CARACTERIZAÇÃO E ANÁLISE DE *SELFIES* E FOTOS COM FACES NO INSTAGRAM

FLÁVIO GONÇALVES HENRIQUES DE SOUZA

CARACTERIZAÇÃO E ANÁLISE DE *SELFIES* E FOTOS COM FACES NO INSTAGRAM

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

Orientador: Virgílio Augusto Fernandes Almeida

Belo Horizonte

Outubro de 2015

FLÁVIO GONÇALVES HENRIQUES DE SOUZA

CHARACTERIZATION AND ANALYSIS OF SELFIES AND PHOTOS WITH FACES ON INSTAGRAM

Dissertation presented to the Graduate Program in Computer Science of the Federal University of Minas Gerais in partial fulfillment of the requirements for the degree of Master in Computer Science.

Advisor: Virgílio Augusto Fernandes Almeida

Belo Horizonte

October 2015

Souza, Flávio Gonçalves Henriques de

S729c Characterization and analysis of selfies and photos with faces on Instagram / Flávio Gonçalves Henriques de Souza. — Belo Horizonte, 2015
xxiv, 74 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais – Departamento de Ciência da Computação
Orientador: Virgílio Augusto Fernandes Almeida

1. Computação – Teses. 2. Redes sociais online – Teses. 3. Instagram. 4. Selfies. 5. Análise espacial (Estatística). 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

Characterization and analysis of selfies and photos with faces on instagram

FLÁVIO GONÇALVES HENRIQUES DE SOUZA

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

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

PROF. FABRÍCIO BENEVENUTO DE SOUZA
Departamento de Ciência da Computação - UFMG

PROFA. MEEYOUNG CHA
Korea Advanced Institute of Science and Technology

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

Belo Horizonte, 16 de outubro de 2015.

For those who believe that a better world is made of better people.

Agradecimentos

Agradeço à minha família pelo suporte constante e imprescindível, que me fez chegar até aqui.

Agradeço à Priscilla pelo apoio incondicional, pelo carinho e pela paciência, em todos os momentos.

Agradeço ao meu orientador, Prof. Virgílio, que acompanhou o trabalho e sempre buscou viabilizar todos os recursos necessários, nos âmbitos técnico e pessoal, para a concretização deste estudo.

Agradeço aos meu colegas de laboratório e aos parceiros de pesquisa pelas inúmeras ideias, pelas empreitadas em conjunto e pelas trocas de experiências, as quais contribuíram de forma significativa para a minha formação.

Agradeço à CAPES e ao CNPq pelo fundamental auxílio financeiro.

Agradeço à Deus, sem o qual nada disso teria sido possível.

"Oh, now. How you know how I can see 'less you can look out my eyes?"
(Hoke, from the movie *Driving Miss Daisy*)

Resumo

A rede social Instagram tem crescido continuamente desde seu lançamento em 2010. Um tipo particular de foto postada na rede que tem atraído atenção nos últimos anos é o *selfie*, um autorretrato tirado com um dispositivo digital e publicado em um website de mídia social. Apesar da existência de alguns trabalhos que exploram certas características do Instagram, poucas pesquisas foram realizadas a respeito dos fatores associados à presença de selfies online. Esta dissertação busca preencher essa lacuna apresentando um estudo sobre as características de selfies, fotos com faces e imagens em geral postadas no Instagram. Um conjunto de dados com mais de 150 milhões de mídias do Instagram, bem como metadados dos usuários, foi coletado usando uma versão modificada de uma ferramenta de coleta especialmente desenvolvida como parte desta dissertação. Diferentes amostras foram cuidadosamente extraídas do conjunto de dados original a fim de executar três tipos de investigação: caracterização, análise temporal e análise espacial. Os resultados mostram, coletivamente, que selfies são muito populares e que são diferentes de outros tipos de conteúdo. Ao capturar quantitativamente essas diferenças, este trabalho contribui para uma melhor compreensão do fenômeno dos selfies, servindo como ponto de partida para outras pesquisas a respeito de importantes tópicos relacionados, como normas culturais digitais e design de plataformas sociais online.

Palavras-chave: Instagram, Selfies, Fotos com Faces, Computação Social.

Abstract

Instagram use is continuously raising since it was launched in 2010. One particular kind of photo posted on Instagram that has attracted attention in the last few years is the *selfie*, a self-portrait taken with a digital device and uploaded to a social media website. Despite the existence of some efforts in exploring Instagram social network characteristics, there has been little research on the factors associated with the presence of selfies online. This thesis tries to fill this gap presenting a study about the characteristics of selfies, photos with faces and general pictures posted on Instagram. A dataset of more than 150 million Instagram media, as well as users metadata, has been crawled using a modified version of a data collection tool specially developed as part of this thesis. Different samples were carefully extracted from this dataset in order to perform three kinds of investigations: characterization, temporal analyses, and spatial analyses. The results collectively show that selfies are very popular and are different from other types of contents. In quantitatively capturing those differences, this work contributes to a better comprehension of the selfie phenomenon, serving as a starting point for other researches about important related topics, such as digital cultural norms and design of social-networking platforms.

Keywords: Instagram, Selfies, Photos with Faces, Social Computing.

List of Figures

List of Tables

Contents

Chapter 1

Introduction

Instagram[1] use is continuously raising since it was launched in 2010 [Duggan et al., 2015]. As a social network focused primarily on interactions via images, Instagram receives millions of photos every day, posted by users from many places around the world [Instagram Press, 2016].

Recent research has shown that photos with faces are 38% more likely to receive likes and 32% more likely to receive comments on Instagram, even after controlling for social network reach and activity [Bakhshi et al., 2014]. Among photos containing faces, one particular kind that has attracted attention in the last few years is the *selfie*, a self-portrait taken with a digital device and uploaded to a social media website. The use of the word "selfie" on the Internet has increased so fast that the term was unanimously elected the Word of the Year 2013 by The Oxford Dictionaries [OxfordWords blog, 2013].

Nonetheless, despite the existence of some efforts in exploring Instagram social network characteristics [Jang et al., 2015; Hosseinmardi et al., 2015; Araújo et al., 2014; Silva et al., 2013], there has been little research on the factors associated with the presence of selfies online [Qiu et al., 2015; Tifentale and Manovich, 2015]. This thesis tries to fill this gap presenting a study about the characteristics of selfies, photos with faces and general pictures posted on Instagram, verifying differences and similarities among these groups of media and among the behaviors of users who post them. The results demonstrate that selfies can be a new window to study collective user behaviors, providing important insights into subjects like digital cultural norms and design of social-networking platforms.

[1] http://www.instagram.com

The main contributions of this work are summarized next:

- Development of a data collection tool for use in OSNs researches;
- A methodology for collecting Instagram data and extracting faces and selfies information;
- A characterization of selfies and users who post them, capturing the current aspects of these subjects;
- A longitudinal analyses presenting how the selfie phenomenon evolved over time, compared to other types of photos;
- A spatial analyses highlighting some characteristics of selfies at a country level;
- New insights about Instagram users and publications in general, which can be used in another researches about this OSN, beyond selfies;
- The starting point for the execution of a complementary project about selfies, which resulted in a publication.

A dataset of more than 150 million Instagram media, as well as users metadata, has been crawled using a modified version of CAMPS Data Collection Tool, a data collection program specially developed as part of this thesis. A second specialized version of the tool was used for Face++ data crawling. This tool, whose source code and documentation is available online, can be modified to suit yet another different data collection scenarios in OSNs researches.

Different samples were carefully extracted from the whole dataset of collected media in order to perform three kinds of investigations: characterization, temporal analyses, and spatial analyses. Photos with faces were identified using the Face++ API [Megvii Inc., 2013] and selfies were identified by the use of hashtags containing the word "selfie".

The characterization shows that selfie users tend to post more, have more relationships, and publish more information about themselves than users who post other types of contents. Besides, selfies also receive more likes than photos with faces and general pictures.

Temporal analyses demonstrate how selfies presented a rapid growth in the thriving period of 2012 and 2013, both in the number of posts and in the number of users. In these early years, young females dominated the presence in selfies, but today a more diverse community of users in terms of gender and age can be seen in the photos. Additionally, temporal analyses confirm that selfies tend to receive more attention than other types of contents on the network, allowing to explore in more detail how their popularity has changed over time.

Finally, spatial analyses give an overview about the distribution of selfies, photos with faces and general Instagram pictures around the world. When considering the relative expected amount of pictures per country, there are some countries which show a higher number of selfies, but low number of photos with faces, while for other countries the opposite is true. Together with the patterns found for gender distribution and smiling tendency, this result indicates that factors such as demography and culture should be taken into account in deeper investigations about selfies.

Some of these cultural factors associated with selfies were explored in a complementary project developed using part of the data collected for this thesis. This project resulted in a paper which was accepted at the ACM Conference on Online Social Networks 2015 (COSN'15) [Souza et al., 2015]. Besides culture, the paper also includes the analysis of photos containing alternative hashtags related to selfies (e.g., `#selca` and `#me`), and the study of selfies as an interaction medium, analyzing gender and age homophily between users who post and like/comment selfies.

This thesis is organized as follows: Chapter 2 details the motivation behind this work and presents the results of other related researches; Chapter 3 comments about the data sources (Instagram and Face++), the data collection process (including CAMPS Data Collection Tool), and the final datasets used in this work; Chapter 4 contains a static analysis of selfies, photos with faces and general Instagram pictures, as well as the users who post them; Chapter 5 presents a dynamic view of the datasets, examining users information in two points in time (December 2014 and June 2015) and photos information for a three-and-a-half-year period, between January 2012 and June 2015; Chapter 6 explores variations in the datasets across countries; Chapter 7 concludes the work, also commenting about some limitations and possible future directions. The complete version of the paper published in the proceedings of ACM COSN'15 can be found in Appendix A.

Chapter 2

Background

Selfies are a ubiquitous phenomenon of digital culture. The term was named Oxford Dictionaries Word of The Year 2013 [OxfordWords blog, 2013] and included in the online version of the dictionary with the following definition: "a photograph that one has taken of oneself, typically one taken with a smartphone or webcam and shared via social media"[1]. This definition gives room for many variations in the content of selfies. In fact, although the classical concept of a selfie generally refers to a single person self-portrait, like the one in Figure 2.1(a), it is common today to find selfies that include other people as well, as exemplified in Figure 2.1(b), and even selfies of body parts other than the face (like arms, legs and thorax) .

Self-portraiture, of course, is not new, and sharing self-portraits likewise pre-dates the Internet [BBC, 2013; Tifentale and Manovich, 2015; Tifentale, 2014]. The rise of selfies, however, is a recent trend in the visual Web, assisted by new technological tools (such as smartphones, webcams and digital cameras) and services (like Flickr, Pinterest, and Instagram) that allow people to better express themselves visually.

2.1 The Meaning of Selfies

Different theories have emerged to explain why people take selfies. Some consider selfies are a mean of self-exploration. As one takes multiple selfies and combine them with different filters, one can re-see herself [Crook, 2014]. A slightly different view is self embellishment, which is grounded in psychological experiments that show people, when exposed to slightly modified pictures of themselves, tend to identify a more attractive version as the original picture [Kilner, 2014]. For others, yet, selfies are a better way of communicating feelings and emotions than text, because they convey facial expressions

[1] http://www.oxforddictionaries.com/us/definition/english/selfie

(a) American actor James Franco, a defender of the selfie culture, and a very active user of Instagram.

(b) Ellen DeGeneres's 2014 famous Oscar selfie attracted so much attention that was worth between $800 million and $1 billion according to an advertising firm working for Samsung[2].

Figure 2.1. Two different selfies. (a) A single person "true selfie". (b) A multiple person selfie.

and place the person into the message [Wortham, 2013]. With the ability to control picture's aesthetics, selfies are a perfect tool for showing the world one's subjective self-image (or a constructed self-image one wants the world to see) [Day, 2013].

A sociological framing recognizes technological possibility to be a necessary condition and also highlights other behavioral factors to be important for selfies [Cole, 2015c]. One is a culture of sharing and belonging fostered by the online environment and transmitted through memes. Another is the constant work of shaping and reaffirming self-identity through social actions. In this perspective, selfies are more than a mere collection of individual pictures but a convention governed by culture and society.

Moving forward, selfies are being interpreted by some as an emerging sub-genre of self-portraiture [Tifentale, 2014]. In fact, others already recognize them as an art form, like the curators of the National #Selfie Portrait Gallery launched at the contemporary video art fair Moving Image London 2013[3], and the creators of Selfeed[4], a website which shows an endless stream of selfies posted on Instagram in real-time. As an art form, selfies pose a stylistic structure, where a face is normally in the foreground and details are in the background to transmit elements of a rhetorical scene. In many selfies, the equipment that captures the photo is present within the frame (see Figure 2.1(a)), which confers credibility by evincing the technological mediation [Losh, 2014].

[2]http://www.nbcnews.com/tech/social-media/ellens-oscar-selfie-worth-1-billion-n75821
[3]http://www.moving-image.info/national-selfie-portrait-gallery
[4]http://www.selfeed.com

2.2 Advocates and Opponents

Selfies are a prominent online culture that have been both criticized and advocated by different parties. Some critics say selfies are vain, narcissistic, and attention-seeking; they argue a wide adoption of selfies by women reflects self-objectification and male gaze [Cole, 2015a]. Self-objectification is also known to be positively correlated with increasing photo sharing activities on Facebook among young women [Meier and Gray, 2014].

Others argue selfies increase demand for plastic surgery. The American Academy of Facial Plastic and Reconstructive Surgery reports that 33% of surgeons have seen an increase in requests for plastic surgery as a result of patients being more self aware of their looks because of social media [Winneberger, 2014]. Another research also demonstrates that adults who own personality traits known as the Dark Triad (narcissism, psychopathy and Machiavellianism) have a higher chance of posting selfies and edited images on social networks [Fox and Rooney, 2015]. This leads to a worry about the loss of control over one's self-image in an increasingly sharing and *hackable* culture, where the mere presence of a person's picture in a photo collection can reveal a large amount of information about that person [Dey et al., 2014]. This kind of worry takes on special relevance given recent news related to the use of facial recognition technology in surveillance operations, raising concerns over privacy issues [Risen and Poitras, 2014].

Defenders of the selfie culture not only deny the previous claims but argue selfies are the pinnacle of control and self-expression; selfies allow people to take control over how they and their peers are represented in the public, which mobilizes the power dynamics of representations and promotes empowerment [Cole, 2015b]. One study interviewed 20 participants who had posted sexual self-portraits and showed how the exchange of such self-portraits can be a transformative experience, increasing their self-awareness in a positive manner [Tiidenberg, 2014]. Being able to reclaim the representation of their bodies, people can rethink their concepts of beauty and dissociates what advertisers want them to believe a beautiful photography is from what they believe a beautiful photography should be [Gervais, 2013].

Finally, in the special case of celebrities, selfies are a way to create a direct channel between them and their fans. A selfie from a celebrity is not only a private portrait of a star, but one also usually composed and taken by the own star [Franco, 2013]. Thus, it brings the public closer to the celebrity than the conventional media does, because the selfie becomes much more than a picture of the private life of a famous person: it reveals the explicit intention of that person to register and share an intimate moment with those who follow her.

2.3 Researches on Selfies

In contrast to the rich body of works on sociological interpretation of selfies, relatively little attention has been given to data-driven analysis of the subject.

In the intersection of computer science and psychology, Qiu et al. [2015] examine the association between selfies and personality, as well as zero-acquaintance personality judgment, by measuring participant's personality traits (agreeableness, conscientiousness, neuroticism, and openness) and coding their selfies using a series of cues. The results show that selfies reflect their owner's personality traits, but observers could only accurately judge selfie owner's degree of openness, which differs from findings of previous researches. The authors discuss the possible relationship between this difference and the impression management carried out by social media users. In a subsequent study, Chandra et al. [2015] enhances the previous framework for selfie owners personality analysis by constructing an automatic personality prediction model. They employ visual features as low-level cues and personality cues as mid-level cues. Low-level cues are extracted from selfies and used to train mid-level cue detectors, which are then used to predict users' personality.

Exploring the content of pictures, Joshi et al. [2014] proposes a method to find clusters of selfies on Twitter. After extracting faces, the photos are clustered using visual similarity and them ranked based on average visual similarity among faces and average size of faces. Yeh and Lin [2014] aim at helping users to take aesthetic better selfies, focusing on angle to evaluate visual quality. They compute patterns from a dataset of profile pictures and combine then with head pose estimation and camera orientation, building an algorithm able to recommend a good look before the photo is captured.

A report by eBay Deals Blog describes the top 25 celebrity accounts on Twitter and Instagram by number of selfies posted. Among the most famous Twitter users, women dominate the field when it comes to selfies. Instagram selfies, however, show a much more even spread of men and women. According to the report, Instagram is used for vastly more celebrities who post selfies than Twitter [eBay Deals Blog, 2013].

A research conduced by TIME looked at how many "selfies per capita" were produced in 459 cities by dividing the amount of users posting selfies by the population of each city. They noticed that it was difficult to find a proper local translation for the hashtag `#selfie`, as different variations were used everywhere [Wilson, 2014].

The largest scale analysis of selfies to date, however, probably was a data visualization project called Selfiecity[5]. As described by Tifentale and Manovich [2015],

[5]http://www.selfiecity.com

the data collection for the project happened in 2013, gathering Instagram photos from different locations. After sampling and manually filtering the original data, the final dataset comprised 3,200 single person self-portraits from five big cities around the world. Aiming at showing that no single interpretation of the selfie phenomenon is correct by itself, they extracted over 20 measurements using computer analysis and built a series of data visualizations upon them. Among the main results, they found people take less selfies than often assumed (only 3–5% of the images in their dataset were actually selfies), females take significantly more selfies than males, and most people in the photos are pretty young (23.7 estimated median age).

Nonetheless, many aspects of selfies have not yet been studied under the perspective of data analysis, remaining as open topics for new researches. This thesis tries to fill this gap presenting a study about the characteristics of selfies, photos with faces and general pictures posted on Instagram. The differences and similarities among these groups of media are explored, as well as differences and similarities in the behaviors of the users who post them.

2.4 Related Work

This section describes several findings from researches on three topics related to this work: photos with faces, Instagram, and online photo sharing.

2.4.1 Photos With Faces

One source of inspiration for this thesis was the research of Bakhshi et al. [2014]. In their study they also use a sample of Instagram photos in conjunction with Face++ to identify photos with faces. They investigate how the presence of a face, it's age and gender might impact social engagement on the photo. They find that photos with faces are 38% more likely to be liked and 32% more likely to be commented on, but that the number of faces, their age and gender do not have significant impact.

On a different line of research, Redi et al. [2015b] design numerous visual features based on portrait literature and extract them from a large annotated dataset of portraits. Next, they study the correlations between features and beauty, and find that facial features are the most significant in guiding portrait aesthetics, while physical/demographic properties such as gender, eye, color, glasses, age, and race show very low correlation with image beauty. A classifier built with the proposed features outperforms a generic classifier in the task of distinguishing beautiful and non-beautiful portraits.

Changing the classification subject to ambiance instead of beauty, Redi et al. [2015a] try to determine what kind of visual cues can be used to infer a place's ambiance, clientele and activities by observing the profile pictures of its visitors. They propose predictors based on aesthetics, colors, emotions, demographics and self-presentation, comparing people classifications with algorithmic classifications. People and algorithm does not always agree in the use of predictors, but while the machine is more accurate at times, humans perform better in other occasions.

2.4.2 Instagram

Previous works have explored Instagram at multiple levels, from a comparative study of differences between teens and adults [Jang et al., 2015] and an investigation of general users practices [Araújo et al., 2014] to a broad analysis of users activities, demographics, social network structure, and user-generated content [Manikonda et al., 2014]. In particular, Hu et al. [2014] adopt a mixed approach, categorizing photos posted on the network and then verifying how do users differ based on the types of images they post. Two of the proposed categories are *Selfies* and *Friends*, which remarkably represent nearly half of the photos in their dataset, with slightly more self-portraits.

Instagram can also be viewed as a proxy to study online user behaviors. For instance, media comments can be examined to detect cyberbullying and cyberaggression incidents [Hosseinmardi et al., 2015], Instagram photos shared on Twitter can be used as sensors to study users characteristics in different cultures [Silva et al., 2013], and the mobile application itself can serve as a tool to investigate how users communicate their experiences while visiting a museum and work to construct their own narratives from their visits. Using the Cultural Analytics framework , Hochman and Schwartz [2012] and Hochman and Manovich [2013] also show how the spatio-temporal visualization of large sets of Instagram images can offer social, cultural and political insights about people's activities in particular locations and time periods.

2.4.3 Online Photo Sharing

Online visual communication and sharing are increasingly gaining attention from the research community. In a recent work, Ottoni et al. [2013] analyze users activities and characteristics on Pinterest, focusing particularly on gender related questions. Data from Pinterest is also used by Totti et al. [2014] to evaluate the power of different

visual features (aesthetical properties and semantical content) on photos popularity, comparing with the predictive power of social cues.

Profile pictures on Facebook are examined by Huang and Park [2013] to demonstrate self-presentation differences between East Asians and Westerns, thus confirming previous findings of social psychology about cultural variations. Kim and Gweon [2014] interview Facebook users in order to uncover privacy preferences of a person in a photo (subject) according to her relationship with the person sharing the information (owner) and the person receiving the information (viewer). Interviews are executed by Miller and Edwards [2007] too, this time with Flickr users, to explore several practices that have evolved around online sharing websites, and how those practices contrast with more traditional digital photo sharing. Flickr photos provide yet the base for Schifanella et al. [2015] in the task of find beautiful pictures from the immense pool of unpopular items aided by computer vision methods, and for Crandall et al. [2009] in their approach to automatically identify places that people find interesting (both at city and landmark scales) and predict these locations from visual, textual and temporal features.

Finally, some studies use photos from various online sources to investigate subjects such as photo filtering [Bakhshi et al., 2015], visual persuasion [Joo et al., 2014], and children's online privacy [Minkus et al., 2015].

Chapter 3

Data Collection

Instagram is a free online social network (OSN) for photo and video sharing, whose main functionalities are available through its mobile application. It was launched on October 2010 for iPhone only and rapidly gained popularity, reaching 10 million users less than a year after. It was named "iPhone app of the year" by Apple on December 2011 and was bought by Facebook 4 months later, right after launching a version for Android. Today, Instagram is also available for Windows Phone and enables its users to share contents on a variety of other social networking platforms, such as Facebook, Twitter and Tumblr. Instagram has recently reached 400 million monthly active users, with more than 40 billion pictures shared on the network [Instagram Press, 2016].

The OSN offers the possibility to view public user profiles on the web, as well as to execute some actions using a web browser after logging in on its website. However, the full range of features is available exclusively in the mobile application, and, likewise, new accounts can be created solely through the app. This enforces Instagram's focus on mobile devices and its nature as an OSN to capture and share moments of life on the go.

Figure 3.1 shows two of the screens that can be accessed on Instagram mobile application. The app lets the user to take a picture or a short video (3 to 15 seconds long), edit its visual and metadata properties and post it on the network, with the option to share it on other OSNs too. In the case of metadata properties, users can add, for instance, hashtags, geolocation information, and a caption for the picture. Among visual properties, Instagram is particularly famous for its rich gallery of filters. Filters are pre-defined modifications a user can apply to a photo when posting it on Instagram. Each filter represents a different set of modifications, changing aspects of the picture like color, tint, shade, exposure, contrast, and saturation, among others. Figure 3.2 shows the result of some filters applied to the same photo.

(a) User profile information.

(b) Photo information.

Figure 3.1. Two example screens of Instagram mobile application.

Social interactions in the network can occur at user level or at post level. At user level, it is possible to "follow" or be "followed by" another users. The relationship is not symmetric, meaning that it is possible to follow someone and not be followed back, and vice-versa. Posts, followers and follows counts can be seen at the top of Figure 3.1(a).

At post level, it is possible to "like" or "comment" a post. In comments and captions, a user can "mention" another user by typing its username preceded by an "@" sign. Moreover, a post can be tagged by writing a word preceded by the hash sign ("#") in its caption field or in its comments, as can be seen in Figure 3.1(b). Posts tagged with the same hashtag can be searched both on the app and on Instagram website. This makes hashtags not only a method to associate ideas to posts but likewise a form of community interaction in the OSN, allowing similar users to group their posts around the same concepts.

(a) Normal (original picture, no filter applied)

(b) Valencia

(c) Amaro

(d) X-Pro II

Figure 3.2. Example of different filters applied to the same photo on Instagram.

All user accounts are public by default, which means that every content published by a user can be viewed by anyone, including people from outside the OSN, because all user's publications are available on Instagram website as well as on the mobile application. In addition, public users can't prevent other users to follow them (although followers can be blocked later). This setting can just be altered through the 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.[1]

Instagram offers an Application Programming Interface (API) that allows developers to access many of the data published on the network. By making requests to Representational State Transfer (REST) endpoints, it is possible to obtain information of users, media (photos and videos), relationships and comments, among other data types.

[1]Instagram 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. Contents sent via "Instagram Direct" can only be viewed by the recipients no matter if the sender's account is public or private.

To use the API it is necessary to log in on Instagram developer's web page[2] using the credentials of a valid user created through Instagram mobile application. Once logged in, it is possible to create a new API *client*, which receives a *client ID* and *client secret* from the system. The ID and secret are required to authenticate the client when connecting to the API.

3.1 Face++ Overview

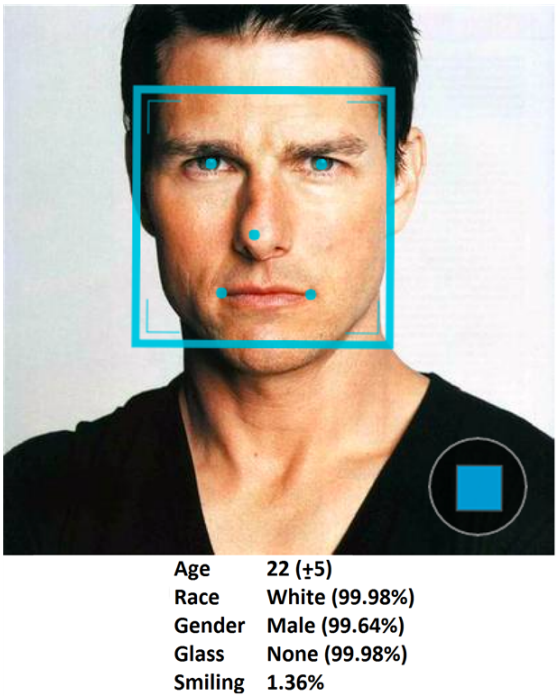
Face++ is a platform that uses computer vision and data mining to provide 3 core vision services (detection, recognition, and analysis), available through an online API. Face++ can detect faces in photos along with some information about each person in the photo, such as age, gender, and smiling, as can be seen in Figure 3.3. Age is given in years along with a confidence range; gender is given as "Male" or "Female", with a confidence value between 0% and 100%; smiling is given as a percentage, where 100% means a large smile and 0% means no smile at all. The high accuracy of Face++'s detection algorithm has been demonstrated by Bakhshi et al. [2014].

Face++ has a similar infrastructure for developers as that of Instagram. After create an account on Face++ website[3], it is possible to set up new *applications* to have access to the API. 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. All Face++ services are available through requests to specific REST endpoints defined in the API.

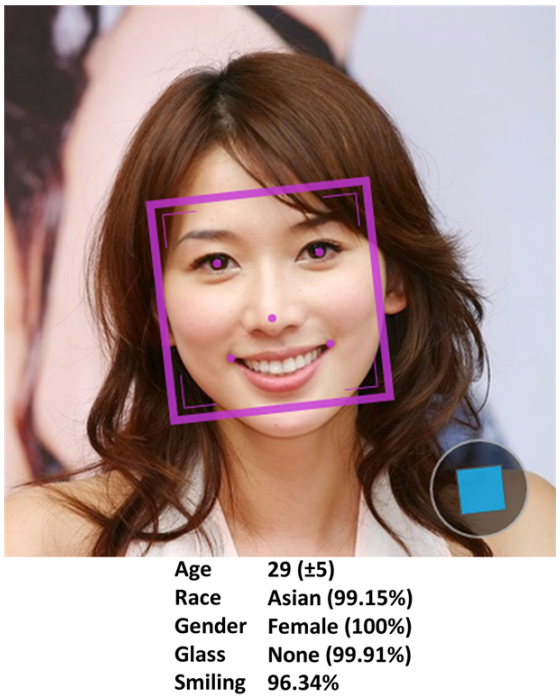
3.2 Challenges and Solutions

To make the data collection process as fast as possible, the ideal strategy is to collect each independent data unit in parallel. This requires, however, some kind of coordination to avoid both data duplication and data loss, guaranteeing that every data unit is collected *once* and *only once*. Taking this into consideration, a simple client-server program was initially used for the first Instagram crawling attempts. This program was an adaptation of other ones developed in previous works [Magno et al., 2012; Ottoni et al., 2013] of our research group. It consisted basically of a crawler script (client side) responsible for collecting the data, and a server script (server side) responsible for distributing data units among crawlers.

[2]http://www.instagram.com/developer
[3]http://www.faceplusplus.com/



(a) Male face with low smiling value.



(b) Female face with high smiling value.

Figure 3.3. Examples of some information returned the by the Face++ API face detection service.

Although that program solved the main problems of data collection distribution, it soon proved to be too restrictive to help with other difficulties that were found along the way. One of these difficulties was related to limitations of Instagram API. For all endpoints, the API establishes a maximum of 5,000 requests per client ID per hour. Each Instagram account can create a maximum of 5 client IDs, giving thus a total of 25,000 requests per hour per account. So, to increase the limit of number of requests per hour it was necessary to create more accounts and more client IDs. But it them brought a new challenge: how to distribute client IDs among crawlers such that the number of available requests was maximized? And, specially, how to do this distribution considering that the number of crawlers was usually greater than the number of client IDs? The best approach seemed to be to centralize this task on the server side, but the current server script at the time was not prepared for that.

A second challenge was the collection of Face++ data. When the adaptation of the data collection program was made, it was intended for Instagram, so the program was entirely customized to the particularities of Instagram API. It was not possible, then, to simply reuse it to crawl Face++ data in a distributed manner, as was desired, without doing a great number of modifications.

These kinds of difficulties led to the consideration that it would be beneficial to spend some time improving the collection program and creating a generic version of it that could be more easily adjusted to different crawling setups. Hence, a step back was taken and a new distributed (client-server) data collection program was developed: the CAMPS Data Collection Tool.

The tool not only solved the previously mentioned issues but also improved the control over the entire crawling process. Besides, other data collection possibilities could be further explored as, for instance, different alternatives for data persistence. Because of its flexible and simple nature, the tool can be very useful for other projects in the future as well, what makes it one of the contributions of this thesis.

3.3 CAMPS Data Collection Tool

Although performance was always a concern during the development of the tool, the main goals were (re)usability, flexibility and extensibility, what explains many of the design and implementation decisions.

Every data unit that has to be collected is called *resource*, and is identified in the program by an ID. For example, the resources could be web pages, which are identified by their URL, or they could be users in a social network, which are usually identified by an user ID string (like in Instagram). The code to do the actual resource crawling must

be written by the user, as this task varies according to resource type, resource origin and other details related to the data in hand. The tool manages the distribution of resources to be collected among multiple clients, enabling and coordinating the crawling of various resources simultaneously.

CAMPS Data Collection Tool source code is available on the Internet, along with usage instructions and modules documentation[4]. Sections 3.3.1 and 3.3.2 present an overview of some aspects of the tool as the time of writing of this thesis.

3.3.1 Usage Overview

To set up the collection program, the first step is to write the crawler code to perform the collection of a resource. After that, it is necessary to adjust the appropriate settings inside a XML configuration file. The XML file holds the values for all configuration options used, both for server and client sides.

The next step is to initialize the server, and, when it is running, start as many clients as needed. A manager program comes with the tool, allowing to monitor the data collection process, as well as to perform some actions upon the server (like remove clients or shut it down).

From a high level point of view, the necessary steps to set up and use the tool could be summarized as follows:

1. Implement the crawling code
2. Create a XML configuration file and adjust the appropriate settings inside it
3. Run the server on the desired machine
4. Run as many clients as needed on the desired machines
5. Monitor and manage the collection process using the manager program
6. Wait for the collection to finish

Usage instructions and details about the set up process can be found in the project's online wiki[5].

3.3.2 Architecture Overview

The program was written in Python 2.7.5 and tested under Linux and Windows. The code is divided in 8 modules: `server`, `client`, `manager`, `serverlib`, `crawler`, `filters`, `persistence` and `common`. The first 3 are executable modules, while the remaining ones are importable modules.

[4] http://www.github.com/fghso/camps-dct
[5] http://www.github.com/fghso/camps-dct/wiki

3.3.2.1 Modules

The 8 modules that comprise the tool can be splitted into 2 different categories: *core modules* and *customizable modules*.

Core modules include the server's library (`serverlib`) and the executable files to run the services provided by the tool (`server`, `client` and `manager`). These modules should not be modified, unless some special requirement is needed that could not be obtained through modifications of the customizable modules.

The customizable modules, on the other hand, hold the most flexible parts of the tool, which can be adapted or extended to suit user needs. The `crawler` module must be modified in order to the program to be useful. Modifications in the other modules (`filters`, `persistence` and `common`) depends on specific details related to the data collection scenario.

A brief description of each module follows:

server

Executable module that just do some initialization procedures before passing control to the `serverlib` module. The procedures include parse command line arguments and load the XML configuration file.

client

Executable module that holds the client side logic of the tool. Besides also parse command line arguments and load the XML configuration file, this module makes all contacts with the server to request a resource and return the collection results. The data collection itself, however, is done by the `crawler` module.

manager

An independent executable module that permits to monitor the collection process and execute some management actions upon the server. This module is not necessary to run the tool, but is nevertheless included in the category of *core modules* because it is not intended, at first, to be modified in an ordinary use of the program.

serverlib

The central module of the tool. The server's main loop resides in this module and is where all the work is coordinated.

crawler

Main customizable module. The user must modify this module, implementing the necessary code to crawl a resource, in order to the tool to be really useful.

filters

The module that stores all filter classes. Filters are segments of code that can be executed in the server side before a resource is sent to a client and/or after a resource has been crawled. Filters are not essential to the regular work of the server though. They were implemented as a way to easily allow pre- and post-processing of resources when the scenario requires it.

persistence

All persistence handlers classes resides in this module. The main purpose of persistence handlers is to load and save information about the resources being processed. The handlers in this module provide a common interface to persistence operations, freeing the server of the duty to know if the resources are being stored in a file or in a database and how to deal with the particular details of each persistence alternative.

common

This module contains shared code. Functions and classes stored within it are meant to be reused in different parts of the program. Some of them were built just for internal use, but others can also be incorporated in user code as well, and that is why it is included in the category of *customizable modules*.

Classes and functions of customizable modules are extensively documented in the project's website[6], with explanations aimed towards developers who want to modify or extended the tool.

3.3.2.2 Workflow

Figure 3.4 shows how the tool's modules are organized around each other (the `common` module is not depicted because it is used by many of the other modules). Blue boxes indicate executable modules. The main server and client libraries appear in green boxes. Filters and persistence modules are represented by yellow and red boxes, respectively. The arrows indicate communication between modules, what happens either when a module import another one or when the modules exchange network messages.

To give a better idea about how the modules work together, one request round of the program is described next:

[6] http://fghso.github.io/camps-dct/

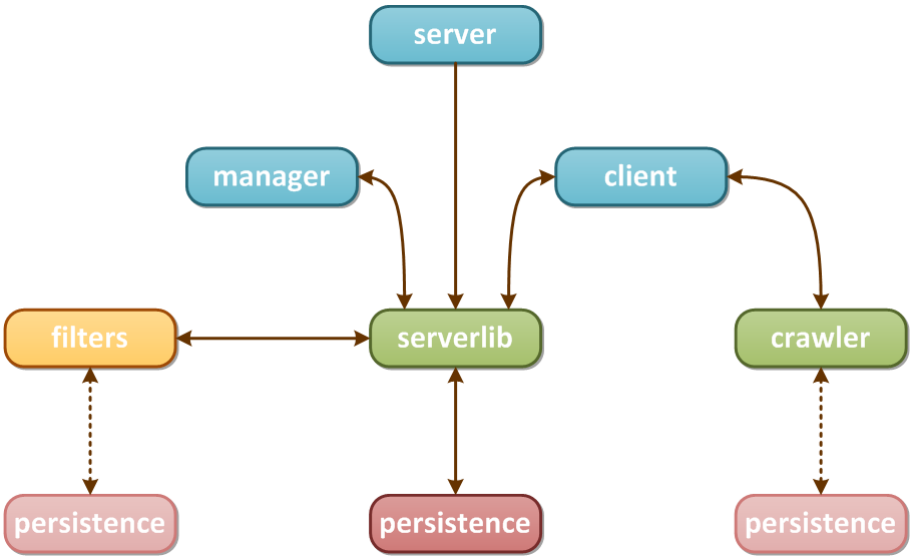


Figure 3.4. CAMPS Data Collection Tool architecture.

1. Server starts
2. Client starts
3. Server receives a client request for connection. It starts a new thread to handle requests for this new client
4. Client requests a new resource ID to crawl
5. Server communicates with persistence handler to obtain a resource ID not yet crawled
6. Having a resource ID to send, server first applies filters (if there are any filters to apply)
7. Server sends the resource ID and filters results to client
8. Client calls user code to do the actual crawling of the resource
9. Client sends the results of the crawling process to server
10. Server calls back filters (if there are any filters to call back)
11. Server communicates with persistence handler to save the status of the crawling process for that resource
12. Back to step 4

In a regular use of the tool, the first module that should be run is the `server` executable module. After the initialization procedures, this module transfers control to the `serverlib` module, where resides the code to handle clients requests. For each new client a new thread is started encapsulating a loop where all transactions between the server and this new client are dealt with.

Clients are started by running the `client` executable module. This module takes care of all interactions with the server, leaving the real process of data collection to

the user code stored in the `crawler` module. The communication link between the `client` module and the `serverlib` module is the backbone of the client-server model employed in the construction of the tool.

When filters are specified in the XML configuration file, `serverlib` is also responsible for applying them before a resource is sent to a client and calling them back after a resource is crawled.

At the bottom level of the architecture is the `persistence` module, containing persistence handlers. The `serverlib` module always uses a persistence handler to load and save resources. However, the `persistence` module could also be used by filters or even by the crawler if the user wants to (this optional use of the `persistence` module is indicated in Figure 3.4 by the dashed arrows).

Finally, management operations are performed by the `serverlib` module in response to requests made by the `manager` module.

3.4 Methodology

The data collection process started by inferring the range of the space of integer identification numbers (IDs) assigned to Instagram users, as well as the distribution of the IDs throughout this space. The IDs space was sampled with batches of 10 thousand random numbers for every contiguous interval of 10 million IDs, starting from ID zero. Although the process was carried out until reach the ID 2 billion, no valid user could be found after the batch 158. This way, it was possible to conclude that user IDs space on Instagram have a range of approximately 1.58 billion. The distribution of IDs in the IDs space can be seen in Figure 3.5. The IDs do not follow a uniform distribution, but it was not possible to find out why. Anyway, no correlation was found between the distribution of IDs and the users information analyzed in this thesis.

Once the user IDs space was identified, a random sample of 16 million IDs ($\sim$1% of the IDs space) was taken as an initial users set, which was inspected at two points in time: once in December 2014 and once in June 2015. In both occasions, 42% of the initial set corresponded to IDs of valid Instagram users. Valid users are further divided into public and private users. Besides, privacy settings can be changed at any time. So, considering the two inspections made, 30% of the 16 million IDs corresponded to public users (users whose accounts were public both in December and June).

The profile information of every public user (4,776,449 in total) was collected, as well as all publications (known as "feeds" on Instagram) of each one of them for a three-and-a-half-year period, between January 2012 and June 2015. The feeds data collection

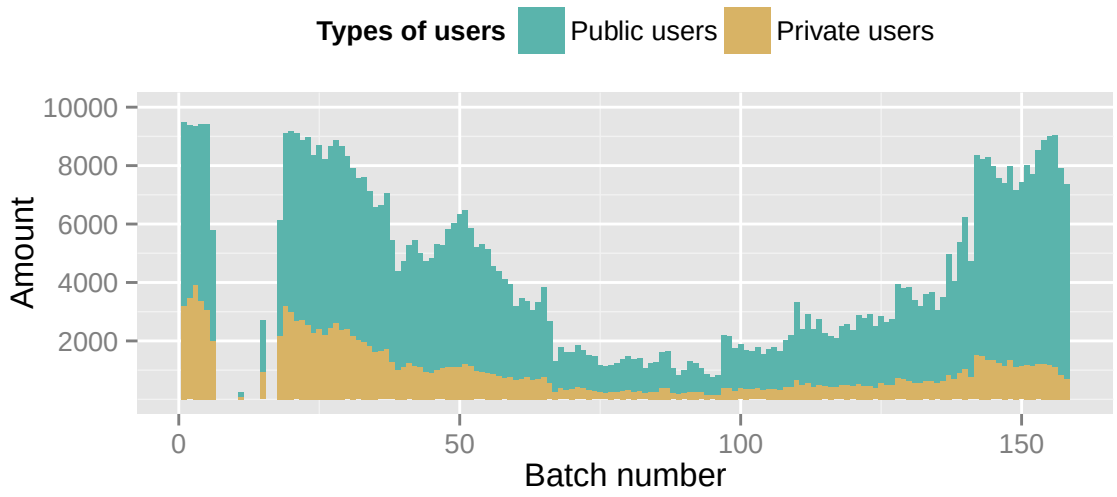


Figure 3.5. Distribution of IDs throughout Instagram user IDs space. Batches are numbered from 1 to 200.

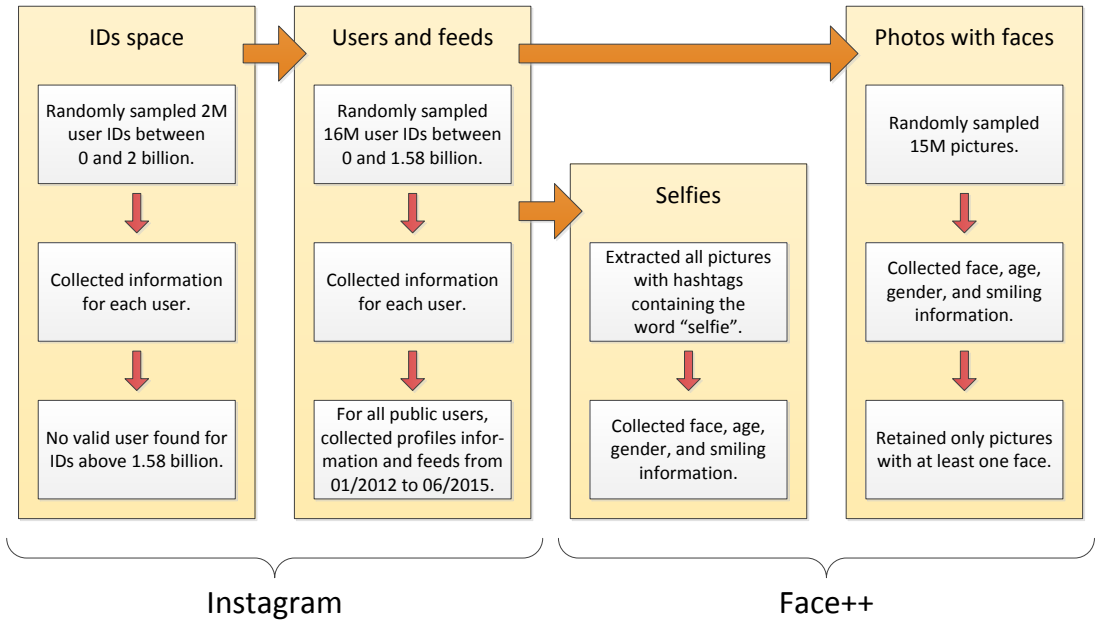


Figure 3.6. Description of the data collection process.

Table 3.1. List of all information collected and used in this thesis.

Users	Photos	Faces
full name	num. comments	num. faces
bio	num. likes	age
website	created time	gender
num. media	image url	smiling
num. followers	users in photo	
num. followees	filter	
	hashtags	
	latitude	
	longitude	
Instagram		Face++

resulted in 169,030,032 media objects, which include photos and videos metadata. For this study, only photos (the majority of the media objects: 164,114,868) were considered.

Selfies were extracted based on hashtags: all pictures with at least one hashtag containing the word "selfie" (for instance, `#selfie`, `#selfietime`, and `#selfiesunday`) were considered as selfies. Face++ API was used to identify photos with faces and to detect faces in selfies, as well as to obtain age, gender and smiling information about each face found. Figure 3.6 summarizes the whole data collection process and Table 3.1 presents details about the data collected and used in this thesis.

3.5 Final Datasets

Three datasets were built from the data collected:

- all: pictures randomly sampled from the set of all photos collected.
- faces: pictures with at least one face, randomly sampled from the set of all photos collected.
- selfies: all photos with at least one hashtag containing the word "selfie".

Table 3.2 presents the number of media, number of faces, and number of users in each dataset. Numbers are shown divided between geotagged pictures and non-geotagged pictures, along with a total. In the case of number of pictures and number of faces, the total represents the sum of geo and non-geo. In the case of users, however,

Table 3.2. Description of the datasets built from the data collected.

		Num. pictures	Num. faces	Num. users
all	Geo	3,516,724	-	426,298
	Non-geo	11,483,276	-	1,442,577
	Total	15,000,000	-	1,543,805
faces	Geo	504,546	1,065,817	221,574
	Non-geo	1,828,855	3,549,882	776,879
	Total	2,333,401	4,615,699	879,718
selfies	Geo	347,950	400,938	66,294
	Non-geo	1,078,193	1,050,522	180,552
	Total	1,426,143	1,451,460	212,119

the total is not equal the sum of geo and non-geo because the sets of users who posted geotagged photos and non-geotagged photos overlap.

To examine the local context in which photos were posted it was necessary to map the location coordinates of geotagged pictures to geographic names. The Global Administrative Areas database (GADM) [Global Administrative Areas, 2012] was used to map all valid latitude and longitude data to actual country names.

Chapter 4

Characterization

To study differences and similarities between selfies, photos with faces and general Instagram pictures, a characterization is conducted in this chapter taking into account all data available in each of the three datasets. The characterization of users explores the number of media they published, the number of followers and followees they have, and the proportion of users who shared some personal information. In the case of photos, the characterization includes interactions (likes and comments), hashtags and filters usage, and faces information (number of faces, age, gender, and smiling values).

4.1 Users

Instagram API provides counts for number of media (photos and videos), number of followers, and number of followees of each user. Figures 4.1, 4.2 and 4.3 show ECDFs and boxplots for these counts, plotted in $\log_{10}$ scale given the skewed nature of the data. Users with no followees and/or followers are not included in the plots, but in both cases they represent less than 3.6% of all users in the three datasets.

It is possible to see that users who post photos with faces have a slightly higher tendency to publish more photos and have more followers and followees than users who post any type of content on the network. For selfies users, however, this tendency is more pronounced. For example, while 77% of users in all have at most 100 published media, this number drops to 42% for users in selfies; and while 50% of users in all have at most 100 followers, this proportion in selfies drops to 19%.

Users can optionally fill some personal information in their Instagram profiles, and these information can be retrieved as part of users metadata. There are 3 available free text fields which users can complete as they wish: full name, bio, and website. Table 4.1 shows the proportion of users in each dataset who have put any kind of information

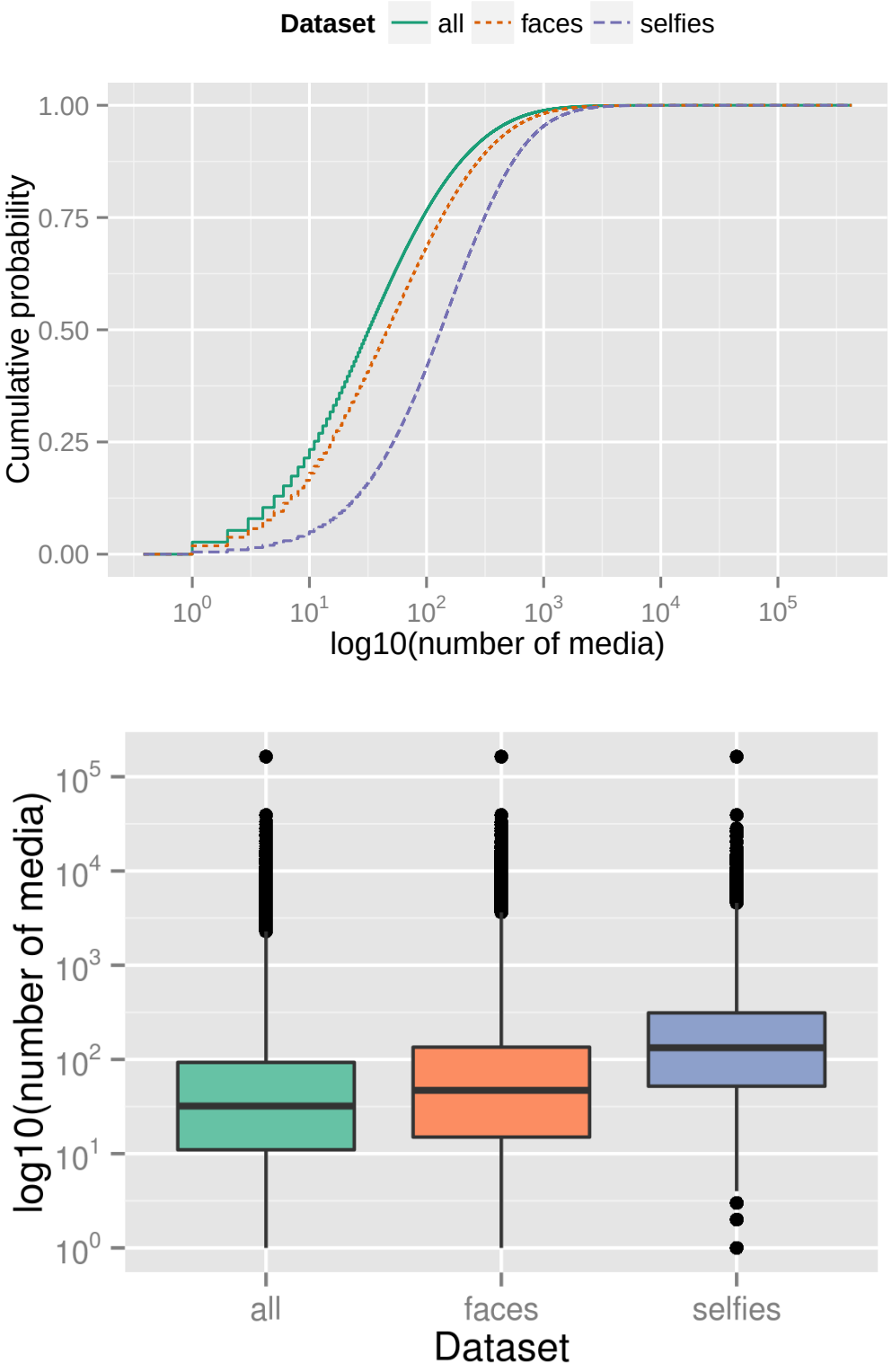


Figure 4.1. ECDF and boxplot for number of media by user in each dataset.

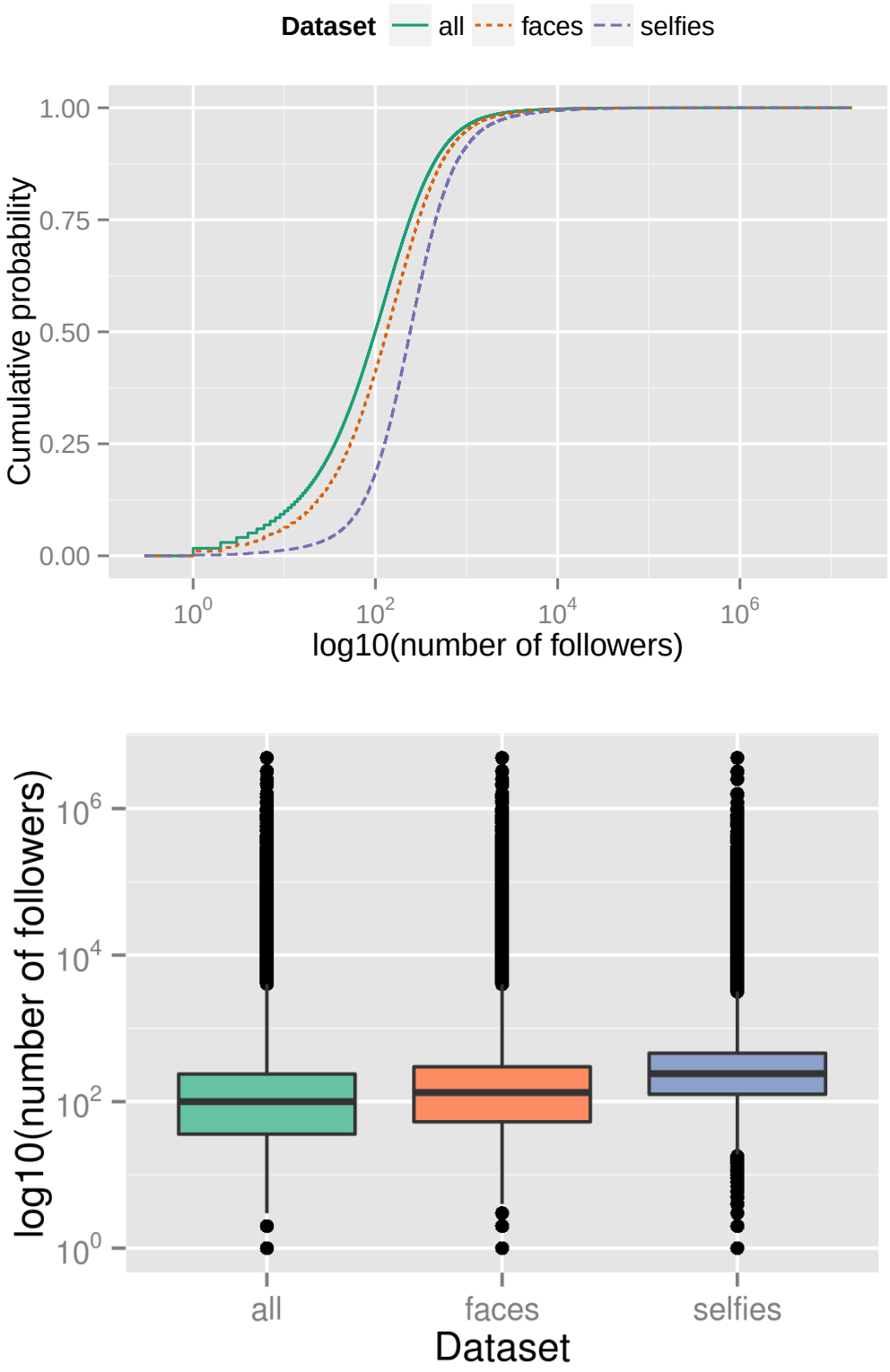


Figure 4.2. ECDF and boxplot for number of followers by user in each dataset.

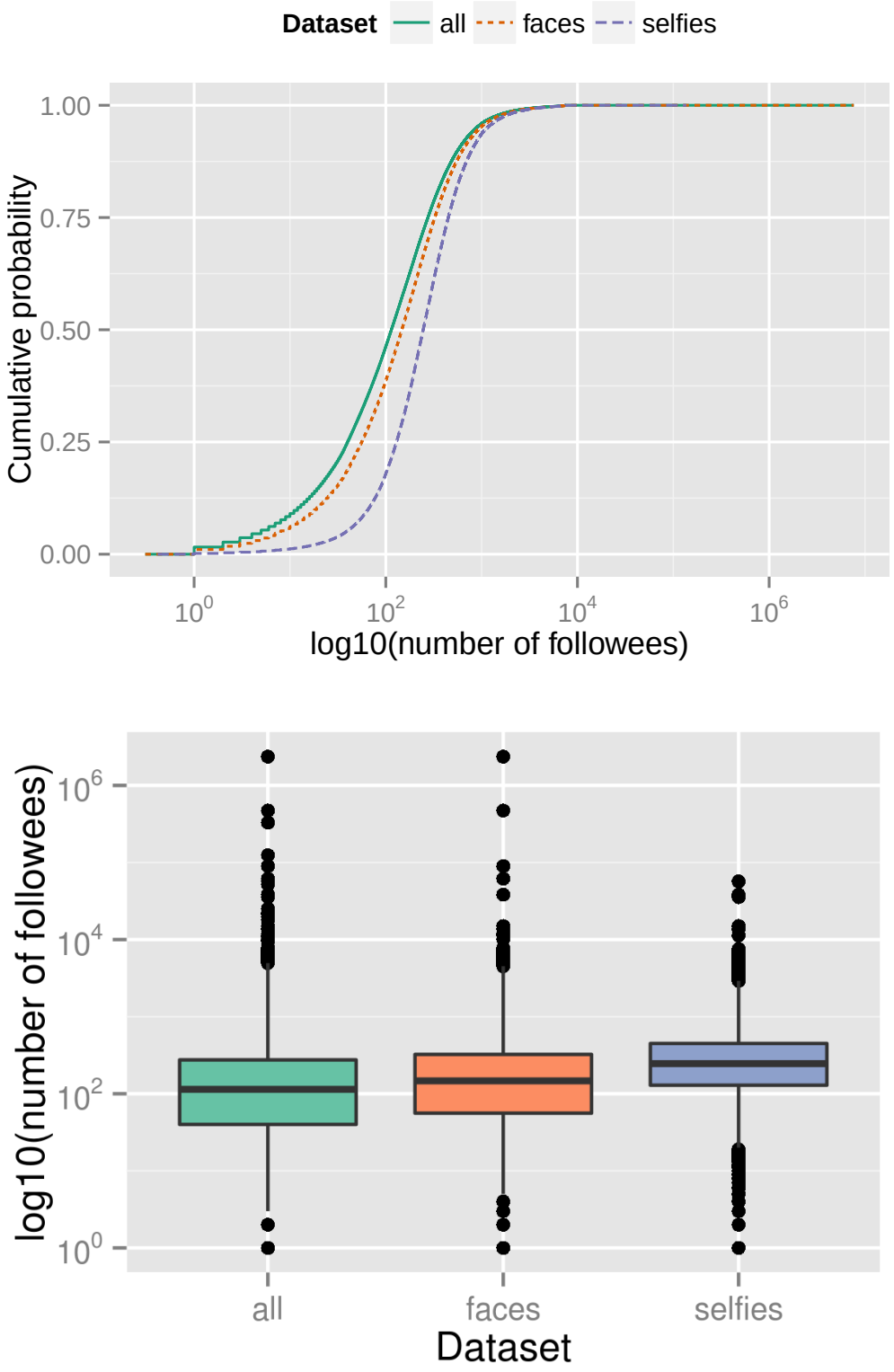


Figure 4.3. ECDF and boxplot for number of followees by user in each dataset.

Table 4.1. Proportions of users with full name, bio, and website fields filled. Differences between datasets are significant at $\alpha = 0.05$ ($p < 10^{-15}$).

	Full name	Bio	Website
all	84%	44%	10%
faces	87%	46%	11%
selfies	90%	74%	19%

in these fields. There is no remarkable difference between users in faces and in all, but users in selfies generally provide more information in their profiles than users in the other datasets.

The comparison of users characteristics in each dataset contributes to a view of selfie users as more active, more connected and more prone to share information than common Instagram users, including those who post general photos with faces.

4.2 Photos

Three categories of data are explored in the characterization of photos. The first one is the interactions category, composed of likes and comments. Figures 4.4 and 4.5 show ECDFs and boxplots for $\log_{10}$ number of likes/comments in each dataset, considering photos with at least 1 like/comment.

The pattern for the number of comments is almost the same in the three datasets. In the case of number of likes, though, the curves indicate that selfies tend to receive more attention: 51% of photos in all (and 46% in faces) have 10 likes or less, but only 26% do so in selfies.

Another important contrast is related to the proportion of pictures with no comments or likes in each dataset. 7.7% of photos in all have no likes. The proportion is similar for faces (6.6%), but much smaller for selfies: just 0.6%. As for comments, 62% of pictures in all and 60% of pictures in faces haven't received any comment, while this proportion is 44% in selfies.

All these statistics highlight that likes happen much more often on Instagram than comments – probably because they are a faster and lightweight form of interaction – but both kinds of interactions occur more frequently for selfies than for other types of contents.

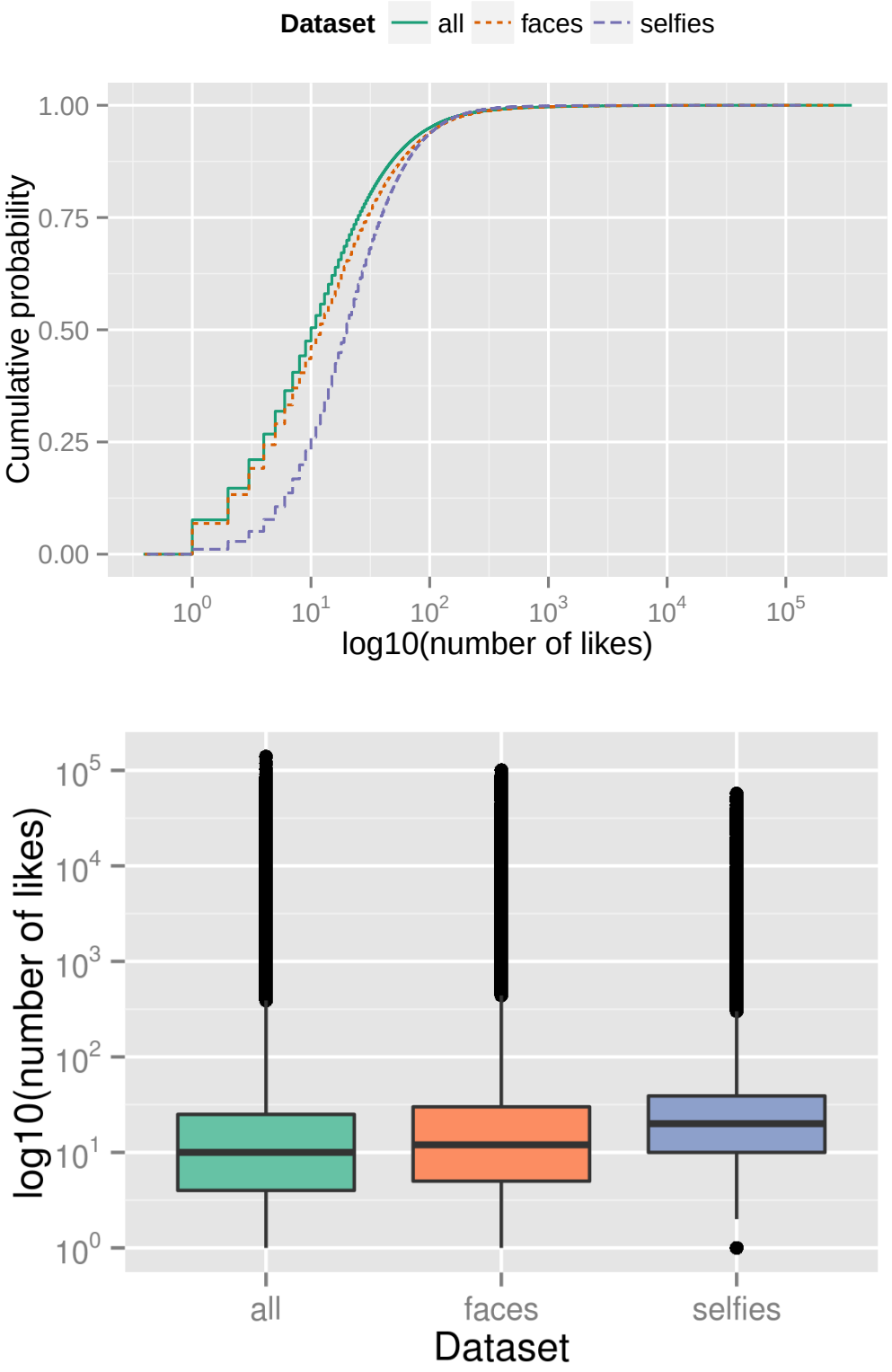


Figure 4.4. ECDF and boxplot for number of likes in each dataset.

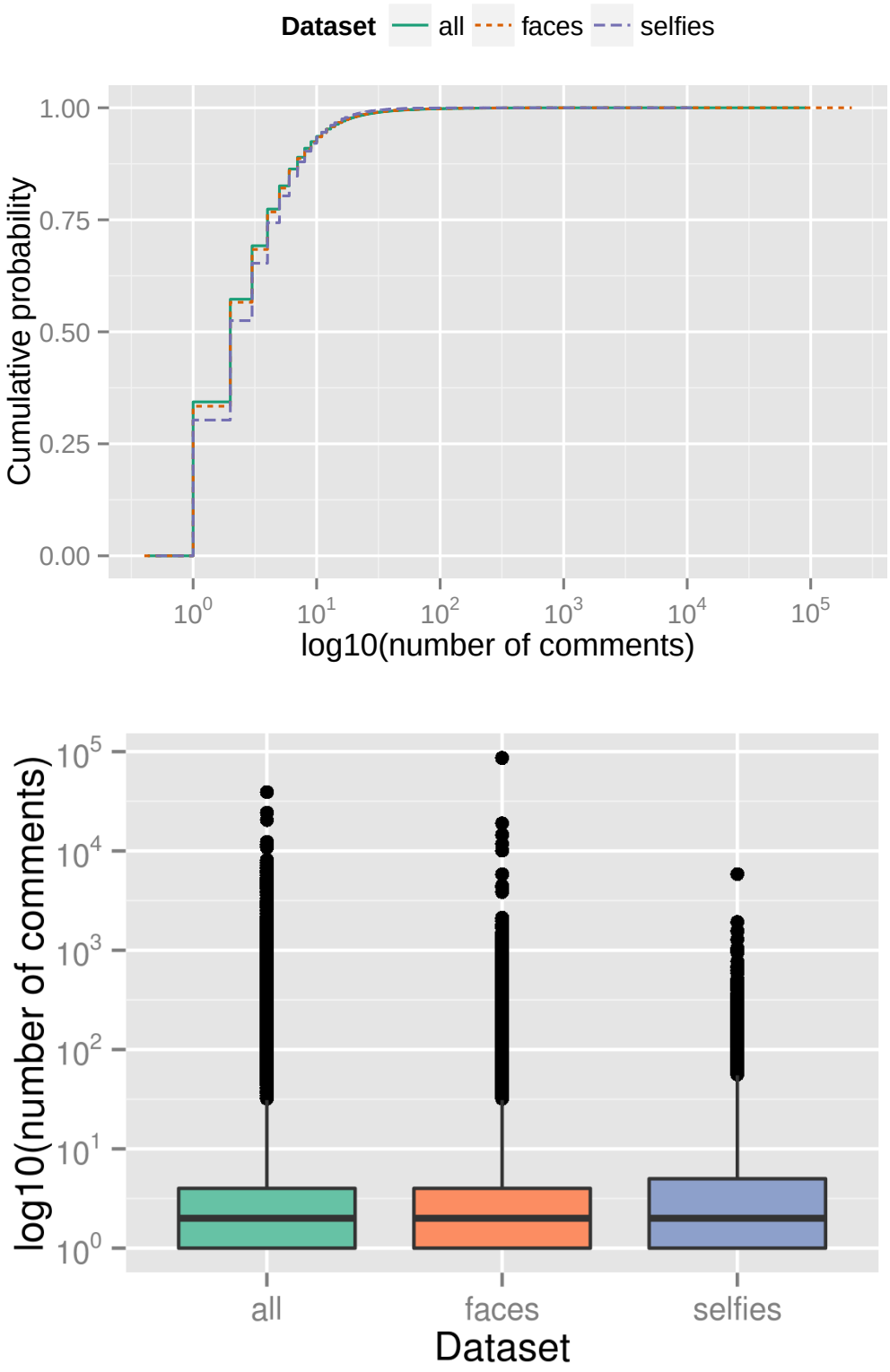


Figure 4.5. ECDF and boxplot for number comments in each dataset.

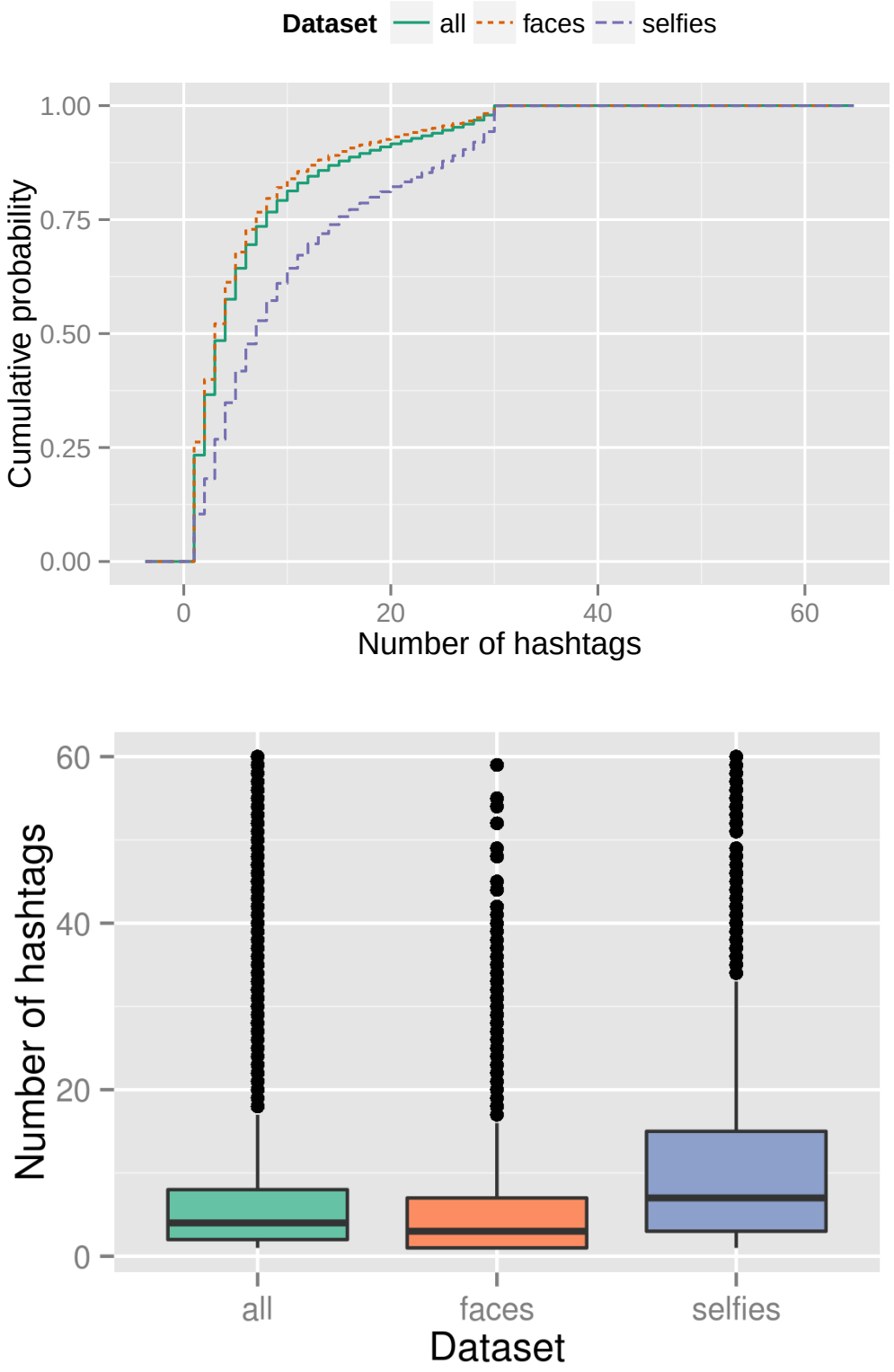


Figure 4.6. ECDF and boxplot for the number of hashtags in each dataset.

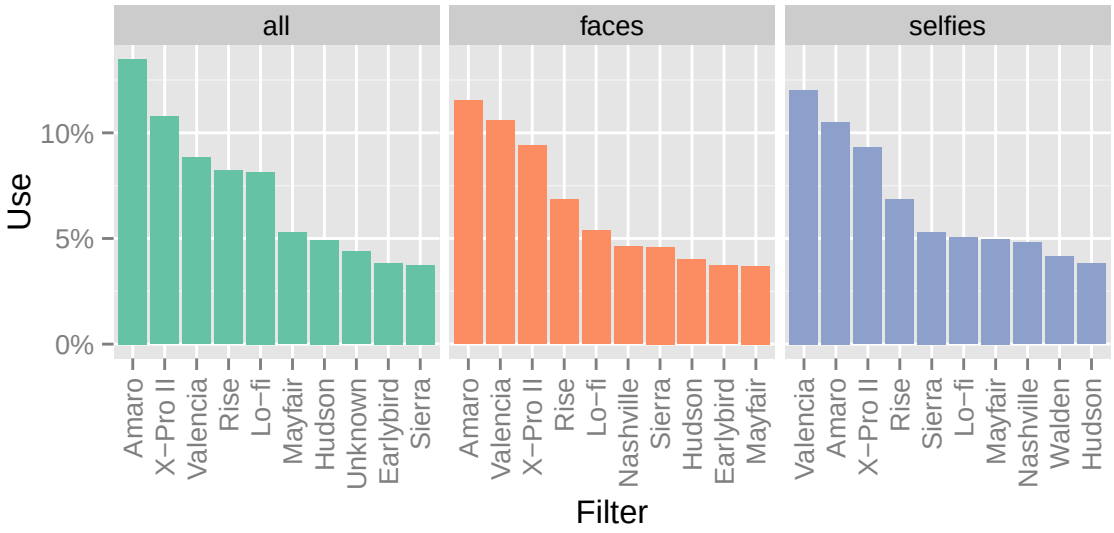


Figure 4.7. Top-10 most used filters in each dataset.

The second category of data studied comprises hashtags and filters. The ECDF and boxplot shown in Figure 4.6 for pictures with at least 1 hashtag reveals a similar pattern for all and faces, with more than 80% of the pictures containing up to 10 hashtags, but a different trend for selfies, where 64% of the photos have 10 hashtags or less. The selfies dataset was generated based on the use of selfie-related hashtags and this way all photos in this dataset have at least one hashtag. For all and faces it was found that the majority of the pictures actually does not contain any hashtag (60% in the former and 63% in the latter).

The three datasets do not seem to vary from each other with respect to filters, except in the order of the most used filters, as exhibited in Figure 4.7. Approximately 50-60% of all pictures in the datasets are not filtered. Among the filtered ones, the two most used filters appear in 10% to 15% of the photos, with other filters appearing in decreasing proportions bellow 10%.

The third category established for characterization is related to data provided by Face++. Because of this, only faces and selfies are compared. Considering the possible different definitions users attribute to selfies, the first fact investigated is the presence of faces in photos with selfie-related hashtags. Figure 4.8 displays a great proportion of photos without faces in selfies, and also the presence of photos with multiple faces. The proportion of pictures identified as "Unknown" represents the photos for which it was not possible to infer the number of faces due to some error in the Face++ API.

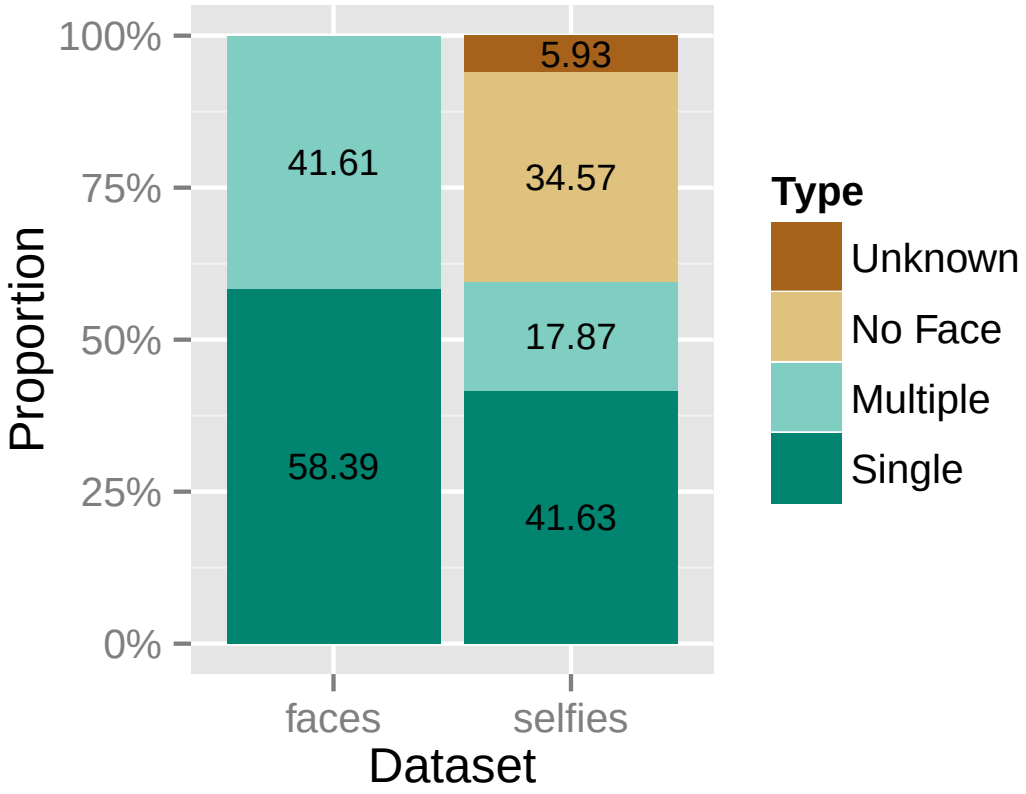


Figure 4.8. Proportion of pictures with a single face, multiple faces, no faces and unknown number of faces.

Considering just the subset of pictures with faces in selfies (∼60%), a comparison with faces dataset shows that selfies are still more connected to the concept of a single person self-portrait, as 69.97% of photos have a single face, in contrast with 58.39% in faces (difference significant at $\alpha = 0.05$, $p < 10^{-15}$). For multiple faces, however, selfies presents a similar pattern with respect to faces, which is observable in the ECDF and in the boxplot of Figure 4.9.

Among pictures metadata returned by Instagram API there is a field named "users_in_photo", which contains information about other Instagram users present in the picture. This field is filled only when the user who posts the picture explicitly choose to mark other people in the photo. There is no automatic recognition, so the user has to manually select points in the picture to mark. Only Instagram users can be marked, though, and it is not possible to identify people outside the network. To test the relation between this information and that provided by Face++, a plot of the number of users in photo returned by Instagram API versus the number of faces returned by Face++ API is displayed in Figure 4.10.

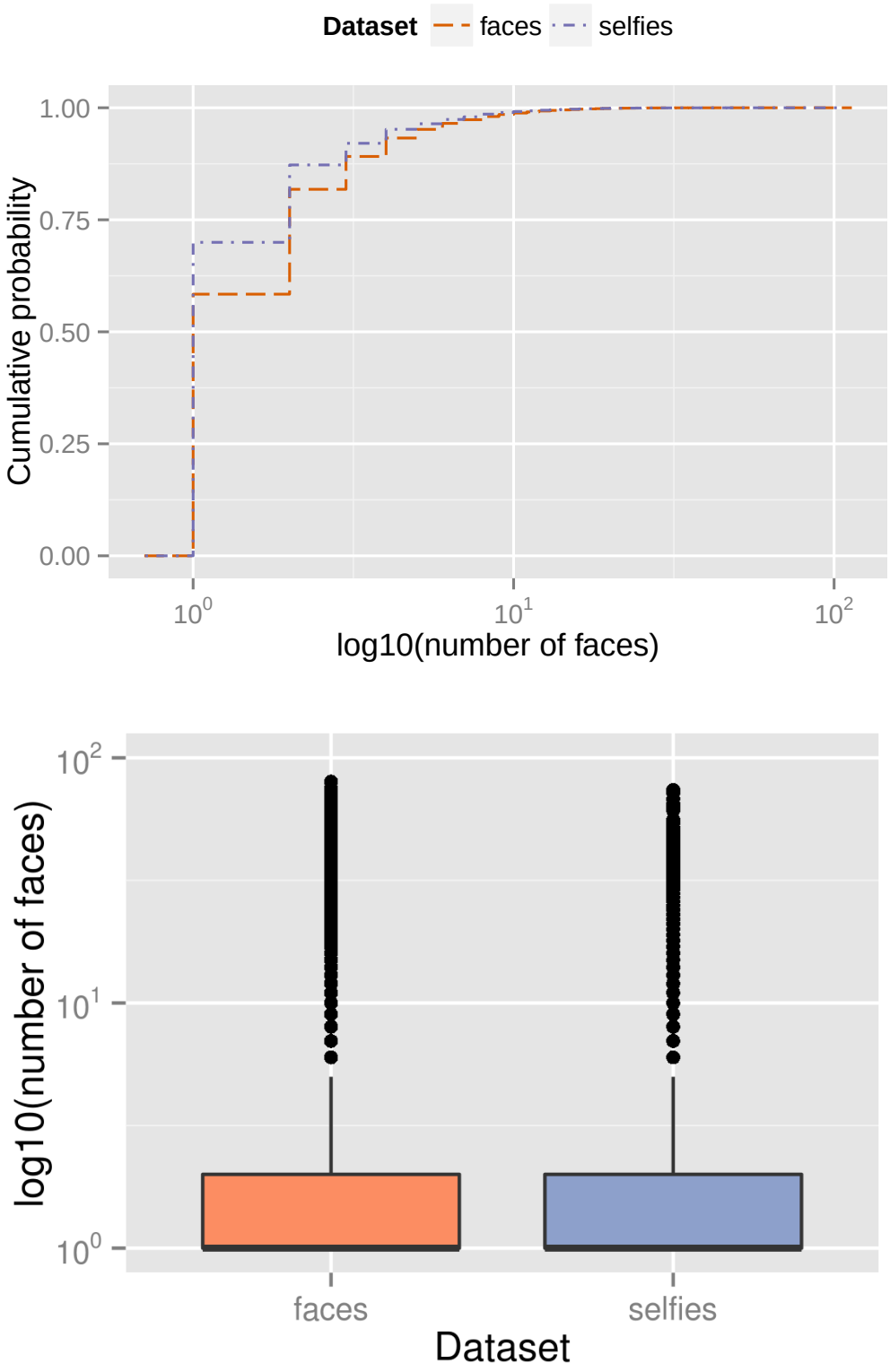


Figure 4.9. ECDF and boxplot for the number of faces in each dataset.

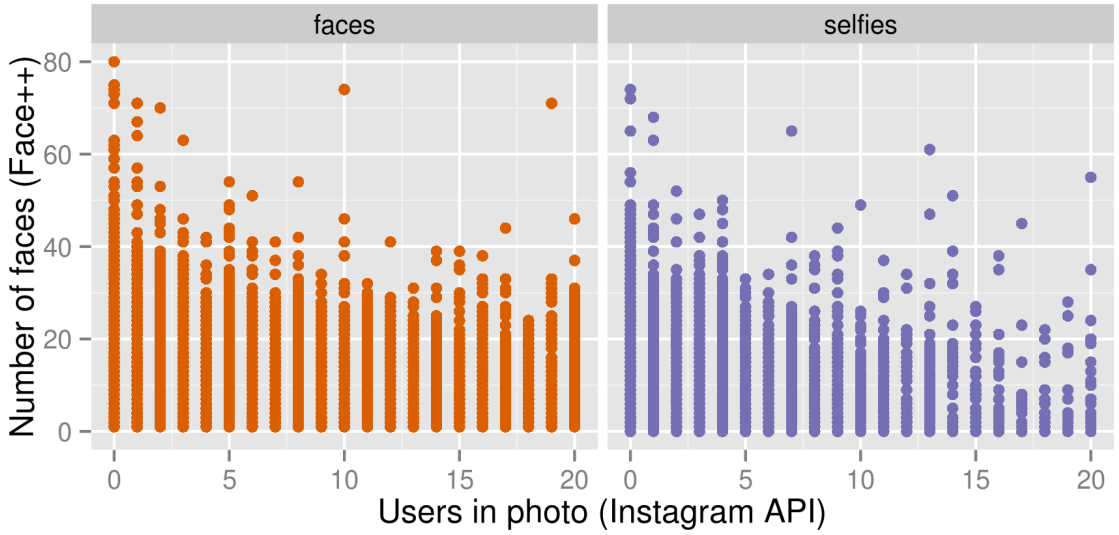


Figure 4.10. Relationship between the number of users in photo, as indicated in images metadata, and the number of faces detected by Face++.

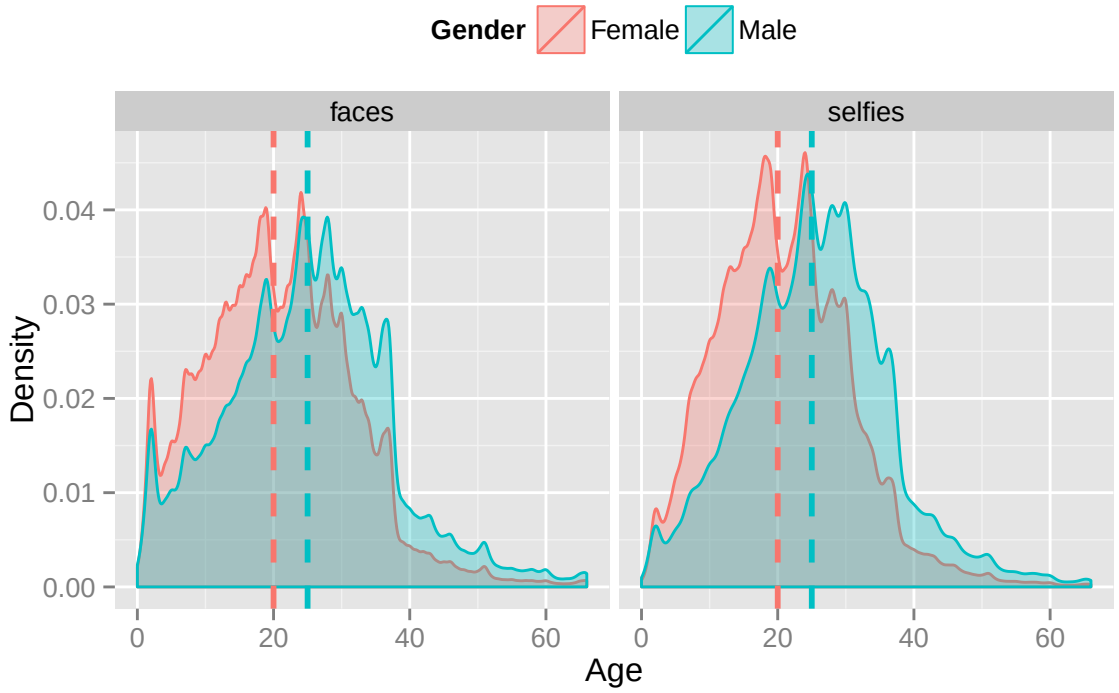


Figure 4.11. Density plot of ages per gender. Median values are indicated by dashed lines.

(a) Density plot of smiling values per gender. Median values are indicated by dashed lines.

(b) Average smiling values per age with 95% confidence intervals.

Figure 4.12. Smiling values per gender and age.

Neither faces nor selfies present a clear relationship between the two variables. Assuming Face++ is likely to give the right number of faces in a picture, due to its high precision face detection algorithm, this lack of relationship with the number of users in photo can be explained either by a low propensity of Instagram users in marking another users when publishing pictures or by the fact that many people who appear in Instagram photos are not current Instagram users and, consequently, cannot be marked. Unfortunately, it is not possible to distinguish these two scenarios without additional data.

Along with face detection, Face++ also provides age and gender information for each face. This information can be used to explore demographics of all people present in the photos of faces and selfies datasets.

In both datasets, the proportion of females is higher than that of males: 58% in faces and 65% in selfies. In Figure 4.11, age and gender are shown together, revealing also that people in the photos tend to be young, specially women. This goes in line with results of other researches showing a prevalence of young females on Instagram [Duggan et al., 2015; Tifentale and Manovich, 2015], and agrees even with information about selfies on another OSN as well [Qiu et al., 2015].

Finally, the smiling value obtained with Face++ can give a sense of possible sentiments/moods conveyed in photos with faces and selfies. Figure 4.12(a) shows smiling values per gender in faces and selfies. The figure indicates females tend to smile more than males, a pattern also found by Redi et al. [2015b] for portrait photographs. The plot of average smiling value per age, displayed in Figure 4.12(b), confirms the previous result and allows to conclude that the highest smiling values for both genders are associated with people between 30 and 40 years old in the two datasets.

Chapter 5

Temporal Analyses

Social networks are naturally very dynamic, so a static view of the network, although informative, possibly only has the power to represent the state of its entities in a short range of time. Fortunately, all media on Instagram have a timestamp associated with them, giving the day and time in UTC of when media objects have been posted. Besides, users metadata were collected on two points in time, apart six months from each other. This longitudinal data provides a unique opportunity to examine trends for users and photos from 2012 to 2015. Some measures are presented using the geometric mean, instead of the arithmetic mean , because of the highly skewed nature of the data.

5.1 Users

To asses how much do users information in each dataset change over time, a comparison is made in this section between some users counters provided by Instagram API. Number of media, number of followers, and number of followees are investigated considering their values in December 2014 and June 2015. The results are shown in Figure 5.1.

The variation in number of media published by each user indicates that selfie users worked more on their feeds (either publishing more photos or deleting existing ones) than general users and users who post photos with faces. The geometric mean of number of added or deleted pictures, displayed in the figure along with 95% confidence intervals, allows to observe that the amount of variation is greater for users in selfies than in the other datasets.

A similar pattern is found for number of followers and number of followees, except that the proportion of users who increased their number of followers is almost the same for users in faces and selfies, both being higher than the proportion found for users in

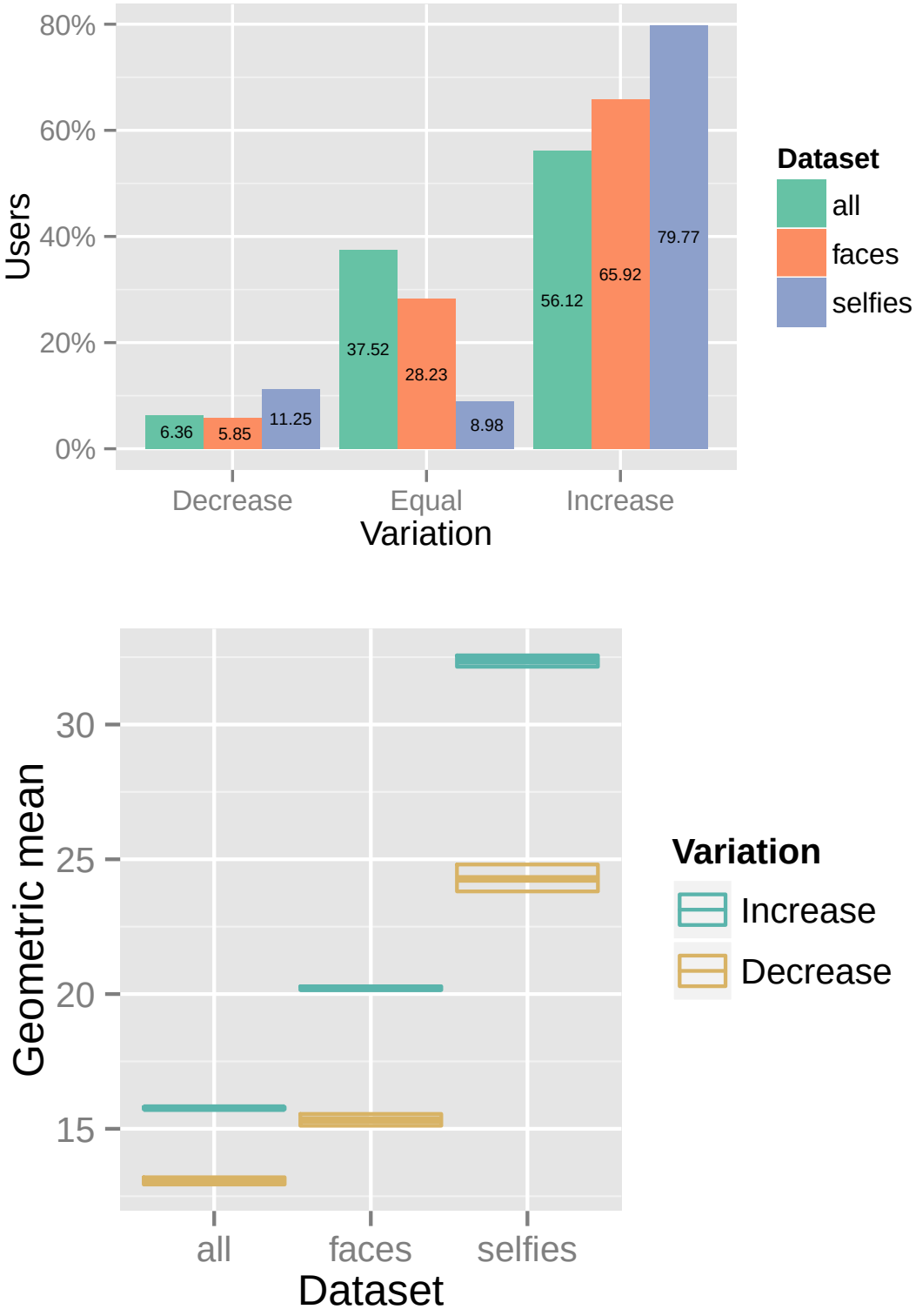


Figure 5.1. Variation in the number of media of users in each dataset.

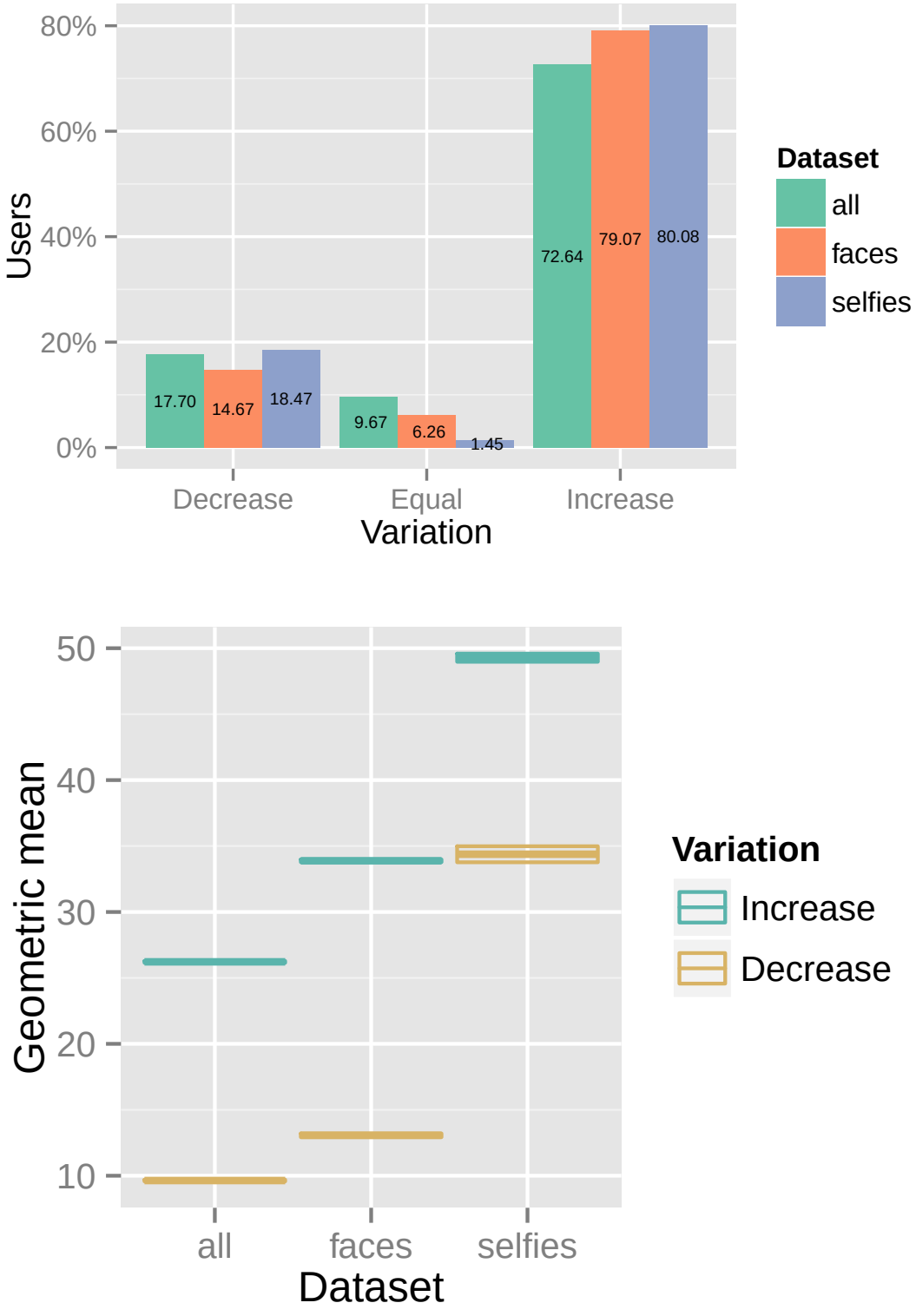


Figure 5.2. Variation in the number of followers of users in each dataset.

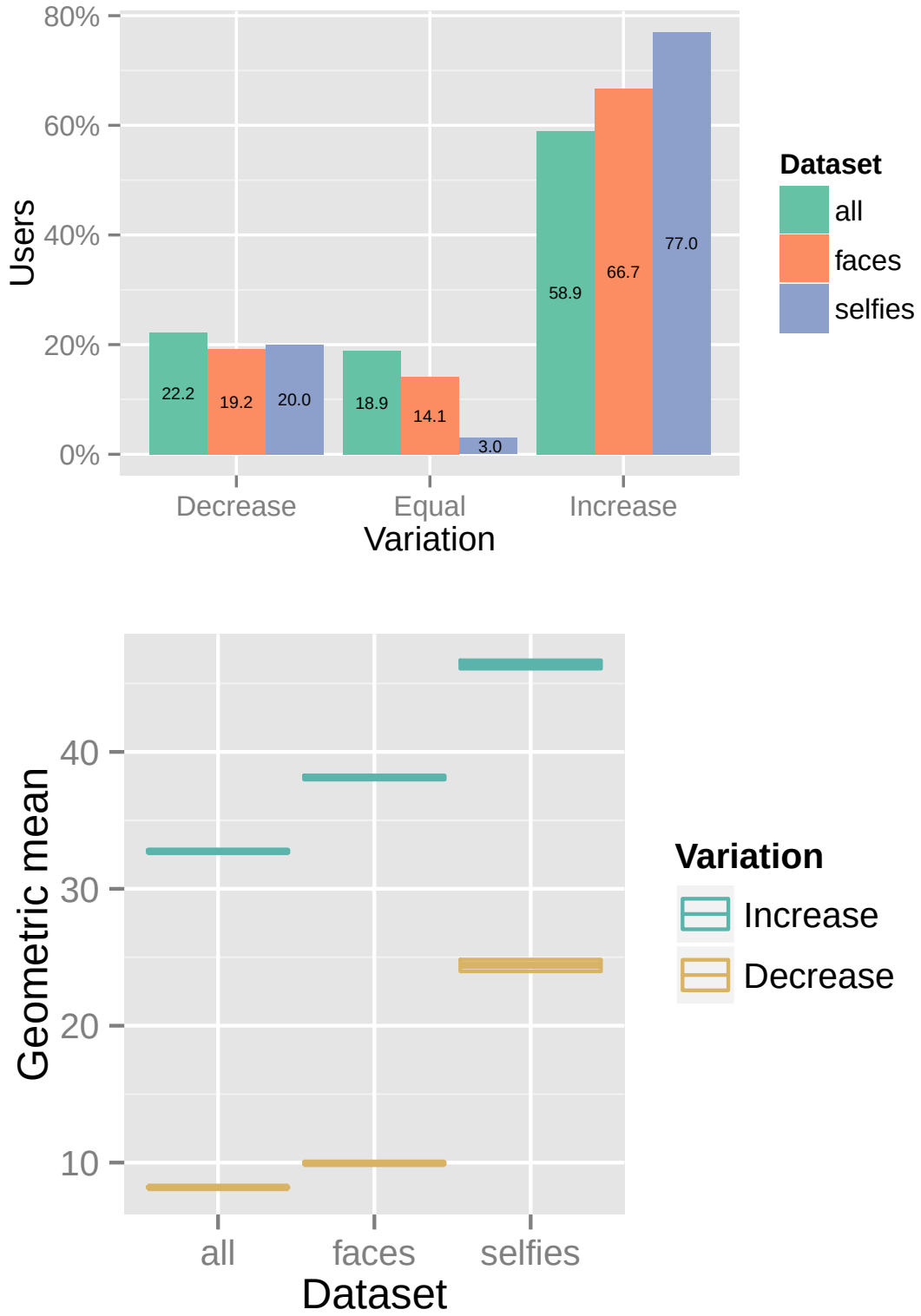


Figure 5.3. Variations in the number of followees of users in each dataset.

all. This contributes to a view of selfies and photos with faces as more engaging than the ordinary content published on the network, although the geometric mean of number of added followers favours selfies as even more engaging than general photos with faces.

5.2 Photos

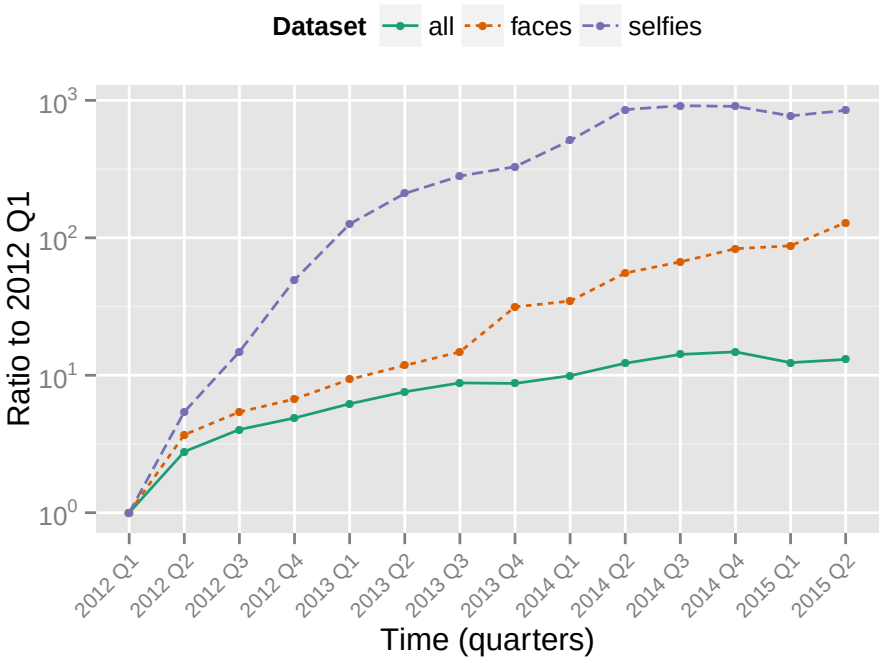
In this section a wide range of photo features are studied over time: number of posts and users, interactions (likes and comments), hashtags and filters, number of faces, demographics (gender and age), and smiling values. Number of faces, demographics and smiling values depend on Face++ data, thus the all dataset was excluded from these analyses. Data are grouped in quarters (i.e., three-month periods) in every case.

5.2.1 Number of Posts and Users

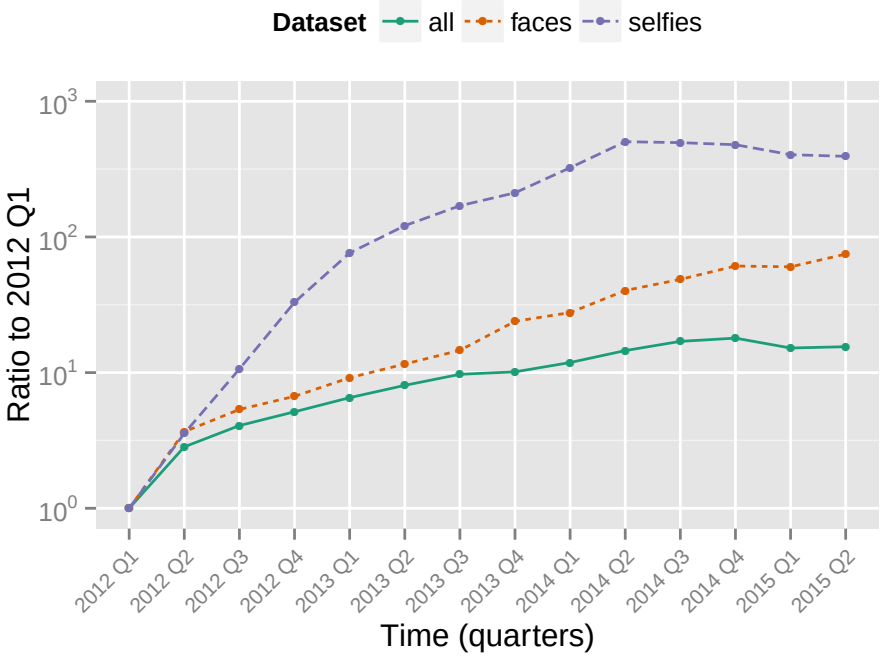
Figure 5.4 shows the variation in number of posts and number of users in each dataset over time, where the x-axis represents time in quarters and the y-axis represents the relative increment or decrement compared to the initial quarter (i.e., the first quarter of 2012). Therefore, a value of 1.0 in this figure means an amount identical to what was measured in the initial quarter and a value of 10.0 means an increment by 10 times.

While the number of posts in all increased 13 times (from 124,473 in the first quarter of 2012 to 1,622,694 in the second quarter of 2015), the number of posts in selfies increased rapidly by 846 times (from 245 to 207,204) over the same time period. faces also became popular compared to all, yet not at the same degree as selfies. When comparing the speed, all and faces show a relatively steady growth in the volume of publications, whereas the volume of selfies grows fast at first and then becomes stagnant by the middle of 2014.

A similar trend is seen in the graph of number of users in each dataset. selfies once more shows orders of magnitude larger growth than the other types of contents. Opposed to faces, though, selfies present a slight decrease in the number of users from the second quarter of 2014 onwards. These trends capture well the rapid rise of selfies on Instagram, which seems to have peaked in the middle of 2014.



(a) Posts



(b) Users

Figure 5.4. Evolution of number of posts and number of users in each dataset relative to the first quarter of 2012.

5.2.2 Interactions

The amount of attention a photo gained can be inferred examining the number of likes and comments. Figure 5.5(a) shows the geometric means of number of likes per picture, which demonstrates that selfies receive nearly 2–3 times more likes than other types of contents. This means pictures with an explicit hashtag marking then as selfies grab more attention from audience than other photos that merely contains faces. Examining closely, however, the relative gap between selfie and all decreases over time from nearly 2.4 times during the thriving initial spread to 1.7 times in the last quarter.

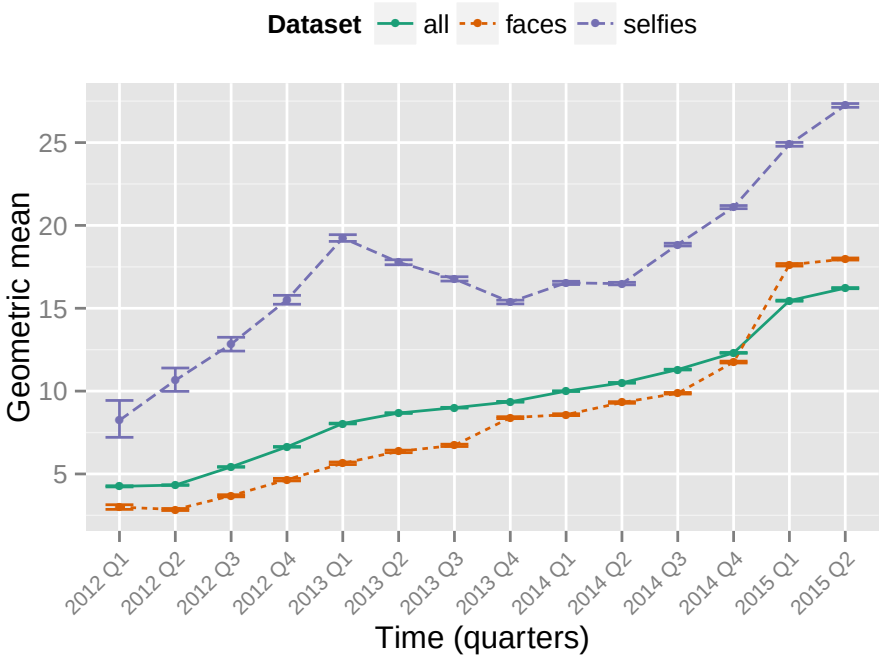
The geometric means of number of comments received in Figure 5.5(b) captures an analogous scenario. Again selfies receive 1.1–1.4 times more comments than all and faces, although this gap decreases over time. This observation indicates that pictures owning a selfie-related hashtag are effective in grabbing attention, yet their engaging effect becomes less pronounced over the years (perhaps as selfies become more mundane).

When compared to the recent literature on the engagement effect of faces in pictures, Bakhshi et al. [2014] demonstrated pictures with faces tend to get 38% more likes and 32% more comments compared to other types of contents on Instagram. While a direct comparison cannot be made, the results in this section further highlight that attributing particular hashtags (such as `#selfie`) to photos on Instagram could incur an even higher level of popularity than merely posting photos with faces.

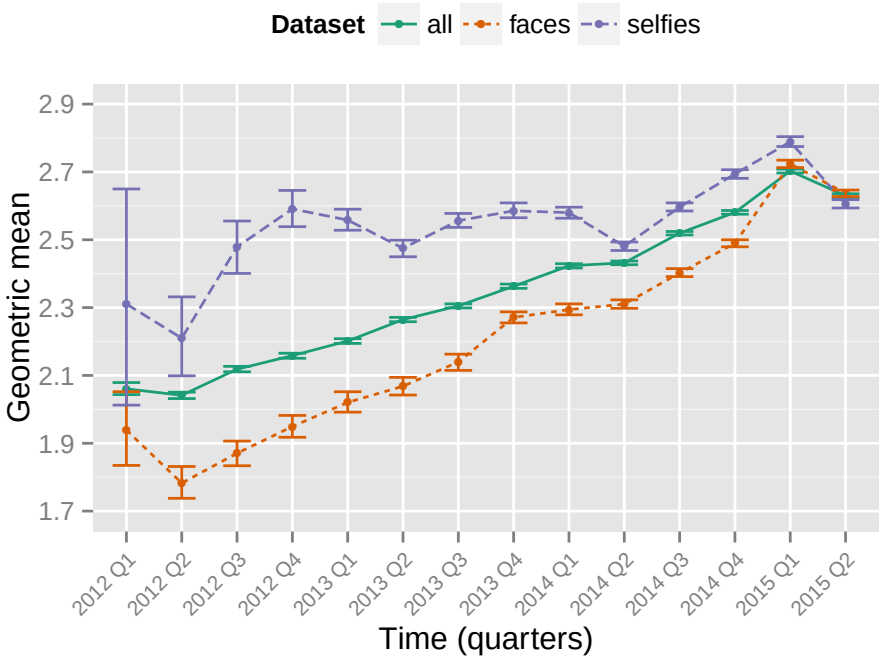
5.2.3 Hashtags and Filters

Hashtags and filters present two different scenarios. In the case of hashtags, the geometric means of number of hashtags, displayed in Figure 5.6(a) reveals a difference between selfies and the other types of contents of approximately 1.6 times from the first quarter of 2012 to the second quarter of 2014. After that, the difference increases, reaching 2.0 between selfies and all in the middle of 2015. Considering, however, that selfies dataset is made up of photos filtered by the use of hashtags it is not possible to know without further investigation if the observed difference in the number of hashtags is precisely attached to selfies (i.e., selfies tend to receive more hashtags) or to the behavior of users who attribute hashtags to their publications (i.e., users who employ hashtags tend to use not only one when they do so, but a higher amount).

In the case of filters, although the proportion of filters usage, shown in Figure 5.6(b), is generally higher over time for selfies than faces and for faces than all, the three datasets present a decreasing trend since the beginning of 2012 until the end of 2014, with a slight increase in first semester of 2015.

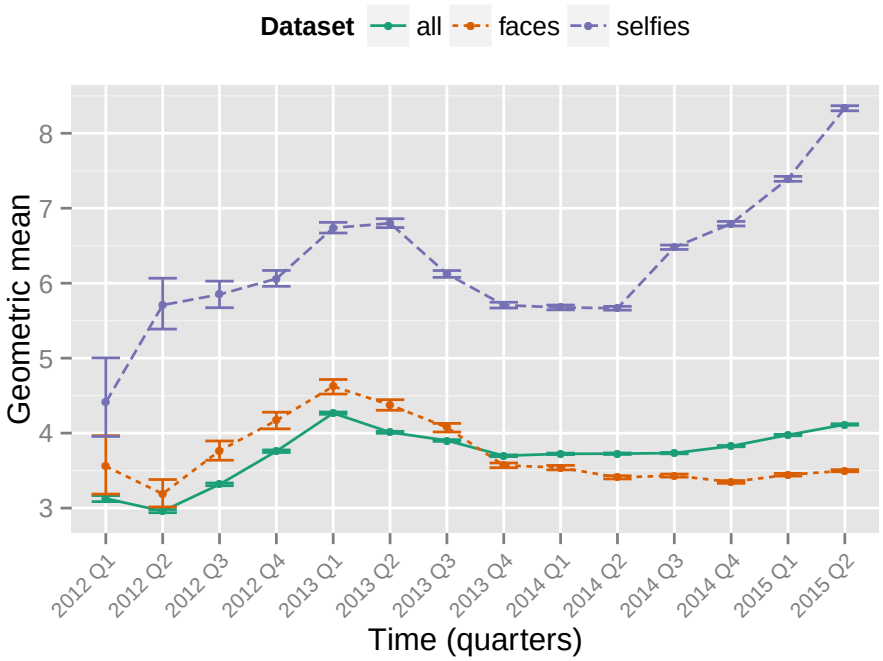


(a) Likes



(b) Comments

Figure 5.5. Evolution of geometric mean number of likes and comments in each dataset.



(a) Hashtags

(b) Filters

Figure 5.6. Evolution of geometric mean number of hashtags and proportion of filtered photos in each dataset.

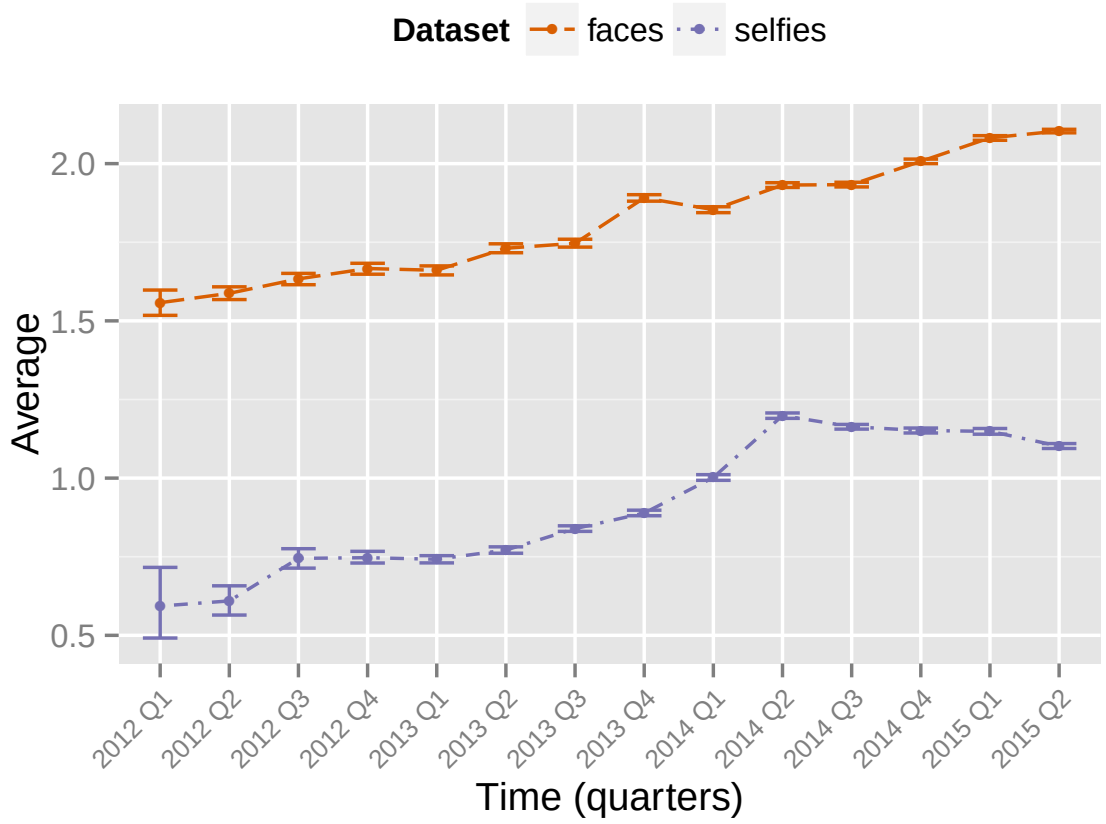


Figure 5.7. Evolution of average number of faces in each dataset.

5.2.4 Number of Faces

How have selfies changed over time with respect to the number of faces? Have the concept of a selfie as a single person self-portrait varied much? To investigate these questions, Figure 5.7 shows the average number of faces per picture for selfies and faces. At a glance, faces contains a higher number of faces (1.5–2.25 faces per picture). In 2012 and 2013, the average number of faces in selfies remained bellow 1.0, which indicates a great proportion of selfies without faces. It then reached 1.0 in the beginning of 2014 and stayed above this mark after that. This variation in the number of faces contained in selfies implies that Instagram users have not been tied to the definition of selfies as single person self-portraits, and rather have used the hashtag for other possible definitions as well.

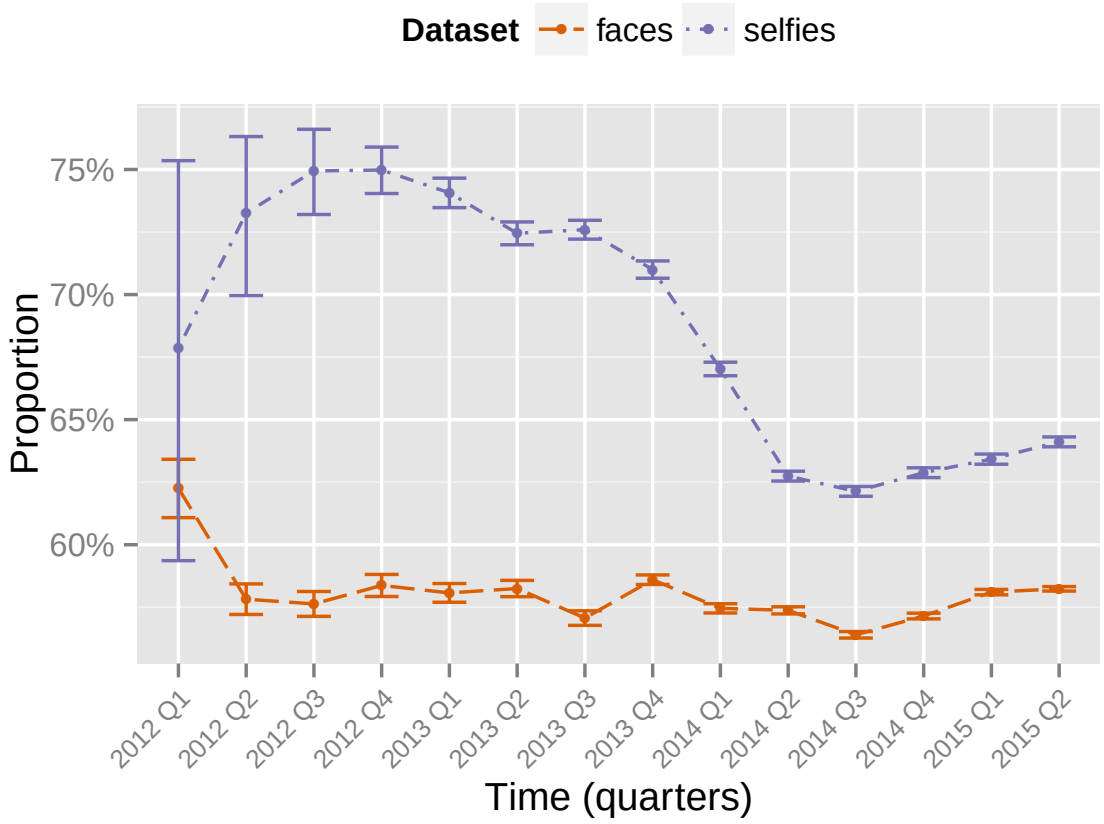
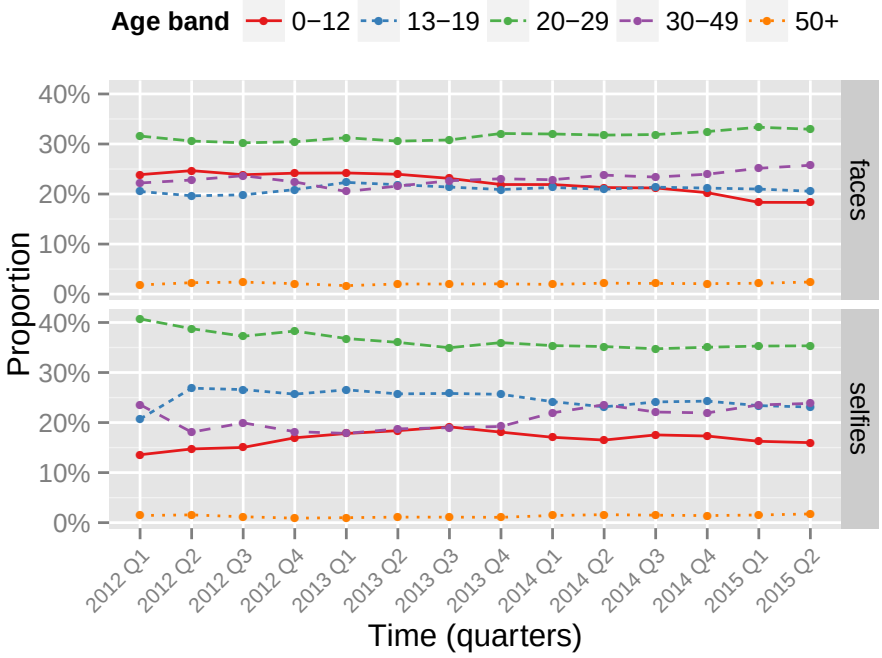


Figure 5.8. Evolution of proportion of female faces in each dataset.

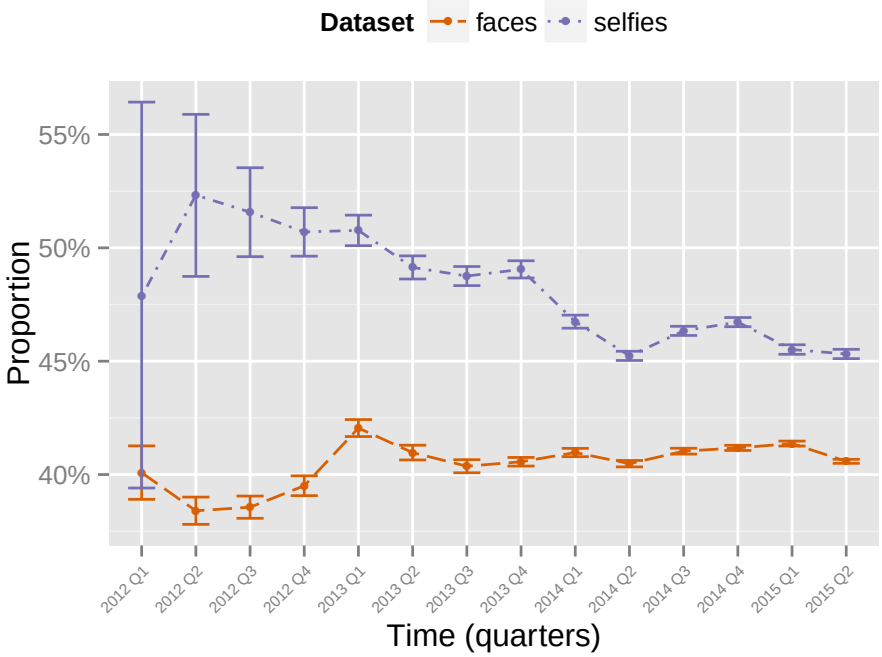
5.2.5 Demographics

To explore variations in the number of people of each gender appearing in selfies and photos with faces, the proportion of females by quarter is plotted in Figure 5.8. faces does not show much change over time, and the proportion of females in this dataset, between 55–65% coincides with the known female prevalence in the network. In selfies, on the other hand, it is possible to see a strong female bias from the second quarter of 2012 to the first quarter of 2014, during selfies thriving period. From the second quarter of 2014 onwards, the proportion dropped, but stayed significantly higher than the proportion found in faces.

Figure 5.9 shows the demographic makeup for different age bands. People between 20 and 29 years old are the most prevalent in both selfies and faces. On the other hand, the proportion of people above 50 years old is almost zero. Taking all bands into account, it is possible to conclude that the majority of people who appear in selfies and photos with faces is young.



(a) All ages



(b) 13–25 year olds vs. others

Figure 5.9. Evolution of proportion of age bands in each dataset.

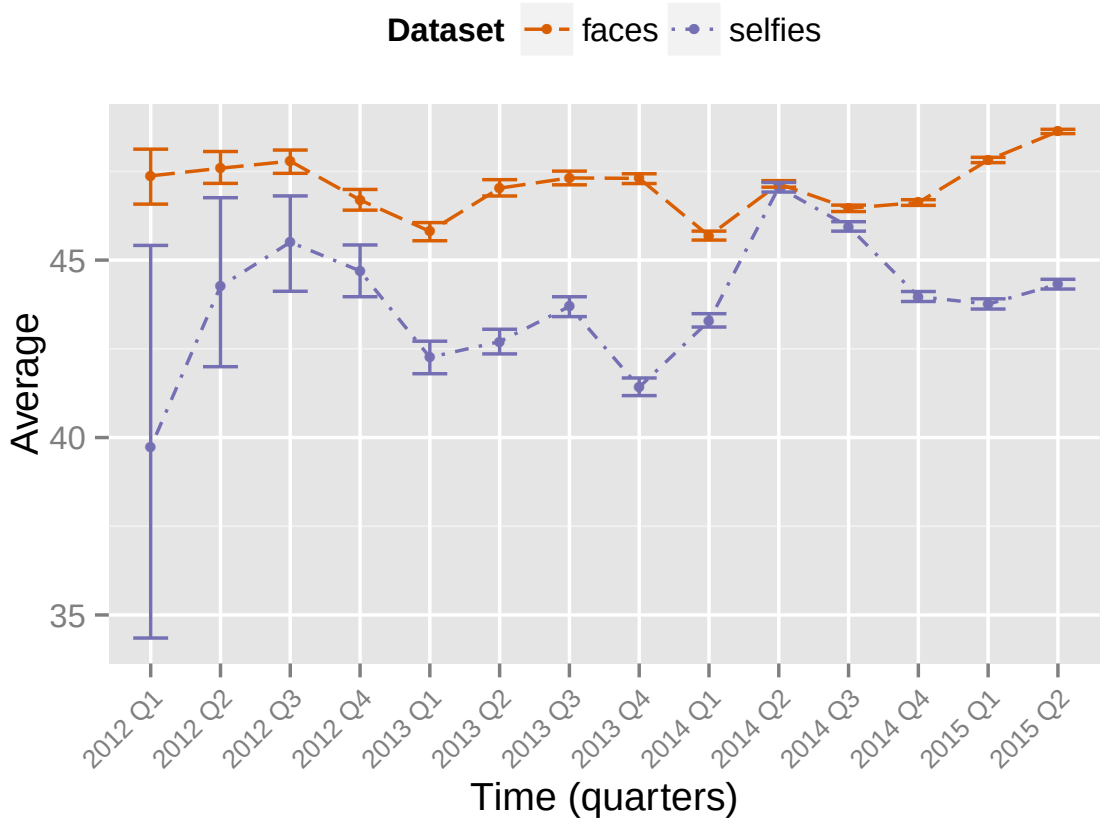


Figure 5.10. Evolution of average smiling values in each dataset.

It is curious to see a reasonable amount of people between 0 and 12 years old, considering Instagram only allows people above 13 years old to sign up and create and account. Admitting no one of these people up to 12 years are actually Instagram users, this result can be related to parents and other adults who share information about children on online social networks, an action that can bring serious concerns about privacy breaches [Minkus et al., 2015].

Another interesting fact arises when the age bands of 13–19 and 20–29 are considered together. When added, the proportions of people in these bands are higher in selfies than faces, leading to a hyphotesis that selfies are linked to even younger people than would be expected for common photos with faces. To investigate this finding, Figure 5.9(b) compares the proportion of people between 13 and 25 years old with people of other ages in the two datasets. The plot indeed confirms a significant difference between selfies and faces with respect to young people in the photos, specially in 2012 and 2013. Gender and age analyses together indicate that young females drove the selfie momentum on Instagram during its initial stage.

5.2.6 Smiling Values

Would pictures tagged as selfies present a more joyful atmosphere? Comparing the average smiling values in Figure 5.10, people in selfies do not show more signs of joy than the ones in faces. Overall, the averages in both datasets do not surpass the value of 50, which represents something like a half smile or a closed-lip smile. Despite some small differences between selfies and faces, it was not possible to find any particular relation neither between the smiling values and the type of data nor between smiling values and time.

Chapter 6

Spatial Analyses

Instagram mobile application has an option for users to specify the location where the photo or video has been taken. These locations are stored as latitude and longitude coordinates. Figure 6.1 displays a world map with photos in all mapped by their coordinates. Each green dot represents a picture. The map clearly shows a high concentration of posts in North America and Europe. There is also a great amount of publications in East Asia, South Africa, and parts of South America and Oceania.

Using GADM [Global Administrative Areas, 2012] to convert coordinates to country names, it was possible to build a rank with the proportion of pictures by country in each dataset. After filtering out countries with least than 30 pictures to avoid distor-

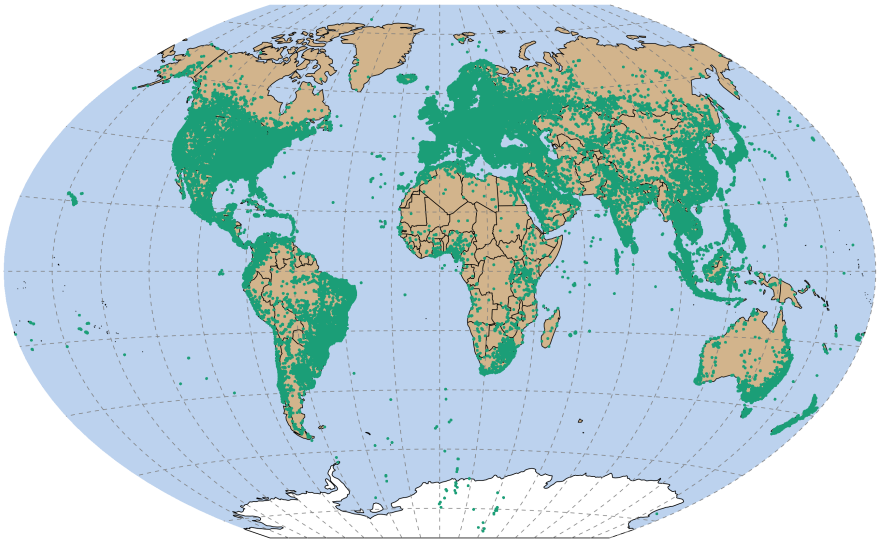


Figure 6.1. Distribution of Instagram pictures (all dataset) around the world.

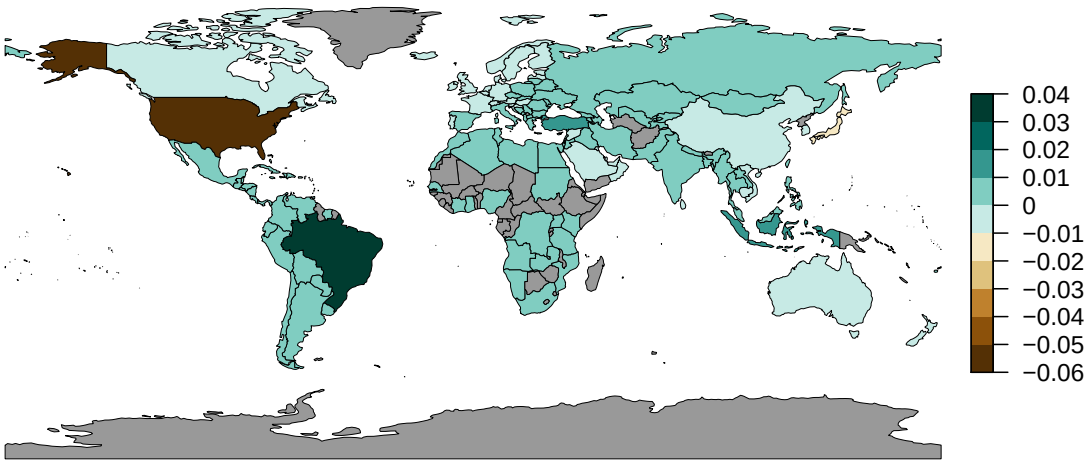
Table 6.1. Top-10 prevalent countries. faces and selfies lists are relative to all.

all	faces	selfies
USA (30.23%)	BRA (2, +3.78%)	ITA (6, +2.53%)
BRA (8.50%)	IDN (8, +1.64%)	GBR (5, +2.02%)
THA (5.99%)	TUR (11, +1.14%)	PHL (13, +1.42%)
RUS (5.66%)	RUS (4, +0.77%)	CAN (10, +1.29%)
GBR (4.06%)	COL (28, +0.47%)	POL (39, +1.20%)
ITA (3.11%)	MEX (14, +0.46%)	KOR (22, +1.18%)
AUS (2.30%)	THA (3, +0.39%)	TUR (11, +0.92%)
IDN (2.25%)	ARG (31, +0.38%)	IND (36, +0.91%)
JPN (2.25%)	PHL (13, +0.33%)	AUS (7, +0.72%)
CAN (2.19%)	IND (36, +0.29%)	MEX (14, +0.64%)
188 countries	144 countries	127 countries

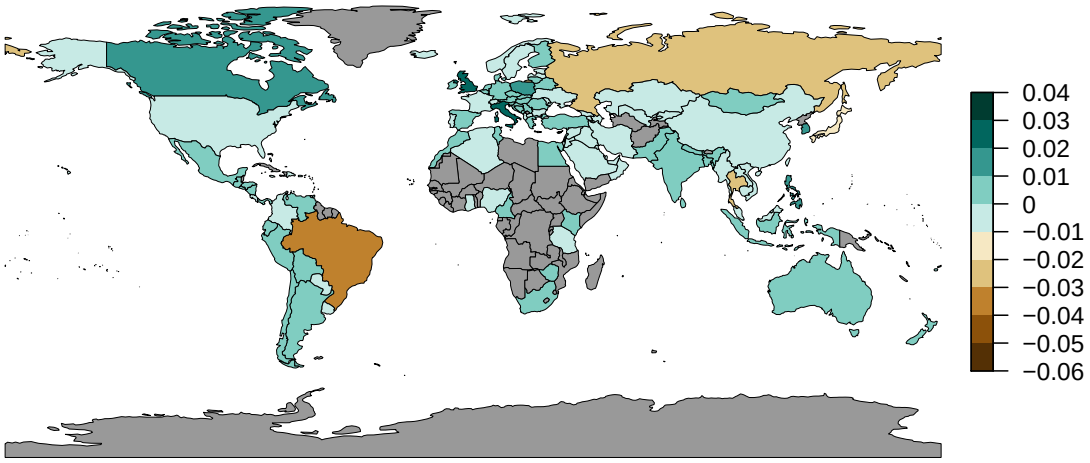
tions, the subtraction of the proportions found in faces and selfies from the ones found in all allowed to discover the countries with the highest relative proportions of selfies and photos with faces, respectively, compared to what would be expected based on the proportions of all Instagram pictures.

The result of this procedure is shown in Figure 6.2. Table 6.1 presents the top-10 most prevalent countries in each dataset, while the bottom-10 least prevalent ones are shown in Table 6.2. For all, the number in parenthesis indicates the proportion of photos in the country. For faces and selfies, the numbers in parenthesis indicate, respectively, the position of the country in the list of all and the relative increase or decrease in the proportion. The total number of different countries found in each dataset is given at the bottom of each table.

Despite the widespread of Instagram posts around the globe, it draws attention the fact that 30.23% of all geotagged photos are concentrated in the United States. In comparison, Brazil, which is the second in the list, holds 8.5% of the pictures, and Canada, the tenth, holds just 2.19%. Besides, the United States figures as the last country in faces list, meaning that a broader range of subjects other than faces can be found in pictures posted in that country. Nonetheless, the proportion of selfies is very similar to that of all, which could indicate that many photos marked as selfies in the United States do not contain faces.



(a) Faces



(b) Selfies

Figure 6.2. Distribution of faces and selfies relative to all. Countries with no data available are colored in grey.

Table 6.2. Bottom-10 prevalent countries. faces and selfies lists are relative to all. Countries in all have a prevalence smaller than 0.0001%.

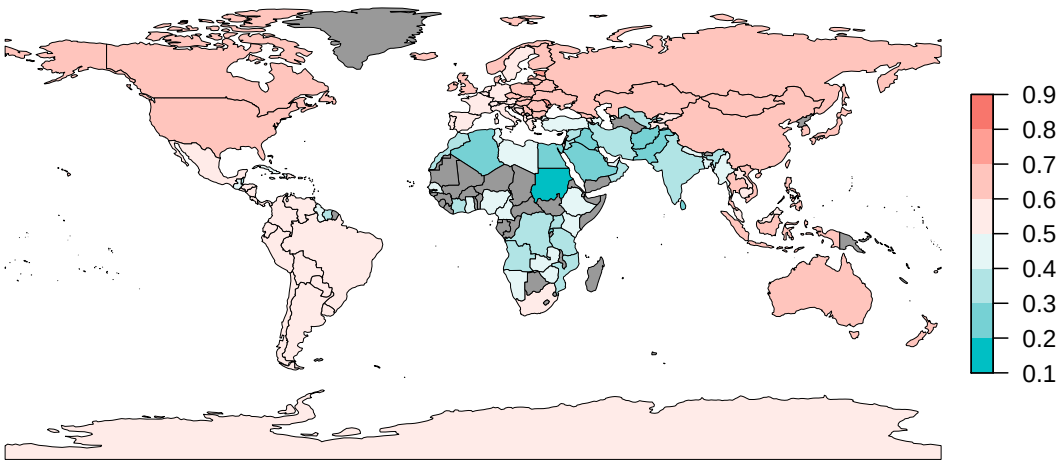
all	faces	selfies
VCT	FRA (15, -0.26%)	KWT (33, -0.38%)
ATA	KWT (33, -0.29%)	ISR (25, -0.43%)
GRL	SAU (19, -0.43%)	TWN (21, -0.62%)
COG	CAN (10, -0.53%)	SAU (19, -0.89%)
MWI	AUS (7, -0.64%)	SWE (16, -0.93%)
SLE	SWE (16, -0.64%)	CHN (17, -0.97%)
FRO	CHN (17, -0.71%)	JPN (9, -1.95%)
GAB	GBR (5, -0.73%)	THA (3, -2.04%)
SJM	JPN (9, -1.10%)	RUS (4, -2.77%)
GIN	USA (1, -5.43%)	BRA (2, -3.29%)
188 countries	144 countries	127 countries

The top-10 lists of faces and selfies share only 4 countries in common (Turkey, Mexico, Philippines, and India), but none of them is in the top-10 list of all. The bottom-10 lists of the two datasets share 5 countries in common (Kuwait, Saudi Arabia, Sweden, China, and Japan), for which selfies and photos with faces occur less frequently than other types of photos.

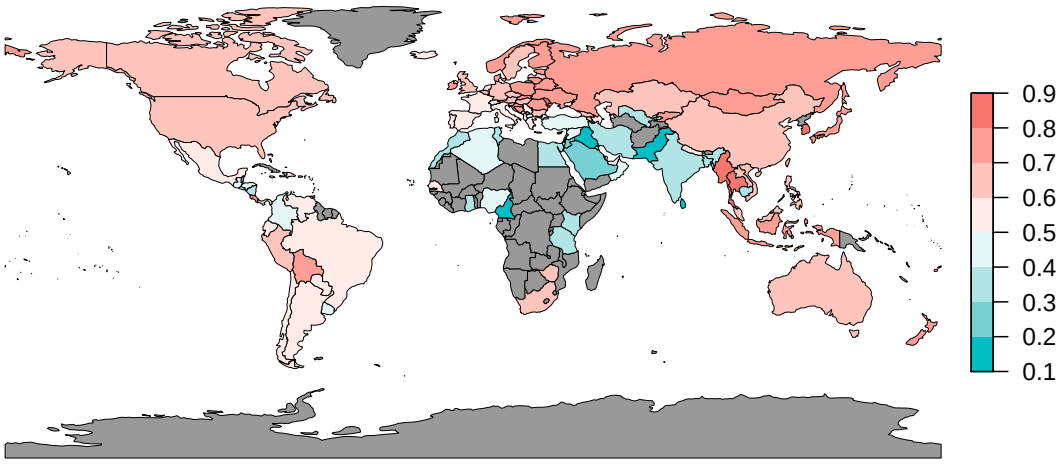
Finally, 6 countries present a peculiar situation: Brazil, Russia, Thailand, United Kingdom, Canada, and Austria. The first three appear in the top-10 list of faces, but are ranked last in the list of selfies. This points to a lower use of selfie-related hashtags in spite of the higher number of photos with faces posted in these countries. United Kingdom, Canada, and Austria, on the contrary, appear in the top-10 list of selfies, but also in the bottom-10 list of faces. As in the case of the United States, this could mean a greater adoption of different concepts of selfies that not necessarily include faces.

6.1 Gender Prevalence

Dividing the number of females faces in each country by the total number of faces, the gender prevalence per country can be calculated for faces and selfies, as displayed in Figure 6.3. The majority of the countries lie in the female range of the scale, which is not surprising given the results in previous chapters. What is interesting is the concentration of male prevalent countries in Africa, Middle East, and South Asia, an outcome possibly linked to demographic and cultural factors.

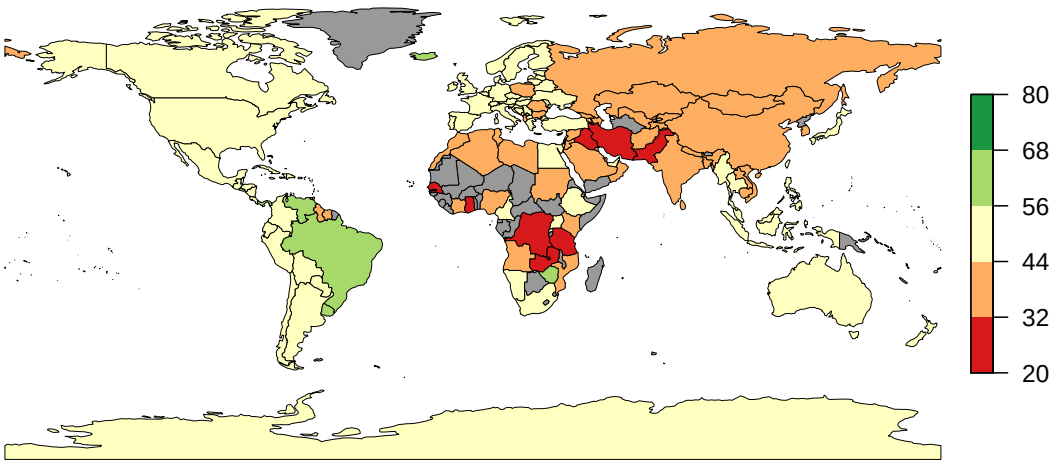


(a) Faces

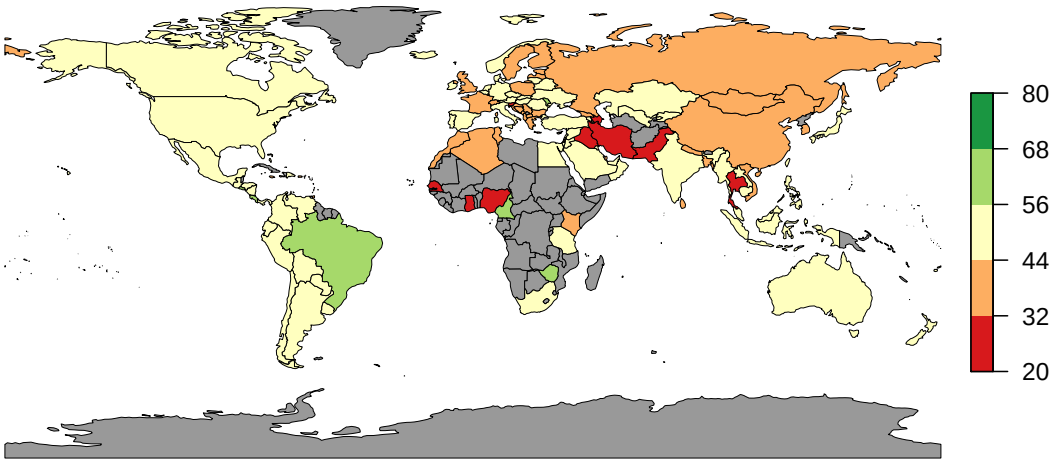


(b) Selfies

Figure 6.3. Gender ratio per country. Countries with no data available are colored in grey.



(a) Faces



(b) Selfies

Figure 6.4. Average smiling value per country. Countries with no data available are colored in grey.

6.2 Smiling Tendency

Using Face++ data to obtain average smiling values at country level, the maps depicted in Figure 6.4 allow to see smiling tendencies around the globe with respect to selfies and photos with faces. Most countries show a neutral to low tendency, with many lower tendency countries concentrated in Asia. Africa shows a mixed panorama, specially in the case of selfies dataset, with some countries in the lower side of the smiling scale and others in the upper side. Brazil presents a consistent pattern, figuring at the upper side of the smiling scale in both datasets.

6.3 Implications

All spatial results in this chapter emphasize the complexity of the selfie phenomenon. Thus, to obtain a deeper level of comprehension about it, the local aspects should be studied as much as the global ones, in order to understand what are the more universal characteristics of selfies and what characteristics are more strongly tied to differences between where and by who they were taken.

Chapter 7

Conclusion

Selfies are ever more present in today's online culture. Nonetheless, many aspects of selfies had not yet been studied under the perspective of data analysis. This thesis tried to fill this gap presenting a measurement study of selfies based on a large amount of data gathered from Instagram, comparing selfies, photos with faces, and general pictures posted on the network.

Many aspects of the publications were explored – such as interactions (likes and comments), hashtags and filters, and faces information (number of faces, gender, age, and smiling values) – in three different ways: characterization, temporal analyses, and spatial analyses. The main findings from each of these analyses can be summarized as follows:

- Selfie users can be seen as more active, more connected and more prone to share information than common Instagram users, including those who post general photos with faces;

- In the period analyzed, selfie users worked more on their feeds (either publishing more photos or deleting existing ones) than general users and users who post photos with faces;

- People in selfies and photos with faces tend to be young, especially women. In fact, young females drove the selfie momentum on Instagram during its initial stage;

- Females tend to smile more in the pictures than males. The highest smiling values for both genders are associated with people between 30 and 40 years old;

- The number of selfies on Instagram increased rapidly from 2012 to 2015, with a growth in the volume of publications more than 65 times greater than the growth of ordinary content published on the network;

- Pictures owning a selfie-related hashtag are effective in grabbing attention, yet their engaging effect are becoming less pronounced over the years;

- Instagram users have not been tied along time to the definition of selfies as single person self-portraits, and rather have used the hashtag for other possible definitions as well;

- There are significant differences between countries with respect to the prevalence of selfies and photos with faces, and with respect to gender and smiling tendencies of people appearing in these types of pictures.

7.1 Contributions

In a recent interview, Instragram founder Kevin Systrom said that "the selfie is something that didn't really exist in the same way before Instagram" [Kubina, 2015]. In quantitatively capturing those ways, this thesis contributes to a better comprehension of the selfie phenomenon, complementing all the psychological and sociological framing already built around them.

This thesis also offers a methodology for collecting Instagram data and extracting faces and selfies information. In a practical point of view, it presents CAMPS Data Collection Tool, a data collection program initially designed to face the challenges found in this work which ended as a robust crawler for use in OSNs researches, easily adaptable to many different scenarios.

The large dataset collected with the proposed methodology (and using the data collection tool) allowed the execution of a complementary work that gave rise to a paper accepted at the ACM Conference on Online Social Networks 2015 (COSN'15). The paper expands the study on selfies addressing some topics beyond the ones explored in this thesis (see Apendix A for details). All the knowledge coming from these efforts contributes significantly to offer important insights to designers of social-networking platforms, online service providers, and mobile devices manufactures on how to improve they work in order to meet the demands of the new selfie generation.

7.2 Limitations

Hashtags can have a dual role of bookmarking content and serving as the symbol of a community membership [Yang et al., 2012]. This work employs selfie-related hashtags to extract selfies from Instagram, differentiating them from other types of contents and, at the same time, identifying users who post selfies. However, considering selfies have different meanings to social media users, it is possible that selfies are published on the network without receiving a selfie-related hashtag. Thus, such selfies (and the users who post them) could not be captured for analyses with the methodology used.

Demographic analyses are tied to information obtained with Face++. Despite the widely employed conceptualization of a selfie as a single person self-portrait, the studies in this work demonstrate that many selfies do not include a face, and there are countries where a higher prevalence of photos with faces is not associated with a higher usage of selfie-related hashtags. Consequently, age and gender analyses for selfies are limited only to selfies that contain faces, restricting somewhat the generalization power of these results.

7.3 Future Directions

Given its multifaceted character, selfies still present much more dimensions not yet investigated in this thesis. The combination of time and space in a spatio-temporal analysis, for instance, could provide a new view on how the selfie phenomenon developed.

A user-oriented approach could allow to study how selfie-related behaviors vary from user to user or even across different classes of users (e.g., celebrities and occasional users), also bringing information on how users characteristics affect the engagement associated with the selfies they publish.

Finally, the combination of different sentiment analysis methods [Gonçalves et al., 2013] based on comments, captions, and hashtags related to selfies could make it possible to explore the attachment between selfies and emotions.

Bibliography

Araújo, C. S., Corrêa, L. P. D., da Silva, A. P. C., Prates, R. O., and Meira Jr., W. (2014). It is not just a picture: Revealing some user practices in instagram. In *Latin American Web Congress (LA-WEB)*, pages 19–23.

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*, CHI '14, pages 965--974, New York, NY, USA. ACM.

Bakhshi, S., Shamma, D. A., Kennedy, L., and Gilbert, E. (2015). Why We Filter Our Photos and How It Impacts Engagement. In *International AAAI Conference on Web and Social Media*.

BBC (2013). Self-portraits and social media: The rise of the 'selfie'. Retrieved July 20, 2015, from `http://www.bbc.com/news/magazine-22511650`.

Chandra, S., Qiu, L., Roy, S., Lin, W., and Jakhetiya, V. (2015). Do others perceive you as you want them to? modeling personality based on selfies. In *ACM Multimedia Conference (Workshop on Affect and Sentiment in Multimedia)*.

Cole, N. L. (2015a). The Selfie Debates, Part I. Retrieved July 20, 2015, from `http://sociology.about.com/od/Ask-a-Sociologist/fl/The-Selfie-Debates-Part-I.htm`.

Cole, N. L. (2015b). The Selfie Debates, Part II. Retrieved July 20, 2015, from `http://sociology.about.com/od/Current-Events-in-Sociological-Context/fl/The-Selfie-Debates-Part-II.htm`.

Cole, N. L. (2015c). Why we selfie. Retrieved July 20, 2015, from `http://sociology.about.com/od/Ask-a-Sociologist/fl/Why-We-Selfie.htm`.

Crandall, D. J., Backstrom, L., Huttenlocher, D., and Kleinberg, J. (2009). Mapping the world's photos. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 761--770, New York, NY, USA. ACM.

Crook, J. (2014). Know thy selfie. Retrieved July 20, 2015, from `http://techcrunch.com/2014/02/24/know-thy-selfie`.

Day, E. (2013). How selfies became a global phenomenon. Retrieved July 20, 2015, from `http://www.theguardian.com/technology/2013/jul/14/how-selfies-became-a-global-phenomenon`.

Dey, R., Nangia, M., Ross, K. W., and Liu, Y. (2014). Estimating heights from photo collections: a data-driven approach. In *Proceedings of the ACM Conference on Online Social Networks*, COSN '14, pages 227--238, New York, NY, USA. ACM.

Duggan, M., Ellison, N. B., Lampe, C., Lenhart, A., and Madden, M. (2015). Social media update 2014. Technical report, Pew Research Center.

eBay Deals Blog (2013). Digital vanity - sizing up the selfie revolution. Retrieved July 20, 2015, from `http://deals.ebay.com/blog/the-selfie-revolution/`.

Fox, J. and Rooney, M. C. (2015). The Dark Triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites. *Personality and Individual Differences*, 76:161--165.

Franco, J. (2013). The meanings of the selfie. Retrieved July 20, 2015, from `http://www.nytimes.com/2013/12/29/arts/the-meanings-of-the-selfie.html?_r=1`.

Gervais, S. (2013). Does instagram promote positive body image? Retrieved July 20, 2015, from `https://www.psychologytoday.com/blog/power-and-prejudice/201301/does-instagram-promote-positive-body-image`.

Global Administrative Areas (2012). GADM database of Global Administrative Areas, version 2.0. `http://www.gadm.org`.

Gonçalves, P., Araújo, M., Benevenuto, F., and Cha, M. (2013). Comparing and combining sentiment analysis methods. In *Proceedings of the First ACM Conference on Online Social Networks*, COSN '13, pages 27--38, New York, NY, USA. ACM.

Hochman, N. and Manovich, L. (2013). Zooming into an instagram city: Reading the local through social media. *First Monday*, 18(7). ISSN 13960466.

Hochman, N. and Schwartz, R. (2012). Visualizing instagram: Tracing cultural visual rhythms. In *International AAAI Conference on Web and Social Media*.

Hosseinmardi, H., Mattson, S. A., Rafiq, R. I., Han, R., Lv, Q., and Mishra, S. (2015). Detection of cyberbullying incidents on the instagram social network. *CoRR*, abs/1503.03909.

Hu, Y., Manikonda, L., and Kambhampati, S. (2014). What we instagram: A first analysis of instagram photo content and user types. In *International AAAI Conference on Web and Social Media*.

Huang, C.-M. and Park, D. (2013). Cultural influences on Facebook photographs. *International Journal of Psychology*, 48(3):334--343. ISSN 0020-7594.

Instagram Press (2016). Press Page - Instagram. Retrieved March 12, 2016, from `http://www.instagram.com/press`.

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*, CHI '15, pages 4039--4042, New York, NY, USA. ACM.

Joo, J., Li, W., Steen, F., and Zhu, S.-C. (2014). Visual persuasion: Inferring communicative intents of images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 216–223.

Joshi, D., Chen, F., and Wilcox, L. (2014). Finding selfies of users in microblogged photos. In *Proceedings of the First International Workshop on Social Media Retrieval and Analysis*, SoMeRA '14, pages 33--34, New York, NY, USA. ACM.

Kilner, J. (2014). The science behind why we take selfies. Retrieved July 20, 2015, from `http://www.bbc.com/news/blogs-magazine-monitor-25763704`.

Kim, A. and Gweon, G. (2014). Photo sharing of the subject, by the owner, for the viewer: Examining the subject's preference. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 975--978, New York, NY, USA. ACM.

Kubina, L. (2015). Selfiesticks, Simplicity and Bourbon - Interview with Instagram Founder Kevin Systrom. Retrieved October 06, 2015, from `http://goo.gl/cpqhhE`.

Losh, E. (2014). Beyond biometrics: Feminist media theory looks at selfiecity. Technical report, Selfiecity.

Magno, G., Comarela, G., Saez-Trumper, D., Cha, M., and Almeida, V. (2012). New kid on the block: Exploring the google+ social graph. In *Proceedings of the 2012 ACM Conference on Internet Measurement Conference*, IMC '12, pages 159--170, New York, NY, USA. ACM.

Manikonda, L., Hu, Y., and Kambhampati, S. (2014). Analyzing user activities, demographics, social network structure and user-generated content on instagram. *CoRR*, abs/1410.8099.

Megvii Inc. (2013). Face++ research toolkit. `http://www.faceplusplus.com`.

Meier, E. P. and Gray, J. (2014). Facebook photo activity associated with body image disturbance in adolescent girls. *Cyberpsychology, Behavior, and Social Networking*, 17(4):199--206.

Miller, A. D. and Edwards, W. K. (2007). Give and take: A study of consumer photo-sharing culture and practice. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 347--356, New York, NY, USA. ACM.

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*, WWW '15, pages 776--786, Republic and Canton of Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Ottoni, R., Pesce, J. a. P., Las Casas, D., Franciscani Jr., G., Meira Jr., W., Kumaraguru, P., and Almeida, V. (2013). Ladies first: Analyzing gender roles and behaviors in pinterest. In *International AAAI Conference on Web and Social Media*.

OxfordWords blog (2013). The oxford dictionaries word of the year 2013 is... Retrieved July 20, 2015, from `http://blog.oxforddictionaries.com/2013/11/word-of-the-year-2013-winner`.

Qiu, L., Lu, J., Yang, S., Qu, W., and Zhu, T. (2015). What does your selfie say about you? *Computers in Human Behavior*, 52:443--449. ISSN 07475632.

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. *CoRR*, abs/1505.07522.

Redi, M., Rasiwasia, N., Aggarwal, G., and Jaimes, A. (2015b). The beauty of capturing faces: Rating the quality of digital portraits. *CoRR*, abs/1501.07304.

Risen, J. and Poitras, L. (2014). N.S.A. collecting millions of faces from Web images. Retrieved July 20, 2015, from `http://www.nytimes.com/2014/06/01/us/nsa-collecting-millions-of-faces-from-web-images.html?_r=1`.

Schifanella, R., Redi, M., and Aiello, L. M. (2015). An Image Is Worth More than a Thousand Favorites: Surfacing the Hidden Beauty of Flickr Pictures. In *International AAAI Conference on Web and Social Media*.

Silva, T. H., Melo, P. O. S. V. d., Almeida, J. M., Salles, J., and Loureiro, A. A. F. (2013). A picture of instagram is worth more than a thousand words: Workload characterization and application. In *Proceedings of the 2013 IEEE International Conference on Distributed Computing in Sensor Systems*, DCOSS '13, pages 123--132, Washington, DC, USA. IEEE Computer Society.

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 Conference on Online Social Networks*, COSN '15, pages 221--231, New York, NY, USA. ACM.

Tifentale, A. (2014). The selfie: Making sense of the "masturbation of self-image" and the "virtual mini-me". Technical report, Selfiecity.

Tifentale, A. and Manovich, L. (2015). Selfiecity: Exploring photography and self-fashioning in social media. In Berry, D. M. and Dieter, M., editors, *Postdigital Aesthetics: Art, Computation and Design*, pages 109--122. Palgrave Macmillan.

Tiidenberg, K. (2014). Bringing sexy back: Reclaiming the body aesthetic via self-shooting. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(1):article 3.

Totti, L. C., Costa, F. A., Avila, S., Valle, E., Meira Jr., W., and Almeida, V. (2014). The impact of visual attributes on online image diffusion. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 42--51, New York, NY, USA. ACM.

Wilson, C. (2014). The selfiest cities in the world: TIME's definitive ranking. Retrieved July 20, 2015, from `http://time.com/selfies-cities-world-rankings/`.

Winneberger, D. (2014). 2013 AAFPRS membership study. Technical report, AAFPRS.

Wortham, J. (2013). My selfie, myself. Retrieved July 20, 2015, from `http://www.nytimes.com/2013/10/20/sunday-review/my-selfie-myself.html?_r=0`.

Yang, L., Sun, T., Zhang, M., and Mei, Q. (2012). We Know What @You #Tag: Does the Dual Role Affect Hashtag Adoption? In *Proceedings of the 21st International Conference on World Wide Web*, WWW '12, pages 261--270, New York, NY, USA. ACM.

Yeh, M.-C. and Lin, H.-W. (2014). Virtual portraitist: Aesthetic evaluation of selfies based on angle. In *Proceedings of the ACM International Conference on Multimedia*, MM '14, pages 221--224, New York, NY, USA. ACM.

Appendix A

Complementary Work

There are many possible ways to complement this study. Some of them were explored in a recent collaborative effort between UFMG and KAIST researchers, in which the author of this thesis took part. Using a subset of the data collected for this thesis, the project brings the following additions to this work:

1. The inclusion of a new dataset in the analyses, comprising photos that include hashtags related to selfies but use variations of the word (e.g., `#selca`, `#selstagram`, `#me`, `#moi`);

2. The study of selfies as an interaction medium, analyzing gender and age homophily between users who post and like/comment selfies;

3. The cultural interpretation of selfies based on correlations with socioeconomic indexes.

The hierarchical clustering of the alternative hashtags based on Pearson's correlation coefficient of their growth trajectories generated groups with different evolution patterns over time with respect to the number of pictures in each group. These differences suggest that underlying mechanisms (i.e., culture, platform) played important roles in how the selfie phenomenon settled. Indeed, the cultural interpretation points to a complex relationship between taking selfies and a country's culture: the chance of using selfie-related hashtags was higher for cultures with stronger local community membership as well as weaker perception of privacy.

One of the alternative hashtags clustering groups in particular, containing hashtags in languages other than English, showed a rapid uptake from the middle of 2013 onwards, indicating the selfie convention has been adopted at different times around the world. Gender and age homophily analyses corroborated with this finding, demonstrating how selfies became more widespread over time.

The details about all experiments and their results can be found in the paper presented in the next pages, which was published in the proceedings of ACM COSN'15 [Souza et al., 2015].

Dawn of the Selfie Era: The Whos, Wheres, and Hows of Selfies on Instagram

Flávio Souza† Diego de Las Casas† Vinícius Flores† SunBum Youn*
Meeyoung Cha* Daniele Quercia* Virgílio Almeida†
†Computer Science Department, UFMG, Belo Horizonte, Brazil
*Graduate School of Culture Technology, KAIST, South Korea

ABSTRACT

Online interactions are increasingly involving images, especially those containing human faces, which are naturally attention grabbing and more effective at conveying feelings than text. To understand this new convention of digital culture, we study the collective behavior of sharing *selfies* on Instagram and present how people appear in selfies and which patterns emerge from such interactions. Analysis of millions of photos shows that the amount of selfies has increased by 900 times from 2012 to 2014. Selfies are an effective medium to grab attention; they generate on average 1.1–3.2 times more likes and comments than other types of content on Instagram. Compared to other content, interactions involving selfies exhibit variations in homophily scores (in terms of age and gender) that suggest they are becoming more widespread. Their style also varies by cultural boundaries in that the average age and majority gender seen in selfies differ from one country to another. We provide explanations of such country-wise variations based on cultural and socioeconomic contexts.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences; H.3.5 [**Information Storage and Retrieval**]: Online Information Services—*Web-based services*

General Terms

Human Factors; Measurement

Keywords

Instagram; Selfies; Cultural Boundaries

1. INTRODUCTION

The amount of rich media is increasing exponentially on the Internet. Online conversations and interactions now involve more images, which are naturally attention grabbing and effective at conveying feelings [21, 39]. Social media in particular has seen a rapid uptake of pictures containing human faces. One notable example is the *selfie* or digital self-portrait, which have become a phenomenal ubiquitous convention of online culture.

Numerous research studies proposed the psychological and sociological framing behind posting selfies, broadly based on narcissism [36], self-exploration [32], self-embellishment [25], and a new genre of art [39]. Other studies approached with the Human Computer Interaction framing to understand pictures with faces and demonstrated their engaging effects [2]. In addition, a project called Selfiecity examined the image traits of single-person self-portraits in five cities across the world [34]. Until now, little effort has been made to quantitatively defining and examining selfies based on a large amount of data.

This paper presents a measurement study of a popular media sharing website, Instagram (www.instagram.com), and characterizes how people appear on Instagram selfies and which patterns emerge from their attention grabbing behaviors. Since selfies are pictures of people, they represent a structured (i.e., social-by-design) form of interaction in social networks. We hence seek to understand whether this new content type can uncover patterns of social interactions. We ask the following two specific questions.

1. **The whos and wheres of selfies:** Can we characterize selfies in terms of age, gender, geography, country, and other cultural variables?

2. **The hows of selfies:** How much attention do selfies receive in terms of likes and comments and to what extent their interactions depend on cultural boundaries?

The first question provides a holistic understanding of what selfies represent in social media. We utilize a subset of photos with hashtags containing the word 'selfie' to determine what kinds of photos are explicitly called as selfies by Instagram users (e.g., how many persons appear in a photo and what kinds of moods these photos contain). Several critical hypotheses related to gender empowerment, group membership, and perceived privacy are tested to better understand the contexts through which users post selfies in a given culture.

Through the second research question, we try to understand how selfie users interact with their audience. Selfies and pictures with faces are more than mere self-expressions; they are phenomenal in grabbing attention and have settled as a popular online practice. By studying the dyadic relationships between selfie users and their audience, we aim to understand what principles rule in pair-wise interactions that involve rich media content. This study tests whether conventional theories such as homophily become strengthened or weakened under the new form of interaction among users.

COSN'15, November 2–3, 2015, Palo Alto, California, USA.

DOI: http://dx.doi.org/10.1145/2817946.2817948.

This paper utilizes a large amount of data gathered from Instagram and carefully selected data about selfies[1] based on three different approaches: (i) pictures containing the word 'selfie' or its immediate variations in hashtags (e.g., #selfie, #myselfie), (ii) pictures containing hashtags related to selfie, but composed only of indirect variations of the word (e.g., #selfcamera, #me), and (iii) pictures containing one or more faces, irrelevant to the choice of hashtags. In this manner, findings in this paper are not dependent on a particular definition of selfies but can provide a holistic view of what people consider selfies. We make the following observations, which are explained throughout the paper.

1. The amount of selfies increased by 900 times over 3 years from 2012 to 2014, which indicates the phenomenon has become a truly ubiquitous convention.

2. Selfies are effective in grabbing attention in social media; they receive 1.1–3.2 times more likes and comments from audience than general posts on Instagram.

3. Young females are the most prominent group who appear in selfies around the world, except for certain countries such as Nigeria and Egypt that show male dominance.

4. There is a complex relationship between taking selfies and a country's culture. The chance of using selfie-related hashtags was higher for cultures with stronger local community membership as well as weaker perception of privacy.

5. Beyond cultural boundaries, selfie-based interactions present homophily variations over time in terms of both age and gender, suggesting that selfies are becoming more mundane.

This work contributes towards better understanding selfies as a popular online phenomenon that have evolved beyond fads, becoming an effective medium of interaction that is attention grabbing and increasing in demand. Our findings demonstrate that selfies are a new window to study collective user behaviors, providing important insights into subjects like perception of privacy, digital cultural norms, and designs of social-networking platforms.

2. BACKGROUND

The rise of selfies is a key trend in the visual Web, assisted by new technological tools and services like Flickr, Pinterest, and Instagram that allow people to better express themselves visually. This section describes several findings from research on self-portrait images, selfies, and Instagram.

2.1 The Meaning of Selfies

Selfies are a ubiquitous phenomenon of modern digital culture. The term was added to the Oxford Dictionaries[2] in 2013, with description: a photograph that one has taken of oneself, typically one taken with a smartphone or webcam and shared via social media.

Different theories emerged to explain why people take selfies. Some state selfies are a mean of self-exploration. As one takes multiple selfies and combine them with different filters, one can re-see herself [9]. A slightly different view is self-embellishment from psychology that states when exposed to slightly modified pictures of themselves, people tend to identify a more attractive version as the original picture [22]. With the ability to control aesthetics of a picture, selfies are a perfect tool for showing the world one's subjective self-image.

A sociological framing recognizes technological possibility to be a necessary condition and also highlights other behavioral factors to be important for selfies [8]. One is a culture of sharing and belonging fostered by the online environment and transmitted through memes. Another is the constant work of shaping and reaffirming self-identity through social actions. In this perspective, selfies could symbolize a convention that is governed by culture and society.

2.2 Advocates and Opponents of Selfies

Selfies are a prominent online culture that have been both criticized and advocated by different parties. Critics say selfies are vain, narcissistic, and attention-seeking; some argue a wide adoption of selfies by female users exacerbates sexual objectification and male gaze [6]. One research demonstrated that adults with the Dark Triad personality trait (e.g., narcissism, psychopathy, and machiavellianism) have a higher chance of posting selfies and editing images on social media [13]. Self-objectification is also known to correlate with increasing photo sharing activities on Facebook among young women [28]. This leads to a worry about the loss of control over one's self-image in an increasingly sharing and hackable culture, where the notion of privacy becomes dependent on the types of interactions that are allowed [31]. The mere presence of an individual's face in a public photo stream can reveal a great detail of information about that person [10].

Defenders of the selfie culture not only deny the above claims but argue selfies are the pinnacle of control and self-expression; selfies allow people to take control over how they and their peers are represented in public, which mobilizes the power dynamics of representations and promotes empowerment [7]. One study interviewed 20 participants who had posted sexual self-portraits and showed how the exchange of such self-portraits can be a transformative experience, increasing their critical self-awareness in a positive manner [35].

2.3 Selfies by Numbers

In contrast to the rich body of work on sociological interpretation of selfies, relatively little attention has been given to data-driven analysis of selfies. A report by eBay Deals states that selfie activity is platform dependent and is well distributed in particular media, for instance Instagram than Twitter [11]. A research conduced by TIME looked at how many "selfies per capita" each city produced by dividing the amount of users posting selfies by the population of each city. They noticed that it was difficult to find a proper local translation for the hashtag 'selfie', as different variations were used everywhere [38].

The largest scale analysis of selfies to date, however, probably was a data visualization project called Selfiecity that aimed at describing features of single selfies (i.e., photos containing a single person's face) in five cities across the world [34]. They investigated demographics, poses, face features, and the moods of 3,200 selfies on Instagram using both automatic and manual methods. Nonetheless, many of the considerations and theories behind selfies (e.g., the contexts, interactions) have not yet been studied under the perspective of data analysis, which is the goal of this paper.

2.4 Studies on Instagram

When it comes to general user behaviors, a number of research utilized the logs gathered from Instagram. For example, researchers examined how color patterns varies between photos posted in two cities [17], how the behaviors of teens and adults on the network differ [20], how users can be grouped based on the types of

[1]Data used in this study are available for research purposes at http://instagram.camps.dcc.ufmg.br/selfies/

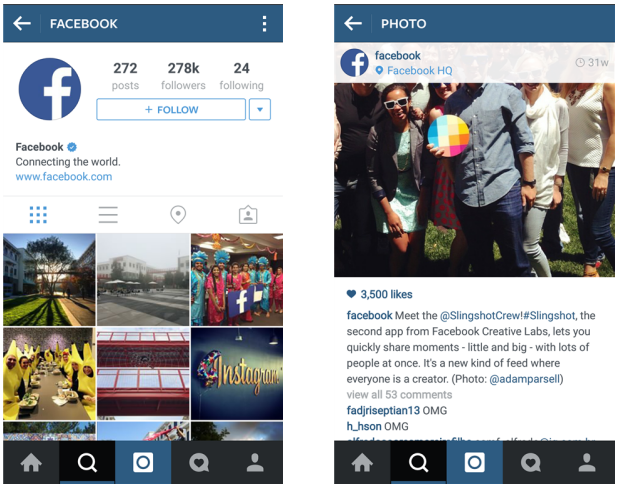
[2]http://www.oxforddictionaries.com

content they share [19], and even how Instagram photos shared on Twitter can be used as sensors to study user characteristics in different cultures [33]. Therefore, the present research can be seen yet as an additional contribution for Instagram characterization efforts, complementing previous works in this direction.

3. INSTAGRAM DATA

We started data collection by inferring ranges of user IDs. This step involved forward sampling batches of 10,000 numeric IDs for every range of 10 million, starting from zero. None of the inspected IDs were valid after the count of 1.6 billion. Through this process, we could identify which specific ranges are valid ID space. Based on these ranges, we next randomly sampled 1% or 16 million IDs to build an initial seed set and found 42% of them to be in use; the remaining IDs were either deleted or not in use. Not all of these in-use accounts could be viewed publicly due to privacy settings; 78% of them were public accounts and the remaining 22% were private accounts, whose profile and feed information could be viewed only by confirmed friends on Instagram.

We gathered profile information of all public users (5,170,062 in total) as well as all of their publications (known as "feeds" on Instagram) for a three-year period between December 2011 and December 2014. There were 153,979,348 data objects called "media", which include a picture or a video along with some metadata such as hashtags, caption, timestamp, and URLs. This paper only focuses on pictures, which takes up a large majority (97%) of all media on Instagram. Figure 1 shows an example profile and feed, where profile includes user-level counts (e.g., posts, followers and following) and feed includes images and picture-level metadata (e.g., likes, caption, hashtags, comments and geolocation, if any). All of these pieces of information can be accessed through Instagram's Application Programming Interface (API).



(a) Profile (b) Feed

Figure 1: Instagram mobile application interface.

One important aspect considered in this paper is geography of selfie users, which were inferred by mapping geolocation tags in the photo content. Instagram is known to have high rates of photos that contain geotags. Among all media gathered, 35,030,356 pictures published by 770,095 users contained any geolocation information. The Global Administrative Areas database [16] was used to map location coordinates to corresponding country and city names.

Another aspect considered is user demographics, which we inferred from photos of users by the Face++ tool [27]. Face++ is an online API that detects faces in a given photo and predicts information about each person in the photo such as age and gender. Its accuracy is known to be over 90% [2]. Age is given in years along with a confidence range; gender is given as 'male' or 'female' with a confidence value between 0 and 100; and smile is given as a score between 0 and 100. We ran Face++ for a random subset of photos and gained demographic information for 2,286,401 pictures posted by 738,901 distinct users.

3.1 Data Validation

To understand potential bias in data, we compare statistics obtained from our data with those of other reports on Instagram. The service reached 300 million active users in 2014 with more than 30 billion photos shared on the network.[3] A research conducted by Pew showed that Instagram is not only increasing its overall user base, but also is seeing a significant growth in almost every demographic group in the United States. Most notably, 53% of young adults between age 18 and 29 used the service in 2014, compared to 37% a year before. The service is also known to have more female users than males [30].

Given its massive scale, findings in this study are bound to insights from a small subset of data. Nonetheless, data we observed had similar properties to what was reported on Instagram. We examined the age and gender distribution of users in our data. We selected a random sample of 100,000 users with at least 10 pictures and examined the profile pictures of such users. The resulting age and gender distribution is shown in Figure 2, where 62% of the sample users are inferred as female and the median ages are 18 and 23 for females and males, respectively. The proportions of different age groups are similar to other reports, like the Pew research and the Selfiecity project [34].

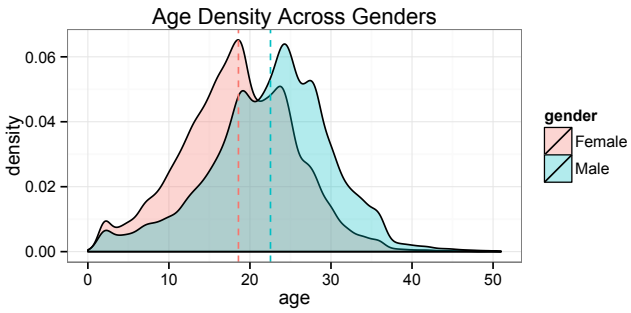
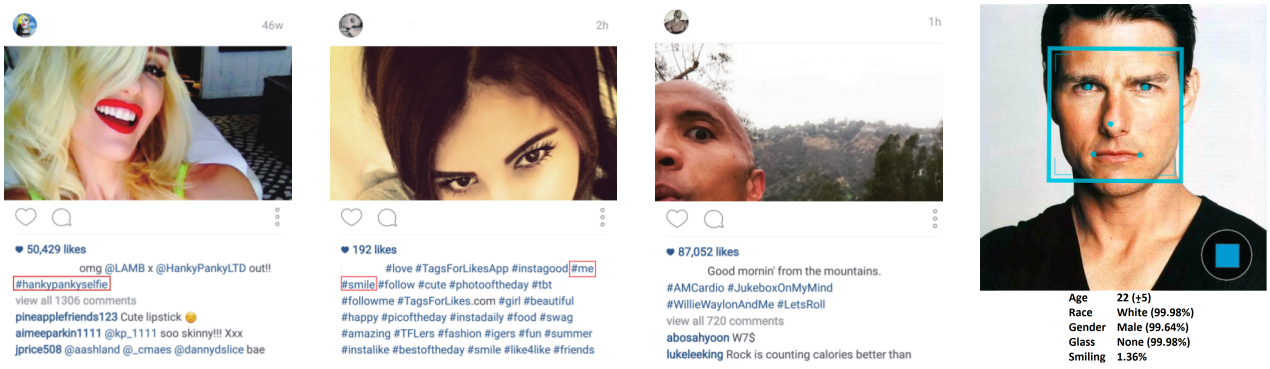


Figure 2: Density plot of users ages, separated by gender.

3.2 Extracting Selfies

We devised three methods to extract selfie posts. Photos with selfie-related tags indicate what Instagram users identify explicitly as selfies. In addition to two datasets found in this manner, we also examine pictures with faces in general. Note that not all photos belonging to this category are selfies (i.e., photos taken of oneself), yet the third dataset will help us understand the engaging effects of faces. Lastly we utilized a random set of photos for comparison. The summary of the four datasets used in the remainder of the paper follows:

[3]http://instagram.com/press/



(a) Selfie (b) Alt (c) Face (d) Face++ Inference

Figure 3: Example pictures of Selfie, Alt, and Face datasets as well as features predicted by Face++.

Dataset	Description	# Pictures	# Users
Selfie	Pictures with hashtags containing 'selfie' (e.g., #selfie, #selfietoday)	1,196,080	214,656
Alt	Pictures with alternative hashtags for 'selfie' (e.g., #selca, #selstagram)	2,453,749	242,650
Face	Pictures with face(s) detected using the Face++ tool	1,921,207	315,751
All	Randomly chosen set of pictures	10,000,019	184,615

Table 1: Number of media and users in each of the four datasets used in this paper.

1. Selfie: a collection of pictures that contain the word 'selfie' in hashtags. Examples include #selfie, #selfietime, and #selfiesunday, which are an explicit indicator.

2. Alt: a collection of pictures that include hashtags related to selfie but use variations of the word. For instance, 'selca' can be used instead of 'selfie' in some contexts.

3. Face: a collection of pictures containing one or more faces detected using the Face++ tool.

4. All: a random collection of 10M Instagram pictures. We compared it with the other three datasets to identify the distinct characteristic of selfies.

Photos in Figure 3 are examples of the three datasets, which were all posted by popular users on Instagram. Figure 3(a) is classified as Selfie due to its hashtag #hankypankyselfie, whereas Figure 3(b) is classified as Alt for its hashtag #me and #smile. The face photo in Figure 3(c) did not contain any selfie-related hashtags, hence it was classified as Face by the Face++ tool. Note that all three types of photos are valid selfie content, which we consider in this paper. Figure 3(d) shows features detected by Face++ on a celebrity photo of Tom Cruise. Table 1 summarizes the description and quantity (the number of pictures and distinct users) of three selfie datasets as well as that of All.

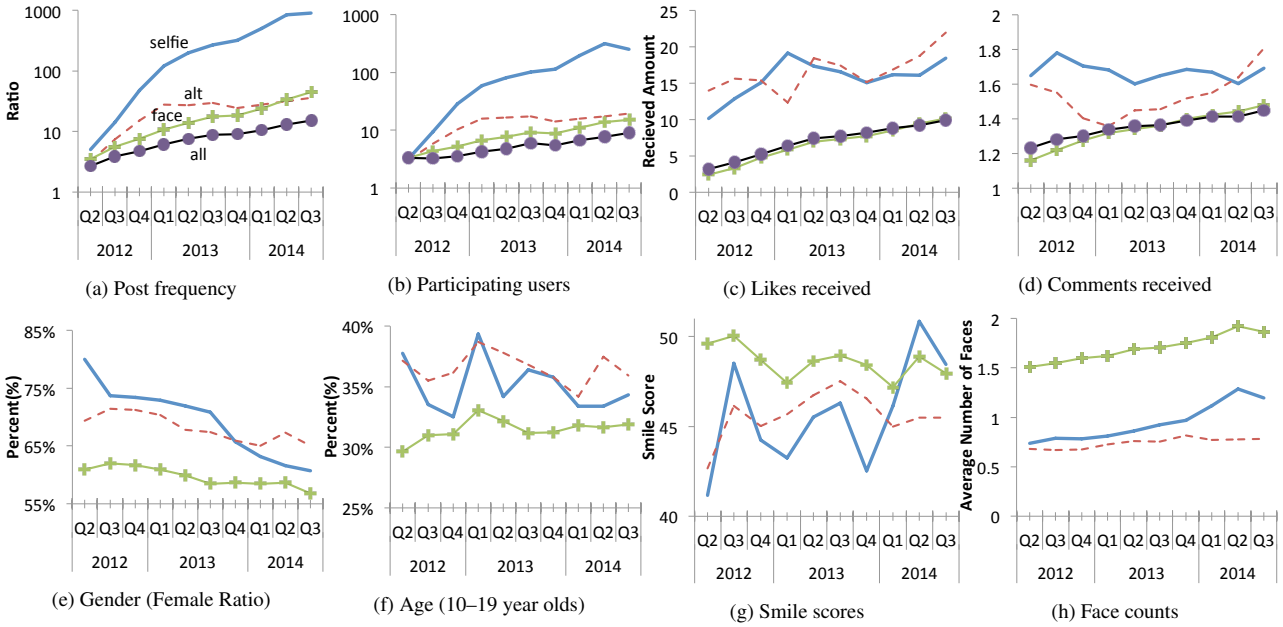
Now we describe our heuristic method to identify Alt photos. For this we first need to examine what users call as selfies on Instagram. The Selfie dataset involved a total of 43,874 distinct hashtags containing the word 'selfie'. To find alternative hashtags for 'selfie', we calculated a similarity score for each hashtag in a way akin to Pointwise Mutual Information [5]. First, we separated all pictures into two sets: one set containing pictures that either have a single-person face or the hashtag #selfie (called True or T) and another set containing pictures that neither have a face nor the hashtag #selfie (called Unknown or U). The similarity score was then designed in an approximate manner to give higher scores to hashtags in the first set, T, as follows:

$$S_h = \frac{f_{h,T} \times u_{h,T}}{f_{h,U} \times u_{h,U}} \quad (1)$$

where S_h is the similarity score for a hashtag h in relation to selfie posts. $f_{h,[T,U]}$ is the frequency of the hashtag h in the set T or the set U and $u_{h,[T,U]}$ is the number of users who use the hashtag h in T or U.

In this manner, we were able to identify words that describe selfies in various languages such as Turkish, Russian, Malaysian, Indonesian, Filipino, etc. Note that the obtained hashtags had high potential to appear with other selfie-related terms, yet they do not cover a complete set of selfie-related terms. The top-10 variant hashtags found with this method were: #shamelessselfie, #gaybeard, #butfirst, #özçekim (the word representing selfie in Turkish), #ethanymotagiveaway, #gaysian, #лифтолук (Russian), #dolledup, #ozcekim (Turkish), as well as #pacute (Malaysian and Filipino). We also included words that are used to describe selfies such as #me and #self in the Alt dataset. The final Alt dataset contained a total of 81 variant hashtags.

When we examined the photo content through Face++, the three selfie datasets varied slightly in terms of user demographics. First, the median age of the users in photos were 22, 20, and 21 for Selfie, Alt, and Face respectively. Alt photos contained the youngest users. The proportion of females to males varied from 64%, 69%, and 59% for the three datasets (in same order as above). Photos in Alt were more likely to contain faces of female users, while Face had a better balance of male and female users. Finally, we compared how many faces appear in a given photo, as sometimes multiple faces may appear in a photo (i.e., groupie). The three datasets contained on average 1.12, 0.76, and 1.75 number of faces for the three datasets (in the same order as above). Some photos in Alt were selfies of pets or body parts, in which case they contained zero human face. These variations, although not prominent, may indicate the differences in base demographics.



(a) Post frequency (b) Participating users (c) Likes received (d) Comments received

(e) Gender (Female Ratio) (f) Age (10–19 year olds) (g) Smile scores (h) Face counts

Figure 4: Longitudinal trend of selfie posts across the datasets.

4. TEMPORAL DYNAMICS

The longitudinal data provides a unique opportunity to examine post and interaction trends from 2012 to 2014.

4.1 Patterns of Selfies Over Time

We first study how a wide range of features changed over time. We examine post frequency, attention (likes and comments), demographics (gender and age), and image (smile score and number of faces in a given photo) for different definitions of selfies. Demographics and image features depend on Face++ data, thus the All dataset was excluded from these analyses.

4.1.1 Post Frequency

Post frequency measures how popular a given photo type is over time. Figure 4(a) shows the post frequency trends over time, where the x-axis represents time in quarters (i.e., three-month periods) and the y-axis represents the relative increment or decrement compared to the initial quarter (i.e., the first quarter of 2012). Therefore, a value of 1.0 in this figure means the post volume is identical to what was measured in the initial quarter and a value of 10.0 means an increment by 10 times.

While the frequency of All increased 15 times (from 103,520 in the first quarter of 2012 to 1,560,697 in the third quarter of 2014), the post frequency of Selfie increased rapidly by 900 times (from 297 to 269,454) over the same time period. Alt and Face also became popular compared to All, yet not at the same degree as Selfie. When we compare the speed, All and Face show a relatively steady growth in volume, whereas the growth of Selfie and Alt is rapid at first and becomes stagnant towards the end of 2014. A similar trend is seen in the graph of participating users who post selfies in Figure 4(b). Selfie again shows orders of magnitude larger growth than any other content type. Selfie and Alt show a stagnant growth towards the end of 2014 as opposed to All and Face. These growth trends capture well the rapid rise of selfies on Instagram, which seemed to have peaked between 2012 and 2013.

4.1.2 Content Popularity

The amount of attention a photo gained can be inferred by examining the number of likes and comments. Figure 4(c) shows the absolute geometric mean of likes per picture, which demonstrates that Selfie and Alt receive nearly 2-3 times more likes than the other content types. This means pictures with an explicit marker about 'selfie' grab more attention from audience than merely containing a face in a photo. Examining closely, however, the relative gap between Selfie and All decreases over time from nearly 3.2 times during the thriving initial spread to 1.3 times in 2014.

The geometric mean of comments received in Figure 4(d) shows a similar trend. Again Selfie and Alt receive 1.1–1.5 times more comments than the other content types, although this gap is decreasing over time. This observation indicates that pictures owning a selfie-related hashtag are effective in grabbing attention, yet their engaging effect becomes less pronounced over the years (perhaps as selfies become widespread and become mundane).

When compared to the recent literature on the effect of containing faces in pictures, the work in [2] demonstrated pictures with faces tend to get 38% more likes and 32% more comments compared to other content on Instagram. While we cannot make a direct comparison, our results further highlights that attributing particular hashtags (such as #selfie) could incur even a higher level of attention than merely posting photo with faces.

4.1.3 Demographics

We next investigate what kinds of users post selfies, by employing the Face++ tool to infer their age and gender based on profile pictures. Figure 4(e) shows the proportion of female users over time. The high female-gender ratio indicates that selfies were initially posted primarily by female users than male users for all three datasets. These rates are high even when we consider the high female prevalence on Instagram. During the 3 year period, however, this difference diminished until the ratio almost reached the base gender ratio of the network, as seen by Face.

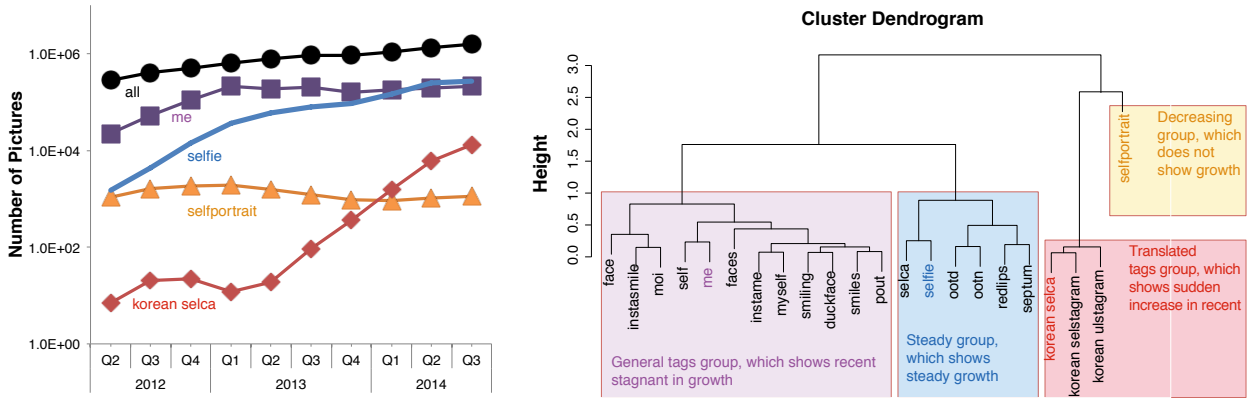


Figure 5: Evolution of tags in Alt and hierarchical clustering based on Pearson correlation.

Figure 4(f) shows the demographic makeup for 10–19 year olds. The age distribution of Face confirms the general perception that young people are the most active users who post photos with faces on Instagram, constituting nearly 32% of all participating users. Nonetheless, this ratio is even larger for Selfie and Alt, meaning that young people are more likely to tag their pictures with selfie-related hashtags. The gender and age analysis together indicates that young female users on Instagram drove the selfie momentum during the initial stage in 2012, which is indeed confirmed by the plot of percentage of young females over time for each dataset (not included here due to space limitations).

4.1.4 *Smiles and Face Counts*

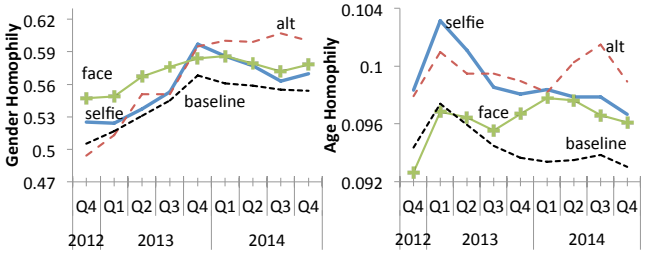
Would face pictures that are tagged as selfies present more joyful atmosphere? The smile score, detected by Face++, indicates the degree of smile in a face, with a score of 100 indicating the highest level of smile. Comparing the average smile scores in Figure 4(g), faces in Selfie and Alt are not more joyful than Face. Overall, the scores of all three datasets are ranged between 40 and 52, which do not necessarily represent a big pleasant smile. We do not see any particular correlation between the smile score and the type of data.

Another question we had was to measure how many faces appear in selfie photos. Would people associate single-person photos as selfies? Figure 4(h) shows the average number of faces per picture for the three datasets. At a glance, Face contains the highest number of faces (1.5–2.0 faces per picture) than Selfie and Alt. The latter types sometimes included zero faces thereby pushing the average below 1.0, where pictures were on parts of body, pets, or other animals. From mid 2013 and onward, there is a gradual increase in face counts for Selfie dataset, which implies that Instagram users increasingly recognize pictures containing multiple faces as selfies.

4.2 Trajectory of Selfie Hashtags

We have so far found similarity between Selfie and Alt, both of which contain hashtags about selfies. It is natural to observe *multiple* variants of the hashtag as they could indicate cultural traits and contexts. In order to observe how different hashtags gained popularity over time, we looked at their adoption trajectory over time. We identified all hashtags in Alt that appeared more than 10,000 times (21 variants) along with #selfie, then calculated the Pearson's correlation coefficient for growth trajectories of all of these variations. A hierarchical clustering approach was applied to identify hashtags that showed similar growth patterns.

Figure 5 shows the result, where hashtags are divided into four groups. The first group, which is the largest in size, contains general description of selfies such as #me and #face. Hashtags in this group initially showed high popularity and then became stagnant over time. The second group, which includes social media specific terms such as #selfie, #selca and #ootd (outfit of the day), shows steady growth over time. The third group, containing hashtags in languages other than English, shows a rapid uptake later in time, indicating that the selfie convention has become widely adopted at different times around the world. The last group contains a single hashtag, #selfportrait, whose growth does not change much over the years. These differences in popularity trajectory of hashtags suggest that underlying mechanisms (i.e., culture, platform) played important roles in how the selfie phenomenon settled.



(a) Gender Homophily (b) Age Homophily

Figure 6: Homophily scores for likes. The *baseline* gives the expected scores for random interactions considering the three datasets.

4.3 Selfies as an Interaction Medium

Since selfies are pictures of people, they represent a structured form of interaction. As shown in the previous subsection, they are an effective medium of communication that incur more likes and comments than other types of content on Instagram. Next, several questions motivate us to examine interaction patterns involving selfies; for instance, how likely males will respond to selfies of other males or other females? Do people tend to interact more frequently with others of the same age? Would these patterns change for pictures explicitly marked as selfies?

One method to examine these questions is homophily, which describes the tendency of individuals to associate and relate with similar others. Homophily is a central hypothesis that can explain user behaviors in various offline and online social networks [26]. This study tested homophily by studying the dyadic relationship between selfie owners and their audience based on likes and com-

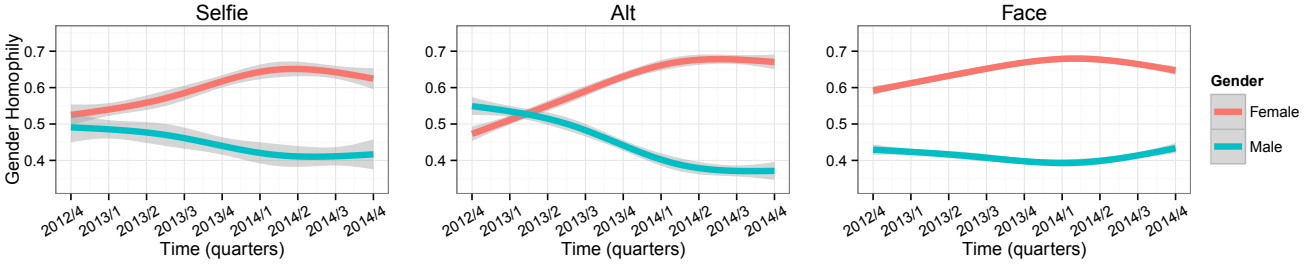


Figure 7: Intra-gender homophily evolution for likes interactions across three datasets: Selfie, Alt, and Face.

ments. In particular, we designed the following experiment: first we built three samples, each containing 200,000 randomly chosen pictures from Selfie, Alt, and Face datasets. We then fetched a random set of likes and comments for each of the pictures and determined the age and gender of such interacting users based on Face++. Next, two measures were defined, following the literature on homophily: Gender Homophily (*H_Gender*) and Age Homophily (*H_Age*). *H_Gender* was calculated as follows:

$$H\_Gender = \frac{F_{mm} + F_{ff}}{F_{interactions}} \quad (2)$$

where $F_{interactions}$ is the total number of interactions of any kind, F_{mm} is the number of male-male interactions, and F_{ff} is the number of female-female interactions. Hence, *H_Gender* measures the rate of same-sex interactions out of all combinations seen in data. Its values range from 0 to 1, where 1 represents the highest level of homophily.

Similarly, *H_Age* was calculated in the following way:

$$H\_Age = \text{RMSE}(A_i, A_j)^{-1} \quad (3)$$

where A_i and A_j are lists containing the ages of interacting users and RMSE is the Root Mean Squared Error of two lists, which represents how close they are related. For example, if $H\_Age = 0.1$, then $1/\text{RMSE}(A_i, A_j) = 0.1$, that is, $\text{RMSE}(A_i, A_j) = 10$, which indicates the mean difference in age between interacting users is roughly 10 years. The smaller the differences in age, the higher *H_Age* will be, indicating greater homophily. We have chosen RMSE instead of mean absolute difference in order to increase the penalty for higher age differences.

Gender and age homophily scores for likes are shown in Figure 6 for each dataset, along with the expected scores for random interactions (labeled as *baseline*) considering the three datasets together. The baseline was built using bootstrap sample in order to get both source and destination users in each quarter, and then calculating the homophily scores for these random pairs of users.

It is possible to observe that Alt and Selfie present more variation in gender homophily scores over the course of three years than Face. To further investigate this finding, we calculated intra-gender homophily scores for likes as follows:

$$H\_Male = \frac{F_{mm}}{F_{m\_all}} \qquad H\_Female = \frac{F_{ff}}{F_{f\_all}} \quad (4)$$

where F_{mm} and F_{ff} are defined as previously, F_{m_all} is the total number of interactions of any kind with males, and F_{f_all} is the total number of interactions of any kind with females. Thus, *H_Male* and *H_Female* are proportions of same-sex interactions calculated for each gender separately. The values of each one range from 0 to 1, where 1 represents the highest level of homophily.

The intra-gender homophily scores of males and females are plotted together for each dataset in Figure 7. Two main facts arise from the graphs. First, there is a clear female bias in likes interactions, which goes in line with female prevalence in the network. Second, both Selfie and Alt present a unique gender homophily evolution. In the beginning, males tend to engage in like interactions more than females, as the scores are close to 0.5 and there are more females in the network. Then over time, their behavior converges to Face's, suggesting that selfies are becoming more mundane. No significant pattern of gender-level homophily emerged for comments in the datasets.

In the case of age homophily in likes interactions, the scores for Selfie follow the pattern of Alt's until the beginning of 2014. After that, Selfie follows the same trend as Face, with decreasing homophily scores, while Alt scores peak and then also decrease. This finding suggests once more that selfies are becoming more widespread. Again, homophily for comments does not present a clear distinction among datasets, although its values are in the same range (0.092 to 0.104) as homophily for likes. The pattern is more pronounced for likes than comments possibly because likes are larger in volume than comments – likes are a lightweight communication form that happen more often on Instagram than comments.

5. CULTURAL INTERPRETATION

Having examined the longitudinal trends, we now provide explanations for the selfie patterns and examine cultural aspects. We investigate whether the new selfie convention portrays any cultural and socioeconomic contexts (i.e., country-wise variations). This is an important question because other online behaviors have been shown to depend on culture [14]. To group selfies by cultural boundaries, we aggregated all geotagged pictures by countries and considered only those countries with at least 20 pictures for analysis. The total number of countries analyzed in each dataset is shown in Table 2 for all indicators used.

Ind.	Selfie	Alt	Face
GGI	111	115	117
PV	54	55	56
IDV	67	68	68
LCS	53	54	55
WCS	53	54	55
Choice	54	55	56
Trust	54	55	56
UAI	67	68	68

Table 2: Number of countries per dataset for each indicator.

As one might expect, selfie patterns differed from one country to another. For instance, the mean age and female-to-male ratio varied as shown in Table 3, which shows the top-5 and bottom-5 countries based on female prevalence of selfies. South Korea is ranked the top with its 71% of selfies shared by female users. Even though there is a general bias towards female users that we have demonstrated in the previous section, several countries such as Nigeria and Egypt present a heavy male bias.

Top 5			Bottom 5		
Country	M.age	F.prev	Country	M.age	F.prev
KOR	16.9	0.71	NGA	23.5	0.31
KAZ	19.3	0.68	EGY	22.7	0.28
PHL	17.9	0.68	SAU	20.4	0.28
CHN	16.6	0.67	KWT	22.0	0.28
UKR	20.9	0.66	IND	23.9	0.20

Table 3: Top and bottom countries by female prevalence.

In order to test whether these country-wise variations can be explained by cultural contexts, we utilized popular international socioeconomic indicators as well as indicators from two important sources: (i) World Values Survey (WVS) that is an individual-level survey probing cultural values of citizens in 59 countries between 2010 and 2014 [1] and (ii) Hofstede's Cultural Dimensions (HCD) that is a five-dimensional model of cultural differences studied since 1971 by Geert Hofstede [18].

5.1 Hypotheses

We set up three hypotheses that could enrich our understanding of country-wise variations in selfie trends:

1. **Gender Empowerment (H_1).** There is no consensus on whether selfies enhance male oppression or allows for a way of asserting agency, although the answer is probably more nuanced [23]. Nevertheless, it is reasonable to expect that differentiation in gender roles within a country will be reflected in the proportion of women or men taking selfies. We hence hypothesize that women in countries with higher gender equality are more comfortable in sharing selfies publicly than in less equal countries.

2. **Self Embellishment & Membership (H_2).** If selfies are indeed a manifestation of self embellishment, it is expected that they will be more prevalent in individualistic societies than in collectivist countries [12]. If, on the contrary, selfies are more widely used as means of belonging and a norm, then they will be more prevalent where citizens feel a strong tie with their local community or with a global connected community.

3. **Intimacy & Privacy (H_3).** If selfies represent one's sense of intimacy and privacy in an online world, then trust in people and the perception of control over one's own life should mediate the behavior. Among relevant socioeconomic indicators, one may consider the level of perceived uncertainty and loss of control. We hypothesize that countries where people are aversive to uncertainty will post comparatively fewer selfies than otherwise.

5.2 Independent and Control Variables

To test the first hypothesis, we compared the proportion of females detected by Face++ in each country with several measures of gender equality. Since the proportion of women in each country varies, we calculated the relative increase or decrease of female prevalence against the observed proportion of women in the World Bank data.[4] We define *GenderBias* as follows:

$$\textit{GenderBias} = \mathrm{P}_{\text{selfies}} - \mathrm{P}_{\text{census}} \tag{5}$$

where $\mathrm{P}_{\text{census}}$ is the proportion of females observed in a country.

We used two relevant socioeconomic measures: (i) the Gender Gap Index (GGI) and (ii) Patriarchal Values (PV). The former is published yearly by the World Economic Forum and measures the relative gaps between women and men across four key areas: health, education, economy, and politics [3]. The score represents how much the gaps has been closed, so a high score means a more equal society. The latter is a scale of four questions from the WVS in which the respondents state whether they agree with values tied to stereo-typical gender roles [24]. A high score here means a less equal society in that cultural values are strongly associated with gender inequality.

To test the second hypothesis, we compared the rate of selfie posts at each country, *Prevalence*, as follows:

$$\textit{Prevalence} = \log \frac{F_T}{F_{All}} \tag{6}$$

where F_T is the frequency of posts in dataset T and F_{All} is the set of posts in the All dataset. We used a logarithmic value since the trend is heavy tailed across countries.

We used three relevant socioeconomic measures: (i) the Individualism score (IDV), (ii) Local Community Score (LCS), and (iii) World Citizen Score (WCS). The first indicator is from Hofstede's Cultural Dimensions and describes how separated is an individual from larger social groups in a country. The second and third are from a recent work [37] and represent the average value of the response to the following propositions: "I see myself as a part of my local community" and "I see myself as a world citizen". A high score indicates a strong community membership. These scores are proxies for how strongly tied citizens of a country are to their local community as well as to the international community at large.

To test the third hypothesis, we resorted to the part of WVS that is used as an indicator of *generalized trust*, i.e., the trust in people outside one's social circle [4]. This question is: "Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?" We used the proportion of citizens who agree with the "Most people can be trusted" answer as our *Trust* indicator. We selected the following question to probe for the perception of choice and control: "How much freedom of choice and control do you feel you have over the way your life turns out?" Responses are situated in a scale from 1 (*no choice at all*) to 10 (*a great deal of choice*), which we averaged per country and used as our *Choice* indicator.

We also selected the dimension *Uncertainty Avoidance* (UAI) from HCD, which indicates a society's tolerance for uncertainty and ambiguity and to what extent the members of a culture feel either uncomfortable or comfortable in unstructured (novel, unknown, unusual) situations. Our hypothesis was that Trust and Choice would be positively correlated and UAI would be negatively correlated to the prevalence of selfies in a country.

Finally, we also considered the following sets of *control* variables to take into account that Instagram is not used evenly across countries. Although the measures reported in the *Hypotheses* subsection make intuitive sense, they are strongly related to confounding factors that varies either in Instagram or between countries.

[4] http://data.worldbank.org/

We chose three variables to control and calculated the *partial correlations* between the variables of interest. Partial correlations allow us to estimate the relationship of two variables X and Y after partialling out the effect of the control variable Z. They are equivalent to constructing two linear models using X and Y as dependent variables and Z as independent variable, then correlating the residuals of each linear model.

The control variables we chose were *log GDP per capita* as an indicator of economic development, *Internet penetration* as a proxy for technology diffusion, and the *average age* of Instagram users (estimated with Face++ data) to account for the different age profiles between countries. Thus, all correlations we report are between the residuals of the variables after their covariance with the control variables has been partialled out. The correlations between measures/indicators and the control variables are not shown here due to space limitations, but are available in our shared repository.

5.3 Results

Results are displayed in Table 4. The positive relationship between GenderBias and gender equality indicators is clear in all datasets, thus confirming H_1. This is true both for the equality measured by the country's socioeconomic structure—parity of gender in social living and access to public institutions—as for the cultural values in which the citizens of a country believe. The presence of an effect in all datasets show that this relationship holds for many definitions of selfies. However, if one is not to consider the Face dataset as representing actual selfies, one can argue that this is the consequence of a broader effect of the presence of women in the network. It is interesting to note, however, that the strongest correlations in each of the indicators was not with the Face dataset, but with the Alt dataset.

We could not detect a meaningful relationship between Individualism Score (IDV) and Prevalence for either direction, and Local Community Score (LCS) and World Citizen Score (WCS) show significant correlations in opposite directions. LCS is moderately correlated to selfies tagged as such, which goes in line with the idea that selfies are tied to a sense of belonging to a community, namely the local community. However, this effect seems exclusively related to the Selfie dataset, as the coefficients are negative (although non-significant) in the other datasets. Moreover, WCS is negatively correlated with the Alt dataset and not meaningfully correlated with the other datasets.

This finding demonstrates a complex relationship between taking selfies and a country's culture of individuation and connectedness. The effect of a country's individualism, if exists, is much smaller than other factors related to belonging to a community, and could not be detected by us. But even these other factors are not easily interpretable and may be related to different conceptions of selfies. The positive relationship between the Selfie dataset and LCS advocate for the idea that taking selfies—and tagging them as such—is related to the importance a culture gives to belonging to a community. However, we expected that the relationship with WCS would follow the same path, which did not. A possible explanation is that, since the Alt dataset includes many hashtags that represent similar concepts of a selfie in a given country, its negative relationship with WCS spans from the attitude of citizens of the country to adapt and transform foreign "memes" into their cultural reality. Thus, a country with citizens that do not strongly identify themselves as world citizens will still have selfies, but adopt different tags. It is worth mentioning that the correlation between the Prevalences of these two datasets (Selfie and Alt) is only moderate ($r = 0.60, p < 0.0001$).

Hypotesis: Measure	Ind.	Selfie	Alt	Face
H_1 : GenderBias	GGI	0.34^{***}	0.41^{***}	0.32^{***}
	PV	$-0.20^{\diamond}$	-0.38^{***}	$-0.19^{\diamond}$
H_2 : Prevalence	IDV	0.09	0.06	-0.04
	LCS	0.22^{*}	-0.13	-0.07
	WCS	0.08	$-0.17^{\diamond}$	-0.03
H_3 : Prevalence	Choice	-0.04	$-0.19^{\diamond}$	0.13
	Trust	$-0.19^{\diamond}$	-0.17	-0.29^{**}
	UAI	0.14	0.21^{*}	0.31^{***}

Stars represent significance values: $p < 0.0001(^{***})$, $p < 0.001(^{**})$, $p < 0.01(^{*})$ and $p < 0.05(^{\diamond})$.

Table 4: Correlations between selfie-related measures and sociocultural measures. There is a complex relationship between taking selfies and a country's culture. The chance of using selfie-related hashtags was higher for cultures with stronger local community membership as well as weaker perception of privacy.

As for H_3, Choice, Trust and Uncertainty Avoidance (UAI) are related to Prevalence, but in the opposite direction that we expected. In countries where the citizens trust each other and feel they have more control over their lives or are not as aversive to uncertain outcomes, people take *fewer* selfies relative to other kinds of pictures. The analysis shows that selfies are not inhibited by a sense of lack of control and certainty but somewhat stimulated by it. We may speculate two (non-excluding) scenarios that could explain our finding: 1) selfies are an assertion of control over one's identity, so they are more important in places where citizens feel they need this; 2) part of what drives selfies is an attitude or set of values that also promotes lack of trust, a sense of lack of control, and aversion to uncertainty. Unfortunately, we cannot distinguish these two scenarios from our results.

6. DISCUSSION AND CONCLUSION

6.1 Implications

Selfies are ever more present in today's online culture. This work presented a measurement study based on a large amount of data gathered from Instagram, and defined selfies through three different ways to understand the whos, wheres, and hows of its patterns. We investigated the distributions of post frequency, likes, comments, age, gender, smile scores, and face counts over the course of three years. These patterns collectively show that selfies have become extremely popular (i.e., spreading to a wider set of users in terms of number, age, and gender bias). We examined how different variants of the selfie hashtag gained popularity over time. The longitudinal study also explored the role of homophily in terms of age and gender in selfie-oriented lightweight interactions.

These temporal patterns showed country-wise variations, some of which could be explained by cultural contexts and others need further investigation. This paper showed that gender equality indicators are tightly related to the proportion of women that appear in different definitions of Instagram's selfies, which goes in line with views that selfies mobilize the power dynamics of representations and promotes empowerment (in this case for women) [7]. This paper also showed that there is a complex relationship between taking selfies and a country's culture of individuation and connectedness. Finally, in contrast to our expectation, selfies were less prevalent in more trustful and not risk averse cultures. These findings show general tendency and we do not claim that there is any causal relationship with culture and selfies.

In a recent interview, Instragram founder Kevin Systrom said that "the selfie is something that didn't really exist in the same way before Instagram."[5] Indeed, selfies take center stage on Instagram, and this work shows that they do so in very specific ways. In quantitatively capturing those ways, we offer two main insights to designers of social-networking platforms. First, the adoption of selfies show a high variability across countries. In countries lacking considerable adoption (because of, e.g., gender issues), designers should think about new ways of encouraging specific segments of the population. Second, it is well known that individuals tend to interact with like-minded others. For selfies on Instagram, however, this tendency is further emphasized. As a result, a filter bubble might well emerge [29]: users become separated from other dissimilar users, effectively isolating them in their own cultural and self-portrait bubbles. In a way similar to what researchers in recommender systems have done [40], designers should build and integrate new algorithmic solutions that partly counter the ominous consequences of self-portrait bubbles.

6.2 Limitations

One limitation of this work is that selfies have different meanings to social media users. Face count varied in that some considered single-person photos as selfies, while others allowed multiple faces to be included. Some explicitly identified photos containing human faces, while others tagged pictures of their pets, animals, personal belongings, as well as body parts as selfies. These examples illustrate the paradigm shift in how people define selfies. The current study tried to capture these diverse meanings by borrowing three different definitions. Nonetheless, our methodology is limited by the use of hashtags and images, as not all selfies will contain such explicit markers.

Another limitation is in the scope of cultural interpretation. Understanding cultural contexts is immensely important, but also very challenging because it is difficult to separate out the complex interplay among socioeconomic factors. This work employed a handful of popular indicators and attempted to provide better explanations for country-wise variations. This, although preliminary, is a meaningful first step towards understanding how a new online phenomenon spread across the world.

6.3 Future Directions

Many questions addressed by this work could be investigated in greater detail to highlight possible nuances not captured by the experiments done. For example, an evaluation of how exactly Face++ accuracy is impacted by the particularities of selfies (close-up, distorted or partial views of a face, etc.) could help to know if there are adjustments to be made in this respect; a detailed analysis of usage patterns and spread of Instagram across countries could reveal how local differences affect the overall temporal dynamics found.

Another possible direction would be to dig further into user-level analyses. For instance, a deeper investigation of general differences in users activities in the network could allow to identify how to appropriately take these differences into account when studying interactions among users. A diverse approach could be to verify how selfie-related behaviors vary from user to user or even across different classes of users (e.g., celebrities and occasional users). This could also bring information about the effect of users characteristics on the engagement associated with the selfies they publish.

Given its multifaceted character, selfies present yet many dimensions not explored in this research. The attachment of selfies with emotions, for example, could be investigated combining different sentiment analysis methods [15] based on comments, captions and hashtags related to selfies.

[5]http://goo.gl/cpqhhE

Acknowledgments

We thank the reviewers and our shepherd, Emre Kiciman, for providing valuable comments that helped us improve the paper.

This work was partially supported by CAPES, CNPq, FAPEMIG, and the Brazilian National Institute of Science and Technology for the Web – InWeb. This work was also supported by the IT R&D program of MSIP/KEIT (R0184-15-1037) and the BK21 Plus Postgraduate Organization for Content Science of Korea.

7. REFERENCES

[1] W. V. S. Association. World values survey wave 6 2010-2014. http://www.worldvaluessurvey.org/, Apr 2015.

[2] S. Bakhshi, D. A. Shamma, and E. Gilbert. Faces engage us: Photos with faces attract more likes and comments on instagram. In *ACM Conference on Human Factors in Computing Systems*, 2014.

[3] Y. Bekhouche, R. Hausmann, S. Zahidi, and L. D. Tyson. The global gender gap report 2014. Technical report, World Economic Forum, 2014.

[4] C. Bjørnskov. Determinants of generalized trust: A cross-country comparison. *Public Choice*, 130(1-2):1–21, 2007.

[5] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, Mar. 1990.

[6] N. L. Cole. The Selfie Debates, Part I. http://sociology.about.com/od/Ask-a-Sociologist/fl/The-Selfie-Debates-Part-I.htm.

[7] N. L. Cole. The Selfie Debates, Part II. http://sociology.about.com/od/Current-Events-in-Sociological-Context/fl/The-Selfie-Debates-Part-II.htm.

[8] N. L. Cole. Why we selfie. http://sociology.about.com/od/Ask-a-Sociologist/fl/Why-We-Selfie.htm.

[9] J. Crook. Know thy selfie. http://techcrunch.com/2014/02/24/know-thy-selfie/, Feb 2014.

[10] R. Dey, M. Nangia, K. W. Ross, and Y. Liu. Estimating heights from photo collections: A data-driven approach. In *ACM Conference on Online Social Networks*, 2014.

[11] Ebay. Digital Vanity. http://deals.ebay.com/blog/the-selfie-revolution/, Oct 2013.

[12] J. D. Foster, W. K. Campbell, and J. M. Twenge. Individual differences in narcissism: Inflated self-views across the lifespan and around the world. *Journal of Research in Personality*, 37(6):469 – 486, 2003.

[13] J. Fox and M. C. Rooney. The dark triad and trait self-objectification as predictors of men's use and self-presentation behaviors on social networking sites. *Personality and Individual Differences*, 76(0):161 – 165, 2015.

[14] R. Garcia-Gavilanes, D. Quercia, and A. Jaimes. Cultural dimensions in twitter: Time, individualism and power. In *International AAAI Conference on Weblogs and Social Media*, 2013.

[15] P. Gonçalves, M. Araújo, F. Benevenuto, and M. Cha. Comparing and combining sentiment analysis methods. In *ACM Conference on Online Social Networks*, 2013.

[16] R. Hijmans. GADM database of Global Administrative Areas, version 2.0. http://www.gadm.org/, Jan 2012.

[17] N. Hochman and R. Schwartz. Visualizing instagram: Tracing cultural visual rhythms. In *International AAAI Conference on Weblogs and Social Media*, pages 6–9, 2012.

[18] G. Hofstede. Cultural dimensions in management and planning. *Asia Pacific Journal of Management*, 1(2):81–99, 1984.

[19] Y. Hu, L. Manikonda, and S. Kambhampati. What we instagram: A first analysis of instagram photo content and user types. In *International AAAI Conference on Weblogs and Social Media*, 2014.

[20] J. Y. Jang, K. Han, P. C. Shih, and D. Lee. Generation like: Comparative characteristics in instagram. In *ACM Conference on Human Factors in Computing Systems*, 2015.

[21] J. Joo, W. Li, F. Steen, and S.-C. Zhu. Visual persuasion: Inferring communicative intents of images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[22] J. Kilner. The science behind why we take selfies. http://www.bbc.com/news/blogs-magazine-monitor-25763704, Jan 2014.

[23] E. Losh. Beyond biometrics: Feminist media theory looks at selfiecity. Technical report, Softtware Studies Initiative, 2014.

[24] K. S. Lyness and M. K. Judiesch. Gender egalitarianism and work–life balance for managers: Multisource perspectives in 36 countries. *Applied Psychology*, 63(1):96–129, 2014.

[25] A. E. Marwick. Instafame: Luxury selfies in the attention economy. *Public Culture*, 27(1 75):137–160, 2015.

[26] M. McPherson, L. Smith-Lovin, and J. M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27(1):415–444, 2001.

[27] Megvii Inc. Face++ research toolkit. http://www.faceplusplus.com/, Dec 2013.

[28] E. P. Meier and J. Gray. Facebook photo activity associated with body image disturbance in adolescent girls. *Cyberpsychology, Behavior, and Social Networking*, 17(4):199–206, 2014.

[29] E. Pariser. *The Filter Bubble: What the Internet Is Hiding from You*. The Penguin Group, 2011.

[30] Pew Research Center. Social media update 2014. http://www.pewinternet.org/2015/01/09/social-media-update-2014/, Jan 2015.

[31] E. Sarigol, D. Garcia, and F. Schweitzer. Online privacy as a collective phenomenon. In *ACM Conference on Online Social Networks*, 2014.

[32] O. Schwarz. On friendship, boobs and the logic of the catalogue: Online self-portraits as a means for the exchange of capital. *Convergence: The International Journal of Research into New Media Technologies*, 16(2):163–183, 2010.

[33] T. H. Silva, P. O. S. V. de Melo, J. M. Almeida, J. Salles, and A. A. F. Loureiro. A picture of Instagram is worth more than a thousand words: Workload characterization and application. In *IEEE International Conference on Distributed Computing in Sensor Systems*, 2013.

[34] A. Tifentale and L. Manovich. Selfiecity: Exploring photography and self-fashioning in social media. Technical report, Softtware Studies Initiative, 2014.

[35] K. Tiidenberg. Bringing sexy back: Reclaiming the body aesthetic via self-shooting. *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, 8(1), 2014.

[36] J. M. Twenge and W. K. Campbell. *The Narcissism Epidemic: Living in the Age of Entitlement*. Free Press, April 2010.

[37] T. Vinson and M. Ericson. The social dimensions of happiness and life satisfaction of australians: Evidence from the world values survey. *International Journal of Social Welfare*, 23(3):240–253, 2014.

[38] C. Wilson. The selfiest cities in the world: Time's definitive ranking. http://time.com/selfies-cities-world-rankings/, Mar 2014.

[39] J. Winston. Photography in the age of facebook. *Intersect: The Stanford Journal of Science, Technology and Society*, 6(2), 2013.

[40] Y. C. Zhang, D. Ó. Séaghdha, D. Quercia, and T. Jambor. Auralist: Introducing serendipity into music recommendation. In *ACM International Conference on Web Search and Data Mining*, pages 13–22, 2012.