A Framework for Measuring Privacy Risks of YouTube

Vanessa Calero

June 26, 2020

MCMASTER UNIVERSITY

MASTER THESIS

A Framework for Measuring Privacy Risks of YouTube

Author:
Vanessa Calero

Supervisor:
Dr. Reza Samavi

A thesis submitted in fulfillment of the requirements for the degree of Master of Science (M.Sc.)

in the

Faculty of Engineering
Department of Computing and Software

June 26, 2020

Declaration of Authorship

I, Vanessa Calero, declare that this thesis titled, "A Framework for Measuring Privacy Risks of YouTube" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

MCMASTER UNIVERSITY

Abstract

Faculty of Engineering
Department of Computing and Software

Master of Science (M.Sc.)

by Vanessa Calero

While privacy risks associated with known social networks such as Facebook and Instagram are well studied, there is a limited investigation of privacy risks of YouTube videos, which are mainly uploaded by teenagers and young adults, called YouTubers. This research aims on quantifying the privacy risks of videos when sensitive information about the private life of a YouTuber is being shared publicly. We developed a privacy metric for YouTube videos called Privacy Exposure Index (PEI) extending the existing social networking privacy frameworks. To understand the factors moderating privacy behaviour of YouTubers, we conducted an extensive survey of about 100 YouTubers. We have also investigated how YouTube Subscribers and Viewers may desire to influence the privacy exposure of YouTubers through interactive commenting on Videos or using other parallels YouTubers' social networking channels. For this purpose, we conducted a second survey of about 2000 viewers. The results of these surveys demonstrate that YouTubers are concerned about their privacy. Nevertheless inconsistent to this concern they exhibit privacy exposing behaviour on their videos. In addition, we found YouTubers are being encouraged by their audience to continue disclosing more personal information on new contents. Finally, we empirically evaluated the soundness, consistency and applicability of PEI by analyzing 100 videos uploaded by 10 YouTubers over a period of two years.

Dedicated to the memory of my father Elias and my aunt Irma

Acknowledgements

First and foremost, I would like to thank God for everything and for giving me the strength, perseverance, and wisdom to continue my research during the most difficult times.

I would like to express my deepest appreciation to my supervisor, Prof. Reza Samavi for giving his invaluable guidance and time during all these years and especially for his support, company, friendship and good humor while I was writing this thesis. Thanks for being such an inspiration to me and other students that have had the opportunity of working with you.

I also thank my committee members, Prof. Rong Zheng and Prof. Wenbo He for providing excellent recommendations during my defense. My appreciation also extends to my research group colleagues, especially to Omar and Anna. Thank you so much for your early insights that helped me a lot.

My thanks are also delivered to all my friends (YouTubers and Subscribers) on YouTube for being so supportive and kind to me. Thanks for submitting the YouTube survey when I asked for and sending me positive and encouraging comments when I was feeling anxious. You guys are my online family and I am grateful to have all of you by my side.

Last, but definitely not least, I am greatly indebted to my family for the unconditional love, support, and prayers. I would like to thank my mom Lilian whose love and guidance are with me in every single aspect of my life. Most importantly, I wish to thank my loving and supportive partner, Angel. Thanks for being with me throughout these years and for being my best friend. You are my true and only love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

YouTube is one of the most popular video-sharing platforms. According to the online portal Statista [119], the number of online video platform *Viewers* will amount to 1.86 billion in 2021, up from 1.47 billion in 2017. In each minute 500 hours of video are uploaded to the platform and 100 million video views are reported. The total number of YouTube videos that are uploaded per hour is astonishing but more astonishing is the fact that many of these videos are revealing sensitive personal information of YouTubers (those who produce the contents of the videos).

During the last years, several YouTubers have uploaded YouTube videos where they share with the audience their negative experiences being a YouTuber [22]. Most of them have experienced stalking, cyber-bullying, impersonation among others due to the exposure of some sensitive information on videos. In 2017, a renowned Spanish speaking YouTuber German Garmendia [37] published a video where he discussed some privacy issues he experienced. He mentioned some privacy implications such as harassment and invasion of his privacy by a group of people who followed his YouTube channel and watched the video content. In this video, the YouTuber revealed how since 2013 he had seen the need to move repeatedly because of his home address had been exposed. Such circumstances emphasize the importance of understanding the privacy risks associated with each YouTube video prior to making it publicly available. This is the research problem that has motivated this thesis.

Understanding privacy implications of what users of social networking websites post, has received significant attentions from computer science research community (e.g.,[35], [75], [74], [29]). An important but not surprising finding of these studies have been the associations between types of information disclosed by the users and the privacy implications. Thus, a challenging aspect of social networking privacy research has focused on understanding the alignment of a user's perception of privacy with the actual privacy behaviour that the user exhibits when perform an activity on a social network [71][39][69]. The researchers have identified that computer scientist should focus on the user experience with a system and aspects of human computer interaction that goes beyond the user interface design to intrigue users to realize if their perception of privacy are aligned with their online behaviour [4]. Therefore, a number of proposals from computer science research community has focused on defining the relationship between an information type and its privacy risk and ultimately developing a method of measuring the subjective concept of privacy. For example, the authors in [8] proposed a privacy framework to quantify the privacy risk of the users when they disclosed sensitive information on their profiles. While all these proposal have focused on classical social networking platforms such

as Facebook and Twitter to evaluate the privacy framework [72] [60], the less interactive YouTube platform which exhibits some social networking characteristics have not been studied. Therefore, the goal of our research is to develop a method of measuring privacy risk on YouTube videos.

The purpose of this thesis is threefold. First, we aim to quantify the privacy exposure of YouTubers based on the type of sensitive information they are disclosing on YouTube videos. We define privacy exposure index (PEI) as a metric for quantifying the privacy risks of a YouTube video. Second, we are interested to empirically investigate the YouTubers' privacy behaviour and concerns along with the factors determining their privacy exposure behaviour. This empirical study also investigates whether or not YouTube viewers are influencing the privacy exposure behaviour of a YouTuber. Third, we aim to evaluate our systematic privacy scoring system (PEI) with YouTuber's perceptional view of privacy (PEI) and develop a set a of major privacy requirements that can speak to the design for privacy aligned with the privacy be design framework [23][111].

The research objectives of this research is formulated as the following two research questions:

RQ1. How privacy risk on YouTube videos can be quantified?

RQ2. How privacy behaviour and concerns of the YouTubers are aligned?

In answering the first research question, we develop a systematic scoring system for a YouTube video, by extending the privacy frameworks proposed for other social networking platforms such as Facebook, Twitter and LinkedIn. In answering the second research question, we conduct an extensive empirical study targeting two groups of YouTube users: YouTubers and Viewers. We will determine the factors affecting the privacy exposure behaviour of YouTubers as well as the consistency between the YouTuber's privacy behaviour and concerns. Likewise, we will evaluate the Viewer's influence on the privacy exposure of a YouTuber.

Finally, we will evaluate our proposed framework for measuring privacy risks of a video to determine the influence of the Viewers through the comment section over time. We also check the consistency of our computed privacy exposure index with the YouTubers' perceived privacy score of selective videos.

1.1 Thesis Contributions

This thesis makes contributions to the Human Computer Interaction (HCI) area of computer science, by advancing our understanding of the privacy determinants of YouTube videos and associated Youtubers' privacy behaviour when uploading a video to the YouTube platform. More specifically, we are making the following contributions:

1. We develop a privacy metric, Privacy Exposure Index (PEI), to quantify the privacy risks of a YouTube video. This metric is designed by extensively investigating the characteristics of the YouTube platform and the applicability of other privacy frameworks designed for similar social networking platforms.

2. With the approval of the McMaster University Research Ethics Board (MREB), we empirically study and report the privacy determinants of YouTube videos from a YouTuber's perspective along with the privacy concerns and behaviour of YouTubers when uploading a video.

3. We also empirically study and report the influence of the Viewers on the YouTuber's privacy exposure behaviour.

4. We publish an open dataset of meticulously indexed using our PEI for 100 videos from 10 YouTubers over two years. This is the first dataset published publicly and can be used by the research community to enhance privacy research on YouTube. The dataset is available at *https://github.com/samavi/YouTubePrivacy.git*.

5. We provide a set of design recommendations for the future automated tool that can help Youtubers make informed decision about their privacy prior to uploading a video.

1.2 Thesis Structure

The thesis structure is described as follows.

Chapter 2 presents the related research in three areas of (1) privacy behavior and concerns of online users, (2) privacy risks and threats in social networks and the frameworks to measure privacy risks in social networks, and (3) the approaches used to quantify privacy risks.

Chapter 3 is dedicated to answering our first research question. We describe the characteristics of YouTube as a social network platform and specify the YouTube privacy risks. We then extensively investigate previously designed methods of quantifying privacy frameworks and develop our metric of quantifying privacy risks of a Youtube video called privacy exposure index (PEI).

Chapter 4 is dedicated to our second research question and reports on two empirical studies that we have conducted to understand privacy concerns and behaviour of YouTube users. The study is conducted as an online survey targeted two groups of YouTube users: YouTubers and Viewers. The study provides answers to the following questions:

1. What are the factors moderating YouTubers privacy behaviour?

2. Are YouTubers' privacy exposure behaviour and concerns consistent in uploading YouTube videos?

3. Do viewers have an internal tendency to influence YouTubers' privacy exposure?

Chapter 5 reports the practicality of the proposed framework for measuring the privacy risks of a YouTube video. We analyze 100 YouTube videos using our developed privacy metric, PEI. These videos are selected from 10 YouTube channels and from 10 known YouTubers with a large viewers base (more than 1 million subscribers) over a period of two years. This study will demonstrate the functionality

of the proposed framework and determine the relationship between the accumulation rate of the privacy exposure index (PEI) and the comments section over time. Finally, the study intends to check the consistency between the YouTuber's privacy perception and the measured PEI using our proposed privacy framework by setting up an experiment where a group homogeneous of YouTubers measure the privacy risk of a group of YouTube videos from our 100 YouTube videos dataset.

Chapter 6 presents the conclusions of this thesis. We also discuss how the proposed framework could be used in future research to predict privacy risks of YouTube videos.

Chapter 2

Related Work

In this chapter, we present the related work in three areas: (i) privacy requirements and human computer interaction (HCI); (ii) privacy risks in online social networks; (iii) existing methods for quantifying privacy risks in online social networks. These three areas are detailed in Sections 2.1 , 2.2, and 2.3, respectively.

2.1 Privacy Requirements and Human-Computer Interaction

Privacy is one of the main concern of the users and practitioners in the digital era. The leakage of personal information had led to risks and threats to the privacy and security of individuals. When a person is interacting with the computer through simple actions like sharing a photo or leaving a comment on a tweet is leaving a trace in his/her digital privacy [3, 47]. Digital privacy is relatively new concept and it refers to the information that a user exposes on the internet, this information could contain personal data like address, political view or sexual preferences and could be in the form of texts, photos, videos, etc. Previous studies [48, 20] have exhibited results regarding self-disclosure using computer-mediated communication, where the users over-share information about themselves resulting in concerns about privacy and security. Thus, researchers are constantly improving the protection of individual's personal information due to the fact that *most privacy threats and vulnerabilities originate from the interaction between the individuals using information systems rather than the actual systems themselves* [43].

In [32, 33], the authors identify four ways to preserve privacy: protection by law, protection by privacy-enhancing technologies, self-regulation for fair information practise, and privacy education of consumers and IT professionals. In order to protect privacy in computing environments, privacy-enhancing technologies (PET) have been proposed. PETs are built on the foundation of a number of digital privacy principles such as informed consent, encryption, data minimization, data tracking, anonymity, and control [93]. In this context, many studies have shown the differences between the privacy behaviour and privacy preferences of the users [14, 4, 16] as a problem in analyzing the privacy techniques that help users protect their privacy. In [5], the author reveals that few people actually take any action to protect their personal information, these behavior of individuals and attitudes cause privacy risks and vulnerabilities.

Human-computer interaction (HCI) is the subfield of computer science that studies the interaction of the people with the new computational technologies, as an interdisciplinary area, HCI helps to create and design privacy mechanisms by understanding individual's need, attitudes and behavior when they interact with the

computer [1].

Understanding online privacy behavior, attitudes, and concerns are necessary for developing good privacy-preserving techniques [98]. In [99], the author mentions the relationship between privacy and other factors that determine the online behaviors of the users. Also in [88, 87] authors investigate the online privacy behaviour and concerns of Personal Health Record (PHR) users. This specific research inspired us to follow the same methodology to study the privacy behaviour of YouTubers, because similarly, the privacy study of PHR users entailed to developing a generic privacy framework to address transparency and accountability of multi parties involved in a PHR context [85, 86].

2.2 Privacy Risks in Social Networks

Online social networks unlike e-commerce and other online environments are different in terms of user behavior and information disclosure [46, 13]. Previous studies have identified that certain types of information when it is disclosed in online social networks can cause potential risks or threats [123]. In [9], the authors present the privacy and security threats associated with online social networks and also discuss the factors behind these privacy threats. They categorized the threats in social networks in four groups:

1. Privacy-related threats such as digital dossier of personal information, face recognition, difficulty with complete account deletion, content-based image retrieval or image tagging and cross-profiling.

2. Information security threats such as spamming, cross site scripting and social network sites aggregators.

3. Identity-related threats such as phishing, information leakage, identity theft.

4. Social threats such as stalking and corporate espionage.

The threats of social networks impact user's privacy and have associated with the personal information disclosure. Hence it is necessary to identify and quantify the personal information considered as sensitive before a user discloses them on their social networks to prevent privacy risks. For example, it has been observed that 87% of the people can be uniquely identified based on their date of birth, gender and zip code [106].

Gross and Acquisti evaluated in their Facebook study the amount of sensitive [38] information that most people disclose on their social networks. In this research the authors presented the relationship between privacy implications and the identifiable information provided by the users in their profiles by analyzing the SNS's of a group of undergraduate, graduate students and staff of the Carnegie Mellon University (CMU). The study gathered 4540 Facebook profiles, the results showed that users provided fully identifiable names, location information, birth dates, phone numbers among others. The authors highlighted the privacy implications associated with the level of identifiability of the information that the users are disclosing on their online social networks. Furthermore, the authors measured the number of Facebook users susceptible to these attacks based on the information divulged in their social networks sites.

According to [15], in online social networks, the control over personal information is negatively associated (in a statistical term) with information disclosure. For example, an individual is disclosing more information in his/her social networks when perceive they have more control over information [2, 50, 82]. Another interesting finding related to self-disclosure in online social networks [52], showed that the privacy concerns of the users of social networks are primarily determined by the perceived likelihood of a threat rather than the expected damage, users behavior in online social networks is inconsistent with their privacy concerns, for instance, users disclose more personal data in their online social interactions, even though they are aware of the privacy threats associated with [45] [87]

A survey that gathers in detail the works on privacy in social networks is presented by E. Zheleva [125]. In their extensive work they described a set of recent techniques for modelling, evaluating and managing privacy risk within the context of Online Social Networks. Some of these papers include studies on how to detect and report unintended information loss in social networks. However, in order to protect the privacy of the users it is necessary to have a metric to measure the privacy risk based on the information disclosure [12].

These previous works have focused on determining the type of information disclosed on traditional online social networks. Therefore, there is a need for a study focused specifically on YouTube that allows researchers to understand what types of information considered sensitive when they are disclosed on YouTube videos and what are the threats associated with those sensitive information. In addition it is necessary to understand how the Youtubers' privacy behaviour can be impacted by the these types of metrics.

2.3 Existing Privacy Risk Scores Frameworks

With the increasing expansion of activities in online social networks, developing a method of quantifying online privacy risk has been a major challenge for information security researchers. Previous studies have proposed frameworks for computing the privacy scores of a user in a social network based on the amount of sensitive information that a user is willing to disclose [62, 28, 68, 60].

Liu et al. [61] are the first to propose the concept of *privacy score* as a quantification of the privacy risk of a user in an online social network. They state that the *privacy score* of a user depends on the type of sensitive information that the user is disclosing in online social networks and the number of users to whom the information are being shared. Authors in [60] are the first to provide a methodology for computing the privacy risk score of a user by introducing the concepts of *S*ensitivity and *V*isibility. According to authors, the privacy score has two premises: *The more sensitive information a user reveals, the higher his or her privacy score is"*, and *"the more people know some piece of information about a user, the higher his or her privacy score is."*

Therefore, the definition of *privacy score* satisfies the following premises:

1. The *privacy scores* vary according to the sensitivity of the information disclosed. Thus, a user's privacy score increases as more sensitive information about the user is revealed in his/her online social network.

2. The *privacy scores* vary according with the visibility of the information disclosed. Thus, the larger number of people see the disclosed information, the higher the privacy risk score of the user is.

The privacy risk score framework assumes that all n users $j\epsilon\ \{1, ..., n\}$ specify in their privacy settings for the same n profile items. These privacy settings are stored in an $n \times n$ response matrix $\mathbf{R}$. This response matrix represents the profile privacy settings (e.g., address, phone number) of a user j. Value $\mathbf{R}(i, j)$ is a value that denotes how willing the user j is to disclose information about item i, the higher the value $\mathbf{R}(i, j)$, the more willing j is to disclose information about item i. Thus, large values in $\mathbf{R}$ imply higher visibility. On the other hand, small values in the privacy settings of an item i are an indication of high sensitivity.

According to the authors, the sensitivity of an item i is denoted as β_i and the visibility $V_{(i,j)}$ of the same item i due to user j. Therefore, the privacy score of a user j for the item i can be any combination of sensitivity β_i and visibility $V_{(i,j)}$. We extend the concept of sensitivity and visibility described here for quantifying privacy scores of Youtube videos.

In [8], the authors extended the previous work by proposing an approach that helps users to measure their privacy disclosure score (PDS) based on the information shared across multiple online social networks. Online social network users generally have multiple social network accounts (e.g., Facebook, Twitter, and Instagram) for different purposes and in each social network they will be disclosing their personal information. Thus, the authors proposed a scoring function to quantify the privacy risks of a user where the inputs of this scoring function are the personal attributes of a user disclosed on multiple social networking platforms. For measuring the PDS, users' attributes (e.g., phone number, email, address, job details, hobbies, and interests) are considered to be obtained from n different sources (online social networks); then the factors (sensitivity and visibility) are calculated in order to measure the PDS. F_{sen} is the sensitivity function and indicates the sensitivity of each attribute of the user. On the other hand, the visibility depends on other three factors known as accessibility to information F_{acc}, difficulty of data extraction of users' information F_{dif} and the data reliability for each attribute F_{rel}, Therefore, to compute the privacy score of a user, the combined sensitivity score and visibility score of the user for several attributes (e.g., name, age, gender, email, hometown, job details, and interest) from different data sources are considered.

The work presented in [101] is an interesting approach to the previous research [61] regarding the calculation of privacy risk score. The authors examine specific information (text message) and extract the sensitive information (e.g., address, location, etc.) using pattern detection in textual data. The data is classified as sensitive or non-sensitive by means of a naive binary classifier. According to the authors, a (text) message may contain sensitive information about the user, the sensitive part of a message is called *item*. They calculate the privacy score using the concepts of sensitivity and visibility previously presented.

While all these previous research proposals have focused on classical social networks e.g., Facebook, Twitter, LinkedIn, on best of knowledge, the privacy risks of non-classical social networks such as YouTube have not been studied. YouTube, as we described in Chapter 1 is also considered as a Social Network but contrary to

the other classical social networks, the user's *"profile items"* cannot be determined he same way. Thus in Chapter 3, we will propose the factors considered as features for determining privacy risk score on YouTube videos and their sensitivity and visibility.

Chapter 3

YouTube Privacy Risk Factors

In this chapter, we will first describe the YouTube platform in Section 3.1, its privacy aspects, and the privacy implications of a YouTube video by analyzing the YouTube users, their roles in the platform, and the characteristics of a YouTube channel. We will then analyze other existing frameworks for calculating privacy score on traditional social networks in Section 3.2. Finally, in Section 3.3, we adapt and extend prior privacy scoring system developed for other social networking platform to develop a privacy exposure index for measuring the privacy risk of an individual video and the accumulative privacy exposure index across the timeline of a YouTuber content generation.

3.1 YouTube Platform

YouTube was founded by Chad Hurley, Steve Chen, and Jawed Karim in February 2005 [121]. Most people do not recognize YouTube as a social network however YouTube is considered as a social network platform, and in fact as one of the most popular and largest ones [119]. YouTube users could like or dislikes videos that they watch and also they can share or comment on a video, and like other social networks, YouTube makes the content recommendations through YouTube user behavior just as Facebook or Twitter recommends friends or contents. In the same way as other social networks, using YouTube has privacy implications and vulnerabilities associated with the YouTube videos [53].

Understanding what the privacy risks are and what are the determinants of privacy for an uploaded video is an important problem that needs attention by the research community. As the first step is important to study the process of uploading a video as there are some terms and definitions commonly used among YouTube users that are important to be known.

3.1.1 YouTube Users and Definitions

In previous research [64], social network users are categorized into five groups based on their role on the platform: Small Community Member, Content Producer, Content Consumer, Producer/Consumer and Other. With the intention of introducing concepts associated with YouTube, these groups are mapped according to [122], and thus producing three main YouTube users: YouTuber, Subscriber, and Viewer. Below we provide the definitions commonly associated with these users.

- *YouTuber*: It refers to a YouTube user who creates content for YouTube also known as **Content Producer** or **YouTube Producer**. As the main role of the YouTuber is to create content and publish videos, they are more vulnerable to privacy risk than other YouTube users when the content is focused on topics related to their personal life.
- *Viewer*: It refers to a YouTube user who consumes the content of a YouTube channel but may or may not be subscribed to the channel. Some YouTube users are not willing to subscribe to a YouTube channel.
- *Subscriber*: A subscriber also known as **Content Consumer** or **YouTube Consumer** is a YouTube user who is subscribed to a YouTube Channel. In a sense, a subscriber is a Viewer that can comment or share the content and stay updated with the latest videos on a YouTube channel. According to YouTube, Subscribers have been shown to watch more videos than non-subscribers.
- *Audience*: It refers to a group of YouTube users that have a YouTube channel in common, it is also known as *Community*.
- *Content*: It refers to a YouTube Video.
- *Vlogger*: It refers to a YouTuber who creates videos focused on the YouTuber's daily life.
- *Beauty Guru*: It refers to a YouTuber who creates videos focused on makeup, skincare, and beauty.
- *Gamer*: It refers to a YouTuber who creates content focused on video games. This type of YouTubers record their screen while playing a video.
- *Vlog*: This term refers to a type of YouTube video, a *Vlog* depicts daily moments from a first-person perspective of the YouTuber.

3.1.2 YouTube Channel Categories

When a YouTube user creates his/her own channel, a YouTube category is enlisted with several channel categories to choose from. Each YouTube channel has associated one specific category that represents the type of content of the channel. The top categories watched by YouTube users are Entertainment, People&Blogs and Gaming [105] as indicated in Table 3.1. In our study, we aim to determine if the YouTube channel category is a factor affecting the exposure of personal information on a video. In addition, by default, the category of a video is the same as the channel category. However, there is an option to choose a different category for a video in particular.

TABLE 3.1: List of the YouTube Channels Category

Categories
Autos and Vehicles
Comedy
Education
Entertainment
Film and Animation
Gaming
Howto and Style
Music
News and Politics
Nonprofits and Activism
People and Blogs
Pets and Animals
Science and Technology
Sports
Travels and Events

3.1.3 YouTube Video Privacy Settings

YouTube has three different privacy settings to make a YouTube video *public, private, or unlisted*. When a video is being uploaded to YouTube, there is an option that allows users to manage the privacy settings of the video, YouTube recommends choosing the option "unlisted" while the uploading process is being completed. After the uploading process is completed, a YouTuber can make the video public. In fact, the YouTube video privacy settings can be updated at any time even after posting a video.

If a video is uploaded as private, then it can only be available for the YouTuber that has uploaded the video. In this case, the video will not be shown to any other user, neither through the feed nor via URL. It is also not possible to interact with the video in any way (e.g., comments or likes) when it is *private*. The next privacy setting is unlisted, videos uploaded as unlisted can be seen, comment on, and shared by any user via URL. Finally, a video is uploaded by default as public, public videos can be seen by, comment on and shared with anyone [120].

As we can see, the privacy settings of YouTube videos are quite crude to the level of a public/private dichotomy. These video privacy settings cannot provide sufficient support for the privacy implications that a video might be associated with, as expressed in the next section.

3.1.4 YouTube Privacy Implications

In general, online social networks exhibit and increasing risks of violating digital privacy. People are increasingly willing to share personal information, their favorite places to go, and politic preferences. This type of information disclosure carries some risk to the privacy and security. Ho et al. [40] classify privacy risks as security risks, reputation&credibility risks, and profiling risks. We have analyzed all these

risks on YouTube Channels. When a person uploads a video on YouTube, the privacy risks are currently not analyzed by default on the platform.

Particularly, sensitive population such as adolescents and young adults are exposing their privacy, their daily routines, places they frequently visit and even where they go on vacation. Having all this information publicly available and by assuming some simple routines, an adversary could infer what a YouTuber will do next week, or whether a YouTuber will be alone or accompanied.

We divide such privacy risks into the categories shown in the Table 3.2. The first category contains *security risks* which includes *stalking*, *kidnapping and robbery* and *cyberbullying*. The second category contains *profiling risks*. This category includes *extortion and ransomware attacks*. The third category contains *reputation and credibility risks* which includes *perceptive discrimination* and *insider threat*. The last is *health risks* which includes *perspective health*.

TABLE 3.2: Privacy Risks in YouTube videos

Privacy Risk Category	
Security Risks	Stalking, Kidnapping and Robbery, Cyberbullying
Profiling Risks	Extortion and Ransomware
Reputation and Credibility Risks	Perceptive Discrimination, Insider Threat
Health Risk	Perspective Health

Security Risk

The security risks that a video might present due to a large amount of information disclosed are *stalking*, *kidnapping and robbery* and *cyberbullying*. All the aforementioned privacy implications pose a threat to the YouTuber that we will explain below.

- *Stalking*: According to Hasib [9] with all the information exposed on video, a person can be stalked easily. By analyzing a particular vlog account on YouTube for two months, we noticed a potential security and latent privacy risk. With all the details of their lives that YouTubers are willing to reveal to their audience, a stalker could put together a pattern of behavior. For example, if a YouTuber in one of their videos or vlogs reveals their favorite place to go out with their friends every month and then in their Instagram or Facebook accounts this aforementioned place is tagged in a photo uploaded by themselves, a stalker would be able to wait until next month and start stalking them.

- *Kidnapping and Robbery*: In videos titled Room Tour or House Tour [31], YouTubers show their houses and all the details of where they live. Some YouTubers share personal addresses in the description box of their videos. This practice usually occurs to allow subscribers to send letters and gifts. The risk involved is not only privacy but security, because combining these two information, someone with bad intentions could plan a robbery or kidnapping. Knowing how the house looks inside and where that house is, the intruder gains advantage over their victims. In 2016, the famous reality-TV star Kim Kardashian was assaulted at her hotel in Paris. In an interview conducted months later, she confessed to posting about jewelries she was carrying on her social networks, and saying where she would be. This made her an easy target [83].

- *Cyberbullying*: Several studies have analyzed various forms of harassment involving social networks. *cyberbullying* is a problem that has gaining attention in recent years [73]. According to Valkenburg [113] there is a relation between the consequences of adolescents' use of friend networking sites and their social self-esteem and well-being. From the results of our YouTube videos analysis explained in Chapter 5, there is a considerable number of negative/malicious comments on a YouTube video. It is necessary to emphasize that YouTube provides the option to delete any comments and even allows to put filters for certain words, that is to say, the YouTubers can block some words in the YouTube channel settings in order to avoid insults or malicious comments.

Profiling Risk

A profile risk such as *extortion and ransomware* is associated when the information disclosed in a video details the behavior of a person.

- *Extortion and Ransomware*: This type of risk in privacy aims to obtain money, property or services from another person through threat or coercion as is the case of Amy Palumbo Miss New Jersey who was extorted with the publication of photos of her social network if she would not given up her crown [77]. By uploading daily videos exposing their life publicly, YouTuber unintentionally can provide certain information through the videos that could be used to determine confidential information, such as address, phone number, name of friends and family. In 2017, a YouTuber from the Dominican Republic with a large audience of more than one million followers shared a video titled "A subscriber harasses me" [84]. In this video she mentions that a subscriber was able to obtain her phone number and work address of her mother. The subscriber wrote her until dawn and later threaten her to make public this information to other subscribers. Although, the video does not explicitly detail the situation as extortion, such a risk is latent and could become common practice on the platform.

Reputation and Credibility Risk

Reputation and credibility risks refer to situations where there is a leak of information that could be exploited by a third party [124]. For example, sharing personal or other people's videos or photos within a social network.

- *Perceptive Discrimination*: This privacy risk examines the discrimination based on a perception of the information. In general, these types of exposures can cause a social and/or career impact on the YouTuber. An example that explains this situation is observed in [36]. In this video, the Youtuber decides to share an event that directly or indirectly involves drug use. According to how this topic is perceived, repercussions can range from family issues to loss of professional confidence in certain cases.

- *Insider Threat*: This is a privacy risk that affects an organization and is caused by individuals whether or not they are associated with the organization. This type of exposure can cause sabotage, theft and fraud, and is one of the biggest problems in cyber and corporate security [80]. On YouTube, this threat is primarily observed with video vlogs. For example, if a YouTuber records a video in a place where access is not authorized to personnel external to the organization or restricted to certain people, simply uploading a video showing the

place carries privacy risks associated with insider threat. In this example, the risk is not associated with the YouTuber but the company or third parties that are involved in the video.

Health Risk

According to an study on the privacy risks in the context of Health Social Networking Sites (HSNS) [59]. It was found that health data could potentially be misused by insurers or prospective employers to deny you policies or employment. In this way, health is also analyzed as a criterion to determine the privacy risks of a YouTube video. As we discussed in Chapter 4, YouTubers tend to be more open to sharing their life and intimacy when an increase of the frequency of use or engagement on the channel is observed (e.g., higher content publication or number of views). As the channel earns subscribers, the audience encourage YouTubers to be more open about their personal life and information. Consequently, YouTubers feel confident enough to reveal more information about their health and other aspects of their personal lives.

3.2 A Survey of Existing Privacy Risk Computation Methods

In this section, we analyze how other existing approaches have proposed calculating privacy scores on traditional social network site. Then, we establish the methodology used for scoring YouTube videos.

Many researchers have proposed frameworks for measuring privacy risk scores of a user in a social network site based on the amount of sensitive information that users are willing to disclose [61, 28, 68, 60]. Authors in [60, 61] are the first to provide a methodology for computing the privacy risk score of a user following two premises: *The more sensitive information a user reveals, the higher his or her privacy score"* and *"The more people know some piece of information about a user, the higher his or her privacy score"*.

In [60], the authors present a framework for computing the privacy score that combines the partial privacy score of each user in a social network. Each user j has a set of associated profile items i (e.g., user's phone number, real name, relationship status, etc). This information is given as input to the framework as a $n \times N$ dichotomous response matrix $\mathbf{R}$ that stores the privacy level of all N users for all n profile items. $\mathbf{R}_{(i,j)}$ refer to the privacy setting of a user j for an item i.

The entries of $\mathbf{R}$ take integer values in [0,1], which means that if $\mathbf{R}_{(i,j)} = 0$ the user j has made the profile item i private. On the other hand, if $\mathbf{R}_{(i,j)} = 1$, then the user j has made the profile item i public. From the response matrix, the authors defined the users' privacy settings of each profile item as $\mathbf{R_i}$ and the profile's privacy settings of each user as $\mathbf{R^j}$.

In this way, the privacy score model proposed by the authors used the response matrix $\mathbf{R}$ to compute a monotonically increasing function of two parameters; the *sensitivity* of the profile items β_i, and the *visibility* these items get $V_{(i,j)}$. The definitions for each parameter are given below.

Sensitivity of the profile item: β_i is a value that depends on the type of disclosed information called *profile items* (e.g., name, gender, phone number), where some profile items are more sensitive than others. For example *address* is considered more sensitive than *gender* due to privacy level of each of the information. In other words, the sensitivity is a characteristic of each profile item i of a user j as is mentioned in [72].

Visibility of the profile item: $V_{(i,j)}$ is the visibility of a profile item i of a user j and depends on the value of $\mathbf{R}_{(i,j)}$ as well as on the particular user j. In practice, there are two types of visibility: the *observed visibility* and the *true visibility*. The observed visibility is computed as shown in Equation 3.1, which simply uses an indicator variable that assigns 1 when the given condition is met. On the other hand, the *true visibility* or simply visibility is computed as shown in Equation 3.2. The visibility considers the users' setting as a random variable of a probability distribution $P_{ij} = \text{Prob}\{\mathbf{R}_{(i,j)} = 1\}$, the latter denotes the probability that a user j select $\mathbf{R}_{(i,j)} = 1$ (i.e., user j has made the item i publicly available).

$$V_{(i,j)} = \mathbf{I}_{ij(\mathbf{R}_{(i,j)}=1)} \tag{3.1}$$

$$V_{(i,j)} = P_{ij} \times 1 + (1 - P_{ij}) \times 0 = P_{ij} \tag{3.2}$$

Thus, the authors proposed a model for calculating privacy score based on the Sensitivity and Visibility. The privacy score is calculated using the Equation 3.3. With this formulation, the privacy risk score can be directly calculated when the values of sensitivity and visibility are specified. For example, the sensitivity can be assigned according to particular domain knowledge of an expert, while the visibility through direct access to the profile settings of the users. On the other hand, in case these parameters cannot be specified directly, it is necessary to use parameter estimation methods to calculate the privacy score, as described in [72, 101, 61].

$$PR_{(j)} = \sum_{i=1}^{n} PR_{(i,j)} = \sum_{i=1}^{n} \beta_i \times V_{(i,j)} \tag{3.3}$$

where:

i	= profile item (e.g., name, gender, birthday)
j	= user
β_i	= sensitivity of profile item i
$V_{(i,j)}$	= visibility of profile item i of user j

All those previous research work focused on measuring privacy risk on traditional social networks. We adapt these approaches to quantify privacy risk in a non-traditional social network like YouTube.

3.3 Privacy Exposure Index (PEI)

Inspired by the prior work outlined in Section 3.2, we define privacy exposure index (PEI) as a metric that quantifies the privacy risk of an individual YouTube video. Equation 4.1 formally describes PEI:

$$PEI = \sum_{k=1}^{3} \sum_{i=1}^{n} \sum_{j=1}^{l_i} m_{kij} w_{ij} V(f_{ij}) \tag{3.4}$$

where:

k = modality index
i = feature type index
j = feature subtype index
n = number of feature types (n=6)
l_i = number of features subtype for the feature type i
m_{kij} = modality k of the feature type i and subtype j
w_{ij} = weight of the feature type i and subtype j
f_{ij} = feature type i and subtype j
$V(f_{ij})$ = visibility of the feature f_{ij}

While we use the concepts of Visibility $V_{(i,j)}$ and Sensitivity β_i described in [61] to compute PEI, there are two specific aspects of privacy exposure in a YouTube video, modality m_{kij} and weigth w_{ij} that need to be considered.

An important consideration in formulating PEI is that different from prior work, in a YouTube video, private features could be disclosed in different ways: (1) throughout the YouTube video where the features could be disclosed from what has been viewed (2) throughout the audio of the YouTube where the features could be disclosed from what has been heard (3) throughout the description of the YouTube video like the title or the description box. Thus, we introduce a new term in our formulation called modality for the measurement of the PEI discussed below.

Modality: Content of a video can be communicated using different modalities: video, audio, and metadata.

The modality m_{kij} refers to the way sensitive information is being disclosed. There are three ways of disclosing sensitive information on a YouTube video that is throughout the video-content, audio-content, or the metadata. For example, if a YouTube video is disclosing the social security number in the description box of the video, then the modality is metadata.

The modality of exposing content-sensitive features could be video, audio, or metadata (correspond to k=1, k=2, k=3 respectively) as we explain below.

1. *video*: It refers to the visual content of a YouTube video. The features disclosed in this portion of the YouTube video are called visual features.

2. *audio*: It refers to the audio content of a YouTube video. The features disclosed in this part of the YouTube content are called audio features.

3. *metadata*: It refers to the part of a YouTube video that contains the information of the video such as Title, Description and Thumbnail. The title of a YouTube video is one of the most important elements of metadata because it explains what the video is about. Description and Thumbnail provide additional context of a video.

Weight: The weight w_{ij} is a value that represents the sensitivity of a content-sensitive feature as is detailed in Section 3.3.1.

The classical sensitive β_i studied in previous work was not defined to include implicit disclosure information. Therefore, we propose to use the concept weight w_{ij} due to the nature of YouTube platform and hence differentiate the concept sensitivity β_i from previous work.

Given the newly added terms to the equation o PEI, the sensitivity and visibility of videos can be defined as follows:

Visibility: The visibility determines if a feature is visible in the YouTube video. The possible values of the visibility of a feature are 0 or 1, 0 if the feature is not visible and 1 if the video is visible.

A YouTube video may contain several sensitive information throughout the video. This type of sensitive information have been called profile items [60] or items [101] in previous studies. Therefore, as part of our approach to measure the PEI of a YouTube video, we group all these items into six groups that we called *content-sensitive features*.

3.3.1 Content-Sensitive Features

Given the above definition of sensitivity, we can define a content-sensitive feature f_{ij} or also referred to as feature is any sensitive information that affects the PEI of a YouTube video. Each feature has a type i, subtype j, and weight w_{ij} as it is described in Section 3.3.1. The type i and subtype j identify the feature, and the weight w_{ij} is a value that indicates the sensitivity of the feature f_{ij}.

In our framework, a content-sensitive feature is any information that contains privacy implications, some features by nature are more sensitive than others, thus each feature have a specific weight value between [1-5] that represents the sensitivity of the feature.

The three main types of features are listed below:

1. *Personal Identifying features:* It is any feature that could potentially be used to identify a particular YouTuber. In Section 3.3.2, we discuss more about the attributes of this category.
2. *Location features:* It is any feature that could expose the YouTuber location. In Section 3.3.3 we provide the list of the features of this type.
3. *Personal Health features:* It is any information related to the YouTuber's health. In Section 3.3.4 we discuss this type of sensitive information as well as the other types of information.

3.3.2 Personal Identifying Features

Definition: Personal identifying feature (PIF) is any information that contains personal data relating to identifying a YouTuber.

According to [107] a person could be identified by three attributes gender, zip code and full date of birth. From YouTube perspective, we could say that a YouTuber is fully identified because they share their names, age, and some personal information. However, we focus on other pieces of information related to personal identifying features that could be protected from others such as social security numbers, drivers license, identity number, etc.

Personal Identifying Features				
Feature	Description	Modalities		
		Video	Audio	Metadata
f_{11}	Full Name	1	1	1
f_{12}	Age	1	1	1
f_{13}	Birthday	2	2	2
f_{14}	Social Security Numbers	5	5	5
f_{15}	Passport Information	5	2	1
f_{16}	Drivers License	5	4	1
f_{17}	Phone	3	1	5
f_{18}	Zip Code	5	4	5
f_{19}	Identity Number	5	5	5
f_{110}	Place of Birth	2	1	1
f_{111}	Hobbies/Interests	1	1	1

TABLE 3.3: Weight Values Related to Personal Identifying Features

Some examples of YouTubers disclosing PIF are:

- YouTuber uploads a video and in the description box of the video he/she mentions his/her birthday. (e.g., if f_{13}=Birthday, then w_{13}=2 when m_{313}=1).
- YouTuber uploads a Vlog where accidentally shows his/her passport (e.g., if f_{15}=Passport Information, then w_{15}=5 when m_{115}=1).
- YouTuber says in a video the place when he/she was born (e.g., if f_{110}=Place of Birth, then w_{110}=1 when m_{2110}=1).

3.3.3 Location Features

Definition: Location feature (LF) is any information that could be used to infer the YouTuber's location.

Sharing location information on YouTube videos raises privacy concerns. According to [9] [95] with all the location and personal information exposed on social network sites, a person can be stalked easily. Typically, the items most commonly disclosed on YouTube videos are home address, building numbers, street name signs and places where the YouTuber usually goes. Table 3.4 provides a list of the items related to location information and how these items can be scored on a YouTube video.

Location Features				
Feature	Description	Modalities		
		Video	Audio	Metadata
f_{21}	Street Signs	3	1	1
f_{22}	Landmarks	2	1	1
f_{23}	Home Address	5	5	5
f_{24}	City	1	1	1
f_{25}	Buildings	2	1	1
f_{26}	Traffic Light	1	1	1
f_{27}	License Plate	4	1	4
f_{28}	Agenda Info	3	3	3

TABLE 3.4: Weight Values Related to Location Features

Some examples of YouTubers exposing LF are:

- YouTuber discloses the home address in the description box (metadata) with the intention of receiving letters from their subscribers (e.g., if f_{23}=Home Address, then w_{23}=5 when m_{323}=1).
- YouTuber uploads a Vlog where he/she shows license plates of vehicles (video-data) that are parked around his/her home (e.g., if f_{27}= License Plate, then w_{27}=4 when m_{127}=1).
- YouTuber Alice shows street signs around his/her neighborhood (e.g., if f_{21}= Street Signs , then w_{21}=3 when m_{121}=1).
- YouTuber fully says his home address (e.g., if f_{23}= Home Address , then w_{23}=5 when m_{123}=1).

3.3.4 Personal Health Features

Definition: Personal health feature (PHF) is any information related to the medical information, mental health conditions, insurance information and other information to identify a YouTuber's health conditions.

According to [59] health data could potentially be misused by insurers or prospective employers to deny you policies or employment. In Table 3.5 we list the items related to PHF.

The most common titles of the videos that contains mental or health information are: *Dealing with Panic Attacks Anxiety, How To Deal With Anxiety, Where I have been, My mental health struggles, my mental health story, my mental illness and my eating disorder story*. We found that many Vloggers and Beauty Gurus frequently discuss with their audience about their mental health issues.

Some examples of YouTubers exposing PHF are:

- YouTuber shares in a talk video her medical details or health information (e.g., if f_{32}=Health Conditions, then w_{32}=4 when m_{232}=1).
- YouTuber uploads a video where she discloses her mental health and she openly talks about her diagnosis) (e.g., if f_{34}=Mental Health, then w_{34}=4 when m_{234}=1).

Personal Health Features				
Feature	Description	Modalities		
		Video	Audio	Metadata
f_{31}	Medical Insurance	3	4	2
f_{32}	Health Conditions	4	4	2
f_{33}	Medical History	1	2	2
f_{34}	Mental Health	4	4	2
f_{35}	Family Medical History	1	2	1
f_{36}	Genetic History	1	2	1

TABLE 3.5: Weight Values Related to Personal Health Features

3.3.5 Personal Finance Features

Definition: Personal finance feature (PFF) is any personal information related to YouTuber's wealth.

The fourth feature is known as PFF, revealing this type of information could have privacy implications for YouTubers such as Kidnapping and Robbery as we discussed in Section 3.1.4. In general, disclosing information related to our finances always represent a risk. The type of information that represents a PFF is show in Table 3.6.

Personal Finance Features				
Feature	Description	Modalities		
		Video	Audio	Metadata
f_{41}	Credit Card	5	2	2
f_{42}	Credit Record	3	3	1
f_{43}	Loan Records	3	3	1
f_{44}	Incomes	2	2	1
f_{45}	Expenses	1	2	1

TABLE 3.6: Weight Values Related to Personal Finance Features

Some examples of YouTubers exposing PFF are:

- YouTuber discloses salary information on a YouTube video while he/she is talking with the audience (e.g., if f_{44}=Incomes, then w_{44}=2 when m_{244}=1).
- YouTuber shows his credit card on a YouTube video (e.g., if f_{41}= Credit Card, then w_{41}=2 when m_{141}=1).

3.3.6 School and Job Features

Definition: School and job feature (SJF) is any academic information or personal information related to the YouTuber's education or job.

Table 3.7 shows sensitive information that are related to school and job features.

Some examples of YouTubers exposing SJF are:

- YouTuber mentions the university he/she attends (e.g., if f_{52}=School Name, then w_{52}=2 when m_{252}=1).
- YouTuber discloses information about one course taken for the current term in the description box (e.g., if f_{55}=Courses Information, then w_{55}=1 when m_{355}=1).
- YouTuber uploads a recording of events in a classroom (e.g., if f_{53}=School Information, then w_{53}=1 when m_{153}=1).
- YouTuber reveals his or her job details (e.g., if f_{56}=Job Information, then w_{56}=1 when m_{156}=1).

School and Job Features				
Feature	Description	Modalities		
		Video	Audio	Metadata
f_{51}	Student Number	5	5	5
f_{52}	School Name	2	2	2
f_{53}	School Information	1	1	1
f_{54}	Degree	1	1	1
f_{55}	Courses Information	2	3	1
f_{56}	Job Information	3	2	1

TABLE 3.7: Weight Values Related to School and Job Features

3.3.7 Family and Friends Features

Definition: Family and friends features (FFF) is any information about YouTuber's family members, relationship, family life, etc.

Table 3.8 shows some sensitive information relation to family and friends features such as marital status or family member information. Disclosing information about families, friends, or sentimental partners increases the YouTuber's privacy risk [19],[44], [25]. After analyzing YouTube videos, we found that most teenagers YouTubers are willing to share this type of information.

Some examples of YouTubers exposing FFF are:

- YouTuber uploads a video and shows his/her boyfriend/girlfriend (e.g., if f_{62}=Partner, then w_{62}=3 when m_{162}=1).
- YouTuber talks about his/her husband/wife (e.g., if f_{62}=Partner, then w_{62}=3 when m_{262}=1).
- YouTuber uploads a video talking about his/her mother's job (e.g., if f_{63}=Family Member, then w_{63}=3 when m_{263}=1).
- YouTuber reveals details about family issues (e.g., if f_{64}=Family Life, then w_{64}=3 when m_{264}=1).

Family and Friends Features				
Feature	Description	Modalities		
		Video	Audio	Metadata
f_{61}	Marital Status	3	2	1
f_{62}	Partner	3	3	3
f_{63}	Family Member	3	2	1
f_{64}	Family Life	3	3	2
f_{65}	Friends	2	1	1

TABLE 3.8: Weight Values Related to Family and Friends Features

Therefore, a feature of a YouTube video has three main characteristics, the type, the subtype that identifies its, and the weight. The values of weight of each feature have been assigned empirically. However, we consider a formal analysis to calculate the values of the weight of the features of a YouTube video as future work.

The information disclosure has been researched in several studies [89] [38]. In [54], the authors provided a classification scheme of the profile items. In a similar way, we provide of a classification of the information disclosure, called content-sensitive features. Having the content features and their weight, YouTubers can refer those tables to have an understanding of the type of information that is considered sensitive in terms of privacy. In the next section, we use the content features in order to compute the PEI of a video.

3.4 Summary

In this chapter, we first presented an overview of the YouTube platform where we described how YouTube videos could potentially affect the privacy of the YouTubers by making an analysis of the privacy risks associated with the information disclosure. Second, we examined previous approaches for measuring privacy risk on traditional social networks such as Facebook, Twitter, and LinkedIn. Third, based on previous research, we proposed a framework for measuring privacy exposure index (PEI) on YouTube videos by analyzing the information.

Chapter 4

YouTubers' Privacy Behaviour

In this chapter, we report on an empirical study of YouTube that we have conducted to understand the factors determining the exposure of personal information on YouTube videos and the consistency between the privacy concerns of the YouTubers and their actual privacy behaviour when they upload a video. Also, we study how extensive is the influence of Viewers on the YouTuber's privacy exposure.

We present the YouTuber's research model and hypotheses development in Section 4.2 followed by the Viewers' research model and hypothesis development in Section 4.3. The research methodology and data collection process are described in Section 4.4. The results of the survey are presented in Section 4.5. The evaluation and interpretation of the hypotheses are presented in Section 4.6 and 4.7, respectively.

4.1 Research Questions of the Empirical Study

In this section, we formulate the research questions associated with the YouTube empirical study. The first two research questions are intended to support empirical evidence regarding YouTuber's privacy behaviour and concerns. The third research question is intended to determine the influence of the Viewers on the privacy behaviour exposure of the YouTubers. In this way, these research questions are formalized as follows:

ERQ1. What are the factors moderating YouTuber's privacy behaviour?

We aim to determine the factors affecting YouTuber's privacy behaviour. For example, the number of subscribers of a YouTube Channel [41], the type of YouTube Channel or the demographic factors [92].

ERQ2. Are YouTuber's privacy concerns consistent with their privacy behaviours when uploading YouTube videos?

In the privacy field, there is a phenomenon called *Privacy Paradox* that states that the privacy concerns and behaviours of an individual not always are aligned. The second research question has the intention to establish how consistent are the YouTubers' privacy behaviours and concerns when they upload a video and also based on the theory of *Privacy calculus* [1] determine the benefits of information privacy on YouTube.

ERQ3. Do viewers have an internal tendency to influence YouTubers' privacy exposure?

We aim to determine the Viewers' influence on YouTuber's privacy exposure behaviour. We believe that Viewers could be influencing the amount of personal information that YouTubers are willing to reveal through the comment section and the Viewers' content engagement [116].

4.2 YouTuber's Research Model and Hypothesis Development

This section describes the YouTuber's research model as well as the associated hypotheses of our empirical study. Six hypotheses are constructed regarding YouTuber's privacy concerns and behaviour when a YouTube video is uploaded by a YouTuber.

In the research model depicted in Figure 4.1, we visualized the relationships between the different factors affecting the YouTuber's privacy behaviour (YPB). The YouTuber's privacy behaviour (YPB) is in the centre of the model and it is considered as a dependent variable, the factors affecting the YouTuber's privacy behaviour (YPB) are presented in the model with incoming arrows to YPB. At the bottom of the model we presented a triangle with the YouTuber's privacy concern (YPC), the YouTuber's self-reported behaviour (YSrB) and the YouTuber's privacy behaviour (YPB) representing the consistency between the YPB and YPC as we aim to determine in our second research question. We describe below the five hypotheses regarding the factors affecting the YouTuber privacy behaviour (YPB) and one hypotheses about the consistency between YouTuber's privacy concerns (YPC) and behaviour (YPB).

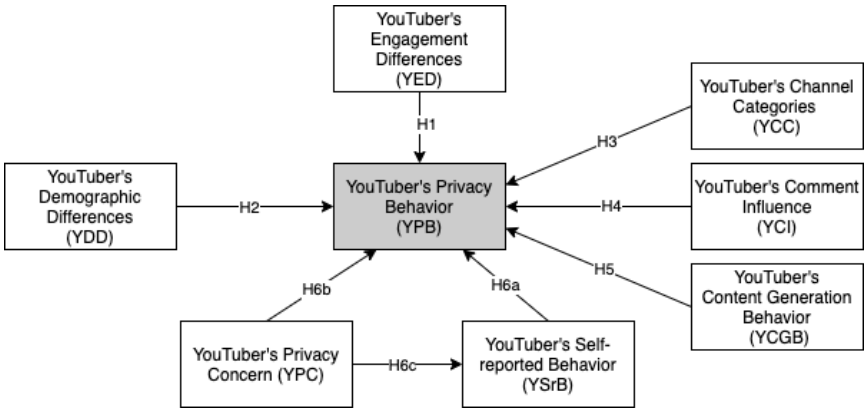


FIGURE 4.1: Research Model for YouTuber

4.2.1 Factors Affecting YouTuber's Privacy Behaviour

After an extensive analysis of YouTube videos throughout the past two years as well as previous surveys addressing privacy issues and implications on social network sites (SNS) [38] the following five factors have been selected for further investigation as potential privacy factors affecting YouTuber's privacy behaviour. Five hypotheses have been developed to answer our first research question.

Engagement differences, such as the number of subscribers of a YouTube Channel or YouTube Views per video could have an impact in the YouTuber's privacy behaviour, we claim that there is a relation between these two measures [55].

H1: The number of subscribers of a YouTube Channel is affecting the YouTuber's privacy behaviour.

Demographic differences such as gender, age, education, and location also have been found to impact individuals' privacy behaviour [91].

H2: YouTuber's privacy behaviour differs based on (a) gender, (b) age (c) location, and (d) education.

YouTube Channel Category differences such as People&Blogs, Gaming, Entertainment could be affecting the privacy exposure behaviour [11] [105].

H3: YouTube's channel category is affecting the YouTuber's privacy behaviour.

The participation of a Viewer has been conceptualized as active participation and passive content consumption [51]. Thus, we intent to determine the influence of the participation of a Viewer through the comments of the *comment section* [76] on the YouTuber's privacy exposure behaviour.

H4: YouTuber's privacy behaviour is influenced by the comment section.

Content generation refers to the uploading frequency of videos on a YouTube channel [94]. We hypothesize that there is a relationship between the YouTuber's content generation behaviour and their privacy behaviour. For example, if a YouTuber uploads more videos than other YouTubers, then eventually this YouTuber is going to tend to disclose more personal information.

H5: YouTuber's content generation behaviour is determining the YouTuber's privacy behaviour.

4.2.2 Consistency Between YouTuber's Privacy Concerns and Behaviour

In order to answer the second research question, we expanded one of the hypothesis into three to better establish the consistency between the self-reported YouTuber's privacy concern and the actual behaviour found when a video is uploaded by the YouTuber.

H6: YouTubers who are more concerned about their privacy, report being more careful when uploading videos.

- **H6a:** YouTuber's privacy behaviour is consistent with the YouTuber's self-reported behaviour.
- **H6b:** YouTuber's privacy behaviour is consistent with his/her privacy concerns.

- **H6c:** YouTuber's privacy concern is consistent with the self-reported behaviour.

4.2.3 YouTuber's Constructs and Measurement Items

The YouTubers' questionnaire is structured as follows: it consists of multiple constructs each corresponding to one of the hypotheses stated in Section 4.2.1 and 4.2.2, and manifested by one or more questions. The questions for each construct are detailed in Table 4.1, and the full questionnaire is in Appendix A.

Questions Q3-Q6 were used to determine the demographic differences of the YouTubers of the hypothesis H2, the construct for the demographic differences is called YouTuber's demographic differences (YDD) [91]. Age-related demographic question Q3 has 7 answers ranges that allow YouTubers of all ages to respond as long as they are not minors thus there are 7 levels of the indicator age. Likewise, gender-related demographic Q4 has 4 answer ranges that allow YouTubers of all genders to respond without exclusions, and the education-related demographic Q6 has a 6 answer ranges [57].

The second construct, YouTuber's engagement differences (YED), was designed to differentiate YouTubers according to the number of subscribers [112] on their YouTube channels as we mentioned in hypothesis H1. Question Q18 captures the number of subscribers [102] and it has five-point Likert scale items that allows the number of subscribers to be presented in a scale 1-5 [1. Less than 10K, 2. 10K-100K, 3. 100K- 500K, 4. 500K- 1M, 5. More than 1M].

In the same way for capturing the type of YouTube's channel categories [18] of each YouTuber participant, as we indicated in hypothesis H3, the YouTuber's channel categories (YCC) construct was created with one question. Q19 was designed with the possible answers of shown in Table 3.1.

The next construct, YouTuber's content generation behaviour (YCGB), was designed to measure the differences of the YouTube usage of each YouTuber [115] using three Likert scale questions Q20-Q22 as we mentioned in hypothesis H5. To capture the number of videos that a YouTuber posted per week [30] we used question Q20 with a scale 1-4 [4. Less than 2, 3. 2-3, 2. 4-5 , 1. More than 5], likewise, we used questions Q21 in a scale 1-5 [1. Never, 2. Seldom, 3. Sometimes, 4. Often, 5. Always] and Q22 in a scale 1-6 [1. Immediately, 2. 1-2 days, 3. 3-4 days, 4. 5-6 days, 5. After one week, 6. Other] to capture the privacy behaviour of a YouTuber before the video is published.

The YouTuber's comment influence (YCI) construct was created to measure the influence of the comments on a YouTuber [96] as we stated in hypothesis H4 using a seven-point Likert scale question. We adapted the method proposed in Equation 4.2 to compute the YouTuber's comment influence. We compute the overall YCI by adding a weight of 0-3 to each possible answers [63] [2. Positive Comments related to the content of the video, 2.Negative Comments related to the content of the video, 1. Positive Comments related to the video/audio quality, 1.Negative Comments related to the video/audio quality, 3. Positive Comments related to my appearance, 3. Negative Comments related to my appearance, 0. I usually do not reply comments].

The last three constructs were used to measure the consistency between the YouTuber's privacy concern (YPC) (Q37) [66] and the YouTuber's privacy behaviour (YPB) (Q27-29, Q31 and Q33) [87] using Likert scale questions. The YouTuber's self-reported behaviour (YSrB) construct is a self-reported privacy behaviour measure with three questions (Q25, Q35 and Q36). The complete questionnaire is listed in Appendix A.

TABLE 4.1: Constructs and Measurement Items for YouTuber Privacy Behaviour

Construct	Question	Ref.
YouTuber's Demographic Differences (YDD)	Q 3 - What is your gender? Q 4 - What is your age? Q 5 - Which country do you live? Q 6 - What is your highest level of education?	[91] [57]
YouTuber's Engagement Differences (YED)	Q 18 - How many subscribers do you have?	[112] [102]
YouTuber's Channel Categories (YCC)	Q 19 - Which of the following categories does your YouTube channel belong to?	[18] [11]
YouTuber's Content Generation Behaviour (YCGB)	Q 20 - How many videos do you post per week? Q 21 - How many times do you review the final edited video before uploading it? Q 22 - Typically, how long after a video is completed do you upload it on the YouTube platform?	[30] [115]
YouTuber's Comments Influence (YCI)	Q 24 - Which type of comments do you usually reply to?.	[96] [63]

Table 4.1 – *Continued on next page*

Table 4.1 – *Continued from previous page*

YouTuber's Self-reported Behaviour (YSrB)	Q 25 - How often does your audience feedback influence the content that you upload to YouTube? Q 35 - Express your agreement with the following sentence: I am comfortable speaking about my private life on my videos? Q 36 - Before uploading a video to YouTube, do you always check for private information?	
YouTuber's Privacy Behaviour (YPB)	Q 27 - Are you willing to share your health information with your audience on your videos? Q 28 - Do you record video clips on places around your home or the place your live or work? Q 29 - Do you speak about your spouse, girlfriend or boyfriend on your videos? Q 31 - Are you willing to share your salary, wealth and other financial information with your audience? Q 33 - How often are your exposing your neighborhood and street signs around your home address on your YouTube videos?	[87] [90] [65]
YouTuber's Privacy Concerns (YPC)	Q 37 - How do you rate yourself with respect to your privacy? >Conservative: I always give priority to my privacy when I'm uploading a video. >Easy going: Privacy is not my primary concern when I'm uploading a video.	[66] [117]

4.3 Viewers' Research Model and Hypothesis Development

In this section, we describe the Viewers' research model and formulate six hypotheses for the Viewers' study. To address our third research question, we aim to understand the Viewers' tendency to influence on the YouTuber's privacy exposure behaviour on videos and also we intend to determine the factors affecting the Viewers' content interaction.

Figure 4.2 conceptualizes the relationship between the Viewers' influence behaviour (VIB) and the Viewers' content interaction (VCI) that we lead us to know whether the Viewers are influencing the YouTuber's privacy exposure, these variables are in the centre of our research model. On the left side, we treat VCI as a dependent variable and the factors affecting VCI are shown in the model with incoming arrows. Then on the right side, a triangle shows the relationship between the Viewers' influence behaviour (VIB) and their desire to influence YouTubers when there is more personal information on videos. We describe below the six hypotheses regarding the Viewers' content interaction and the Viewers' influence behaviour.

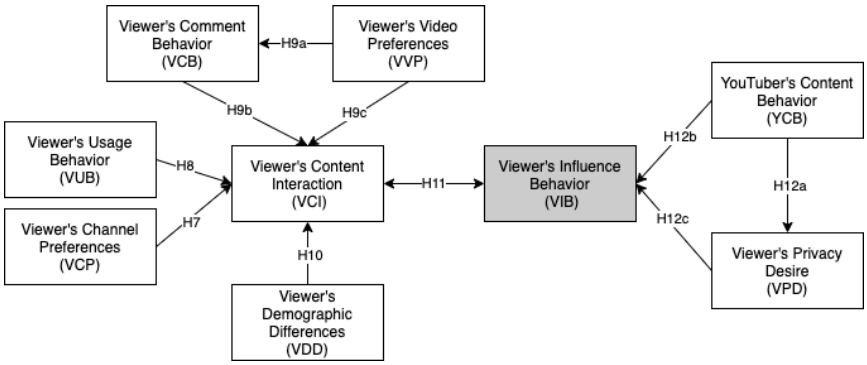


FIGURE 4.2: Research Model for Viewer

4.3.1 Viewers' Content Interaction and Influence Behaviour

In order to establish the consistency between the Viewers' content interaction and their influence behaviour, we determine the factors affecting content interaction as well as the factors modeling the Viewers' influence behaviour.

There is a YouTube category associated with each YouTube Channel as we explained in Section 3.1.2. We state that there is a relationship between the type of YouTube channel category that Viewers watch more frequently and the Viewers' content interaction such as comment on a video [11].

H7: Viewers' content interaction is being affected according to the most-watched type of the YouTube channel categories.

Viewers' content interaction is determined by the number of YouTube videos *content usage* that they watch every day . We state that if a Viewer watch more YouTube videos then we can assume that the interaction with the YouTubers are stronger.

H8: Viewers' content interaction is determined by the Viewers' usage behaviour.

Commenting on YouTube videos could be influenced by the amount of personal information on videos [109].

H9: Viewers commenting more on videos with personal information exposure is determining the Viewers' content interaction.

- **H9a:** Viewers leave more comments on videos that disclose personal information than in other videos.
- **H9b:** Viewers' number of comments per day is determining the Viewers' content interaction.
- **H9c:** The type of video is determining the Viewers' content interaction.

Demographic differences such as age, gender, education have been found to impact in the Viewers' interaction behaviour [91].

H10: The Viewers' interaction behaviour changed based on (a) gender (b) age (c) location (d) education.

Viewers usually have a strong interaction with their favorite YouTubers through self-disclosure comments, content-related comments, or YouTuber-related comments, our goal is to determine if there is some internal intentions to influence on the privacy exposure behaviour of the YouTuber when this *viewer-youtuber* interaction becomes stronger [114].

H11: The Viewers' content interaction is consistent with the Viewers' influence behaviour.

H12: The number of videos with personal information uploaded by a YouTuber is determining the Viewers' desire to influence the privacy exposure.

- **H12a:** The number of videos with personal information disclosed on a YouTube channel is determining the Viewers' privacy desire.
- **H12b:** The number of videos with personal information is determining the Viewers' influence behaviour.
- **H12c:** The Viewers' desire is affecting the Viewers' influence behaviour.

4.3.2 Viewers' Constructs and Measurement Items

The Viewer questionnaire is structured as follows: it consists of multiple constructs each corresponding to one of the hypotheses stated in Section 4.3.1 and manifested by one or more questions. The questions for each construct are detailed in Table 4.2 where we present the questions for Viewer participants, Q8 to Q17 were designed to capture a self-evaluation of YouTube usage of the Viewers, these questions are about YouTube's usage statistics of a Viewer participant [78].

We had question Q8 to capture how a Viewer rates him/herself with respect to the participation in commenting on a YouTube video in a scale 1-5 (1- Uninvolved, 2- Somewhat uninvolved, 3- Neither uninvolved or participative, 4- Somewhat participative, 5- Participative) [51]. We used Q9 and Q11 to measure the YouTube Usage Behaviour [65] [34]. To capture how often the viewer comments on the videos we use Q10. Likewise, we had Q12 to determine the channel categories preferences of the Viewers and Q13 to capture the interest of a Viewer when they leave a comment. For measuring the number of Vlogs that the Viewers tend to watch we used Q14. We used Q15 and Q16 to identify if there is a relationship between the engagement of a participant with the disclosed personal information of the YouTuber. Lastly, the demographics questions (Q3-Q6) were used to measure the effect of the Viewers' demographic differences [57] [91].

TABLE 4.2: Constructs and Measurement Items for Viewers' Content Interaction and Viewers' Influence Behaviour

Construct	Question	Ref.
Viewers' Demographic Differences (VDD)	Q 3 - What is your gender? Q 4 - What is your age? Q 5 - Which country do you live? Q 6 - What is your highest level of education?	[91] [57]
Viewers' Content Interaction (VCI)	Q 8 - As a YouTube viewer, how would you describe yourself? > participative: I always like and comment on the video. > uninvolved: I only watch the video without interaction.	[51]
Viewers' Usage behaviour (VUB)	Q 9 - How many Vlogs videos do you watch per day? Q 11 - How many videos do you watch per day?	[65] [34]
Viewers' Comment behaviour (VCB)	Q 10 - In a regular week, how many videos do you comment?	[78]

Table 4.2 – *Continued on next page*

Table 4.2 – *Continued from previous page*

Viewers' Channel Preferences (VCP)	Q 12 - What type of YouTube Channels do you prefer to watch?	[92] [97] [11]
Viewers' Influence behaviour (VIB)	Q 13 - Complete the sentence as best describes your interests: When I comment on a video, typically I like to know	[7]
YouTuber's Content behaviour (YCB)	Q 14 - In the past 6 months, how many Vlogs have been uploaded by your favorite YouTuber?	[108] [114]
Viewers' Videos Preferences (VVP)	Q 15 - In which type of videos do you comment on more frequently?	[49] [58] [24] [70]
Viewers' Privacy Desire (VPD)	Q 16 - Express your agreement with the following sentence: The most likable videos are the ones where the YouTuber speaks about her/his personal life and issues?	[100]

4.4 Research Methodology

This research was conducted as an online survey presented to YouTubers and Viewers. In this section, we report the development of the questionnaire followed by explaining the data collection process and survey administration.

4.4.1 Scale Development

The questionnaire is based on prior literature in the field of privacy implications previously studied in Chapter 2. We measure the privacy behaviour and concerns on YouTube's users in order to define the privacy factors affecting the privacy risks on YouTube videos.

The questionnaire includes 38 questions divided into two parts as we have two types of YouTube users (YouTuber and Viewers). Where 14 questions were designed for the Viewers and 26 were intended for YouTubers. The first two questions (Q1 and

Q2) are introductory/criteria questions. The next four questions (Q3-Q6) are demographics questions. We used question Q7 in identifying whether the participant is a YouTuber or a Viewer. Thus, the first seven questions are intended to be answered for both groups of participants.

4.4.2 Survey Administration

This survey is part of a study that has been reviewed and cleared by the McMaster Research Ethics Board (MREB) on March 20, 2020. The MREB protocol number associated with this survey is 3635. Before collecting data a focus group of 4 YouTubers and Viewers evaluated the questions of the survey.

As we required to collect information regarding the privacy behaviour and concerns of YouTube's users then a viable solution to collect potential YouTube users (YouTubers and Viewers) was to recruit participants throughout a YouTube Channel with a large and balanced audience. Therefore we decided to upload a recruitment video on a YouTube channel of 390,000 subscribers called VaneVane [21]. We made the survey available in the description box of the recruitment video for a seven-day period.

Participation in this study was voluntary, the first page of this survey was a criteria question to participate in this survey, participants had to meet the following criteria: be over 18, speak fluent English or Spanish and be a YouTube user. After following the criteria, YouTube participants were required to read the opening paragraph about the purpose of the study and the informed consent, and then respond if they understood the informed consent and were willing to continue with the survey questions. At any time even after answering the criteria and consent, the participant could leave the survey and still be considered as an entry valid for a draw for a prize draw. However, if participants dropped the survey, then these unfinished surveys are not considered in this study. In gratitude for answering the survey, we offered participants a chance to enter a draw to win 5 PayPal cards valued at $20.

This survey was confidential, all responses were anonymized. Any contact information we collected from participants to entry for the draw were stored separately from their answers to the survey questions and were deleted once the draw was completed.

4.4.3 Survey Validity and Reliability

In order to increase the validity of our survey we included a number of questions throughout the survey to cross validate other questions. For example we used Q26 as cross validation for Q29. Table C.1 describes all original questions and their cross validation questions. We avoided including these complementary questions when measuring our constructs. Instead for each pair of cross validation questions we ran a correlation test to find out the relationship. If the results of two questions are highly correlated, we conclude the validity of the measurement. We further discuss this test in Section 4.6 of this chapter.

Reliability of the Questionnaire

We have also measured the internal consistency of the construct that we are measuring in our questionnaire using reliability test. Since we did not ask the same questions from the participants twice to check the reliability (test-retest), We computed the Cronbach's Alpha coefficients for all constructs used in measuring YouTuber's privacy concerns (YPC), YouTuber's privacy behaviour (YPB) and Viewers' influence behaviour (VIB). The results are reported in Table C.2 and indicates a good scale reliability.

4.5 Analysis and Results

This section summarizes the data that was collected on this study and the statistical analyses that were performed. First, we describe our sample in terms of demographic based on age, gender, education, and residency country.

4.5.1 Descriptive Analysis and Sample Characteristics

There were 2,875 participants in our study. However, not all participants completed the survey or followed the requirements criteria. The number of participants who finished the survey were 2,460 out 2,875 (85.57%) and 415 out 2,875 (14.43%) dropped the survey. The number of participants who followed the criteria and consent (Q1 and Q2) were 2,412 out 2,460. Thus, 2,412 responses were usable for this study as shown in Table 4.3.

TABLE 4.3: Survey Population

Partial	**Full**	**Usable**	**Total**
415	2,460	**2,412**	2,875
14,43%	85.57	83.90%	100%

As we mentioned previously, our study has targeted two groups as shown in Table 4.4 where most of our participants 2,319 out of 2,412 usable responses (96.14%) were Viewers. 93 out of 2,412 usable responses (3.86%) were YouTubers and the rest of the participants that declined to answer were not considered in our study.

TABLE 4.4: YouTube's user Population

Viewer	**YouTuber**	**Usable**
2,319	93	**2,412**
96.14%	3.86%	100%

The distribution of viewer participants were highly skewed in terms of gender where 2,244 out of 2,319 (96.77%) participants were female and 68 out of 2,319 (2.93%) were male, 4 out of 2,319 (0.22%) participants declined to answer their gender and 2 out of 2,319 (0.09%) answered the option Others. The distribution of YouTuber participants were less skewed in terms of gender where 71 out of 93

(76.34%) were female and 16 out of 93 (17.20%) were male. Both groups of our participants for this study were mostly female as illustrated in Table 4.5.

TABLE 4.5: Participants Gender Distribution

Answer	**Viewer**		**YouTuber**	
	Count	**% Total**	**Count**	**% Total**
Male	68	2.93%	16	17.20%
Female	2,244	96.77%	71	76.34%
Others	2	0.09%	6	6.46%
Prefer not to disclose	5	0.22%	0	0.00%
Total (Gross)	**2,319**	100%	**93**	100%

In terms of age, more than 80% of Viewer participants were less than 34 years old as shown in Table 4.6. 870 out of 2,319 (37.52%) participants were in the range of 18-24 years old and 1,023 out of 2,319 (44.11%) were in the range of 25-34 years old, the rest of the participants were older than 35 years old, this is consistent with the age groups of the YouTubers [103]. Likewise, YouTuber participants mainly were less than 34 years old . 36 out of 93 (38.71%) participants were in the range of 18-24 years old, 45 out of 93 (48.39%) were in the range of 25-34 years old, and 6 out of 93 (6.45%) were in the range of 35-44 years old, while the rest of the YouTuber participants are older than 45 years old.

TABLE 4.6: Participants Age Distribution

Answer	**Viewer**		**YouTuber**	
	Count	**% Total**	**Count**	**% Total**
18-24	870	37.52%	36	38.71%
25-34	1,023	44.11%	45	48.39%
35-44	288	12.42%	6	6.45%
45-54	94	4.05%	3	3.23%
55-64	29	1.25%	1	1.07%
65+	4	0.17%	2	2.15%
do not want to disclose	11	0.47%	0	0.00%
No answer	0	0.00%	0	0.00%
Not completed or Not displayed	0	0.00%	0	0.00%
Total (Gross)	**2,319**	100%	**93**	100%

Both groups of our participants were highly educated. For viewer participants 947 out of 2,319 (40.84%) completed graduate school and another 674 out 2,319 (29.06%) graduated from some school as shown in Table 4.7. Likewise, more than 65% of the YouTuber participants were completed graduate school or graduated from some graduate school or College. 52 out 93 (55.91%) completed graduate school, 22 out of 93 (23.66%) were some graduate school, 7 out of 93 (7.53%) graduated from college.

TABLE 4.7: Participants Education Distribution

Answer	**Viewer**		**YouTuber**	
	Count	**% Total**	**Count**	**% Total**
Some High School	97	4.18%	3	3.23%
Graduated from High School	354	15.27%	8	8.60%
Graduated from College	185	7.98%	7	7.53%
Some Graduate School	674	29.06%	22	23.66%
Completed Graduate School	947	40.84%	52	55.91%
do not want to disclose	62	2.67%	1	1.08%
No answer	0	0%	0	0.00%
Not completed or Not displayed	0	0%	0	0.00%
Total (Gross)	**2,319**	100%	**93**	100%

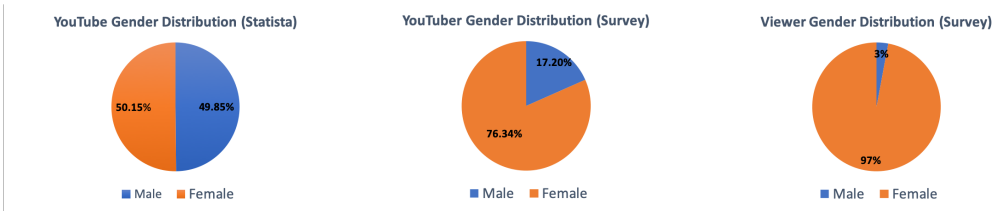
Participants in this study mainly lived in South American countries and the United States as shown in Table 4.8. 543 out of 2,319 (23.42%) Viewer participants lived in Mexico, 445 out of 2,319 (19.19%) lived in Ecuador, and 171 out of 2,319 lived in the United States. In the case of our YouTuber participants, most of them lived in Mexico, Colombia, Ecuador, and the United States. Similarly, YouTuber participants mainly lived in South American countries and the United States. The descriptive analysis for the rest of the questions are reported in Appendix B.

TABLE 4.8: Participants Location Distribution

Answer	**Viewer**		**YouTuber**	
	Count	**% Total**	**Count**	**% Total**
Mexico	543	23.42%	17	18.28%
Ecuador	445	19.19%	19	20.43%
United States	171	7.37%	12	12.90%
Colombia	139	5.99%	9	9.68%
Others	1,021	44.03%	36	38.71%
Total (Gross)	**2,319**	100%	**93**	100%

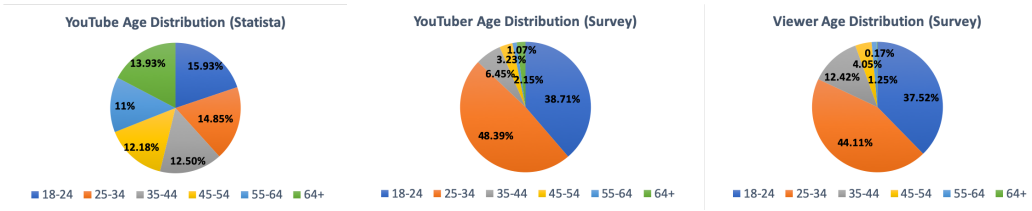
4.5.2 Comparison of the sample with YouTube General Population

When comparing our sample with the YouTube population in terms of gender, our survey population is highly skewed. We used the United States YouTube users as the population to compare [104] as shown in Figure 4.3. Likewise, the distribution of age for our sample also reflects a similar statistics as we can see in Figure 4.4 where our participants were not well distributed in terms of age. According to [110], in contrast with the general population in terms of education level, YouTube users are more likely to have a college degree. This can be reflected in the distribution of our participants in terms of their education as shown in Figure 4.7 where 947 out of 2319 (40.84%) Viewers and 52 out of 93 (55.91%) YouTubers had completed graduate school. Given these limitations, we are cautious in making interpretations in terms of gender, age, and education. This is an area of future work to study a more diverse group of YouTube population in terms of gender, age, and education.



(a) (b) (c)

FIGURE 4.3: Comparison Between the General Population of YouTube Users and Our Survey Population in Terms of Gender



(a) (b) (c)

FIGURE 4.4: Comparison Between the General Population of YouTube Users and Our Survey Population in Terms of Age

4.5.3 Factors Affecting YouTuber's Privacy Behaviour

In order to analyze the relationship between the YouTuber's privacy behaviour (YPB) and other YouTube privacy factors as depicted in Figure 4.1, we tested hypotheses H1 through H5. In order to measure the YouTuber's privacy behaviour (YPB), we computed a score called Privacy Exposure Index (PEI) as described in Chapter 3 based on the questions Q27-Q29, Q31 and Q33. Equation 4.1 shows how we aggregated PEI for each individual participant.

Measuring Privacy Exposure Index. We measured Youtubers' privacy behaviour using Q27-Q29, Q31 and Q33. Each question collects self-reported behaviour in terms of different aspects of privacy including health, location, personal finance, relationship as we mentioned in Chapter 3. Since the impact of each question on the YouTuber privacy behaviour (YPB) construct can be different, we used PEI, introduced in Chapter 2 of this thesis, to aggregate the YouTubers Behaviour Exposure over multiple indicators.

$$PEI = \sum_{i=1}^{n} \beta \, w_i \tag{4.1}$$

where:

β = sensitivity score
w_i = feature normalized value
n = number of features

TABLE 4.9: Method of deriving scoring systems to sensitive information for the construct YPB

Feature	β	**Question**
Address	3	Q33
Name	1	–
Age/Date of Birth	2	–
Interests/Hobbies	1	–
Health Information	2	Q27
Places	2	Q28
School/ Job	1	–
Relationship	2	Q29
Families/Friends	1	–
Finance	2	Q31

Note that we normalized answers for each question (Q27-29, Q31, Q33 and Q36) before using Equation 3.3 to compute the aggregated score, the results of the data normalization of Q27-29, Q31 and Q33 were considered as the feature weights (w_i).

Higher values of PEI mean that YouTuber's privacy behaviour are exposing more sensitive information than YouTubers with a lower score of PEI. The results show that the minimum and maximum value of PEI of our collected data were 3.5 and 11 respectively.

Similarly, we measured YouTuber's comment influence (YCI) by computing a score called Privacy Comment Index (PCI) based on the question Q24. Equation 4.2 shows how we aggregated PCI for each individual participant.

Measuring Privacy Comment Index. We computed the PCI score using Q24, each of the answers have a sensitivity score called γ between 0-3 as described in Table 4.10.

$$PCI = \sum_{i=1}^{n} \gamma \, \mathrm{f}_i \tag{4.2}$$

where:

γ = sensitivity score of each feature
f_i = feature normalized value
n = number of features

TABLE 4.10: A naïve approach for deriving scoring systems to sensitive information for the construct YCI

Feature	γ
Positive Comments related to the content of the video	2
Negative Comments related to the content of the video	2
Positive Comments related to the video/audio quality	1
Negative Comments related to the video/audio quality	1
Positive Comments related to my appearance	3
Negative Comments related to my appearance	3
I usually do not reply comments	0

Before we built our regression model for testing our hypothesis we calculated the Spearman correlation matrix between our variables and also run a normality test using Shapiro-Wilk test as illustrated in Appendix D.1. For the former test, the correlation matrix results (as reported in Appendix D.2 show there were a correlation between our variables. For the latter test, the null hypothesis states that the sample comes from a normal distribution. The results indicated that there was no evidence to suggest normality in any of the constructs.

For hypothesis H1, the Spearman's rank-order correlation test indicated that there was a negative correlation between the YouTuber's engagement differences (YED) and the YouTuber's privacy behaviour (YPB) constructs as can be seen from Table D.3. Therefore, we tested the hypothesis H1 using a linear regression model where the YouTuber's privacy behaviour (YPB) was the dependent variable and the YouTuber's engagement differences (YED) was the independent variable. We aimed to understand how engagement differences may contribute to predicting the variability of YouTuber's privacy behaviour and whether the contribution was significant. We ran the regression model and the overall results ($F = 5.600$, $Prob(F) = 0.020$, $R^2 = 0.058$) predicted the YouTuber's privacy behaviour significantly well as reflected in Table 4.11. In examining the coefficient **B** of the model, YED, the YouTuber's engagement differences ($t(93) = -2.366, p < .05$) indicated that was statistically significant because its $p - values$ equal 0.020, accepting the hypothesis H1.

TABLE 4.11: Linear Regression Results for Hypothesis H1

Variable	B	SE B	t	p	95% Conf.	
(Constant)	8.407	0.427	19.703	0.000	7.560	9.255
YED	-0.513	0.217	-2.366	0.020	-0.943	-0.082
$F = 5.600$, $Prob(F) = 0.020$, $R^2 = 0.058$						

To test our hypothesis H2, we performed multiple Kruskal-Wallis tests to determine whether there were significant differences in YouTuber's privacy behaviour (YPB) associated to some personal characteristics identified as the YouTuber's demographic differences (YDD) as reported in Table 4.1. The results stated that the population median of location and education groups are equal as reflected in Table D.10 of the Appendix D.

In order to test the hypothesis H3, we first analyzed the relation between the YouTuber's channel categories (YCC) and the YouTuber's privacy behaviour (YPB) constructs using the Spearman's rank-order correlation, results indicated that there was a correlation between the variables as can be seen in Table D.4. Consequently, we tested hypothesis H3 using a multi-variable linear regression model where the dependent variable was the YouTuber's privacy behaviour (YPB) construct and the YouTuber's channel categories (YCC) construct were the independent variable. Results of the regression model ($F = 3.521$, $Prob(F) = 0.010$, $R^2 = 0.138$) were presented in Table 4.12. The regression output indicated that the Entertainment ($t(93) = 5.104, p < .01$), Howto and Style ($t(93) = 2.686, p < .01$) and People and Blogs ($t(93) = 5.335, p < .01$) predictor variables were statistically significant because their $p - values$ equal 0.000. On the other hand, Education ($t(93) = -0.453, p > .05$) and Other ($t(93) = 0.891, p > .05$) channel categories were not statistically significant because theirs $p - values$ were greater than the usual significance level of 0.05.

TABLE 4.12: Multi-variable Linear Regression Results for Hypothesis H3

Variable	B	SE B	t	p	95% Conf.	
(Constant)	6.047	0.222	27.193	0.000	5.605	6.489
YCC	1.998	0.392	5.104	0.000	1.220	2.776
	1.239	0.461	2.686	0.009	0.3222	2.155
	2.674	0.501	5.335	0.000	1.678	3.669
	-0.333	0.734	-0.453	0.651	-1.792	1.126
	0.470	0.527	0.891	0.376	-0.578	1.357
$F = 3.521, Prob(F) = 0.010, R^2 = 0.138$						

For hypothesis H4, according to Spearman's correlation test results shown in Table D.5 there was a correlation between the variables of the hypothesis. Thus, we performed a multi-variable linear regression analysis where the YouTuber's privacy behaviour (YPB) construct was considered as our dependent variable and YouTuber's comment influence (YCI) construct as the independent variables as shown in Table 4.13. The regression model overall predicted the YouTuber's privacy behaviour (YPB) significantly well ($F = 12.560$, $Prob(F) = 0.000$, $R^2 = 0.218$). The regression output suggested that the Privacy comment index (PCI) of the YouTuber's comment influence (YCI) computed using Q24 as can be seen from the Table 4.1 ($t(93) = -1.900, p > .05$) was not statistically significant because its $p - value$ was greater than the significance level of 0.05. Moreover, question Q26 of the YouTuber's comment influence (YCI) construct ($t(93) = 4.139, p < .01$) was statistically significant because its $p - values$ was equal to 0.000.

TABLE 4.13: Multi-variable Linear Regression Results for Hypothesis H4

Variable	B	SE B	t	p	95% Conf.	
(Constant)	5.683	0.733	7.758	0.000	4.227	7.138
YCI	-0.145	0.076	-1.900	0.061	-0.296	0.007
	0.704	0.170	4.139	0.000	0.366	1.041
$F = 12.560, Prob(F) = 0.000, R^2 = 0.218$						

We tested H5 using a multi-variable regression model where the YouTuber's privacy behaviour (YPB) construct was our dependent variable and the YouTuber's content generation behaviour (YPGB) construct were our independent variables. Before we built our regression model, we created a correlation matrix for all dependent and independent variables as shown in Table D.6 reported in Appendix D. This result shows that the variables are correlated with the YouTuber's privacy behaviour. Therefore, we fitted our multi-variable regression model with all the dependent and independent variables as can be seen from Table 4.14. The regression model overall predicted considerably well YouTuber's privacy behaviour (YPB) ($F = 4.603$, $Prob(F) = 0.003$, $R^2 = 0.273$). The coefficients **B** of the model indicated that questions Q21 ($t(93) = 2.888, p < .01$)), Q22 ($t(93) = -2.187, p < .01$)) and Q23 ($t(93) = -3.713, p < .01$)) of the YouTuber's content generation behaviour (YPGB) construct were statistically significant. On the other side question Q20 of the YouTuber's content generation (YPGB) construct ($t(93) = 1.840, p > .05$)) was not statistically significant because its $p - value$ was greater than the usual significance level of 0.05.

TABLE 4.14: Multi-variable Regression Results for Hypothesis H5

Variable	B	SE B	t	p	95% Conf.	
(Constant)	5.404	1.364	3.962	0.000	2.693	8.115
YCGB	0.551	0.299	1.840	0.069	0.223	1.953
	0.532	0.184	2.888	0.005	0.166	0.898
	-0.323	0.148	-2.187	0.031	-0.616	-0.030
	-0.561	0.151	-3.713	0.000	-0.862	-0.261
$F = 17.650$, $Prob(F) = 0.000$, $R^2 = 0.445$						

4.5.4 Consistency Between YouTuber's Privacy Concerns and Behaviour

To determine the consistency between YouTuber's privacy behaviour (YPB) and concerns (YPC) as shown in Figure 4.1, we tested hypothesis H6a through H6c. The first step was to determine the relation between YouTuber's self-reported behaviour (YSrB) and the YouTuber's privacy behaviour (YPB). The next step was to establish whether the YouTuber's privacy concern (YPC) could predict the YouTuber's privacy behaviour (YPB) and the third step was to identify the relation between the YouTuber's privacy concern (YPC) and YouTuber's self-reported behaviour (YSrB).

Before we built our regression model, we created a correlation matrix for our dependent and independent variables for each hypothesis and also ran a number of diagnostic tests (as reported in Appendix D). For H6a, the Spearman's rank-order correlation test results show that there was a correlation between the YouTuber's privacy concern (YPC) and the YouTuber's self-reported behaviour (YSrB) meaning there was a good chance that YPC could predict a significant variation of YSrB. We tested hypothesis H6a using a linear regression model where the independent variable was the question Q36 of the YouTuber's self-reported behaviour (YSRB) and the YouTuber's privacy behaviour (YPB) was the dependent variable. We ran our model where it significantly predicted the outcome of the dependent variable (YSrB) ($F = 8.704$, $Prob(F) = 0.004$, $R^2 = 0.087$) as can be seen in Table 4.15.

TABLE 4.15: Linear Regression Results for Hypothesis H6a

Variable	B	SE B	t	p	95% Conf.	
(Constant)	2.243	0.428	5.244	0.000	1.394	3.093
YSrB (Q36)	0.299	0.101	2.950	0.004	0.098	0.500
$F = 8.704$, $Prob(F) = 0.004$, $R^2 = 0.087$						

Likewise, we performed the Spearman's rank-order correlation test for the variables of the hypothesis H6b and overall there was evidence of correlation between the YouTuber's privacy concern (YPC) and the YouTuber's privacy behaviour (YPB) as illustrated in Table D.8. Thus, we ran a linear regression model where Q27 of the YouTuber's privacy behaviour (YPB) construct was the independent variable and Q37 of the YouTuber's privacy concern (YPC) construct was the dependent variable. The results of the regression model suggested that the YouTuber's privacy concern (YPC) could predict the outcome of the dependent variable ($F = 12.25$, $Prob(F) = 0.001$, $R^2 = 0.119$) as indicated in Table 4.16. The coefficient **B** of the model indicated that the YouTuber's privacy concern (YPC) $(t(93) = 3.500, p < .001)$ was contributing to the model as shown in Table 4.16.

TABLE 4.16: Linear Regression Results for Hypothesis H6b

Variable	B	SE B	t	p	95% Conf.	
(Constant)	1.803	0.486	3.706	0.000	0.837	2.769
YPC (Q37)	0.396	0.113	3.500	0.001	0.171	0.621
$F = 12.25, Prob(F) = 0.001, R^2 = 0.119$						

In order to test the hypothesis H6c, we first analyzed the relation between the YouTuber's privacy concern (YPC) and the YouTuber's self-reported behaviour using the Spearman's rank-order correlation as shown in Table D.9. The correlation matrix suggested that there was no evidence of correlation between the variables rejecting the hypothesis H6c.

4.5.5 Viewer's Content Interaction Determining the Viewer's Influence Behaviour

To investigate the factors determining the Viewers' content interaction (VCI), we tested the hypothesis H7 through H10. Likewise, the factors affecting the Viewers' influence behaviour (VIB) through hypothesis H12. The consistency between the Viewers' content interaction (VCI) and the Viewers' influence behaviour (VIB) is analyzed in hypothesis H11.

Prior to testing each hypothesis, we ran a normality test for each of our constructs shown in Table 4.2 using Shapiro-Wilk test as detailed in Table D.2 of the Appendix D. The results suggest that there was no evidence of normality in our constructs.

After the normality test was completed, we used Spearman's rank correlation test to find the simple correlation between our variables for each hypothesis, the results are shown in Appendix D.4. In hypothesis H7, our preliminary results suggest that there is no correlation evidence between the Viewers' channel preferences (VYP) and the Viewers' content interaction (VCI), rejecting hypothesis H7.

For hypothesis H8, the correlation test indicates that there was a positive correlation between the Viewers' usage behaviour (VUB) and the Viewers' content interaction (VCI) constructs as can be seen in Table D.12. Thus, we tested our hypothesis H8 with a multi-variable regression model where the Viewers' content interaction (VCI) was the dependent variable and the Viewers' usage behaviour (VUB) our independent variables. The regression model overall predicted the Viewers' content interaction (VCI) significantly well as can be seen in Table 4.17 ($F = 43.55, Prob(F) = 0.000, R^2 = 0.036$). In examining the coefficients **B** of the model, question Q9 of the Viewers' usage behaviour (VUB) construct ($t = 6.991, p < .001$) was contributing more to the model than question Q11 of the Viewers' usage behaviour (VUB) construct ($t = 3.459, p < .001$).

TABLE 4.17: Linear Regression Results for Hypothesis H8

Variable	B	SE B	t	p	95% Conf.	
(Constant)	1.833	0.102	17.937	0.000	1.633	2.034
VUB	0.220	0.032	6.991	0.000	0.159	0.283
	0.099	0.029	3.459	0.001	0.043	0.156
$F = 43.55, Prob(F) = 0.000, R^2 = 0.036$						

Before testing our hypothesis H9, we analyzed the preliminary results of the Spearman's test. The test suggested that there is a strong correlation between the variables of H9a, H9b, and H9c as shown in Tables D.13, D.14 and D.15. Thus, these correlation results led us to perform a linear regression for each of the hypotheses.

We tested hypothesis H9a using a linear regression model where the dependent variable was the Viewers' comment behaviour (VCB) construct and the independent variable was the Viewers' videos preferences (VVP) construct. The regression model overall predicted the Viewers' comment behaviour (VCB) ($F = 531.00, Prob(F) = 0.000, R^2 = 0.186$). Table 4.18 presents the tests for the linear regression model for the Viewers' comment behaviour (VCB) and the coefficients of the predictors. The coefficient **B** of the Viewers' videos preferences (VVP) of the model $(t = -15.006, p < .001)$ shows that there was a negative relation between the variables.

TABLE 4.18: Linear Regression Results for Hypothesis H9a

Variable	B	SE B	t	p	95% Conf.	
(Constant)	3.690	0.025	147.228	0.000	3.641	3.739
VVP	-0.118	0.008	-15.006	0.000	-0.134	-0.103
$F = 225.2, Prob(F) = 0.000, R^2 = 0.089$						

For the hypothesis H9b, we ran a linear regression model where the Viewers' comment behaviour (VCB) construct was the independent variable and the Viewers' content interaction (VCI) construct was the dependent variable. Our model overall predicted the Viewers' content interaction ($F = 531.00, Prob(F) = 0.000, R^2 = 0.186$) as detailed in Table 4.19. The coefficient **B** of the model indicated that the Viewers' comment behaviour (VCB) $(t = 23.043, p < .001)$ was significantly contributing to the model.

TABLE 4.19: Linear Regression Results for Hypothesis H9b

Variable	B	SE B	t	p	95% Conf.	
(Constant)	0.488	0.101	4.825	0.000	0.290	0.687
VCB	0.662	0.029	23.043	0.000	0.606	0.719
$F = 531.00, Prob(F) = 0.000, R^2 = 0.186$						

Likewise, we tested our hypothesis H9c using a linear regression model where the Viewers' content interaction (VCI) construct was the dependent variable and the Viewers' videos preferences (VVP) construct was the independent variable. The results indicated in Table 4.20 show the model overall predicted the Viewers' content interaction (VCI) ($F = 633.9, Prob(F) = 0.000, R^2 = 0.215$). The coefficient **B** of the Viewers' videos preferences (VVP) $(t = -25.178, p < .001)$ suggested that there was a negative relation between the variables.

TABLE 4.20: Linear Regression Results for Hypothesis H9c

Variable	B	SE B	t	p	95% Conf.	
(Constant)	3.411	0.036	95.538	0.000	3.341	3.481
VVP	-0.283	0.011	-25.178	0.000	-0.305	-0.261
$F = 633.9, Prob(F) = 0.000, R^2 = 0.215$						

To test our hypothesis H10, we performed multiple Kruskal-Wallis test to determine whether there were significant differences in Viewers' interaction behaviour (VIB) measures associated to some demographic characteristics as reported in Table 4.2. The results concluded that the population median of all groups (gender, age, location and education) were not equal as reflected in Table D.20 of the Appendix D, accepting our hypothesis.

Before testing our hypothesis H11, we evaluated the results of the Spearman's rank correlation test. The test indicated that there was a significant correlation between the Viewers' content interaction (VCI) construct and the Viewers' influence behaviour (VIB) construct. Then, we tested the hypothesis H11 using a linear regression model where the Viewers' influence behaviour (VIB) was the dependent variable and the Viewers' content interaction (VCI) was the independent variable. The model overall predict the Viewers' influence behaviour (VIB) ($F = 400.8, Prob(F) = 0.000, R^2 = 0.147$). The coefficient **B** of the model indicated that the Viewers' content interaction (VCI) ($t = 20.020, p < .001$) was contributing to the model as shown in Table 4.21.

TABLE 4.21: Linear Regression Results for Hypothesis H11

Variable	**B**	**SE B**	**t**	**p**	**95% Conf.**	
(Constant)	2.181	0.061	35.952	0.000	2.062	2.300
VCI	0.398	0.020	20.020	0.000	0.359	0.437
$F = 400.8, Prob(F) = 0.000, R^2 = 0.147$						

Before testing our hypothesis H12, we examined the results of the Spearman's test. The test indicated that there is a strong correlation between the variables of H12a, H12b and H12c as illustrated in Table D.13, D.14 and D.15. These results can be used to continue performing a linear regression for each hypothesis.

We tested hypothesis H12a using a linear regression model where the YouTuber's content behaviour (YCB) was the independent variable and the Viewers' privacy desire (VPD) was the dependent variable. We ran the regression and found that the model overall predicted the Viewers' privacy desire (VPD) ($F = 10.19, Prob(F) = 0.001, R^2 = 0.004$). The coefficient **B** of the model indicated that the YouTuber's content behaviour (YCB) ($t = 3.193, p < .001$) was contributing to the model as detailed in Table 4.22.

TABLE 4.22: Linear Regression Results for Hypothesis H12a

Variable	**B**	**SE B**	**t**	**p**	**95% Conf.**	
(Constant)	2.778	0.046	60.994	0.000	2.688	2.867
YCB	0.044	0.014	3.193	0.001	0.017	0.070
$F = 10.19, Prob(F) = 0.001, R^2 = 0.004$						

For hypothesis H12b, we used a linear regression model, the Viewers' influence behaviour (VIB) was the dependent variable and the YouTuber's content behaviour (YCB) was the independent variable as seen in Table 4.23. The results indicate that the model predicted the Viewers' influence behaviour (VIB) significantly well ($F = 2.121, Prob(F) = 0.145, R^2 = 0.001$). The coefficient **B** of the model suggested

that the YouTuber's content behaviour (YCB) $(t = 1.456, p < .001)$ was contributing positively to the model.

TABLE 4.23: Linear Regression Results for Hypothesis H12b

Variable	B	SE B	t	p	95% Conf.	
(Constant)	3.199	0.060	52.956	0.000	3.081	3.318
YCB	0.026	0.018	1.456	0.145	-0.009	0.062
$F = 2.121, Prob(F) = 0.145, R^2 = 0.001$						

We tested hypothesis H12c using a linear regression model, the Viewers' privacy desire (VPD) was the independent variable and the Viewers' influence behaviour (VIB) was the dependent variable as described in Table 4.24. The coefficient **B** of the Viewers' privacy desire (VPD) $(t = 6.209, p < .001)$ was significantly contributing to the model.

TABLE 4.24: Linear Regression Results for Hypothesis H12c

Variable	B	SE B	t	p	95% Conf.	
(Constant)	2.785	0.084	33.097	0.000	2.620	2.950
VPD	0.169	0.027	6.209	0.000	0.116	0.223
$F = 38.55, Prob(F) = 0.000, R^2 = 0.016$						

4.6 Discussion and Implications

The YouTube study we conducted targeted two YouTube user groups - YouTubers and Viewers - to understand YouTuber's privacy behaviour and concerns as well as the Viewers' influence on YouTuber's privacy exposure as outlined in our research questions discussed in Section 4.1.

4.6.1 Factors Affecting YouTuber's Privacy Behaviour

For the YouTuber analysis, we aimed to determine the factors affecting the YouTuber's privacy behaviour by testing hypothesis H1 through H5. In Section 4.5.3, the regression models built suggested that the number of subscribers of a YouTube channel is a factor determining the YouTuber's privacy behaviour. Based on this statistic YouTubers with more subscribers are exposing less sensitive information than those with a small audience. This finding may suggest that YouTubers who are more experienced and maintain a larger audiences are more cognizant of their privacy. This implies that YouTubers in their early days of joining Youtube and start sharing contents might be more vulnerable to exposing their privacy. Therefore, a privacy tool for YouTubers should consider small YouTubers (e.g., those with less than 100K subscribers) as its target group.

Similarly, we find that the YouTube channel category is affecting the YouTuber's privacy behaviour as it is reflected in Table 4.12. Based on the results, YouTubers with channels categories such as Entertainment, Howto&Style and People&Blogs are exposing more susceptible information on their YouTube videos. Consequently, we could theorize that YouTubers with channel categories where the type of YouTube videos more commonly uploaded are Vlogs, Q/A and/or Storytime videos, are

disclosing more sensitive information. This can be explained by the fact that in these types of videos the content is always related to the YouTuber's personal life. We suggest that in order to keep the privacy exposure at minimum, YouTubers have to develop self-checking behaviour before uploading their videos on YouTube. Thus, a self-checking tool for YouTubers with channel categories of Entertainment, Howto&Style and People&Blogs could help to keep privacy exposure levels on videos at appropriated levels.

According to the results of the regression model indicated in Table 4.13, the YouTuber's privacy behaviour is being affected by the comments section. YouTubers with an audience requesting to know more about their personal life through the comments section are exposing more sensitive information on their YouTube videos. This finding confirms the theory that YouTubers are willing to share a wide range of information about themselves as long as their audience request them as was suggested in previous privacy studies on social networks [27, 96, 6]. The implication of this result is that in order to predict a privacy risk on YouTube videos, the comment section have to be analyzed as a predictor of privacy exposure on YouTube videos.

Another interesting finding is how the YouTuber's content generation behaviour is affecting the YouTuber's privacy behaviour as shown in Table 4.14. The regression model suggests that the number of uploaded videos per week on a YouTube channel is not determining the YouTuber's privacy behaviour but the types of videos like Vlogs or Q/A. Consequently, we state that not all YouTube videos are affecting the privacy exposure of a YouTuber just videos where the YouTuber disclosed more personal information. Once again it is proven that we can predict more privacy exposure in Vlogs. In addition, we found that when a YouTuber waits longer before uploading an edited video on the YouTube platform and also reviews the edited video then the YouTuber's privacy exposure decreases.

Regarding the demographic differences and its impact on the YouTuber's privacy behaviour, we did not perceive evidence to suggest that the YouTuber's income and education were impacting the YouTuber's privacy behaviour. On the other hand, the results for age and gender show there was a difference between the groups. An interesting approach would be to test this study in YouTubers under 18. However due to the requirements of our survey, the participants were not minors. Thus, we could not suggest if there is an evidence that minor YouTubers are exposing more sensitive information on YouTube video than YouTuber adults.

To conclude the analysis of the factors affecting the YouTuber's privacy behaviour, we find that most of our participants (82.79%) independently of the demographic, channel categories, engagement or content generation behaviour differences reported that a YouTube privacy tool that may provide them recommendations about their privacy exposure on their videos could be helpful. Thus, YouTubers participants are expecting an improvement of privacy experts in the YouTube platform.

4.6.2 Consistency Between YouTuber's Privacy Concerns and Behaviour

The second question we were intending to answer was to determine the consistency between the YouTuber's privacy concern and behaviour. Therefore we tested H6a through H6c. In Section 4.5.4 the regression models suggested that the YouTubers who always check for private information before uploading the video on YouTube

are not willing to share private information on their videos. This may suggest that YouTubers who are more aware of their privacy are self-analyzing their videos. This implies that YouTubers who are not analyzing their videos before uploading them are more vulnerable to sensitive information leakage.

According to the results of the regression model referred in Table 4.16, there is a relation between the YouTuber's privacy behaviour and the YouTuber's privacy concern. This result suggests that when a YouTuber considers himself/herself as conservative with respect to his/her privacy, then the actual privacy behaviour of the YouTuber indicates that he/she is more likely to look for private information leakage when uploading a video.

On the other hand, the results of the hypothesis H6c as shown in Table D.9 suggested an inconsistency between the YouTuber's privacy concern and his/her self-reported privacy behaviour. Some YouTubers were answering they consider themselves easy-going then they were reporting that they always check for private information or answering they never check for private information but then considering themselves as conservative. This result indicates that there is a lack of privacy education for YouTubers. Thus, a privacy tool for YouTubers should consider to include a privacy-awareness information manual with some privacy basics in order to educate the YouTubers with basic knowledge about the type of sensitive information that could cause privacy risks.

4.6.3 Viewers' Content Interaction determining the Viewers' influence Behaviour

In our previous YouTuber study where we analyzed the YouTuber's privacy behaviour we found that the comment section was a factor affecting the YouTuber's privacy exposure on videos. Thus an analysis of the Viewers' behaviour was necessary in order to understand the relationship Viewer-YouTuber and how this relationship could affect the exposure of sensitive information on YouTube videos.

In answering our third research question, the hypotheses H7 through H10 were tested. The results of Section 4.5.5 regarding hypothesis H7 suggested that the YouTube channel category of a YouTuber is not determining the Viewers' participation on the comments section. However, we believe the sample size and diversity were a limiting factor for taking a conclusive stance in our study.

On the other hand, according to the results of the regression model shown in Table 4.17, the number of Vlogs videos uploaded by a YouTuber is determining the Viewers' content interaction. This suggests that a Viewer is commenting more on Vlogs than another type of videos. Thus, Vlogs videos are determining if a Viewer is or not participative on a YouTube channel. However, this finding does not nullify the hypothesis that the number of videos uploaded on a YouTube channel, regardless of whether these are Vlogs, could also increase the participation of a Viewer.

An interesting finding of the Viewers' study was that Viewers who consider themselves as participative are commenting more regularly in types of videos where the YouTuber is disclosing personal information such as Vlogs, Makeup Tutorial, Unboxing/Haul and Tag or Challenge videos. Consequently, we state that Viewers were influencing YouTubers more on videos where the YouTubers are discussing

topics related to their personal life. Thus, a privacy tool for YouTubers should consider these types of videos as target videos exposing sensitive information.

4.7 Summary

In this chapter, we reported the results of a privacy YouTube study focused on two YouTube users group: YouTubers and Viewers. We built two research models for the YouTubers and Viewers and tested several hypotheses to understand the factors affecting the exposure of personal information on YouTube videos and the consistency between the YouTuber's privacy concern and behaviour. The Viewer study intended to determine the Viewers' influence on YouTubers to expose more private/personal information.

We concluded this chapter with an interesting finding from the results of the YouTuber study, most of our YouTubers participants (82.79%) reported in the last question of the questionnaire that a YouTube privacy tool that may provide them recommendations about their privacy aspects on the YouTube video could be very helpful for their privacy. Therefore, there is a need to provide a measurement system that helps YouTubers to understand which factors are determining the privacy risk on a YouTube video.

Chapter 5

Evaluation of PEI

In this chapter, we report the empirical evaluation of our framework for measuring privacy exposure of YouTube videos proposed in Section 3.3. We evaluated PEI's soundness, consistency and functionality. To complete the evaluation, we analyzed 100 videos (10 videos from 10 different YouTubers) over a period of two years. We evaluated the soundness of our scoring system by meticulously measuring PEI of each 100 videos. We then evaluated consistency of our privacy score with the perceived privacy scores assigned by three professional YouTubers. We have also evaluated the functionality of our scoring system when a comment sentiment analysis tool is available to be used as a precautionary determinant of privacy exposure.

In Section 5.1, we described the experimental setup by explaining the criteria for selecting 100 YouTube videos and describing the YouTube video dataset features. The evaluation of the soundness of our privacy exposure index (PEI) is presented in Section 5.2. The evaluation of the consistency of our privacy score with the experts' privacy score perception is detailed in Section 5.3. In Section 5.4, we report our evaluation of the functionality of the scoring system using a comments sentiment analysis tool.

5.1 Experimental Setup

5.1.1 YouTube Video Dataset

We collected hundred YouTube videos from ten YouTubers uploaded over a period of two years. In general, the YouTube videos evaluated in our dataset are coming from YouTubers with YouTube channels that have a high volume of subscribers in a range of [250K-10M]. We made this criterion as this group of YouTubers tend to have a higher frequency of content publication required in our two-years analysis. The additional criteria set out below in Table 5.1 were used to locate an overall group of potential YouTubers.

TABLE 5.1: The Criteria for Selecting YouTube Videos for Evaluating the Privacy Score Framework

Criteria for Video Dataset	
Criterion	**Description**
Age	12 to 17 years
Channel Categories	Entertainment, Howto&Style and People&Blogs
Video Type	Vlogs, Makeup and Lifestyle Videos
Video Title Keywords	Tour, Storytime, Vlogs, Routine, Review, Room, House, Makeup, School, Haul, Shopping, Trip, Challenge.

The first criterion for selecting a YouTube video was the age of the YouTubers, we considered a range age between 12-17 years old. We focused exclusively on adolescents, as they are a demographic group that spends most of the day using computers and/or cell phones [56] but more importantly they are a vulnerable population group because of their age.

Our next criterion was the type of YouTube channel categories. Three types of YouTube Channel categories were chosen due to the nature of the videos that YouTubers are uploading on their channels as we found in our privacy behavior and concerns study in Chapter 4. We selected YouTube channels that belongs to the categories: Entertainment, HowtoStyle and PeopleBlogs. These channel categories are exposing more vulnerable information of the YouTuber on YouTube videos according to our finding in Section 4.6.1.

Another criterion were the type of YouTube videos. We analyzed Vlogs videos since this type of video is more frequently uploaded by teenagers. Vlogs are a vulnerable type of YouTube video where the YouTubers record what they do daily and upload these videos and share them with their subscribers. Therefore, this type of videos is more susceptible of disclosing personal information. Also, Makeup and Lifestyle videos were selected as part of the type of videos analyzed due to YouTubers in these type of videos frequently discuss their personal life with their audience.

The last considered criterion for selecting a YouTube video was the title of the YouTube video. We selected titles that contained certain keywords that give us indications or hints that the YouTube video could potentially disclose personal sensitive information. The considered keywords were: *Tour, Storytime, Vlogs, Routine, Review, Room, House, Makeup, School, Haul, Shopping, Trip, Challenge*.

Once the selection criteria have been set up, we proceeded to select 100 YouTube videos. we chose 10 YouTube videos from 10 YouTubers over a period of 2 years as we wanted to determine the variation and differences of the privacy exposure index (PEI) of a YouTuber over time. We wanted also to determine if the YouTuber through the time is exposing more or less personal sensitive information on videos.

5.1.2 YouTube Video Dataset Features

The dataset is structured according to the following categorization: video-identifier features, content-sensitive features and comment-section features.

TABLE 5.2: Description of the Features of the YouTube Dataset

YouTube Dataset Features		
Type of Features	**Description**	**Features**
Video-identifier Features	These features identify each video.	date, channel, url, youtuber-age
Content-sensitive Features	These features determine the amount of sensitive information of each video.	address, name, city, age,hobbies, health, places, school, relationship, fam&friends
Comment-section Features	These features are related to about the comment section.	n_comments, anger, anticipation,disgust, fear, joy, sadness, surprise, trust, negative, positive

1. **Video-identifier Features:** This group of features contains information related to the video as shown in Table 5.3, these features are extracted manually. The following are Video-identifier features:

   - date: It refers to the date of publication of the YouTube video.
   - channel: It refers to the name of the YouTube channel.
   - url: It refers to the video's URL.
   - youtuber-age: It refers to the age of the YouTuber. This feature is frequently found on the channel information.

TABLE 5.3: Features that contain information related to the YouTube video

Video-identifier Features	
Feature	**Description**
date	It refers to the publication of the YouTube video.
channel	It refers to the name of the YouTube channel.
url	It refers to the YouTube video's URL.
youtuber-age	It refers to the age of the YouTuber.

2. **Content-sensitive Features:** This group of features are used for calculating the privacy exposure index PEI and representing the type of sensitive information disclosed on the YouTube video. These features are found in the visual, audio or metadata content as we discussed in Section 3.3. The following content-sensitive features were selected for our analysis:

   - address: It refers to any information disclosed on the video related to the YouTuber's address as we mentioned in Section 3.3.3.
   - name: It refers to any information disclosed on the video related to the YouTuber's name as we mentioned in Section 3.3.2.
   - city: It refers to any information disclosed on the video related to the city where the YouTuber lives as we mentioned in Section 3.3.3
   - age: It refers to any information disclosed on the video related to the YouTuber's age as we mention in Section 3.3.2.

- hobbies: It refers to any information disclosed on the video related to the YouTuber's hobbies/interests as we mentioned in Section 3.3.2.
- health: It refers to any information disclosed on the video related to the YouTuber's health information as we mentioned in Section 3.3.4.
- places: It refers to any information disclosed on the video related to the places such as a mall, stores, or other places where the YouTuber frequently goes as we mentioned in Section 3.3.3.
- school: It refers to any information disclosed on the video related to the YouTuber's school information as we mentioned in Section 3.3.6.
- relationship: It refers to any information disclosed on the video related to the YouTuber's love life as we mentioned in Section 3.3.7.
- fam&friends: It refers to any information disclosed on the video related to the YouTuber's family and friends as we mentioned in Section 3.3.7.

TABLE 5.4: Features that represent the sensitive information on a YouTube video

Content-sensitive Features	
Feature	**Description**
address	It refers to any information disclosed on the video related to the YouTuber's address.
name	It refers to any information disclosed on the video related to the YouTuber's name.
city	It refers to any information disclosed on the video related to the city where the YouTuber lives.
age	It refers to any information disclosed on the video related to the YouTuber's age.
hobbies	It refers to any information disclosed on the video related to the YouTuber's hobbies/interests.
health	It refers to any information disclosed on the video related to the YouTuber's health information.
places	It refers to any information disclosed on the video related to the places where the YouTuber frequently goes.
school	It refers to any information disclosed on the video related to the YouTuber's school information
relationship	It refers to any information disclosed on the video related to the YouTuber's love life
fam&friends	It refers to any information disclosed on the video related to the YouTuber's family and friends

3. **Comment-section Features:** This group of features contains information related to the comment section of a YouTube video. These features are extracted using a sentiment analysis tool. The following are the comment-section features:

   - n_comments: It refers to the number of comments of a YouTube video up to the login date.
   - anger: It refers to the number of comments that indicate anger sentiment.

- anticipation: It refers to the number of comments that indicate anticipation sentiment.
- disgust: It refers to the number of comments that indicate disgust sentiment.
- fear: It refers to the number of comments that indicate fear sentiment.
- joy: It refers to the number of comments that indicate joy sentiment.
- sadness: It refers to the number of comments that indicate sadness sentiment.
- surprise: It refers to the number of comments that indicate sentiment sentiment.
- trust: It refers to the number of comments that indicate trust sentiment.
- negative: It refers to the number of comments that indicate negative sentiment.
- positive: It refers to the number of comments that indicate positive sentiment.

TABLE 5.5: Features that contain information related to the comment section of a YouTube video

Comment-section Features	
Feature	**Description**
n_comments	It refers to the number of comments of a YouTube video up to the login date
anger	It refers to the number of comments that indicate anger sentiment.
anticipation	It refers to the number of comments that indicate anticipation sentiment.
disgust	It refers to the number of comments that indicate disgust sentiment.
fear	It refers to the number of comments that indicate fear sentiment.
joy	It refers to the number of comments that indicate joy sentiment.
sadness	It refers to the number of comments that indicate sadness sentiment.
surprise	It refers to the number of comments that indicate sentiment sentiment.
trust	It refers to the number of comments that indicate trust sentiment.
negative	It refers to the number of comments that indicate negative sentiment.
positive	It refers to the number of comments that indicate positive sentiment.

5.2 Evaluating Soundness of PEI

We evaluated the soundness of the proposed PEI by systematically following the steps below to extract sensitive features, identify the visibility score and finally compute PEI for every single video.

Extraction of Features

We manually extracted the content-sensitive features (e.g., address, name, city, etc.) required to analyze all privacy aspects of a video described in Section 3.3.1 and the features that were identifying a YouTube video as we shown in Table 5.3.

The analysis process of the YouTube video took approximately two hours per video as we meticulously extracted privacy aspects of each video. This process was divided in three phases as described below:

1. *Video-identifier Features Extraction:* The Video-identifier features shown in Table 5.3 were extracted the first time that we watched the YouTube video. The date of publication of the video was found in the metadata. Likewise, the name of the YouTube channel was extracted from the YouTube page but also we could found it in the metadata. Then, the video's URL was extracted. The age of the YouTuber was extracted from the YouTube channel profile. This phase took approximately 20 minutes to complete.

2. *Content-sensitive Features Extraction:* This phase required that the video to be watched/analyzed at least three times in order to carefully extract the sensitive information. If a sensitive feature from Table 5.4 was found on the video, audio or metadata, we proceeded to add the corresponding weight on the YouTube video dataset. This phase took approximately 60-90 minutes to complete.

3. *Comment-section Extraction*: The extraction of the comment-section features shown in Table 5.5 required that we download the comments of a video using the YouTube Data API. Then through a sentiment analysis tool explained in Section 5.4.1, we extracted the sentiment polarities. This phase completed for all videos in our YouTube video dataset and it took approximately 5 minutes to complete for each video.

The extraction of features of a YouTube video is an exhaustive process that requires an understanding of YouTube and the privacy aspects of a YouTube video. In total, we spent 200 hours completing the manual extraction of the features of the YouTube video dataset.

Identifying the Visibility

The Visibility of a content-sensitive feature f_{ij} indicates whether or not the feature is disclosed in the YouTube video. The video is watched entirely and proceeded to look for sensitive information disclosed on the modality of the YouTube video (visual, audio and metadata) and using the weight tables indicated in Section 3.3.1. We assigned to each feature a weight for computing the PEI.

Computing the PEI
Finally, after the video-identifier and content-sensitive extraction was performed, we proceeded to compute the privacy exposure index (PEI) of each video using our proposed framework, as described in Section 3.3. In the manual extraction of the content-sensitive features, we found our developed PEI can sufficiently capture all privacy sensitive information that we observed in these videos. Therefore, we had reasonable ground to consider the privacy exposure index (PEI) a sound formulation.

We understand the limitation of our empirical approach to evaluate the soundness of PEI. Therefore, in the next section, we describe an evaluation on the consistency of our privacy score with the perceived privacy scores assigned by the experts.

5.3 Evaluating Consistency of PEI

In order to check the consistency between the YouTuber's privacy perception and the measured PEI using our proposed privacy framework, we set up an experiment where a group homogeneous of YouTubers measure the privacy risk of a group of YouTube videos from our 100 YouTube videos dataset. This experiment had two phases: the YouTube expert-interview phase and the YouTuber perception phase.

In the first phase, we randomly selected 5 YouTube videos from our 100 YouTube video dataset with different PEI scores stratified in five groups (1.very low PEI, 2.low PEI, 3.medium PEI, 4. high PEI, 5. very high PEI). Then we designed an expert-interview form as we detailed in Appendix E with the links of each of the 5 videos. In the expert-interview form, we included the same content-sensitive features of the YouTube video dataset as we detailed in Section 5.1.1. Finally, at the end of the last column of the expert-interview form, we included the YouTuber's identified privacy risk to each video.

In the second phase, we contacted three YouTubers as experts. This was a homogeneous group with similar skill sets (number of subscribers, channel categories and demographics). We sent the expert-interview form to those YouTubers and we asked them to evaluate the YouTube videos in terms of privacy risks looking for leakage of sensitive features (e.g., address, health). Then, YouTubers provided the identified privacy risk score to each video according to their perception. The expected value of the privacy score for each video was between 1-5 consistent with out PEI strata. Our goal in this experiment was to see if the score that the YouTubers assign to each YouTube video is consistent with our score using our proposed privacy framework.

The results of the expert-interview are detailed in Table 5.6 and summarize the sensitive features found in five YouTube videos. The number of checkmarks represents the number of experts who identified the presence of such sensitive features in the video. For example, in video 1, all experts noted that the name was revealed, while in video 3, two of the experts identified a relationship leakage.

TABLE 5.6: Expert Interview Results of the Sensitive Features

Videos	address	name	city	age	hobby	health	street signs	school	relationship	fam&friends
1		xxx			xxx					x
2		xxx		xxx	xx		x	xxx		x
3		xxx	x		xxx		xx	xxx	xx	xxx
4		xxx	x	xx						
5	xx	xxx	xxx	xxx			xxx			xxx

The perceived risk of each video indicated by the experts is detailed in Table 5.7. The perceived scores of the experts (E1-E3 Score) are consistent with the score obtained from our framework (PEI Score). In addition, we measured the relationship of these results by performing a correlation test between the score from our framework (PEI Score) and the majority score of the experts (MA Score). The results of the Spearman's rank correlation test (r =0.936, $p - value$ =0.019) show that there is a strong correlation between the perceived privacy score of the experts and the privacy score obtained from our framework.

TABLE 5.7: Expert Interview Results of the Identified Privacy Risk

Videos	E1 Score	E2 Score	E3 Score	MA Score	PEI Score
1	1	1	1	1	2
2	3	1	4	3.5	3
3	3	4	1	3.5	4
4	1	1	1	1	1
5	5	4	3	4.5	5

In summary:

1. The high correlation between our systematic scoring system and the perception scoring of the YouTuber indicates that our scoring system is consistent with the scoring of the YouTubers when they are explicitly asked for privacy leakage. This can be explained by the checkmarks of each sensitive feature (e.g., address, city) identified in the video.

2. In addition, the correlation from our systematic scoring system with the perceptional scoring of the experts demonstrate that when YouTubers are aware of the sensitive information on a video, they can recognize the privacy implications associated with the information.

3. The results also confirm the overall finding of our conducted survey where the majority of the YouTubers (82.79%) reported that a privacy tool could be very helpful for their privacy by providing recommendations about the privacy aspects of a video. We observed that expert YouTubers identify the privacy risks when they have been explicitly asked to find sensitive information in a video.

5.4 Evaluating Functionality of PEI for Sentiment Analysis

In the previous Chapter 4, we determined the relationship between the privacy exposure of a YouTube video and the comment section by analyzing the answers of the Viewers and YouTubers. We found evidence that there is a relationship between the amount of sensitive information disclosed on a YouTube video and the comments of the Viewers. Thus, we evaluated the functionality of our framework when a comment sentiment analysis tool is available to be used as a precautionary determinant of privacy exposure.

5.4.1 Sentiment Analysis on Viewers' Comment

As we discussed in Section 4.6.3, the comment section is determining the privacy exposure on YouTube videos. There is evidence that the YouTuber's privacy exposure behavior is influenced by the comments that Viewers leave on videos. Thus, we also analyze the comment section on each of the videos by extracting some features of the comment section of a YouTube video such as the number of comments of videos among others.

The extraction of the comment-section features is performed by a sentiment analysis tool developed by the authors in [17][10] [26]. Our goal was to analyze the opinions and feelings in each comment of a YouTube video. Using the sentiment analysis, we were able to extract positive and negative opinions, emotions, and feelings in the comment section of each of those 100 uploaded videos.

The steps to perform the sentiment analysis is described below. We first downloaded the YouTube comments using the YouTube Data API [118] which is available on Google Developers Console. We downloaded the comments for each video in a CSV format. Then, for the sentiment analysis, we used a library in Python called VADER (Valence Aware Dictionary and sEntiment Reasoner) [79]. This is a lexicon and rule-based sentiment analysis tool and is built to work for sentiment analysis in social media texts [42] [67]. Finally, we extracted the number of comments with positive sentiment and negative sentiment separately as described in our comment-section features in Table 5.5. We also extracted other sentiment polarities such as anger, anticipation, etc.. However, we decided not to analyse other polarities as we could not collect evidence through our survey for the impact of other polarities on the privacy behaviour of Youtubers. To avoid overwhelm survey participants, we limited the number of question to focus more on the privacy behaviour determinants and included only one question related to the influence of negative and positive comments (Q24).

5.4.2 Relationship Between Comment Section and PEI

In order to determine how the comment section is affecting the PEI over time. We calculated the percentage of negative and positive comments with respect to the total number of comments. Then, we proceeded to accumulate the values of PEI and percentage of negative and positive comments within the time period of data collection (2 years). These cumulative values were analyzed using their rate of changes.

Since the leakage of sensitive information of a YouTube video could impact the YouTuber's privacy over time (e.g., months to years), the concept of accumulated privacy exposure or privacy budget is considered in this YouTube analysis. The privacy budget of a YouTuber is the privacy exposure index accumulated from each video uploaded by a YouTuber in a period of time. For example, Alice uploads her first video v_1 in her channel in time t_1. Thus, the PEI and comments of the video v_1 correspond to the initial accumulated privacy of Alice. Then, when Alice uploads a second video v_2 in time t_2, the PEI and comments of the new video v_2 are accumulated. This process updates the accumulated privacy of Alice that already had the leakage amount of sensitive information from the first video.

In order to determine the influence of the comments on the accumulated PEI. We calculated the percentage of the negative and positive comments of each video with respect to the total number of comments. Then, we accumulated the values of the PEI and percentage of negative and positive comments for each period of time. These cumulative values were normalized due to the large variation of the scales of these variables. The results of these steps are summarized in Figure 5.1, which shows the normalized cumulative PEI and negative and positive comments for each YouTuber. Additionally, we computed the point-to-point average of the normalized cumulative values obtained for all YouTubers, shown in Figure 5.2.

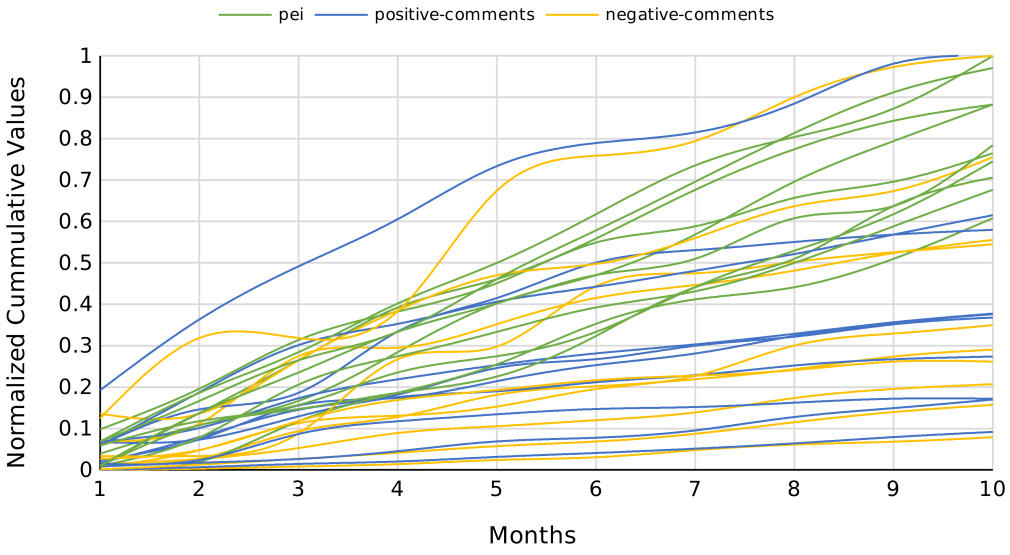


FIGURE 5.1: Cumulative Values of PEI, Positive and Negative Comments Percentage of each YouTuber

As we can observe from Figure 5.2, all three variables are increasing as we expect because they are cumulative values, but in particular, the trend of comments shows that the lines representing positive and negative comments are crossing each other, which may suggest that negative comments are slightly more impactful to encourage YouTubers to expose more personal information over time.

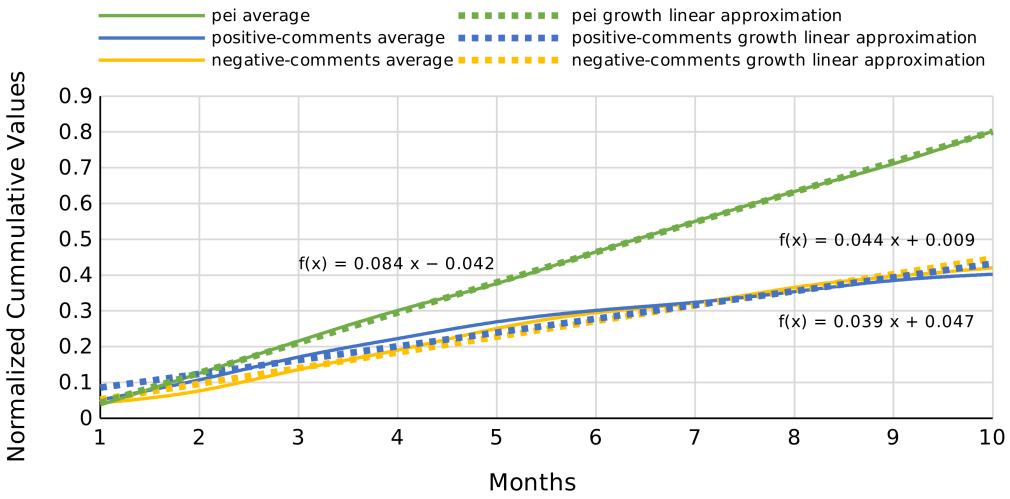


FIGURE 5.2: Average Values of PEI, Positive and Negative Comments Percentage and Trend lines

5.5 Summary

In this chapter, we evaluated the PEI in terms of its soundness, consistency, and functionality. To complete the evaluation in terms of soundness, we analyzed a group of 100 YouTube videos over a period of time of two years for measuring the privacy exposure index (PEI) of each video. Then, in order to evaluate the consistency of our privacy scoring system, we set up an experiment where three expert YouTubers measured the perceived privacy risk of a group of video from the 100 YouTube videos previously analyzed. Finally, we evaluated the functionality of our framework using a sentiment analysis tool as a precautionary determinant of privacy exposure.

Chapter 6

Conclusions

In this chapter, we summarize the thesis major contributions, some recommendations and discussions for future work.

6.1 Summary of Contributions

This thesis aimed to measure privacy risks on YouTube videos by analyzing the leakage of sensitive information. Our major contributions in this research work were threefold. We developed a privacy risk framework for YouTube videos inspired by the existing frameworks that quantify privacy risks in other social networks. Then, we conducted a YouTube study for identifying the privacy behavior and concerns of YouTube users in order to determine the factors affecting the privacy exposure of the YouTubers. Finally, we provided a YouTube video dataset (as an open source with create common copyright) to evaluate the practicality of the proposed framework for measuring the privacy risk of an individual video and the accumulative privacy exposure index across the timeline of a YouTuber content generation.

In Chapter 3, we described the YouTube platform and its privacy aspects, we found that YouTube as a social network has several privacy implications. Then based on the existing methods of quantifying privacy scores on social networks, we developed a privacy framework for measuring privacy risks on YouTube videos. Our privacy risk score that quantifies the risks of a YouTube video is called privacy exposure index (PEI). PEI is a privacy score of each YouTube video that helps YouTubers to determine if and to what extent a video have privacy implications for the YouTuber.

In Chapter 4, we studied the YouTuber's privacy behavior and concerns for two reasons. First to determine the factors affecting the privacy exposure of YouTubers. Second, to see whether the privacy concerns of YouTubers are consistent with their privacy behavior when they upload a video. As part of our contribution, we conducted a survey targeted the YouTube Viewers as well, to evaluate the influence of the comment section of a YouTube video on the privacy exposure index (PEI). We concluded this chapter with an interesting finding from the results of the YouTuber study that most of our YouTuber participants (82.79%) reported that a YouTube privacy tool that provides them with some recommendations about their privacy aspects on the YouTube video could be very helpful for their privacy.

In Chapter 5 we evaluated the soundness, consistency and functionality of the proposed framework. We produced a rich dataset of 100 YouTube videos with their computed PEIs. This dataset contains three types of features of a YouTube video to measure the PEI of a video and the influence of the comment section of a YouTube

channel throughout the time. We found that the privacy exposure index (PEI) of the videos of a YouTuber is increasing over time and the negative comments of the comment section is slightly more impactful to encourage YouTubers to expose more personal information over time.

6.2 Recommendations

Based on the findings of this thesis, we make the following recommendations to address the YouTubers' privacy challenges. When a YouTuber uploads a video to YouTube, there is no tool that verifies if the video has privacy implications for the YouTuber. We found that YouTube privacy policies are lacking boundaries that allow YouTubers to know the privacy implications of the information that they are disclosing on videos. Therefore, a Youtube-oriented privacy policy recommender tool could be beneficial for YouTubers to protect their privacy. In our conducted survey as indicated in Appendix A we found that most of our participants (82.79%) independently of the age, gender, channel categories or number of subscribers reported that a YouTube privacy tool that may provide them recommendations about their privacy could be helpful. They consider a privacy tool necessary for them to protect their privacy.

Consequently, the implementation of a privacy tool should consider two privacy aspects: a YouTuber's privacy guide and a privacy quantifying tool. The YouTuber's privacy guide would help YouTuber to self-evaluate their YouTube videos before uploading to the YouTube platform. The privacy quantifying tool would predict the privacy risks of a YouTube video.

The YouTuber's privacy guide must consider asking YouTubers what type of videos they are uploading (e.g., vlogs, makeup tutorial, haul). In this research work, we found evidences that Vlogs videos contain more sensitive information than others types of YouTube videos. Thus, the YouTuber's privacy guide for self-evaluation should include the type of content (e.g., vlogs, makeup tutorials, hauls) criterion.

Other findings of this thesis that speaks to the design of a privacy tool for YouTubers are explained below. First, the engagement differences have to be considered. In Chapter 4, we found that YouTube channels with a small number of subscribers are exposing more personal information. Since YouTubers with less experience in the platform are more vulnerable to expose sensitive information, we suggest that a privacy tool have to consider YouTube channels with a small number of subscribers as well as YouTubers with less experience in the platform as the target group. Another aspect to consider is the comment section of the YouTube videos. The comment section is influencing the YouTuber's privacy exposure as we have seen in Chapter 4 and 5. Therefore it should also be considered as a factor influencing the YouTuber's privacy exposure behavior. The last aspect to consider for quantifying the PEI of a YouTube video is the privacy exposure index accumulated from other YouTube videos uploaded previously as we have seen in Section 5.4.

6.3 Future Work

The major future work would be to develop an automated tool for computing in an efficient manner the privacy exposure index (PEI) of a YouTube video. The tasks

of object recognition and detection have to be implemented in order to automatically detect the features that are considered as sensitive on the video modality of the YouTube videos. The author of this thesis [81] has found that extracting features on YouTube videos is possible and made several video data sets available with sensitive feature indexed. The future research can exploit this dataset to train a model of automated or semi-automated detection of these sensitive features. We found the current object detection techniques in videos are far from maturity to identify sensitive information in a video. Therefore a research in this direction will enhance privacy research. However, we realize that complete this research need of an extensive amount of resources because the accuracy of video feature extraction is not mature enough and also in order to automatically compute the PEI. Future research can also include automated analysis of other YouTube modalities such as audio, metadata and comment section for detecting sensitive information.

In addition the current method of quantifying the privacy exposure index (PEI) can benefit if it can be evaluated in the further studies with a more diverse and larger group of YouTubers and reviewers.

Appendix A

Survey Questionnaire

YouTube Privacy Perception & Behavior

There are 38 questions in this survey.

SURVEY CRITERIA

∗ 1 To participate in this survey you must meet the following criteria: Be over 18, speak fluent English or Spanish and be a YouTube user.

Please choose **only one** of the following:

- ◯ Yes, I meet the criteria.
- ◯ No, I do not meet the criteria.

This survey is administered by Vanessa Calero, Department of Computer and Software. The purpose of the survey is to understand Privacy Behavior and Perception from YouTube's user. Information gathered during this survey will be written up as a thesis. What we learn from this survey will help us understand YouTube's privacy risk.

To learn more about the survey and the researcher's study, particularly in terms of any associated risks or harms associated with the survey, how confidentiality and anonymity will be handled, withdrawal procedures, incentives that are promised, how to obtain information about the survey's results, how to find helpful resources should the survey make you uncomfortable or upset, etc., please read the accompanying letter of information.

This survey should take approximately 8 minutes to complete. People filling out this survey must be *18 years of age or older*.

This survey is part of a study that has been reviewed and cleared by the McMaster Research Ethics Board (http://reo.mcmaster.ca/) (MREB). The MREB protocol number associated with this survey is 3635.
You are free to complete this survey or not. If you have any concerns or questions about your rights as a participant or about the way the study is being conducted, please contact:
McMaster Research Ethics Secretariat
Telephone 1-(905) 525-9140 ext. 23142
C/o Research Office for Administration, Development and Support (ROADS)
E-mail: ethicsoffice@mcmaster.ca (mailto:ethicsoffice@mcmaster.ca)

∗ 2 Consent to participate

Having read the above, I understand that by clicking the "Yes" button below, I agree to take part in this study under the terms and conditions outlined in the accompanied letter of information.

Please choose **only one** of the following:

- ◯ Yes, I agree to participate
- ◯ No, I do not agree to participate

Demographic Questions

This group of questions is intended to collect demographic information about YouTube viewers

∗ 3 What is your gender?

Please choose **only one** of the following:

- ◯ Male
- ◯ Female
- ◯ Others
- ◯ Prefer not to disclose

∗ 4 What is your age?

Please choose **only one** of the following:

- ◯ 18-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55-64
- ◯ 65+
- ◯ do not want to disclose

∗ 5 Which country do you live?

Please choose **only one** of the following:

∗ 6 What is your highest level of education?

Please choose **only one** of the following:

- ◯ Some High School
- ◯ Graduated from High School
- ◯ Graduated from College
- ◯ Some Graduated School
- ◯ Completed Graduated School
- ◯ do not want to disclose

Type of User

There is two types of participants: Subscriber and YouTuber.

7 *What type of YouTube's user are you ?*

>YouTuber (refers to a person who creates content on YouTube).

>Viewer (refers to a person who watches the content on YouTube Channels and can be a subscriber to the channel or watch it anonymously).

Please choose **only one** of the following:

- ◯ YouTuber
- ◯ Viewer

Youtube Usage

These questions are about YouTube's usage statistics

∗ 8 As a YouTube viewer, how would you describe yourself?

> participative: I always like and comment on the video.

> uninvolved: I only watch the video without interaction.

Please choose **only one** of the following:

- ◯ Uninvolved
- ◯ Somewhat uninvolved
- ◯ Neither uninvolved or participative
- ◯ Somewhat participative
- ◯ Participative

∗ 9 How many Vlogs videos do you watch per day?

Please choose **only one** of the following:

- ◯ 0
- ◯ Less than 2
- ◯ 2-5
- ◯ 5-10
- ◯ More than 10

∗ 10 In a regular week, how many videos do you comment?

Please choose **only one** of the following:

- ◯ 0
- ◯ Less than 2
- ◯ 2-5
- ◯ 5-10
- ◯ More than 10

* 11 How many videos do you watch per day?

Please choose **only one** of the following:

- ◯ less than 5
- ◯ 5-10
- ◯ 10-15
- ◯ more than 15

*12 What type of YouTube Channels do you prefer to watch?

Please choose **all** that apply:

- ☐ Autos and Vehicles
- ☐ Comedy
- ☐ Education
- ☐ Entertaiment
- ☐ Film and Animation
- ☐ Gaming
- ☐ Howto and Style
- ☐ Music
- ☐ News and Politics
- ☐ Nonprofits and Activism
- ☐ People and Blogs
- ☐ Pets and Animals
- ☐ Science and Technology
- ☐ Sports
- ☐ Travels and Events

Comments associated with Privacy Exposure

These questions intend to answer our first research question.

* 13 Complete the sentence as best describes your interests: When I comment on a video, typically I like to know___

Please choose **only one** of the following:

- ◯ More about the content
- ◯ Only about the content
- ◯ More about the YouTuber
- ◯ Only about the YouTuber
- ◯ Equally both
- ◯ This question doesn't apply to me

*14 In the past 6 months, how many Vlogs have been uploaded by your favorite YouTuber?

Please choose **only one** of the following:

- ◯ 0
- ◯ Less than 5
- ◯ 5-10
- ◯ 10-15
- ◯ 15-20
- ◯ More than 20

* 15 In which type of videos do you comment on more frequently?

Please choose **only one** of the following:

- ◯ Vlogs
- ◯ Makeup Tutorial videos
- ◯ Gaming videos
- ◯ Tag or Challenge videos
- ◯ Product Review videos
- ◯ Comedy videos
- ◯ Unboxing/Haul videos
- ◯ Educational videos
- ◯ Others
- ◯ I do not comment on any video

*16 Express your agreement with the following sentence: The most likable videos are the ones where the YouTuber speaks about her/his personal life and issues?

Please choose **only one** of the following:

- ◯ Strongly Agree
- ◯ Agree
- ◯ Undecided / Neutral
- ◯ Disagree
- ◯ Strongly Disagree

*17 In the past 6 months, how often have you found location-sensitive (e.g. addresses, streets, neighborhood names) leakage on YouTube videos?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

Youtube Creation

These questions are intended to ask YouTubers about the statistics of the uploaded videos

* 18 How many subscribers do you have?

Please choose **only one** of the following:

- ◯ Less than 10K
- ◯ 10K-100K
- ◯ 100K- 500K
- ◯ 500K- 1M
- ◯ More than 1M

* 19 Which of the following categories does your YouTube channel belong to?

Please choose **only one** of the following:

- ◯ Autos and Vehicles
- ◯ Comedy
- ◯ Education
- ◯ Entertaiment
- ◯ Film and Animation
- ◯ Gaming
- ◯ Howto and Style
- ◯ Music
- ◯ News and Politics
- ◯ Nonprofits and Activism
- ◯ People and Blogs
- ◯ Pets and Animals
- ◯ Science and Technology
- ◯ Sports
- ◯ Travels and Events

* 20 How many videos do you post per week?

Please choose **only one** of the following:

- ◯ Less than 2
- ◯ 2-3
- ◯ 4-5
- ◯ More than 5

* 21 How many times do you review the final edited video before uploading it?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

* 22 Typically, how long after a video is completed do you upload it on the YouTube platform?

Please choose **only one** of the following:

- ◯ Immediately
- ◯ 1-2 days
- ◯ 3-4 days
- ◯ 5-6 days
- ◯ After one week
- ◯ Other

* 23 In the past 6 months, how many vlogs or Q/A type of videos have you upload on your YouTube Channel per month?

Please choose **only one** of the following:

- ◯ None
- ◯ Less than 2
- ◯ 2-3
- ◯ 3-4
- ◯ More than 5

Comments associated with YouTuber Privacy Exposure

These questions are intended to answer our first research question.

* 24 Which type of comments do you usually reply to?

Please choose **all** that apply:

- ☐ Positive Comments related to the content of the video
- ☐ Negative Comments related to the content of the video
- ☐ Positive Comments related to the video/audio quality
- ☐ Negative Comments related to the video/audio quality
- ☐ Positive Comments related to my appearance
- ☐ Negative Comments related to my appearance
- ☐ I usually do not reply comments

* 25 How often does your audience feedback influence the content that you upload to YouTube?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

* 26 How often does your audience through your section comments requesting your love life information?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

YouTuber Privacy Perception and Behavior

These questions are intended to answer our second and third research questions

* 27 Are you willing to share your health information with your audience on your videos?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

* 28 Do you record video clips on places around your home or the place your live or work?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

∗29 Do you speak about your spouse, girlfriend or boyfriend on your videos?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

∗30 If you are making a Vlog video outside your home, which places do you usually go to? ∗

Please choose **all** that apply:

- ☐ Other family member's home
- ☐ Other friend's home
- ☐ The nearest mall to my home
- ☐ My favorite places that are near my home
- ☐ My favorite places that are not near my home
- ☐ None of the above apply

∗31 Are you willing to share your salary, wealth and other financial information with your audience?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

∗32 In the past 6 months, how many videos have included topics about your health?

Please choose **only one** of the following:

- ◯ 0
- ◯ Less than 5
- ◯ 5-10
- ◯ 10-15
- ◯ More than 15

∗33 How often are your exposing your neighborhood and street signs around your home address on your YouTube videos?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

∗34 In the past 6 months, have you shared information about buying a property (e.g. a car) with your audience?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

∗35 Express your agreement with the following sentence: I am comfortable speaking about my private life on my videos?

Please choose **only one** of the following:

- ◯ Strongly Agree
- ◯ Agree
- ◯ Undecided / Neutral
- ◯ Disagree
- ◯ Strongly Disagree

∗36 Before uploading a video to YouTube, do you always check for private information?

Please choose **only one** of the following:

- ◯ Always
- ◯ Often
- ◯ Sometimes
- ◯ Seldom
- ◯ Never

∗37 How do you rate yourself with respect to your privacy?

>Conservative: I always give priority to my privacy when I'm uploading a video.

>Easy going: Privacy is not my primary concern when I'm uploading a video.

Please choose **only one** of the following:

- ◯ Conservative
- ◯ Somewhat conservative
- ◯ Neither conservative nor easy going
- ◯ Somewhat easy going
- ◯ Easy going

∗38 Express your agreement with the following sentence: A tool that gives you recommendations about the privacy aspects of your videos could be helpful?

Please choose **only one** of the following:

- ◯ Strongly Agree
- ◯ Agree
- ◯ Undecided / Neutral
- ◯ Disagree
- ◯ Strongly Disagree

Thank you for taking this survey. Your answers are a valuable part of this research.

Appendix B

Descriptive Analysis Results of the Questionnaire

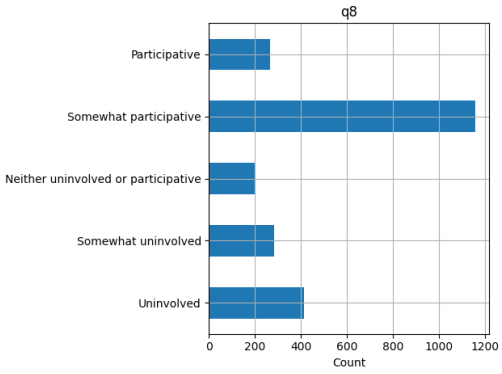


FIGURE B.1: Survey question 8

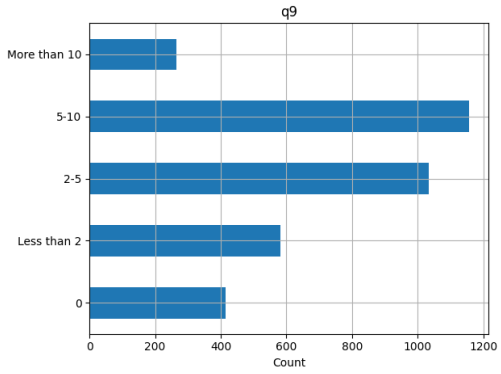


FIGURE B.2: Survey question 9

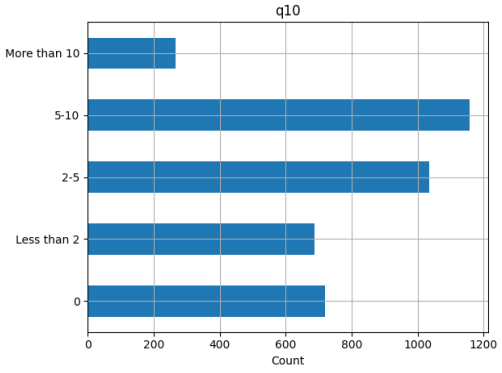


FIGURE B.3: Survey question 10

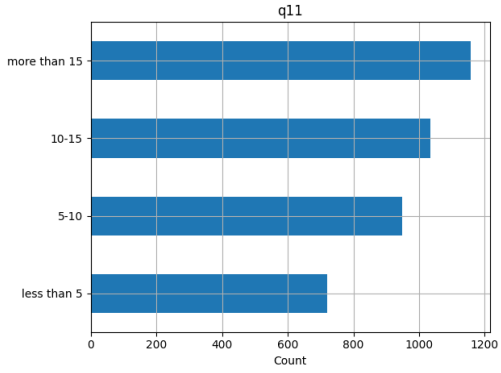


FIGURE B.4: Survey question 11

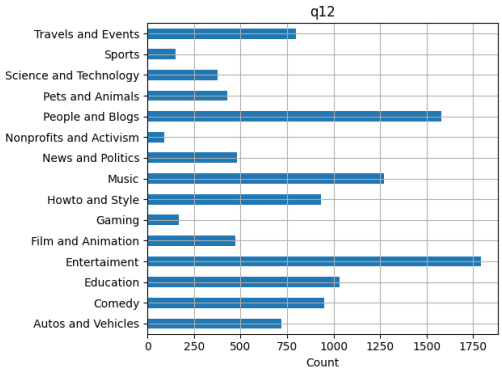


FIGURE B.5: Survey question 12

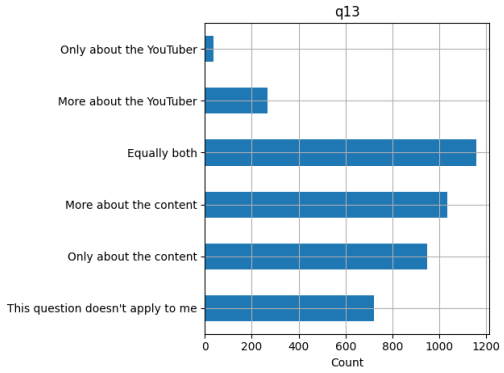


FIGURE B.6: Survey question 13

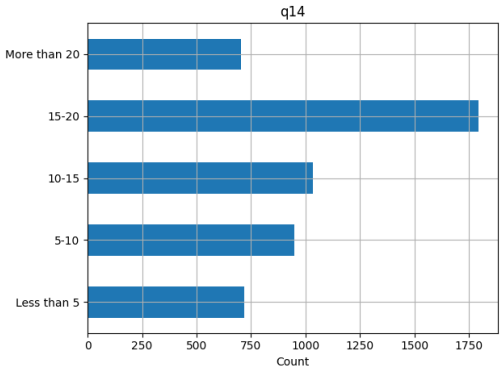


FIGURE B.7: Survey question 14

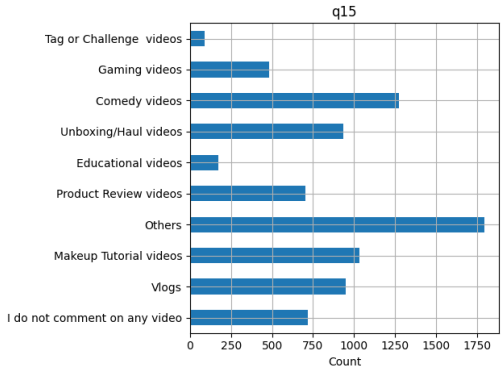


FIGURE B.8: Survey question 15

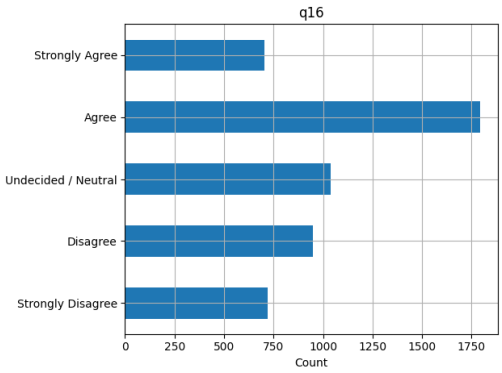


FIGURE B.9: Survey question 16

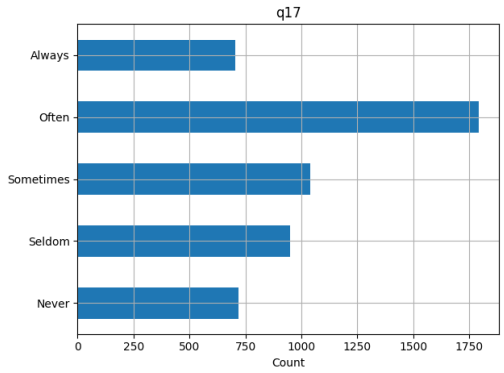


FIGURE B.10: Survey question 17

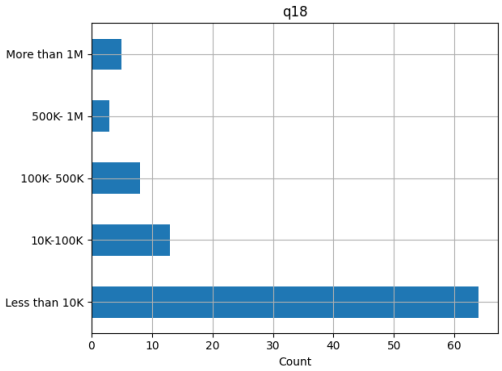


FIGURE B.11: Survey question 18

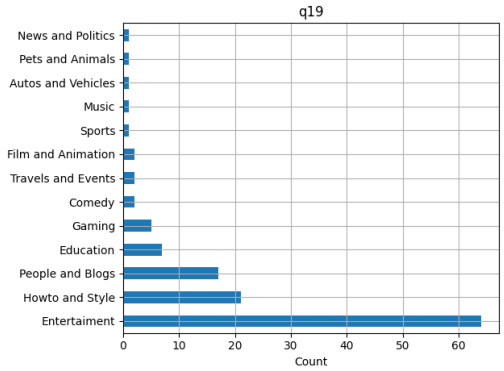


FIGURE B.12: Survey question 19

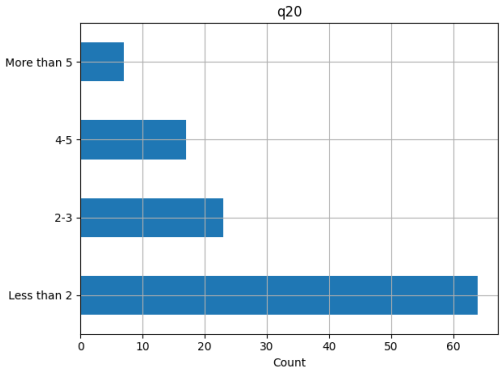


FIGURE B.13: Survey question 20

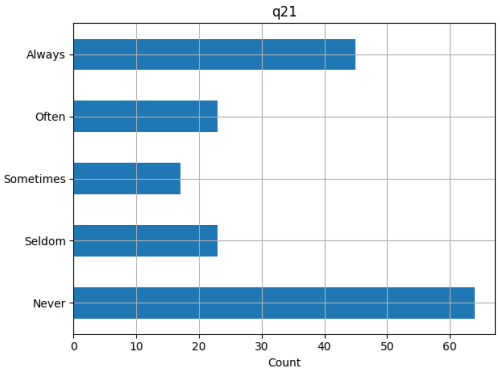


FIGURE B.14: Survey question 21

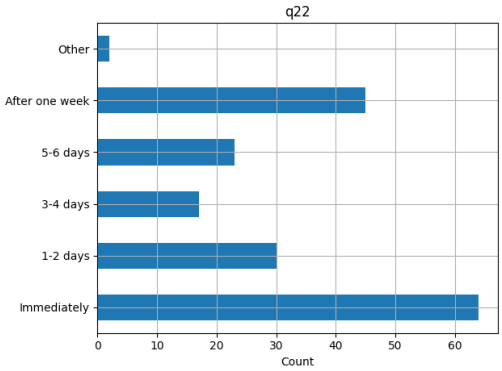


FIGURE B.15: Survey question 22

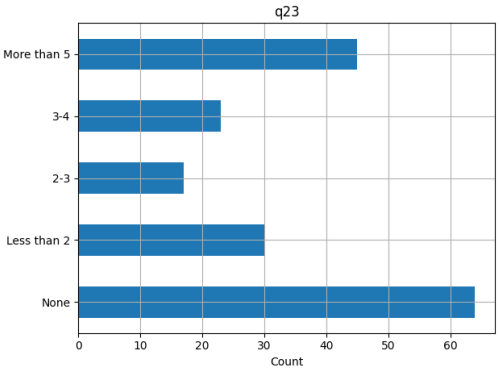


FIGURE B.16: Survey question 23

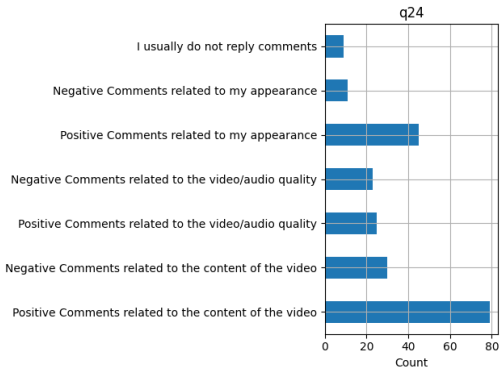


FIGURE B.17: Survey question 24

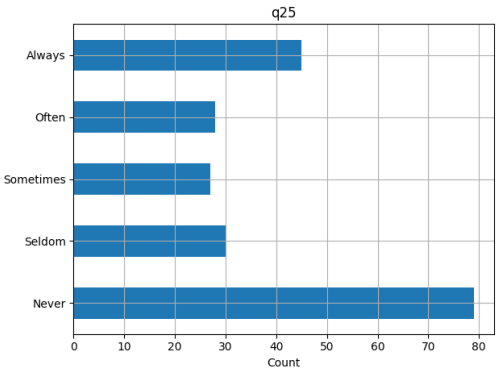


FIGURE B.18: Survey question 25

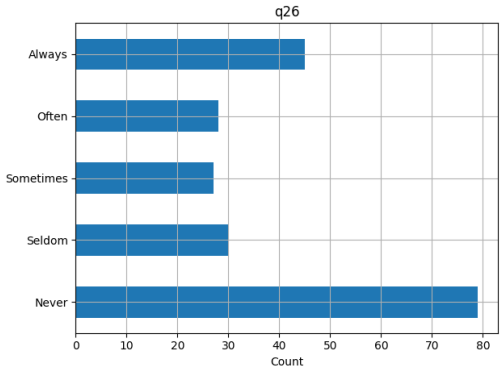


FIGURE B.19: Survey question 26

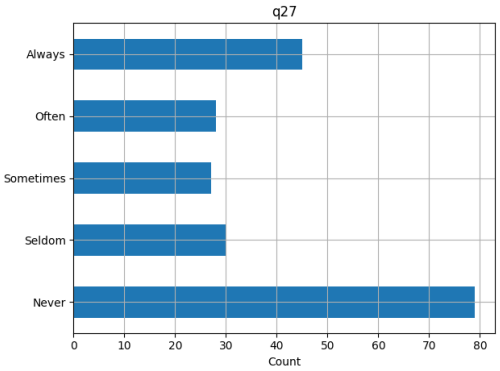


FIGURE B.20: Survey question 27

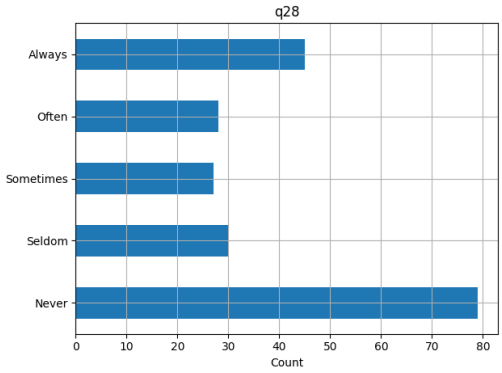


FIGURE B.21: Survey question 28

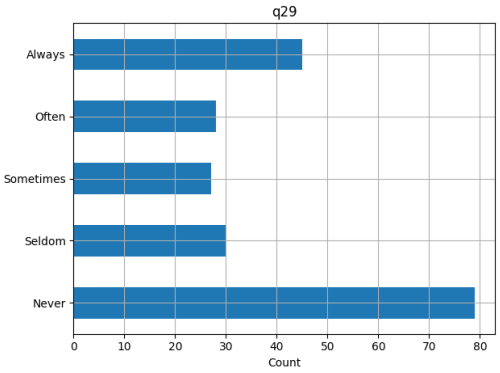


FIGURE B.22: Survey question 29

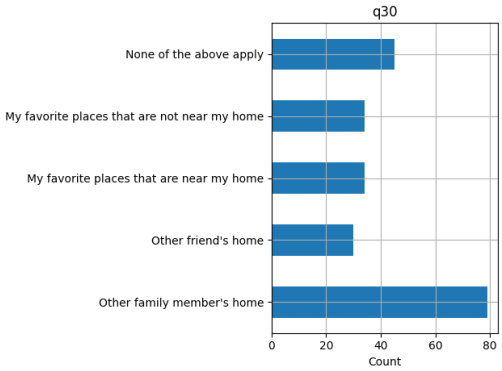


FIGURE B.23: Survey question 30

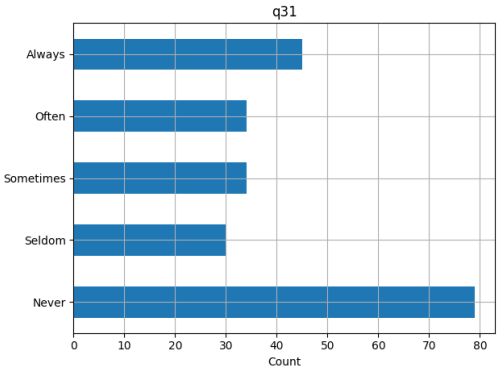


FIGURE B.24: Survey question 31

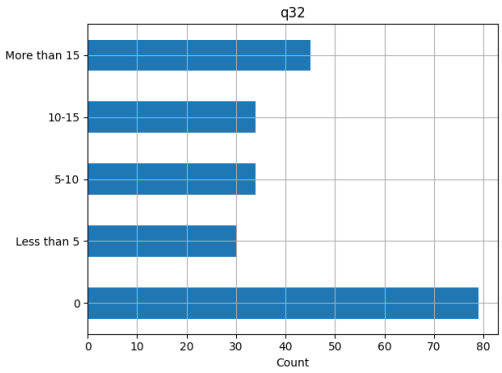


FIGURE B.25: Survey question 32

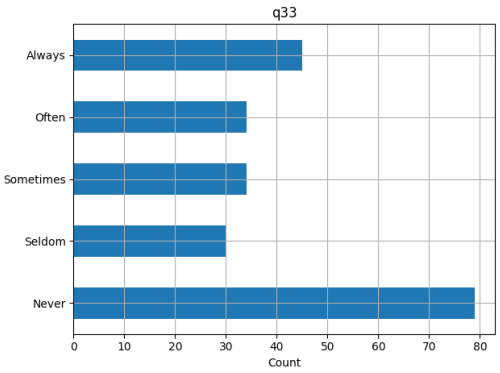


FIGURE B.26: Survey question 33

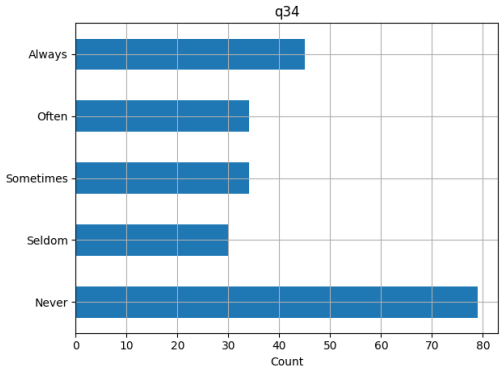


FIGURE B.27: Survey question 34

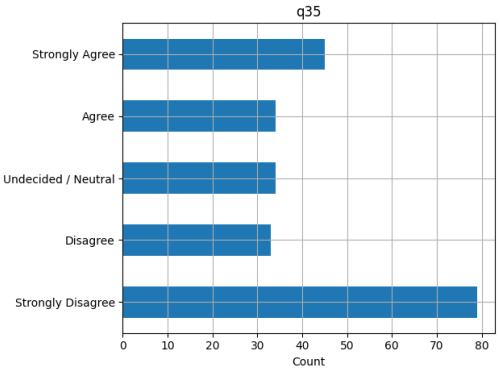


FIGURE B.28: Survey question 35

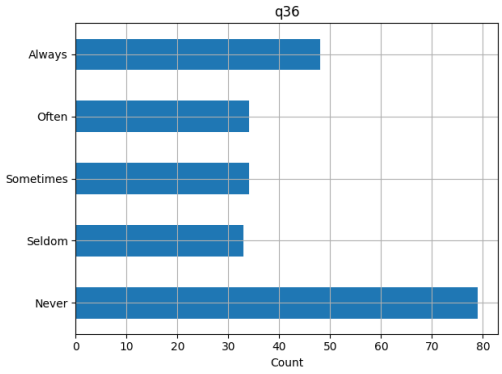


FIGURE B.29: Survey question 36

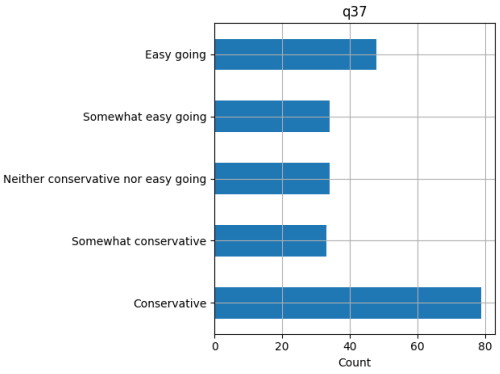


FIGURE B.30: Survey question 37

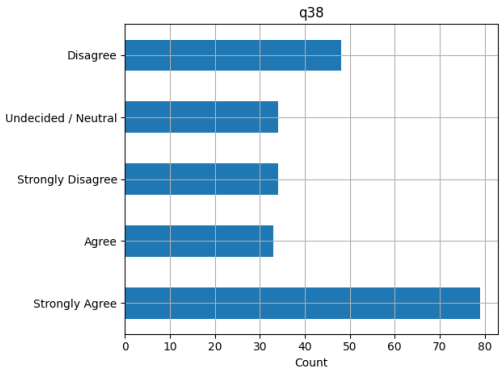


FIGURE B.31: Survey question 38

Appendix C

Survey Validity and Reliability

TABLE C.1: Survey Cross-validation Questions

Construct	Question	Cross-validation Question
YouTuber's Privacy Behavior (YPB)	Q 27 - Are you willing to share your health information with your audience on your videos?	Q 32 - In the past 6 months, how many videos have included topics about your health?
	Q 31 - Are you willing to share your salary, wealth and other financial information with your audience?	Q 34 - In the past 6 months, have you shared information about buying a property (e.g. a car or house) with your audience?

TABLE C.2: Survey Reliability Questions

Construct	Question	Cronbach's Alpha
YouTuber's Privacy Behavior (YPB)	Q 27, Q 28, Q 29, Q 31, Q 33	0.71

Appendix D

Details of the Statistical Analysis

D.1 Checking Assumptions of Normality

The $p - values$ of the Shapiro-Wilk tests for each construct of YouTubers and Viewers as indicated in Table D.1 and D.2 show that we do not have enough evidence to conclude that our data do not follow a normal distribution.

TABLE D.1: Normality test results for the YouTuber's Construct

Construct		Score	P-Value
YED	Q18	0.615	0.000
YCC	Q19a (Entertainment)	0.605	0.000
	Q19b (Howto)	0.516	0.000
	Q19c (People)	0.470	0.000
	Q19d (Education)	0.290	0.000
	Q19e (Other)	0.442	0.000
YCGB	Q20	0.627	0.000
	Q21	0.790	0.000
	Q22	0.760	0.000
	Q23	0.790	0.000
YCI	PCI	0.775	0.000
	Q26	0.875	0.000
YSrB	Q25	0.900	0.000
	Q35	0.878	0.000
YPB	PEI	0.766	0.000
YPC	Q37	0.766	0.000

TABLE D.2: Normality test results for the Viewer's construct

Construct		Score	P-Value
VCI	Q8	0.811	0.000
VUB	Q9	0.840	0.000
	Q11	0.864	0.000
VCB	Q10	0.696	0.000
VCP	Q12a (Entertainment)	0.517	0.000
	Q12b (Howto)	0.622	0.000
	Q12c (People)	0.587	0.000
	Q12d (Education)	0.616	0.000
	Q12e (Other)	0.479	0.000
VIB	Q13	0.878	0.000
YCB	Q14	0.841	0.000
VVB	CVI	0.850	0.000
VPD	Q16	0.903	0.000

D.2 Correlation Matrix for YouTuber Privacy Behavior

Given the results of the normality test suggest that the distributions of our constructs are non-normal, we performed a correlation test using a non-parametric statistic such as the Spearman's rank correlation. We created the Spearman correlation matrix for all dependent and independent variables. Results are presented in Table D.3, D.4, D.5, D.6, D.7, D.8 and D.9.

TABLE D.3: Correlation Test for H1

Constructs	YED	YPB
YED	1.0	
YPB	-0.387	1.0

*Not significant at 0.05 level (2-tail)

TABLE D.4: Correlation Test for H3

Constructs	YCC					YPB
YCC	1.0					
	-0.401	1.0				
	-0.351	-0.255	1.0			
	-0.212	-0.154*	-0.135*	1.0		
	-0.325	-0.237	-0.207	-0.125*	1.0	
YPB	0.115*	-0.086*	0.244	-0.224	-0.147*	1.0

*Not significant at 0.05 level (2-tail)

TABLE D.5: Correlation Test for H4

Constructs	YCI		YPB
YCI	1.0 -0.250	1.0	
YPB	-0.195*	0.385	1.0

*Not significant at 0.05 level (2-tail)

TABLE D.6: Correlation Test for H5

Constructs	YCGB				YPB
YCGB	1.0 0.136* -0.073* -0.394	1.0 -0.182* -0.176*	1.0 0.344		
YPB	0.199*	0.223	-0.379	-0.535	1.0

*Not significant at 0.05 level (2-tail)

TABLE D.7: Correlation Test for H6a

Constructs	YSrB (Q36)	YPB (Q27)
YSrB (Q36)	1.0	
YPB (Q27)	0.324	1.0

*Not significant at 0.05 level (2-tail)

TABLE D.8: Correlation Test for H6b

Constructs	YPC (Q37)	YPB (Q27)
YPC (Q37)	1.0	
YPB (Q27)	0.318	1.0

*Not significant at 0.05 level (2-tail)

TABLE D.9: Correlation Test for H6c

Constructs	YPC (Q37)	YSrB (Q36)
YPC (Q37)	1.0	
YSrB (Q36)	0.144*	1.0

*Not significant at 0.05 level (2-tail)

D.3 Demographic Differences and Privacy Behaviour

TABLE D.10: Kruskal-Wallis H1-Test: H2

Construct		**Score**	**P-Value**
YDD	Q3	21.999	0.000
	Q4	17.836	0.000
	Q5	6.408	0.171
	Q6	6.111	0.191

D.4 Correlation Matrix for Viewers' Privacy Influence

We perform a set of correlation test between all the independent and dependent variables to investigate the dependence between the variables of each hypothesis of our Viewers' research model. Results of the Spearman's rank-order correlation are found in Tables D.11D.12 D.13 D.14 D.15 D.16 D.17D.18 D.19.

TABLE D.11: Correlation Test for H7

Constructs	VCP					VCI
VCP	1.000 0.054 -0.009 0.097 0.203	1.000 0.078 0.070 0.132	1.000 0.035 0.114	1.000 0.237	1.000	
VCI	*-0.029	*-0.027	-0.057	-0.084	-0.078	1.000

* Not significant at 0.05 level (2-tail)

TABLE D.12: Correlation Test for H8

Constructs	VUB		VCI
VUB	1.000 0.303	1.000	
VCI	0.192	0.134	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.13: Correlation Test for H9-a

Constructs	VVP	VCB
VVP	1.000	
VCB	-0.381	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.14: Correlation Test for H9-b

Constructs	VCB	VCI
VCB	1.000	
VCI	0.512	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.15: Correlation Test for H9-c

Constructs	VVP	VCI
VVP	1.000	
VCI	-0.511	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.16: Correlation Test for H11

Constructs	VIB	VCI
VIB	1.000	
VCI	0.376	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.17: Correlation Test for H12a

Constructs	YCB	VPD
YCB	1.000	
VPD	0.069	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.18: Correlation Test for H12b

Constructs	YCB	VIB
YCB	1.000	
VIB	0.033	1.000

*Significant at 0.00001 level (2-tail)

TABLE D.19: Correlation Test for H12c

Constructs	VPD	VIB
VPD	1.000	
VIB	0.128	1.000

*Significant at 0.00001 level (2-tail)

D.5 Demographic Differences and Viewer's Interaction Behavior

TABLE D.20: Kruskal-Wallis H1-Test: H10

Construct		**Score**	**P-Value**
YDD	Q3	6.730	3.456e-02
	Q4	46.682	2.165e-08
	Q5	34.779	7.175e-02
	Q6	32.848	4.0342e-06

Appendix E

Expert Interview

E.1 Expert Interview Form

The expert interview aim to evaluate the privacy risks perception on YouTube videos by asking three YouTubers to evaluate 5 YouTube videos in terms of privacy risk in order to determine the consistency between the YouTuber's privacy perception and the calculated PEI of each video.

The expert-interview form provides of a list of sensitive features (e.g., address, name,city, etc.) that the expert have to look for throughout the entire YouTube video (video, audio and metadata). The last column contains the privacy risk score perception assigned by the experts.

Hola amigos,

Espero que estén teniendo un excelente comienzo de semana.
Como muchos de ustedes sabrán, estoy trabajando en una investigación de privacidad que busca calcular riesgos de privacidad en videos de YouTube, y como parte de la investigación para el ultimo capitulo de mi tesis, necesito su valiosa ayuda y experiencia, ustedes al ser YouTubers conocen o tienen una idea de los peligros de exponer información personal en un video. Yo he elegido al azar 5 videos que están revelando cierta información personal en diferente grados de riesgos.

Si tu observas que el video revela algunas de las informaciones indicadas en la tabla de abajo, por favor marcalo con una X. Por ejemplo, yo completé la primera columna en donde indico que el video que estoy viendo revela la dirección del YouTuber, alguna información sobre la salud del YouTuber y algo sobre su vida amorosa y sentimental. Yo dejé las otras columnas vacias porque en mi video EXAMPLE no encontré que se revelarán esas informaciones personales en el video. Al final de la tabla en la ultima columna, por favor asigna a cada video un numero entre el 1 y 5. Selecciona 1 si consideras que el video tiene un riesgo de privacidad bajo, 5 si el riesgo es alto y 2-4 para valores que esten entre estos valores. Por ejemplo para mi video EXAMPLE yo le he dado un puntaje de 3 en el riesgo de privacidad.

Tema: YouTubers expertos califican videos en terminos de privacidad

#	Enlace	direccion	nombre	ciudad	edad	hobby	salud	señales de calle	escuela / trabajo	relacion amorosa	familia / amigos	tu riesgo de privacidad (1-5)
EXAMPLE	EXAMPLE	X					X			X		3
Video 1	https://goo.gl/PSkIG8											
Video 2	https://goo.gl/JXBtHQ											
Video 3	https://goo.gl/yjV66x											
Video 4	https://goo.gl/Bf6xks											
Video 5	https://goo.gl/CeLJQi											

Marca con una X la información que aparece en el video

Hi friends,

I hope you are having a great weekend.
As all of you already know, I am working on a privacy risk score system for YouTube video, and as part of my research, I need to know regarding your expertise on YouTube which of these following 5 videos are disclosing personal information such as an address, health information, relationship, etc. Also, I would like to know how you would rank these videos in terms of privacy.

If you observe the video reveals any of the items on top of the table, please simply put an X there. For example, I completed the first row indicating that my example video reveals address of the youtuber, some information about her health and some information about the relationship. I left the other columns empty.

At the end for the last column, please assign to each video a number between 1 to 5. Please select 1 very low privacy risk, 5 very high privacy risk and 2-4 something in between that you feel appropriately reflects the privacy risk of the video. For example for my video I have selected score 3 for my observed video.

Subject: Expert YouTubers ranking videos in terms of privacy

#	Link	address	name	city	age	hobbies	health	Street signs	School / Job	Relationship	Family / Friends	Your identified Privacy Risk (1-5)
EXAMPLE	EXAMPLE	X					X			X		3
Video 1	https://goo.gl/PSklG8											
Video 2	https://goo.gl/JXBtHQ											
Video 3	https://goo.gl/yjV66x											
Video 4	https://goo.gl/Bf6xks											
Video 5	https://goo.gl/CeLJQi											

Mark with a X the information that appears in the video

E.2 Expert Interview Answers

YouTube Dataset												
#	link	address	name	city	age	hobby	health	street signs	school	relationship	family/friends	pei
Video 1	https://goo.gl/PSklG8		x	x	x	x						4
Video 2	https://goo.gl/JXBtHQ		x	x	x	x					x	7
Video 3	https://goo.gl/yjV66x		x	x	x	x		x			x	10
Video 4	https://goo.gl/Bf6xks		x		x							2
Video 5	https://goo.gl/CeLJQi		x	x	x	x	x	x			x	13

YOUTUBER 1: Abril Martinez												
#	link	address	name	city	age	hobby	health	street signs	school	relationship	family/friends	privacy risk identified
Video 1	https://goo.gl/PSklG8		x			x						1
Video 2	https://goo.gl/JXBtHQ		x		x	x			x			3
Video 3	https://goo.gl/yjV66x		x			x			x	x	x	3
Video 4	https://goo.gl/Bf6xks		x									1
Video 5	https://goo.gl/CeLJQi	x	x	x	x			x			x	5

YOUTUBER 2: Mirianny de los Santos												
#	link	address	name	city	age	hobby	health	street signs	school	relationship	family/friends	privacy risk identified
Video 1	https://goo.gl/PSklG8		x			x					x	1
Video 2	https://goo.gl/JXBtHQ		x		x				x			1
Video 3	https://goo.gl/yjV66x		x			x		x	x		x	4
Video 4	https://goo.gl/Bf6xks		x		x							1
Video 5	https://goo.gl/CeLJQi	x	x	x	x			x			x	4

YOUTUBER 3: Paul Lando												
#	link	address	name	city	age	hobby	health	street signs	school	relationship	family/friends	privacy risk identified
Video 1	https://goo.gl/PSklG8		x			x						1
Video 2	https://goo.gl/JXBtHQ		x		x	x		x	x		x	4
Video 3	https://goo.gl/yjV66x		x	x		x		x	x	x	x	5
Video 4	https://goo.gl/Bf6xks		x	x	x							1
Video 5	https://goo.gl/CeLJQi		x	x	x			x			x	3

Bibliography

[1] M. S. Ackerman. "Privacy in pervasive environments: next generation labeling protocols". In: *Personal and Ubiquitous Computing* 8.6 (2004), pp. 430–439.

[2] A. Acquisti, I. Adjerid, and L. Brandimarte. "Gone in 15 seconds: The limits of privacy transparency and control". In: *IEEE Security & Privacy* 11.4 (2013), pp. 72–74.

[3] A. Acquisti, S. Gritzalis, C. Lambrinoudakis, and S. di Vimercati. *Digital privacy: theory, technologies, and practices*. CRC Press, 2007.

[4] A. Acquisti and J. Grossklags. "Privacy and rationality in individual decision making". In: *IEEE security & privacy* 3.1 (2005), pp. 26–33.

[5] A. Acquisti and J. Grossklags. "Privacy attitudes and privacy behavior". In: *Economics of information security*. Springer, 2004, pp. 165–178.

[6] K. Adlassnig, L. Fernandez, and N. Elahi. "An analysis of personal medical information disclosed in YouTube videos created by patients with multiple sclerosis". In: *Medical Informatics in a United and Healthy Europe: Proceedings of MIE 2009, the XXII International Congress of the European Federation for Medical Informatics*. Vol. 150. Los Press. 2009, p. 292.

[7] N. Aggarwal, S. Agrawal, and A. Sureka. "Mining YouTube metadata for detecting privacy invading harassment and misdemeanor videos". In: *2014 Twelfth Annual International Conference on Privacy, Security and Trust*. IEEE. 2014, pp. 84–93.

[8] E. Aghasian, S. Garg, L. Gao, S. Yu, and J. Montgomery. "Scoring users' privacy disclosure across multiple online social networks". In: *IEEE access* 5 (2017), pp. 13118–13130.

[9] A. Al Hasib. "Threats of online social networks". In: *IJCSNS International Journal of Computer Science and Network Security* 9.11 (2009), pp. 288–93.

[10] M. Z. Asghar, S. Ahmad, A. Marwat, and F. M. Kundi. "Sentiment analysis on YouTube: A brief survey". In: *arXiv preprint arXiv:1511.09142* (2015).

[11] M. Bärtl. "YouTube channels, uploads and views: A statistical analysis of the past 10 years". In: *Convergence* 24.1 (2018), pp. 16–32.

[12] J. L. Becker. *Measuring privacy risk in online social networks*. University of California, Davis, 2009.

[13] F. Belanger, J. S. Hiller, and W. J. Smith. "Trustworthiness in electronic commerce: the role of privacy, security, and site attributes". In: *The journal of strategic Information Systems* 11.3-4 (2002), pp. 245–270.

[14] V. Bellotti and A. Sellen. "Design for privacy in ubiquitous computing environments". In: *Proceedings of the Third European Conference on Computer-Supported Cooperative Work 13–17 September 1993, Milan, Italy ECSCW'93*. Springer. 1993, pp. 77–92.

[15] V. Benson, G. Saridakis, and H. Tennakoon. "Information disclosure of social media users". In: *Information Technology & People* (2015).

[16] B. Berendt, O. Günther, and S. Spiekermann. "Privacy in e-commerce: stated preferences vs. actual behavior". In: *Communications of the ACM* 48.4 (2005), pp. 101–106.

[17] H. Bhuiyan, J. Ara, R. Bardhan, and M. R. Islam. "Retrieving YouTube video by sentiment analysis on user comment". In: *Signal and Image Processing Applications (ICSIPA), 2017 IEEE International Conference on*. IEEE. 2017, pp. 474–478.

[18] J.-I. Biel and D. Gatica-Perez. "Call me Guru: user categories and large-scale behavior in YouTube". In: *Social Media Modeling and Computing*. Springer, 2011, pp. 167–188.

[19] D. Boyd. "Facebook's privacy trainwreck: Exposure, invasion, and social convergence". In: *Convergence* 14.1 (2008), pp. 13–20.

[20] L. E. Buffardi and W. K. Campbell. "Narcissism and social networking web sites". In: *Personality and social psychology bulletin* 34.10 (2008), pp. 1303–1314.

[21] V. Calero. *VaneVane Channel*. 2020. URL: https://www.youtube.com/user/VaneVaneFabulosity (visited on 05/06/2020).

[22] M. Caryn. *The Dark Side of Being a YouTuber Channel*. 2019. URL: https://www.youtube.com/watch?v=A6N5DhRLtH4 (visited on 05/06/2019).

[23] A. Cavoukian. "Privacy by design". In: *Take the challenge. Information and privacy commissioner of Ontario, Canada* (2009).

[24] B. J. Chambers and S. L. Bichard. "Public opinion on YouTube: A functional theory analysis of the frames employed in user comments following Sarah Palin's 2008 acceptance speech". In: *International Journal of E-Politics (IJEP)* 3.2 (2012), pp. 1–15.

[25] C. F. Choi and Z. Jiang. "Trading friendship for value: An investigation of collective privacy concerns in social application usage". In: (2013).

[26] A. A. L. Cunha, M. C. Costa, and M. A. C. Pacheco. "Sentiment Analysis of YouTube Video Comments Using Deep Neural Networks". In: *International Conference on Artificial Intelligence and Soft Computing*. Springer. 2019, pp. 561–570.

[27] L. A. Cutillo, R. Molva, and T. Strufe. "Safebook: A privacy-preserving online social network leveraging on real-life trust". In: *IEEE Communications Magazine* 47.12 (2009), pp. 94–101.

[28] S. J. De and A. Imine. "Privacy Scoring of Social Network User Profiles through Risk Analysis". In: *International Conference on Risks and Security of Internet and Systems*. Springer. 2017, pp. 227–243.

[29] S. Du, X. Li, J. Zhong, L. Zhou, M. Xue, H. Zhu, and L. Sun. "Modeling privacy leakage risks in large-scale social networks". In: *IEEE Access* 6 (2018), pp. 17653–17665.

[30] N. B. Ellison, C. Steinfield, and C. Lampe. "The benefits of Facebook "friends:" Social capital and college students' use of online social network sites". In: *Journal of computer-mediated communication* 12.4 (2007), pp. 1143–1168.

[31] Z. Epifani. *YouTube Home Tour | vivo da sola*. 2020. URL: https://www.youtube.com/watch?v=prH1rIZ7iOw (visited on 05/31/2020).

[32] S. Fischer-Hübner. "Privacy and security at risk in the global information society". In: *Information Communication & Society* 1.4 (1998), pp. 420–441.

[33] S. Fischer-Hübner. "Privacy-Enhancing Technologies (PET)". In: *Course Description, Karlstad University Division for Information Technology, Karlstad, Sweden* (2001).

[34] B. Fu, E. Chi, P. Cao, R. Yang, and S. Singh. "Video WatchTime and Comment Sentiment: Experience from YouTube". In: (2016).

[35] K. Fulpagare Priya and N. N. Patil. "RESOLVING PRIVACY CONFLICT FOR MAINTAINING PRIVACY POLICIES IN ONLINE SOCIAL NETWORKS". In: *Journal of Computer Engineering and Technology* 10.3 (2019), pp. 94–101.

[36] V. Gabriella. *CAUGHT DOING COKE IN PUBLIC (STORYTIME W/LIVE FOOTAGE)*. 2016. URL: https://www.youtube.com/watch?v=ql300P0yg5s (visited on 05/21/2017).

[37] G. Garmendia. *This has to stop... please*. 2017. URL: https://www.youtube.com/watch?v=cfmrMKX0G6k (visited on 03/26/2017).

[38] R. Gross and A. Acquisti. "Information revelation and privacy in online social networks". In: *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. ACM. 2005, pp. 71–80.

[39] A. Heravi, S. Mubarak, and K.-K. R. Choo. "Information privacy in online social networks: Uses and gratification perspective". In: *Computers in Human Behavior* 84 (2018), pp. 441–459.

[40] A. Ho, A. Maiga, and E. Aımeur. "Privacy protection issues in social networking sites". In: *Computer Systems and Applications, 2009. AICCSA 2009. IEEE/ACS International Conference on*. IEEE. 2009, pp. 271–278.

[41] W. Hoiles, A. Aprem, and V. Krishnamurthy. "Engagement and popularity dynamics of youtube videos and sensitivity to meta-data". In: *IEEE Transactions on Knowledge and Data Engineering* 29.7 (2017), pp. 1426–1437.

[42] C. J. Hutto and E. Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text". In: *Eighth international AAAI conference on weblogs and social media*. 2014.

[43] G. Iachello and J. Hong. "End-user privacy in human-computer interaction". In: *Foundations and Trends in Human-Computer Interaction* 1.1 (2007), pp. 1–137.

[44] J. L. Jensen and A. S. Sørensen. ""Nobody has 257 friends": Strategies of friending, disclosure and privacy on Facebook". In: *Nordicom Review* 34.1 (2013), pp. 49–62.

[45] Z. Jiang, C. S. Heng, and B. C. Choi. "Research note—privacy concerns and privacy-protective behavior in synchronous online social interactions". In: *Information Systems Research* 24.3 (2013), pp. 579–595.

[46] L. K. John, A. Acquisti, and G. Loewenstein. "Strangers on a plane: Context-dependent willingness to divulge sensitive information". In: *Journal of consumer research* 37.5 (2011), pp. 858–873.

[47] A. N. Joinson and C. B. Paine. "Self-disclosure, privacy and the Internet". In: *The Oxford handbook of Internet psychology* 2374252 (2007).

[48] A. N. Joinson, U.-D. Reips, T. Buchanan, and C. B. P. Schofield. "Privacy, trust, and self-disclosure online". In: *Human–Computer Interaction* 25.1 (2010), pp. 1–24.

[49] M. Kandias, V. Stavrou, N. Bozovic, and D. Gritzalis. "Proactive insider threat detection through social media: The YouTube case". In: *Proceedings of the 12th ACM workshop on Workshop on privacy in the electronic society*. 2013, pp. 261–266.

[50] M. J. Keith, S. C. Thompson, J. Hale, P. B. Lowry, and C. Greer. "Information disclosure on mobile devices: Re-examining privacy calculus with actual user behavior". In: *International journal of human-computer studies* 71.12 (2013), pp. 1163–1173.

[51] M. L. Khan. "Social media engagement: What motivates user participation and consumption on YouTube?" In: *Computers in Human Behavior* 66 (2017), pp. 236–247.

[52] H. Krasnova, E. Kolesnikova, and O. Guenther. "" It won't happen to me!": self-disclosure in online social networks". In: (2009).

[53] H. Krasnova, S. Spiekermann, K. Koroleva, and T. Hildebrand. "Online social networks: Why we disclose". In: *Journal of information technology* 25.2 (2010), pp. 109–125.

[54] C. A. Lampe, N. Ellison, and C. Steinfield. "A familiar face (book) profile elements as signals in an online social network". In: *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2007, pp. 435–444.

[55] P. G. Lange. "Publicly private and privately public: Social networking on YouTube". In: *Journal of computer-mediated communication* 13.1 (2007), pp. 361–380.

[56] A. Lenhart, K. Purcell, A. Smith, and K. Zickuhr. "Social Media & Mobile Internet Use among Teens and Young Adults. Millennials." In: *Pew internet & American life project* (2010).

[57] R. Lennon, R. W. Rentfro, and J. M. Curran. "Exploring relationships between demographic variables and social networking use". In: *Journal of Management and Marketing Research* 11 (2012), p. 1.

[58] S. P. Lewis, N. L. Heath, M. J. Sornberger, and A. E. Arbuthnott. "Helpful or harmful? An examination of viewers' responses to nonsuicidal self-injury videos on YouTube". In: *Journal of Adolescent Health* 51.4 (2012), pp. 380–385.

[59] J. Li. "Privacy policies for health social networking sites". In: *Journal of the American Medical Informatics Association* 20.4 (2013), pp. 704–707.

[60] K. Liu and E. Terzi. "A Framework for Computing the Privacy Scores of Users in Online Social Networks". In: *Proceedings of the 2009 Ninth IEEE International Conference on Data Mining*. ICDM '09. Washington, DC, USA: IEEE Computer Society, 2009, pp. 288–297. ISBN: 978-0-7695-3895-2. DOI: 10.1109/ICDM.2009.21. URL: https://doi.org/10.1109/ICDM.2009.21.

[61] K. Liu and E. Terzi. "A framework for computing the privacy scores of users in online social networks". In: *ACM Transactions on Knowledge Discovery from Data (TKDD)* 5.1 (2010), p. 6.

[62] K. Liu and E. Terzi. "Towards identity anonymization on graphs". In: *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 2008, pp. 93–106.

[63] A. Madden, I. Ruthven, and D. McMenemy. "A classification scheme for content analyses of YouTube video comments". In: *Journal of documentation* (2013).

[64] M. Maia, J. Almeida, and V. Almeida. "Identifying user behavior in online social networks". In: *Proceedings of the 1st workshop on Social network systems*. ACM. 2008, pp. 1–6.

[65] D. Maity and M. Racat. "The Role of Audience Comments in YouTube Vlogs: An Abstract". In: *Academy of Marketing Science Annual Conference*. Springer. 2018, pp. 551–551.

[66] N. K. Malhotra, S. S. Kim, and J. Agarwal. "Internet users' information privacy concerns (IUIPC): The construct, the scale, and a causal model". In: *Information systems research* 15.4 (2004), pp. 336–355.

[67] H. Malik and Z. Tian. "A framework for collecting youtube meta-data". In: *Procedia computer science* 113 (2017), pp. 194–201.

[68] E. M. Maximilien, T. Grandison, T. Sun, D. Richardson, S. Guo, and K. Liu. "Privacy-as-a-service: Models, algorithms, and results on the facebook platform". In: *Proceedings of Web*. Vol. 2. 2009.

[69] T. Mendel and E. Toch. "Susceptibility to social influence of privacy behaviors: Peer versus authoritative sources". In: *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. 2017, pp. 581–593.

[70] E. D. Miller. *Content Analysis of YouTube Comments From Differing Videos: An Overview and Key Methodological Considerations*. SAGE Publications Ltd, 2018.

[71] M. H. Millham and D. Atkin. "Managing the virtual boundaries: Online social networks, disclosure, and privacy behaviors". In: *New Media & Society* 20.1 (2018), pp. 50–67.

[72] R. K. Nepali and Y. Wang. "Sonet: A social network model for privacy monitoring and ranking". In: *2013 IEEE 33rd International Conference on Distributed Computing Systems Workshops*. IEEE. 2013, pp. 162–166.

[73] J. W. Patchin and S. Hinduja. "Bullies move beyond the schoolyard: A preliminary look at cyberbullying". In: *Youth violence and juvenile justice* 4.2 (2006), pp. 148–169.

[74] R. G. Pensa and G. Di Blasi. "A privacy self-assessment framework for online social networks". In: *Expert Systems with Applications* 86 (2017), pp. 18–31.

[75] R. G. Pensa, G. Di Blasi, and L. Bioglio. "Network-aware privacy risk estimation in online social networks". In: *Social Network Analysis and Mining* 9.1 (2019), p. 15.

[76] PewDiePie. *YouTube My comment section is the worst on YouTube LWIAY 0099*. 2019. URL: https://www.youtube.com/watch?v=prH1rIZ7iOw (visited on 11/30/2019).

[77] E. Pilkington. *Blackmail claim stirs fears over Facebook*. 2017. URL: http:%20//www.guardian.co.uk/international/%20story/0,,2127084,00.html (visited on 09/02/2017).

[78] E. Poché, N. Jha, G. Williams, J. Staten, M. Vesper, and A. Mahmoud. "Analyzing user comments on YouTube coding tutorial videos". In: *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*. IEEE. 2017, pp. 196–206.

[79] Python. *Python Library VADER*. 2016. URL: https://pypi.org/project/vaderSentiment/ (visited on 01/12/2019).

[80] M. R. Randazzo, M. Keeney, E. Kowalski, D. Cappelli, and A. Moore. *Insider threat study: Illicit cyber activity in the banking and finance sector*. Tech. rep. DTIC Document, 2005.

[81] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. "Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 5296–5305.

[82] N. J. Rifon, R. LaRose, and S. M. Choi. "Your privacy is sealed: Effects of web privacy seals on trust and personal disclosures". In: *Journal of Consumer Affairs* 39.2 (2005), pp. 339–362.

[83] E. Roberts. *Kim Kardashian West robbery: Everything we know so far*. 2016. URL: http://edition.cnn.com/2016/10/03/entertainment/kardashian-west-robbery-what-we-know/ (visited on 03/14/2017).

[84] Y. Rodriguez. *Un Suscriptor me acosa*. 2017. URL: https://www.youtube.com/watch?v=RKBdONBUnqU (visited on 05/21/2017).

[85] R. Samavi and M. P. Consens. "Publishing L2TAP Logs to Facilitate Transparency and Accountability." In: *LDOW*. 2014.

[86] R. Samavi and M. P. Consens. "Publishing privacy logs to facilitate transparency and accountability". In: *Journal of Web Semantics* 50 (2018), pp. 1–20.

[87] R. Samavi, M. P. Consens, and M. Chignell. "PHR user privacy concerns and behaviours". In: *Procedia Computer Science* 37 (2014), pp. 517–524.

[88] R. Samavi and T. Topaloglou. "Designing privacy-aware personal health record systems". In: *International Conference on Conceptual Modeling*. Springer. 2008, pp. 12–21.

[89] J. Schrammel, C. Köffel, and M. Tscheligi. "How much do you tell? Information disclosure behaviour indifferent types of online communities". In: *Proceedings of the fourth international conference on Communities and technologies*. 2009, pp. 275–284.

[90] P. Shah, E. T. Loiacono, and H. Ren. "Video Blogs: A Qualitative and Quantitative Inquiry of Recall and Willingness to Share". In: *International Conference on Social Computing and Social Media*. Springer. 2017, pp. 234–243.

[91] K. B. Sheehan. "An investigation of gender differences in on-line privacy concerns and resultant behaviors". In: *Journal of Interactive Marketing* 13.4 (1999), pp. 24–38.

[92] K. B. Sheehan and M. G. Hoy. "Dimensions of privacy concern among online consumers". In: *Journal of public policy & marketing* 19.1 (2000), pp. 62–73.

[93] Y. Shen and S. Pearson. "Privacy enhancing technologies: A review". In: *HP Laboratories* 2739 (2011), pp. 1–30.

[94] J. Shibchurn and X. Yan. "Information disclosure on social networking sites: An intrinsic–extrinsic motivation perspective". In: *Computers in Human Behavior* 44 (2015), pp. 103–117.

[95] R. Shokri, G. Theodorakopoulos, J.-Y. Le Boudec, and J.-P. Hubaux. "Quantifying location privacy". In: *2011 IEEE symposium on security and privacy*. IEEE. 2011, pp. 247–262.

[96] S. Siersdorfer, S. Chelaru, W. Nejdl, and J. San Pedro. "How useful are your comments? Analyzing and predicting YouTube comments and comment ratings". In: *Proceedings of the 19th international conference on World wide web*. 2010, pp. 891–900.

[97] H. J. Smith, T. Dinev, and H. Xu. "Information privacy research: an interdisciplinary review". In: *MIS quarterly* 35.4 (2011), pp. 989–1016.

[98] H. J. Smith, S. J. Milberg, and S. J. Burke. "Information privacy: measuring individuals' concerns about organizational practices". In: *MIS quarterly* (1996), pp. 167–196.

[99] H. Smith, T. Dinev, and H. Xu. "Information Privacy Research: An Interdisciplinary Review". In: *MIS Quarterly* 35 (Dec. 2011), pp. 989–1015. DOI: 10.2307/41409970.

[100] J. Smith. *Sharing intimate moments on YouTube: Women who vlog and their sense of community, friendship and privacy*. Gonzaga University, 2012.

[101] A. Srivastava and G. Geethakumari. "Measuring privacy leaks in online social networks". In: *Advances in Computing, Communications and Informatics (ICACCI), 2013 International Conference on*. IEEE. 2013, pp. 2095–2100.

[102] J. Staddon, D. Huffaker, L. Brown, and A. Sedley. "Are privacy concerns a turn-off? Engagement and privacy in social networks". In: *Proceedings of the eighth symposium on usable privacy and security*. 2012, pp. 1–13.

[103] Statista. *YouTube Age Group Statistics*. 2019. URL: https://www.statista.com/statistics/296227/us-youtube-reach-age-gender/ (visited on 01/13/2020).

[104] Statista. *YouTube Percentage of U.S. internet users who use YouTube as of 3rd quarter 2019, by gender*. 2019. URL: https://www.statista.com/statistics/810461/us-youtube-reach-gender/ (visited on 05/16/2020).

[105] Statista. *YouTube Top YouTube content categories*. 2018. URL: https://www.statista.com/statistics/1026914/global-distribution-youtube-video-content-by-category/ (visited on 05/16/2020).

[106] L. Sweeney. "Achieving k-anonymity privacy protection using generalization and suppression". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 571–588.

[107] L. Sweeney. "k-anonymity: A model for protecting privacy". In: *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems* 10.05 (2002), pp. 557–570.

[108] K. Talvitie-Lamberg. "Confessions in social media: performative, constrained, authentic and participatory self-representations in vlogs". In: *Publications of the Department of Social Research* 10 (2014).

[109] M. Thelwall, P. Sud, and F. Vis. "Commenting on YouTube videos: From Guatemalan rock to el big bang". In: *Journal of the American Society for Information Science and Technology* 63.3 (2012), pp. 616–629.

[110] ThinkwithGoogle. *YouTube Education Level of YouTube Audience*. 2016. URL: https://www.thinkwithgoogle.com/data/education-level-of-youtube-audience/ (visited on 06/16/2020).

[111] K. C. Toth, A. Cavoukian, and A. Anderson-Priddy. "Privacy by Design Architecture Composed of Identity Agents Decentralizing Control over Digital Identity". In: *Open Identity Summit 2020* (2020).

[112] S. Valenzuela, N. Park, and K. F. Kee. "Is There Social Capital in a Social Network Site?: Facebook Use and College Students' Life Satisfaction, Trust, and Participation1". In: *Journal of Computer-Mediated Communication* 14.4 (July 2009), pp. 875–901. ISSN: 1083-6101. DOI: 10.1111/j.1083-6101.2009.01474.x. eprint: https://academic.oup.com/jcmc/article-pdf/14/4/875/22317792/jjcmcom0875.pdf. URL: https://doi.org/10.1111/j.1083-6101.2009.01474.x.

[113] P. M. Valkenburg, J. Peter, and A. P. Schouten. "Friend networking sites and their relationship to adolescents' well-being and social self-esteem". In: *CyberPsychology & Behavior* 9.5 (2006), pp. 584–590.

[114] J. Warmbrodt, H. Sheng, R. Hall, and J. Cao. "Understanding the video bloggers' community". In: *Technical, Social, and Legal Issues in Virtual Communities: Emerging Environments*. IGI Global, 2012, pp. 63–79.

[115] D. J. Welbourne and W. J. Grant. "Science communication on YouTube: Factors that affect channel and video popularity". In: *Public Understanding of Science* 25.6 (2016), pp. 706–718.

[116] L. Xie. "Subscriber Number is Not Everything: YouTube Community Engagement Measurement". PhD thesis. Middle Tennessee State University, 2016.

[117] H. Xu, T. Dinev, H. J. Smith, and P. Hart. "Examining the formation of individual's privacy concerns: Toward an integrative view". In: *ICIS 2008 proceedings* (2008), p. 6.

[118] YouTube. *Comments YouTube API developer*. 2014. URL: https://developers.google.com/youtube/v3/docs/comments (visited on 01/02/2019).

[119] YouTube. *YouTube Global YouTube Audience Statistics*. 2020. URL: https://www.statista.com/statistics/805656/number-youtube-viewers-worldwide/ (visited on 05/16/2020).

[120] YouTube. *Youtube privacysettings*. 2017. URL: https://support.google.com/youtube/answer/157177?co=GENIE.Platform%3DDesktop&hl=en (visited on 03/14/2017).

[121] YouTube. *Youtube Statistics*. 2017. URL: https://www.youtube.com/yt/press/statistics.html (visited on 03/14/2017).

[122] YouTube. *YoutubeUsers Statistics*. 2017. URL: https://support.google.com/youtube/answer/9314416 (visited on 10/01/2019).

[123] C. Zhang, J. Sun, X. Zhu, and Y. Fang. "Privacy and security for online social networks: challenges and opportunities". In: *IEEE Network* 24.4 (2010).

[124] N. Zhang, C. Wang, and Y. Xu. "Privacy in online social networks". In: (2011).

[125] E. Zheleva and L. Getoor. "Privacy in social networks: A survey". In: *Social network data analytics*. Springer, 2011, pp. 277–306.