LEVERAGING ONLINE ADVERTISING PLATFORMS TO MEASURE AND CHARACTERIZE DIGITAL INEQUALITIES

by

Yonas Mitike Kassa

in partial fulfillment of the requirements for the degree of Doctor of Philosophy in

Telematics Engineering

Universidad Carlos III de Madrid

Advisor:

Dr. Rubén Cuevas Rumín

December 2019
Leveraging Online Advertising Platforms to Measure and Characterize Digital Inequalities

Prepared by
Yonas Mitike Kassa
contact: yonas.kassa@imdea.org

Under the advice of:
Dr. Rubén Cuevas Rumín
Departamento de Ingeniería Telemática, Universidad Carlos III de Madrid

This work has been supported by:

Unless otherwise indicated, this thesis is distributed under a Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA).
— If I have seen further it is by standing on the shoulders of Giants.
Sir Isaac Newton

Acknowledgements

Being able to reach at this moment of scholarly progress is not an easy journey. And I believe the tradition of having this section is an indication of complete agreement between generations of scholars’ desires to pay respect and gratitude to those great minds and hearts behind their achievement. I was fortunate to have the support of some of the few selfless, dedicated, and caring people one would ask for.

I am grateful to my supervisor Dr. Ruben Cuevas for his continuous encouragement throughout these years, for giving me this opportunity, for always telling me that I have a lot of potential to do a lot more. I hope to internalize the exemplary dedication and hard work that I learned from him. I was also fortunate to learn from great researchers through our collaborative work. Dr. Angel, Dr. Roberto, Dr. Manuel, Dr. David, Prof. Esteban, Prof. Iyad – thanks a lot, I feel really humbled and grateful to have the opportunities to working with you. To my colleagues Costas, Patri, Toni, Pelayo, and Jose, thanks for the interesting collaborations! To my friends in the Netcom group – desde F05 hasta arriba (Victor, Borja, Jorge, Luis, Oscar, Winnie, Gines, Chris, Sergi, Kiril, Pedro, Estefno, Milan, Nuria – to name a few, and all my friends that have been my mates throughout the lab’s evolution under my watch, and to their supervisors). Also to the nice people I often meet in the corridors and during lunch time. I would like to extend my gratitude to IMDEA networks for giving me this opportunity, it is an honor to be part of this flourishing research institute.

My deepest gratitude for my family and friends that always stood by me providing all the support, guidance, and the rationale to persevere, for always checking how I am going, and more importantly for laying the strong foundation that keeps me going – thank you again, I am forever indebted.
Published and submitted content

This thesis covers contributions from the following literature:

  - this paper is fully included in chapter 5 of this thesis.
  - The material from this source included in this thesis is not singled out with typographic means and references.

- A large-scale analysis of Facebook’s user-base and user engagement growth. YM Kassa, R Cuevas, A Cuevas. IEEE Access, 2018. (ranked Q1 with Impact Factor of 4.089 according to JCR 2018.)
  - this paper is fully included in chapter 6 of this thesis.
  - The material from this source included in this thesis is not singled out with typographic means and references.

- A price comparison system for online advertising systems. YM Kassa, R Cuevas, A Cuevas. Submitted, Scientific reports.
  - this paper is partially included in chapter 4 of this thesis.
  - The material from this source included in this thesis is not singled out with typographic means and references.

  - this paper is partially included in chapter 4 of this thesis.
  - The material from this source included in this thesis is not singled out with typographic means and references.

1https://jcr.incites.thomsonreuters.com/

  • this paper is partially included in chapter 4 of this thesis.
  • The material from this source included in this thesis is not singled out with typographic means and references.

• Automated web compliance testing with pyObserver. YM Kassa, R Cuevas, A Cuevas. working paper.

  • this paper is partially included in chapter 2 of this thesis.
  • The material from this source included in this thesis is not singled out with typographic means and references.

**Other research merits**

Below we list some of software tools implemented during the development of this thesis (available at http://dataportal.netcom.it.uc3m.es/webtools/).

• Web tool to assess the value of different audiences in online advertising.

• Crawling and data collection tools for YouTube, Twitter, Facebook, Alexa, Instagram.

• Chrome browser extension that identifies and classifies third parties in a website.

• Chrome browser extension that extracts Facebook tokens for accessing the API for research purposes.

• A system to extract third party domains and third party cookies associated with a website.

• Web platform to preemptively identify trackers in a given URL.

• A web service for revealing intermediaries involved in serving an ad.

• A data driven machine learning based customer value attribution modeling system.

• A software tool for automated web compliance testing.
Selected Media Coverage

Below we list media coverage based on some of the contributions mentioned above. These reports indicate the societal impact of our contributions.

- Scientific American - Facebook Use Linked to Gender Equality, September 2018
- Mic - Researchers are using Facebook data to study gender inequality, July 2018.
- Inverse - Why Do More Men Use Facebook in These Countries?, June 2018.
- ORF - Geschlechterkluft auf Facebook, June 2018.
- El Pais - La brecha de género en Facebook refleja la desigualdad real, October 2017.
- El Pais - Un artículo científico desglosa por primera vez las cifras de Facebook, October 2018.
Abstract

As the Internet is becoming a fundamental aspect of the society serving as a de facto platform for social and business activities, traditional offline activities and services have remarkably migrated to the web. The presence and interaction of users on these platforms has created a large amount of digital trace which is being effectively exploited by businesses targeting users on these platforms. One of the main players in this regard is online advertising which is the underneath business that drives the majority of the most important online services such as social media, search engines, map services, etc. This has made online advertising a crucial Internet service in its own right. While the benefit of using these digital resources has been accepted widely pushing governments and organizations to improve their Internet coverage, there are major challenges that limit the society from enjoying their benefits. Together with transparency and privacy, digital inequality is the main challenge that the society faces today.

In this thesis we propose a set of inexpensive and large scale methodologies that leverages datasets from online advertising systems to measure and characterize digital inequality on the web. Our methodologies consider various demographic, geographic and interest categories at global scale that advances the knowledge of the scientific community to better understand the challenges in the interplay between online services and users, specifically digital inequalities.

In particular we present three main contributions in this context:

(1) **A methodology to measure the price variability assigned to users by the online advertising system.** We created an advertising price comparison system that leverages the bidding data from four online advertising platforms that contributes to the transparency efforts to understand the economic value that online advertising system assigns to user profiles. Using this data we show that advertising price assigned to user profiles varies depending on the profile of the targeted user.

(2) **A methodology to leverage social media data for gender based digital inequality research.** Efforts to understand global prevalence of gender based digital inequality and its interplay with socioeconomic inequalities are limited by lack of large scale representative dataset. Towards solving this problem we developed inexpensive methodologies using the Facebook online advertising platform to quantify the extent of digital inequality in access to social media and its relation with existing inequality indicators.

(3) **A methodology to measure and characterize user representation and growth variability**
ity on Facebook. Facebook is the most widely used social media platform connecting billions of people globally. However, little is known about composition and growth dynamic of this advertising driven social media giant. The main challenge in measuring these phenomena is due to lack of large scale representative dataset that describe the actual number of users and their activities on the social media. We address this problem by leveraging its online advertising platform to measure and characterize its global composition and growth based on age, gender, and country of location.

In summary, the work presented in this thesis contributes to advance our knowledge of digital inequalities measured through social media and to develop methodologies for understanding important socioeconomic issues based on digital traces from social media and the Internet in general. It will motivate further research in the area to understand the extent and causes of digital inequalities in various aspects of the Internet considering different groups of societies. The reported findings and methodologies will also help lay foundation for informed policy development and research to close the gender gap and reduce digital inequalities worldwide.
# Table of Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acknowledgements</td>
<td>5</td>
</tr>
<tr>
<td>Research contributions</td>
<td>7</td>
</tr>
<tr>
<td>Abstract</td>
<td>11</td>
</tr>
<tr>
<td>Table of Contents</td>
<td>13</td>
</tr>
<tr>
<td>List of Tables</td>
<td>17</td>
</tr>
<tr>
<td>List of Figures</td>
<td>22</td>
</tr>
<tr>
<td>List of Acronyms</td>
<td>23</td>
</tr>
<tr>
<td>I Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1. Introduction</td>
<td>3</td>
</tr>
<tr>
<td>1.1. Thesis Overview</td>
<td>4</td>
</tr>
<tr>
<td>II Background</td>
<td>5</td>
</tr>
<tr>
<td>2. Online Advertising and the Web</td>
<td>7</td>
</tr>
<tr>
<td>2.1. The World Wide Web</td>
<td>7</td>
</tr>
<tr>
<td>2.2. Online Advertising</td>
<td>8</td>
</tr>
<tr>
<td>2.3. Tracking and Third Party Services</td>
<td>10</td>
</tr>
<tr>
<td>III Literature Review</td>
<td>17</td>
</tr>
<tr>
<td>3.1. Collection of Personal Data and Privacy Concerns</td>
<td>17</td>
</tr>
<tr>
<td>3.2. Leveraging Digital Traces for Social Good</td>
<td>18</td>
</tr>
<tr>
<td>3.3. Digital Trace Based Digital Inequality Studies</td>
<td>19</td>
</tr>
</tbody>
</table>
### III Measuring and Characterizing Digital Inequality

4. Measuring Value of Online Users
   4.1. Introduction ............................................. 25
   4.2. Related work .......................................... 26
   4.3. PriCom System Design ................................. 27
       4.3.1. priCom Design Constraints ....................... 27
       4.3.2. priCom Implementation ............................. 28
       4.3.3. Usability Assessment ............................... 30
   4.4. Analysis of prices of user profiles ................. 30
       4.4.1. Overall comparison of advertising platforms .. 31
       4.4.2. Influence of geographic location in profiles’ prices . 31
       4.4.3. Gender based price variability .................... 33
       4.4.4. Price variability across users’ interests .......... 35
       4.4.5. Temporal Variability of Prices .................. 36
   4.5. Discussion ................................................ 38
   4.6. Conclusion ............................................... 39

5. Digital Gender Inequality in Access to Online Services ......... 47
   5.1. Introduction .......................................... 47
   5.2. Related Work ........................................... 48
   5.3. Datasets ................................................ 49
       5.3.1. The Facebook Global Dataset ..................... 49
       5.3.2. Gender Equality and Development Datasets ....... 49
   5.4. Analysis methodology ................................... 50
       5.4.1. Computing the Facebook Gender Divide .......... 50
       5.4.2. Validating the FGD ................................. 51
       5.4.3. Regression Models .................................. 53
   5.5. Results .................................................. 54
       5.5.1. Gender Inequality ................................. 54
       5.5.2. Explaining Gender Inequality ....................... 54
       5.5.3. Network Effect ..................................... 56
       5.5.4. Relation of FGD with Economic Gender Inequality Changes 56
   5.6. Discussion ................................................. 57
   5.7. Conclusion ............................................... 59

6. User Representation and Growth on the Largest Social Media ... 61
   6.1. Introduction .......................................... 61
   6.2. Related Work ........................................... 62
   6.3. Key Insights .............................................. 63
<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.4. Data Collection to Measure Facebook User-base and User Engagement Evolution</td>
<td>65</td>
</tr>
<tr>
<td>6.5. Recommendations on using the API for research purposes</td>
<td>67</td>
</tr>
<tr>
<td>6.6. Analysis Methodology</td>
<td>69</td>
</tr>
<tr>
<td>6.7. Results</td>
<td>70</td>
</tr>
<tr>
<td>6.7.1. Facebook’s Growth Capacity</td>
<td>70</td>
</tr>
<tr>
<td>6.7.2. Facebook Growth Evolution</td>
<td>72</td>
</tr>
<tr>
<td>6.7.3. External Datasets to Characterize the Observed Growth Variability:</td>
<td>77</td>
</tr>
<tr>
<td>6.8. Conclusion</td>
<td>80</td>
</tr>
<tr>
<td>IV Conclusion and Future work</td>
<td>81</td>
</tr>
<tr>
<td>7. Conclusion and Future Work</td>
<td>83</td>
</tr>
<tr>
<td>V Appendices</td>
<td>85</td>
</tr>
<tr>
<td>Appendices</td>
<td>87</td>
</tr>
<tr>
<td>References</td>
<td>131</td>
</tr>
</tbody>
</table>
List of Tables

4.1. Dataset Description .................................................. 30
4.2. Most expensive countries on All platforms. The displayed prices are in euro cents. 33
4.3. Least expensive countries on All platforms: the displayed prices are in euro cents. 33
4.4. Pearson and spearman correlation between advertising price and GDP of the country 33
4.5. Geographic price correlation (pearson/spearman) between ad-platforms .......... 34
4.6. Interests showing highest advertising prices (in euro cents) on each platform: Interests are listed in increasing order. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube. .......................... 36
4.7. Interests showing cheapest advertising prices (in euro cents) on each platform: Interests are listed in increasing order. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube. .......................... 37
4.8. Interest categories ordered by their decreasing price in each platform. Median prices in euro cents are listed in brackets. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube. .......................... 37

6.1. Counts of median Facebook total users (TOT) and daily active user (DAU) estimates, in July 2015 and June 2018 showing results for both genders, and results for US and India (countries having the largest user counts). ..................... 63
6.2. Regression coefficients (Signif. codes: 0 ‘***’, 0.005 ‘**’, 0.05 ‘*’) ............ 78
6.3. Variance Inflation Factors (VIF) of independent variables in the user base and user engagement growth models. For both models VIF of all factors is below 5. . 79
List of Figures

2.1. Figure showing multiple players in Digital advertising ecosystem (source LUMA partners) ......................................................... 9
2.2. Analysis of third parties associated with elpais.com (a) list of intermediaries behind an ad on the webpage. (b) list of third parties present on the webpage. (c) preemptively detected list of third parties on a webpage linked from the current webpage ......................................................... 10
2.3. Analysis of top 50 Alexa websites (a) Number of first party and third party cookies stored by a website. (b) Number of third party domains detected by block-lists. (c) Total number of redirects per website ...................................................................... 15
4.1. Architectural diagram of the price comparison system .............................................................................................................. 28
4.2. Distribution of beta testers: the left panel shows the demographic distribution of users, the middle panel shows average time spent on internet showing that majority of users spend less than 5 hours per day. The right panel shows self reported Internet user level ........................................................................ 30
4.3. General advertising price variability across four platforms .................................................................................................. 31
4.4. Geographic advertising price variability on the four major advertising platforms: Facebook (CPC), Google (CPC), YouTube (CPV) and Instagram (CPC). Countries are colored from blue (indicating the cheapest countries) to green (indicating the most expensive countries). Countries with no data are colored in red ................................................................. 30
4.5. General advertising price variability across US states: the left figure shows the price variability in CPC billing scheme, while the figure on the right shows advertising price variability in CPM. States are colored from purple (indicating the cheapest state) to red (indicating the most expensive state), the inset figures show the prices in Alaska and Hawaii respectively ........................................ 41
4.6. Gender based advertising price variability on Facebook and Instagram .......................................................................................... 42
4.7. Interest based advertising price variability across four advertising platforms: the values on the y-axis show the median prices per category. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube ........................................................................ 43
4.8. Temporal advertising price variability across four advertising platforms: the values on the y-axis show the normalized median prices per month ........................................................................................................ 44
4.9. Temporal Price variability across user profiles: the box plots show the difference between the 95 percentile price and the 5 percentile price of a profile. 

5.1. Diagram representing architecture of the crawling tool used for the measurement: The measurement tool is composed of a master program that first contacts the Facebook authentication server (step 1), after getting authenticated on step 2, it then divides the data collection task between individual crawler programs (shown on step 3), on step 4 each assigned crawler uses the authentication token to establish a connection with the marketing end point from Facebook graph API. After establishing connection, each crawler starts to query the graph API for information by iterating over the list of queries assigned to it by the master crawler. For each valid request the end point returns total number of registered users and number of daily active users together with other information including the cost to reach that desired demographic. On step 5, upon receiving response each crawler writes the response with the queried parameters on a log file.

5.2. The Facebook Gender Divide across 217 countries. This figure shows worldwide FGD values. The FGD values range from red which represents highly skewed values towards males, to green (not visible) which represents countries with skewed values of FGD values towards females. Countries with a balanced FGD are colored blue. The left inset shows the scatter plot of male and female activity ratios across all countries, revealing a spread along the diagonal. The right inset shows the histogram of FGD values in bins of width 0.2. While the mode of countries is slightly below zero, there is significant skewness towards high FGD values (bias towards males).

5.3. Regression results of FGD as a function of gender equality. A) Model predictions versus rank of FGD, where rank 1 is the country with the highest FGD. The model achieves a high $R^2$ above 0.74, explaining the majority of the variance of the FGD ranking. Some countries are labeled, from high FGD (Liberia, India, and Saudi Arabia) to low FGD (Finland, Norway, and Uruguay), as well as some outliers (Dominican Republic, Austria, and Sri Lanka). B) Coefficient estimates and 95% CI of the terms of the regression fit (excluding intercept). Education (Edu), health (Heal), and economic gender equality (Eco) are significantly and negatively associated with the FGD, but political gender equality (Pol) is not. From the control variables, Internet penetration (IP) is negatively associated with FGD, but the rest are not. The main role of education equality in FGD can be observed on panel A, where dots are colored according to the rank of education gender equality, showing that countries with low FGD are ranked high on education gender equality.
5.4. Gender differences in network externalities on Facebook. Scaling of Facebook activity ratio per gender versus total Facebook penetration. Solid lines show fit results and shaded areas show their 95% confidence intervals. Both male and female activity ratios grow superlinearly with Facebook penetration ($\alpha > 1$), indicating positive network externalities. These network externalities are stronger for female than for male users ($\alpha_F > \alpha_M$).

5.5. Analysis of changes in economic gender equality and FGD. Coefficient estimates of the regression model of changes in economic gender equality as a function of FGD and control terms (excluding intercept, top left), and of the model of changes in FGD as a function of economic gender equality and control terms (excluding intercept, bottom left). Right panels show the bootstrap distributions of partial $R^2$ of $FGD_{2015}$ in the first model and of $Eco_{2015}$ in the second one, with dashed vertical lines showing the median $R^2$ values: 0.027 in the first model and 0.002 in the second one. The FGD explains changes in economic gender equality much better than economic gender equality explains changes in the FGD.

6.1. Distribution of days per month used in the analysis.

6.2. Global map of Facebook Potential growth: countries are colored based on their available room for Facebook growth ranging from dark blue (highest penetration hence less opportunity to grow) to light green (highest potential). Countries with no data are colored red.

6.3. Global map of Facebook engagement capacity: countries are colored based on their observed engagement ranging from dark blue (least engagement among total users implying least engagement capacity) to light green (highest observed engagement). Countries with no data are colored red.

6.4. Difference between “home” and “recent” location types and comparison with Facebook quarterly report. The left panel shows comparison between daily active users, the center panel shows comparison between total users. The right panel shows comparison of the two targeting results with Facebook global users as reported in June 2018.

6.5. Correlation between Facebook number of users at country level using country name as a targeting option vs partitioned demographic targeting with (country,age,sex). The left panel shows the correlation between daily active users, the right panel shows correlation between total users.

6.6. Comparison of the dataset collected though the marketing API with Facebook quarterly report. The left panel shows comparison between daily active users, the right panel shows comparison between total users. In both figures the median value of each month is compared with the Facebook quarterly report.
6.7. Global population pyramid of Facebook (median value of June 2018): The outer pyramid (darker color) represents demographic distribution of total users, inner pyramid (lighter color) shows daily active users. ............................ 71
6.8. Monthly evolution trend of Facebook user-base and user engagement based on age groups. The left figure shows the evolution trend of user-base measured via total users, the right figure shows the trend in user engagement. ................. 72
6.9. The default setting of Facebook registration page showing 1993 as a default year of birth. .......................................................... ...................... 73
6.10. Growth rate of Facebook’s user-base (total users) and user engagement (daily active users) for different age groups: Adolescence (13-19), Early Adulthood (20-39), Adulthood (40-64) and Maturity (65+). .............................. 73
6.11. Growth rate of Facebook’s user-base (growth in total users) and user engagement (growth in daily active users) for male and female users on Facebook. ......... 74
6.12. Growth rate of number of total users (top) and number of daily active users (down). The range moves from lowest growth rate (yellow; not visible) to highest growth rate (light green). .................. .................. 75
6.13. Comparison of geographic, gender, and age group dimensions in determining user base and user engagement growth variability calculated using random forest regression method. For user base growth geographic dimension is highly influential covering 90.3% of variability followed by age group and gender with values 7.2%, and 2.5%, respectively. Geographic dimension is also highly influential in user engagement growth variability with 71.6% followed by age group, and gender based variability with values 71.6%, 22%, and 6.4%, respectively. ............. 76
6.14. Heat map showing the correlation between each pair of external socioeconomic indicators, the FB user-base (1st column) and FB user engagement (2nd column). The correlation ranges from blue (negatively correlated) to red (positive correlation). 79
List of Acronyms

CDNs  Content Distribution Networks
DAU  Daily Active User
GDPR  General Data Protection Regulation
PII  personally identifiable information
FGD  Facebook Gender Divide
WWW  World Wide Web
HTTP  HyperText Transfer Protocol
URL  Uniform Resource Locator
RTB  Real Time Bidding
DNS  Domain Name System
NAT  Network Adress Translation
CPC  Cost per click
CPM  Cost Per Mille
CPV  Cost Per View
ICT  Information Communication Technologies
GWI  Global Web Index
SNS  social networking services
VIF  Variance Inflation Factor
GUI  Graphical User Interface
Part I

Introduction
Introduction

Following the rapid advancement of computing technologies and their increasing adoption by the society, a significant portion of societal activity is taking place online. This has led to the emergence of globally prevalent Internet services, such as social media. These services offer open and publicly available platforms where people can communicate, share their feelings and experiences, exchange relevant information with other people or organizations worldwide. Despite the challenges such as inequality, bias, privacy, and transparency, the digital traces left by these users has opened new research directions and business opportunities enabling creation of personalized algorithms for product recommendation [1], content placement [2], customer profiling [3], and targeted advertising [4]. Motivated by these opportunities a large number of traditional businesses have also migrated to online form with online advertising being a typical example.

On the other hand existence of unequal access to these technologies might hinder parts of societies from realization of the full potential offered by these technologies for social and economic benefits, which is often limited by the number of users having access to these technologies. Existence of digital inequality is a huge problem since it has the potential to exacerbate inequalities in the offline world such as social, political and economic aspects of society. One of the key research challenges in understanding digital inequalities such as digital divide is the lack of large scale and publicly available datasets that can enable comparative study between different groups such as demographics, gender, and countries.

In this thesis we aim to contribute to this collective effort by providing methodologies for collecting and analyzing large scale datasets form online advertising platforms to measure and characterize digital inequalities.

— everything is related to everything else, near things are more related than distant things.

Waldo Tobler
1.1. Thesis Overview

The remainder of the thesis is organized as follows: Chapter 2 provides a necessary background about online advertising and the world wide web. It provides an explanation of the operation of the online advertising ecosystem with a focus on the techniques used to track users and categorize them. Important aspects such as targeted advertising, real time bidding, tracking and third party services are introduced. Note that state of the art will be reviewed across the different chapters that will focus on the existing work on the specific topic of the chapter. After laying out this background, in chapter 3 we will cover some of related literature regarding developments and challenges in using digital traces specifically traces recorded by online advertising systems showing the progress of this upcoming interdisciplinary field. Analysis of the value of users and their generated content is presented in Chapter 4. Results of leveraging online advertising platform for digital gender inequality study in the access to online services specifically on Facebook are presented in chapter 5. The composition and growth in access to social media, using Facebook as example which is the largest and most important online service is presented in Chapter 6. Finally conclusions and some of potential future research lines are presented in chapter 7.
Part II

Background
Online Advertising and the Web

This chapter highlights the necessary technical background on the World Wide Web (WWW), online advertising, and specialized third party services that we often find embedded in websites. First we start with world wide web, which is the most popular application layer system that enables accessing and browsing a hyper linked connection of services distributed across computers on the Internet. Later, we highlight the working principle of the major financial backbone of the web, which is online advertising. Finally we discuss the third party services that constitute the modern dynamic web enabling tracking and collection of personal data, among other things.

2.1. The World Wide Web

The World Wide Web has progressively changed both functionally and structurally since its initial implementation as a large collection of hypertext documents. The web of today is a dynamic integration of multiple services which are triggered upon a user visit to a certain web application using a particular web client (e.g. browsers like Mozilla Firefox, Google Chrome, Internet Explorer, and Safari) via the stateless application layer HyperText Transfer Protocol (HTTP) or the encrypted version of it (HTTPS). Unlike their predecessors today’s web-pages are dynamically generated composed of dynamic scripts, text, and multimedia documents serving various popular applications on the Internet such as finance, health, entertainment, and social media. Each user interaction with a service on the Internet generates a series of procedures that involve cooperation of different devices via a set of sophisticated protocols typically following the TCP/IP protocol stack [5]. Here we briefly describe a high-level view of this interaction. When a user re-
quests a specific service on the WWW using the service’s Uniform Resource Locator (URL), the browser sends the request to the identified address of the web server hosting that particular web service. The web server hosting the requested service executes the application and responds with a document that includes executable scripts and addresses of third party services that are included by the publisher of that web-service. These third party applications might be related to online advertising, Content Distribution Networks (CDNs), tracking, and multimedia services that will be further integrated with the application on the user device. This process involves communication between different services at different layers of the TCP/IP such as Domain Name System (DNS) look-ups, Network Address Translation (NAT), routing, and switching [5]. It is important to note that each communication between the user device and the remote services mentioned above involves significant data communication. Analyzing end user’s mobile traffic Papadopoulos et al. have shown that the cost paid by end users to download data packets related to an ad is higher than the cost paid by advertisers to deliver them.

2.2. Online Advertising

As the number of Internet users grows and the web technologies evolve to accommodate various societal and business activities to the digital platform, advertisers have also turned into the online industry to reach the large and diverse potential users. As a result the online advertising system is able to become the backbone of the open Internet where majority of websites are accessible openly. Online advertising is now a multibillion-dollar industry. In 2017 alone, online advertising generated $88B worth of revenue only in US, representing an increase of 21.4% with respect to 2016 [6]. It is this impressive size and growth that permits online advertising to emerge as the main source of revenue for most of Internet services such as maps, search, social media, and user generated content platforms which are at the forefront of innovation in the Internet and have generated millions of direct and indirect jobs in Europe alone [6].

Since the introduction of static banner ads in early 1990s the technology of online advertising has evolved to include many players that helps to generate dynamic ads based on a mechanism referred to as programmatic advertising [7]. These technologies include web publishers that sell slots (also called inventories) in their web pages to advertisers that want to reach potential customers; and intermediaries that facilitate this process. Intermediaries include ad-networks and ad-exchanges that also get information from data brokers. They serve as a marketplace for automatic trading of ad inventory. Demand Side Platforms (DSPs) and Supply Side Platforms (SSPs) are other intermediaries lately incorporated into the ad-serving ecosystem, they operate as a unified interface for advertisers and publishers to interact with multiple ad-exchanges respectively. Advertising system also leverages third party data brokers to get additional profiling data about users which are not available via tracking. To identify the interests of each user, the advertising system leverages a sophisticated tracking mechanism that monitors users’ online activities (visited websites, search queries, social media activity, etc.) and gradually assigns each user to an audience
bucket (a set of interests that best describe the user) \[8\]. As opposed to traditional advertising, the current online advertising system has created an opportunity for advertisers to selectively target specific users based on their demographic and interest profiles. While this opportunity has attracted advertisers to move to online advertising, it has also increased the complexity of the advertising ecosystem. Figure 2.1 shows a bird’s eye view of the complexity of the ecosystem. Advances in advertising techniques has resulted in the following three major types of targeted advertising.

A. Contextual advertising: in this form of advertising advertisers select inventories based on the contents of the web page oblivious of the type of users visiting it.

B. Re-targeting: this form of advertising is based on the shopping history of users across websites, using this technique an advertiser follows potential customers who showed interest in a particular product but left without conversion (e.g. buying a product, filling a membership form etc.).

C. The other most advanced and effective targeting mechanism is behavioral advertising that utilizes personal information about users based on advanced user tracking techniques. In this scenario when a user consumes a given internet service the tracking intermediaries integrated in the service expose the collected information about her to advertisers which in turn decide the value of that particular user and offer a monetary value through Real Time Bidding (RTB) - if they find that user potential user of their product or service.

\[RTB\] is an automatic auction mechanism on which a variety of advertisers can decide the monetary value of advertising to a user visiting a certain website, if that user-website combination meets the target audience of their advertising campaigns. Even though user profile information such as

\[www.lumapartners.com/resource-center/lumascapes-2\]
their behavioral data is not a necessary requirement in RTB, this additional information improves the targeting capacity for advertisers letting them to fine-grain their campaigns.

There exist different pricing schemes used in the online advertising industry such as: Cost per click (CPC), Cost Per Mille (CPM), Cost Per View (CPV), which are described briefly as follows:

- **Cost-Per-Mille (CPM):** This is a price tag assigned to serve 1000 impressions of the advertisement associated to an ad campaign. It usually ranges between few cents of euros to few euros.

- **Cost-Per-Click (CPC):** This is the price associated when a user clicks on the advertisement and she is redirected to the advertiser’s website or a landing page defined by the advertiser such as app store. ranges between few cents of euros to euros.

- **Cost-Per-View (CPV):** This pricing scheme is used in video advertising. The advertisers are charged when their video ad is viewed by the user. Typical CPVs ranges between 0.01 and 5 euros.

Note that in CPM pricing no involvement from end users is required, while in the rest of pricing schemes user involvement is assumed. In CPA and CPC an action is required from end users while in CPV pricing users are required to be delayed from their desired action for a certain number of seconds (for instance it is very typical to have a delay of two to five seconds in pre-roll ads shown before the video plays).

![Figure 2.2: Analysis of third parties associated with elpais.com](image)

Figure 2.2: Analysis of third parties associated with elpais.com (a) list of intermediaries behind an ad on the webpage. (b) list of third parties present on the webpage. (c) preemptively detected list of third parties on a webpage linked from the current webpage.

### 2.3. Tracking and Third Party Services

The technology of online advertising relies on a sophisticated dynamic system of third parties with the objective of identifying information about users such as their demographics and behav-
ioral interests to match relevant advertisements to a target user. These tracking mechanisms are able to infer locations frequented by the user, types of devices used, their age and gender groups, the types of websites they visit, and their behavioral and personality types [8,9]. To achieve this objective, third party trackers leverage distributed tracking techniques to gather information about the online activities of each user and generate user profiles based on the collected personal data. As part of our research we have developed tools that help users get information about third parties (both preemptively before the user visits a webpage, and also on the webpage the user is currently visiting). These tools detect and classify third party companies behind websites. For example, screenshots from these tools presented on figure 2.2 show the list of intermediaries behind an ad displayed on elpais.com, third party domains detected on this website, and preemptively detected list of third party domains behind a webpage of moneygram.com linked from this website and classified into tracker, ad-serving, and other categories. As it can be seen from the figures both elpais.com (a major news portal in Spanish world) and moneygram.com (one of the major money transfer institutions) leverage third party services. We can also see that many intermediaries are involved behind an advertisement that is shown on the figure (first panel). Note that due to rising privacy concerns detecting and understanding how third party websites are integrated within websites is becoming crucial. In fact, the new General Data Protection Regulation (GDPR) law requires website owners to get consent from users regarding their privacy and transparency in use of third parties [10]. Recently these laws and regulations are being enforced in different countries resulting in heavy financial loss to non-compliant website owners [11,13]. Given the above circumstances identifying which third parties are included, the behavior of cookies and transfer of data between the first parties (also called data controllers) and third parties (data processors) is crucial for any organization offering web services. As a result, web compliance testing should be integrated as a principal dimension to the existing automated test workflows for performance, usability, and security tests performed by website owners.

To help website owners create applications that comply with these privacy regulations we have developed an advanced version of the above tools called pyObserver, which enables collection of three important data logs gathered during browser’s interaction with a website. The tool collects logs listed below which are particularly useful to assess the compliance requirements emphasized in recent laws and regulations such as GDPR and the cookie law:

- Integrated third party domains in the given website,
- Time-stamped URL redirect traffic with parameters between redirects including intermediaries,
- Cookies stored on the client by the first party and all third party domains included by the first party domain.

Besides, the developed tool also classifies the list of URLs that will be identified for blocking as tracking/advertising by major ad-blocking software. It is important to note that although many
large organizations and enterprises may be able to handle the tight and growing laws and regulations such as GDPR, many SMEs do not have the financial, legal, or technical resources to assess their compliance and take necessary measures in short time. While we can find several related tools and platforms developed for studying privacy issues in websites [14–18], these tools are not suitable for advertisers mainly because they are developed for privacy researchers from the end user perspective and have limitations making them unsuitable to be integrated as part of the existing testing workflows performed by website publishers. Specifically, we developed pyObserver with the following goals to solve previous limitations:

- Should not significantly alter existing test workflows,
- Should not introduce unnecessary traffic and workload,
- Should be able to integrate with existing test automation using real browsers
- Should list all third party domains included by the website under study,
- Should collect first party cookies and all third party cookies,
- Should record http(s) redirects together with data passed as URL parameters,
- Should return results in a widely used data structure format that can easily be analyzed or exported.

To achieve the above objectives pyObserver was developed on top of the most popular test automation tool called selenium and the chromium browser. Selenium is a tool used for automated testing of web applications by automating browser interactions with websites [19]. Chromium on the other hand is the most popular browser engine by market share [20,21]. Note that the default selenium automation framework does not provide all of the functionalities mentioned above, particularly a functionality that logs the third party cookies, third party web domains, and also the HTTP(S) redirect traffics. One of the most popular way of recording redirect traffics used by previous research is using an external proxy server such as browsermob proxy to log the traffic and store it in a HAR file format; however, this introduces an extra overhead [22]. To provide these functionalities that are not available in the selenium framework, we have developed an integrated browser extension that records these values and returns the results to the main selenium test workflow as JSON object. pyObserver is implemented in python programming language and inherits all the functionalities of chrome based selenium webdriver (webdriver.Chrome). As a result a web tester can easily use pyObserver following the usual procedures. The inspection results can be obtained by calling the inspect method which returns a JSON object containing the following keys: GIVEN_URL which indicates the URL of the website inspected, cookies which returns all the cookies stored on the client by first party website and the third party domains, redirects which includes the http(s) redirects between URLs contacted by the browser together with URL parameters during the inspection session, thirdparties which is the list of third parties included by the inspected website as recorded by pyObserver, and inspection_duration which records the time taken to complete the inspection of a website. For each cookie object stored, details of the cookie including the domain that stored the cookie, cookie name, and value are recorded. Details of each redirection record in redirects list is also included such as the following: frame
### 2.3 Tracking and Third Party Services

id, initiator, redirect URL with the URL parameters between redirects, and redirection IP. These values will be useful in later analysis of the log results. The `classify_third_parties` module available in pyObserver provides a functionality to check which third parties will be detected by major adblockers as advertising related or tracking related using the easylist and easyprivacy filter lists \[23\]. These lists are a sets of rules widely used by popular adblockers such as Adblock Plus and ublock origin \[24\]. This module classifies third parties into four categories as TRACKER, AD, BOTH, OTHER based on whether the given URL is detected as tracker related, advertising related, both, or if it is not detected by both lists. The developed tool is part of our working paper about improving compliance and transparency of websites, and the detail installation and usage instruction is available at [http://dataportal.netcom.it.uc3m.es/pyobsver/](http://dataportal.netcom.it.uc3m.es/pyobsver/). In what follows we show the usability of pyObserver by inspecting Alexa top 50 websites.

**pyObserver use-case**

To demonstrate the usability of the tool we collected Alexa’s top 50 websites as of August 3, 2019. For each website in the list we performed automated loading of the website using pyObserver and closing the session once the loading is finished without further interaction with the website from the browser side.

**Cookies** After extracting the cookies stored by each website from the JSON object returned by pyObserver, we first filtered out session cookies. Session cookies, also known as transient cookies, are cookies that are temporarily stored on the browser. We then grouped the remaining cookies (i.e. permanent cookies) into first party cookie and third party cookie based on the domain name of the party that stored the cookie. Figure 2.3 (a) shows the number of cookies stored by each website. As shown in the figure majority of the websites in this list store permanent cookies when the website loads up on user’s first visit to their website without active involvement from the user. The results show that these websites store a median of 7 cookies with 5 first party cookies. Using the results from this tool website owners can further inspect the contents of the cookies to check their contents and structure to address issues such as compliance.

**Third party URLs** After getting the list of third party requests from the returned JSON object, we applied `classify_third_parties` method to classify the third parties recorded. Figure 2.3 (b) shows the number of third party URLs detected as advertising, tracking, or both on each website. Using this tool website owners can inspect the detected third party URLs and take measures to ensure their compliance.

**Redirects** To summarize the results of redirects per website, we first counted the number of redirects per each website, then we inspected if these redirects also contain URL parameters. URL parameters are data that are passed between web entities for purposes such as user tracking notification, advertising bidding in both clear test and encrypted form \[25\]. Figure 2.3 (c) shows the number of redirects per website, with average of 23 URL parameters while the median number

---

\[2\] Alexa top sites: www.alex.com/topsites
of parameters is zero, showing that some of these websites utilize URL parameters as a means to transfer data. These preliminary results indicate the usefulness of the tool enabling website owners to further investigate the content of URL parameters, the third parties that took part when their website loads, and the nature of redirects and cookies that took place during first interaction with the website.
Figure 2.3: Analysis of top 50 Alexa websites (a) Number of first party and third party cookies stored by a website. (b) Number of third party domains detected by block-lists. (c) Total number of redirects per website.
Thanks to the growing ubiquity of Internet and relative availability of affordable devices, every day a large amount of digital data is produced from users during their search for information, during their interaction with websites, or during their interaction with other users. This large scale availability of data about users has brought its own opportunities and challenges, leading to the parallel explosive boom of competing Internet services specializing in collection and processing of this big data, often refereed to as the new oil \[26\], and rising privacy and transparency issues that are a topic of discussion both in academic and non academic settings. In this chapter we first discuss some of previous research on the opportunities and challenges of collecting and leveraging digital traces of users, later we highlight research covering the problem of digital inequality.

3.1. Collection of Personal Data and Privacy Concerns

The collection and processing of user data from Internet users has been proven effective for various aspects, such as improving user experience by offering users personalized products and services \[27\], maximizing revenue by leveraging user data to find potential customers \[28\], and for various societal and technological purposes \[29, 30\]. However, this aggressively collecting personal data is often exposing user privacy which has resulted in undesired effects, including exposure of sensitive user interests and psychological targeting for digital mass persuasion. The public has responded to these aggressive tracking and advertising practices by using different tools and services to block and circumvent ads and trackers from the advertising based Internet \[31\]. Similarly, the research community has been actively investigating these issues, such as
discrimination based on user personally identifiable information (PII) \[32,34\], algorithmic bias and filter bubbles \[35,37\], privacy and exposure of sensitive topics \[38,40\], extent of tracking practices \[8,9\], alternative solutions to bring transparency and protecting user privacy on the web \[41,42\].

The online tracking ecosystem uses a range of multiple sophisticated tracking and fingerprinting techniques integrated with majority of websites \[9\], using these techniques online advertisers are able to assign users a set of demographic and personality categories \[8\] which will enable them to reach specific target audience. Based on experimental study, Mikians et al. \[32\] empirically demonstrated the existence of signs of both price and search discrimination on the Internet based on user PII such as location and personal information. Ali et al. \[34\] have also shown that a combination of targeting parameters based on large collected data in advertising platforms could lead to potentially discriminatory ad delivery. Even though under various legislations such as GDPR it is prohibited to target users based on sensitive information such as their sexual orientation, Cabañas et al. \[38\] were able to show that it is possible to reveal identity of Facebook users with potentially sensitive interest at a cost of a fraction of cents per user. On the other hand Matz et al. \[39\] have demonstrated a methodology for political mass persuasion based on analyzing psychological behavior of users on social networking sites. Note that a similar methodology was used by Cambridge Analytica during the US Election \[37\]. To remediate privacy concerns arising from extensive tracking mechanisms various researchers have proposed different solutions including proposing privacy protection and discrimination prevention algorithms \[43,44\], creating a common identity to a group of users to reduce exposure of individual browsing patterns \[41\], and letting users decide the types of tracking and ad categories they want \[42\]. Concerned with the challenges associated with big data driven algorithms, such as algorithmic bias, researchers have also called for a research discipline to study the confounding behavior of algorithms and their societal impact \[35\]. Note that this is not an exhaustive list of literature on personal data and privacy concerns, for a more exhaustive list of related work in the area we encourage readers to refer to review papers such as \[45,46\].

3.2. Leveraging Digital Traces for Social Good

Despite the above limitations, large scale digital traces collected for various tasks have been innovatively re-purposed to study different aspects of societal challenges ranging from understanding societal knowledge gaps and needs \[47,50\], to understanding discrimination \[51,53\], to disease prediction and prevention \[54,57\]. Even though there is a promising progress in leveraging these digital traces for social good, research in this direction does not benefit fully from the big-data revolution. Coulton et al. \[58\] pointed that the main reason behind slow progress in the "social sector" is that a large portion of useful information remains trapped in the piles of legacy information systems or in possession of large commercial enterprises which makes it inaccessible for researchers interested in applying these resources for social good, such as developing
3.3 Digital Trace Based Digital Inequality Studies

Informed policies. They further noted that due to such inaccessibility, social work professionals are limited to rely on survey and other primary data collection methods. Below we highlight some of research progress in this effort.

A recent example of using social data for social good is the Facebook disaster map which utilizes Facebook user data in areas affected by hazards [59]. The system leverages aggregate user data such as historical usage information, Facebook population, power availability (based on number of users charging their phones), network coverage and user displacement to assist post disaster recovery and response in areas impacted by natural hazards at Facebook scale. Nazer et al. [60] have provided a survey of related research approaches in using social media data for disaster relief and response, they have identified four phases of disaster and the research progress at each phase. A recent study based on anonymized search query logs has shown that health related search logs can offer insight about health information gaps in developing regions enabling development of informed health policy and targeted education campaigns [47]. In a related research Araujo et al. [50] leveraged the Facebook advertising platform to track non-communicable diseases globally offering insight into health awareness across demographic groups. Digital traces from short tweet conversations have also been used to detect and monitor air pollution [56]. Santillana et al. [61] have presented a machine learning based surveillance methodology that combines data from search queries, social media, and traditional sources such as aggregated hospital visit records for influenza monitoring and prediction obtaining better results than the baseline methods constructed independently with each data source. Satoshi et al. [62] used a monthly user visit data to a food recipe-sharing website to create a recommender framework for a seasonal food recipe diversification scheme. Bookmarking activities of users in online food recipe portals has also been leveraged to develop a methodology to monitor obesity prevalence [63]. Social media data has also been instrumented to generate more aesthetically pleasing walking routes in a city [64]. The potential of data collected by advertising platforms has also been demonstrated to be a promising tool for global development statistics enabling a wider reach [65].

3.3. Digital Trace Based Digital Inequality Studies

As social Internet applications and the Internet in general is becoming influential in everyday life, the very existence of inequality in access to this innovation and its benefits is a limiting factor for local development and the global development goals [65]. As a result digital inequality studies have been one of pressing issues among researchers and policy makers. However, the expensive and difficult procedure to perform traditional methodologies on a large scale has hindered advancement of research in this area. In their attempt to perform interview based digital inequality study in use of social media by marginalized societies in Brazil, Nemer [66] mentioned that performing ethnography in such areas was very challenging and risky.

Even though very little progress, recently digital inequality studies are also being benefited from large availability of digital traces, especially social media data. Fatehkia et al. [67] have
shown that gender gaps in Internet and mobile phone access can be indirectly measured using Facebook advertising platform enabling tracking the gap in inaccessible regions in low-income countries. Similarly Mejova et al. have demonstrated the feasibility of this approach to monitor digital gender inequality among states in India [68]. In a loosely related study about investigating patterns of Internet usage by Goel et al. [69], analysis of web activity log of 250K individuals who participated in a panel found that the frequency with which individuals turn to the web for research, news, and health-care is strongly related to educational background. Indaco et al. [70] created a methodology to study social media inequality using meta-data in Instagram images. Applying their methodology in Manhattan district the authors found that inequality in Instagram images is larger than socioeconomic inequalities they considered (levels of income, rent, and unemployment rate). A geolocated network measurement study on the allocation of Internet coverage across ethnic groups was used to quantify inequality in Internet coverage across ethnic groups worldwide [71]. The results show that politically excluded groups also have significantly lower Internet penetration rates, concluding that governments play a key role in its allocation. Johnson et al. [72] compared Internet usage between developed countries and rural Africa based on digital traces collected from a wireless network in rural Zambia. Their results showed a significant difference with traffic usage from developed countries. Based on their studies online social networks were the most dominant applications, while a large portion of the network traffic was occupied by system updates. The authors also noted a strong feedback loop between user behavior and networks performance. Micheli et al. [73] proposed to extended the problem of digital inequality study to include digital footprint inequality which studies how digital footprints vary according to socioeconomic variables. In their work they argued that as artificial intelligence and machine learning algorithms heavily rely on large-scale personal data, digital footprint inequality will have huge impact in training such algorithms. In fact algorithmic fairness studies have shown the impact of training data induced biases in algorithms [74]. Analyzing the historical logs from an online labor market platform Galperin et al. have shown the existence of inequality in getting contracts based on the online worker’s country of origin, suggesting the reason for the higher attrition rate for workers from developing countries [75]. Geolocated hashtags from Twitter and Instagram were used to measure the relation between income inequality and female sexualization in social media [76]. The authors found that this association is stronger in developed nations. Schaab et al. [77] performed a gender based comparative content analysis of user comments on YouTube videos of males and females performing vocal covers of extreme metal music. Their results show some interesting findings: female grunters are extremely underrepresented with about 3% of their random sample, females were met with more extreme sentiment (such as surprise and stereotypical comments). They also found that females were more likely to receive technical advice on their performance. Analyzing their overall results authors suggest that the Internet might be a viable tool to circumvent offline gender inequality in individual music production.

We believe that a lot has to be done to narrow the wide gap of digital inequality. Leveraging digital traces to better understand pressing societal issues, such as digital inequalities to help come
up with far reaching solutions is not only an inexpensive solution but also a promising approach to perform a large scale representative study. We hope that the contributions in this thesis will help advance the research effort in this direction.
Part III

Measuring and Characterizing Digital Inequality
Measuring Value of Online Users

4.1. Introduction

In this chapter, we focus on methods for analysis of the economic value of end-users in the online advertising ecosystem. Since behavioral targeting based online advertising is the source of revenue for majority of web-based services, understanding the difference in value of users assigned by online advertising is a crucial step to get the full picture of the Internet dynamics. The behavioral advertising model is in theory beneficial for end-users, since they are exposed to ads related to their interest instead of random ads. However, in practice, the online advertising ecosystem operates as a black box. This lack of transparency raises a series of privacy concerns which open several questions: how the tracking is conducted? [8, 80], which data is collected and with whom it is shared? [81], what is the actual value of the end-users’ data? [25], etc. Each of these questions is hard to answer and thus they require specific analyses to address them.

As a contribution to the development of this area we present advertising price comparison system called priCom, a system that collects the daily value of thousands of user-profiles across some of the most popular advertising platforms, namely Facebook, Instagram, Google and YouTube and comparatively informs users about the prices of these profiles. With this publicly available online service users can configure targeting parameters using demographic and geographic keywords with the desired bidding mechanism for which the tool returns the price variability between the configured profiles. The tool is designed to benefit various stakeholders,

http://dataportal.netcom.it.uc3m.es/betaC2/
including: end users, regulators, and advertisers. It helps to educate end users so they are aware of what their data and privacy is worth. Regulators will also benefit from this extensible tool since it enables them to oversee existence of advertising price variability based on gender, age, or geographical location and identify potential price discrimination practices [81][82]. The tool is also important for advertisers themselves. It can serve advertisers as a decision support system to compare between profile tags and advertising platforms and target the cheapest/optimal ones satisfying their requirements. Based on the dataset collected using this tool we present initial analysis on price variability on user profiles considering geographic and demographic analysis dimensions. The rest of the chapter is organized as follows. In section 4.2 we review related literature in this specific area. Section 4.3 presents the architecture and measurement methodology of PriCom. Section 4.4 shows the cross-platform analysis of price variability across geographic, demographic, and behavioral attributes of users using the dataset collected from PriCom. Section 4.5 discusses the results. Finally, Section 4.6 concludes the chapter.

4.2. Related work

The research community has recently started investigating the monetary value of personal information collected from users as they browse the web [83–86]. For example on [84] authors developed a browser plug-in that provides individual Facebook users with a realistic estimation of the revenue they generate for Facebook, the survey results in the paper also reveal a general lack of awareness among Internet users regarding the monetary value of their personal information. Even though the theme of the tool developed in this paper is revealing the value of a user’s personal information, it is designed for individual users and differs from our work in scope and diversity of covered advertising markets. In [86] authors analyzed a web traffic dataset to study the monetary cost of delivering an advertisement to mobile devices, the authors concluded that in a typical data plan the cost paid by a user to download an ad is higher than the RTB costs paid by advertisers. Another related work to ours is [83], which explored the value of Facebook users using a dataset of 50K users from New Orleans and proposes a model based on the actions of a user in Facebook. Liu et al. [87] analyzed the CPM value of Facebook users leveraging the same API we used to study Facebook prices. Actively running ad campaigns on and tapping on the http(s) traffic traffic [25] shows that advertising to mobile devices costs higher than desktop prices, with IOS being more costly advertising target than Android devices. Our current work differs from the above related works in the sense that the object of these papers is targeted on a single platform and therefore do not offer an online service covering major platforms at large scale nor an analysis of price variability across the three dimensions we covered: geography, demography and categories. Moreover, to the best of our knowledge there is no previous work targeting development of a system offering unified platform for online advertising price comparison service of valuable to various stake holders starting from end-users and marketers to policy makers.
4.3 PriCom System Design

In this section we describe PriCom, a price comparison system that enables to assess the value different configurable audiences profiles across four major advertising platforms: Google AdWords, YouTube, Facebook and Instagram. A preliminary prototype of PriCom is described in [88] whereas its final version is publicly available here[2]

4.3.1. priCom Design Constraints

The final goal of priCom is to provide users with a user friendly platform that will allow them to understand the value that the online advertising market assigns to a persona described by a combination of targeting parameters. This ambitious goal has to deal with the following challenges.

A) The online advertising market is not transparent enough, and it is not easy to access the collective bidding information associated by the advertisers. Therefore, the first challenge is to find appropriate and relevant sources of information that serve as input for the tools we develop. To the best of our knowledge, only a few popular advertising platforms widely provide information regarding aggregated bidding information for particular audiences. The following are some of the major advertising systems that provide bid estimate for a particular user profile defined by a combination of targeting parameters: Facebook, Instagram, Google, YouTube.

B) Creating sophisticated crawlers that allow us to access first-hand bidding information available through the referred data sources. Note, that each company provides distinct mechanisms to access the bidding information intended for their potentially paying customers, and thus each information source will require unique implementation of specific crawlers. Whenever it is possible we exploit APIs made available by the information providers.

C) Storage of the crawled information has to be efficient and robust DBMSs that has to fulfill the following requirements: (i) large storage capacity to store time-based (e.g., daily, hourly, etc.) bidding information for a large number of audiences, (ii) easy access to user queries via a Graphical User Interface (GUI), (iii) flexibility to extend the data sources in an easy manner, and (iv) efficient design that allows an easy integration with backend crawlers that periodically collect information and the front-end interface to end-users.

D) Design and implement simple and accessible front-end UIs targeting average users in order to maximize the adoption of the developed tools. The front-end of the tools developed in this task should be very simple and intuitive in order to maximize the number of users that can use the tool. In addition, it needs to be portable to many operative systems, web
browsers, mobile environments, etc. so as to maximize the number of users with access to the platform.

4.3.2. priCom Implementation

Most advertising platforms offer their (potential) advertisers web interfaces or APIs to configure their ad campaigns that let them reach their desired set of users. A campaign is defined by a specific target audience. These web interfaces and APIs provide advertisers information with respect to the audience they have configured including a reference price required to show ads to such audience with the potential size of the targeted audience. PriCom implements ad-hoc crawlers that leverage the APIs or web interfaces of the four major platforms mentioned above to retrieve the daily prices reported by them for each of the configured profiles.

Figure 4.1: Architectural diagram of the price comparison system.

In the rest of this section we explain in more detail the design of PriCom, that as depicted in Figure 4.1 is formed by the following four main components:

- **Configuration module**: This module serves as the user interface, where PriCom’s users can define up to two different audiences to retrieve their price in parallel. In particular, the user first selects an advertising platform (among the four available). After that, the PriCom’s user can set-up a profile among the different geographic, demographic and behavioural parameters offered by the selected advertising platform. Finally, the user can choose among the available pricing schemes offered by the platform (CPM, CPC, or CPV).
Therefore the configuration set-up is formed by the advertising platform, the user profile characteristics and the pricing scheme. Upon the configuration of these parameters, the user can submit her request to the system. Such request will be handled by the "Storage and processing module" and the results will be shown by the "Presentation module".

- **Crawling module:** It is a distributed master-slave system using several crawling micro-services. The master contacts each advertising platform and automatically identifies the list of geographic, demographic and behavioral targeting parameters provided by each platform. Then it schedules the crawling micro-services (i.e., the slaves) by assigning to each of them an advertising platform as well as a list of profiles to crawl (defined by specific geographic, demographic and behavioral parameters). Each micro-service will run a crawling process to retrieve a daily price for the list of assigned profiles. The retrieved information is passed to the "Storage and processing module".

- **Storage and processing module:** This module serves two purposes. First, it handles the information received from the Crawling module and stores it in a central database. The information is processed and stored in the appropriate format for its presentation as time series of prices by the "Presentation module". Second, it provides an API service to the "Configuration" and "Presentation" modules. This API transforms the configuration set-up defined by the user in the "Configuration Module" into a query format to retrieve information from the database. Similarly, the information extracted from the database is sent in a response format to the "Presentation module", which uses the response to present the results of price variability of the considered profile in a time-series format.

- **Presentation module:** This module is responsible to generate a friendly visualization of the time-series data produced by the "Storage and processing module" API. The resulting visualization allows the user to browse through the price of the considered audience in the full time window where data for such profile is available.

The developed platform is deployed in the openstack infrastructure using a custom built API to transfer data between crawler instances developed with python programming language and the webserver instance which hosts the presentation module and the data repository which is a mysql database. The presentation module leverages amcharts\(^5\) to graphically display the price charts which are fetched from the data repository in a JSON format.

The system is used to monitor prices of overall more than 441K audiences across the four targeted advertising platforms. Table 4.1 shows detailed information with respect the number of geographic and behavioural inputs considered for each platform as well as the overall number of audiences monitored as a combination of the different data inputs. Finally, the table also presents the pricing metrics collected for each platform.

\(^5\)www.amcharts.com/
30 Measuring Value of Online Users

(a) Demography of beta testers
(b) average time spent on Internet
(c) Internet user level

Figure 4.2: Distribution of beta testers: the left panel shows the demographic distribution of users, the middle panel shows average time spent on internet showing that majority of users spend less than 5 hours per day. The right panel shows self reported Internet user level.

Table 4.1: Dataset Description

<table>
<thead>
<tr>
<th>AD platform</th>
<th>behavioral</th>
<th>geographic</th>
<th>transaction</th>
<th>#audiences</th>
<th>#data points</th>
</tr>
</thead>
<tbody>
<tr>
<td>Google</td>
<td>712</td>
<td>191</td>
<td>CPC</td>
<td>28.8K</td>
<td>1.28M</td>
</tr>
<tr>
<td>YouTube</td>
<td>133</td>
<td>129</td>
<td>CPV</td>
<td>10.9K</td>
<td>0.21M</td>
</tr>
<tr>
<td>Facebook</td>
<td>290</td>
<td>240</td>
<td>CPC, CPM</td>
<td>201.6K</td>
<td>68.4M</td>
</tr>
<tr>
<td>Instagram</td>
<td>289</td>
<td>236</td>
<td>CPC, CPM</td>
<td>199.8K</td>
<td>65M</td>
</tr>
</tbody>
</table>

4.3.3. Usability Assessment

The described design provides a high scalability performance, allowing the prototype to collect information for more than 441K audiences daily using two virtual machine crawling instances running on open stack platform with 20GB HDD and 2GB RAM.

Moreover, the tool has been used as part of the European H2020 Project TYPES\(^6\) to improve awareness of users about value of personal data in online advertising. Surveys conducted to more than 191 PriCom users show that 78.8% of them find the tool useful and 27.1% stated that they were very surprised by the discrepancy in the values that different audiences generate for the online advertising stakeholders. The aggregate demography of users and their Internet expertise level is shown on Figure 4.2. These observations present a clear indication of the usability as well as the value of PriCom from a social and educational perspective.

4.4. Analysis of prices of user profiles

In this section we analyze the price of different audiences in the online advertising system. To this end, we leverage the data collected by PriCom for more than 441K audiences between Oct 2016 and June 2018. We specifically dissect our dataset based on geographic, demographic and behavioural parameters in order to understand what parameters define existing biases in users’ profiles prices. The analyses unless stated otherwise: (i) will be conducted for the four consid-

\(^6\)http://www.types-project.eu
4.4 Analysis of prices of user profiles

4.4.1. Overall comparison of advertising platforms

We start our analysis by a high-level comparison of prices across platforms to check the existence of price variability and to understand which of them are more expensive. Figure 4.3 shows the distribution of absolute reference prices for Instagram, Facebook, Google, and YouTube respectively with geographic and interest based targeting. Note that while Google adwords offers a single price estimate value, other platforms return maximum and minimum price for the specified profile. In this analysis we have used the average values between these price estimates (the maximum and minimum returned for a profile configuration) as reference price value of a profile. If we compare the median values of the boxplots distributions, we can see that Instagram is the platform where advertisers are willing to pay most for clicks followed by Facebook, Google and finally YouTube (remind that in YouTube we are considering CPV as reference price).

In the following sections we will provide further details of our cross platform analysis as we analyze the influence of geographic, demographic and behavioural parameters on profile prices. Moreover, we will also compare the temporal variability of prices in the different advertising platforms.

4.4.2. Influence of geographic location in profiles’ prices

Advertising and marketing campaigns typically operate with a predefined geographical scope. Based on this, it is common that campaigns are setup to target specific countries although we can

---

Note that the primary pricing scheme for YouTube video ads is CPV.
find campaigns with broader or narrower scope targeting several countries (e.g. Latin America countries), a region or a city [89].

We first focus our analysis at the country level. Figure 4.4 presents in a map the median price of profiles grouped by country locations for each of our four advertising platforms. The figure shows the median reference price for each country across the time period considered in our dataset. A visual inspection of the results suggest that countries in North America, Northern and Central Europe as well as UK, Australia and New Zealand present the highest prices. On the contrary, African and Asian countries seem to show the lowest prices. To provide further detail, Tables 4.2 and 4.3 show the list of 10 countries showing the highest and lowest prices in each advertising platform, which reinforce our observation. Based on these results, we conjecture that users’ value may be correlated with the wealth level of their country. In other words, showing ads to users in rich countries is more expensive than in poor countries. To validate this hypothesis, we have computed the Spearman and Pearson correlation between the profile’s price and the GDP for each country in our dataset with p-values < 0.001. The results are shown in Table 4.4. These correlation results show that geographical advertising prices are highly correlated with the GDP of countries.

Our visual inspection also seems to indicate a geographical correlation of prices across the four studied platforms. To validate this observation we have computed the Spearman and Pearson correlation of the country prices for each pair of the considered ad platforms with p-values < 0.001. Table 4.5 shows high correlation values for every pair of platforms, confirming that the set of expensive and cheap countries are in general similar across all analyzed platforms. Moreover, a high Pearson indicates that countries showing very expensive prices (outliers) are expensive in all studied ad-platforms.

Now, we narrow down our geographical analysis to the level of states within a country using USA as example. Considering Facebook, we repeated the country level experiments described above with region level dataset collected in June 2018. Figure 4.5 shows the median price of profiles across USA states. As in the case of countries, we also observe an important variability in prices (up to 2.60 € in median CPC, and 6.69 € in median CPM) across states. We have computed the correlation between the GDP per capita of states and their associated advertising price. We observe a non-negligible but significantly smaller correlations at the state-level compared to country-level. In particular, the pearson (spearman) correlations are 0.51 and 0.31 (0.53 and 0.32) for CPC and CPM prices respectively with p-values < 0.001 for correlations with CPC and <0.03 for correlations with CPM respectively.

Advertising campaigns typically target a specific geographic region and in many occasions are country-based. Therefore, the results of our country-based analysis are informative to citizens and society agents. Furthermore, the results from our state-level analysis show that using information at state-level may be very useful for advertisers in order to optimize the configuration of their

---

4.4 Analysis of prices of user profiles

Table 4.2: Most expensive countries on All platforms. The displayed prices are in euro cents.

<table>
<thead>
<tr>
<th>Country</th>
<th>Google Price</th>
<th>YouTube Price</th>
<th>Instagram Price</th>
<th>Facebook Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>63.0</td>
<td>11.5</td>
<td>2475.0</td>
<td>United States 218.0</td>
</tr>
<tr>
<td>Germany</td>
<td>64.0</td>
<td>11.5</td>
<td>2590.0</td>
<td>Denmark 219.5</td>
</tr>
<tr>
<td>Norway</td>
<td>65</td>
<td>Canada 12.0</td>
<td>UAE 2607.5</td>
<td>Germany 222.0</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>72</td>
<td>Netherlands 12.0</td>
<td>Canada 2674.0</td>
<td>Sweden 226.0</td>
</tr>
<tr>
<td>New Zealand</td>
<td>77</td>
<td>Germany 12.5</td>
<td>Finland 2709.0</td>
<td>South Georgia 230.0</td>
</tr>
<tr>
<td>Switzerland</td>
<td>81</td>
<td>United Kingdom 13.0</td>
<td>Sweden 2778.0</td>
<td>Australia 236.0</td>
</tr>
<tr>
<td>Canada</td>
<td>91</td>
<td>United States 14.0</td>
<td>Australia 2836.0</td>
<td>Finland 237.0</td>
</tr>
<tr>
<td>United Kingdom</td>
<td>94</td>
<td>Norway 14.5</td>
<td>Japan 2974.5</td>
<td>Switzerland 274.0</td>
</tr>
<tr>
<td>Australia</td>
<td>102</td>
<td>Switzerland 14.5</td>
<td>Norway 3003.0</td>
<td>Norway 298.5</td>
</tr>
<tr>
<td>United States</td>
<td>136</td>
<td>New Zealand 16.50</td>
<td>United States 3015.5</td>
<td>Japan 349.0</td>
</tr>
</tbody>
</table>

Table 4.3: Least expensive countries on All platforms: the displayed prices are in euro cents.

<table>
<thead>
<tr>
<th>Country</th>
<th>Google Price</th>
<th>YouTube Price</th>
<th>Instagram Price</th>
<th>Facebook Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Christmas Island</td>
<td>1</td>
<td>Gabon 1.50</td>
<td>Venezuela 57.5</td>
<td>Libya 4.0</td>
</tr>
<tr>
<td>Central African Rep.</td>
<td>2</td>
<td>Fiji 1.50</td>
<td>W. Sahara 66.0</td>
<td>Afghanistan 5.0</td>
</tr>
<tr>
<td>Comoros</td>
<td>2</td>
<td>Ethiopia 1.50</td>
<td>Tuvalu 73</td>
<td>Algeria 5.0</td>
</tr>
<tr>
<td>Turkmenistan</td>
<td>2</td>
<td>Uganda 1.50</td>
<td>Tunisia 83.0</td>
<td>Timor-Leste 5.5</td>
</tr>
<tr>
<td>Burundi</td>
<td>2</td>
<td>Tuvalu 1.50</td>
<td>Cabo Verde 87.5</td>
<td>Madagascar 6.0</td>
</tr>
<tr>
<td>Eritrea</td>
<td>2</td>
<td>Tunisia 1.50</td>
<td>Libya 91.0</td>
<td>Nepal 5.5</td>
</tr>
<tr>
<td>Cook Islands</td>
<td>2</td>
<td>Trinidad &amp; Tobago 1.50</td>
<td>Turkmenistan 92.5</td>
<td>Bangladesh 5.5</td>
</tr>
<tr>
<td>Guinea-Bissau</td>
<td>2</td>
<td>Togo 1.50</td>
<td>N. Korea 97.5</td>
<td>Myanmar 6.0</td>
</tr>
<tr>
<td>Antarctica</td>
<td>2</td>
<td>Benin 1.50</td>
<td>Algeria 100.5</td>
<td>Guinea 6.5</td>
</tr>
<tr>
<td>Western Sahara</td>
<td>3</td>
<td>Tonga 1.50</td>
<td>Somalia 103</td>
<td>Tunisia 6.5</td>
</tr>
</tbody>
</table>

campaigns inside a country, for instance, reducing the budget dedicated to extremely expensive states.

4.4.3 Gender based price variability

One of the most relevant demographic properties in marketing and online advertising is gender enabling advertisers target a specific gender type such as showing advertisements on Facebook to females only. In marketing there is a very well known marketing concept known as "pink tax". This refers to the extra price that women pay for the same product in comparison to men [90,91]. In this section we analyze whether this "pink tax" translates into the profile prices of men vs.

Table 4.4: Pearson and spearman correlation between advertising price and GDP of the country

<table>
<thead>
<tr>
<th>AD platform</th>
<th>Spearman correlation</th>
<th>Pearson correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>0.81</td>
<td>0.78</td>
</tr>
<tr>
<td>Instagram</td>
<td>0.73</td>
<td>0.79</td>
</tr>
<tr>
<td>YouTube</td>
<td>0.67</td>
<td>0.78</td>
</tr>
<tr>
<td>Google</td>
<td>0.71</td>
<td>0.72</td>
</tr>
</tbody>
</table>
Table 4.5: Geographic price correlation (pearson/spearman) between ad-platforms

<table>
<thead>
<tr>
<th></th>
<th>Facebook</th>
<th>Instagram</th>
<th>YouTube</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>Facebook</td>
<td>1</td>
<td>0.89/0.828</td>
<td>0.87/0.69</td>
<td>0.81/0.69</td>
</tr>
<tr>
<td>Instagram</td>
<td>-</td>
<td>1</td>
<td>0.91/0.77</td>
<td>0.85/0.83</td>
</tr>
<tr>
<td>YouTube</td>
<td>-</td>
<td>-</td>
<td>1</td>
<td>0.86/0.69</td>
</tr>
<tr>
<td>Google</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>1</td>
</tr>
</tbody>
</table>

women. We conduct our analysis for two of the considered advertising platforms, Facebook and Instagram, that offer this gender based targeting parameter through their APIs. In particular, we compare two pricing metrics CPM and CPC. Figure 4.6 shows the distribution of Facebook and Instagram CPM and CPC prices associated to men and women profiles in our dataset.

If we consider median values, we observe a slight overall bias of higher prices for men than women, with the exception of Facebook CPM that is slightly higher for women than men. These results seem to indicate that the ”pink tax” in the price of products is not present as a system bias in the price advertisers pay for female vs. male profiles in the online advertising ecosystem. While the existence of a systemic pink tax is not apparent in the online advertising ecosystem, if we focus on specific interests we can identify substantial differences between male and female prices based on their behavioral interests. We observe that price differences can be up to 1.46x and 1.41x for interests such as Strategy games and Dating. Note that to compute this price difference first the price differences are calculated as difference between female and male and then this difference is normalized in a scale [0,1] using the maximum difference between female and male advertising prices. This is to avoid bias towards more expensive profiles (e.g. male 400 euro cents vs female 450 euro cents and male 1 euro cents vs female 2 euro cents). The prices of interests also vary depending on the platform. Also note that these price inequalities also differ based on the pricing scheme on each platform. This reinforces our previous argument that advertisers may benefit from price variability by adjusting pricing methods on different platforms so that they can reach more audiences with cheaper budgets. For example an advertiser planning to target Facebook users with interest in non-fiction books could start with the observation that CPC pricing scheme is more expensive for females while CPM pricing is expensive for males. Hence, instead of assigning equal budget, the advertiser could add more campaign budget for the CPM pricing scheme to reach females and CPC pricing scheme to reach males. Note that this kind of approach might not always work since some categories are specifically more interesting to certain gender. Examples of such categories are dresses and women’s closing which are often specific to females.
4.4 Analysis of prices of user profiles

4.4.4. Price variability across users’ interests

The online advertising ecosystem operates as an open market where prices of certain profiles increase or decrease as function of the demand of such profile among advertisers and the budget of interested advertisers. Therefore, profiles of interest for a large number of advertisers with large budgets are more expensive than profiles of interest to small budget advertisers.

In this subsection, we analyze the price variability of the online advertising market across profiles defined by end users’ interests in the four analyzed advertising platforms. In particular, Figure 4.7 shows the distribution of the median value of the end users’ interests for our advertising platforms. In particular, the distribution is obtained from 290, 289, 712, and 133 interests in Facebook, Instagram, Google, and YouTube, respectively.

In order to quantify the price variability of each platform we have defined a metric $V$ computed as the ratio IQR/Median of the distribution. Note that this is a common metric used to quantify variability. The values of $V$ for each of the platforms are: $V_{FB} = 0.16$, $V_{Instagram} = 0.09$, $V_{Google} = 1.89$ and $V_{YouTube} = 0.71$. These results indicates that Google is the platform presenting a significantly larger price variability, 2.7 times larger than YouTube, which shows the second largest variability. Next we look at variability of user interests on these platforms. Table 4.6 shows the list of the 10 most expensive interests per advertising platform whereas Table 4.7 shows the list of the 10 cheapest interests. These results indicate that there seems to be a low overlapping among different platforms. This may represent an opportunity for advertisers and marketers since they can optimize their ad campaigns by selecting cheaper ad platforms for the specific audience of the campaign or vice versa.

To confirm this lack of overlapping, we have grouped the interests of each platform in to 7 parent categories. We have used the Leacock-Chorodow semantic distance [8] to find similarity between categories that were not exactly matched[9] and computed the median price of each parent category in each platform. Table 4.8 shows the list of these parent categories sorted from most expensive to least expensive in each advertising platform. We observe that the relative cost of a given category significantly vary across platforms. For example, the results indicate that technology is among most expensive category on Instagram while it is the cheapest category after Food & drink on Facebook. On the other hand, "fitness & wellness" is the most expensive category on Facebook and YouTube, while this category is the second cheapest category on Google.

The analysis conducted in this section reinforces the already mentioned usefulness or our tool for advertisers. The ability to compare the cost of similar categories across platforms will help advertisers to optimize their campaign planing from a budgetary point of view.

---

[9] Note that Facebook and Instagram share a common set of configurable interest names which differ from the set shared by Google and YouTube. This makes a head-to-head comparison at the level of individual interest very difficult.
Table 4.6: Interests showing highest advertising prices (in euro cents) on each platform: Interests are listed in increasing order. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube.

<table>
<thead>
<tr>
<th>Facebook</th>
<th>Instagram</th>
<th>Google</th>
<th>YouTube</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comedy films</td>
<td>Web hosting</td>
<td>Technical Support</td>
<td>E-Commerce Services</td>
</tr>
<tr>
<td>(34.0)</td>
<td>(425.5)</td>
<td>(574.0)</td>
<td>(7.0)</td>
</tr>
<tr>
<td>Science fiction films</td>
<td>Motorhomes</td>
<td>Real Estate Listings</td>
<td>Service Providers</td>
</tr>
<tr>
<td>(34.0)</td>
<td>(433.5)</td>
<td>(607.0)</td>
<td>(7.0)</td>
</tr>
<tr>
<td>Anime films</td>
<td>GPS devices</td>
<td>Aerospace &amp; Defense</td>
<td>Professional &amp; Trade Associations</td>
</tr>
<tr>
<td>(34.5)</td>
<td>(434.5)</td>
<td>(624.0)</td>
<td>(7.5)</td>
</tr>
<tr>
<td>Dancehalls</td>
<td>Camcorders</td>
<td>Enterprise Resource Planning</td>
<td>Food Service</td>
</tr>
<tr>
<td>(34.5)</td>
<td>(440.0)</td>
<td>(639.5)</td>
<td>(7.5)</td>
</tr>
<tr>
<td>Vietnamese cuisine</td>
<td>Email marketing</td>
<td>Enterprise Technology</td>
<td>Bars, Clubs &amp; Nightlife</td>
</tr>
<tr>
<td>(35.0)</td>
<td>(444.0)</td>
<td>(666.0)</td>
<td>(7.5)</td>
</tr>
<tr>
<td>Bodybuilding</td>
<td>Investment banking</td>
<td>Exchange Traded Funds</td>
<td>Health Insurance</td>
</tr>
<tr>
<td>(35.5)</td>
<td>(454.5)</td>
<td>(734.5)</td>
<td>(8.0)</td>
</tr>
<tr>
<td>Online advertising</td>
<td>Audio equipment</td>
<td>Customer Relationship Management</td>
<td>Live Sporting Events</td>
</tr>
<tr>
<td>(36.5)</td>
<td>(466.0)</td>
<td>(869.5)</td>
<td>(8.0)</td>
</tr>
<tr>
<td>Role-playing games</td>
<td>eBook readers</td>
<td>Teleconferencing</td>
<td>Weather</td>
</tr>
<tr>
<td>(37.0)</td>
<td>(499.5)</td>
<td>(954)</td>
<td>(8.0)</td>
</tr>
<tr>
<td>German cuisine</td>
<td>Display advertising</td>
<td>Stock Photography</td>
<td>Programming</td>
</tr>
<tr>
<td>(39.0)</td>
<td>(508.5)</td>
<td>(1079)</td>
<td>(8.5)</td>
</tr>
<tr>
<td>Non-fiction books</td>
<td>Projectors</td>
<td>Project Management Software</td>
<td>Thriller, Crime &amp; Mystery Films</td>
</tr>
<tr>
<td>(47.0)</td>
<td>(511.5)</td>
<td>(1276)</td>
<td>(11.5)</td>
</tr>
</tbody>
</table>

4.4.5. Temporal Variability of Prices

The aforementioned market nature of the online advertising ecosystem, makes that the prices of users’ profiles fluctuate along time depending on the advertisers’ demand on those profiles. In this subsection, we analyze the temporal evolution of profiles’ prices in the four considered advertising platforms over the range of 21 months between Oct 2016 and June 2018 covered by our dataset based on 13 months for YouTube, 14 months for Google and 16 months for Facebook and Instagram.

First, we analyze the overall price trend of each of the four considered platforms. To this end, we consider the time series of daily median price on each platform between Oct 2016 and June 2018 to estimate the advertising price trend.\textsuperscript{10} To unveil the time series trend we aggregate the price dataset at monthly level calculating the median value of each month, then we normalize

\textsuperscript{10}Note that the median value is computed from the user profiles shown in Table 4.1
Table 4.7: Interests showing cheapest advertising prices (in euro cents) on each platform: Interests are listed in increasing order. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube.

<table>
<thead>
<tr>
<th>Facebook</th>
<th>Instagram</th>
<th>Google</th>
<th>YouTube</th>
</tr>
</thead>
<tbody>
<tr>
<td>Network storage</td>
<td>Hair products</td>
<td>Track &amp; Field</td>
<td>Documentary Films</td>
</tr>
<tr>
<td>(22.5)</td>
<td>(259)</td>
<td>(1.0)</td>
<td>(1.5)</td>
</tr>
<tr>
<td>Audio equipment</td>
<td>Women’s clothing</td>
<td>Public Safety</td>
<td>Dance &amp; Electronic</td>
</tr>
<tr>
<td>(23.0)</td>
<td>(267)</td>
<td>(1.0)</td>
<td>Music (1.5)</td>
</tr>
<tr>
<td>GPS devices</td>
<td>Dresses</td>
<td>Livestock</td>
<td>Music Videos</td>
</tr>
<tr>
<td>(23.0)</td>
<td>(267.5)</td>
<td>(1.0)</td>
<td>(1.5)</td>
</tr>
<tr>
<td>Portable media</td>
<td>Tattoos</td>
<td>Legal Forms</td>
<td>Country Music</td>
</tr>
<tr>
<td>player (23.0)</td>
<td>(269)</td>
<td>(1.0)</td>
<td>(1.5)</td>
</tr>
<tr>
<td>Retail banking</td>
<td>Do it yourself</td>
<td>Law Enforcement</td>
<td>Animated Films</td>
</tr>
<tr>
<td>(23.0)</td>
<td>(DIY) (270.5)</td>
<td>(1.0)</td>
<td>(1.5)</td>
</tr>
<tr>
<td>Children’s clothing</td>
<td>Cosmetics (272)</td>
<td>Geology (1.5)</td>
<td>Movies (2.0)</td>
</tr>
<tr>
<td>(23.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ecotourism</td>
<td>Chocolate</td>
<td>Campaigns &amp;</td>
<td>Bollywood &amp; South</td>
</tr>
<tr>
<td>(23.5)</td>
<td>(272.5)</td>
<td>Elections (2.0)</td>
<td>Asian Film (2.0)</td>
</tr>
<tr>
<td>Display advertising</td>
<td>American football</td>
<td>Cameras (2.0)</td>
<td>Horror Films</td>
</tr>
<tr>
<td>(23.5)</td>
<td>(272.5)</td>
<td></td>
<td>(2.0)</td>
</tr>
<tr>
<td>Fast casual</td>
<td>Singing</td>
<td>Table Games</td>
<td>Classical Music</td>
</tr>
<tr>
<td>restaurants</td>
<td>(272.5)</td>
<td>(2.0)</td>
<td>(2.0)</td>
</tr>
<tr>
<td>(23.5)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Discount shops</td>
<td>Rhythm and blues</td>
<td>Public Policy</td>
<td>Martial Arts Films</td>
</tr>
<tr>
<td>(23.5)</td>
<td>music (272.5)</td>
<td>(2.0)</td>
<td>(2.5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

the prices to a scale \([0,1]\) to enable cross platform comparison, finally we apply regression on the normalized monthly level dataset to find the slope of the trend which shows us the direction of the price trend. Figure 4.8 shows the obtained results. The figures seem to indicate that Facebook and Instagram show a relatively increasing trend as compared to Google and YouTube. To quantify this trend we have computed the slope of the trend line. The results indicate that Facebook doesn’t have a clear trend (slope value 0.001 with r-squared 0.001 and p-value 0.91. While YouTube seems to have a decreasing trend with slope -0.026 the low statistical r-squared and p-value (0.31, 0.05) show that the trend is not very strong. The results for Instagram showed a

Table 4.8: Interest categories ordered by their decreasing price in each platform. Median prices in euro cents are listed in brackets. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube.

<table>
<thead>
<tr>
<th>Facebook</th>
<th>Instagram</th>
<th>Google</th>
<th>YouTube</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fitness &amp; wellness</td>
<td>Business and industry</td>
<td>Business and industry</td>
<td>Fitness &amp; wellness</td>
</tr>
<tr>
<td>(32.5)</td>
<td>(332.0)</td>
<td>(89.0)</td>
<td>(6.5)</td>
</tr>
<tr>
<td>Entertainment</td>
<td>Technology</td>
<td>Family &amp; relationships</td>
<td>Business &amp; industry</td>
</tr>
<tr>
<td>(30.0)</td>
<td>(331.0)</td>
<td>(68.5)</td>
<td>(6.0)</td>
</tr>
<tr>
<td>Family &amp; relationships</td>
<td>Entertainment</td>
<td>Technology</td>
<td>Family &amp; relationships</td>
</tr>
<tr>
<td>(29.5)</td>
<td>(304.5)</td>
<td>(49.5)</td>
<td>(5.5)</td>
</tr>
<tr>
<td>Business and industry</td>
<td>Fitness &amp; wellness</td>
<td>Food and drink</td>
<td>Technology</td>
</tr>
<tr>
<td>(29.0)</td>
<td>(303.0)</td>
<td>(49.0)</td>
<td>(3.5)</td>
</tr>
<tr>
<td>Hobbies &amp; activities</td>
<td>Food &amp; drink</td>
<td>Entertainment</td>
<td>Hobbies &amp; activities</td>
</tr>
<tr>
<td>(26.0)</td>
<td>(298.0)</td>
<td>(48.5)</td>
<td>(4.75)</td>
</tr>
<tr>
<td>Technology</td>
<td>Family &amp; relationships</td>
<td>Fitness &amp; wellness</td>
<td>Food &amp; drink</td>
</tr>
<tr>
<td>(26.0)</td>
<td>(296.6)</td>
<td>(41.0)</td>
<td>(4.5)</td>
</tr>
<tr>
<td>Food &amp; drink</td>
<td>Hobbies &amp; activities</td>
<td>Hobbies &amp; activities</td>
<td>Entertainment</td>
</tr>
<tr>
<td>(25.5)</td>
<td>(296.5)</td>
<td>(33.5)</td>
<td>(3.0)</td>
</tr>
</tbody>
</table>
sloe of 0.026 and r-squared 0.37 with p-value 0.01 indicating an increasing price trend. Google on the other hand showed a decreasing price slope of -0.039 and r-squared values of 0.57 and p-value 0.001. Interestingly, we see that different advertising platforms present different price trends. This information is of high interest not only for marketers and advertisers, while defining their media investment strategy, but also for investors.

After analyzing the overall trend of advertising platform prices, we now focus on analyzing the price variability of individual users’ profiles. To this end we consider as reference the peak-to-valley price difference metric defined as the difference between the highest and lowest price for a given profile in the time series. However, this metric is very sensitive to outliers (i.e., extremely high or low prices) resulting from errors in platform reporting system or our own measurement infrastructure. Therefore, we consider a proxy metric to the peak-to-valley price difference which is robust to outliers, this is the 95-to-5 percentile price difference. This metric measures the difference between the 95 and 5 percentile price for a given profile. Figure 4.9 shows the normalized distribution of this metric across all profiles in the considered advertising platforms. We observe that the 95-to-5 percentile prices vary in all platforms with Instagram showing highest variability followed by Facebook and YouTube, with Google showing lowest variability. These results show that the price fluctuation of certain profiles may be significant in the online advertising market. These results confirm the usefulness of priCom for advertisers and marketers to plan their advertising campaigns. In addition, priCom can be used by regulators to monitor the temporal evolution of prices to assess, among other things, whether the free competition required by law in many countries of the world is being altered, for instance by non-public agreement among platforms to increase the advertising costs.

We observe that advertising prices vary not only based on location and gender but they also vary temporally.

4.5. Discussion

The advertising price comparison system PriCom and the analysis of its associated data generate added value to main actors around the online advertising ecosystem, namely users, regulatory bodies, and advertisers/marketers. PriCom’s primary goal is creating awareness among citizens about the fact that online personal information has a value and let them get a sense of what this value might be. To this end we are in communication with different associations to get PriCom integrated as part of awareness creation services. A first achievement in this direction is the integration of PriCom as part of the TestYourPrivacy portal.11

As we have shown in previous sections PriCom can be used to assess differences based on demographic parameters such us gender what may serve to trigger warnings with respect to the presence of discriminatory practices by the online advertising industry. For instance, it can be checked whether certain type of ads (e.g., job, education, etc.) present strong gender biases

11http://www.testyourprivacy.eu
(i.e., price differences), which may reflect an interest in the industry on delivering ads related to that topic to one specific gender.

Third, marketers and advertisers can benefit from the methodology developed by PriCom and analysis of data collected in order to optimize their advertising strategies in several dimensions: for creating an optimized campaign by leveraging priCom for comparison between different user interests located in same geographic locations, cross-platform comparison of prices, insight on temporal evolution of prices across audiences and platforms, etc.

Finally, a new approach to online advertising is gaining momentum. It proposes to give back to users part of the revenues they generate from online ads [93]. The methodology developed in PriCom can serve as an auxiliary service in this context that will help users to estimate which is the approximate value they generate and thus have a rough idea of the compensation they should receive. Online advertising intermediaries adopting this new concept may use PriCom as a reference to define the monetary compensation plans to offer to users.

### 4.6. Conclusion

In this chapter we present a price comparison system for online advertising by considering four major advertising platforms. The developed system allows to see the monetary values of different attributes representing user information from the actual advertising market perspective. The system leverages its crawling modules to collect data provided by the advertising platforms for their intended advertisers or potential advertisers. Analyzing the data collected using this system we show that user profiles are valued differently based on geography, user interest and gender in all of four platforms. The developed price comparison system is available at [http://dataportal.netcom.it.uc3m.es/betac2/](http://dataportal.netcom.it.uc3m.es/betac2/)
Figure 4.4: Geographic advertising price variability on the four major advertising platforms: Facebook (CPC), Google (CPC), YouTube (CPV) and Instagram (CPC). Countries are colored from blue (indicating the cheapest countries) to green (indicating the most expensive countries). Countries with no data are colored in red.
Figure 4.5: General advertising price variability across US states: the left figure shows the price variability in CPC billing scheme, while the figure on the right shows advertising price variability in CPM. States are colored from purple (indicating the cheapest state) to red (indicating the most expensive state), the inset figures show the prices in Alaska and Hawaii respectively.
Figure 4.6: Gender based advertising price variability on Facebook and Instagram
4.6 Conclusion

Figure 4.7: Interest based advertising price variability across four advertising platforms: the values on the y-axis show the median prices per category. Indicated prices are in CPC for Google, Facebook, and Instagram and CPV for YouTube.
Figure 4.8: Temporal advertising price variability across four advertising platforms: the values on the y-axis show the normalized median prices per month.
Figure 4.9: Temporal Price variability across user profiles: the box plots show the difference between the 95 percentile price and the 5 percentile price of a profile.
— if you educate a woman you educate a whole nation.

Dr. James Emman Kwegyir Aggrey

5

Digital Gender Inequality in Access to Online Services

5.1. Introduction

The growing proliferation of the Internet, and Information Communication Technologies (ICT) at large, has created an epidemic of promising Internet services some of which have effectively become globally ubiquitous. However, despite its widespread adoption, this resource is characterized by uneven distribution creating disparate demographic groups with unequal access to the technology – a scenario referred to as digital divide.

Since the first popularization of the term by the then US vice president Al Gore in 1996, the unrelenting digital divide has become a major concern for governments and global organizations including the UN [94–96]. For example, the 2016 World Bank report has suggested that unequally distributed growth in Internet penetration might exacerbate socioeconomic inequalities [97]. Indeed this situation has also brought about some of the audacious projects such as internet.org1 and project loon2 aiming to connect the 4 billion people that remain offline [97,98].

Although the digital divide problem has been widely acknowledged by the research community, efforts to understand the global state of digital inequality, its prevalence in Internet applications and services such as social media, its relation with existing socioeconomic conditions, demographic heterogeneity in engagement on major online platforms are challenged by the lack

1www.internet.org
2google loon project [www.google.com/loon/]

47
of large scale data sources and difficulties to generate unbiased survey data that can provide large-scale measurements across multiple demographics and countries [97].

In this chapter we discuss our methodology and results on study of some of the fundamental questions on digital divide in social media at a global scale from gender perspective. As reported by the [WWW] Foundation, gender inequality is one of the key elements in digital divide [99]. To overcome the limitations of finding global scale representative dataset, we have developed a system to collect large-scale data from the Facebook online social network through its marketing API, the methodology is explained in detail in the Analysis methodology section. The Facebook marketing API has been useful in previous research in different innovative ways such as: to estimate the value of user data [84], to approximate the size and integration of migrant populations [100–102], and to generate estimations of Internet and mobile phone gender gaps that explain 69% of the variance of ITU measurements [67]. Using anonymous dataset collected with this tool, we study gender inequality in social media and its relationship with other economic, education, health, and political gender inequalities, and effect of network externalities. The generated anonymous dataset contains statistical information about the total number of registered users and daily active users of each gender in 217 countries with a total of more than 1.4 Billion users.

5.2. Related Work

Given that digital inequality is a widely recognized issue, various existing works have explored the problem of digital inequality from different perspectives such as: gender differences in Internet and mobile phone addiction [103], studying differences in people’s online skills [104], extent of digital divide in higher education students [105], relation between gender, physical activity, and Internet use [106]. Some works also studied proposals for possible changes in order to bridge the digital divide, providing suggestions such as: the development of light-weight web applications that could run on low-end mobile devices aimed at low-income and developing populations [107], and to concentrate on overcoming the barriers for skills and usage instead of material to fill the digital divide gap within developed countries [108]. Using surveys from five countries Buchi et al. [109] identified four core Internet usage types concluding that the digital divide has shifted from gaps in having access to inequalities in usage in the five countries studied. Based on survey data obtained from the U.S. Federal Communication Commission, Dobransky and Hargittai [110] studied usage of digital media among people with disabilities identifying key differences with people without disabilities, which they called the digital disability divide. Alam et al. [111] studied social inclusion and the digital divide among refugee migrants in Australia using focus group discussions and measured digital divide among refugees attributing the reasons to the inequalities in physical access, skills, and ability to pay for the service. The authors stated the limitation of their approach as it is confined to specific region. However most of these research works seem to miss a global large scale dataset on actual access and usage, as a result they concentrate on small scale social survey data and case studies performed on a specific demography.
or location. Recently the digital inequality is also benefiting from developments in computational approaches leveraging digital traces. Using data from Facebook advertising platform, Zagheni et al. [102] demonstrated the feasibility to study assimilation of migrant populations using the data from the advertising platform. In a related study Fatehkia et al. [67] have demonstrated the potential of data from the Facebook advertising platform to estimate global gender gaps in Internet and mobile phone accesses. Analyzing editing activity logs of users on Wikipedia, Antin et al. [112] measured the extent of gender difference in Wikipedia editors which let them elaborate previous survey based studies [113].

5.3. Datasets

5.3.1. The Facebook Global Dataset

The Facebook dataset with number of Facebook users and Daily Active User (DAU)s per age and gender in each country was collected leveraging the Facebook marketing graph API, which is a primarily used API end point to collect data for advertising purposes. Among other services, this API delivers data for its advertising clients to provide targeted advertising. When supplied with a specific target population, the API returns the total audience size and the price to reach that target audience through Facebook. We iterated each combination of age and gender values, retrieving the total number of users and the number of DAUs for each segment in each country. Figure 5.1 diagrammatically explains the architecture and measurement procedure. Our dataset contains the number of male and female registered users and DAUs for all available countries (the API does not deliver data for certain countries; e.g., Syria, Iran, and Cuba). After removing entries of small countries with missing values or low resolution, our dataset contains the total number of users and DAUs segmented by age and gender for 217 countries. Age data in the API has a granularity of one year and start at 13 years old, increasing by 1 year up to a last bin that contains all users starting from age 65 years old and above (65+). Note that our analysis of the collected country level data from the Facebook marketing API includes only the aggregated public statistical data at country level and hence we have no access to any PII of any user, the procedure had also been approved by different ethical boards (details available at the main paper at [114]).

5.3.2. Gender Equality and Development Datasets

To measure relations with external socioeconomic and demographic aspects we have used the following datasets. To normalize the number of active users over the total population of each country, we use the data collected by the US Census Bureau International Database. This dataset contains estimates of the resident population by age and gender for more than 226 countries. We combine this data with gender equality indices measured by the World Economic Forum.

[https://www.census.gov/programs-surveys/international-programs/about/idb.html](https://www.census.gov/programs-surveys/international-programs/about/idb.html)
Figure 5.1: Diagram representing architecture of the crawling tool used for the measurement: The measurement tool is composed of a master program that first contacts the Facebook authentication server (step 1), after getting authenticated on step 2, it then divides the data collection task between individual crawler programs (shown on step 3), on step 4 each assigned crawler uses the authentication token to establish a connection with the marketing end point from Facebook graph API. After establishing connection, each crawler starts to query the graph API for information by iterating over the list of queries assigned to it by the master crawler. For each valid request the end point returns total number of registered users and number of daily active users together with other information including the cost to reach that desired demographic. On step 5, upon receiving response each crawler writes the response with the queried parameters on a log file.

Gender Gap reports of 2015 and 2016 [115]. This dataset quantifies the magnitude of gender equality in 145 countries, measuring it with respect to four key areas: health, education, economic opportunity, and politics. This report updates the values for education, economic, and political gender equality on a yearly basis, allowing us to measure changes between 2015 and 2016. To account for additional economic and development indicators, we include data from the World Bank and the Human Development Index [116], measuring control variables of gross domestic product at purchasing power parity per capita in 2012, economic inequality as the quintile ratio, and Internet penetration.

5.4. Analysis methodology

5.4.1. Computing the Facebook Gender Divide

This section discusses the methodology used to calculate Facebook Gender Divide (FGD). FGD is calculated based on a comparison of the activity ratios between genders. We define the
Facebook Gender Divide in country $c$ as:

$$FGD_c = \log \left( \frac{R_{\text{Male},c}}{R_{\text{Female},c}} \right)$$

The result of $FGD_c$, which compares male and female Facebook activity rates over the population of country $c$, has the following characteristics.

$$X = \begin{cases} 
\text{Female biased inequality}, & \text{if } FGD_c < 0 \\
\text{Equal activity tendencies}, & \text{if } FGD_c \approx 0 \\
\text{Male biased inequality}, & \text{if } FGD_c > 0 
\end{cases}$$

The rates of activities $R_{\text{Male},c}$ and $R_{\text{Female},c}$ are computed from the ratios of DAU in a given age and total population in that age segment. The DAU measures how many users have logged into Facebook at a given day. We use the segmented data from 13 to 65 years old to normalize the DAU over the total population of a country in those ages, truncating all data that is not included in that age range. This helps us avoid introducing a bias with life expectancy and average age across countries. To have a stable estimate of the DAU, we use as the median value over the month of July 2015, replicating over other months afterwards. With this, for each country $c$ and gender $g \in \{\text{Female}, \text{Male}\}$, we have a measurement of the number of active users $A_{g,c}$ between 13 and 65 years old. Additionally, this allows us to calculate the mean user age for a country to include it in our models. Facebook offers three types of gender based targeting codes regarding sex, which are 0 for all, 1 to target only males, and 2 to target only females. For simplicity, we take gender as birth sex, i.e. male or female. Normalization is done using the US Census Bureau data. We first calculate the total population of each gender between the ages of 13 and 65 years old in each country, which we denote as $P_{g,c}$. Then we normalized the total activity in Facebook over the population in the same age ranges, calculating the activity ratios $R_{g,c} = A_{g,c}/P_{g,c}$. We further compute the Facebook penetration as the ratio between user accounts between 13 and 65 years old reported by the API (total users regardless of activity and gender) and the total population of the country between those ages.

5.4.2. Validating the FGD

Since the primary purpose of the Facebook marketing API is for advertising purposes and also it is a relatively new research tool the results obtained through this measurement need to be validated. For this purpose, we have used the following three reference survey datasets and analyzed their correlation with the Facebook dataset obtained using our measurements:

- The survey of the Global Web Index (GWI). We used the survey responses of the Global Web Index panel during the two last quarters of 2015 and the two first quarters of

https://developers.facebook.com/docs/marketing-api/buying-api/targeting/
https://www.globalwebindex.net/
2016. For this period, the GWI contains responses from 99,338 panelists in 34 countries, providing re-scaled estimates of survey responses that generalize to the population as a whole. We take the responses about the frequency of Facebook use and other social media (Whatsapp, LinkedIn, Twitter, Instagram, and YouTube).

- **Pew Research Center Spring 2016 Global Attitudes (Pew Global)** This dataset includes responses from 23,462 panelists in 19 countries. We use the positive answer rate to question 82 as way to measure the penetration of social networking services (SNS) in general.

- **Pew Research Center Internet & Technology survey** This US centered questionnaire dataset includes questions about Facebook use in the US with gender and age data from 1,601 respondents.

First a correlation test between the total Facebook penetration for both genders based on our API measurements versus the value estimated from the GWI survey was performed. The result indicates a high positive correlation (Pearson 0.89, CI [0.78; 0.94], Spearman 0.86, CI [0.72; 0.94]). Next, the same correlation test against the Pew Global survey was performed. The correlation between these measurements is also positive and high (Pearson: 0.71, CI [0.39; 0.88], Spearman: 0.77, CI[0.43; 0.94]). Finally, the comparison between penetration estimates based on GWI and Pew Global survey data is also positive and significant (Pearson: 0.63, CI [0.17; 0.86], Spearman: 0.7, CI [0.38; 0.91]). Similar procedures were followed to validate Facebook gender divide estimates. Comparison of FGD of Facebook marketing API estimate with the same estimate using GWI data has showed high positive correlation (Pearson: 0.83, CI [0.68; 0.91], Spearman: 0.63, CI [0.27; 0.87]). Similar comparison with Pew survey data for all SNS also reveals high positive correlation (Pearson: 0.85, CI [0.65; 0.94], Spearman: 0.74, CI [0.35; 0.91]). Comparison of FGD estimates using GWI and Pew also showed a positive correlation (Pearson: 0.85, CI [0.6; 0.95], Spearman: 0.49, CI [-0.11; 0.86]). From these results we were able to conclude that the estimate of the FGD using the marketing API is consistent with GWI and Pew survey metrics. Even though these survey datasets are targeted at small scale population, they have allowed us to validate our calculation of the Facebook penetration and FGD through the Facebook marketing API. As we can see the individual correlation of these datasets with our Facebook measurement dataset are as good as the correlations between the survey datasets, showing that the Facebook API data has comparable quality enabling a much higher coverage in terms of countries and better temporal resolution. The results have also showed that, while we should not take Facebook as representative for all social media, there is certain similarity in gender inequality that can motivate future research. The details of validation exercises are reported in detail in Supplementary Text 1 of the Appendix.

5.4 Analysis methodology

5.4.3. Regression Models

After computing the FGD, we have examined dependence relationship between FGD and the gender equality and development datasets. Linear regression was used to model dependence relation between the gender equality indicators and the FGD. Before the regression was performed a rank transformation was applied to all variables such that rank 1 is the highest possible value of the variable, this allows a monotonic dependence relation robust to outliers to a strictly linear relation. We define this FGD model as:

\[ FGD = a_f \cdot Q + b_f \cdot C + c_f + \epsilon \]

where \( Q \) is a matrix with the ranks of economic, health, education, and political gender equality in each country and \( C \) contains control variables such as Internet penetration (IP), income inequality (Ineq), total population (Pop), Facebook penetration (FBP) and mean user age (Age). \( c_f \) is the intercept and \( \epsilon \) denotes the residuals as the normally distributed, uncorrelated error of the model.

The relationship between changes and levels in economic gender equality and of FGD was analyzed with two models.

First, with an equality changes model:

\[ \Delta Eco_{2016} = a_o \cdot Eco_{2015} + b_o \cdot FGD_{2015} + c_o \cdot O + d_o + \psi_o \]

And second with, a FGD changes model:

\[ \Delta FGD_{2016} = a_q \cdot FGD_{2015} + b_q \cdot Eco_{2015} + c_q \cdot O + d_q + \psi_q \]

In the above equations \( \Delta Eco_{2016} \) and \( \Delta FGD_{2016} \) represent the changes in economic gender inequality and FGD between 2015 and 2016. Both models include a control for autocorrelation as a term with the unranked value of the variable in the 2015, and a main term of the re-scaled ranked value of the other variable. Following previous Economics research on Facebook data [117], we include various ranked controls in the matrix \( O \), first with a simple correction for GDP, later with extensions with other controls as for the FGD model.

Network externalities were modeled as a power-law relationship between the activity ratio of a gender \( R_{g,c} \) and the total Facebook penetration for both genders together \( P_c \) in a joint model that includes an intercept for gender and interaction with gender. We define the network externalities model as follows:

\[ \log(R_{g,c}) = \alpha \cdot \log(P_c) + \beta + \delta_{g,Female}(\alpha_F \cdot \log(P_c) + \beta_F) + \phi \]

Where \( \alpha \) measures the scaling relationship between the Facebook presence ratio and the activity ratio of male users, \( \alpha_F \) the difference in that relationship for female users, and \( \phi \) the residuals. The Kronecker delta function \( \delta_{g,Female} \) takes value 1 when \( g = Female \) and 0 otherwise. All the
above models show a Variance Inflation Factor (VIF) value of below 5 which indicate robustness of the models to multicollinearity [118].

In results section we report the FGD model fit and the network externalities model with Markov Chain Monte Carlo sampling in JAGS [119]. To test the validity of the assumptions of our models after fitting, the normality of residuals were verified through Shapiro-Wilk tests [120], and residuals were checked to make sure they are uncorrelated with fitted values and independent variables.

Figure 5.2: The Facebook Gender Divide across 217 countries. This figure shows worldwide FGD values. The FGD values range from red which represents highly skewed values towards males, to green (not visible) which represents countries with skewed values of FGD values towards females. Countries with a balanced FGD are colored blue. The left inset shows the scatter plot of male and female activity ratios across all countries, revealing a spread along the diagonal. The right inset shows the histogram of FGD values in bins of width 0.2. While the mode of countries is slightly below zero, there is significant skewness towards high FGD values (bias towards males).

5.5. Results

5.5.1. Gender Inequality

Figure 5.2 shows a world map with countries colored according to their FGD, revealing that many countries are very close to gender equality in Facebook (blue color). The red scale shows countries with positive FGD—that is, higher proportion of males on Facebook. The range of values towards FGD below zero (more tendency for women to be on Facebook) is much narrower than above zero, as can be seen in the scatter plot with the activity ratios of each gender (Fig. 5.2, left inset), and in the skewness of the distribution of FGD across countries (Fig. 5.2, right inset).

5.5.2. Explaining Gender Inequality

Countries with high FGD are located around Africa and South West Asia, as shown on Fig. 5.2. This suggests that variations in socio-economic factors of gender inequality across regions
could be explanatory of the FGD. We test this observation using a linear regression model of the FGD as a function of the four indices of gender equality measured by the World Economic Forum (economic opportunity, education, health, and political participation), plus five non-gender-based controls of Internet penetration, population size, economic inequality, Facebook penetration, and mean Facebook active user age. The left panel of Fig. 5.3 shows the quality of the model fit, comparing empirical values of FGD rank versus model predictions. As results indicate, the model can explain well the ranking of FGD \( R^2 = 0.74 \), with very few points far from the diagonal. While this result might be partially explained by Facebook using vital statistics in their calculations, it is nevertheless consistent with replications of the model using limited survey samples from the Pew Research center and the Global Web Index as described in the model validation discussed above. This indicates that the performance of the model is not an artifact of the Facebook marketing API.

Figure 5.3: Regression results of FGD as a function of gender equality. A) Model predictions versus rank of FGD, where rank 1 is the country with the highest FGD. The model achieves a high \( R^2 \) above 0.74, explaining the majority of the variance of the FGD ranking. Some countries are labeled, from high FGD (Liberia, India, and Saudi Arabia) to low FGD (Finland, Norway, and Uruguay), as well as some outliers (Dominican Republic, Austria, and Sri Lanka). B) Coefficient estimates and 95% CI of the terms of the regression fit (excluding intercept). Education (Edu), health (Heal), and economic gender equality (Eco) are significantly and negatively associated with the FGD, but political gender equality (Pol) is not. From the control variables, Internet penetration (IP) is negatively associated with FGD, but the rest are not. The main role of education equality in FGD can be observed on panel A, where dots are colored according to the rank of education gender equality, showing that countries with low FGD are ranked high on education gender equality.

The right panel of Fig. 5.3 shows the estimate of the coefficients of our model of FGD. The strongest coefficient is that of education gender equality, which can also be observed on the colors of the left panel of Fig. 5.3. Specifically, countries with high rank in this index have, on average, lower FGD. Health and economic gender equality also have significant negative coefficient estimates, showing that the FGD captures more than one type of inequality. Note that the index for political gender equality does not have a significant relationship with FGD when the other indices are considered in the model.
Among gender-independent controls, only Internet penetration is negatively associated with FGD. Nevertheless, the FGD is also correlated with GDP per capita (Spearman correlation $-0.57$, $p < 10^{-6}$). For that reason, we repeated the model using GDP as a control variable, finding similar results. These results evidence that the relationship between gender equality indices and FGD is observable when development metrics are considered.

5.5.3. Network Effect

The theory of network effect, also called network externalities, states that the value of a given communication channel increases as the number of users increase, adding value to each member as more users join. As the above results suggest the value of being active in social media might vary across genders. The general penetration of a communication channel can increase the value that individuals get for using it, which is an example of a feedback mechanism driven by (positive) network externalities [121], this concept is also known as Metcalfe’s law [122]. If there are network externalities on Facebook, the activity ratio of countries should scale superlinearly with the Facebook penetration in each country. This scaling relationship with Facebook penetration might vary for the activity ratios of different genders, which would signal an additional marginal benefit of using Facebook for one gender.

Fig. 5.4 shows the scaling relationship per gender between the activity ratio and the total Facebook penetration in each country. Lines show the result of a power-law fit between both variables with an intercept and an interaction term for gender. The estimate of the scaling exponent for each gender is clearly above one for both genders, revealing a superlinear trend consistent with network externalities in Facebook. This exponent is significantly stronger for female users ($\alpha_F = 1.45$ CI=[1.41, 1.49]) than for male users ($\alpha_M = 1.20$ CI=[1.16, 1.24]), suggesting that the network externalities in Facebook are stronger for women than for men.

5.5.4. Relation of FGD with Economic Gender Inequality Changes

Given the network externalities shown above, could the FGD be related to changes in economic gender inequality? We test this possibility by analyzing the change in FGD and Economic gender equality between 2015 and 2016. We fitted two regression models, one of changes of Economic gender equality as a function of FGD ($FGD_{2015} \rightarrow \Delta Eco_{2016}$), and the converse one ($Eco_{2015} \rightarrow \Delta FGD_{2016}$), including controls for autocorrelation and GDP as explained in the Methods section. The coefficient estimates, shown in Fig. 5.5, reveal a significant positive relationship between the FGD rank and changes in economic gender inequality, but not vice versa: there is no significant relationship between Economic gender equality and the changes in FGD.

The partial $R^2$ value of $FGD_{2015}$ in the first model is much higher than the equivalent of $Eco_{2015}$ in the second model (median bootstrap values of 0.027 and 0.002 respectively), as shown in the second column of Fig. 5.5. This suggests the existence of an association between FGD and changes in Economic gender equality such that countries with a low value of FGD (i.e. high rank
5.6 Discussion

By leveraging the Facebook advertising platform to gather data about its total users and DAU, we were able to characterize and quantify world-wide gender inequality on the social media. To measure this inequality we instrumented a methodology which we call the Facebook Gender Divide. By quantifying the Facebook Gender Divide among 1.4 Billion Facebook users, we demonstrate a number of phenomena that deserve further investigation. First, we show that the FGD is heterogeneous across countries and is associated with other types of gender inequality, including economic, health, and education inequality. While the mechanisms behind this connection and its generalizability to other social media remain an open question, this work is an example of how publicly accessible social media data can be used to understand such important social phenomenon. Second, we found evidence of gender-dependent network externalities—that is, women might receive higher marginal benefit than men from the general adoption of Facebook in a country. While we only observe traces of this phenomenon at an aggregate level, in the differences between activity rates across countries, these results point towards a new research direction: using observational data to understand the value of social networking sites across demographic attributes. Third, the FGD methodology provides an inexpensive and accessible way to compute gender inequalities in social media that can be tracked over time and across the vast majority of countries. This allowed us to identify a relationship between the FGD in 2015 and changes in.

Figure 5.4: Gender differences in network externalities on Facebook. Scaling of Facebook activity ratio per gender versus total Facebook penetration. Solid lines show fit results and shaded areas show their 95% confidence intervals. Both male and female activity ratios grow superlinearly with Facebook penetration ($\alpha > 1$), indicating positive network externalities. These network externalities are stronger for female than for male users ($\alpha_F > \alpha_M$).

number) tend more on average to approach economic gender equality.
economic gender inequality in 2016. This relationship could be produced by three mechanisms: i) a causation path between the FGD and changes in economic gender equality, ii) a more complex causation from economic gender equality on changes in FGD, or iii) by the prevalence of a third factor of cultural gender norms that drive both the FGD and economic gender equality. While we find evidence for the first explanation, we must note that the real interplay between the FGD and economic gender equality is probably a combination of all three mechanisms, and only future research with more detailed data can answer how.

Our results show trends across a wide range of countries, but caution should be taken when extrapolating to the future or when predicating about individual countries. Before doing so, we need longitudinal models of changes in development factors in a wide range of countries, to find the role of the FGD in broader development sequences that include economic development, health, education, and inequality [123] before formulating policy suggestions. Nevertheless, our results allow us to speculate that social media can be an equalizing force that counteracts other barriers—e.g. those that limit women’s mobility [124]—by providing access to greater economic
opportunities and social capital. In a similar way as mobile phones increased the life quality of fishermen in India [125], social media might work as a digital provide that helps disfavored groups despite the still generalized inequalities in access to ICT and in adoption of social media technologies.

Our results also show that non-personal data, e.g. anonymous and aggregated data produced by billions of Facebook users can be used for social good, in particular to understand the issue of gender inequalities in society at large. This approach of using such large-scale aggregate data of high granularity also motivates research approaches that are less prone to privacy concerns as evidenced by Cambridge Analytica’s use of Facebook user data for political campaigns [126] or other privacy concerns such as possible construction of shadow profiles of non-users [127, 128].

5.7. Conclusion

In this chapter we reported our study on the applicability of leveraging the Facebook online advertising platform to measure gender inequality in access to social media. The results of analysis using a set of measurement tools and methodologies demonstrate the existence of world-wide heterogeneity in having access to the social media. Further investigation into the observed heterogeneity using inequality indicator datasets indicates that the FGD is associated with existing inequalities particularly gender inequality indices in education, health, and economic opportunity. In addition, gender based analysis of network externalities indicate a higher positive effect on females. This result, coupled with relation of FGD with above inequality indices, suggests that access to social media has an added value for women. Given the huge influence and the transformative power the social media offers, the existence of inequality in having access to and the usage of such information resources is a huge obstacle to development and transformation of a society, especially to those on the vulnerable side of the inequality. In conclusion, these results suggest that closing the digital divide is critical to social and economic development.
6

User Representation and Growth on the Largest Social Media

6.1. Introduction

Having access to social media platforms empowers people with direct access to social capital on the Internet. These resources offer people new opportunities to engage in societal, governmental, and business interactions on these platforms. Due to this fact, having the knowledge about user representation, the growth, and their engagement in such online services is critical for various stakeholders including policy makers, researchers, digital equality and development organizations. Besides, understanding the user-base of an online service and its user engagement is also an important factor for various entities such as investors, marketers the operators. Therefore, a large user-base is a key factor for the financial success of these online services [129][131]. Being aware of this, the now leading companies such as Facebook, Google, and Twitter, had avoided generating revenues to solely focus on increasing the user-base in their initial days. Since the main business drivers behind most of these services are marketing and advertising based on the rich data they collect about their users, online services must also aim at having the largest possible number of users engaging with the service on a daily basis. Typical metrics used by the industry to assess such level of engagement include the number of daily active users and the average daily active time users spend in the service. The temporal evolution of these metrics needs to be monitored to properly assess the health of an online service. This would help identify growth potentials or to anticipate the loss of users and/or the reduction of engagement so that investors
and customers can make informed decisions with respect to the service. In addition, the online service operator needs to go one step further by identifying the reasons behind the observed trend in the evolution of user-base and user engagement. This would allow the service operator to reinforce the practices leading to positive trends and take corrective measures on negative trends before its users, investors, and customers lose their interest in the service. From social benefit perspective understanding the growth evolution of these metrics is also a practical measure to answer important questions such as: how the society’s access to information is distributed, how user demographics are represented in online communities, how people’s access to digital information platforms is growing over time.

6.2. Related Work

The research community has understood the importance of this issue. We find several works analyzing the evolution of the user-base in MySpace, Twitter or Google+ [132–134] while just few works are able to analyze the evolution in user engagement [133,135]. These works use, in general, crawling techniques to collect a sample of user profiles that help them estimate the overall user-base as well as its evolution. While these works have been enlightening and of high value, it is hard to assess the accuracy of their estimations due to the lack of ground truth data, which is proprietary data owned by the online service operator. Moreover, these techniques depend on the “friendliness” of the online service to be crawled. In particular, Facebook, the service we consider in this research, presents enormous barriers for large-scale crawling techniques. Due to this, previous research attempts to investigate Facebook growth or discontinued use is limited to using a very small sample of users (usually confined to a small demographic or geographical area) [136,138]. Cannarella et. al [139] used a different approach by combining search queries and contagion epidemic models to study the growth evolution of Facebook’s global user-base. This work concluded that the OSN will see a rapid decline and eventually die out. In fact, it wrongly predicted a reduction of up to 80% of Facebook’s user-base between 2015 and 2017. Considering Facebook as an epidemic, this model is based on the following three possible states of users: Susceptible, Infected, Recovered. Other relevant theories include the network effect in technology adoption which states a network becomes more valuable to its users as number of users in a network increase [140], and diffusion of innovation theory which classifies adopters into innovators, early adopters, early majority, late majority, and laggards (very conservative and techno phobic groups) based on their degree of innovativeness and their willingness to adopt new technology involving five stages of persuasion offering distinct strategies to appeal to each category [141]. Rogers further argues that for an innovation to self-sustain it has to be widely adopted and reach a critical mass [141]. Even though these models have their own strengths and weaknesses in their applicability to OSNs [140,142], discussing how Facebook fits these models is out of the scope of this chapter.

While we can find research works addressing the composition and growth of users in online
6.3 Key Insights

<table>
<thead>
<tr>
<th></th>
<th>July 2015</th>
<th>June 2018</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>DAU</td>
<td>TOT</td>
</tr>
<tr>
<td>Female</td>
<td>63764454</td>
<td>656408750</td>
</tr>
<tr>
<td>Male</td>
<td>377748364</td>
<td>799575020</td>
</tr>
<tr>
<td>Users in US</td>
<td>124634859</td>
<td>190430000</td>
</tr>
<tr>
<td>Users in India</td>
<td>46390720</td>
<td>131810000</td>
</tr>
</tbody>
</table>

Table 6.1: Counts of median Facebook total users (TOT) and daily active user (DAU) estimates, in July 2015 and June 2018 showing results for both genders, and results for US and India (countries having the largest user counts).

services such as Twitter, MySpace or Google+, such detailed analysis is missing for Facebook, which is currently the largest online social network. Facebook, often synonymous to the Internet in many parts of the world [143], has proved to be one of the main Internet services that are fundamentally changing the way people use the Internet globally having tremendous impact in business, political discourse, and day to day societal interactions. In fact, Facebook penetration is found to show a close approximation to Internet penetration [65].

In this chapter we analyze, for first time, the demographic and geographic composition and growth of Facebook’s user-base and user engagement by taking datasets spanning three years and across different countries and demographic groups. We also characterize the observed growth across countries using external datasets from sources such as World Bank and CIA world factbook showing that observed heterogeneity can be explained by countries’ existing socioeconomic and demographic standing. To this end, we define a methodology which leverages the Facebook marketing API offered to advertisers to configure their ad campaigns. This API allows to define queries including a specific target population based on demographic (age and gender) and geographical parameters (e.g. country or region). The API returns (among other information) the number of total users (user-base) and daily active users (user engagement) that match the parameters defined in the query. This methodology overcomes the limitations of previous approaches since it is scalable and directly obtains the actual data provided by the service operator itself. Moreover, the ability to measure the user-base and user engagement slicing per country and age groups allows us to analyze potential factors affecting the evolution of these important variables. Even though leveraging this API has been shown to be an effective alternative source of statistical information about users of the OSN [67,114,144,145], re-purposing this data source should be taken with caution, such as doing repeated measurements, and taking median value, comparing with alternative sources, when available. Section 6.5 presents details and recommendations.

6.3. Key Insights

Our analysis of Facebook’s user-base and user engagement evolution over a period of three years and across 230 countries and age groups ranging between 13 and 65+ years has served to explain some fundamental aspects of Facebook growth in the last 3 years:
Our Analysis indicates that Facebook is still growing (unlike previous studies that concluded otherwise) but at a very slow rate, where only half of its users are active on a daily basis.

The evolution of user-base as well as user engagement is heterogeneous across age, gender and location. For instance, our analysis explains that the growth rate among adolescents is lower than other age groups (around 2.3 times smaller than the case of adults), while the growth rate of women is lower than for men (around 1.26 times smaller in user engagement growth and 1.1 times smaller in user-base growth). Moreover, Facebook shows a low to moderate growth in most analyzed countries. In particular, developed countries show a plateau in the growth trend, whereas the most important growth takes place in Africa and Central Asia.

We have developed an explanatory model that considers socioeconomic factors related to Facebook growth across countries. This model shows that Facebook growth potential (measured by Facebook penetration) does not directly imply Facebook growth. In particular, (i) the user-base grows faster in areas having high urbanization rate, presenting a higher employment rate, and less infrastructure to access the Internet and thus Facebook (measured through Internet broadband access per 100 users), having higher gender inequality, and where the OSN is not among the top services in the country. (ii) the number of daily users presents a higher increase in countries showing a faster population growth, a decreasing unemployment rate, and where Facebook is not the most popular Internet service. The first three characteristics are representative to a high extent of emerging and pre-emerging countries with population resembling stage two of population pyramid [146, 147].

The application of a random forest regression analysis on the three considered growth variability factors (i.e. geographic, age group, and gender) explains that geography is the variable influencing most of Facebook’s growth in both user-base and engagement, fol-
6.4 Data Collection to Measure Facebook User-base and User Engagement Evolution

The Facebook dataset was collected directly from the OSN leveraging its marketing graph API. Intended for its customers, Facebook offers this feature-rich API [148] to enable them to reach a target audience defined by a range of demographic and behavioral targeting parameters. When queried with these pre-configured parameters, the API end point returns a JSON response which includes, among others, the number of total users, and the number of daily active users satisfying the targeted parameters. We can query this API using geographic (i.e., countries or regions) and demographic (age and gender) parameters and without considering behavioral targeting parameters. Recently the research community has leveraged this API as an alternative way to extract actual datasets from the OSN [83, 84, 87, 114, 149–153]. These works address different research questions than the topic of this chapter. Using this API we developed a distributed measurement system able to monitor the actual number of total users (user-base) and daily active users (user engagement) across ages and genders in all supported countries. The system is composed of a master program that handles authentication and load-balancing the data collection task among individual agent programs. Each agent contacts the Facebook API and starts querying the API, iterating over the list of targeting queries assigned to it by the master. Agents store the obtained responses in a central repository. Note that it is crucial to make sure that individual queries being parallelized do not contain overlapping target audiences. To preserve uniqueness of targeting

Figure 6.2: Global map of Facebook Potential growth: countries are colored based on their available room for Facebook growth ranging from dark blue (highest penetration hence less opportunity to grow) to light green (highest potential). Countries with no data are colored red.
queries, before assigning it to agents, the master recursively partitions our targeting parameters first based on countries, then based on gender (male vs female), finally each resulting query is partitioned into 53 groups based on age (from age = 13, the minimum age supported, through 65). We would like to note that age 65, which is the maximum possible targeting age, is actually an age group that includes users with age 65 and above, i.e. 65+. Moreover, Facebook offers three location types ("recent location", "home location", and "travel in") to target specific users. From these available options we have used the home location of users to determine their geographical location since it is a more reliable way to identify "more permanent" location of users that use both mobile and desktop devices. As stated in their API documentation [148], Facebook uses a combination of techniques to reliably identify the "home location" of a user. These techniques include information based on IP address, "current city" in user’s profile and from their friends’ stated profile locations. In section 6.5 we include a quantitative comparison analysis between home and recent location options that proves the former is a better choice. Using the described measurement system we have obtained snapshots of the total number of users and daily active users for each age and gender group, in 230 available countries, extended over three years since July 2015 and collected in two periods between July 2015 to February 2017, and between October 2017 and June 2018. Every month in this dataset has a complete snapshot of at least 4 daily samples with a median of 19 days in a month. Figure 6.1 shows the number of days per month used in our analysis as we will see in Figure 6.8.

Figure 6.3: Global map of Facebook engagement capacity: countries are colored based on their observed engagement ranging from dark blue (least engagement among total users implying least engagement capacity) to light green (highest observed engagement). Countries with no data are colored red.
6.5 Recommendations on using the API for research purposes

The Facebook marketing API provides a rich set of targeting options to help estimate the number of target Facebook populations; however, it is worth to remind that the primary purpose of this API is marketing, not exactly a well-documented research tool as such. Due to this reason, data collection and analysis should be taken with caution. We recommend the following guidelines in data collection.

- State the level of targeting parameters along with results. In our measurements the targeting parameters used were gender, location, and age, without any behavioral or interest based targeting. We found that user targeting at a country level without any detail targeting specifications does not yield exact number as partitioning by age and country and summing the results. To perform this analysis, we queried the Facebook API with (age, gender, country) as targeting parameters for 230 countries and we collected total users and daily active users for each query. Later, the API was queried with only country as a targeting parameter not including age and gender parameters. Our results show that even though the results are very similar with very high correlation (Pearson = 0.9992, Spearman = 0.9998) between daily active users and (Pearson = 0.9974, and Spearman = 0.9996) between total users per country, the results are not exactly the same. Figure 6.5 shows these relations. As a final note, due to lack of detailed documentation and API’s primary purpose being a marketing tool, the data from this API should not be taken as a ground truth data. Its use as a viable large scale target population estimation tool should be taken with caution.

- State the location type used for the research purpose. Facebook offers three locations types: “recent”, “home”, and “travel in”. While “travel in” and ‘recent’ are used to target users based on their recent location, the “home” option offers a “more permanent” location based targeting option. It is also important not to confuse “home town” with “home location” as the later is a more general term to refer to user’s permanent location as can be stated in current city of living. Figure 6.4 shows the difference between "home” and ”recent” location types. The right panel of the figure shows comparison of global users measured using the two targeting options with Facebook’s June 2018 official quarterly report. These results provide solid evidence that “home location” is a better choice to measure user-base and user-engagement values.

- Compare with external datasets. Finding such a large scale evolution dataset is difficult, the closest we found was the global number of users from Facebook quarterly report. We compared our results with Facebook quarterly report showing that both datasets present similar growing trend and but not offering exact numbers.
Figure 6.4: Difference between “home” and “recent” location types and comparison with Facebook quarterly report. The left panel shows comparison between daily active users, the center panel shows comparison between total users. The right panel shows comparison of the two targeting results with Facebook global users as reported in June 2018.

Figure 6.5: Correlation between Facebook number of users at country level using country name as a targeting option vs partitioned demographic targeting with (country, age, sex). The left panel shows the correlation between daily active users, the right panel shows correlation between total users.

Comparison with Facebook quarterly report

We compared the number of daily active users and total users which was used to measure the Facebook growth that we collected through the marketing API with Facebook’s quarterly report about its global total active users and daily active users (reported as median MAU and median DAU of the last month of each quarter, respectively). We gathered these quarterly numbers from the quarterly dissemination report documents on its investor relations page\footnote{https://investor.fb.com/}. As shown on figure 6.6 the trends from both sources are qualitatively similar; moreover, both trends have very high correlation (> 0.99). In addition the figures also show that the number of total users and daily active users has been steadily increasing in the last three years.
6.6 Analysis Methodology

Figure 6.6: Comparison of the dataset collected through the marketing API with Facebook quarterly report. The left panel shows comparison between daily active users, the right panel shows comparison between total users. In both figures the median value of each month is compared with the Facebook quarterly report.

6.6. Analysis Methodology

Our goal is to measure the evolution of the user-base and user engagement in Facebook over the considered period of time. The number of total users obtained from our measurement is represented by the monthly active users variable reported by Facebook, we define this variable as the user-base. Moreover, the number of daily active users provides an aggregate measure of the extensively used user engagement. Therefore, we will consider these two variables for our analysis in the rest of this chapter. To compute the evolution of the proposed metrics over time, we use a reference value of the metric, $x_0$, which represents the median number of Facebook total users (or daily active users) of a target population on the first month from our dataset. Then we compute the growth rate on a given month $i$, $(U_i)$, as the ratio of median number of added total users (daily active users) in that month and the reference value as:

$$U_i = \frac{x_i - x_0}{x_0} \times 100$$

The above $U_i$ values were calculated for both total users ($U_{tot}$) and daily active users ($U_{dau}$) of the target population. Note that the target population under study could be Facebook users of certain demographic or geographic group such as users per country, age group or gender. To derive the growth trends of the user-base and user engagement defined metrics, we analyze the temporal series of $U_{tot}$ and $U_{dau}$ as follows. First, a monthly temporal series of $U_{tot}$ and $U_{dau}$ is calculated based on the median values of each month. At the end of this step we obtain the evolution trend on monthly scale. Using this result we apply a regression model on the monthly

---

2In real world measurements median is more robust to outliers than mean. [154].
growth values. In the model the growth value $U_i$ is set as dependent variable on month number $i$ which is calculated as the number of months since June 2015. At the end of this step we obtain model coefficients, which represent the gradient of the regression line, which in turn indicates the rate of monthly growth of the metric as compared to the initial value $x_0$. For each metric, $U_i$ indicates by how much the given value ($U_{tot}$ or $U_{dau}$) has changed as compared to $x_0$, for example a $U_{tot} = 2.0$ tells that the number of total users has shown a 2% monthly increase as compared to $x_0$. An additional advantage of using a normalized metric $U$ is that it allows a head-to-head comparison between targeted groups (e.g. comparison across demographic groups or countries).

6.7. Results

Using the above described methodologies and datasets we analyze the evolution of Facebook user-base and user engagement between July 2015 and June 2018. For reference Table 6.1 shows the median total users and DAU for first month and last months in our dataset. First we analyze a snapshot of our dataset in June 2018 to get a sense of Facebook’s health status and potential growth capacity across countries and demographic groups. Then we analyze the evolution of the user-base and user engagement for the aggregated OSN as well as for different countries and age groups.

6.7.1. Facebook’s Growth Capacity

According to our measurement, Facebook has 2.2 billion total users, and 1.3 billion daily active users as of June 2018, making it the most populous online community in the world. As a reference, the world population as of June 2018 is 7.6 billion with China, India, and the US being the most populous countries with 1.39, 1.33, and 0.32 billion inhabitants, respectively. Figure 6.2 shows the growth capacity of Facebook in each country as of June 2018. Growth capacity measures the fraction of the population in a given country which is still not on Facebook and is computed as $1 - FB_P$, where $FB_P$ is Facebook penetration in each country. Note that the Facebook penetration is computed as the ratio between the user-base of Facebook as of June 2018 and the country population as reported by US census bureau (www.census.gov). Note that our methodology is subject to errors due to measurements errors by FB in computation of their user-base or DAU as well as errors in the computation of each country’s population census over 13. Moreover, Figure 6.3 shows the ratio between number of daily active users and total users in each country as of June 2018. This metric captures engagement capacity of Facebook in different countries. The results suggest that the OSN’s growth potential lies in Asia and Africa. However, the currently existing engagement capacity in these geographical areas indicates that even though Facebook has a large room to grow in these regions, the engagement there may not reach the level of other areas with a larger penetration such as North America, South America or Europe. Note that Africa shows a higher engagement capacity than China due to the fact that Facebook is
officially blocked in China, and relative the unavailability of other strongly competing services in Africa (e.g. WeChat of China has more than 900 million users).

We now analyze the demographic population pyramid for the same dataset snapshot which is shown on Figure 6.7. Literature on demographic studies [147] classifies demographic pyramid of populations into three types: expansive pyramid, stationary pyramid, and constrictive pyramid. It is worth noting that Facebook’s population pyramid does not exactly fit any of these three types. If we consider the number of births in Facebook as the population size at age 13 (minimum age to join the OSN) it can be characterized by a very low birth rate, and an increasing “immigration” (joining the site at a later age) up to the age of 19. If we take the pyramid above the age of 20, it resembles an expansive population. In general, the smaller percentages of people in the younger age cohorts make the pyramid more similar to a Constrictive population (a pyramid constricted at the bottom with a lower percentage of younger people). A strict interpretation of this type of pyramid would suggest that the long-term survival of the social network may be questionable since the OSN is becoming less appealing to the younger generation, which will not use the service. An alternative explanation of the lower presence of young population at the bottom of the pyramid could be that the social network lacks features tailored to a younger population, and people join the site as they become older and find the features of the service more appealing. The stability of the pyramid shape, which shows a similar shape over the two analyzed years, suggests that the latter explanation is more plausible. If we consider the group G formed by the total users of age X in July 2015 and thus ages X + 1 in July 2016, X + 2 in July 2017, and X + 3 in June 2018 for the first hypothesis to hold, the size of this group should (at most) remain the same. Instead, we observe a growth of such groups for ages ranging between 13 and 18, suggesting that as teenagers
become older they find Facebook more interesting. Another interesting observation is that the pyramid does not follow a smooth trend at some age, typically showing a large spike on age 25, further investigation shows that this age group is related with the default year of birth (1993) put at Facebook registration page (Figure 6.9), which led us assume that many users just proceeded without modifying the default birth date set by the platform. Looking at the pyramid also indicates a male-biased disparity between genders in the platform as was recently revealed [114].

6.7.2. Facebook Growth Evolution

6.7.2.1. Overall Growth:

Let us start by considering the evolution of number of total users and daily active users over the analyzed period. Applying the previously described methodology, we conclude that Facebook is growing, contrary to the prediction of previous studies [139]. During our observation time, Facebook has grown from 1.45 billion users with 746 million daily active users in July 2015, to 2.3 billion users with 1.3 billion daily active users in June 2018. Table 6.1 shows comparative statistics between these two months. These are doubtless impressive numbers. However, our analysis reveals that the growth rate in number of total users and active users is still male-biased indicating that gender inequality as seen through this social media is relentlessly increasing.

6.7.2.2. Growth Across Age Groups:

Based on Erikson’s stages of psycho-social development [155] we classify the age range into the following groups: adolescence (ages 13-19), early adulthood (ages 20-39), adulthood (ages 40-64), and maturity (ages 65+). Using this classification we analyze and compare the evolution of the Facebook user-base and user engagement in each stage. The computed metric values on
6.7 Results

Figure 6.9: The default setting of Facebook registration page showing 1993 as a default year of birth.

Each age group are presented on Figure 6.10. The overall observed trend is shown on Figure 6.8. The results show that Facebook growth rate is heterogeneous across age groups, with user-base growth of 0.98%, 1.78%, 2.30%, and 2.46% for adolescent, early adolescent, adulthood and maturity age groups, respectively. Having user engagement growth of 1.43%, 2.44%, 3.03%, and 3.27% for adolescent, early adolescent, adulthood and maturity age groups, respectively. This shows the growth among adolescents is significantly smaller, by a factor of (at least) two, than for adult groups. This result reinforces our conclusion from the demographic pyramid analysis, showing that Facebook seems to be less appealing for adolescents, that become more interested in the OSN as they become adults. Note that this might have implications from a marketing perspective since Facebook might not be the most appropriate venue to engage with adolescents. Given that Facebook is an OSN first popularized by US college students which could be considered as its innovators and early majority (according to Rogers technology adoption model [141]), Facebook has already gained momentum after astonishing growth over the past decade. We can observe that maturity age groups, the seemingly late majority and laggards, are joining the OSN at a high rate.
6.7.2.3. Growth across gender groups:

Here we discuss growth evolution based on gender. As a result for gender-wise comparison we study user-base and user engagement growth for men vs. women. The computed metric values for both genders are presented on Figure 6.11. As shown in the figure, user engagement growth in both genders shows a higher growth rate than growth in user-base, suggesting that engagement in existing users grows as new users join the system. However, the growth in males is higher in both user-base (by a factor of 1.1) and user engagement (by a factor of 1.26) growth. Note that these results demonstrate that the Facebook gender inequality is only increasing. Gender inequality in Facebook is thoroughly discussed in [114], this chapter showed that gender inequality is associated with socioeconomic inequalities where higher gender inequality is found in countries around Africa and southwest Asia.

![Figure 6.11: Growth rate of Facebook’s user-base (growth in total users) and user engagement (growth in daily active users) for male and female users on Facebook.](image)

6.7.2.4. Growth across countries:

Here we use our described methodology to analyze the growth in the number of total users and daily active users across the 230 countries over the considered period of three years. Figure 6.12 shows the results.

A first look at the results reveals an overall higher growth in engagement than in total users, what can be interpreted as a sign of health suggesting Facebook’s potential as a business platform (e.g., for marketing or advertising) is viable. If we consider individual countries, we observe that Facebook has slower growth and almost plateaued in most developed regions (US, Canada, EU, Scandinavian countries and Australia) whereas it is experiencing its most significant growth in Africa and Central Asia. This suggests that from social and business perspectives Facebook has reached an (almost) stable status in developed countries, where it has established itself as a de-facto social platform that connects a considerable fraction of the population. Instead, Facebook is
6.7 Results

Figure 6.12: Growth rate of number of total users (top) and number of daily active users (down). The range moves from lowest growth rate (yellow; not visible) to highest growth rate (light green).

(a) Growth in user-base

(b) Growth in user engagement
6.7.2.5. Measuring Factors Influencing shift in Usage/Engagement

As we have seen above Facebook growth variability is demonstrated across geographic and demographic groups. Next, we collectively take age group, gender, and geographic dimensions and determine the influence of each dimension in growth variability. To measure influences we first calculate growth metric values for each group sliced per gender, country, and age group for both user engagement and user-base growth. We then apply random forest regression to identify the importance of each dimension in determining growth metric [154]. This methodology has been widely used in various disciplines [156][157]. Our results show that in determining user-base growth, geographic dimension is the most influential variable following age group and gender with values 90.3%, 7.2%, and 2.5%, respectively. User engagement growth variability is also highly influenced by geographical growth followed by age group, and gender with values 71.6%, 22%, and 6.4%, respectively. The pie charts on Figure 6.13 show these values calculated using feature importance measure of random forest regression (with r-squared scores 0.97 and 0.94, respectively). The left panel shows the influences in user base growth metric variability while the right panel shows influences in user engagement growth metric variability.

![Pie charts showing influences in user base and user engagement growth variability](image)

Figure 6.13: Comparison of geographic, gender, and age group dimensions in determining user base and user engagement growth variability calculated using random forest regression method. For user base growth geographic dimension is highly influential covering 90.3% of variability followed by age group and gender with values 7.2%, and 2.5%, respectively. Geographic dimension is also highly influential in user engagement growth variability with 71.6% followed by age group, and gender based variability with values 71.6%, 22%, and 6.4%, respectively.
6.7 Results

6.7.3. External Datasets to Characterize the Observed Growth Variability:

The previous visual interpretation is valuable to gather a first intuition. However, it does not reveal the underlying relations behind the observed heterogeneity of Facebook growth across countries. Our hypothesis is that this heterogeneity may be a reflection of the socioeconomic and demographic composition of countries. To explore the validity of this hypothesis, we leverage regression analysis techniques to measure the correlation between representative socioeconomic metrics and the growth of Facebook for different countries. To conduct our analysis we used the following datasets:

(1) **Demographic Metrics**: We leverage the population distribution dataset from census.gov\(^3\). This dataset includes the number of inhabitants of each sex and age country-wise. Using this dataset we aim to investigate the relation between Facebook’s growth and its penetration in each country. We define Facebook penetration as the ratio between number of Facebook total users and the total population of a country. We also used this data source to get the population size and population density per sq. km. in each country. Moreover, datasets on urbanization rate and annual population growth were collected from The World Bank\(^4\) to answer the following questions: “Does the sparsity of the population affect Facebook’s growth?”, “does rise in urbanization lead to a higher Facebook growth?, presuming a higher ICT infrastructure availability in urban areas”. To measure relations with gender inequality we introduced ratios of male Facebook penetration and female Facebook penetration as gender-divide in a country.

(2) **Birth, Death, and Immigration Metrics**: We leverage datasets from CIA’s world fact book\(^5\) on birth and death rates per 1000 inhabitants of each country to investigate their relation with Facebook’s growth. Moreover, we collect the net migration rate from the same dataset. One of the factors that contribute to population change in a country is migration. Even though the immediate impact of migration is insignificant to the overall population of a country, we can argue that migration is a net effect of seeking better social, economic and political opportunities not available in one’s own country. Using this dataset we also aim to cover these unobserved factors.

(3) **Economic Metrics**: To account for economic metrics we leverage the GDP (gross domestic product) and GNI (gross national income) per capita which are primary indicators of a country’s economic standing. We leverage the World Bank Report on GNI and GDP growth rate of world countries as of 2016\(^6\). The premise of using these datasets is to explore whether countries having a higher Facebook growth is correlated with economic prosperity of inhabitants.

(4) **Availability and Accessibility to Internet**: To explore the existence of correlations with the status of Internet in a country, we consider three signals collected from the World Bank: Internet affordability, Internet availability (Internet users per 100 inhabitants), and Quality of Internet

---

\(^3\)www.census.gov
\(^4\)data.worldbank.org
\(^5\)cia.gov/library/publications/the-world-factbook/
\(^6\)https://data.worldbank.org/indicator/
(Fixed broadband Internet subscriptions per 100 inhabitants). Finally, we collected Alexa\(^7\) ranking of Facebook in each country. Alexa Internet Rank is a relative measure of the popularity of an Internet service in a country as compared to other websites. Globally, Facebook ranks third right after Google and YouTube. Using this ranking, we aim to identify if Facebook is growing in countries where it is already popular or if it is expanding to new places where it is not.

### 6.7.3.1. Characterizing Facebook Evolution:

Here our goal is to model the multi-linear dependence relationship between the Facebook growth metrics of total users and daily active users per country computed using methodology in Section III and the explanatory variables described above. As discussed in Section III, growth metrics per country are calculated using the evolution dataset between July 2015 and June 2018. To this end, we use a linear regression model where multi-collinearity between variables is filtered using the variance inflation factor \([154]\) (Table 6.3 reports the VIF values, and Figure 6.14 shows the correlation matrix between considered variables and growth metrics). The linear relationship models below show the dependence relation between evolution metrics and linear combination of most significant independent variables for each metric. Table 6.2 shows the coefficients of resulting models.

\[
\text{User base growth} \sim \text{broadband penetration} + \text{urbanization growth} + \text{Unemployment rate} + \text{FB alexa rank},
\]
\[
( \text{with p-value } < 3.91\text{e-13, Adjusted R-squared } = 0.63 \text{ and R-squared } = 0.64 )
\]

\[
\text{User engagement growth} \sim \text{birthrate} + \text{gender divide} + \text{Unemployment rate} + \text{FB alexa rank},
\]
\[
( \text{with p-value } < 2.83\text{e-17, adjusted R-squared } = 0.61 \text{ and R-squared } = 0.62 )
\]

<table>
<thead>
<tr>
<th>Explanator</th>
<th>User base</th>
<th>User engagement</th>
</tr>
</thead>
<tbody>
<tr>
<td>broadband penetration</td>
<td>-0.33 ***</td>
<td>—</td>
</tr>
<tr>
<td>birthrate</td>
<td>—</td>
<td>0.38 ***</td>
</tr>
<tr>
<td>urbanization growth</td>
<td>0.18 *</td>
<td>—</td>
</tr>
<tr>
<td>Unemployment rate</td>
<td>-0.24 ***</td>
<td>-0.24 ***</td>
</tr>
<tr>
<td>gender divide</td>
<td>—</td>
<td>0.24 ***</td>
</tr>
<tr>
<td>FB alexa rank</td>
<td>3.81 ***</td>
<td>3.35 ***</td>
</tr>
</tbody>
</table>

Table 6.2: Regression coefficients
(Signif. codes: 0 ’***’, 0.005 ’**’, 0.05 ’*’)

As it can be seen from the coefficients of determination, the relations between Facebook’s growth metrics and the socioeconomic factors are relevant (adjusted R-square > 0.6). Therefore, we can safely reject the null hypothesis that there is no relation between Facebook’s growth metrics and socioeconomic factors.

\(^7\)www.alexa.com
Figure 6.14: Heat map showing the correlation between each pair of external socioeconomic indicators, the FB user-base (1st column) and FB user engagement (2nd column). The correlation ranges from blue (negatively correlated) to red (positive correlation).

First, the model suggests that Facebook has a higher user-base growth in areas with the following characteristics: where Facebook is not among the most popular Internet service, with higher urbanization growth, decreased unemployment rate, but with less Internet penetration (measured through broadband subscriptions per 100 people). Facebook seems to have noticed this potential; its recent “basic service” and the internet.org initiative efforts to expand Internet coverage to developing countries and areas with scarce connectivity seem to justify that [158].

On the other hand, user engagement growth shows a higher increase in countries characterized by: higher birthrate, decreased unemployment rate, but with higher gender inequality, and where
the OSN is not in the top list of popular sites. The collective characteristics of countries showing
the first three indicators represent a high extend of emerging or pre-emerging country with a stage
two expansive pyramid [146,147].

Even though correlation does not imply causation, and it is possible that other factors are
responsible for the observed correlations, the examined attributes and associated results can be
considered as an insight for further study into the observed phenomenon instead of as a prediction
model of OSN growth in countries.

6.8. Conclusion

This chapter presents a novel methodology to measure user representation on Facebook and
growth of its user-base and user engagement with detailed demographic (per gender and ages
ranging between 13 and 65+) and geographic granularity of 230 countries. The obtained results
with this methodology are available at: http://track.netcom.it.uc3m.es/fb_viz.

This information is of high interest for digital inequality researchers, policy makers, Face-
book’s customers (mainly advertisers), social media analysts, and investors, which will enable to
assess the composition and evolution of Facebook with such level of detail. Based on our ex-
perience, we have also included recommendations in using the API for research purposes. At
the same time the collected data available through the mentioned system has important value for
researchers in multiple disciplines (computational social science, sociology, politics, etc). In par-
ticular, in this chapter we present an initial analysis of the collected data over a period of three
years, between July 2015 and June 2018, that report interesting results: Overall Facebook shows a
growing trend. However, both Facebook user-base and engagement growth show gender bias with
a higher growth rate in males than females (1.26 times smaller in user engagement growth and
1.1 times smaller in user-base growth). The growth pattern presents a clear heterogeneity across
demographic groups and geographic regions. Facebook’s growth rate is twice smaller among ado-
lescents than adults. Facebook is reaching a plateau situation in developed regions of the world,
while presenting the largest growth rate in developing countries mainly concentrated in Africa
and Central Asia.
Part IV

Conclusion and Future work
7

Conclusion and Future Work

Large amounts of digital traces are being generated from users’ interaction in popular Internet services such as social media platforms. These digital traces have a huge potential to advance research efforts in improving different aspects of the society such as closing the digital divide and reducing digital inequalities. However, given that the majority of digital platforms are corporate owned, large scale aggregate data collected by these platforms is not widely available for societal benefit. Hence, developing methodologies to leverage and re-purpose large scale aggregate datasets available in these platforms to understand important societal issues is crucial to the development and advancement of the society.

In this thesis, we develop a set of methodologies to leverage the dataset from large social media platforms that help answer important research questions. First, we have developed a methodology to understand the monetary value that the online advertising system assigns to different user profiles. Using this methodology, we aim to contribute the research efforts to bring transparency on the workings of Internet services specifically online advertising services. Second, we have developed a methodology to measure the digital gender inequality across 217 countries and its relation with existing social inequality indicators. Gender inequality being one of the fundamental societal issues, the results and methodologies developed will contribute towards understanding gender and other forms of inequalities in access to and usage of Internet services and their relation with offline inequalities. Third, we have developed a methodology to measure and understand how Facebook is growing worldwide across age groups, genders, and countries. This has enabled us to understand how access to this social media service is distributed globally and across demographics. The promising results obtained from these research contributions will help advance
gender and inequality studies by offering different perspective and a new set of resources in this area. As future work, the developed tools and methodologies could be extended to answer related societal issues such as to measure socioeconomic gaps and resource allocations, to monitor adoption of Internet policies and regulations, to identify bias inherent in online interaction platforms, big data algorithms, and their societal impact with a special focus on advertising platforms and to understand other forms of inequalities. The methodologies of using the explored platforms could be extended to other related platforms which would enable us to answer similar questions and also to perform a comparative study which will further allow us to develop generalized theories about inequalities in social media, and the Internet in general.
Part V

Appendices
Appendices
**Supplementary Table 1**

<table>
<thead>
<tr>
<th>Gender</th>
<th>Facebook accounts</th>
<th>Median DAU</th>
<th>Total population</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female</td>
<td>646,953,680</td>
<td>375,148,380</td>
<td>2,422,856,560</td>
</tr>
<tr>
<td>Male</td>
<td>800,026,950</td>
<td>391,673,365</td>
<td>2,485,459,305</td>
</tr>
<tr>
<td>Both</td>
<td>1,446,980,630</td>
<td>766,821,745</td>
<td>4,908,315,865</td>
</tr>
</tbody>
</table>

Table 1: Total counts of Facebook accounts, DAU estimates, and total population covered by the dataset.
Supplementary Text 1 - Validating the FGD

We validate the measurement of the FGD against three reference datasets:

1. **Global Web Index (GWI).** We use the survey responses of the Global Web Index panel during the period of our study (the two last quarters of 2015 and the two first quarters of 2016). For this period, the GWI contains responses from 99,338 panelists in 34 countries, providing rescaled estimates of survey responses that generalize to the population as a whole. We take the response to the question about the frequency of use of Facebook, calculating the weighted fraction of respondents for each gender and age segment that report to use Facebook daily or more than once a day. Furthermore, we repeat the analysis for other social media (Whatsapp, LinkedIn, Twitter, Instagram, and YouTube), calculating gender divide values outside Facebook.

2. **Pew Research Center Spring 2016 Global Attitudes (Pew Global).** This dataset includes responses from 23,462 panelists in 19 countries and can be found online. We use the positive answer rate to question 82 “Do you ever use online social networking sites like Facebook, Twitter...?” as way to measure the penetration of SNS in general. We take the respondent weights reported in the dataset to rescale the frequency of positive responses taking into account the self-reported gender of survey respondents.

3. **Pew Research Center Internet & Technology, March 7-April 4, 2016 (Pew US).** This US questionnaire includes questions about Facebook use in particular (act135, “Do you ever use the internet or a mobile app to use Facebook?”), as well as gender and age data from 1,601 respondents. We use the respondent weights to compute Facebook use rates across gender and age groups.

1. [https://www.globalwebindex.net/](https://www.globalwebindex.net/)
Comparison of Facebook penetration estimates

Figure 1: Total Facebook penetration for both genders as reported by the marketing API versus Facebook penetration as estimated in the GWI survey (left) and penetration of all Social Networking Sites in Pew Global survey (right). The red dashed line shows a linear regression profile, with its prediction standard errors in the shaded area.

The left panel of Fig. 1 shows the relationship between the total Facebook penetration for both genders as measured by us through the marketing API versus the value estimated from the GWI survey. There is a high positive correlation between both measurements of Facebook Penetration (Pearson 0.89, CI [0.78, 0.94], Spearman 0.86, CI [0.72, 0.94]). The right panel of Fig. 1 shows the same evaluation against the Pew Global survey. The correlation between both measurements is positive and high (Pearson: 0.71, CI [0.39, 0.88], Spearman: 0.77, CI [0.43, 0.94]).

The left panel of Fig. 2 shows the comparison between estimates based on GWI and Pew Global survey data. The correlation is also positive and significant, even though samples are of limited size (Pearson: 0.63, CI [0.17, 0.86], Spearman: 0.7, CI [0.38, 0.91]). Both in the right panel of Fig. 2 and in left panel of Fig. 1 it can be seen that China is a clear outlier. This stems from the difference between comparing Facebook penetration versus penetration for Social Networking Sites in general, which is the precise question of the Pew survey. If we focus on the rest of countries, where we can expect a priori that Facebook is more representative of social media
in general, the correlation reaches higher values when comparing to the Facebook marketing API (Pearson: 0.84, CI [0.62, 0.94], Spearman: 0.87, CI [0.66, 0.96]) and to the GWI survey data (Pearson 0.84, CI [0.55, 0.95], Spearman 0.78, CI [0.43, 0.94]). To make a fair comparison, we should not include China when comparing Facebook penetration with penetration in SNS in general.

When we focus only on the set of countries available in all three datasets, we can compare the correlation of each pair of data sources to assess the validity of the Facebook API data in terms of Facebook penetration. In this smaller sample, the Facebook penetration calculated through the marketing API is highly correlated with both the Pew Global survey values (Pearson: 0.87, CI [0.64, 0.96], Spearman: 0.86, CI [0.59, 0.97]) and with the GWI survey (Pearson: 0.89, CI [0.68, 0.96], Spearman: 0.83, CI [0.53, 0.97]). The point estimates of these two values are higher than the correlation between the Pew Global and GWI datasets, as reported above. Nevertheless, these differences are not significant, as confidence intervals overlap and bootstrap sampling shows that estimates are indistinguishable (Fig. 1). From this analysis we conclude that the estimate of
Facebook penetration from the marketing API has comparable quality to the values reported in the high-quality, representative surveys of GWI and the Pew Research Center.

**Facebook gender divide estimates**

![Figure 3: Measurement of the FGD in the API versus estimates using GWI data (left) and a gender divide for all SNS in PEW (right). The red dashed line shows a linear regression profile, with its prediction standard errors in the shaded area.](image)

We calculated surrogates of the FGD based on the GWI and Pew global survey datasets. The left panel of Fig. 3 shows the comparison of the Facebook marketing API estimate with the same estimate using GWI data, revealing high positive correlation (Pearson: 0.83, CI [0.68, 0.91], Spearman: 0.63, CI [0.27, 0.87]). The right panel of Fig. 3 shows the comparison of the Facebook marketing API estimate with Pew survey data for all SNS (including China), also revealing high positive correlation (Pearson: 0.85, CI [0.65, 0.94], Spearman: 0.74, CI [0.35, 0.91]).

A comparison between estimates of the FGD using GWI versus using Pew data is shown on the left panel of Fig. 4 also revealing positive correlations (Pearson: 0.85, CI [0.6, 0.95], Spearman: 0.49, CI [−0.11, 0.86]). As with Facebook Penetration, the correlation between the FGD using the Facebook marketing API and the other two (with Pew: Pearson: 0.77, CI [0.43, 0.92], Spearman: 0.68, CI [0.14, 0.93]; with GWI: Pearson 0.96, CI [0.890.99], Spearman 0.83, CI
Figure 4: Left: Comparison of gender divide measurements in PEW and GWI data. The red dashed line shows a linear regression profile, with its prediction standard errors in the shaded area. Right: Bootstrapping distributions of Spearman’s correlation coefficient between all pairs of gender divide measurements.

[0.47, 0.97]) estimates is comparable to the correlation within estimates, as evidenced in bootstrapping samples reported in the right panel of Fig. 4. We can conclude that the estimate of the FGD using the marketing API is consistent with GWI and Pew survey metrics, opening the study of the FGD to a much larger sample of countries.

We measured the absolute difference between the FGD in the marketing API and in each survey dataset. As expected, the correlation between this absolute difference and the Facebook penetration across countries is negative (with GWI: Pearson: $-0.3$, CI $[-0.59, 0.05]$, p-value = 0.09; with Pew: Pearson: $-0.27$, CI $[-0.65, 0.21]$, p-value = 0.26), but its value is weak and not significant. Nevertheless, we include controls for Facebook penetration in our further analyses, to make sure that our results are not an artifact of a correlation between penetration and measurement error in the marketing API.

Furthermore, the GWI survey allows us to compare measurements of the FGD in other social networks with our measure based on the Facebook marketing API. We get moderate to high Pearson correlation coefficients with other sites, such as Whatsapp (0.67, CI [0.43, 0.82]), LinkedIn (0.65, CI [0.40, 0.81]), Twitter (0.69, CI [0.46, 0.84]), Instagram (0.79, CI [0.62, 0.89]),
and YouTube (0.89, CI [0.79, 0.94]). While we cannot generalize to all social networks based only on Facebook data, we can see that, to some extent, the difference in activity across genders also appears in other SNS. This is particularly interesting when comparing Facebook, a very private social network, with YouTube or Twitter, which are much more public but still display substantial correlations in terms of FGD.

**Comparison across age groups**

![Comparison of FB presence ratio versus PEW gender categories for both genders together (left) and gender-wise (center). Replication of the same validation versus GWI estimates (right).](image)

We compared Facebook penetration estimates in the US across the four age groups reported in the Pew US dataset. The left panel of Fig. 5 shows the comparison for both genders together, which have a Pearson correlation coefficient of 0.96, CI [0.04, 0.99]. The central panel of Fig. 5 shows the same comparison by taking the gender-wise estimates, which also have positive Pearson correlation (0.94, CI [0.72, 0.99]). This also appears when surveying the GWI dataset for US respondents in similar age categories, as shown on the left panel of Fig. 5 which has a high and significant Pearson correlation coefficient of 0.92, CI [0.69, 0.98]. We can conclude that data provided by the Facebook marketing API is consistent across ages, but to be sure that our further analyses are robust we take two action: 1) we add a mean user age control to our regression models, and 2) we stratify our analyses across age categories, using in each stratum a measurement of the FGD in the corresponding age range.
We retrieved data from the Facebook marketing API on a daily frequency, starting our retrieval at 3 AM Central European Time. To validate the consistency of our measurement with any other times of the day, we checked the consistency of our construction of the FGD with hourly values for 24 hours in December 2017.

Fig. 6 shows the hourly measurement for a sample of large countries, revealing high consistency with very small fluctuations. When comparing the measurement of the FGD at 3AM CET with any other time in the same day, we get extremely high pearson correlation coefficients (0.9942654, CI [0.9939277, 0.9945844]), as also evidenced in Fig. 7. This also extends to the measurement of Daily Active Users for male (0.9999431, CI [0.9999397, 0.9999462]) and female (0.9999378, CI [0.9999341, 0.9999412]), as well as the total number of accounts for male users (0.9999926, CI [0.9999921, 0.9999930]) and female users (0.9999968, CI [0.9999966, 0.9999970]). Any fluctuation can be attributed to the rounding that Facebook does to preserve individual user anonymity and to the inter day changes in the number of Daily Active Users.
Figure 7: Facebook API measurements at different hours of the day. The left panel shows a comparison of measurements of the FGD for all countries in the dataset at 3AM CET versus hourly measurements at other times of the day. The right panels show the comparison between the measurement at 3AM CET and at other times of the day for the number of DAU and of present users (TOT) per gender.

**Supplementary Text 2 - FGD as a function of other inequalities**

**Regression diagnostics**

The Variance Inflation Factors of the variables in the FGD model are below 5, allowing us to discard collinearity in the linear model of FGD as a function of other inequalities. Table 2 reports the detailed results of the FGD model fit and Table 3 reports the results of the same model when fitted with a robust regression method. Table 4 shows a fit with HC correction for heteroskedasticity. All results are qualitatively similar, revealing that the FGD model result is robust to outliers and heteroskedasticity.

Fig. 8 shows the normal Q-Q plot and the histogram of residuals, which are distributed very close to normality. This is confirmed by a Shapiro-Wilk normality test, with a statistic of 0.99 and unable to reject the null hypothesis that residuals are normally distributed ($p = 0.63$). Furthermore, residuals are uncorrelated with all gender equality variables (Near-zero Pearson correlation coefficients, with p-values above 0.9) and the square root of absolute residuals are not
<table>
<thead>
<tr>
<th>Term</th>
<th>Median estimate</th>
<th>95% Credible Interval</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>135.8</td>
<td>[119.9, 152.1]</td>
<td>$p &lt; 0.01$</td>
</tr>
<tr>
<td>Education Equality Rank</td>
<td>$-0.54$</td>
<td>[−0.67, −0.41]</td>
<td>$p &lt; 0.01$</td>
</tr>
<tr>
<td>Health Equality Rank</td>
<td>$-0.27$</td>
<td>[−0.37, −0.17]</td>
<td>$p &lt; 0.01$</td>
</tr>
<tr>
<td>Economic Equality Rank</td>
<td>$-0.16$</td>
<td>[−0.27, −0.06]</td>
<td>$p &lt; 0.01$</td>
</tr>
<tr>
<td>Political Equality Rank</td>
<td>0.05</td>
<td>[−0.05, 0.14]</td>
<td>0.19</td>
</tr>
<tr>
<td>Internet Penetration Rank</td>
<td>$-0.27$</td>
<td>[−0.44, −0.09]</td>
<td>$p &lt; 0.01$</td>
</tr>
<tr>
<td>Income Inequality Rank</td>
<td>0.01</td>
<td>[−0.09, 0.10]</td>
<td>0.44</td>
</tr>
<tr>
<td>Population Rank</td>
<td>0.01</td>
<td>[−0.09, 0.11]</td>
<td>0.40</td>
</tr>
<tr>
<td>Facebook Penetration Rank</td>
<td>0.03</td>
<td>[−0.12, 0.18]</td>
<td>0.33</td>
</tr>
<tr>
<td>Mean User Age Rank</td>
<td>0.02</td>
<td>[−0.08, 0.11]</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 2: Regression results of FGD model. Estimates of p-values are based on the posterior of parameter estimates after 10,000 iterations.

significantly correlated with predicted values. In addition, Facebook penetration is uncorrelated with residuals of the model (Pearson $-0.0028$, p-value $= 0.97$), showing no signs of bias due to the variance of Facebook penetration rates across countries.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>137.4</td>
<td>9.14</td>
<td>15.04</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
<tr>
<td>Education Equality Rank</td>
<td>$-0.55$</td>
<td>0.07</td>
<td>$-7.47$</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
<tr>
<td>Health Equality Rank</td>
<td>$-0.29$</td>
<td>0.05</td>
<td>$-4.98$</td>
<td>$p &lt; 10^{-5}$</td>
</tr>
<tr>
<td>Economic Equality Rank</td>
<td>$-0.13$</td>
<td>0.06</td>
<td>$-2.03$</td>
<td>0.04</td>
</tr>
<tr>
<td>Political Equality Rank</td>
<td>0.05</td>
<td>0.05</td>
<td>0.95</td>
<td>0.34</td>
</tr>
<tr>
<td>Internet Penetration Rank</td>
<td>$-0.29$</td>
<td>0.10</td>
<td>$-2.89$</td>
<td>$p &lt; 0.01$</td>
</tr>
<tr>
<td>Income Inequality Rank</td>
<td>0.001</td>
<td>0.05</td>
<td>0.01</td>
<td>0.99</td>
</tr>
<tr>
<td>Population Rank</td>
<td>$-0.003$</td>
<td>0.05</td>
<td>0.05</td>
<td>0.96</td>
</tr>
<tr>
<td>Facebook Penetration Rank</td>
<td>0.05</td>
<td>0.09</td>
<td>0.62</td>
<td>0.54</td>
</tr>
<tr>
<td>Mean User Age Rank</td>
<td>0.01</td>
<td>0.05</td>
<td>0.21</td>
<td>0.84</td>
</tr>
</tbody>
</table>

Table 3: Robust regression results of FGD model.
Table 4: Coefficient estimates using HC corrected estimates.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>95% HC CI</th>
<th>Standard error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>136.8</td>
<td>[122.5, 151.2]</td>
<td>7.32</td>
<td>( p &lt; 10^{-10} )</td>
</tr>
<tr>
<td>Education Equality Rank</td>
<td>-0.54</td>
<td>[-0.68, -0.40]</td>
<td>0.07</td>
<td>( p &lt; 10^{-10} )</td>
</tr>
<tr>
<td>Health Equality Rank</td>
<td>-0.27</td>
<td>[-0.36, -0.18]</td>
<td>0.05</td>
<td>( p &lt; 10^{-8} )</td>
</tr>
<tr>
<td>Economic Equality Rank</td>
<td>-0.17</td>
<td>[-0.26, -0.07]</td>
<td>0.05</td>
<td>( p &lt; 0.001 )</td>
</tr>
<tr>
<td>Political Equality Rank</td>
<td>0.04</td>
<td>[-0.04, 0.13]</td>
<td>0.04</td>
<td>0.32</td>
</tr>
<tr>
<td>Internet Penetration Rank</td>
<td>-0.27</td>
<td>[-0.44, -0.10]</td>
<td>0.09</td>
<td>( p &lt; 0.01 )</td>
</tr>
<tr>
<td>Income Inequality Rank</td>
<td>0.004</td>
<td>[-0.09, 0.10]</td>
<td>0.05</td>
<td>0.92</td>
</tr>
<tr>
<td>Population Rank</td>
<td>0.01</td>
<td>[-0.08, 0.10]</td>
<td>0.05</td>
<td>0.87</td>
</tr>
<tr>
<td>Facebook Penetration Rank</td>
<td>0.03</td>
<td>[-0.12, 0.18]</td>
<td>0.08</td>
<td>0.7</td>
</tr>
<tr>
<td>Mean User Age Rank</td>
<td>0.02</td>
<td>[-0.07, 0.10]</td>
<td>0.04</td>
<td>0.66</td>
</tr>
</tbody>
</table>

Figure 8: Left: Normal Q-Q plot of residuals of the FGD model. Right: Histogram of residuals.

The role of GDP

Fig. shows the relationship between the FGD and the GDP per capita. There is a significant negative correlation between both (Spearman \(-0.57, p < 10^{-6}\)), motivating a replication of the above model with GDP as a control. Since GDP is highly correlated with Internet Penetration and leads to high Variance Inflation, we replace Internet Penetration with GDP in our model. Coefficient estimates are reported on Fig. [10] revealing that the main result is robust to controlling for the wealth of countries.
Figure 9: Relationship between FGD and GDP (log scale).

**Model stability across monthly measurements**

We repeated the fit of the FGD model for measurements of the DAU in twelve months between 2015 and 2016. Fig. 11 shows the results of the fit for these alternative periods. The coefficient estimates barely depend on the period when the DAU are calculated and the $R^2$ of the fits range between 0.727 and 0.758, confirming that our results are robust to fluctuations in the reporting of DAU through the Facebook API.
Figure 10: Coefficient estimates of the FGD model with GDP instead of Internet Penetration as control.

Model stability across age groups

Figure 12 shows the replication of the model for segments of different age groups. Results are qualitatively similar to those of the whole population, with significant effects of gender equality variables and high $R^2$ values.
Figure 11: Results of repetition of the fit of the FGD model for 12 months between 2015 and 2016. The results are qualitatively stable across months.

Model test with GWI and Pew approximations to the FGD

We repeated the model using approximations of the FGD using the limited sample of GWI. The performance of the model is similar, as shown in Figure 13, with $R^2 = 0.77$. While the sample size of GWI is too small to test the role of all equality variables, the Pearson correlation between the rank of education equality and the rank of FGD in GWI is $-0.52$ ($p-value < 0.01$). Similarly, the model for the approximation of the FGD with data from Pew for all SNS gives similar $R^2 = 0.63$ and a significant negative Pearson correlation between the rank of education equality and the rank of FGD in PEW ($-0.73$, $p-value < 0.001$).
Figure 12: Replication of the model for data segmented into age groups.

Figure 13: Replication of the model with the GWI estimate of the FGD.

Supplementary Text 3 - Network externalities

The results of the network externalities model are shown on Fig. 14. The model achieves a $R^2 = 0.96$ on the logarithmic scale and a $R^2 = 0.89$ on the linear scale of activity ratios per gender. Table 5 shows the detailed results of the model, evidencing the superlinear scaling ($\alpha = 1.2$) and the difference between genders ($\alpha_F = 0.25$).

Fig. 15 shows the analysis of the residuals of the model and the error in the linear scale of activity ratios per gender. Some small deviations from normality can be observed at the tails, corresponding to significant Shapiro-Wilk statistics of 0.94 and 0.95. Both types of residuals are
uncorrelated with Facebook penetration and do not appear to have a structure across predicted values. We identified some of the residual outliers, such as China and Tajikistan, which when removed do not have a qualitative impact in the results of the model fit and lead to residual distributions closer to normality.

Table 6 shows the results when correcting for heteroskedasticity. All results remain qualitatively unchanged.

We repeated the fit using a robust regression method, reporting the results on Table 7.
estimates slightly change, the qualitative results of a superlinear relationship that is stronger for female users still hold. This shows that our conclusions are robust to the influence of outliers.

As with previous models, we evaluated the model of network externalities over twelve months following our initial measurement. Fig. 16 reports the overall results, showing no relevant decrease in $R^2$ and generally the same result, where the parameter $\alpha_F$ is significantly larger than zero and the parameter $\alpha$ is significantly larger than one.

We stratified the analysis, fitting the network externalities model using calculations of the
<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>95% Confidence Interval</th>
<th>Standard Error</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>$-0.57$</td>
<td>$[-0.61, -0.53]$</td>
<td>0.02</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$1.198$</td>
<td>$[1.17, 1.23]$</td>
<td>0.02</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
<tr>
<td>$\beta_F$</td>
<td>$0.15$</td>
<td>$[0.07, 0.24]$</td>
<td>0.045</td>
<td>$p &lt; 0.001$</td>
</tr>
<tr>
<td>$\alpha_F$</td>
<td>$0.25$</td>
<td>$[0.18, 0.33]$</td>
<td>0.038</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
</tbody>
</table>

Table 6: HC corrected results of network externalities model.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>Standard Error</th>
<th>t-value</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta$</td>
<td>$-0.523$</td>
<td>0.033</td>
<td>$-16.25$</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
<tr>
<td>$\alpha$</td>
<td>$1.19$</td>
<td>0.02</td>
<td>60.88</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
<tr>
<td>$\beta_F$</td>
<td>$0.25$</td>
<td>0.05</td>
<td>5.43</td>
<td>$p &lt; 10^{-7}$</td>
</tr>
<tr>
<td>$\alpha_F$</td>
<td>$0.37$</td>
<td>0.03</td>
<td>12.59</td>
<td>$p &lt; 10^{-10}$</td>
</tr>
</tbody>
</table>

Table 7: Robust regression results of the network externalities model.

Figure 16: Results of repetition of the fit of the network externalities model for 12 months between 2015 and 2016. The results are qualitatively stable across months.
FGD using only data from a set of age categories. Fig. [17] shows the model results, evidencing that the female intercept, which measures the surplus of the exponent for female users, is positive and significant for all age categories.

Figure 17: Results of repetition of the fit of the network effects model for age segments.
Supplementary Text 4 - Gender equality changes

Table 8 reports the Variance Inflation Factors of the variables in the model of economic gender equality changes and in the model of changes of FGD. All factors are low enough to discard multicollinearity.

<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>FGD Rank</td>
<td>1.833</td>
<td>FGD</td>
<td>1.66</td>
</tr>
<tr>
<td>Economic Gender Equality 2015</td>
<td>1.276</td>
<td>Rank Economic Gender Equality 2015</td>
<td>1.15</td>
</tr>
<tr>
<td>GDP Rank</td>
<td>1.505</td>
<td>GDP Rank</td>
<td>1.493</td>
</tr>
</tbody>
</table>

Table 8: Variance Inflation Factors of independent variables in the economic gender equality changes model.

Table 9 presents the detailed results of both models of changes. Before fitting, we rescaled the ranked variables to have a value between zero and one to allow a better comparison of their relationships, controlling for autocorrelation by including the unranked value of the variable in the previous year. The results of Table 9 are confirmed by ANOVA tests of the FGD rank in the $\Delta Eco$ model $F = 10.195, p < 0.01$, and the non-significant result for the Eco rank in the $\Delta FGD$ model $F = 0.003, p > 0.9$.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
<th>Term</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.011</td>
<td>0.015</td>
<td>0.437</td>
<td>Intercept</td>
<td>0.019</td>
<td>0.010</td>
<td>0.069</td>
</tr>
<tr>
<td>FGD Rank</td>
<td><strong>0.039</strong></td>
<td>0.01</td>
<td>$&lt; 0.01$</td>
<td>FGD</td>
<td>-0.002</td>
<td>0.011</td>
<td>0.848</td>
</tr>
<tr>
<td>Eco</td>
<td><strong>-0.061</strong></td>
<td>0.021</td>
<td>$&lt; 0.005$</td>
<td>Rank Eco</td>
<td>-0.001</td>
<td>0.015</td>
<td>0.94</td>
</tr>
<tr>
<td>GDP Rank</td>
<td><strong>0.052</strong></td>
<td>0.01</td>
<td>$&lt; 10^{-6}$</td>
<td>GDP Rank</td>
<td>0.007</td>
<td>0.015</td>
<td>0.65</td>
</tr>
<tr>
<td>Multiple $R^2$</td>
<td>0.1501</td>
<td></td>
<td></td>
<td>Multiple $R^2$</td>
<td>0.0009</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 9: Results of robust regression for the model of changes in economic gender equality and of changes in the FGD.

The residuals of the model of changes in economic gender equality are distributed close to normality, as shown on Fig. 18, with a significant Shapiro-Wilk statistic of 0.97 and only some small deviations from normality at the tails. Residuals are uncorrelated with all independent
variables and do not show signs of heteroskedasticity.

![Graph showing normal Q-Q plot and histogram of residuals](image)

Figure 18: Analysis of residuals of the economic gender equality changes model. The left panel shows the normal Q-Q plot of residuals of the model, and the right panel their histogram. Some minor deviations from normality can be observed in both.

We tested the robustness of the positive association between FGD and $\Delta E_{co}$ in two new fits including the same controls as for the FGD model (VIF of all factors below 5). The results are shown on Table 10, evidencing that the observed association between FGD and $\Delta E_{co}$ is robust to other socio-economic indicators, and to other possible confounds such as Facebook penetration or mean user age.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.007690</td>
<td>0.01828</td>
<td>0.67473</td>
<td>0.006294</td>
<td>0.01796</td>
<td>0.72667</td>
</tr>
<tr>
<td>FGD Rank</td>
<td><strong>0.02431</strong></td>
<td>0.01096</td>
<td>&lt; 0.05</td>
<td><strong>0.02640</strong></td>
<td>0.01091</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Eco</td>
<td>-<strong>0.07252</strong></td>
<td>0.02479</td>
<td>&lt; 0.01</td>
<td>-<strong>0.07597</strong></td>
<td>0.02437</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>Ineq Rank</td>
<td>-<strong>0.01643</strong></td>
<td>0.00798</td>
<td>&lt; 0.05</td>
<td>-0.013190</td>
<td>0.00792</td>
<td>0.09833</td>
</tr>
<tr>
<td>Pop Rank</td>
<td><strong>0.01791</strong></td>
<td>0.00836</td>
<td>&lt; 0.05</td>
<td><strong>0.01937</strong></td>
<td>0.00809</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Mean Age Rank</td>
<td>0.005872</td>
<td>0.01213</td>
<td>0.62919</td>
<td>0.006279</td>
<td>0.01178</td>
<td>0.59494</td>
</tr>
<tr>
<td>FB Penetration Rank</td>
<td>0.015300</td>
<td>0.01353</td>
<td>0.26029</td>
<td>0.015453</td>
<td>0.01238</td>
<td>0.21442</td>
</tr>
<tr>
<td>GDP Rank</td>
<td>0.024484</td>
<td>0.01664</td>
<td>0.14359</td>
<td>0.024763</td>
<td>0.01539</td>
<td>0.11000</td>
</tr>
<tr>
<td>Internet Penetration Rank</td>
<td>0.2009</td>
<td></td>
<td></td>
<td>0.1966</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 10: Results of the $\Delta E_{co}$ model including additional controls.

We further tested the possible role of other equality indices in the relationship between FGD
and $\Delta Eco$. We added all other three gender equality scores as controls (VIF below 5), and repeated the fit. The result, shown on Table 11, shows that the positive association between FGD and $\Delta Eco$ is robust to the possible effect of other kinds of gender inequality.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-0.029304</td>
<td>0.021736</td>
<td>0.179917</td>
</tr>
<tr>
<td>FGD Rank</td>
<td>0.039560</td>
<td>0.016828</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Economic Score rank</td>
<td>-0.043693</td>
<td>0.025742</td>
<td>0.091983</td>
</tr>
<tr>
<td>GDP rank</td>
<td>0.045928</td>
<td>0.012187</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Education Score rank</td>
<td>0.001291</td>
<td>0.013318</td>
<td>0.922924</td>
</tr>
<tr>
<td>Political Score rank</td>
<td>0.014899</td>
<td>0.009621</td>
<td>0.123862</td>
</tr>
<tr>
<td>Health Score rank</td>
<td>0.004209</td>
<td>0.009031</td>
<td>0.641953</td>
</tr>
<tr>
<td>N</td>
<td>139</td>
<td></td>
<td>Multiple $R^2$</td>
</tr>
</tbody>
</table>

Table 11: Results of the $\Delta Eco$ model including controls for other gender equality indices.
We tested whether the association between FGD and $\Delta Eco$ could be explained by general cultural differences. We combined our dataset with Hofstede’s cultural dimensions: Power Distance Index (PDI), Individualism (IDV), Uncertainty Avoidance Index (UAI), and Masculinity (MAS) (VIF below 5). This limits the analysis to a set of 66 countries common to both datasets, with results reported on Table 12. The positive association between FGD and $\Delta Eco$ is still significant, suggesting that the relationship between both variables goes beyond what Hofstede’s model captures in terms of culture.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.04743</td>
<td>0.03102</td>
<td>0.1316</td>
</tr>
<tr>
<td>FGD Rank</td>
<td>0.03792</td>
<td>0.01764</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td>Eco</td>
<td>-0.1243</td>
<td>0.03779</td>
<td>&lt; 0.01</td>
</tr>
<tr>
<td>PDI</td>
<td>3.217*10^{-4}</td>
<td>1.759*10^{-4}</td>
<td>0.0725</td>
</tr>
<tr>
<td>IDV</td>
<td>1.754*10^{-4}</td>
<td>-1.708*10^{-4}</td>
<td>0.3088</td>
</tr>
<tr>
<td>MAS</td>
<td>-2.490*10^{-4}</td>
<td>1.546*10^{-4}</td>
<td>0.1126</td>
</tr>
<tr>
<td>UAI</td>
<td>3.148*10^{-5}</td>
<td>1.366*10^{-4}</td>
<td>0.8185</td>
</tr>
<tr>
<td>N</td>
<td>66</td>
<td>Multiple $R^2$</td>
<td>0.362</td>
</tr>
</tbody>
</table>

Table 12: Results of the $\Delta Eco$ model including controls for cultural dimensions.

Following the same methodology as for previous models, we stratified our analysis with FGD measured only in a variety of age groups. The results are shown on Fig. [19] revealing the positive and significant role of FGD in the model for all age segments.
Figure 19: Replication of the model stratifying for different age groups.

Our data offers the opportunity to measure the role of FGD in the changes of other gender equality measures, but the indices for Political and Health gender equality have negligible changes between 2015 and 2016. For that reason we can only evaluate the role of FGD for changes in Education gender equality. We find no significant effect of FGD, as reported on Table 13.

<table>
<thead>
<tr>
<th>Term</th>
<th>Estimate</th>
<th>s.e</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>0.0613210</td>
<td>0.0613210</td>
<td>&lt;10^{-7}</td>
</tr>
<tr>
<td>FGD rank</td>
<td>-0.0004546</td>
<td>0.0023720</td>
<td>0.848</td>
</tr>
<tr>
<td>Edu</td>
<td>-0.0606702</td>
<td>0.0099782</td>
<td>&lt;10^{-6}</td>
</tr>
<tr>
<td>GDP rank</td>
<td>-0.0012934</td>
<td>0.0022238</td>
<td>0.562 &lt; 0.01</td>
</tr>
<tr>
<td>N</td>
<td>139</td>
<td>Multiple $R^2$</td>
<td>0.08</td>
</tr>
</tbody>
</table>

Table 13: Results of the $\Delta Edu$ model.
<table>
<thead>
<tr>
<th>Variable</th>
<th>VIF</th>
</tr>
</thead>
<tbody>
<tr>
<td>Education Gender Equality Rank</td>
<td>2.200318</td>
</tr>
<tr>
<td>Economic Gender Equality Rank</td>
<td>1.500195</td>
</tr>
<tr>
<td>Health Gender Equality Rank</td>
<td>1.309959</td>
</tr>
<tr>
<td>Political Gender Equality Rank</td>
<td>1.268285</td>
</tr>
<tr>
<td>Internet Penetration Rank</td>
<td>4.036525</td>
</tr>
<tr>
<td>Income Inequality Rank</td>
<td>1.167592</td>
</tr>
<tr>
<td>Total Population Rank</td>
<td>1.221762</td>
</tr>
<tr>
<td>Facebook Penetration Rank</td>
<td>2.950804</td>
</tr>
<tr>
<td>Mean User Age Rank</td>
<td>2.630176</td>
</tr>
</tbody>
</table>

Table 14: Variance Inflation Factors of independent variables in the FGD model.
References


[13] “French data watchdog dishes out largest gdpr fine yet,” [https://www.theregister.co.uk/2019/01/21/google_50m_cnil_gdpr/](https://www.theregister.co.uk/2019/01/21/google_50m_cnil_gdpr/).


[25] P. Papadopoulos, N. Kourtellis, P. R. Rodriguez, and N. Laoutaris, “If you are not paying

[26] “The world’s most valuable resource is no longer oil, but data,” May 2017. [Online]. Available: [https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data](https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data)


REFERENCES


REFERENCES


REFERENCES


[142] K. Lyytinen and J. Damsgaard, “What’s wrong with the diffusion of innovation theory?” in


