we were able to collect from our survey and from the platform company with which we collaborated.

to raise prices to provide a better income for workers, as this would price some clients out of the market, resulting in less work being available overall. In addition, while platforms highlight the flexibility they offer workers, this flexibility is limited in practice by client demand, which determines the volume, location, and timing of bookings. Workers also face safety issues that stem from the fact that many on-demand gigs take place in private households. The large distances between gigs exposed workers to additional security risks, including robbery and assault, which were accentuated further by poor public transport and early working hours. Gig workers also lack social insurance that would provide protections, for example in the case of ill health or childbirth. The latter proved a particular hardship, as workers often had limited means of coping during maternity periods in which they could not work.

Figure 2. Weekly hours spent working on paid work and unpaid care and domestic work.



Workers also struggled to balance gig work with childcare and other unpaid household work. While women's unpaid care was inversely correlated with their paid work, nearly a third of all gig workers did at least four hours of unpaid work daily (Figure 2). Moreover, the flexibility that gig work offers is often not used in practice because workers are reluctant to turn down bookings and forego badly needed income. Women relied on family, neighbors, and friends to provide childcare, but when such support was lacking, they relied on high-risk strategies, such as leaving young children alone or in the care of slightly older children.

Implications

To make the gig economy work for marginalized people, we argue that policymakers and platform companies have a responsibility to improve economic security, support unpaid childcare, give workers more flexibility over schedules, and ensure worker safety. Governments ought both to focus on the broader contextual factors that influence working conditions (through better infrastructure for transport, childcare etc.) and to put in place regulations that ensure that platform companies provide a minimum set of protections to their workforce, in line with relevant national legislation on labor rights.

Economic security requires:

1. Greater investment in training, certification, and skills development;
2. The promotion of workers by emphasizing their education, skills, and experience in addition to customer ratings;
3. Cost analysis to ensure that workers are being paid at least the national minimum wage (and that the statutory minimum equates to a living wage);
4. Active participation in initiatives to raise standards in the gig economy.

Supporting unpaid childcare includes:

1. Analysis of the costs and benefits of quality childcare services;
2. Canvassing worker demand for and preferences about childcare;
3. Advancing public and private initiatives that provide affordable childcare for working families.

Giving workers flexibility entails:

1. User-friendly platform features that allow workers to select and cancel gigs with ease;
2. New strategies to incentivize clients to accept work at hours preferred by workers;
3. Addressing the issue of client underreporting of the amount of work required.

The converse of this is that workers need some assurance of **stability in their work**, such as a guarantee of a minimum number of hours and compensation when clients cancel gigs they are expecting to do.

Ensuring worker safety requires:

1. Government investment in transport infrastructure (with support from companies, where feasible);
2. Providing workers with information on violence hotspots;
3. Creating platform features for workers to access support in the event of safety threats;
4. Taking reports of client abuses seriously;
5. Tailored workplace safety laws that recognize the specific risks gig workers face.

Finally, we note that our research efforts were complicated by data access issues. These include difficulties in reaching a dispersed worker population and in accessing platform data itself. While we were able to collaborate effectively with platform companies, there are sound ethical and legal reasons for gig platforms to limit access to data, including worker privacy considerations. However, newly available data sharing models could allow researchers to query data through specific algorithms, running on the servers of the gig platforms, that would perform analytical tasks without sharing or exposing data. This would also help resolve the intrinsic tension between understanding and improving the experience of vulnerable populations, preventing further disenfranchisement, and managing relationships with data custodians. Such a system would also help to involve gig workers further in all future research using data about them.

Introduction

Our Approach

1) Retrieval and Classification of Gender-related Content

Figure 1. The tool used by volunteers to annotate text.

BIG DATA, BIG IMPACT? TOWARDS GENDER-SENSITIVE DATA SYSTEMS

2) Sentiment Analysis of the Collected Data

Sentiment analysis (SA) is the computational study of people's opinions, sentiments, emotions, and attitudes, as expressed in text. It provides a way to automatically evaluate and capture change in public opinion towards women and their rights. SA of Arabic poses some unique challenges given its structural and stylistic properties.

We tried two different approaches to analyze sentiments in YouTube comments and Twitter posts. First, we took two publicly available "dictionaries" containing lists of words annotated with polarity scores. We extracted the words in YouTube comments and Twitter posts, got their corresponding polarity scores from the dictionary, and then aggregated these scores to obtain the overall score of the initial text. We also considered the context in which a text appears — for example, whether the comment is associated to a video promoting GBV. Our second approach was machine learning-based sentiment analysis. We built a training dataset by asking volunteers to manually annotate thousands of text samples (Figure 1). Based on this training set, we built various classification tools and applied them to over one hundred thousand gender-related YouTube comments and tweets. Again, the classifiers performed very well, identifying nearly all features more than 90% of the time.

3) Gender and Geographic Disaggregation

Finally, we automated the disaggregation of text by sex and location. Textual sex-disaggregation relies on the assumption that, within a particular cultural context, stylistic differences exist between men and women. To build a training dataset, we used two name-sex inference databases and then built a model that predicted sex accurately in 90% of text samples. Our approach to automated location relied on the existence of geographically distinct dialects of Arabic. We were able to build models to accurately identify dialects for all regions except Gulf countries.

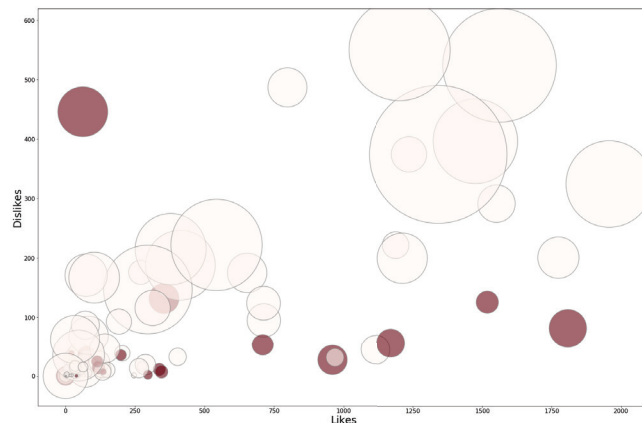
With respect to Twitter, we searched a gender-relevant keyword list to identify relevant tweets, and then applied our sentiment analysis algorithm to detect polarity. Twitter data allowed us to study relationships between positive and negative opinion holders.

These steps resulted in datasets of YouTube comments and tweets illustrating user attitudes about violence against women, classified according to sentiment polarity (positive or negative), sex of the user, and geographical location.

Results

Based on the articles defined in the Convention on the Elimination of All Forms of Discrimination against Women (CEDAW), we examined 12 topics in YouTube videos: women's work, women's education, beating women, female refugees, female genital mutilation, women in power positions, women in decision-making positions, women and politics, the eligibility of women to make decisions, women and leadership, and women and driving. Broadly, we found that negative videos were much more "liked" than positive ones. Many positive videos had high viewership (Figure 2), but the "likes to views" ratio was much higher for negative videos compared to the positive ones. "Likes to views" is a metric often used to measure popularity of videos in YouTube, and therefore we can conclude that negative videos are generally more popular. Negative sentiments were especially dominant in male-authored comments related to domestic violence and sexual harassment. We found that YouTube and Twitter sub-networks overlapped, representing ways by which both positive and negative opinions flow across populations.

Figure 2. Ratio of likes to dislikes for examined videos. The size of circles represents the number of views. Negative videos are colored dark, positive videos are light.



Implications

Automated mining of content on the World Wide Web is a cost-efficient way to learn about personal attitudes and social norms around VAW. To maximize the value of this information, data access, representativeness, and reliability should be more carefully considered.

Public APIs (application program interfaces) give access only to a small percentage of global data, and many provide access only to real-time data collection, making post-hoc analysis difficult. Moreover, APIs' usage conditions are continuously changing. These restrictions are understandable given the market importance of such data, but they represent serious barriers to research of the kind profiled in this brief.

The representativeness of online data is important for two reasons: does the online population represent the population as a whole and does the publicly available sample of web data represent the online population as a whole? As internet access becomes more universal, the first source of bias will ease and, in the meantime, can be quantified relatively straightforwardly. The second source of bias is not a major issue for YouTube, as search techniques can compile all videos relevant to a topic, however, it is problematic with respect to Twitter. Currently, researchers cannot assess the representativeness of the sample of free tweets made available through Twitter's API; more open algorithms would be useful.

Finally, the reliability of information on social media is a concern. Most false data is about demographic information (e.g., name, age, location); false opinions are less common. Sarcasm detection techniques are also rapidly evolving, allowing more nuanced analysis of user intent.

Computational approaches hold great potential for measuring attitudes and norms. In the short-term, this research group will create an interactive dashboard to make the usage of the tools described in this report more intuitive for all users. In the medium-term, we will add a forecasting tool to predict trends in attitudes and norms. We will also analyze how opinions flow across networks, as well as how to characterize the different kinds of personalities that hold positive and negative opinions about VAW.

In-Progress Case Studies

The following two case studies are in progress. We summarize their design and objectives below.

Case study 9: Uptake and Usage of Financial Services to Advance Women's Financial Inclusion

In order to improve the overall financial health of their low-income customers, policy-makers, financial institutions, and non-profit organizations must have a better understanding of savings behavior. As part of efforts to promote financial inclusion among women, Diamond Bank (as of December 2018, acquired and merged with Access Bank) and Women's World Banking, supported by Visa Inc. and Enhancing Financial Innovation and Access (EFInA), designed and launched the BETA Savings account. BETA encourages customers to save towards their goal consistently and is mainly targeted at female market entrepreneurs and traders. BETA includes a transactional savings account, a commitment savings account (Target Savers), value-added services (balance inquiry, transfers, airtime top-up), and a mobile credit product. Agents, known as BETA Friends, visit a customer's business to open accounts and handle transactions, including deposits and withdrawals, using a mobile phone application. Customers can conduct value-added services through their mobile phones.

Through its ongoing relationship with Diamond Bank, Women's World Banking was granted access to the transaction-level data of all Diamond BETA customers, representing over 600,000 accounts, including ~40% women, as well as the BETA Friend agents. We analyzed Diamond Bank's BETA individual-level data on women customers' enrollment and usage of the account to understand how they engage with digital financial services, as well as analyzing agent individual-level data to understand how well agents are serving customers.

Case study 10: Gender-Differentiated Credit Scoring Algorithms Using Call Detail Records and Machine Learning

Low-income women disproportionately lack access to credit, often because they lack credit histories, property rights, and formal earnings. This, in turn, leads to a cycle of exclusion from formal credit markets, as a lack of data to assess their creditworthiness prohibits them from building a credit history. Big data from mobile phones provide an opportunity to overcome the gender gap in access to credit, as this data has been shown to be strongly predictive of asset ownership and repayment behavior. We are partnering with one of the largest banks in the Dominican Republic to implement a gender-differentiated credit scoring model using mobile phone data and machine learning and test whether it can increase women's access to credit.

Specifically, we are testing a new approach to credit scoring that allows for men and women to have different determinants of loan eligibility. The model uses machine learning algorithms to sift through and transform a broad range of characteristics from existing clients' mobile phone data (including recently obtained high-frequency data on clients' use of La Nacional's mobile banking app) to determine the best predictors of creditworthiness separately for men and for women.

