ESTIMATING AGE AND GENDER IN INSTAGRAM USING FACE RECOGNITION: ADVANTAGES, BIAS AND ISSUES.

DIEGO COUTO DE. LAS CASAS

ESTIMATING AGE AND GENDER IN INSTAGRAM USING FACE RECOGNITION: ADVANTAGES, BIAS AND ISSUES.

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais – Departamento de Ciência da Computação como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

Orientador: Virgílio Augusto Fernandes de Almeida

Belo Horizonte

Fevereiro de 2016

DIEGO COUTO DE. LAS CASAS

ESTIMATING AGE AND GENDER IN INSTAGRAM USING FACE RECOGNITION: ADVANTAGES, BIAS AND ISSUES.

Dissertation presented to the Graduate Program in Ciência da Computação of the Universidade Federal de Minas Gerais – Departamento de Ciência da Computação in partial fulfillment of the requirements for the degree of Master in Ciência da Computação.

Advisor: Virgílio Augusto Fernandes de Almeida

Belo Horizonte

February 2016

Ficha catalográfica elaborada pela Biblioteca do ICEx - UFMG

Las Casas, Diego Couto de.

L337e Estimating age and gender in Instagram using face recognition: advantages, bias and issues. / Diego Couto de Las Casas. – Belo Horizonte, 2016.
xx, 80 f. : il.; 29 cm.

Dissertação (mestrado) - Universidade Federal de Minas Gerais – Departamento de Ciência da Computação.

Orientador: Virgílio Augusto Fernandes de Almeida.

1. Computação - Teses. 2. Redes sociais on-line. 3. Computação social. 4. Instagram. I. Orientador. II. Título.

CDU 519.6*04(043)

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Estimating age and gender in instagram using face recognition: advantages, bias and issues

DIEGO COUTO DE LAS CASAS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. VIRGÍLIO AUGUSTO FERNANDES ALMEIDA - Orientador
Departamento de Ciência da Computação - UFMG

PROF. ADRIANO ALONSO VELOSO
Departamento de Ciência da Computação - UFMG

DR. DANIELE QUERCIA
University of Cambridge

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 22 de março de 2016.

Acknowledgments

Gostaria de agradecer a todos que me fizeram chegar até aqui.

À minha família, pelos conselhos, pitacos e por todo o suporte ao longo desses anos. Aos meus colegas do CAMPS(-Élysées), pelas colaborações, pelas risadas e pelo companheirismo. Aos meus amigos, pelos bons momentos e pelos momentos edificantes. À Ana, pelo apoio, pelas conversas, pelo carinho, pela inspiração e por estar sempre comigo. Ao meu orientador Virgílio, pela sabedoria, paciência, e por acreditar no meu trabalho. E finalmente, aos professores e as secretárias do DCC, por me ajudar a resolver todas as confusões e pepinos.

Resumo

Estudos em computação social frequentemente se utilizam de atributos pessoais dos usuários de serviços on-line a fim de entender melhor o seu comportamento. Como esses atributos são muitas vezes indisponíveis para pesquisadores e desenvolvedores, esforços recentes têm se dedicado a estimá-los através da combinação de outras fontes de informação. Além de oferecer *insights* sobre como usuários se relacionam com plataformas online, tais metodologias de estimação de atributos também podem contribuir na compreensão de quão expostas estão as informações do usuário a terceiros. Nesta dissertação eu proponho estudar o uso de tecnologias de reconhecimento facial para a estimação do gênero e da idade dos usuários de uma rede social online baseada em imagens, o Instagram. Esta abordagem é inspirada pela crescente riqueza de informações de dados de imagem em redes sociais on-line, bem como os recentes avanços no reconhecimento facial.

Abstract

Studies in social computing often take into account personal attributes of users of specific online services in order to better understand their behavior. As these attributes are often unavailable for researchers and developers, recent efforts have been devoted to estimate them by combining other sources of information. Besides offering insights to how users relate to online platforms, such attribute estimation methodologies can also contribute to understanding how exposed is user information to third parties. In this master thesis I propose to study the use of face recognition technologies to estimate the age and the gender of the users of a popular, image-based online social network, Instagram. This approach is inspired by the increasing wealth of information from image data in online social networks, as well as recent advances in face recognition.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

The internet has acquired a major role in the daily lives of an increasing number of people in the world. The website Internet World Stats, for example, reports that the web has 3,366,261,156 users, equivalent to 46.4% of the world population, 832.4% of what it had a decade ago[1]. Product of the explosion of innovations brought by the Internet, Online Social Networks (OSNs) allowed users to interact with each other in unprecedented ways, and new methods of analysis and information processing are transforming these forms of interactions into insights and applications that are increasingly customized and geared towards certain demographic groups [Boyd, 2013].

Due to this influx of innovation, discussions once considered unrelated to the field of Computer Science must now be addressed, in order to enable the design of solutions that tackle the new demands posed by these systems. These new demands spanned a wide range of studies that combine data collection algorithms with statistical techniques in order to harvest thousands or millions of user profiles and test hypotheses that could never be tested before.

Traditionally – perhaps in continuation with studies of the topology and properties of the Internet – these studies have treated social networks like a huge graph, focusing in the modeling of graph properties. For example, Mislove et al. [2007] studied theoretical characteristics of friendship networks in Orkut, Flickr, LiveJournal and YouTube, demonstrating interesting properties like a small average path length between nodes (*small-world*) and a strongly uneven distribution of edges (*scale-free*). In fact, in recent years, virtually all global-scale OSNs and many regional OSNs have been studied in this fashion. Examples are Google+ [Magno et al., 2012], Twitter [Kwak et al., 2010], Orkut [Benevenuto et al., 2009], Facebook [Ugander et al., 2011] and Cyworld [Ahn et al., 2007a].

[1] `http://www.internetworldstats.com/stats.htm` (Visited Jan 2016)

Recently, the attention has shifted from simply estimating graph properties to developing methods of systematically characterizing the content produced in the network [Ottoni et al., 2014; An et al., 2011; Zhao et al., 2011] and the personal attributes of the users who interact within a network [Cunha et al., 2014; Gong et al., 2012]. This allows researchers and designers to better understand different patterns of behavior within a network, and to develop theories and intuitions based on existing literature in the social and behavioral sciences.

In the present work, I aim to explore a novel method of estimating and analyzing age and gender in OSNs. There are three main motivations for such objective.

First, age and gender are well known predictors of user behavior. Recent studies show a consistent pattern of gender differences in relevant online behavior, as in the choice of hashtags in Twitter [Cunha et al., 2014], the expression of positive feelings in micro-blogging networks [Kivran-Swaine et al., 2012], and organization of interaction networks in massive online games [Szell and Thurner, 2013]. I was fortunate to contribute in the area with a study on information disclosure in Facebook, in which we showed how self-reported gender and age were important predictors of exposure both to the community at large and to the user's social circle [Quercia et al., 2012]. In another work, we showed how men and women differed significantly in how they share images in Pinterest [Ottoni et al., 2013].

Second, gender and age are sensitive attributes of digital life. As mentioned above, users of different ages and genders differ in the way they share information, and for good reasons – users with different age/gender profiles have different concerns on how they use OSNs in their social life. For example, Boyd [2007] describes how young people use OSNs publicly but control their level of exposure through implicit codes and impression management – what she calls *social stenography*. Similarly, Hargittai et al. [2010] shows that although the youth notably expose more personal information on the network than adults, they care and control the public who have access to that information.

In fact, the relationship between OSNs and the youth has been central to the public debate concerning Internet services. An example of a heated debate in the early days of OSNs was centered on the risks that children in MySpace faced by exposing themselves to sexual predators. In this sense, the work of Marwick [2008] shows how the public perception of this risk has motivated many public policies from the U.S. government and design decisions from MySpace's owners. Lewis et al. [2008] also cite MySpace's media scandals to explain their observation that women are more concerned about their privacy than men.

This concern about privacy is important because users are frequently not aware of

the level of exposure they have in a network. Indeed, personal exposure online is often underestimated, since users do not usually take into account the various strategies used by online services and third parties to discover and infer their personal data - like the so-called *privacy attack algorithms*. In a recent work, my research group and I showed how photo-tags – pointers to other users in photos in Facebook – can be leveraged to uncover the user's age and gender with good accuracy [Pesce et al., 2012]. Knowledge about which methods can be used to uncover socially sensitive information is important to spread awareness of the level of exposure users are submitting themselves.

The third motivation for this work is that understanding the relationship between OSN usage and attributes such as gender and age has a high social impact, since it is related to socially relevant debates such as gender equality and generational differences. In fact, a number of recent studies combined data collection methods with theories from the humanities and social sciences to study the pervasiveness of issues such as women's lack of representation in highly prestigious positions [Wagner et al., 2016; Terrell et al., 2016], the biased depiction of women in their community-written biographies [Graells-Garrido et al., 2015] and the lack of efficiency of impeding kids from using social networks in order to protect their safety [Dey et al., 2013; Minkus et al., 2015].

Following this trend, my research group and I managed to show that the proportion of women that appeared in photos and "selfies" posted in Instagram in different countries correlated with gender equality indicators [Souza et al., 2015]. We also investigated how users who declared they were neither male or female behaved in Google+, raising the hypothesis that other gender categories allow for non-binary gender identities to be expressed in the network [de Las Casas et al., 2014].

The growing perception of the sensitivity and value of these two attributes means that they are becoming increasingly more difficult to obtain. There is an effort of both the users and the OSNs of hiding them from public exposure, and sometimes they are not even available as a field to be filled by the user. This is not to say that the OSN is simply ignoring this information. For example, Twitter does not have a user-fed gender field, and the age field is not publicly exposed. However, both their campaign targeting tools and their analytics tools offer filters of age and gender, which means they do track each user's age and gender, but chose not to expose it. This is expected of any OSN with business models that rely heavily in advertising, exactly because age and gender are strong predictors of user – and consumer – behavior, and therefore primary variables to be used at user segmentation strategies. But this is also expected of any OSN that make efforts of segmenting its user base in order to offer customized services.

This emerging scenario, in which gender and age information is increasingly dif-

ficult to obtain, but also increasingly relevant, calls for a need of estimating these attributes in novel ways. Methods developed for such task must be thoroughly evaluated in order to avoid unexpected biases and effects. Additionally, these methods must be publicized to all parts involved whenever possible, and methods using proprietary technologies must present a way of properly evaluating the "black boxes" they encompass.

In this work, I propose to use face recognition technology to estimate age and gender. Although person recognition and attribute estimation using facial data is still considered an open problem, there have been much recent progress, which motivated many social computing researchers to incorporate these technologies in their work [Bakhshi et al., 2014; Redi et al., 2015b,a; Jang et al., 2015; Polakis et al., 2012; Li et al., 2014].

Using face data for this task makes intuitive sense – humans recognize themselves mostly by looking at their own faces, having a specialized brain area for such task, and incorporating such data into an attribute estimation methodology can help us approach, or even surpass, a human-level recognition performance. Moreover, using data in images uploaded by the users goes in line with recent trends in Internet use, as users have been massively uploading and publishing personal photographs, and social media have been increasingly more adorned with pictures of the persons related to the content they share. This shift from textual content to pictorial content in the Internet – the Visual Web – has been observed and debated by specialists[2]. It is only logical to expect that novel methods in social computing will try to leverage this new data-rich environment.

However, the use of face recognition for estimating user attributes has not been rigorously evaluated, only validated in a case-by-case fashion – when validated at all. More specifically, no study has tackled possible sources of bias in these automatic attribute estimation methods. Thus, I propose to evaluate and validate a popular, proprietary facial recognition system in a reproducible way, and then show how its bias can be corrected or at least taken into account when analyzing user behavior.

I intend to use Instagram to test this approach. Instagram is the OSN that managed to dominate the Visual Web. Although other services such as Google+, Flickr and Facebook offer photo sharing capabilities and allow for social interactions in photos, Instagram is a network solely dedicated to social interactions centered around photo (and short video) sharing.

Thus, my full proposal is to study the use of face recognition technologies for at-

[2] `http://om.co/2014/12/10/weaving-a-very-visual-web/` (Visited Jan 2016)

tribute estimation using Instagram as a case study. To be more specific, by **attribute** I mean any relevant characteristic that can be retrieved from information available in the OSN that is directly and individually related to each user. This can be either explicitly stated in the OSN's interface, or can be inferred from other explicit information. In the later case, the attribute is **estimated**. Thus, I intend to use explicitly available information in Instagram's interface (the content of the photos posted by the users) to infer attributes of the posts (the number, gender and age of the persons depicted in the photos) and eventually attributes of the users (the gender and age of the users who posted the photos). In doing so, I intend to open a discussion about the benefits and risks of using face recognition technologies for such tasks, as well as to present a framework to evaluate the reliability of such methodology.

In Chapter 2, I will explain how I collected data from the OSN and briefly describe attributes of the users I managed to collect. In Chapter 3 I will briefly walk through the state of the art in face recognition and present my evaluation of the face recognition system of choice, `Face++`, using a benchmark dataset that best approaches the kind of images that are expected to be found in Instagram. In Chapter 4, I will present my analysis of relevant aspects of user behavior in Instagram using `Face++` in conjunction to other well-known methods of analysis. Finally, in Chapter 5, I will round up the decisions and findings presented here and present a final discussion on the topic.

Chapter 2

Instagram

Instagram is a free OSN for photo and video sharing with over 400 million active users and 40 billion photos shared as of January 2016[1]. It has a wide international projection: only 25% of its user base is from the service's country of birth (USA). It was launched as an iPhone app at October 2010 and rapidly gained popularity, reaching 10 million users in September of 2011. It was named "iPhone app of the year" by Apple on December 2011 and was bought by Facebook four months later. The Android version was launched in April 2012, and the Windows Phone version of Instagram was only launched in November of 2013, after much pressure from Nokia executives[2].

Instagram was one of the first popular mobile-first applications: although it has mobile and web interfaces, users can only post and interact in the mobile interface. It is also neatly integrated with other services, and allows users to share content to networks, such as Facebook, Twitter and Tumblr.

Next, I will briefly describe how the service is structured, introducing the concepts that I will use throughout the text to describe the data. In Section 2.1, I will describe the main concepts related to the network. In Section 2.2, I will describe how the Instagram's data was collected and the methodological decisions made during the collection step. Then, in Section 2.3, I will explain how geolocation was handled to map each geotagged media to its country of origin. Finally, in Section 2.4, I will present an exploratory analysis of the collected data that aims to examine how the data is distributed.

[1] `https://www.instagram.com/press` (Visited Jan 2016)
[2] `http://techcrunch.com/2013/11/20/instagram-windows-phone/` (Visited Jan 2016)

2.1 Core concepts

Instagram allows the user to take a picture or a short video, edit its visual and metadata properties and post it to the network. Edition of the media content can be made by a variety of pre-defined filters - considered one of the main features of the service. Moreover, users can add captions, hashtags and geolocation metadata. After posting a media, other users can interact with it through comments and signs of approval ("likes"). The user can also share this content in other online social networks, which extends the post's reach to beyond Instagram.

Anyone with a smartphone running Android, Windows or iOS can download the App and create a user account tied to a **username** and a **user ID**. A **user profile** is assigned to each user account. After created, the user account can be used to visit the user's own profile or other profiles visible to her.

Users can create **posts** in their profiles, which are displayed as a list, sorted in descending order (from last post to first post). A post is a **media** object, which can contain either a photo or a 3 to 15 second long video, a list of **comments** and **likes** of other users and extra metadata detailing when and where the post was created and textual data describing (and enhancing) the post's content. The data and metadata from the media object is best described in Table 2.1.

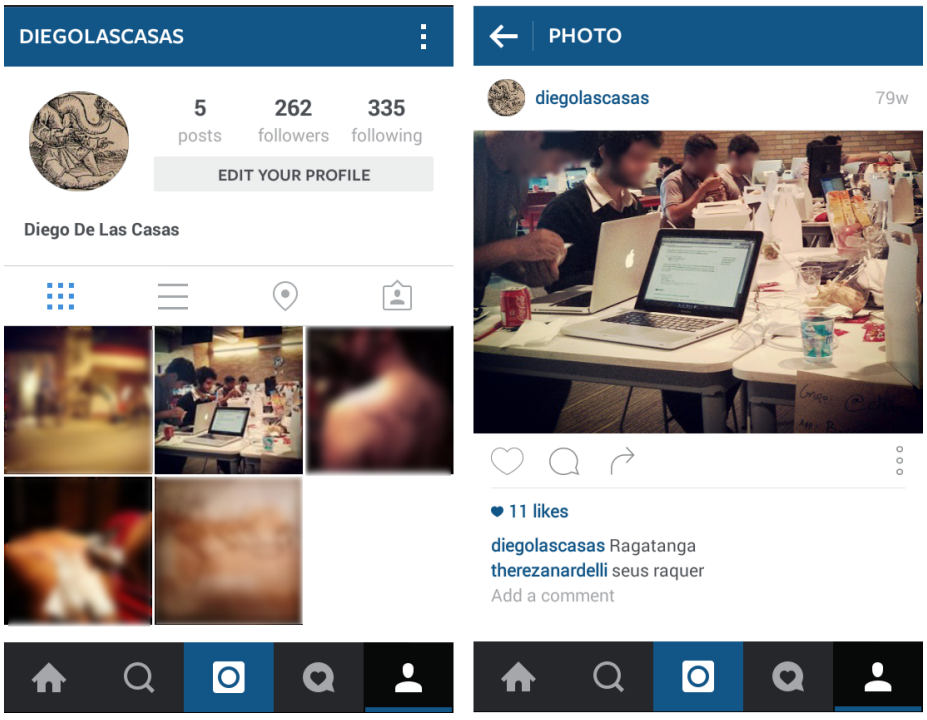


Figure 2.1: Screenshots of a user profile (left) and a post (right) in Instagram. The faces and remaining posts were blurred.

`media`	Either a photo or a 3 to 15 second long video and all its associated metadata
`filter`	One in a predefined set of image transformations suggested by Instagram to make the posts prettier
`timestamp`	Time of creation of the post. Added automatically by the service.
`caption`	A field of text, which is displayed below the media and above the comments.
`hashtags`	A set of words preceded by the hash sign ("#") added in the caption. They are hyperlinks to other posts with the same tag, and can also be queried in a search engine provided by the service.
`geotag`	A pair of a GPS coordinate and a location name, referring the post to a physical location. The user can pick any existing (*i.e.* named) location for the post, but only one per post. Geotags are hyperlink to a page with the location on a map, along with the most recently posted public media using the tag.
`users_in_post`	A set of links to user profiles who are presumably in the photo. The user who made posted the photo provides these tags, and they are not equivalent as another user being in a photo, *i.e.* a photo with a tag of an user does not necessarily depict the person whose user account was tagged, and a person that appears on a photo will not necessarily have her user profile tagged.

Table 2.1: Attributes of media objects in Instagram

Social interactions in the network can occur at *media level* or at *user level*. At media level, it is possible to *like* or *comment* a post.

Liking is a form of lightweight interaction in which a user signals that he approved (or supports) the content of the post. If at most 10 users liked a post, their username is displayed in it. If more users liked it, the amount of likes is displayed. **Commenting** has a higher social cost, and means appending a short text in the post, signed with the user's username. The profile owner can delete any comment made in its media. In comments and captions, a user can **mention** another user by typing its username preceded by an "@". Comments can also have hashtags.

At user level, it is possible to **follow** or be **followed** by other users. The relationship is not symmetric, meaning that it is possible to follow someone and not be followed back, and vice-versa. Following a user means that everything posted by her

will appear in the main page (**feed**) of the followee.

Instagram also offers some options to control the profile's visibility. All user accounts are **public** by default, which means that every content published by a user can be viewed by anyone without the need of following that user. This setting can be altered in the mobile app, where a user can configure its account as **private**. Private users' publications are only visible to their followers, and they cannot be followed without their consent.

Instagram does not offer finer grain control of the posts visibility, such as posting a content only to a predefined list of followers (such as in Google+) or posting something public while keeping the account private (such as in Facebook). However, it has a service called **Instagram Direct**, available for all users in the mobile application, that allows a user to send photos or videos directly to other specific users. Content sent via Instagram Direct can only be viewed by the recipients no matter if the sender's account is public or private.

2.2 Data collection

Instagram offers an Application Programming Interface (API) that allows developers to access many of the data published in the network. By making requests to Representational State Transfer (REST) endpoints, it is possible to obtain information of users, media, relationships and comments, among other data types. However, the API imposes tight limits to the number of requests allowed per client. To address this issue, my research team and I built a distributed crawler to make requests from multiple clients, the *CAMPS Data Collection Tool*. Details of its implementation can be found in de Souza [2015].

2.2.1 A brief review on sampling methodologies

Methodologies for sampling OSNs can be informally divided in two approaches: **crawling** and **selection sampling**. The proper choice of approach depends on the sampling goal, *i.e.* what is being modeled and which attributes are being estimated.

In the **crawling** approach, the network is treated like a big, unknown graph, and Graph Sampling Algorithms are used. These either are a simple implementation of a graph traversal algorithm (*e.g.* Bread First Search, Depth First Search, etc), which may produce biased samples but when used correctly can reasonably approach the graph's characteristics [Mislove et al., 2007; Ahn et al., 2007b]; or are more sophisticated sam-

pling methods that correct for bias either at collection time or at a posterior correction step [Kurant et al., 2011; Gjoka et al., 2010].

In general, in the case of simple, non-corrective algorithms, a proper sample must contain a reasonable component of the graph (*e.g.* the biggest connected component), while for corrective algorithms this is not necessary, but at expense of a significant increase in the cost of data collection.

The alternative approach, **selection sampling**, involves finding a way of selecting a group of nodes independent of their position in the network. This is normally done by using a feature such as a feed of top posts [De Choudhury et al., 2010]. The problem of this strategy is that the collection is strongly biased towards users with higher visibility, such as celebrities.

Alternatively, one can often query the network for user IDs, and some authors have used the ID selection strategy to collect the *whole* network [Magno et al., 2012; Cha et al., 2010]. However, these authors have relied on idiosyncrasies of the collected networks, such as sequential ID numbers with few gaps, or availability of the full ID list. Since data collection can provide third parties important information about a network service, OSNs will often try to increase the cost of this type of collection by making the generated IDs more difficult to guess – normally by hiding the ID list and making the ID space sparse, with a low hit-to-miss ratio, so that many possible IDs must be guessed before a valid node is discovered.

This makes collecting the whole network unfeasible in most cases, but still allows for smaller samples. When done at random, selection sampling is equivalent to classical statistical sampling, and has the advantage of always producing unbiased samples with respect to node labels. One disadvantage of this method is that, since edges are not taken into account, the network loses most of its topological structure unless a big proportion of it is sampled or the selection is made by edge instead of by node [Lee et al., 2006; Leskovec and Faloutsos, 2006].

2.2.2 Sampling Instagram

For our project, we did not intend to model Instagram as a graph, so node (*i.e.* user) relationships are not of primary interest. Although attributes such as user interactions are of interest, the *links* between users can be safely ignored. I will only work with the labels assigned to nodes, which can be collected by the API. This has the additional advantage of using node attributes calculated taking the whole network into account, since they are provided by Instagram, thus avoiding distortions due to sampling. Moreover, it is of interest to have a representative sample of Instagram's population. Therefore,

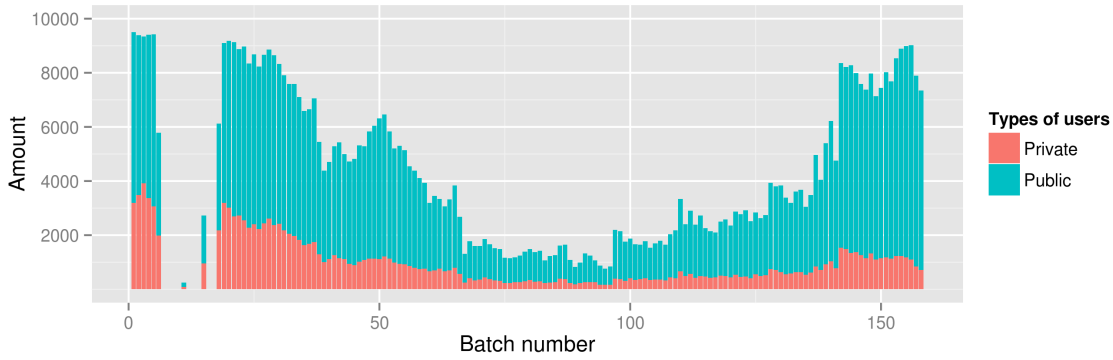


Figure 2.2: Instagram's ID space

selection sampling with a random batch of user IDs was chosen.

To do so, one needs to know how the IDs were distributed. Thus, the following methods was devised: (a) The ID space was divided in batches of 10^7, *i.e.* Batch 1 would hold the IDs $\{1, 2, \cdots, 9999999\}$, Batch 2 would hold the IDs $\{10000000, 10000001, \cdots, 19999999\}$, and so on. (b) For each batch, 10^4 IDs were chosen randomly and the hits (*i.e.* IDs that existed) were counted.

Using this method, no hit after Batch 158 could be found, which means that all valid IDs were between 1 and 1.58×10^9. We can also get an idea of how the IDs were distributed, as displayed in Figure 2.2. It can be seen that the IDs are not uniformly distributed across the whole ID space. Instead, the distribution is U-shaped, with a gap starting at 500M, reaching its "valley" at around 700M, and then slowly rising.

After knowing the ID space and assuring that it is completely covered, 16 million IDs were sampled from the valid ones and the user profiles were collected in December 2014. 30% of the sampled profiles were public, while the remaining profiles were either private, blocked or deleted.

Activity was collected retroactively up to three and a half years, giving information from January 2012 to December 2014. This data traces Instagram's activity to a time period when it had just reached its 10M user mark, a few months before its acquisition by Facebook, and in a time when its use was restricted to iPhone users.

The first post of most of the users can be captured by looking at the post with the minimum `created time` for each user ID. Figure 2.3 shows the first post of each user ID. It is possible to affirm that the IDs follow a somewhat chronological order, due to the lack of any data point in the upper left corner of the graph[3]. However, the presence of lower ID numbers in more recent "first posts" suggests that (1) many

[3]Notice that this method will not be reliable for the first 10M users of the service, but they will be more unlikely to appear in the dataset as the ID number increases.

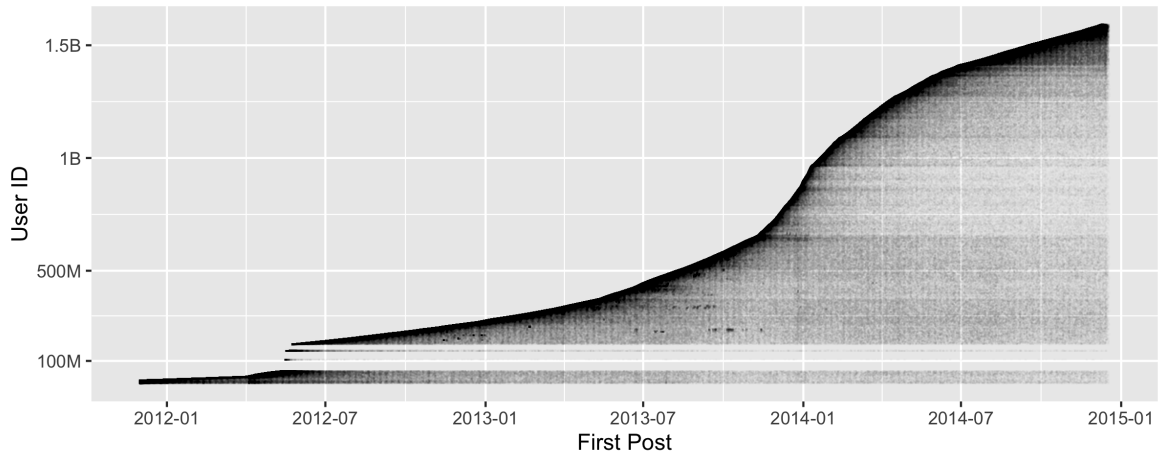


Figure 2.3: Scatterplot showing the relationship between ID and date of the user's first post. Dots were set to an alpha level of 1/100 in order to highlight overlaps.

IDs were reused along Instagram's existence; (2) some IDs were "skipped" and only attributed to a user later or (3) many users created their profiles and did not post anything until years later. Notice that these possible explanations are not mutually exclusive.

The ID space's growth may have followed the growth of Instagram, but considering that the U shape of the ID space coincides with the steeper slope in the ID growth, it seems there was an implementation decision taken around July 2013 to generate sparser ID numbers. Thus, the steeper slope does not seem to represent solely a higher adoption rate of Instagram.

It may be interesting to point out that the "leap" in IDs seems to have happened around June 2012. This was a few months after the acquisition by Facebook (April) and at the same moment that they announced having reached 80 million users[4].

Table 2.2 shows some interesting descriptives on the results of the collection of the public profiles, broken down by different patterns of user activity.

2.3 Resolving geolocation

Many posts of the collected users were geotagged. As stated in Section 2.1, geotags are a pair of name and geographical coordinates. The name is arbitrary and unpredictable[5], but it is possible to map each coordinate to a location defined in a geographical database. There is a number of freely available geographical databases with

[4]http://blog.instagram.com/post/28067043504/the-instagram-community-hits-80-million-users

[5]Names are Foursquare locations. Foursquare is a location-based social network that partnered up with Instagram in 2010: `foursquare.com`

Statistic	Absolute	Relative
Total number of valid users	5,170,062	
... with at least 1 followee	3,801,988	74%
... with at least 1 follower	3,797,961	73%
... with at least 1 post	2,860,421	55%
... without followers and followees	956,813	19%
Total number of posts	153,979,348	
... that are photos	150,088,274	97%
... with at least one like	141,087,975	91%
... with a least one comment	57,699,726	37%
... without likes and comments	12,252,832	8%
... with hashtags	58,794,786	38%
... with geotags	35,392,626	22%

Table 2.2: Collection Descriptives

differing levels of resolution. I opted to use the Global Administrative Areas (GADM) Database – a high-resolution public database of country administrative areas, with a goal of mapping "the administrative areas of all countries, at all levels"[6]. Version 2.8, released in November 2015, was used to map coordinates to country labels.

Figure 2.4 shows a sample of all the geotags scattered around the world, and Figure 2.5 shows the 310,067 media (0.876 % of the tagged media) that could not be mapped using GADM. Most of these geotags are located in the coastlines of countries, which means that they were either taken over the sea, or taken at a seashore and displaced by measurement error. Moreover, 585 locations were invalid (had a latitude over 90, under -90, or longitude over 180 or under -180). By looking at the name given by Instagram, I could notice that some of those were inverted (latitude was considered longitude), and the remaining of them had impossibly large values (*e.g.* longitude of 999 degrees). I could not reproduce the error, and since the proportion of erroneous cases was negligible, I simply filtered them out.

Due to the low number of posts in certain countries, some must be disconsidered from analysis: first, because Instagram use may not be widespread in the region, and geotagged media from there may be solely due to travelers and tourists from other countries; second, because even if this is not true, the low sample size for the country will yield estimates that are too far from their true value.

Thus, a cutoff point was determined. A cutoff as low as 10 medias is possible and will likely include all possible countries, but the number is too small for reliably

[6] `http://www.gadm.org/about` (Visited Jan 2016)

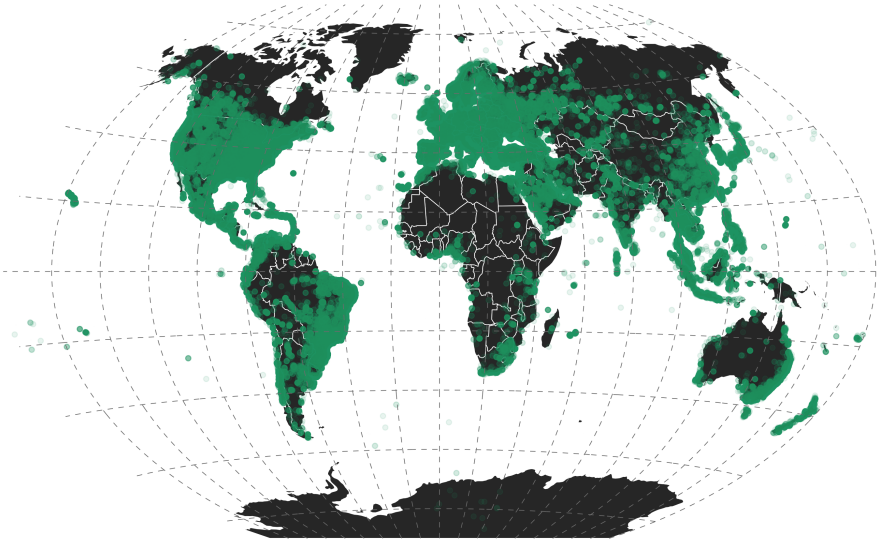


Figure 2.4: A random sample of 1M of the geotags mapped to their position in the world. The data points were set to transparency with an alpha value of 0.01 to highlight areas where they overlap.

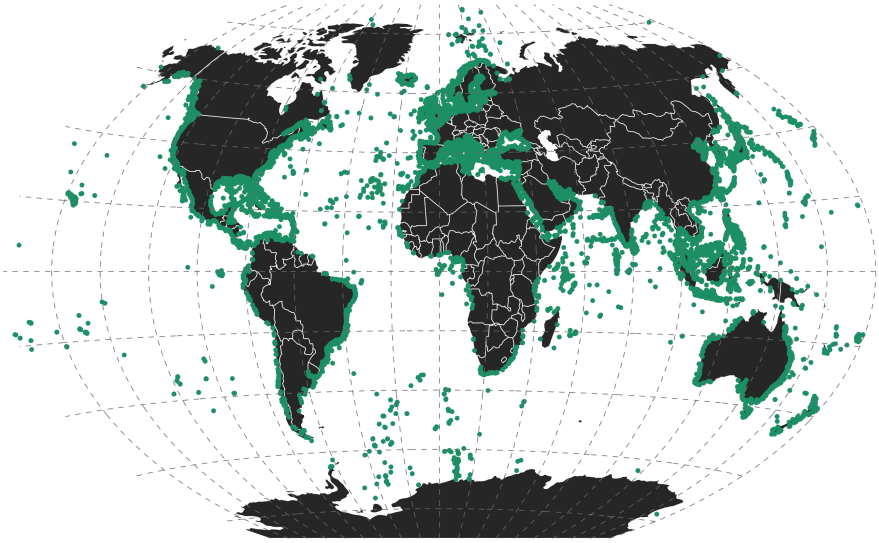


Figure 2.5: Location of geotags that could not be mapped to any country, with no transparency.

estimating any measure of interest. On the other hand, a cutoff of at least 1000 medias will exclude a big number of countries that can have interesting information. A cutoff of a minimum of 100 medias proves to be a reasonable middle term, in which only a small set of countries are excluded, and the sample size is big enough so as to avoid noisy estimates. Figure 2.6 shows the set of countries that are eliminated with a cutoff

of at least 100 medias, and shows that most of the world's countries are still included.

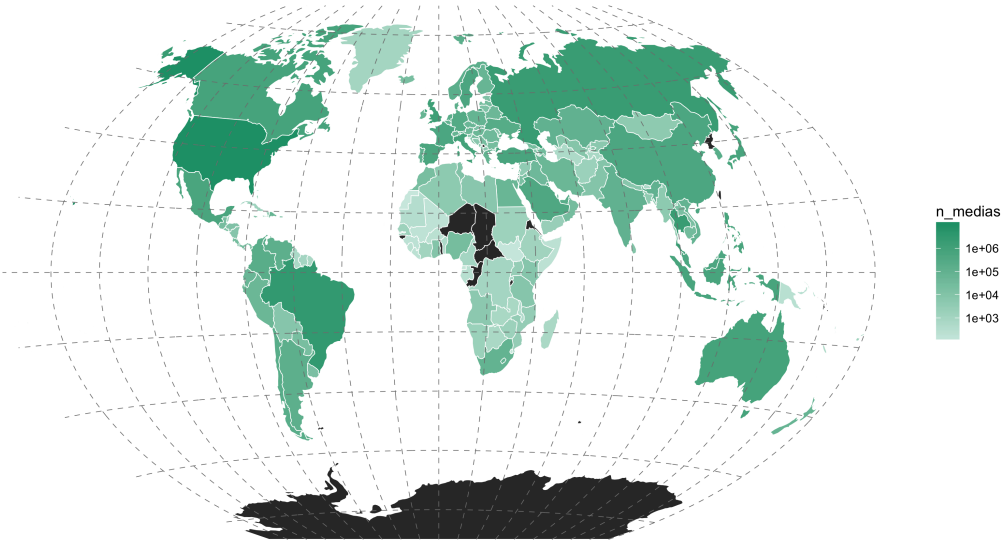


Figure 2.6: Countries included (green) and excluded (dark gray) when the cutoff value is set to 100. The "darkness" of the included countries is proportional to the amount of medias located in that country, log scaled.

2.4 Distribution of Instagram's attributes

A well-grounded quantitative analysis of Instagram must take into account how the attributes are distributed. OSNs, and social systems in general, have important characteristics that rule these distributions.

First, OSN attributes are sometimes subject to arbitrary cutoffs that distort the distribution. A social network service has limitations related to network bandwidth and data storage that pushes their developers to impose reasonable limits to their usage. The best known example is Twitter, which imposes a 140 character limit to all tweets (posts) produced in the network. Another example is Facebook, which limits user profiles to at most 5000 friends. These limits can sometimes be lifted after the network has grown and is able to support a better infrastructure, but the effects on the distributions take time to fade away.

Second, social graphs are known to exhibit heavy-tailed distributions in its node degrees [Mislove et al., 2007]. The "tail", in this sense, is the probability of observing a very large value (relative to the expected value), and it is considered heavy if there is a significant amount of probability density in it – that is, extremely large values are likely enough to appear even in a small set of observations. For us, this means that, for

some attributes, the majority of users (or medias) will have small to moderate values, but a few will have extremely high values, and the distribution will be heavily skewed because of that.

There are different distributions that can generate a heavy tail. Here, I will focus in the two most commonly found in OSN modeling: the power law distribution (also known as Pareto or Zipf distribution) and the Log-normal distribution. Other distributions, such as the Weibul, or a combination of distributions, can eventually provide a better fit to the data. However, a precise and rigorous modeling of each attribute in Instagram is beyond the scope of this work. A detailed treatment on the procedures used to estimate the parameters, as well as a definition of the method used for deciding among the two distributions are given in Appendix A.

Table 2.3 shows the results of the MLE estimates for all attributes, along with the best fit and test statistic. Figures 2.7, 2.8 and 2.9 show the log-log plots of each attribute, along with a log-scaled density plot to better describe the probability density in each point. Due to memory restrictions, attributes with more than 10^7 observations had to be downsampled to this number in the density plot (but the log-log plots were generated using the whole dataset). Next, I will briefly examine the log-log plots of each attribute.

Entity	Attribute	x_0	α	μ	σ	$\mathcal{R}$	Best Fit
Media	# likes	37	**2.29**	-674.55	22.91	56.99	Power law
	# comments	8	**2.73**	-655.51	19.51	78.64	Power law
User	# follows	503	2.57	**-0.57**	**2.24**	-17.85	Log-normal
	# followees	401	**2.36**	-933.67	26.23	6.98	Power law
	# media	976	3.23	**2.87**	**1.47**	-4.64	Log-normal
Hashtag	Frequency	4	1.75	**-44.97**	**8.04**	-24.50	Log-normal

Table 2.3: Estimates. $\mathcal{R}$ was standardized by its estimated standard deviation in order to make it comparable between tests. All $\mathcal{R}$ values were significant under $p < 0.001$

2.4.1 Media Interactions: Likes and Comments

Figure 2.7 shows the distributions of likes and comments for all the medias collected. It is not possible to visually discriminate which of the distributions fits better to the data, but the power law yields a better fit in the likelihood ratio test. Although the comments distribution lay reasonably straight in the log-log plot, the likes distribution shows a weaker decay at values between 100 and 5000, and a stronger decay for higher

values. Most of the medias received up to 10 comments and 100 likes, but the plots shows that some media received more than $10,000$ comments and likes.

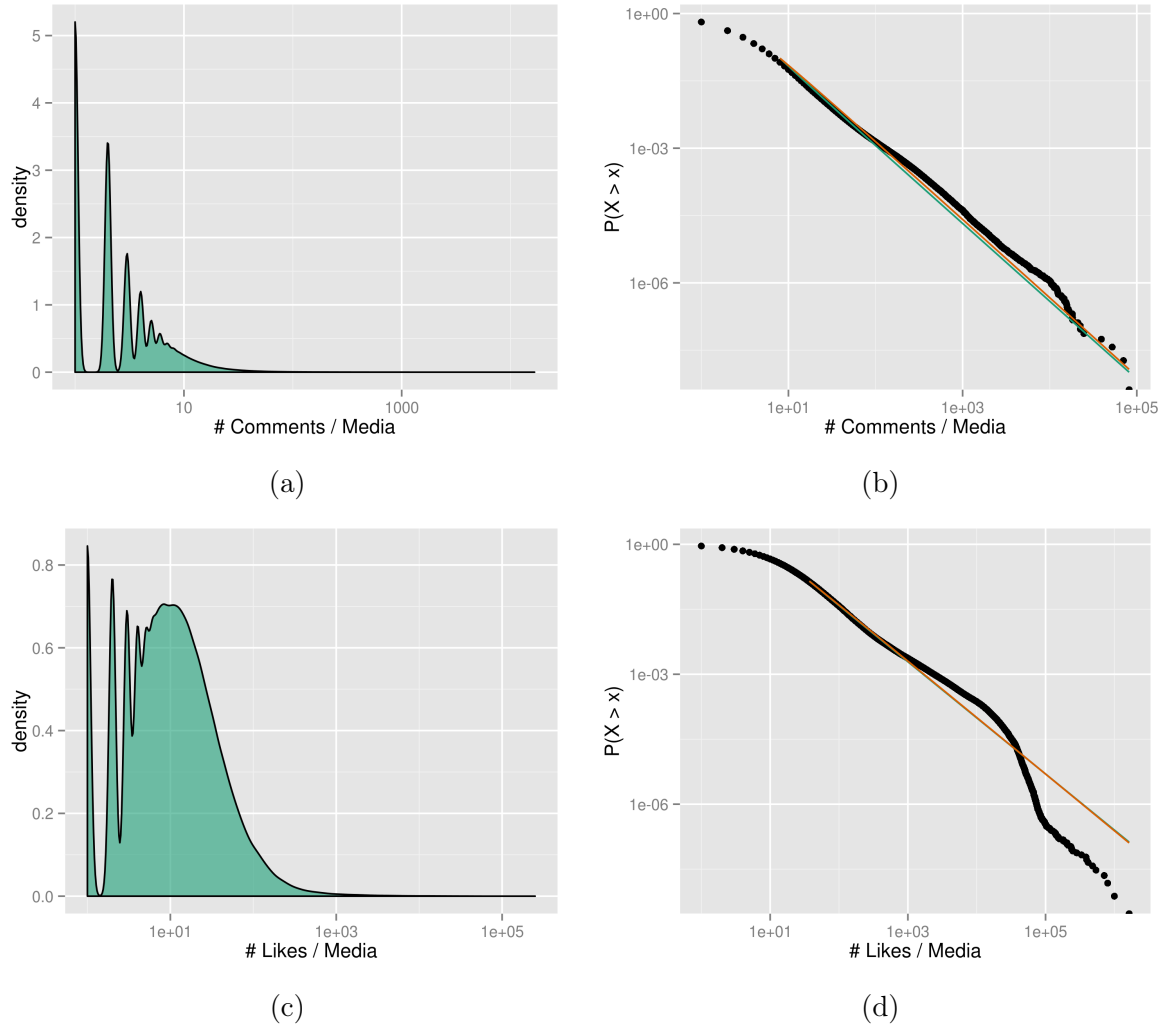


Figure 2.7: Density and log-log plots of the distributions of attributes related to media interactions

2.4.2 Hashtags

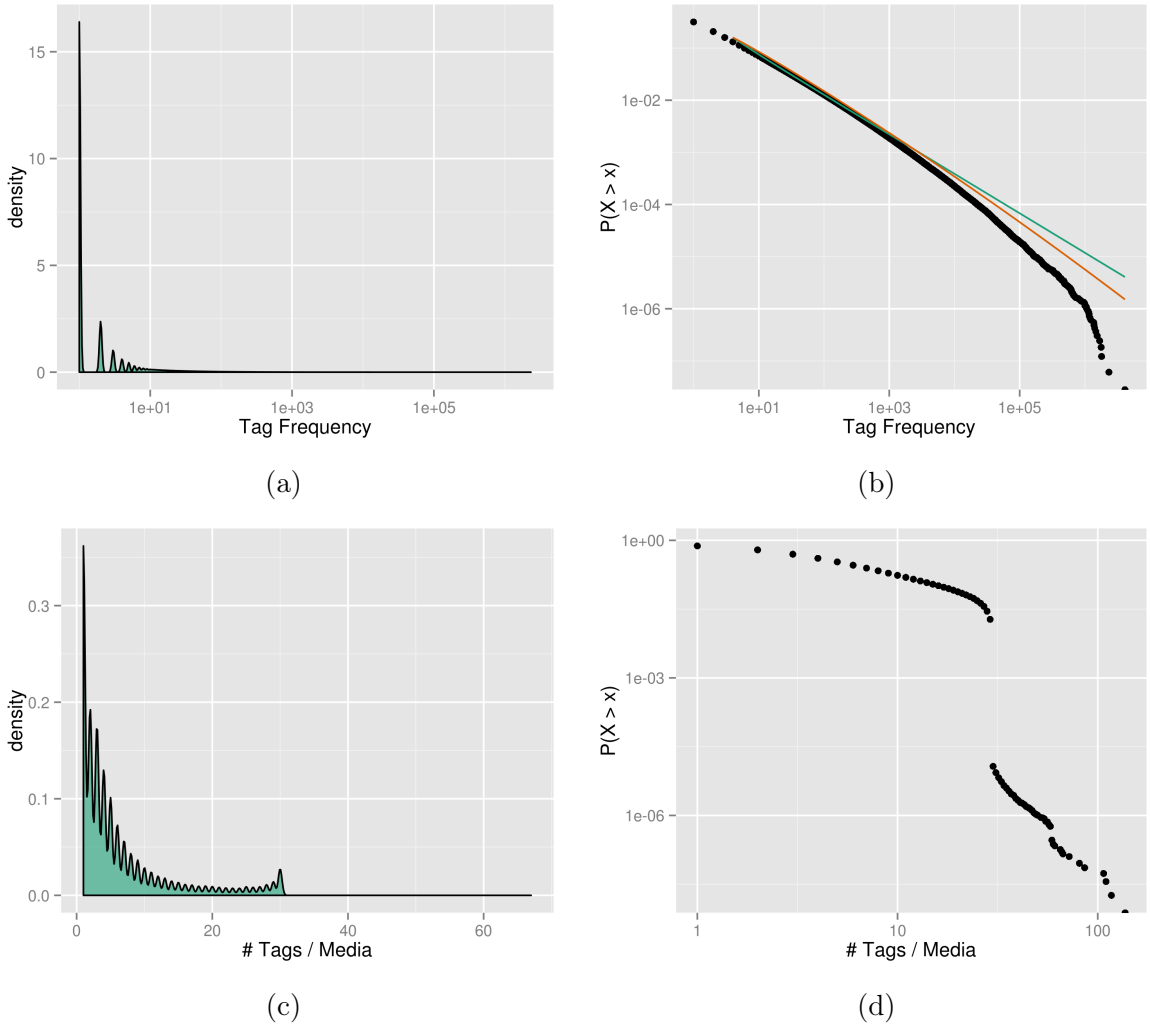
Figure 2.8 shows distributions related to the use of hashtags: the number of hashtags per media and the frequency of each individual hashtag.

The number of hashtags per media has a gap in the probability of values above 30, which is a limit imposed by Instagram. Numbers above 30 are actually due to "cheats", *e.g.* deleting a hashtag after it was listed[7].

[7] `http://www.justin.my/2012/05/instagram-hashtags-cheat-and-tips/` (Visited Jan 2016)

The hashtag frequency has a stronger decay than each of the two proposed distributions, but can still be fairly well approximated by a log-normal distribution.

Although the comments distribution lay reasonably straight in the log-log plot, the likes distribution shows a weaker decay at values between 100 and 5000, and a stronger decay for higher values. Most of the medias received at up to 10 comments and 100 likes, but the plots shows that some media received more than 10,000 comments and likes.



(a) (b)

(c) (d)

Figure 2.8: Density and log-log plots of the distributions of attributes related to the use of hashtags. Since the number of hashtags per media does not follow a power law, the x axis of the density plot is not log-scaled.

2.4.3 User Attributes: Followers, Followees and Posts

Figure 2.9 shows the distributions of followers, followees and posts.

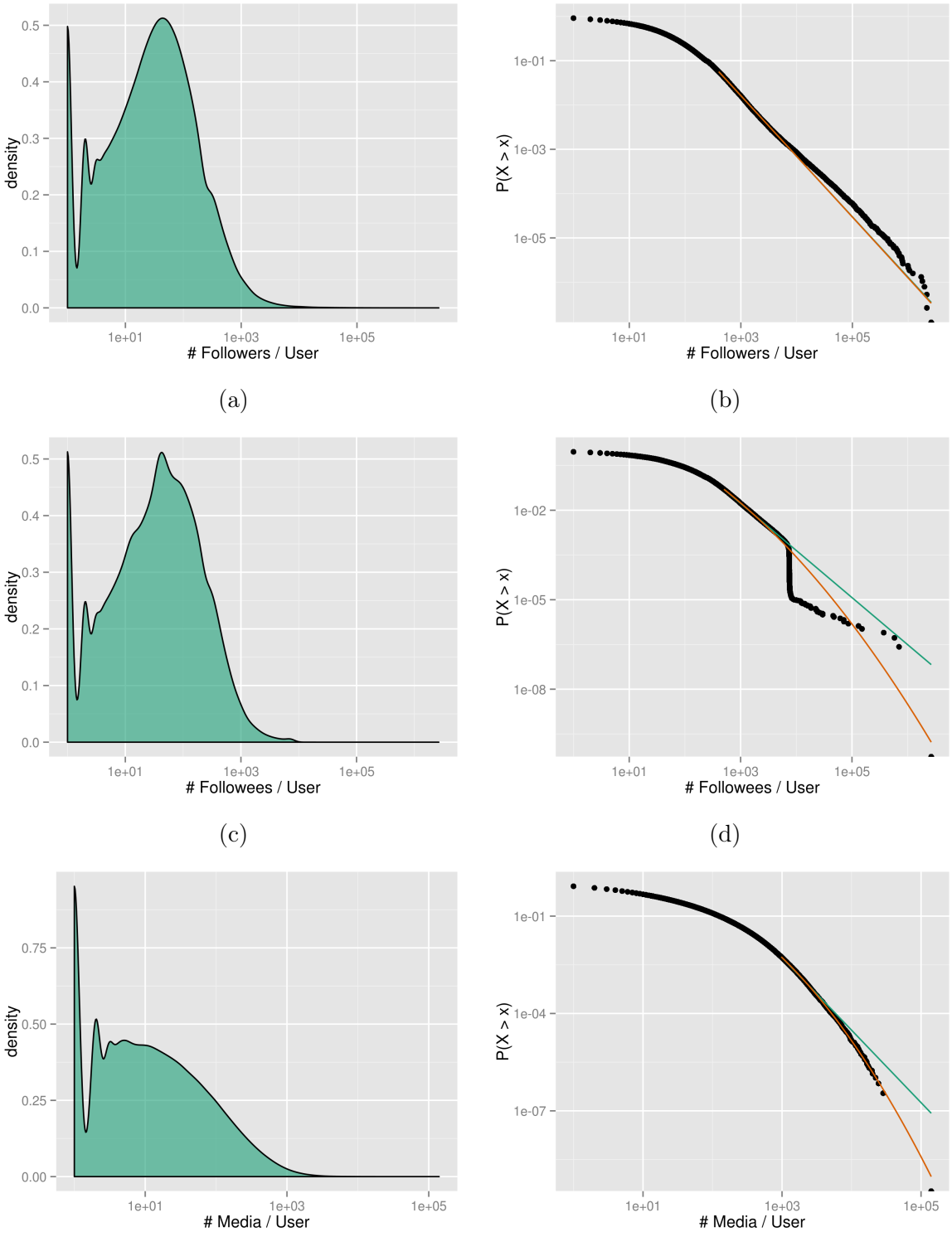
The number of followers has a well behaved power law tail, although with a weaker decay than expected for high values. Again, a log-normal distribution is visually indistinguishable for a power law, but the fit for a power law is better. It can be seen that around 99% of the users have less then 1000 followers, but some users more than a million.

As with the number of hashtags, the number of followees also has an unusual shape that can be explained by the service's limit policies. Instagram established a limit of $7,500$ users in June 2012 due to increased spam after its acquisition by Facebook. However, users who already were above this limit were not affected by the change, and can still follow unlimited users[8]. Interestingly, the log-normal fit shows a possible projection of how the attribute would be distributed if this limit were not set.

As with followers, 99% of the users follow at most 1000 other users, but even with Instagram's limits this count can get to a million.

The number of posts has a curved shape that fits a log-normal distribution.

[8] `http://ubuntulife.net/instagram-follow-limit-you-cant-follow-anymore-people/` (Visited Jan 2016)



(a) (b) (c) (d) (e) (f)

Figure 2.9: Density and log-log plots of the distributions of attributes related to user interactions

Chapter 3

Face Recognition

3.1 Attribute estimation from face data

Computer Vision and Biometrics researchers have been historically greatly interested in extracting information about faces in image data. Although there is no agreed-upon, cross-disciplinary terminology, this topic is commonly referred to as Face Recognition. It normally involves three steps [Huang et al., 2007]: (1) a **detection** step, in which faces are highlighted in the picture through a face detection algorithm; (2) a **normalization** step, in which faces are transformed (rotated and re-scaled) so that all faces are in a standard position; and finally, (3) a **classification** step, in which the aligned face is converted to a vectorized representation and assigned a class depending on the task at hand. This task is normally an **identification** task, in which a face is mapped to a member of a known set of persons (or none, in case of *open set* tasks), or a **verification** task, in which a decision is made of whether two faces represent the same person [Zhao et al., 2003].

Note that some researchers reserve the term **face recognition** exclusively for the classification step. However, the classification task can also use sets of labels that are not linked to individual identities. For clarity and simplicity, I will use the term "Face Recognition" meaning the task of, given a face, outputting a label related to the person to which the face belongs. A system that solves this task will be called a Face Recognition System (FRS).

In a research setting with benchmark datasets, normally the detection step is disconsidered – the regions with faces in the picture are already given with the data. However, there is a difference between extracting facial features for classification in **constrained** and **unconstrained** settings. In constrained settings, the extraction problem is considerably simpler – faces in input images are expected to be well-aligned,

so the alignment step can (also) be disconsidered, and one can expect that background, lighting conditions and facial expressions will be controlled. Unconstrained settings, exemplified by family pictures and newspaper photos, demand systems that are invariant to scale, rotation, alignment, different facial expressions, etc. FRSs that manage to overcome these challenges are able to cover a much wider range of situations: while constrained situations can be assumed in tasks such as extracting information from passport pictures and mugshots, they cannot be assumed for most photographs and pictures taken from videos. Therefore, a high performance in unconstrained settings is highly desirable.

In the present work, I will explore a task that is related, but distinct from identification/verification: **Attribute Estimation**. It differs from the former in that the outcome of the system is not a decision over the face's identity, but over describable aspects of visual appearance [Kumar et al., 2009]. More specifically, here it is either the age or the gender of the person.

Although recent work has tried to estimate gender and age using an unique method [Han and Jain, 2014; Kumar et al., 2009], historically, these two tasks have been approached differently.

3.1.1 Perceived gender classification

Deciding whether a face belongs to a male or female can be posed either as a two-class problem (`{Male,Female}`) or a one-class problem, when one gender is assumed in positive cases and the alternative in negative cases. Two considerations must be made. First, I chose to use the term *Perceived Gender Classification*, as the class is related to which gender the face *appears* to be and has nothing to do with a self-assigned gender identity. Second, it should be noted I am aware of no study that takes into consideration other possibilities besides male or female, in spite of the assumption of binary gender roles having been criticized in the past decades [de Las Casas et al., 2014].

Normally, classification is done by extracting features using standard image descriptors – such as SIFT, HOG, Local Binary Patterns, Gabor features, Biologically Inspired Features and Color Histograms [Santarcangelo et al., 2015] – and feeding it to a state of the art classifier such as Support Vector Machines or Random Forests. Some methods use dimensionality reduction techniques, such as Independent Component Analysis or Principal Component Analysis [Santarcangelo et al., 2015], while others rely only in regularization to control for model complexity.

3.1.2 Age estimation

The problem of visually estimating the age of a person through her face can be understood as either a classification problem, in which each age group is treated as a category and a binary encoding is used, or a regression problem, when age is treated as a real valued outcome problem [Levi and Hassner, 2015]. When taken as a regression problem, a common metric of performance is Mean Absolute Error(MAE):

$$MAE = \frac{1}{n}\sum_{i}^{n} \|x_i - y_i\|$$

where x is the predicted age of the i-th face, y is the true age of the i-th face, and n is the number of faces in the test set. An alternative metric, the Cumulative Score (CS), considers a success if the predicted age is within a range from the true age and calculate its accuracy (*e.g.* . a $CS_5(x)$ will calculate the proportion of predictions that missed the true age by 5 years or less).

Early methods of representing faces for age estimation are based on the layout of facial features (eyes, nose, mouth) in the face region. This proved to be unsuitable for unconstrained settings, as accurately finding facial features in these environments is a challenge by itself [Levi and Hassner, 2015]. Moreover, proportions between facial elements give considerably less information about age after adulthood, when shape changes become less prominent and texture changes start be more relevant [Fu et al., 2010; Guo et al., 2009]. An alternative approach was to model the aging process as a subspace that aggregates individual aging patterns encoded by a model, or as a manifold, that learns a low-dimensional trend from faces with different ages. However, these two approaches are limited due to problems of generalization to faces different from the ones in the database, and also due to bad performance for faces in unconstrained settings [Levi and Hassner, 2015].

Finally, a number of image descriptors combining local and global information of shape and texture were suggested [Levi and Hassner, 2015; Fu et al., 2010; Guo et al., 2009].

3.1.3 Performance in Face Recognition

The *de facto* standard benchmark dataset for face verification in unconstrained settings is the Labeled Faces in the Wild (LFW) dataset [Huang et al., 2007]. It is composed of 13,233 annotated pictures gathered from news articles on the Web of 5,749 different individuals, 4,069 of whom have just one picture assigned to them. Besides the public

availability of performance metrics for many algorithms[1], the dataset features the performance of human annotators, contributed by Kumar et al. [2009]. Recently, a number of advanced algorithms surpassed human annotators [Taigman et al., 2014][2] and virtually reached a plateau in LFW [Zhou et al., 2015], achieving over 99% accuracy. Two major factors that collaborated for this were the development of sophisticated Deep Neural Networks (DNNs) and the availability of external, big datasets (*i.e.* with millions of faces) for training the models.

DNNs are machine learning systems that can be understood as many interconnected layers of parallel processing units (called *neurons* or *units*). They are high capacity classifiers, which means that they can learn extremely complex data transformations if they are trained with enough data. This allows them to learn highly discriminative features from raw input, which enables the modeler to avoid engineering domain-dependent features – such as the aforementioned biologically inspired features made for face processing. Moreover, they can be architectured in a way that handles image data extremely well. More specifically, Convolutional Neural Networks are especially powerful for computer vision tasks [Krizhevsky et al., 2012]. All these properties make DNNs especially fit to be used for Face Recognition tasks. However, in order to explore all the potential of DNNs, they must be trained with a huge number of labeled data – in the order of millions.

Creating this kind of dataset through manual annotation would be extremely expensive. Fortunately, the rise of online social platforms allowed for a huge influx of publicly available, semi-annotated image data from millions of internet users. Although the indiscriminate collection of this data may pose ethical concerns, it has become an increasingly common practice in order to circumvent the need for data for high capacity algorithms such as DNNs [Taigman et al., 2014; Zhou et al., 2015]. This, however, means that high-performance face recognition is increasingly dependent on mass collection of data.

3.1.4 Bias in Face Recognition

Bias is an important, but ambiguous concept. There is a number of definitions for it, depending on which kind of literature is taken into account.

In Machine Learning and Statistics, bias represents "the systematic difference between a random variable and a particular value" [James, 2003]. For example, sam-

[1] `http://vis-www.cs.umass.edu/lfw/results.html` (Visited Jan 2016)

[2] According to the authors, human performance is about 97.5%. It must be noted, however, that this is only the case for tightly cropped faces. When the full picture is shown, humans get a 99.2% accuracy.

pling bias means deviating from a random sample of the population, a biased estimator is a method of calculating estimates that systematically over- or underestimates the quantity of interest, and a learning algorithm with high bias is one that systematically misses its target, irrespective of the amount of data used to train it.

The simple definition of bias according to the Miriam-Webster dictionary website is "a tendency to believe that some people, ideas, etc., are better than others that usually results in treating some people unfairly"[3]. This definition shows the importance of considering fairness when working with a concept of bias that relates to social matters.

The concept of **algorithmic bias** has recently emerged in the literature. When describing a framework to understand algorithmic bias, Friedman and Nissenbaum [1996] define that "a system discriminates unfairly if it denies an opportunity or a good or if it assigns an undesirable outcome to an individual or group of individuals on grounds that are unreasonable or inappropriate". Again, the authors point out that the discrimination alone cannot be considered bias unless it occurs systematically, *i.e.* it follows a consistent pattern based on attributes of the system or the environment in which the system is embedded. Note that their definition lacks a proper description of what are "reasonable and appropriate" outcomes. For my work, it suffices to consider that the outcome is the probability of the FRS yielding a correct result for a given individual, as most applications that use a FRS rely on its accurate performance to make decisions. Thus, I will consider that a FRS is biased if its performance is consistently different for different groups of individuals in a manner that is not publicly disclosed and justified by the system's engineers. Of especial interest here is bias of FRSs based on different age groups and genders.

A number of factors can impact the performance of a FRS. Depending on the algorithm used for recognition, visual factors, such as pose and illumination have a strong influence in identification/verification performance [Zhao et al., 2003]. However, the high degree of success on the LFW datasets suggests that these factors are impacting progressively less the algorithms. In contrast, external factors, such as factors related to the demographical distribution, still seem to affect state of the art algorithms.

The most comprehensive study about this is the series of reports produced by the American National Institute of Standards and Technology using the Face Recognition Vendor Test (FRVT) dataset. The FRVT is a large-scale project created to test commercial and prototype-level academic FRSs in constrained settings using over 7 million images taken from passports and mugshot images from the United States government. Although the dataset is not publicly available, they allow companies to

[3] `http://www.merriam-webster.com/dictionary/bias` (Visited Jan 2016)

subscribe to their tests and publish their results in open reports, along with comparisons with unconstrained datasets such as the LFW and others. The latest version of the dataset is from 2013, and reports were made separately concerning Age Estimation, Gender Classification and Face Identification. The dataset format is especially suited for investigating bias due to demographic profile: since it is from a constrained setting, confounding factors are unlikely to play a role. Moreover, since it uses government-certified data, it has very reliable annotations containing each face's true age, gender and nationality.

Their reports present interesting findings. The following effects are observed in all algorithms tested, unless explicitly noted:

Biases in Gender Classification [Ngan and Grother, 2015] All algorithms have a good overall performance, with the most accurate correctly classifying the gender of a person 96.5% of the time. Males have their gender classified with more accuracy than females in all age bands. The peak of classification accuracy for females is for young adults (21-30 age band), while for males is for older adults (31-60 age band). For males, the performance stabilizes after 20 years old and only drops slightly after 81 years old. However, for four of the nine algorithms, the performance for the 0-10 age band for males is significantly worse than for older ages. In contrast, for females, performance quickly reaches the peak for young ages, but drops steadily after 40 years old, reaching its worst accuracy at the 81-90 age band (with an average accuracy of 57% across the algorithms, almost random guessing). Finally, the performance for males is the most affected in unconstrained settings.

Biases in Age Estimation [Ngan and Grother, 2014] Algorithms have a satisfactory performance, but with room for improvement, with the most accurate algorithm yielding a MAE of 4.3 years and correctly estimating the age of 63% of the participants within 5 years. The report shows that age estimation is not independent of the target age group: different algorithms show different patterns of precision in estimating age across age groups. Most algorithms are most precise estimating the age of adults (18-55), which is also the most operationally relevant age group. Some algorithms are especially good at estimating the age of kids and teens (0-18), some are especially bad for this group, and most algorithms show the highest estimation error (accuracy and MAE) in the senior age group (56-99 y.o.). Moreover, the estimation error is not centered at zero in all cases, which means that the algorithms systematically overestimate or underestimate certain groups. More specifically, the seniors tends to have their age underestimated,

while the youth age group tend to have its age overestimated. Women have their age consistently underestimated in all age groups. Finally, they show that ethnicity has an impact on age estimation: South Americans tend to have their age overestimated, and Asians underestimated.

Besides the FVRT reports, other researchers also point out the presence of age/gender bias in FRSs:

- Guo et al. [2009] show that gender classification accuracy is 10% higher for adult faces than for young or senior faces when using a SVM classifier and biologically inspired features. However, they use an unbalanced dataset, with twice the number of adult faces than for the other age groups.

- Dago-Casas et al. [2011] train SVM and LDA classifiers in LFW and in the GROUPS dataset (see Section 3.3.1) and show that age estimation accuracy for adults (20-65) is much higher than for other age groups. They point out that this effect can be at least partly due to adults being much more prevalent in their training set than the other groups.

It can be seen that FRSs do show signs of algorithmic bias, and specifically to minorities – women, seniors and US immigrants. It is also noticeable that some of this bias is consistent across many different algorithms. The reasons for this are yet unclear.

One possibility is an inherent limitation of facial features to estimate such attributes. For example, the precise age of a person may not be as clearly discernible by her facial features as she gets older. If this is the case, then there is an upper bound in performance that cannot be overcome by any FRS that relies only on facial features. In fact, if humans were to be considered a "gold standard" of face recognition that manages to extract all possible information about age and gender that a face possesses, one could argue that human performance is heavily biased, especially towards age [Voelkle et al., 2012].

However, it may also be possible that exogenous factors are at play. First, FRS engineers may not be economically motivated to assure that their algorithms are not biased for people whose demographics are not commercially or operationally relevant for their services. Moreover, the training sets used in the FRSs may be unbalanced for some specific demographic groups. As mentioned in Section 3.1.3, state of the art FRSs often rely on data collected on the Internet to generate the algorithms' training set. This means that the gender and age distribution of the faces in the training set is the distribution available in the Internet. Moreover, to find data that is already reliably

labeled, engineers will often restrict to well known faces – such as pictures of celebrities, a cohort with very specific demographics. The LFW Dataset is an example of such Internet collected database, and its gender distribution is heavily skewed towards males (70% of the faces are male).

Although the precise source of algorithmic bias is a delicate subject and beyond the scope of this work, it is possible the access the presence/absence of such bias. This can be done even when the internals of the FRS are not accessible. As Diakopoulos [2015] points out:

> Algorithms are often described as black boxes, their complexity and technical opacity hiding and obfuscating their inner workings. At the same time, algorithms must always have an input and output, two openings that can be manipulated to help shed light on the algorithm's functioning. It is not essential to understand the code of an algorithm to begin surmising something about how the algorithm operates in practice.

Thus, it is possible – even desirable – to use an off-the-shelf, widely adopted algorithm, even if its implementation is proprietary. What matters in evaluating algorithmic bias is the relationship between input and output.

In the next Section, I will describe the algorithm I chose for this work, `Face++`.

3.2 Attribute estimation with `Face++`

`Face++` is a FRS with a publicly available web service based in China with an endpoint in the USA. It is trained in millions of images downloaded from the Internet. As of 2015 they match the state of the art in face recognition, with 99.5% accuracy in the LFW dataset [Zhou et al., 2015]. `Face++` has been used in a number of publications [Bakhshi et al., 2014; Redi et al., 2015b,a; Jang et al., 2015] and boasts dozens of partnerships with big companies like Intel and Lenovo[4].

All `Face++` services are available through requests to specific REST endpoints defined in its Application Programming Interface (API). After creating an account on its website, it is possible to create new applications to have access to the services. A `Face++` API application is analogous to an Instagram API client and receives an API key and an API secret that are also used for authentication.

In this work, I used only one of its resources, located at the endpoint `/detection/detect`. A GET request pointing to an image URL and appropriate

[4] `http://www.faceplusplus.com/` (Visited Jan 2016)

parameters returns a JSON-formatted text with the position, pose and attributes of all faces detected in the image.

The attributes returned are age, gender, race, whether the person is smiling and whether the person is wearing glasses. In this work I will focus only in age and gender.

An example query would be:

```
https://apius.faceplusplus.com/v2/detection/detect?url=http%3A%2F%2Ff
aceplusplus.com%2Fstatic%2Fimg%2Fdemo%2F1.jpg&api_secret=YOUR_API_SEC
RET&api_key=YOUR_API_KEY&attribute=pose,gender,age
```

The precise method used by Face++ for attribute estimation is not clear, although it is likely built on top of its patented DNN face representation technology [Fan et al., 2015]. The format of its output, however, provides some information on some characteristics of its estimation method.

For the perceived gender classification, their API outputs a gender value ({Male,Female}) and a confidence value ranging from 50 to 100. With this information, one can assume that they model the problem as a binary classification task with a continuous outcome, in which values above 50 are set to be from one class and values under 50 are set to be from another class.

For age estimation, their API outputs an integer age value and an "age range" that takes discrete values from 4 to 18. There is no information on how this range is calculated or to what it refers. It is possibly an interval related to the estimation error, but since there is no documentation[5] and no clear interpretation of how it is meant to be handled, I decided to ignore it.

3.2.1 Face++ performance

In its website[6] Face++ reports more than 96% accuracy, but have no official benchmarks for age estimation or face detection. However, existing work by independent research teams provide some clues. Bakhshi et al. [2014] use crowdsourcing to validate the Face++ results in its ability to detect at least one face when there is one (97% accuracy), the ability to correctly classify the gender of the faces (96% accuracy for both genders) and its ability to classify a face in one of three age groups: under 18-, 18-34 and 35+.

Yadav et al. [2014] compare Face++ results with that of human raters. They find that accuracy is very high for humans and Face++ in 0-5 and 6-10 age groups, and then drops significantly. They also find that female faces have their age more easily

[5]The API developers also did not respond to contact.

[6]http://www.faceplusplus.com/tech_gender/ (Visited Jan 2016)

estimated by either gender, which goes in the opposite direction of what was reported with other algorithms in the FRVT tests.

3.3 Validation of Face++'s output

Since Face++ is proprietary and does not offer detailed information about its performance in attribute estimation, it is important to validate it with ground truth data to access how it can be biased. Besides providing an objective way to access the algorithm's behavior, a ground truth can also yield information that can be used to **calibrate** Face++'s gender confidence score, yielding a powerful tool to measure the uncertainty of the perceived gender classification for out of sample data.

There is no source of information in Instagram that can be reliably used as a ground truth for this task. A good alternative is to use benchmark datasets with age and gender labels. Unfortunately, although there are labels available for the LFW dataset, their labeling methodology is not well specified. Thus, I recurred to another well known dataset presented by Gallagher and Chen [2009]: the GROUPS dataset.

The GROUPS dataset is composed of images of groups of people in unconstrained environments, and is well established in the literature, despite it being relatively new. It is built from a collection of 5080 Flickr Images containing 28 231 faces, all labeled by human annotators using crowdsourcing. It uses discretized age labels and binary gender labels. The dataset is very well balanced for gender: it has 13 445 (52.3 %) female faces and 12 273 (47.7 %) male faces. The age label distribution is skewed towards adults, as can be seen in Table 3.1.

Age Band	[0,2]	(2,7]	(7,13]	(13,19]	(19,36]	(36,65]	(65,99]
Frequency	757	1440	790	1560	13 893	6193	1085

Table 3.1: Number of faces in each age band in the GROUPS dataset.

3.3.1 Evaluation Method

In order to evaluate Face++'s performance, I fed all pictures of the GROUPS Dataset to Face++ using its API. The coordinates of the faces found by Face++ were matched to a corresponding face in the Dataset's Ground Truth (GT) using a point matching method:

- For each image, I found the midpoint between the eyes of each face for both the GT and Face++, generated all possible pairings between the GT and F++

sets, sorted them by their euclidean distance, and selected all non-overlapping pairs, starting from the pair with smallest distance. Here "non-overlapping" pairs means those that did not share a member with another selected pair (*i.e.* with a smaller distance).

- I excluded the pairs that were above a given distance threshold. Given that the images were of varying sizes, this threshold was set to be 10% of the maximum possible distance between two points in the image, *i.e.* $\sqrt{w^2 + h^2}$, where w is the image's width and h is the image's height. The number 10% was found empirically. 53 pairs were excluded in this manner.

For each pair of (F++, GT) face, a record was generated containing the attributes estimated by `Face++` and the attributes stated in the GT. With this I managed to match 25 752 of the 28 231 faces (91.21 %). This is a reasonable detection accuracy compared to the state of the art on unconstrained datasets[7]. 270 faces found by `Face++` could not be matched. By manual examination I could observe that many of those were treated as background by the annotators of the dataset.

Overall accuracy for gender and age were 88 % and 65 %, respectively[8]. Performance for gender did not differ substantially for males and females.

To evaluate `Face++`'s performance in age estimation I grouped all age values that were inside one of the age bands specified by GROUPS as belonging to that age band. Table 3.2 shows the Confusion Matrix of the errors for age group estimation, separated by gender. It is important to highlight that the ground truth for these estimates is hand labeled and is subject to error – especially in the age estimation task.

The pattern of error is similar in both genders, although slightly higher for males in all age bands. The adolescent (`(13,19]`) and senior (`(65,99]`) age groups had worse performance in both genders, and kids (`[0,2]`) and young adults (`(19,36]`) had the best performance.

Most age groups have their age significantly underestimated in both genders: the youth (`(7,13]`), seniors and old adults (`(36,65]`). However, adolescents are strongly overestimated: in fact, more adolescents were estimated as young adults than estimated as adolescents.

This pattern of error may be due to the arbitrary boundaries established by the GROUPS labelers. For example, the difference between someone with 19 and 20 years is smaller then that of someone between 20 and 30, but in this evaluation framework

[7] `http://vis-www.cs.umass.edu/fddb/results.html` (Visited Jan 2016)

[8] Notice that the probability of the algorithm getting a right answer by chance is 50% and 14%, respectively.

Gender	Predicted Band	[0,2]	(2,7]	(7,13]	(13,19]	(19,36]	(36,65]	(65,99]
Female	[0,2]	**0.75**	0.14	0.02				
	(2,7]	0.21	**0.56**	0.30	0.03	0.01	0.01	
	(7,13]	0.02	0.22	**0.44**	0.20	0.06	0.03	0.01
	(13,19]	0.01	0.04	0.10	**0.29**	0.17	0.06	0.02
	(19,36]	0.01	0.03	0.14	0.45	**0.66**	0.47	0.14
	(36,65]				0.03	0.11	**0.42**	0.67
	(65,99]						0.01	**0.16**
Male	[0,2]	**0.80**	0.21	0.03				
	(2,7]	0.14	**0.55**	0.32	0.05			
	(7,13]	0.04	0.16	**0.43**	0.20	0.03	0.01	
	(13,19]	0.01	0.04	0.11	**0.29**	0.10	0.03	
	(19,36]	0.01	0.03	0.11	0.43	**0.68**	0.36	0.06
	(36,65]			0.01	0.03	0.18	**0.58**	0.72
	(65,99]						0.02	**0.22**

Table 3.2: Proportion of faces classified in each age band. Zero valued-cells were omitted for readability, and boldface values are correct results.

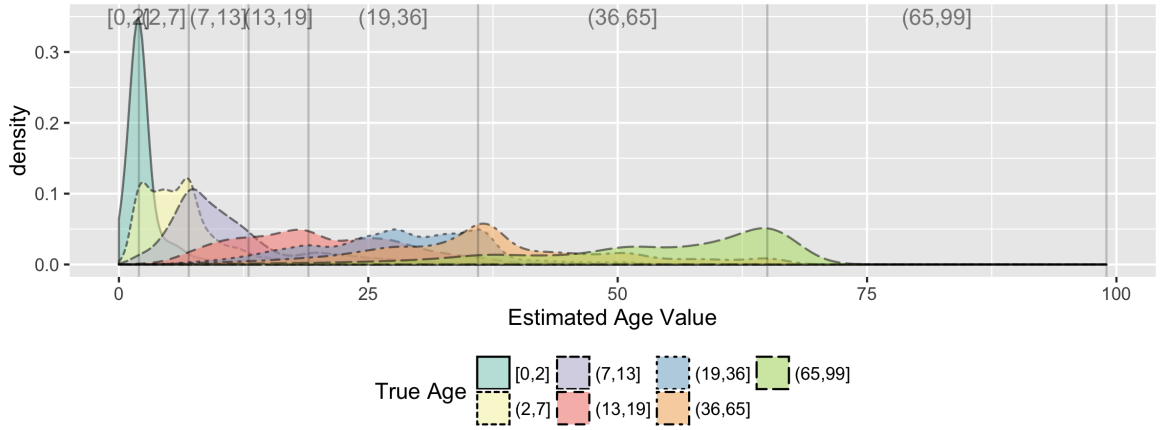


Figure 3.1: Density plot of age values separated by its true age group. Vertical gray lines separate each of the groups. Note that the fact that the density is less spread in lower age bands may be at least partly due to smaller ranges between the cutoff points in these bands.

the former would count as a mistake, while the latter would not. Indeed, the good performance in the adult age groups can be accounted at least partially by the fact that the age bands are much wider. Moreover, the fact that most of the errors were in the neighbor class suggests that collapsing some of the age bands into broader bands would greatly increase performance.

Unfortunately, fine grained age annotation is not available in GROUPS, so it

is not possible to use better measures of error, such as MAE or CS[9]. However, it is possible to visualize the spread of the age values for each band. This can be seen in Figure 3.1. Indeed, most of the density curves of the age groups are centered near the lower cutoff points of the age bands, which confirm our three main conclusions: (1) Face++ systematically underestimates most of the age groups, especially the youth, seniors and old adults, (2) Face++ fails to accurately estimate the 13-19 age band and, (3) the arbitrary cutoff points lead to arbitrary accuracies

With this in mind, I propose merging the categories (2,7] with (7,13] and (36,65] with (36,99]. Judging by the accuracies alone, it would be interesting to merge the (13,19] band to the (19,36]. However, this would throw away important information, since these two are the primary demographics in Instagram. Thus, I will treat them separately, with a caution note that Face++ cannot reliably separate these two age bands for individual faces. This new subgroup yields accuracy estimates stated in Table 3.3.

Predicted Band	[0,2]	(2,13]	(13,19]	(19,36]	(36,99]
[0,2]	**0.78**	0.12			
(2,13]	0.20	**0.75**	0.24	0.05	0.02
(13,19]	0.01	0.07	**0.29**	0.14	0.04
(19,36]	0.01	0.06	0.44	**0.67**	0.37
(36,99]			0.03	0.14	**0.56**

Table 3.3: Proportion of faces classified in the newly proposed age bands. Zero-valued cells were omitted for readability.

Moreover, although it is not possible to correctly evaluate the accuracy of age estimation in finer grain, I will also use the raw age values. The evaluation made in this section suggests that the age values are systematically underestimated, except for young adults. However, the discretization of such values throws away important information. It is reasonable to assume that, although the age estimates will individually show a high degree of error, they will on average approach a value close to the actual age, with a bias to lower values. When this is taken into account, analyses can yield important insights.

[9]These measures were briefly reviewed in Sec. 3.1.2

3.4 Calibrating the gender confidence

In binary classification, classifiers normally output a class *score*, which can be later discretized to get the predicted class. This score reflects the degree of "confidence" the classifier has over the class. Thus, a low predicted score would mean that the referred observation is not likely from the class.

Classifiers such as DNNs and logistic regressions output a score with a value ranging between 0 and 1, which is commonly understood as a probability of being from the target class. However, this interpretation must be taken with caution. Although this score can, indeed, be interpreted as a probability, it is not guaranteed that this will be the posterior probability of an instance belonging to the class – which is normally what is desired. If this is indeed the case, then this probability score should approximate the proportion of positive instances of a class for each confidence level. Thus, for example, approximately 80 % of the predictions with a score of 0.8 should be positive and 20 % should be negative. When the score of a classifier satisfies this condition, it is said to be **well-calibrated**

Well-calibrated classifiers are of interest because the scores can be directly interpreted. This allows for much more analytical power: if the average score of a classifier for a given situation is p, we can expect that the classifier will miss $n(1-p)$ cases in n.

Although some families of classifiers already output fairly well-calibrated scores, one can enhance the calibration of a classifier by passing its scores through a model trained on external data. There are many methods for classifier calibration, but two of them are the most widely adopted:

Sigmoid Method (Platt Scaling) Pass the output of the classifier to a sigmoid with parameters fitted using maximum likelihood estimation. Let the score of a classifier be s:

$$P(y = 1|s) = \frac{1}{1 + \exp(\alpha s + \beta)}$$

where α and β are the parameters. This is akin to fitting a logistic regression on the classifier score and using the regression model to map any new score to the calibrated probabilities.

Isotonic Method Find a monotonically increasing (isotonic) function that maps the score to the probability scores. This method is more general, as it only assumes that the mapping function is isotonic. Given a training set (s, y), where s is the vector of scores and y is the vector of true classes, the Isotonic Regression

problem is finding the isotonic function $\hat{m}$ such that

$$\hat{m} = \arg\min_{z} \sum_{i} (y_i - z(s_i))^2$$

The isotonic method is less constrained than the sigmoid method, which makes it more prone to overfit [Niculescu-Mizil and Caruana, 2005].

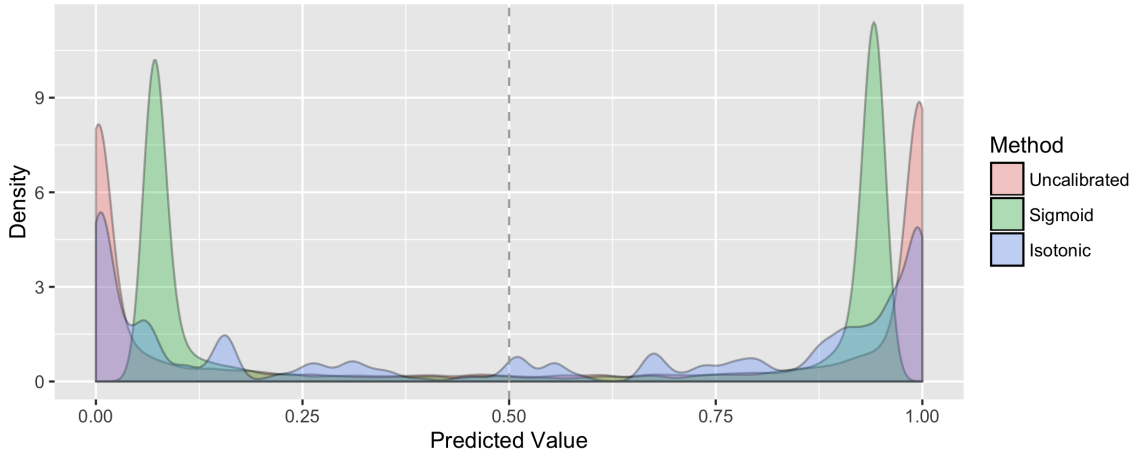
It is only possible to calibrate the gender confidences, as `Face++` does not provide an equivalent metric for age. To do so, I transformed the `(gender value, gender confidence)` tuple into an unique score that would reflect its binary classification score, and thus the predicted posterior probability of a positive class. Hence, the probability $P(y = \texttt{Female}|x)$ is defined as:

$$P(y = \texttt{Female}|x) \approx s = \begin{cases} \frac{c}{100} & \text{if } \hat{y} = \texttt{ Female} \\ \frac{100-c}{100} & \text{if } \hat{y} = \texttt{ Male} \end{cases}$$

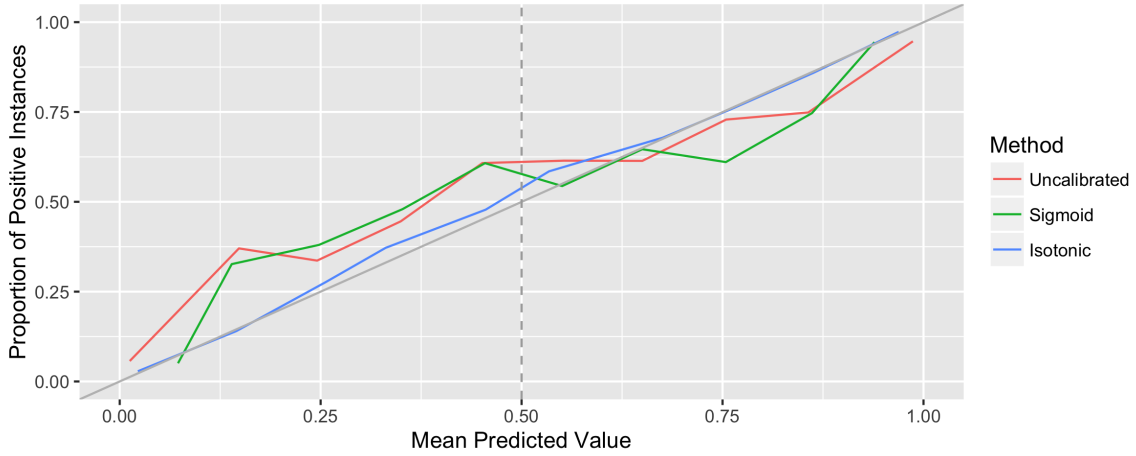
where $\hat{y}$ is the predicted class (`gender value`) of the observation, and c is the confidence score (`gender confidence`) attributed to the classification. The choice of the `Female` class as the positive instance was arbitrary, but does not impact the results. Notice that this transformation is easily undone – if the predicted class is `Female`, then $c = 100 \times s$, otherwise $c = 100 \times (1 - s)$.

Both the Isotonic and the Sigmoid method were used for calibration. The models were trained with the same training set: a random sample of 80% of the GROUPS dataset (the remaining data was used as the test set). To evaluate the performance of each scoring procedure, one must use a Reliability Plot – the scores are discretized into 10 equally sized bins from 0 to 1, and the mean value of the predicted scores in each bin is plotted against the proportion of positive instances of the class in that bin. A perfectly calibrated classifier should produce results that lie in the diagonal line of the plot. Data points above the diagonal line indicate that the model is underestimating the probability of a positive instance for that predicted value (*i.e.* the probability is actually higher), and results bellow the diagonal line indicate the opposite.

Figure 3.2a shows the distribution of scores using each method, while Figure 3.2b shows the reliability plot (both generated with the test set). Most of the scores concentrate in the 90 % (or 10 %) confidence bin. The uncalibrated scores are fairly good at estimating the posterior probabilities, which may be attributed to the learning method employed – neural networks are known to yield well-calibrated results [Niculescu-Mizil and Caruana, 2005]. However, it can be seen that `Face++` systematically underesti-



(a)



(b)

Figure 3.2: (a) Density plot of the distribution of the scores for each method. (b) Reliability plot of the uncalibrated scores of `Face++` and the calibrated scores using two different methods. The scores were transformed to reflect the probability of the face being female. The dashed lines in the plots represent the threshold for deciding whether the face is `female` or `male`.

mates its confidence of the face being male and overestimates the confidence of it being female.

The sigmoid method does not correct this bias. Possible reasons are: (1) the score distribution violates the assumption of normality (conditioned to the class value), as it is highly skewed towards extreme values; (2) As Niculescu-Mizil and Caruana [2005] argue, the sigmoid method is best fit when the reliability curve is sigmoid shaped, and in this case it is shaped as an inverse sigmoid; (3) neural networks normally apply a sigmoid function (or a generalization of it) at its last layer for classification, and this may already leverage all the benefits that the sigmoid method could bring to the

probability estimation.

The isotonic method is able to correct the bias, except for small underestimation errors in the mid-range of the probability estimates. The density plot also shows how it manages to redistribute scores in intermediate values. The errors are probably due to overfit – the isotonic regression has many parameters to estimate, and has been observed to only saturate after tens of thousands of examples in common classification datasets [Niculescu-Mizil and Caruana, 2005].

Due to the superiority of using the isotonic correction to estimate the posterior probabilities of a class, I will use it whenever possible instead of the raw `gender confidence`. To do so, I created a model using all the GROUPS dataset (not only the training set) and generated a mapping function to convert the gender confidence score to its calibrated score.

Chapter 4

Gender and Age in Instagram

Much of the work in the relationship between face recognition and Internet data has focused in accessing the risks incurred by exposing face data in OSNs (*facial disclosure*). Li et al. [2014] suggest that facial disclosure is increasingly common, which can affect the security of systems that use face validation as authentication mechanisms, and propose a method for estimating the risk. A more specific example is offered by Polakis et al. [2012], who present an attack strategy towards Facebook's Social Authentication system that combines automatic face recognition and publicly accessible information on a user's friends list to solve these questions, rendering the user vulnerable to identity theft.

Alternatively, some authors suggest the use of face recognition to enhance user privacy. Xu et al. [2014] propose a mechanism that gives a user control of her personal exposure by automatically identifying photos in which she is involved and giving her management access to these photos. In a similar line of work, Ilia et al. [2015] propose a method for detecting and blurring faces of users who do not want to disclose their information.

Given the richness of context information in OSNs, some authors suggest methods that use the additional data available in social networks as means to get information that enhances the performance of FRSs [Stone et al., 2010; Taigman et al., 2014]. Recently, however, some researchers have followed the opposite trend, and employed face recognition as a tool to extract information to complement their analyses. For example, Redi et al. [2015b] and Vonikakis et al. [2014] combine face detection with other visual descriptors to model aesthetic qualities of images containing persons. Redi et al. [2015a] goes even further and predict "ambiance" ratings from Foursquare locations by analyzing the profile pictures of users who frequented the locations.

A number of such studies used Instagram as the OSN of choice. Bakhshi et al.

[2014] show that users tend to engage more to photos with faces than other kinds of photos, and that gender, age and the number of faces does not impact the engagement as significantly as the presence or absence of faces. Jang et al. [2015] compared the behavior of adults and adolescents in Instagram, using automatic face recognition to sample its data and further validating the dataset manually. They find the adolescents post less photos, but get more likes and use more hashtags than adults. They also find that adolescents tend to remove more posts they already shared, their posts are more directed to stating their mood and asking for likes/followers, and that they post more selfies.

Interestingly, most of the authors cited in the last two paragraphs used `Face++` as the FRS of choice (the exception is Vonikakis et al. [2014], who used the free computer vision library `OpenCV`), and Bakhshi et al. [2014] and Jang et al. [2015] used it specifically for estimating gender and age. While Jang et al. [2015] used it only for finding adults and validated the output manually afterwards, Bakhshi et al. [2014] validated their method using crowd sourcing. Unfortunately, they transformed the data before submitting it to validation, and their results do not generalize for `Face++`'s raw output.

Now that I have presented the results of collecting data in Instagram in Chapter 2 and how to estimate personal attributes using face recognition systems in Chapter 3 – along with the biases implied in this method – I will show how `Face++` can be used in Instagram to yield insightful information. In order to do so, I used the *CAMPS Data Collection Tool* to make `Face++`'s API scan all medias from the users in the Instagram dataset.

This chapter is organized as follows. In Section 4.1, I will investigate how biased is `Face++` with Instagram's sample through the calibrated scores described in Chapter 3. After addressing algorithmic bias, in Section 4.2, I will analyze how the faces detected in Instagram are distributed, and what they can say about the network. Then, in Section 4.3, I will present a simple methodology to map from attributes of individual faces found in the network to attributes from users. Finally, in Section 4.4, I will present some analyses that use these newly estimated user attributes.

4.1 Assessing estimation bias

The investigation of bias in face estimation follows two hypotheses, based on the literature review in Section 3.1.4:

Perceived gender classification bias depends on age In Section 3.1.4 I argued that there is mounting evidence of different confidence levels for gender classifi-

cation in different age bands. This can be reproduced in the Instagram dataset after `Face++`'s confidence scores are converted to probability estimates.

Bias depends on location There is some evidence that FRSs are biased towards some ethnicities. Thus, it is reasonable to postulate that some countries will have different expected confidence levels for gender estimation. I will test this by grouping medias per country and analyzing how gender confidence varies for each country, for each gender.

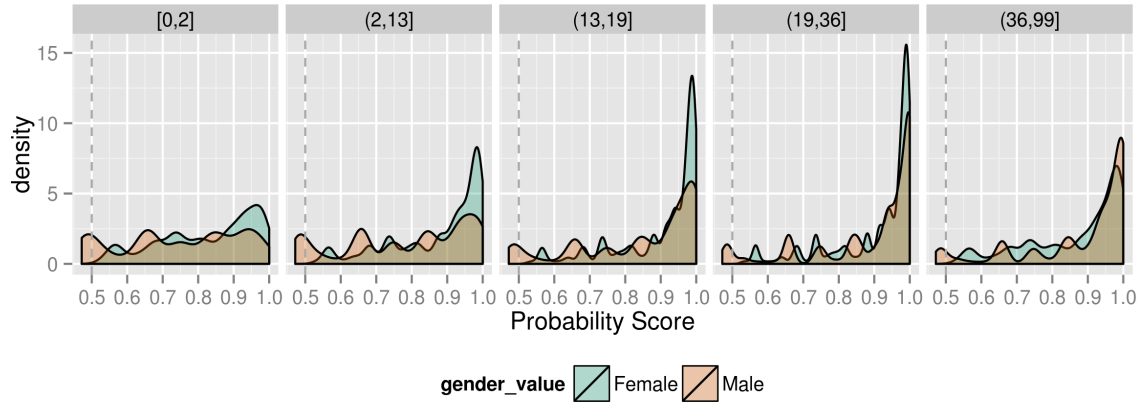
Notice that age bias is not being assessed. Unfortunately, due to the absence of a reliable confidence score or ground truth, this will not be possible. However, the age estimate can be used to better understand the perceived gender classification bias.

In Section 3.4 I described a method of obtaining the probability of a given face being a female face using calibration from the GROUPS dataset. This resulted in a model that converts the `gender confidence` scores returned by `Face++` to a probability score. The expected frequency of female faces in a sample of n faces randomly drawn from a population with a probability score of s is $n \times s$. Thus, the closer this probability score is to 1, the more likely it will be a female face. Combining this information with the `gender value` attribute returned by `Face++`, it is possible to calculate the expected **precision** of `Face++` in estimating a gender value. If the face is predicted to be female, then the algorithm is expected to be right with a probability of s. Conversely, if the face is predicted to be male, then the algorithm is predicted to be right with a probability of $1 - s$. If this precision varies systematically with another attribute – in our case age and location – we can say that the algorithm is biased relative to this attribute.

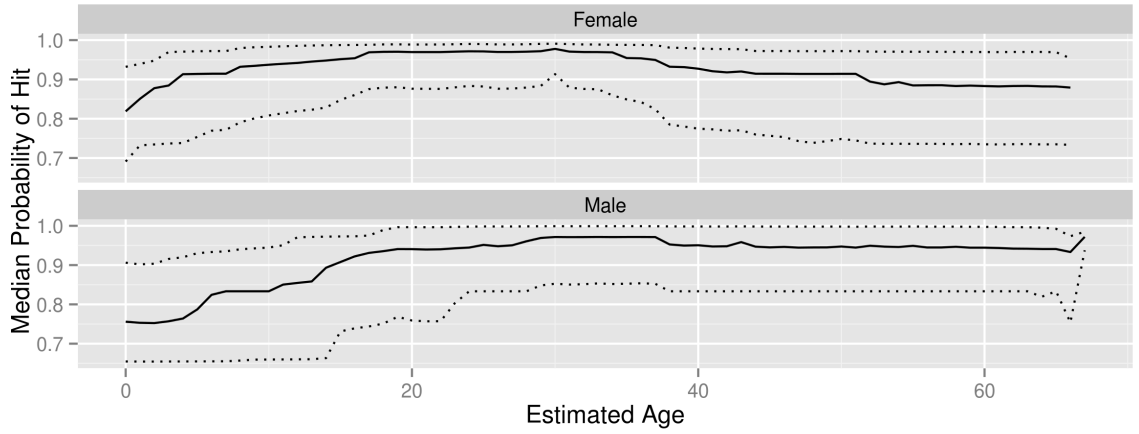
Figure 4.1a explores the relationships between the estimated age and the confidence of the gender estimation. It can be seen that for some age bands (*e.g.* $(19, 36]$), the probability is much more concentrated in one gender than for other bands (*e.g.* $[0, 2]$). This can be further verified in Figure 4.1b, which shows the median and the first and third quartiles of the probability score values for each age value[1]. The probability scores for males were transformed ($s' = 1 - s$) to ease interpretation. The algorithm's expected precision for both females and males is fairly high in across ages, varies consistently along a person's lifespan.

Two effects observed in Ngan and Grother [2015] can also be observed in this dataset, although much subtler. First, the peak of classification accuracy for males occurs in later ages than for females – the probability score reaches its plateau for 18

[1] I opted to use robust statistics instead of mean and standard deviation because, as can be seen in Figure 4.1a, the distribution of the probability score is skewed. However, using the mean instead of median yields the same conclusions.



(a)

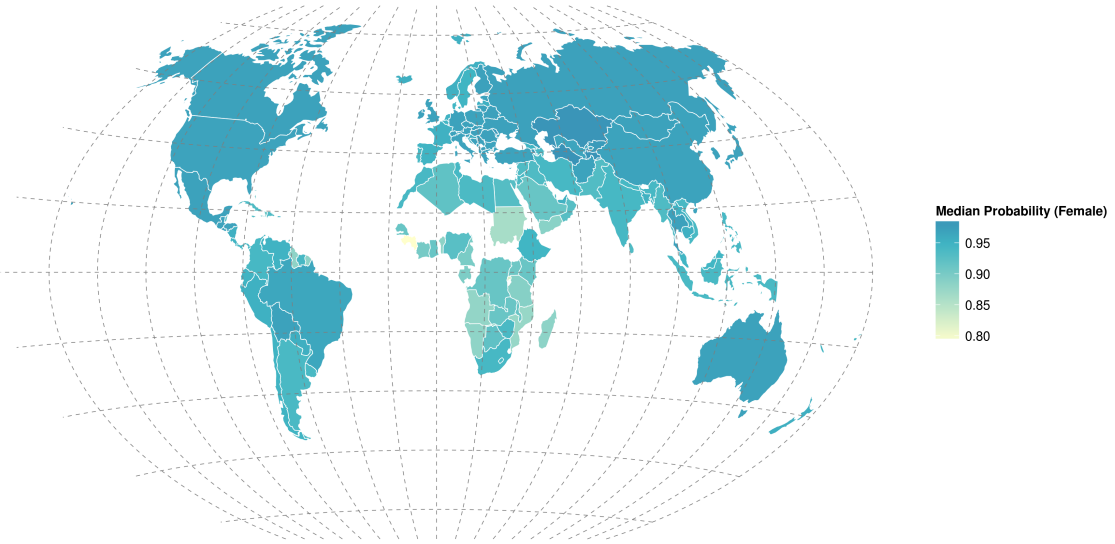


(b)

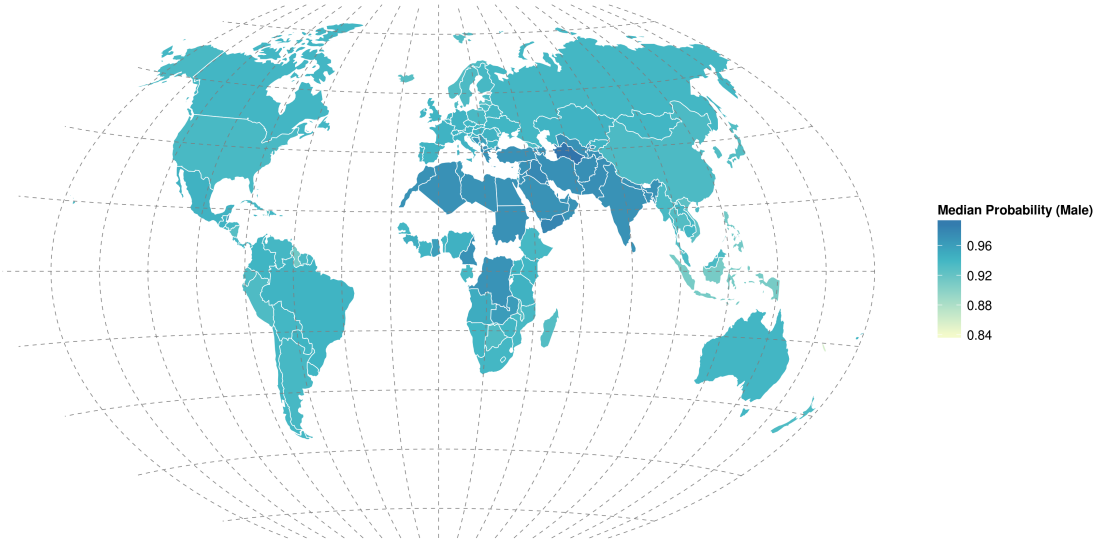
Figure 4.1: Relationship between age and probability scores in Instagram. (a) Density plot of the probability scores by age ranges. (b) Line plot of the median probability score and first and third quartiles per age value, separated by gender.

year old females, but only for 30 year old males. Second, performance for 40 year old males "stabilizes" and stays at the same level for older ages, whereas it keeps dropping for females. At 60 years old, median performance for males is 94.4%, while for females is 88.3%. Conversely, in Ngan and Grother [2015] performance for male classification was consistently better in all algorithms and ages, while in this dataset the average expected precision is slightly higher for females – 89.4% versus 86%.

Notice that some probability scores for users classified as males are lower than 0.5. This means that, after the isotonic correction, they should be classified as females instead. I opted not to change their label, since what is being assessed is the output of `Face++`, and the isotonic correction was a step taken only for defining the expected precision of `Face++`'s output in Instagram.



(a)



(b)

Figure 4.2: Relationship between country and probability scores in Instagram.

The variation along countries can be seen in Figure 4.2. Again, the scores for males were transformed. It can be seen that Western and Asian countries show a better precision for females, while countries in Africa and in South and Western Asia show a worse precision. Conversely, the highest precision for males is achieved in countries in North Africa and in South and Western Asia, indicating that there is a big overlap between low precision for females and high precision for males (and vice-versa), with the exception being South Africa.

4.2 The faces of Instagram

Figure 4.3 shows the distribution of ages estimated by `Face++`, separated by gender. It can be seen that the algorithm estimated much more females than males, and the female population is slightly younger on average than the male population.

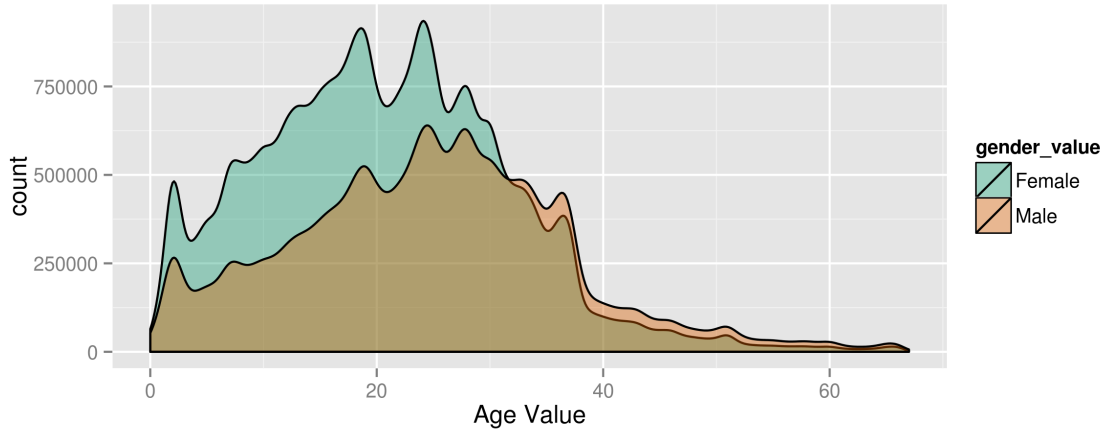


Figure 4.3: Density of the age of the found faces separated by gender. Notice that the marginal densities were preserved to enable comparison between genders.

To better understand the demographics of faces in Instagram, it is worth observing how it varies by country. To do so, I calculated the median age and proportion of female faces for each country, based on the resolved geolocations described in Section 2.3. In order to control for country-specific demographics, I subtracted the proportion of females from freely available estimates. More specifically, I defined a new metric, **female representation** (FR), as the difference between the proportion of females seen in Instagram from a given country from the country's proportion of females:

$$\mathrm{FR(c)} = F(c)_{instagram} - F(c)_{census}$$

where $F(c)$ is the proportion of females observed in Instagram, and in a reliable census estimate, respectively. Here, I used census estimates from the World Bank's DataBank [2].

I opted not to apply the same correction to the median age. This decision was motivated by the fact that the median age across countries in census data varies much more than the median age across countries in Instagram. In fact, using data from the United Nations[3] from 2014, one can observe that the median age varies from 15

[2] `databank.worldbank.org`
[3] `data.un.org`

years old to 45 with a standard deviation of 8.32). In contrast, the median age from Instagram varies from 17 to 31, with a standard deviation of 2.26. This means that simply subtracting the two measures would be insufficient – they would have to be re-scaled, hindering the interpretation of the results.

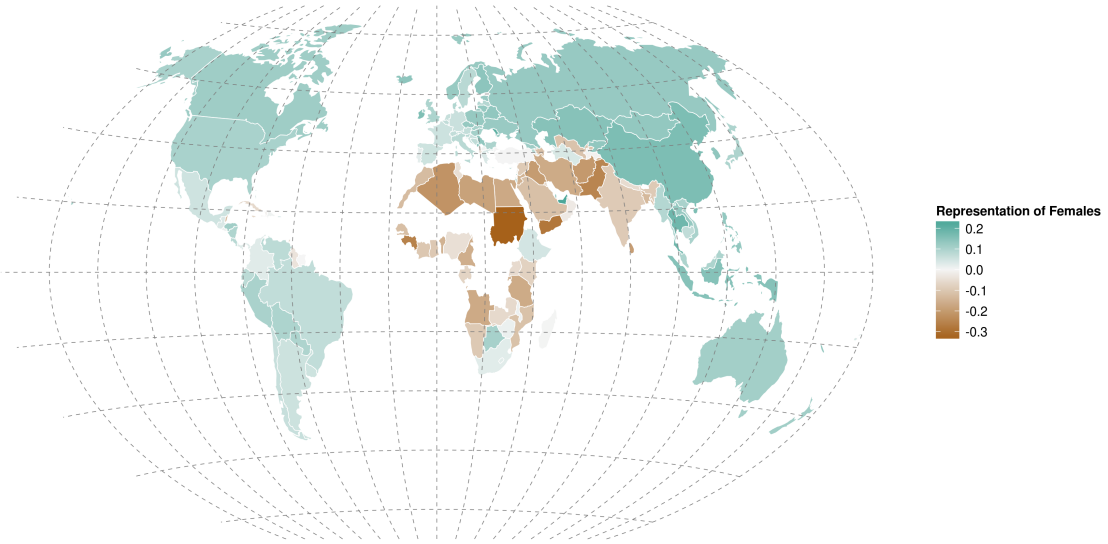


Figure 4.4: Female representation by country. Countries that were either below the cutoff of at least 100 medias or did not have enough data in World Bank's DataBank were omitted.

Results can be seen in Figures 4.4 and 4.5. It can be seen that women are overrepresented in almost all of the world, with the exception of North Africa and South and Western Asia. Notice that these are also the regions that have the most bias in gender estimation, as shown in Section 4.1. One could suggest that this effect is due to this bias differential. However, the high bias for females and low bias for males in these regions mean that many faces predicted to be females will actually be male, while most of the faces predicted to be male will be in fact male. This means that the amount of "true females" in these regions is expected to be even lower than what has been estimated – *i.e.* females in these regions are overestimated.

The median age of Instagram in all countries stays in the 20-30 year old band. As stated before, the median age of different countries vary considerably, and this stability in Instagram's median age is probably due to the service targeting a specific demographic group. Asian countries have the youngest population, especially China. This can be due to bias from face estimation, as Ngan and Grother [2014] noticed that Asians tend to have their age underestimated. However, Ngan and Grother analyzed FRSs from the USA, and `Face++` is a Chinese service with many Chinese clients, which

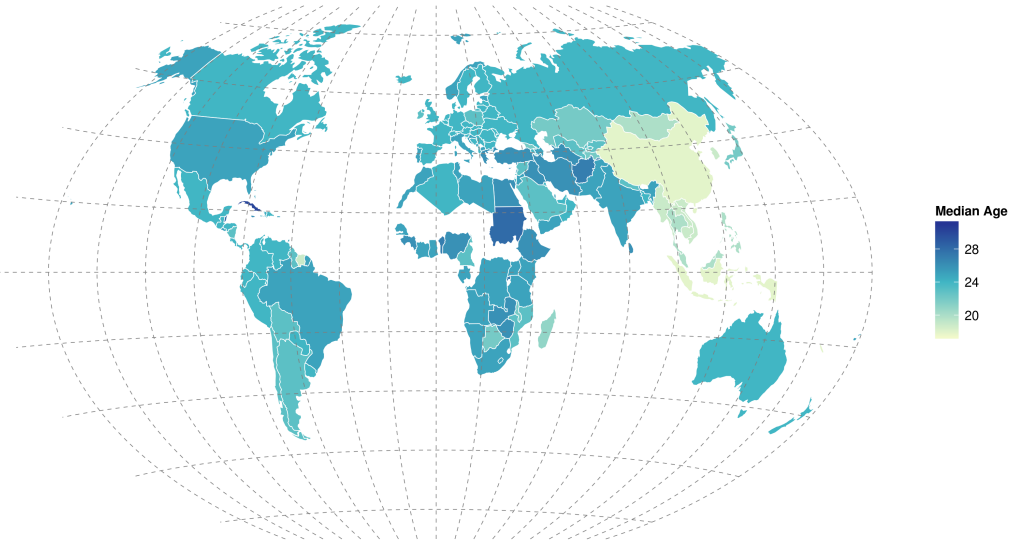


Figure 4.5: Median age by country. Countries that were either below the cutoff of at least 100 medias, or did not have enough data in World Bank's DataBank were omitted.

means that they have a strong economic incentive to be precise with this demographic profile. Unfortunately, this hypothesis cannot be verified in the current work.

4.3 Connecting medias to users

`Face++` returns information about all faces it can detect in a photo, but looking at each media alone does not allow us to identify which face is the user's face, or even if the user is depicted in all her posts.

One could assume that the face of the user in the profile picture is the user's face, and thus assign each user the gender and age found in the profile picture. However, many profile pictures have either many or zero faces: 1 511 631 faces were found in the profile pictures, while 3 119 742 profile pictures had no face detected – 1 455 529 of which were from users who did not change their picture from the default (*i.e.* their url points to `https://instagramimages-a.akamaihd.net/profiles/anonymousUser.jpg`). Therefore, this approach will have a very low coverage and biases the results.

This approach can be enhanced by combining information of the profile picture with information in the user's posts. Due to homophily – a tendency for people to have friends that share their own traits [McPherson et al., 2001] – a person is expected to have slightly more same-gender friends, and one can assume this would translate

to more pictures with same-gender faces. The same applies to age: the average age of a person's friends is her age, and it is reasonable to assume that the average age of the people appearing in her pictures will be the user's age. In fact, Pesce et al. [2012] showed that one can achieve good classification performance for estimating both age and gender in Facebook by simply averaging the ages and genders of a user's photo-tagged friends.

Therefore, even without knowing the user's face, it is possible extract information from all the faces from the user profile and develop a method to discover the user's gender and age.

There are a variety of ways this problem can be tackled, but I chose to approach it by modeling the problem as finding a set of thresholds in the user's known attributes that can be used to decide the user's unkown attributes. This can be done by training a Decision Tree model from a ground truth and applying the model to the dataset.

A Decision Tree recursively splits the dataset in a way that best reduces an *impurity* measure $I(A)$. In this case I used the Gini Impurity[4]:

$$I(A) = \sum_i p(1-p) \tag{4.1}$$

where A is the subset of the data representing a given split and p is the proportion of positive instances of the class being learned. Thus, a set of instances with high impurity has a proportion of positive instances near 50%, and a set with low impurity has a proportion near 0 or 100%, meaning that this method manages to separate positive and negative instances efficiently. The Decision Tree Algorithm selects the attribute and split point that best achieve this goal [Breiman et al., 1984]. Additionally, it selects surrogate variables that can be used when the selected variable is missing. Thus, Decision Trees have a natural way of dealing with missing data.

The first splits of a decision tree are highly informative of the dataset's inner structure, but afterwards the tree quickly increases in complexity without generalizing well to unseen data (in other words, it overfits the training data). In order to avoid that, the tree is *pruned* after a given number of splits. More specifically, a subset of the training set can be used as a validation set (in which the tree is not trained) and the tree can be pruned when the error in the validation set stops decreasing.

More formally, the best decision tree is found by cross validation: the training set is separated into 10 parts, the tree is trained in 9 of these parts and the remaining part is used as the validation set. A classification error is calculated for each point the

[4]Not to be confused with the *Gini Coefficient*, a popular measure in the social sciences to describe income inequality.

tree splits the data. This procedure is repeated, leaving a different 10^{th} of the training set as the validation set, until all the 10 parts were used for validation. The average and standard error of the classification error are estimated using the error found for each iteration. Then it is possible to find the number of splits that has the smallest classification error. However, splits within 1 standard error from this minimum point are equally good candidates for pruning. Thus, for parsimony, the optimal point for pruning is the one with the smallest number of splits in the interval within one standard error from the point with the smallest classification error.

In order to maintain this methodology as general as possible, I limited the feature set to the output given by `Face++` over each user, combined by simple transformations: the proportion of faces of each gender, the average age, and the number of faces found for the user. The precise definition of the features can be seen in Table 4.1. A diagram of the whole approach can be seen in Figure 4.6.

`n_faces_m`	Sum of the number of faces identified in all of the user's medias
`prop_females_m`	Proportion of female faces in the users's medias
`prop_females_p`	Proportion of female faces in the user's profile picture
`avg_age_m`	Average age of faces seen in the posted medias
`avg_age_p`	Average age of faces seen in profile picture

Table 4.1: Features used in the Decision Tree

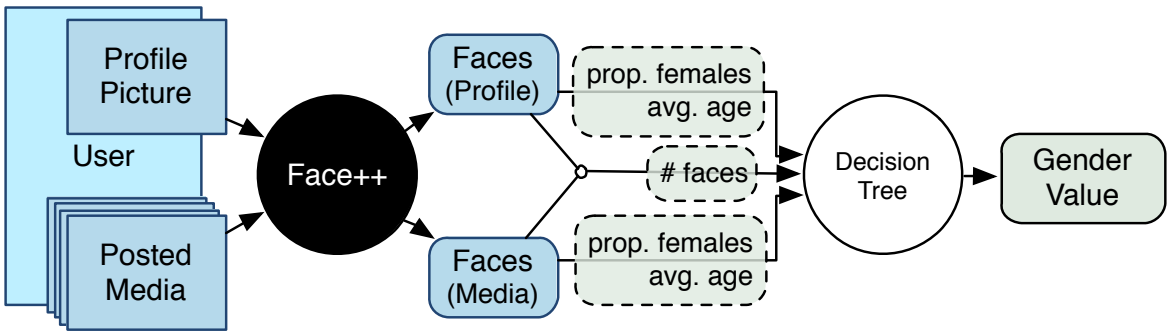


Figure 4.6: Diagram of the pipeline for estimating the gender of a user.

4.3.1 A ground truth for gender

To train and test this approach, a ground truth is needed. There is no publicly available dataset that resembles Instagram profiles, but it is possible to leverage other sources of information from certain users to discover their gender, and then predict the gender of the remaining users using machine learning methods.

A highly precise method of estimating gender in social media is through **name lists**. Tang et al. [2011] used a name list to estimate the gender of Facebook users from the New York and Boston networks, and achieved 96.8% accuracy and 96.3% coverage in gender estimation. A particular advantage of using name lists in Facebook is that the service enforces a Real Name Policy[5], which allows the method to achieve such a high coverage. Even if the name list does not contain all possible names of the users, since names are distributed according to a power law [Tang et al., 2011], the most frequent names will account for a significant proportion of all user names. In other OSNs (such as Instagram) this is not normally the case – users are identified by a unique nickname and can optionally fill their real name in a field. This exposes the main disadvantage of using name lists: when users cannot be identified by their real names, the method's coverage is hindered. However, the accuracy is not severely impacted, as shown in Burger et al. [2011] in a study on Twitter.

Another important factor to consider is that the relationship of name and gender is culture-specific, so a given name list is only valid for the location where it is generated. Moreover, many names cannot be properly attributed to a specific gender (*i.e.* they are assigned to both females and males) and are usually removed from the list. Alternatively, when information about the frequency of times that a name is assigned to either males or females is available, it is possible assign a probability score of a name being from each gender. That is, let F_n and $M_n n$ be the count of the times that a given name n is assigned to females and males (respectively) and M and F be the sum of the frequencies of all known name assignments for males and females. Then, the name's score s can be calculated as:

$$s(n) = \frac{F_n/F}{F_n/F + M_n/M}$$

and a threshold value can be set to determine when the name can be reliably considered male or female.

When all these observations are taken into account, name lists are one of the most precise methods for gender estimation using profile information. Although attempts have been made to use names to estimate age [Gallagher and Chen, 2008], it is normally only possible to do so with the help of external data.

I used three name lists from different sources:

Census 1990 Names: a list of 5163 American names and their relative frequencies, compiled from the Census of 1990, which is freely available in the US Census'

[5] `https://www.facebook.com/help/112146705538576` (Visited in February 2016)

Dataset		[1]	[2]	[3]
Brazilian Names	[1]	–	0.81	0.83
Census 1990	[2]	0.81	–	0.97
Facebook NYC	[3]	0.83	0.97	–

Table 4.2: Correlation between overlapping names in each dataset. The numbers in brackets are only shorthands that refer to each dataset, added for readability.

site[6].

Brazilian Names: A list of 33 866 known brazilian names used by Cunha et al. [2014]. No frequency information is available, but names with an ambiguous gender assignment were removed.

Facebook NYC Names: A list of 23 405 names and their frequencies, compiled from public profiles of Facebook's New York network, presented by Tang et al. [2011].

The datasets overlapped moderately. Census 1990 and Facebook NYC had an overlap of 2210 names, and Brazilian Names overlapped with Facebook NYC by 4200 names and with Census 1990 by 2210 names. Interestingly, even for a 0.5 decision threshold, all datasets agree on all the gender assignments for each name. The correlations of the probability scores in these overlapping regions can be seen at Table 4.2. Notice that in the Brazilian Names list the name frequencies are not available and the scores are binary: 1 for females and 0 for males.

Female names are more diverse than male names in all datasets. Figure 4.7 shows the distribution of scores for each list. Following Tang et al. [2011], to achieve the highest precision I used a threshold of 0.8 for determining whether a name was for a given gender, *i.e.* a name was considered male if its score was at most 0.2 and female if it was at least 0.8. The intermediary results were discarded.

Instagram provides the option of filling a `full name` field in a users's profile. The first word of this field can be considered the first name, and then compared to the name lists to assign the proper gender. To account for cultural specificity, I restricted the search to the country where the name list was generated (*i.e.* either Brazil or USA). Since users do not have a `country` field, I considered a user being from a country if the user had most of her geotagged photos in that country. Thus, my method for generating the ground truth was the following:

For each named list:

[6] `http://www.census.gov/topics/population/genealogy/data/1990_census.html` (Visited December 2015)

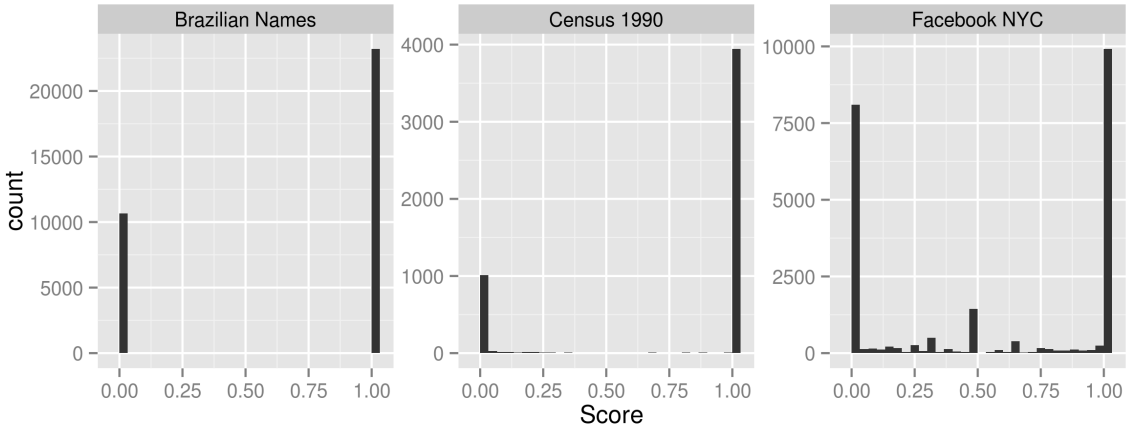


Figure 4.7: Histogram of the distribution of name scores for each dataset.

1. Determine the target country
2. Select all users whose post mostly in the target country
3. Extract the first word of the `Full Name` field of all these users
4. Check if this word matches any of the names in the name list

Not all users could be assigned to a country, since geotagging is optional. 746 001 had a top country, 85 659 of whom had Brazil as their top country, and 181 037 had the USA as their top country. From these users, 50 569 Brazilian users could be matched to the Brazilian name list, 85 659 American users could be matched to the Census name list and 65 551 users could be matched to the Facebook NYC name list.

4.3.2 Inference

Having established a ground truth, I extracted the relevant features and trained the Decision Tree. For comparison, I also tried two naive methods that did not require learning: considering all users with `prop_fem_profile` over 0.5 as female, and cosidering all users with `prop_fem_media` over 0.5 as female. Additionally, I compared the Decision Tree with another highly efficient classier: a Support Vector Machine (SVM). Since the SVM does not handle missing values naturally, I imputed these missing values with the mean of each attribute. The SVM hyperparameters were fine-tuned using grid search and cross-validation (the best paramerters were $\gamma = 1$ and $C = 0.1$)

The naive methods yielded surprisingly good results, which can be seen in Table 4.3, with an accuracy of 0.70 and 0.82 and coverage of 0.99 and 0.44, respectively. It must be noted that using media information yielded better accuracy than using the

	Accuracy	Precision		Coverage
		Male	Female	
Naive (media)	0.82	0.79	0.84	0.99
Naive (profile)	0.81	0.75	0.86	0.44
SVM	0.83	0.82	0.84	1.00
Decision Tree	0.84	0.82	0.85	1.00

Table 4.3: Performance of the different methods of classification

profile picture, even though one would expect that a profile picture would be more important to determine the users's gender.

The decision tree and SVM perform similarly, with the decision tree performing slightly better. Besides the natural handling of missing data, one advantage of using a Decision Tree is that its thresholds can be plotted and inspected, as in Figure 4.8. It can be seen that most of the training set (92%) receives a label at the second split, when the tree has levaraged the "raw" information of the genders in the medias and profile picture. These are also the cases with highest purity. The remaining, "hard" cases are successively split and secondary information is used, like the number of scanned faces and the average age in the profile picture. However, the tree does not manage to achieve a good level of purity in most of the leafs.

Since the Decision Tree was trained with users from only two countries, and considering that there is a considerable variation in the proportion of female faces and average age depending on the country, it is reasonable to worry that the classification rules will not generalise well for users of other cultural backgrounds. In order to evaluate that, I randomly sampled 10 users from each of the 16 countries with most users in the dataset (which contain 75 % of the geotagged media) and manually labeled them as female or male based on their profile[7]. Since what is of interest is the country-wise variation, I avoided "hard cases" by sampling only users with more than 10 faces detected in their profiles. The results are in Table 4.4. Although the sample size for each country is modest, the average accuracy and precisions is comparable with the results in Table 4.3. It can also be seen that the classifier performs reasonably well regardless of the country.

[7]I did not have access to the predicted gender during labeling.

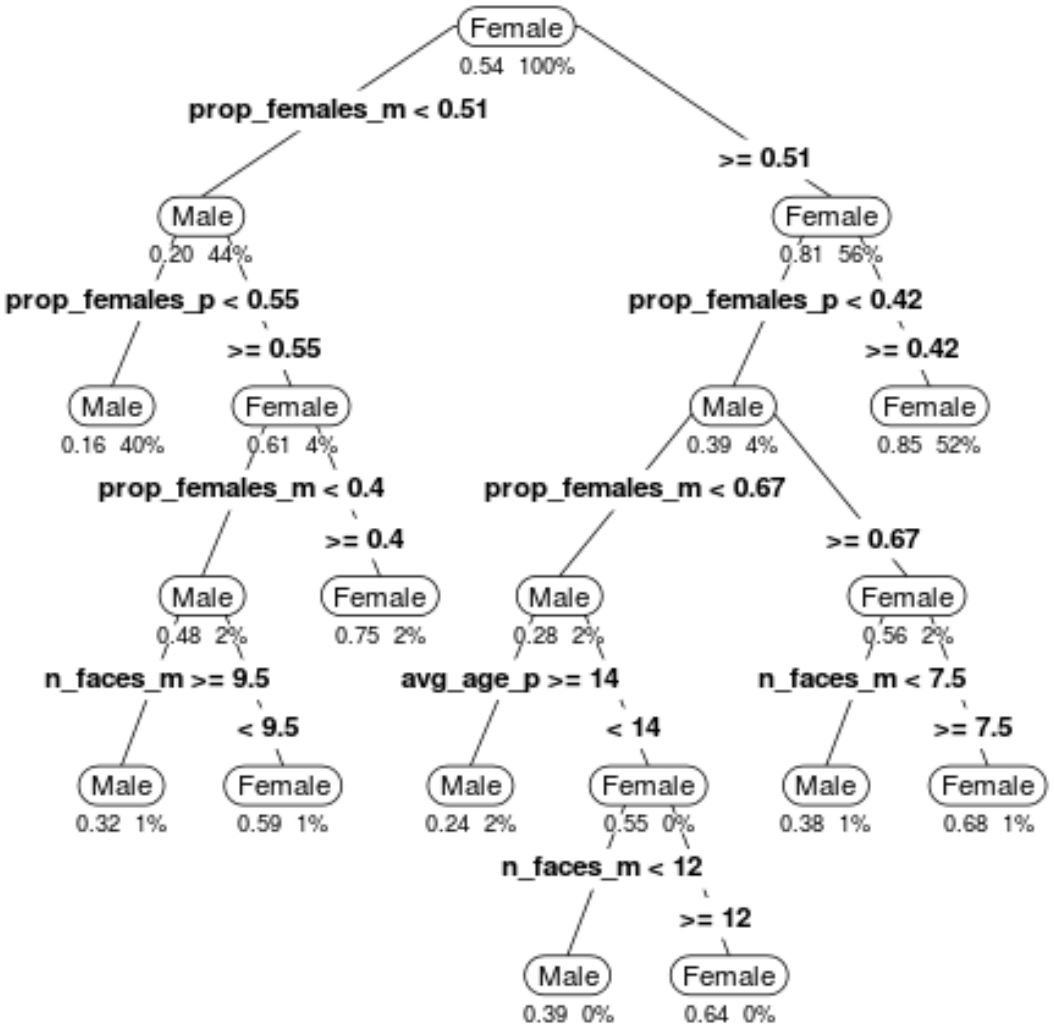


Figure 4.8: First 10 splits of the gender Decision Tree. Each node states the dominant class of its split side. The numbers below the node represent the probability of an observation inside the node being Female (left), and the proportion of the training set that "falls" into that split (right). The rule for each split is stated in the branches of the tree.

Country	Accuracy	Precision		Frequency	
		Mal.	Fem.	Mal.	Fem.
USA	0.8	1.00	0.75	4	6
BRA	1.0	1.00	1.00	4	6
AUS	1.0	1.00	1.00	3	7
CAN	1.0	1.00	1.00	3	7
CHN	0.9	1.00	0.87	3	7
DEU	1.0	1.00	1.00	4	6
ESP	0.8	1.00	0.71	5	5
FRA	1.0	1.00	1.00	4	6
GBR	0.9	0.80	1.00	4	6
IDN	0.9	0.75	1.00	3	7
ITA	0.9	0.83	1.00	5	5
MEX	1.0	1.00	1.00	6	4
MYS	0.8	1.00	0.71	5	5
PHL	0.9	0.80	1.00	4	6
RUS	0.9	0.83	1.00	5	5
SAU	1.0	1.00	1.00	7	3
THA	0.7	0.50	0.83	3	7
TUR	0.9	0.83	1.00	5	5
Average*	0.91	0.90	0.94	4.31	5.68

Table 4.4: Performance in different countries.
*The average does not consider BRA and USA.

4.4 User-level analysis

4.4.1 Gender balance in Instagram

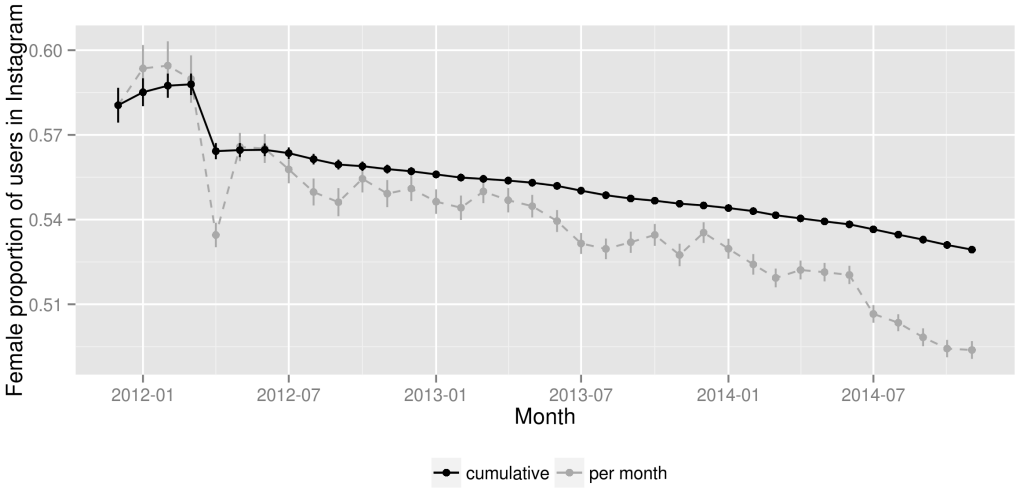
It is possible to investigate the gender balance in Instagram by analyzing the proportion of female (or male) users in the network. The proportion will be different depending on how the data is analyzed. In order to estimate the user's gender, the profile must have at least one media or a profile picture with an identifiable face. The more medias the user has, the more reliable will be the estimates. However, if the user's gender has an effect in user activity, the proportion of female users in different strata of activity may also differ substantially.

	Females	Males
All users	53 %	47 %
At least 10 posts	55 %	45 %
... 50 posts	59 %	41 %
... 100 posts	61 %	39 %
... 500 posts	64 %	36 %
... 1000 posts	62 %	38 %
... 10000 posts	55 %	45 %
At least 10 followers	53 %	47 %
... 50 followers	53 %	47 %
... 100 followers	54 %	46 %
... 500 followers	56 %	44 %
... 1000 followers	55 %	45 %
... 10000 followers	51 %	49 %

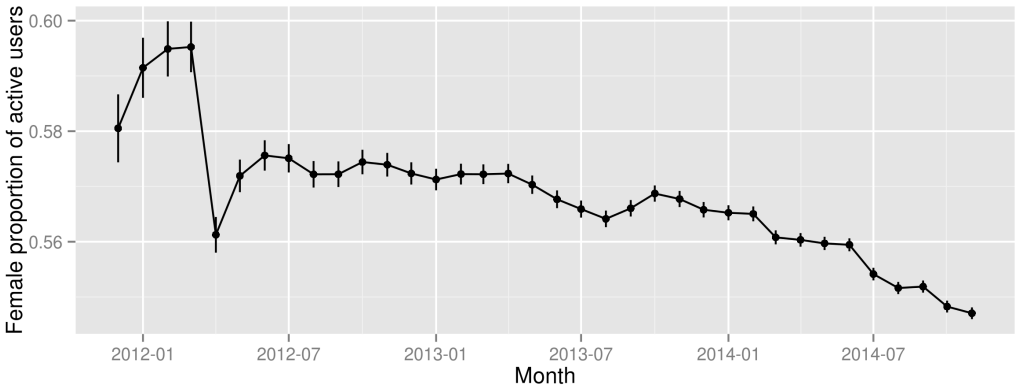
Table 4.5: Proportion of users of each gender for different levels of engagement

Table 4.9 shows the gender proportions for different cutoff criteria. There is a small variation in the gender proportion when different cutoffs in the number of posts or followers are considered. This variation does not stabilize for users with more than a few posts or a few followers, and differs depending on the variable chosen for filtering. In any case, it stays between 55 % and 64 % of females, and varies less when the cutoff variable is the number of followers. It is possible that there is a relationship between the number of posts and the user's gender – this will be further explored in the next section.

It is possible to follow the evolution of the gender balance in Instagram by looking at the timestamps of the medias posted by the users. Figure 4.9 shows this evolution,



(a)



(b)

Figure 4.9: Change in proportion of users in Instagram over time. The ranges are confidence intervals. Figure (a) refers to new users (whose first post is in month x) and Figure (b) refers to active users (who posted in month x).

both for new users, as measured by their first post in the network, as for active users, measured by the number of users who posted something in a given month. The figures show that Instagram was much more unbalanced in the beginning of 2012, but is becoming progressively more balanced. In fact, slightly more male users seem to be joining the network than female users, and the number of active male users is almost approaching the number of active female users.

This is a smooth trend that has been happening consistently throughout the years, except for a sharp difference in new and active users in April 2012. I could not find the reason for such difference, but it is interesting to notice that it occurred at approximately the same time as the "jump" in user IDs occurred, described in

Section 2.2.2.

4.4.2 Gender differences in Instagram's attributes

Having estimated the gender of the users in Instagram, it is possible to know how different is the behavior of users of each gender, by comparing the distribution of the user attributes conditioned on the user gender.

An intuitive non-parametric effect-size measure for comparing distributions is the **probability of superiority** (PS), based on Mann-Whitney-Wilcoxon's U statistic. The U statistic represents the number of times an observation from a group is bigger then an observation of another group in all possible pairings of the two groups. It can be calculated in large samples by obtaining the rank of all observations (the smallest observation's rank is 1, the second smallest observation's rank is 2, and so on), and summing these ranks for each group.

For example, consider a set of observations of a random variable X paired with groups A and B. If the observed values of variable X for group A are $X_A = \{2, 5, 70\}$ and for group B are $X_B = \{1, 3, 6\}$, then the ranks of the observations are $\text{Ranks}_A = \{2, 4, 6\}$ and $\text{Ranks}_B = \{1, 3, 5\}$. Thus, the sum R of the ranks are $R_A = 2+4+6 = 12$ and $R_B = 1 + 3 + 5 = 9$. These rank sums carry information on how the values of X for one group are "lagged behind" another group without considering the magnitude of their differences. In other words, R_A is proportional to the number of times that the values of group A are smaller than the values of group B considering all possible pairings of groups A and B.

Let R_f and R_m be the sum of the ranks for females and males, respectively, and n_f and n_m be the number of observations for males and females. The U statistic for females and males can be calculated by the following formula:

$$U_f = R_f \frac{n_f(n_f + 1)}{2} \qquad\qquad U_m = R_m \frac{n_m(n_m + 1)}{2}$$

And the chosen U statistic is the smallest from the two groups: $U = \min(U_f, U_m)$. The U statistic is normally distributed, with:

$$\mu_U = \frac{n_f n_m}{2} \qquad\qquad \sigma_U = \sqrt{\frac{n_f n_m (n_1 + n_2 + 1)}{12}}$$

Thus, a p-value is normally computed by looking up the probability of U in the null hypothesis of no difference between the two groups. Moreover, the female

probability of superiority PS_f can also be easily calculated as $PS = \frac{U_f}{n_f n_m}$. Thus, the PS_f of an attribute is the probability that a randomly chosen female will have a value higher in this attribute than a randomly chosen male. A value of 0.5 means that the gender has no effect in the attribute, and values close to 1 or 0 mean that gender have a strong effect.

	Median		PS_f	p-value
	Fem.	Male		
# Followees	60	61	49.46 %	< 0.0001
# Followers	52	51	50.58 %	< 0.0001
# Media Posted	15	12	53.57 %	< 0.0001
Avg. Tags used	0.0028	0.00001	53.30 %	< 0.0001
Avg. Likes	4	3.99	50.45 %	< 0.0001
Avg. Likes per Follower	0.073	0.074	51.14 %	< 0.0001
Avg. Comments	0.31	0.29	49.75 %	< 0.0001
Avg. Comments per Follower	0.0031 x	0.0033	51.01 %	< 0.0001

Table 4.6: Effect size of gender in the attributes

Results can be seen in Table 4.6, along with the median value for each gender. Besides attributes from the user profile, I also calculated the average number of likes and comments each user received, as well as the ratio of the average of likes/comments per follower. Moreover, I calculated the average amount of hashtags used by the user. The extent of the differences of the distributions can also be compared in Figures 4.10, 4.11, 4.12, 4.13, 4.14, 4.15 and 4.17.

Contrary to other OSNs[8], the differences in the behavior of male and female users are not very pronounced. Although all results are statistically significant to a high degree, this is likely to be due to the big sample size. The exceptions are average hashtag use and number of posts – females post slightly more and use slightly more hashtags than males. This higher effect size for number of posts explains the unstable results for Table 4.5. Also, although the distributions for the average of comments and likes are practically the same for both genders, females have a slightly higher ratio of average likes and comments per follower.

However, even the strongest the effect sizes are very close to 50 %. This shows that the gender differences in Instagram are not as pronounced as in other networks. Unfortunately, I cannot analyze differences in content use – this is beyond the scope

[8]see Ottoni et al. [2013], for an example on Pinterest

of this work, as methods of natural language processing and computer vision must be used in order to extract quantitative measures of content production.

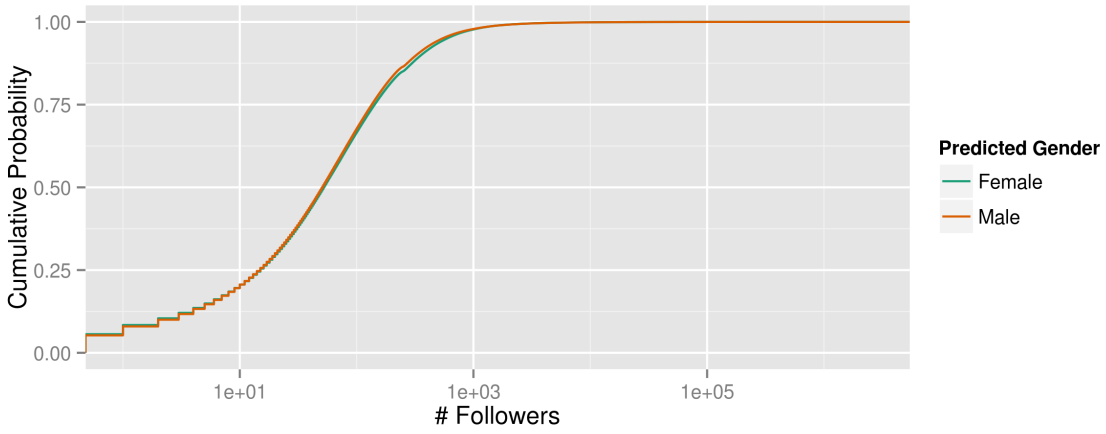


Figure 4.10

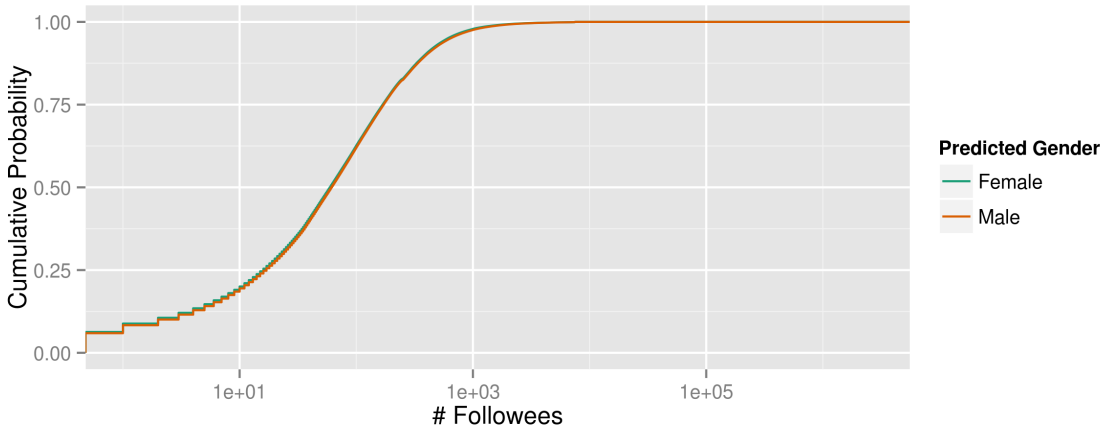


Figure 4.11

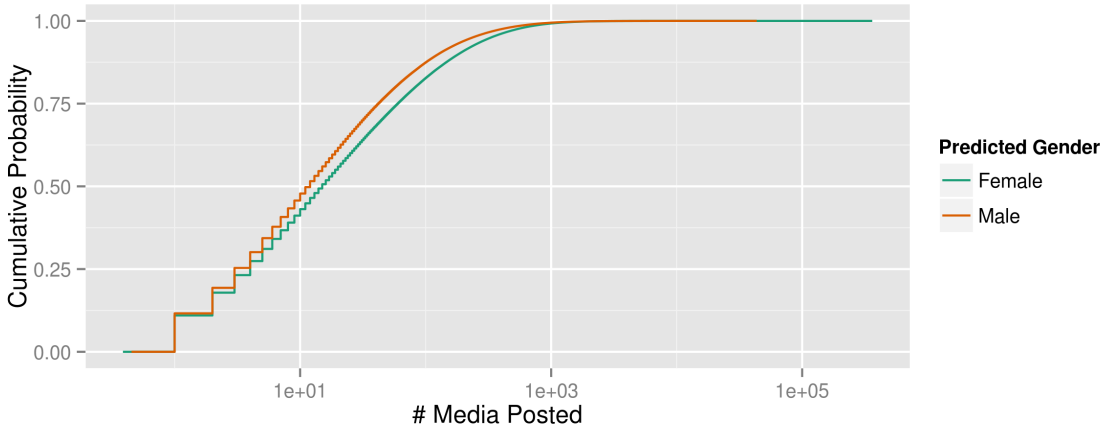


Figure 4.12

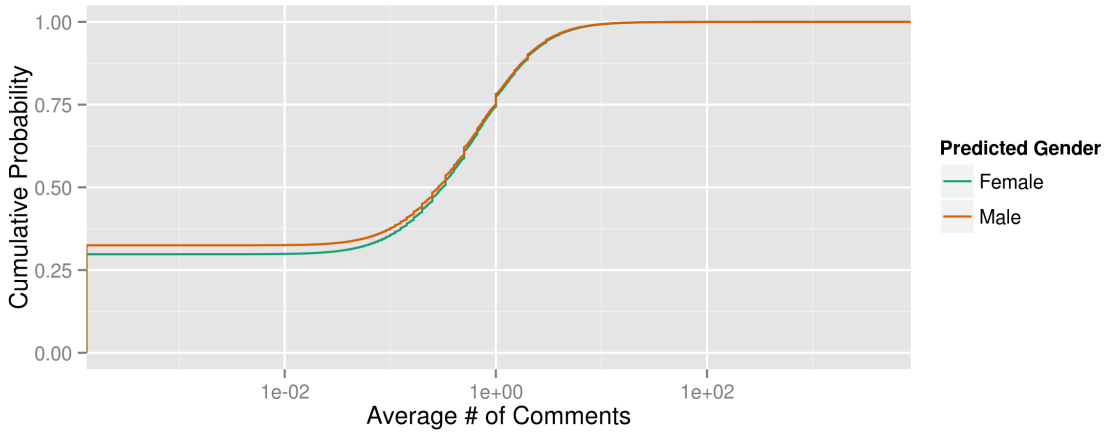


Figure 4.13

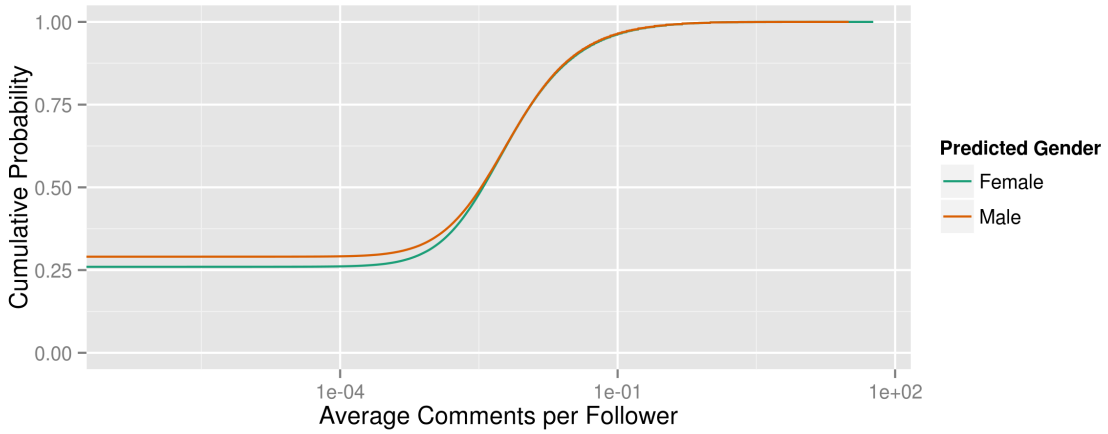


Figure 4.14

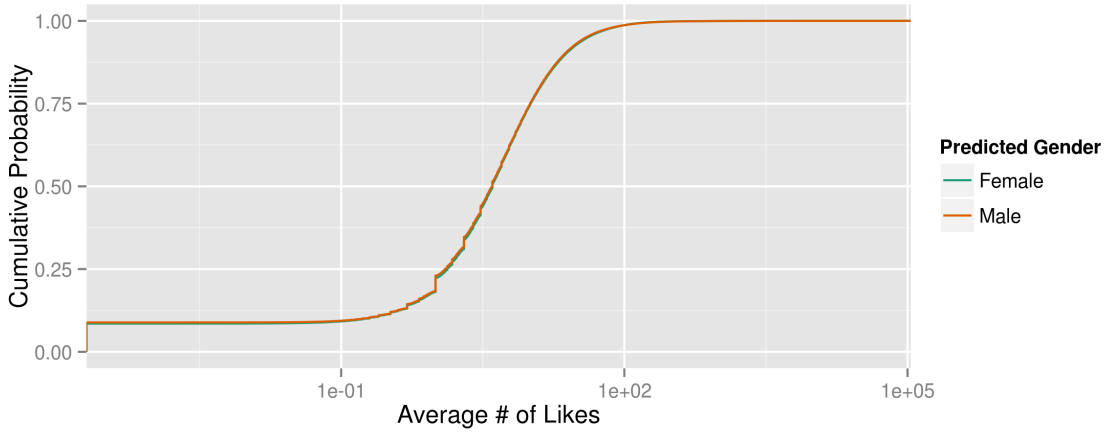


Figure 4.15

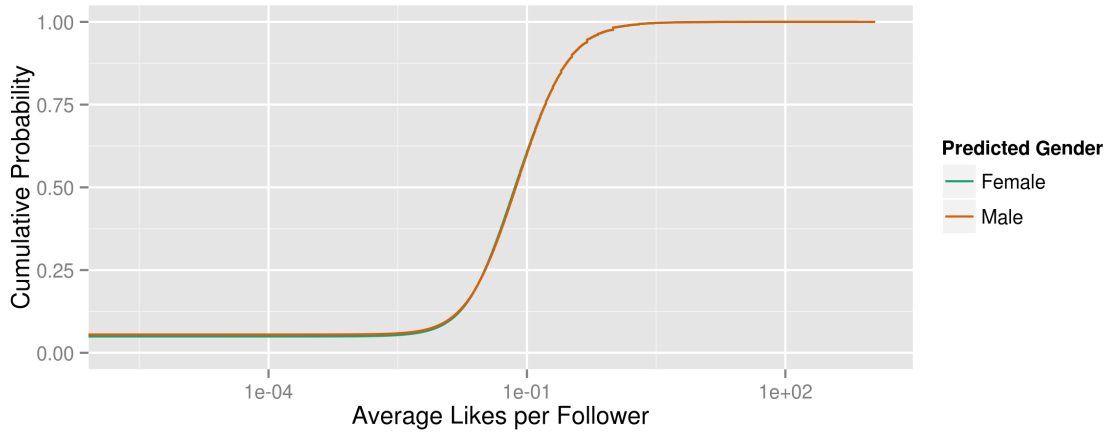


Figure 4.16

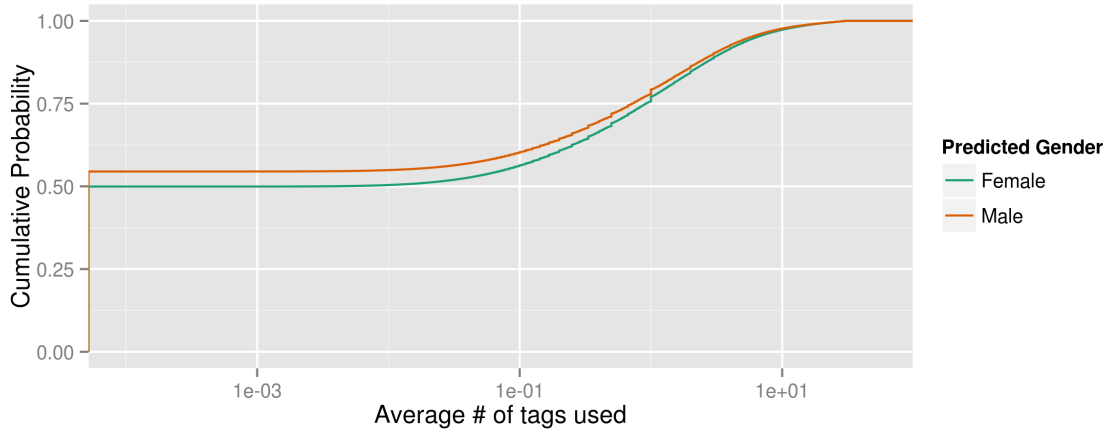


Figure 4.17

Chapter 5

Conclusions

The estimation of the age and gender of users in OSNs is an important step to define factors on which the behavior of these users can vary, as well as to define how the demographics of each online service is given in different points in time. In this dissertation, I investigated how face recognition can be used for estimating these attributes using images posted in Instagram.

To do so, I collected a random sample of public user profiles and their activity in the network using its REST API, and their posts were scanned using a state of the art, freely available face recognition system, `Face++`. The use of such system allowed me to achieve high precision in face recognition, while keeping the research steps reproducible.

I presented the current state of face recognition in the literature, and argued that although very advanced, it is not without bias towards certain groups of people. Thus, I defined a method of identifying algorithmic bias in the output of `Face++` by using a well known dataset, with pictures similar to the ones that are expected in Instagram, to generate estimates of how it should behave in the network. This method allowed me to generalize measures of bias to datasets in which human annotation is not feasible, and thus derive the bias that should be expected for Instagram as a whole. The results showed that the reliability of the estimation of age and gender is indeed conditioned to endogenous aspects of the task at hand. More specifically, male and female genders are not estimated with the same precision, and gender estimation depends on the age of the person whose face is being scanned and the geographical location where the face was photographed – probably due to different performance for different ethnicities.

Nonetheless, I showed that attribute estimation methods based on face data are still very reliable, and when bias is taken into account it is possible to spot differences in how women and men of different age ranges use Instagram. This approach can also be combined with and additional classification step that maps all the faces found in an

user profile to the user's gender – and possibly to the user's age, if a ground truth can be established for training the classifier.

This has the advantage of not needing to determine which of the detected faces is the user's face. Instead, it leverages the fact that users of one gender are more likely to appear in photos of users of their own gender – a well known phenomenon in social science called homophily. In fact, the strategy of relying in homophily of the posted media is so powerful that it shows performance comparable to cases where the profile picture of the user is available and has face data in it, which is a good indicator that the user's face is in fact available.

When the gender of users is estimated, it can be used for a wide range of tasks already defined in the social computing literature. Here, I limited myself to two of the most basic methods of characterization: describing how the gender balance changed during the network's evolution, and calculating how different are the behaviors of male and female users in the network as a whole. Contrary to what is observed in other OSNs, it can be seen from the results that gender only affects the user attributes slightly, and mostly in the production of content – female users posted more media and used more hashtags. Additionally, by looking at the date of the posts of the users I reconstructed how Instagram was in the past. By doing that, I showed that Instagram used to be more imbalanced towards females, but the proportion of female users is trending towards 50 %, and the tendency for the future is for Instagram to have a more balanced user base.

This characterization is merely a small demonstration of the techniques that can be employed to answer interesting research questions. Numerous examples of more sophisticated techniques that depend on knowing the user's gender are readily available in the literature, covering topics such as the categorization of content [Ottoni et al., 2013], linguistic style [Cunha et al., 2014; de Las Casas et al., 2014], gender discrimination [Terrell et al., 2016] and gender bias appearing in discursive patterns [Garcia et al., 2014].

However, this also signals a risk of undesired exposure of personal information. Although users who publicly upload pictures in the network are aware that third parties can infer their gender and age by looking at their pictures, most users do not know that this can be done automatically with high precision. This risk of exposure can extend beyond Instagram pictures and be combined with other attack strategies – all one needs is access of a set of pictures associated with a person. Moreover, if a user wants to explicitly hide information about her age and gender, she might assume that it is safe to upload pictures of her friends, as long as none of these uploads contain her face. However, results shown here suggest that the estimation method does not

need to depend on that availability of the person's face to reliably infer her gender. Future work could explore this question further by explicitly targeting users who do not expose their own face, and by exploring other ways of associating a set of pictures to a person that do not depend on extracting it from a user profile in an OSN.

My work also explores the topic of algorithmic bias. The method for investigating algorithmic bias presented here can also be generalized to any computational system employed for automatic decision making, as long as one can find a method of carefully inspecting the relationship between input and output of the system. Thus, every system can be treated as a "black box", which enables researchers to include proprietary algorithms in their investigation. The limitation for this method is when the input cannot be controlled or the output cannot be inspected. In this work, I leveraged the fact that `Face++` offers confidence scores for its classification of the perceived gender, as well as the fact that `Face++` exposes a public API with a high quota of detections per hour – which allows for a large number of inputs to be fed to the system. If this method were to be employed to investigate Facebook's face recognition system, for example, it would need to be extended, since Facebook embeds its system in its own platform, with no public API, and gives no information on the confidence for each classification.

The algorithmic bias reported here has been found before in other FRSs. Its origin is still unknown and can be subject of future investigation. Possibilities range from either an inherent limitation in using face data to estimate the perceived gender and age; economic and operational incentives for the FRS to respond better to certain demographic groups; or idiosyncrasies in the data collection methods employed to train the algorithms that will perform the task.

My research has some limitations. First, it is restricted to only one platform, and only publicly available data. It is perfectly reasonable to assume that users will behave differently in different OSNs, and users who make their profile private – almost 70 % of the IDs found during data collection – will have a different pattern of behavior. The public exposure of user profiles is likely to be related to the user's gender and age, since women incur in more social risks when exposing personal information publicly, and people from different age groups have different attitudes towards privacy. A more general characterization of user behavior could benefit from using all the data contained in an OSN, including private profiles. However, this is normally impossible to be done without access privileges given by the service's administrators.

A second limitation is in the method of estimating bias. Although images in the GROUPS dataset are drawn from Flickr, which should be very similar in format to what is expected to appear in Instagram, there is no guarantee that this is the case. A better option would be to generate a labeled dataset from pictures directly

drawn from Instagram, perhaps using crowdsourcing. However, this would demand extra resources, and would mean sacrificing reproducibility if this dataset could not be shared. Future work could explore the differences in pictures across multiple image-based online services in order to establish objective criteria of how transferable are the features of images found in one service to another.

A third limitation is in the scope of the analyses produced here. To limit my research subject, I chose not to analyze the content of the posts made by the users, and focused only in the most salient attributes in the OSN: followers, followees, interactions, hashtag use and the number of posts. It is possible that differences in behavior for different genders and age groups are more strongly manifested in the content instead of raw activity counts.

Finally, I did not manage to explain the unusual pattern of IDs in Instagram's ID space, and this could affect how representative is the sample collected. Investigating whether users in different ranges of the ID space show systematic differences in behavior that cannot be explained by their time in the network is a research question on itself.

I believe that these limitations do not hinder the contributions made by this work. Although some systematic variations in the data were probably not accounted for, there is no reason to believe that they would invalidate the conclusions.

The use of face recognition technologies as a tool to describe how users relate to the Visual Web is becoming increasingly more common, and a proper understanding of its uses and limitations is important. The exact comprehension of the mechanisms employed by the algorithms used for these tasks is desirable but, as I argue, not necessary. Instead, researchers can focus on the effects of using such algorithms in the groups they are willing to investigate. In this work I chose age and gender as personal attributes to be estimated, due to their social and theoretical relevance. However, the methods presented here can be extended to cover other topics, and as machine learning becomes more sophisticated, more inferences will be possible to be made by looking at the information uploaded and exposed by users in online services. It is up to independent researchers to identify, investigate and publicize these methods of inference in order to ensure public awareness of the risks and benefits incurred in embracing these services.

Bibliography

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007a). Analysis of topological characteristics of huge online social networking services. In *Proc. of the World Wide Web Conference (WWW)*.

Ahn, Y.-Y., Han, S., Kwak, H., Moon, S., and Jeong, H. (2007b). Analysis of topological characteristics of huge online social networking services. In *Proceedings of the 16th international conference on World Wide Web*, pages 835--844. ACM.

An, J., Cha, M., Gummadi, K., and Crowcroft, J. (2011). Media landscape in twitter: A world of new conventions and political diversity. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Bakhshi, S., Shamma, D. A., and Gilbert, E. (2014). Faces engage us: Photos with faces attract more likes and comments on instagram. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 965--974. ACM.

Benevenuto, F., Rodrigues, T., Cha, M., and Almeida, V. (2009). Characterizing user behavior in online social networks. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, IMC '09, pages 49--62. ACM.

Boyd, D. (2007). Why youth (heart) social network sites: The role of networked publics in teenage social life. *The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning*, pages 119--142.

Boyd, D. (2013). White flight in networked publics? *How Race and Class Shaped American Teen Engagement with Myspace and Facebook." Race after the Internet. Ed. Lisa Nakamura and Peter A. Chow-White. vols: Routledge*, pages 203--22.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and regression trees*. CRC press.

Burger, J. D., Henderson, J., Kim, G., and Zarrella, G. (2011). Discriminating gender on twitter. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1301--1309. Association for Computational Linguistics.

Cha, M., Haddadi, H., Benevenuto, F., and Gummadi, P. K. (2010). Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30.

Clauset, A., Shalizi, C. R., and Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM review*, 51(4):661--703.

Cunha, E., Magno, G., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014). He votes or she votes? female and male discursive strategies in twitter political hashtags. *PloS one*, 9(1):e87041.

Dago-Casas, P., González-Jiménez, D., Yu, L. L., and Alba-Castro, J. L. (2011). Single- and cross-database benchmarks for gender classification under unconstrained settings. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 2152--2159. IEEE.

De Choudhury, M., Lin, Y.-R., Sundaram, H., Candan, K. S., Xie, L., Kelliher, A., et al. (2010). How does the data sampling strategy impact the discovery of information diffusion in social media? *ICWSM*, 10:34--41.

de Las Casas, D. C., Magno, G., Cunha, E., Gonçalves, M. A., Cambraia, C., and Almeida, V. (2014). Noticing the other gender on google+. In *Proceedings of the 2014 ACM conference on Web science*, pages 156--160. ACM.

de Souza, F. (2015). Caracterização e análise de selfies e fotos com faces no instagram. Master's thesis, Universidade Federal de Minas Gerais.

Dey, R., Ding, Y., and Ross, K. W. (2013). The high-school profiling attack: How online privacy laws can actually increase minors' risk. In *Proc. of Internet Measurement Conference*, volume 13.

Diakopoulos, N. (2015). Algorithmic accountability: Journalistic investigation of computational power structures. *Digital Journalism*, 3(3):398--415.

Fan, H., Cao, Z., Jiang, Y., Yin, Q., and Doudou, C. (2015). Learning deep face representation. US Patent Application 20150347820.

Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330--347.

Fu, Y., Guo, G., and Huang, T. S. (2010). Age synthesis and estimation via faces: A survey. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(11):1955--1976.

Gallagher, A. and Chen, T. (2009). Understanding images of groups of people. In *Proc. CVPR*.

Gallagher, A. C. and Chen, T. (2008). Estimating age, gender, and identity using first name priors. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1--8. IEEE.

Garcia, D., Weber, I., and Garimella, V. R. K. (2014). Gender asymmetries in reality and fiction: The bechdel test of social media. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Gjoka, M., Kurant, M., Butts, C. T., and Markopoulou, A. (2010). Walking in facebook: A case study of unbiased sampling of osns. In *INFOCOM, 2010 Proceedings IEEE*, pages 1--9. IEEE.

Gong, N. Z., Xu, W., Huang, L., Mittal, P., Stefanov, E., Sekar, V., and Song, D. (2012). Evolution of social-attribute networks: measurements, modeling, and implications using google+. In *Proceedings of the 2012 ACM conference on Internet measurement conference*, pages 131--144. ACM.

Graells-Garrido, E., Lalmas, M., and Menczer, F. (2015). First women, second sex: Gender bias in wikipedia. In *Proceedings of the ACM conference on Hypertext*.

Guo, G., Dyer, C. R., Fu, Y., and Huang, T. S. (2009). Is gender recognition affected by age? In *Computer Vision Workshops (ICCV Workshops), 2009 IEEE 12th International Conference on*, pages 2032--2039. IEEE.

Han, H. and Jain, A. (2014). Age, gender and race estimation from unconstrained face images. *Dept. Comput. Sci. Eng., Michigan State Univ., East Lansing, MI, USA, MSU Tech. Rep.(MSU-CSE-14-5)*.

Hargittai, E. et al. (2010). Facebook privacy settings: Who cares? *First Monday*, 15(8).

Huang, G. B., Ramesh, M., Berg, T., and Learned-Miller, E. (2007). Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report, Technical Report 07-49, University of Massachusetts, Amherst.

Ilia, P., Polakis, I., Athanasopoulos, E., Maggi, F., and Ioannidis, S. (2015). Face/off: preventing privacy leakage from photos in social networks. In *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, pages 781--792. ACM.

James, G. M. (2003). Variance and bias for general loss functions. *Machine Learning*, 51(2):115--135.

Jang, J. Y., Han, K., Shih, P. C., and Lee, D. (2015). Generation like: Comparative characteristics in instagram. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 4039--4042. ACM.

Kivran-Swaine, F., Brody, S., Diakopoulos, N., and Naaman, M. (2012). Of joy and gender: emotional expression in online social networks. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, CSCW '12, pages 139--142. ACM.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097--1105.

Kumar, N., Berg, A. C., Belhumeur, P. N., and Nayar, S. K. (2009). Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365--372. IEEE.

Kurant, M., Markopoulou, A., and Thiran, P. (2011). Towards unbiased bfs sampling. *Selected Areas in Communications, IEEE Journal on*, 29(9):1799--1809.

Kwak, H., Lee, C., Park, H., and Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, pages 591--600. ACM.

Lee, S. H., Kim, P.-J., and Jeong, H. (2006). Statistical properties of sampled networks. *Physical Review E*, 73(1):016102.

Leskovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 631--636. ACM.

Levi, G. and Hassner, T. (2015). Age and gender classification using convolutional neural networks. In *IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) workshops*.

Lewis, K., Kaufman, J., and Christakis, N. (2008). The taste for privacy: An analysis of college student privacy settings in an online social network. *Journal of Computer-Mediated Communication*, 14(1):79--100.

Li, Y., Xu, K., Yan, Q., Li, Y., and Deng, R. H. (2014). Understanding osn-based facial disclosure against face authentication systems. In *Proceedings of the 9th ACM symposium on Information, computer and communications security*, pages 413--424. ACM.

Magno, G., Comarela, G., Saez-Trumper, D., Cha, M., and Almeida, V. (2012). New kid on the block: Exploring the google+ social graph. In *Proc. of the ACM Internet Measurement Conference (IMC)*, pages 159--170.

Marwick, A. E. (2008). To catch a predator? the myspace moral panic. *First Monday*, 13(6).

McPherson, M., Smith-Lovin, L., and Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual review of sociology*, pages 415--444.

Minkus, T., Liu, K., and Ross, K. W. (2015). Children seen but not heard: When parents compromise children's online privacy. In *Proceedings of the 24th International Conference on World Wide Web*, pages 776--786. International World Wide Web Conferences Steering Committee.

Mislove, A., Marcon, M., Gummadi, K. P., Druschel, P., and Bhattacharjee, B. (2007). Measurement and analysis of online social networks. In *proc. of ACM IMC*.

Ngan, M. and Grother, P. (2014). Face recognition vendor test (frvt) performance of automated age estimation algorithms. Technical report, National Institute of Standards and Technology.

Ngan, M. and Grother, P. (2015). Face recognition vendor test (frvt) performance of automated gender classification algorithms. Technical report, National Institute of Standards and Technology.

Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning*, pages 625--632. ACM.

Ottoni, R., Las Casas, D., Pesce, J. P., Meira Jr, W., Wilson, C., Mislove, A., and Almeida, V. (2014). Of pins and tweets: Investigating how users behave across image-and text-based social networks. *AAAI ICWSM*.

Ottoni, R., Pesce, J. P., Las Casas, D. B., Franciscani Jr, G., Meira Jr, W., Kumaraguru, P., and Almeida, V. (2013). Ladies first: Analyzing gender roles and behaviors in pinterest. In *ICWSM*.

Pesce, J. P., Las Casas, D., Rauber, G., and Almeida, V. (2012). Privacy attacks in social media using photo tagging networks: a case study with facebook. In *Proceedings of the 1st Workshop on Privacy and Security in Online Social Media*, page 4. ACM.

Polakis, I., Lancini, M., Kontaxis, G., Maggi, F., Ioannidis, S., Keromytis, A. D., and Zanero, S. (2012). All your face are belong to us: breaking facebook's social authentication. In *Proceedings of the 28th Annual Computer Security Applications Conference*, pages 399--408. ACM.

Quercia, D., Las Casas, D. B., Pesce, J. P., Stillwell, D., Kosinski, M., Almeida, V., and Crowcroft, J. (2012). Facebook and privacy: The balancing act of personality, gender, and relationship currency. In *ICWSM*.

Redi, M., Quercia, D., Graham, L. T., and Gosling, S. D. (2015a). Like partying? your face says it all. predicting the ambiance of places with profile pictures. *arXiv preprint arXiv:1505.07522*.

Redi, M., Rasiwasia, N., Aggarwal, G., and Jaimes, A. (2015b). The beauty of capturing faces: Rating the quality of digital portraits. In *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*, volume 1, pages 1--8. IEEE.

Santarcangelo, V., Farinella, G. M., and Battiato, S. (2015). Gender recognition: Methods, datasets and results. In *Multimedia & Expo Workshops (ICMEW), 2015 IEEE International Conference on*, pages 1--6. IEEE.

Souza, F., de Las Casas, D., Flores, V., Youn, S., Cha, M., Quercia, D., and Almeida, V. (2015). Dawn of the selfie era: The whos, wheres, and hows of selfies on instagram. In *Proceedings of the 2015 ACM on Conference on Online Social Networks*, pages 221--231. ACM.

Stone, Z., Zickler, T., and Darrell, T. (2010). Toward large-scale face recognition using social network context. *Proceedings of the IEEE*, 98(8):1408--1415.

Szell, M. and Thurner, S. (2013). How women organize social networks different from men. *Scientific Reports*, 3(i):20--22. ISSN 2045-2322.

Taigman, Y., Yang, M., Ranzato, M., and Wolf, L. (2014). Deepface: Closing the gap to human-level performance in face verification. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1701--1708. IEEE.

Tang, C., Ross, K., Saxena, N., and Chen, R. (2011). What's in a name: A study of names, gender inference, and gender behavior in facebook. In *Database Systems for Adanced Applications*, pages 344--356. Springer.

Terrell, J., Kofink, A., Middleton, J., Rainear, C., Murphy-Hill, E., and Parnin, C. (2016). Gender bias in open source: Pull request acceptance of women versus men. Technical report, PeerJ PrePrints.

Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. *CoRR*, abs/1111.4503.

Voelkle, M. C., Ebner, N. C., Lindenberger, U., and Riediger, M. (2012). Let me guess how old you are: Effects of age, gender, and facial expression on perceptions of age. *Psychology and Aging*, 27(2):265.

Vonikakis, V., Subramanian, R., Arnfred, J., and Winkler, S. (2014). Modeling image appeal based on crowd preferences for automated person-centric collage creation. In *Proceedings of the 2014 International ACM Workshop on Crowdsourcing for Multimedia*, pages 9--15. ACM.

Wagner, C., Graells-Garrido, E., and Garcia, D. (2016). Women through the glass-ceiling: Gender asymmetries in wikipedia. *arXiv preprint arXiv:1601.04890*.

Xu, K., Guo, Y., Guo, L., Fang, Y., and Li, X. (2014). Control of photo sharing over online social networks. In *Global Communications Conference (GLOBECOM), 2014 IEEE*, pages 704--709. IEEE.

Yadav, D., Singh, R., Vatsa, M., and Noore, A. (2014). Recognizing age-separated face images: Humans and machines. *PloS one*, 9(12):e112234.

Zhao, W., Chellappa, R., Phillips, P. J., and Rosenfeld, A. (2003). Face recognition: A literature survey. *ACM computing surveys (CSUR)*, 35(4):399--458.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., and Li, X. (2011). Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338--349. Springer.

Zhou, E., Cao, Z., and Yin, Q. (2015). Naive-deep face recognition: Touching the limit of lfw benchmark or not? *arXiv preprint arXiv:1501.04690*.

Appendix A

Estimating heavy-tailed distribution parameters

Here I will describe well-known methods of estimating distributional parameters, as well as differentiating between different possible heavy-tailed distributions for the empirical data.

An important heavy-tailed distribution that constantly emerges in OSNs is the Power-law distribution. The probability of high values in a power law distribution decays polynomially. That is, for a given value of $\alpha > 1$,

$$p(x) = Cx^{-\alpha} \tag{A.1}$$

Where C is given by the normalization requirement that

$$1 = \int_{x_0}^{\infty} p(x)\mathrm{d}x = C\int_{x_0}^{\infty} x^{-\alpha}\mathrm{d}x = \frac{C}{1-\alpha}\left[x^{-\alpha+1}\right]^{\infty}$$

This gives $C = (\alpha - 1)x_0^{\alpha-1}$, which when plugged to Equation A.1 yields the normalized expression:

$$p(x) = \frac{\alpha - 1}{x_0}\left(\frac{x}{x_0}\right)^{-\alpha}$$

Here, x_0 is the lowest possible value of x, and is the tail's offset, above which the distribution follows a power law. This means that it is possible (and indeed fairly common) that an attribute follows some distribution at lower values, and above a given threshold it becomes power-law distributed.

Another interesting and recurrent distribution with a heavy tail is the log-normal distribution. In the log-normal distribution, the logarithm of the attribute values are

normally distributed:

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}x} \exp\left[-\frac{(\ln x - \mu)^2}{2\sigma^2}\right]$$

Identifying heavy tailed distributions is important because this class of distributions must be dealt with special methods that are not normally used, for example, when an attribute follows a Gaussian distribution. For example, power laws with $\alpha < 2$ have an infinite expected value, and thus simply comparing the means of two groups of observations distributed in such a way makes no sense.

The Maximum Likelihood Estimation (MLE) method is a reliable and unbiased method to estimate the parameters for each of these two distributions. It consists of finding the set parameters θ that generate a distribution that maximizes the likelihood function $L(\theta)$, defined as

$$L(\theta) = \prod_{i}^{n} p(x_i|\theta)$$

where x_i is the i^{th} observation seen in the empirical data.

It is normally simpler to deal with the logarithm of the likelihood function, represented by $\ell(\theta)$, which is monotonically related to the L and handles the same results. Thus, a MLE estimator $\hat{\theta}$ can be defined as:

$$\hat{\theta} = \arg\max_{\theta} L(\theta) = \arg\max_{\theta} \ell(\theta)$$

The MLE estimator for both the power-law and the log-normal distributions can be found analytically.

For power laws, consider that

$$L(\alpha) = \prod_{i=1}^{n} \frac{\alpha - 1}{x_0} \left(\frac{x_i}{x_0}\right)^{-\alpha}$$

Then,

$$\ell(\alpha) = \sum_{i=1}^{n} \left[\ln(\alpha - 1) - \ln x_0 - \alpha \ln \frac{x_i}{x_0}\right]$$

Setting $\frac{\partial \ell}{\partial \alpha} = 0$,

$$\frac{n}{\hat{\alpha}-1} - \sum_{x=1}^{n} \ln \frac{x_i}{x_0} = 0$$
$$\implies \hat{\alpha} = 1 + n \left[\sum_i \ln \frac{x_i}{x_0} \right]^{-1}$$

There are many methods suggested to determine x_0. Normally it is informally set by visually inspecting the distribution. A more rigorous method, suggested by Clauset et al. [2009], is to calculate the Kolgovorov-Smirnov (KS) statistic to various possible values (or all values) of x_0 and pick the one that yields the best fit. The KS statistic D is the maximum distance between the CDFs of the data ($S(x)$) and the fitted model ($P(x)$):

$$D = \max_{x \geq x_0} |S(x) - P(x)|$$

Thus, in order to find good values for $\hat{\alpha}$ and x_0, one must calculate the MLE of α for all the candidate values of x_0 and pick the one that minimizes D.

For the log-normal distribution, consider that:

$$L(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma x_i} \exp\left[-\frac{(\ln x_i - \mu)^2}{2\sigma^2} \right]$$
$$\ln(\mu, \sigma^2) = -\frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} (\ln x_i - \mu)^2}{2\sigma^2}$$
$$= -\frac{n}{2} \ln(2\pi\sigma^2) - \sum_{i=1}^{n} \ln x_i - \frac{\sum_{i=1}^{n} \ln x_i^2}{2\sigma^2} + \frac{\sum_{i=1}^{n} \ln(x_i^2)\mu}{2\sigma^2} - \frac{n\mu}{2\sigma^2}$$

Setting the gradient to 0, with respect to μ:

$$\frac{\partial \ell}{\partial \mu} = \frac{\sum_{i=1}^{n} \ln x_i}{2\hat{\sigma}^2} - \frac{2n\hat{\mu}}{2\hat{\sigma}^2} = 0$$
$$\implies n\hat{\mu} = \sum_{i=1}^{n} \ln x_i$$
$$\implies \hat{\mu} = \frac{\sum_{i=1}^{n} \ln x_i}{n}$$

And with respect to σ^2,

$$\begin{aligned}\frac{\partial \ell}{\partial \hat{\sigma}^2} &= -\frac{n}{2\hat{\sigma}^2} + \frac{\sum_{i=1}^{n}(\ln x_i - \hat{\mu})^2}{2(\hat{\sigma}^2)^2} = 0 \\ \implies n &= \frac{\sum_{i=1}^{n}(\ln x_i^2 - \hat{\mu})^2}{\hat{\sigma}^2} \\ \implies \hat{\sigma}^2 &= \frac{\sum_{i=1}^{n}(\ln x_i^2 - \hat{\mu})^2}{n}\end{aligned}$$

A good way to examine the tail of a distribution is by a plot of its Complementary Cumulative Distribution Function (CCDF). When both the x and y axes are put in log scale, a CCDF that follows a power law is expected to lay in a straight line. This can be easily be seen from Equation A.1, since

$$\ln p(x) = \ln C - \alpha \ln x,$$

which is a linear equation with slope $-\alpha$.

Log-normal distributions have higher decay that can also be seen in a log-log plot:

$$\begin{aligned}\ln p(x) &= -\ln x - \frac{(\ln x - \mu)^2}{2\sigma^2} \\ &= -\frac{(\ln x)^2}{2\sigma^2} + \left[\frac{\mu}{\sigma^2} - 1\right] \ln x - \frac{\mu^2}{2\sigma^2}\end{aligned}$$

which is quadratic in $\ln x$, so a quadratic curve is expected.

Despite this expected visual distinction, in practice the log-normal parameters can be tuned in a way that the curve "looks like" a flat line – its curve should only be noticeable if a broader proportion of the probability space were shown. Thus, visual inspection is not a good way to access which distribution fits best to the data. A better method is to calculate the logarithm of the ratio of the likelihood of each observation given each model (log likelihood ratio). This can be expressed as

$$\mathcal{R} = \ln \prod_i \frac{p_1(x_i)}{p_2(x_i)} = \sum_i [\ln p_1(x_i) - \ln p_2(x_i)] = \sum_i [\ell_i^{(1)} - \ell_i^{(2)}]$$

where p_1 and p_2 are the probabilities of the two distributions and $\ell^{(1)}$ and $\ell^{(2)}$ are their respective log-likelihoods. Since each value of x is independent, so are $\ell^{(1)}$ and $\ell^{(2)}$, and by the central limit theorem, $\mathcal{R}$ must be normally distributed. Thus, it is possible to estimate the p-value of a given value of $\mathcal{R}$ and determine the best fit. A positive and statistically significant result means that p_1 is the best fit, while a negative statistically significant result means p_2 is the best fit. More details in Clauset et al. [2009].