Profiling professional and regular users on popular Internet services based on implementation of large scale Internet measurement tools

Reza Farahbakhsh

Doctor of Philosophy (PhD) Thesis
Institut Mines Telecom, Telecom SudParis
and Université Pierre & Marie Curie - Sorbonne Universités

Specialization

Computer Networks

presented by

Reza Farahbakhsh

Profiling professional and regular users on popular Internet services based on Implementation of large scale Internet measurement tools

21^{th} May 2015
Thesis no: 2015TELE0012

Committee:

David Hausheer	Reviewer	Professor, TU Darmstadt, Germany
Danny De Vleeschauwer	Reviewer	Assoc. Professor/Senior Researcher, UGent/Alcatel-Lucent - Belgium
Pierre Sens	Examiner	Professor, UPMC - Sorbonne Universités - France
Xiaoming Fu	Examiner	Professor, University of Goettingen - Germany
Nishanth Sastry	Examiner	Professor, Kings College London - UK
Jean-Michel Portugal	Examiner	Area Director, Orange Labs - France
Leandros A. Maglaras	Invited	Assoc. Professor, De Montfort University - UK
Noël Crespi	Advisor	Professor, Institut Mines Telecom/Telecom SudParis - France
Ángel Cuevas	Co-Advisor	Assoc. Professor, Universidad Carlos III de Madrid - Spain

Thèse de Doctorat (PhD) de
Institut Mines Telecom, Telecom SudParis et l'université Pierre & Marie Curie - Sorbonne Universités

Spécialité

SYSTÈMES INFORMATIQUES

présentée par

Reza Farahbakhsh

Profilage d'usagers professionnels et non-professionnels de services Internet basés sur l'implémentation d'outils de mesure Internet á grande échelle

21^{th} May 2015
Thése no : 2015TELE0012

Jury composé de :

David Hausheer	Rapporteur	Professeur, TU Darmstadt, Germany
Danny De Vleeschauwer	Rapporteur	Maître de Conférences/Chercheur, UGent/Alcatel-Lucent-Belgium
Pierre Sens	Examinateur	Professeur, UPMC - Sorbonne Universités - France
Xiaoming Fu	Examinateur	Professeur, University of Goettingen - Germany
Nishanth Sastry	Examinateur	Professeur, Kings College London - UK
Jean-Michel Portugal	Examinateur	Directeur Terrain de Conquéte, Orange Labs - France
Leandros A. Maglaras	Invité	Maître de Conférences, De Montfort University - UK
Noël Crespi	Directeur de Thése	Professeur, Institut Mines Telecom - France
Ángel Cuevas	Co-encadrant	Maître de Conférences, Universidad Carlos III de Madrid - Spain

Declaration

This Thesis :

- is the result of my own research work and contains nothing which is outcome of work other studies except where specifically indicated in the text by references.

- has not previously been submitted for a degree or diploma, or other qualification at any other university.

Reza Farahbakhsh
May 2015

Dedication

To My Family

Acknowledgements

This thoroughly enjoyable research is largely a result of the interaction that I have had with my supervisor Prof. Noël Crespi who I feel very privileged to have worked with him. I owe a great debt of gratitude for his patience, support, inspiration and friendship. I owe my gratitude to my Co-Supervisor, Ángel Cuevas. I received his personal support and I would like to thank him for his continuous efforts to show me the path to follow. Sincere thanks to both of them for having always been there to listen to and comment upon whatever I had thought out or wanted to discuss about and patiently reviewing all my work.

This thesis would not have been possible without many collaborative efforts and support of a lot of people whose help has been fundamental during the years of my PhD, especially Ruben Cuevas, Reza Rejaie, Xiao Han, Roberto Gonzales and Reza Motamedi. I would like also thank Walter Willinger, Bruce Maggs, Elie Najm, Djamal Zeghlache and Joerg Widmer, who gave me constructive comments on different presented studies in the manuscript. I also would like thank the reviewers of my thesis who patiently read this dissertation ; Prof. David Hausheer and Prof. Danny De Vleeschauwer and many thanks go to Prof. Pierre Sens for serving as the chairman of the jury committee and other member of the committee including Prof Xiaoming Fu, Mr. Jean-Michel Portugal and Prof. Nishanth Sastry.

A special thank to all my colleagues (and former colleagues) in Service Architecture Lab (SAL) at TSP specially Rebecca Copeland for her constructive comments on my manuscript.

During my Phd I had a great chance to be involved in two international projects, FP7 eCOUSIN and ITEA TWIRL, and I would like also to address special thanks for people who were involved in.

Finally, I would like to express my deep gratitude to my family and friends, and especially my parents, for their love and support during the ups and downs of graduate school. I am grateful beyond words for all that they have given me.

Reza Farahbakhsh
May 2015

Abstract

Popular Internet services are fundamentally shaping and reshaping the traditional ways of people communication, thus having a major impact on their social life. Two of the very popular Internet services with this characteristic are Online Social Networks (OSNs) and Peer-to-Peer (P2P) systems. OSNs provide a virtual environment where people can share their information and interests as well as being in contact with other people. On the other hand, P2P systems, which are still one of the popular services with a large proportion of the whole Internet traffic, provide a golden opportunity for their customers to share different type of content including copyrighted content.

Apart from the huge popularity of OSNs and P2P systems among regular users, they are being intensively used by professional players (big companies, politician, athletes, celebrities in case of OSNs and professional content publishers in case of P2P) in order to interact with people for different purposes (marketing campaigns, customer feedback, public reputation improvement, etc.).

In this thesis, we characterize the behavior of regular and professional users in the two mentioned popular services (OSNs and P2P systems) in terms of publishing strategies, content consumption and behavioral analysis. To this end, five of our conducted studies are presented in this manuscript as follows:

- "The evolution of multimedia contents", which presents a thorough analysis on the evolution of multimedia content available in BitTorrent by focusing on four relevant metrics across different content categories: content availability, content popularity, content size and user's feedback.

- "The reaction of professional users to antipiracy actions", by examining the impact of two major antipiracy actions, the closure of Megaupload and the implementation of the French antipiracy law (HADOPI), on professional publishers behavior in the largest BitTorrent portal who are major providers of online copyrighted content.

- "The amount of disclosed information on Facebook", by investigating the public exposure of Facebook users' profile attributes in a large dataset including half million regular users.

- "Professional users Cross Posting Activity", by analyzing the publishing pattern of professional users which includes same information over three major OSNs namely Facebook, Google+ and Twitter.

- "Professional Users' Strategies in OSNs", where we investigate the global strategy of professional users by sector (e.g., Cars companies, Clothing companies, Politician, etc.) over Facebook, Google+ and Twitter.

The outcomes of this thesis provide an overall vision to understand some important behavioral aspects of different types of users on popular Internet services and these contributions can be used in various domains (e.g. marketing analysis and advertising campaign, etc.) and different parties can benefit from the results and the implemented methodologies such as ISPs and owners of the Services for their future planning or expansion of the current services as well as professional players to increase their success on social media.

Keywords

Online Social Networks, User Behavior Analysis, Content Consumption, Publishing Strategies, Facebook, Google+, Twitter, Peer-to-Peer (P2P), BitTorrent, Privacy.

Table of contents

Chapter 1

Introduction

Contents

1.1 Motivation

The rapid expansion of Online Social Networks (OSN) has made a profound influence on our global community's communication patterns especially on the Internet which tends to reshape its utility and even the future design of it. This exponentially increasing number of users of social network services such as Facebook, Google+ and Twitter is creating a potentially dramatic change in people behavior and is bringing a huge change on traditional industries of content, media, and communications.

On the other hand, Peer-to-Peer services (P2P) are still getting a significant part of the whole Internet traffic and with the huge growth of end-users accessibility in terms of bandwidth, new type of services based on P2P technologies attract their customers by providing updated services such as online TV and games based on P2P phenomena.

In this era, different group of users are utilizing those services both from regular and normal users to very professional users with clear business strategies behind their activities. Regular users are getting the basic services from BitTorrent such as download the available contents and rarely share contents with others as well as doing normal activities in social networks such as building personal profiles, share posts in the profile and follow their friends and interests. On the other hand, professional users are benefiting from these popular internet services with other strategic goals. In term of BitTorrent, there are many professional users that actively publish contents (mostly copyrighted contents) and try to attract regular users to their activities. In social networks, the presence of professional users from different sectors is noticeable and there are provided services for those type of users, such as "Fan Page" in Facebook.

Analysing the behavior of users in each of these two systems is an important track of research in different communities from ISPs' Internet architecture designers to business and management sectors. Having the knowledge from end-users interests and predicting their behavior provide the opportunity to target right communities of users for different aspects such as advertising and recommendation systems.

The main goal of this thesis is to analyze the professional and regular user's behavior in these two main arena of Internet, online social networks and Peer-to-Peer systems. To this end, we study some aspects of professional and regular users behavior such as their content consumption, publishing strategies.

1.2 Thesis Contributions

The contributions and innovations of this thesis can be categorized in four parts: i) Main Outcomes, ii) Measurement Methodologies, iii) Implemented Measurement Tools, and iv) Available Datasets. Next, a summary of each mentioned categories will be presented.

1.2.1 Summary of the main outcomes

This part presents a summary of different studies that is presented in this manuscript as the main outcomes of this thesis. In P2P domain, the main focus is on "ThePirateBay" which is one the most popular bit-torrent portals and we study two group of users; i. Regular users who use this system to download contents and ii. Professional users who are content publishers with business strategies behind their activities. Next we present a short abstraction of the two studies in this domain:

- **Multimedia Evolution on P2P:** Today's Internet traffic is mostly dominated by multimedia content and the prediction is that this trend will intensify in the future. Therefore, main Internet players, such as ISPs, content delivery platforms (e.g. Youtube, BitTorrent, Netflix, etc.) or CDN operators, need to understand the evolution of multimedia content availability and popularity in order to adapt their infrastructures and resources to satisfy clients requirements while they minimize their costs. This study presents a thorough analysis on the evolution of multimedia content available in BitTorrent. Specifically, we analyze the evolution of four relevant metrics across different content categories: content availability, content popularity, content size and user's feedback. To this end we leverage a large-scale dataset formed by four snapshots collected from the most popular BitTorrent portal, namely The Pirate Bay, between Nov. 2009 and Feb. 2012. Overall our dataset is formed by more than 160k content that attracted more than 185M of download sessions.

- **Reaction to Antipiracy actions:** Based on the economic and social impact of copyrighted content infringement in recent years, few countries have put in place online antipiracy laws and there have been some major enforcement actions against violators. This raises the question *to what extent antipiracy actions have been effective in deterring online piracy?* This is a challenging issue to explore because it is difficult to capture user behavior, and to identify the subtle effect of various underlying (and potentially opposing) causes. In this study, we tackle this question by examining the impact of two major antipiracy actions, the closure of Megaupload and the implementation of the French antipiracy law, on publishers in the largest BitTorrent portal who are major providers of copyrighted content online. We capture snapshots of Bit-

Torrent publishers at proper times relative to the targeted antipiracy event and use the trends in the number and the level of activity of these publishers to assess their reaction to these events. Our investigation illustrates the importance of examining the impact of antipiracy events on different groups of publishers and provides valuable insights on the effect of selected major antipiracy actions on publishers' behavior.

In the OSN part of my thesis, the focus is on three major OSNs (Facebook, Twitter and Google+) and we analyze two types of users: i. Regular profiles who are normal users that use social networks for their everyday social activities. ii. Professional users who are interested in social networks for some business aspects and behind these users are usually companies or individual figures and celebrities with huge number of fans. In this thesis, three of our studies in this domain are presented.

- **Disclosed Information on Facebook:** Facebook, as the most popular online social network which according to Alexa [1], is the 2nd most popular website in the world at the time of writing this thesis. There is a large amount of personal and sensitive information publicly available that is accessible to external entities/users. In this study we look at the public exposure of Facebook profile attributes to understand what type of attributes are considered more sensitive by Facebook users in terms of privacy, and thus are rarely disclosed, and which attributes are available in most Facebook profiles. Furthermore, we also analyze the public exposure of Facebook users by analyzing the number of attributes that users make publicly available in average. To complete our analysis we have crawled the profile information of 479K randomly selected Facebook users. Finally, in order to demonstrate the utility of the publicly available information in Facebook profiles we show in this study three case studies. The first one carries out a gender-based analysis to understand whether men or women share more or less information. The second case study depicts the age distribution of Facebook users. The last case study uses data inferred from Facebook profiles to map the distribution of worldwide population across cities according to its size.

- **Cross Posting Activity:** On-line Social Networks (OSNs) are being intensively used by professional players (e.g., big companies, politician, athletes, celebrities, etc.) in order to interact with a huge number of regular OSN users with different purposes (marketing campaigns, customer feedback, public reputation improvement, etc.). Hence, due to the large catalog of existing OSNs, professional players are usually involved in different OSNs. In this context an interesting question is whether professional users publish the same information across their OSN accounts, or actually they use different OSNs in a different manner. We define as cross-posting activity

the action of publishing the same information in two or more OSNs. In this study we aim at characterizing the cross-posting activity of professional OSN users across three major OSNs, Facebook, Twitter and G+. To achieve this goal we perform a large-scale measurement-based analysis across more than 2M posts collected from 616 professional users with active accounts in the three referred OSNs.

- **Professional Users' Strategies in OSNs:** The intensive use of professional players led to an increasing research interest that aims at understanding what are the strategies of professional users in OSNs. In this study we investigate the global strategy of professional users by sector (e.g., Cars companies, Clothing companies, Politician, etc.). To perform that analysis we have to first validate that users belonging to the same sector/category present a similar strategy in their use of OSNs. To find whether there are some sectors fulfilling that requirement, we use a dataset of 616 professional users with active accounts in the three most popular OSNs: Facebook (FB), Twitter (TW) and Google+ (G+). We find 8 categories in which users present similar behavioural elements: Athletes, Cars, Media News, Movie, Musician-Band, News Website, Politician, and Sports Teams. We describe the behaviour for these categories across FB, TW and G+ highlighting those elements that differentiate each strategy. Finally, we present a simple methodology that allows us to estimate the success of each strategy based on the number of reactions per post that a category is able to attract.

1.2.2 Measurement Methodologies

Apart from the mentioned outcomes of this thesis, there are some measurement methodologies that were implemented during this thesis as follows:

- Methodology to evaluate the effect of an antipiracy action on the publishers and consumers of P2P contents in a country.

- Methodology to find cross posting activity of professional users across multi social networks.

- Methodology to find and classify different strategies that professional users are following in different sectors and evaluate the success of their strategies.

- Methodology to evaluate the architecture of a large CDN to understand how different services are serving to its customers from different locations.

1.2.3 Implemented Large Scale Measurement Tools

This part of thesis includes a list of implemented data collection tools during this thesis. The implemented tools are available for further research collaboration.

- **"Facebook Fan Page crawler"** This tool is able to collect popularity, activity and attracted reactions for a list of Facebook professional users (Fan pages).
- **"Facebook Fan Page Popularity monitoring"** This tool is able to monitor periodically a list of Fan Pages and collect the popularity and its' relevant parameters.
- **"Facebook regular users' profile crawler"** This tool is able to collect Facebook regular users' profile including the general information of the users public profile and the activity of a users that contains published/shared posts in the wall pages.[1]
- **"Facebook physical network discovery tool"** This tool is able to monitor reachability to a list of Facebook servers by using planetlab infrastructure distributed in the world. This tool can be modified and utilized for other large scale networks such as Google or YouTube network.
- **"Facebook Traffic analyzer"** Main function of this tool is to monitor the network packet level traffic and collect the packets of Facebook sessions. By running this tool in the gateway of a network, we are able to collect all users packets for their Facebook activities and sessions to see what is the amount of each individual FB users' traffic and generally what type of activity have been performed by users.
- **"BitTorrent Trackers Crawler"** This tool is able to connect to the BitTorrent trackers and collect downloads' detail information as well as many other useful data per torrent.
- **"Movie industry data collector"** This tool collects movies' general and business related information from three online resources including IMDB portal.

1.2.4 Available Datasets

- **"Facebook Fan pages dataset"** includes popularity, activity and reactions of around 300K very popular Facebook Fan pages.
- **"Facebook regular profiles"** includes around 500K users profile information as well as their social connections and their public activity.

[1]This tool has been implemented in collaboration with my colleague Xiao HAN

- **"Facebook services' server reachability"** includes more than one year ping and trace route data (6 times per day) for 47 facebook services (servers from Akamai and FB) from 473 planet lab nodes across the world.
- **"Facebook Fan pages popularity evolution"** includes evolution of #Fans for 10K of top Fan Pages more than 18 months (snapshots captured 6 times per day).

1.2.4.1 Ethics Considerations

Although we only collected publicly available data from both regular and professional users, we enforced a few steps to protect user privacy specially for social network data. During our analysis, all data were encrypted and not re-distributed, and no personal and sensitive information was extracted, and we only analyzed aggregated statistics.

1.3 Publication List

The following papers, (published, submitted or under submission) are the partial outputs during my PhD.

Journal Articles:

- **R. Farahbakhsh**, A. Cuevas, R. Cuevas, R. Gonzalez, N. Crespi, "Understanding the evolution of multimedia content in the Internet through BitTorrent glasses", IEEE Networks Magazine, Volume:27 Issue:6, Dec. 2013.
- M. Mani, W. Seah, N. Crespi, **R. Farahbakhsh**, "P2P IP Telephony over Wireless Ad-hoc Networks A Smart Approach on Super Node Admission", Peer-to-Peer Networking and Applications Journal, Springer, June 2012.
- **Reza Farahbakhsh**, Angel Cuevas, Antonio Ortiz, Xiao Han, Noel Crespi, "How Far is Facebook from Me!, Facebook Network Architecture Analysis", Under revision on IEEE Communication Magazine, 2015.
- Xiao Han, **Reza Farahbakhsh**, Angel Cuevas, Ruben Cuevas and Noel Crespi, "Community Similarity Degree: Finding Similarity to Improve Recommendations in Online Social Networks", submitted to Expert Systems with Applications, Elsevier, 2015.
- **Reza Farahbakhsh**, Angel Cuevas, Ruben Cuevas, Noel Crespi, "Characterization of Professional Users Strategies in major OSNs", submitted to Communication of ACM, 2015.
- **Reza Farahbakhsh**, Xiao Han, Angel Cuevas, Noel Crespi,"Privacy Evolution of Publicly Disclosed Information in Facebook Profiles", submitted to IEEE Security & Privacy magazine, 2015.

Conference Papers:

- **Reza Farahbakhsh**, Angel Cuevas, Ruben Cuevas, Reza Rejaie, Michal Kryczka, Roberto Gonzalez and Noel Crespi, "Investigating the Reaction of BitTorrent Content Publishers to Antipiracy Actions", IEEE P2P, Trento, Italy, Sep. 2013.
- **Reza Farahbakhsh**, Xiao Han, Angel Cuevas and Noel Crespi, "Analysis of publicly disclosed information in Facebook profiles", IEEE/ACM ASONAM, Niagara fall, Canada, 2013.
- W. Chanthaweethip, X. Han, N. Crespi, Y. Chen, **R. Farahbakhsh** and A. Cuevas, "Current City Prediction for Coarse Location Based Applications on Facebook", IEEE Globecom, USA, Dec. 2013.
- Xiao Han, Leye Wang, Son N. Han, Chao Chen, Noel Crespi, **Reza Farahbakhsh**, "Link Prediction for New Users in Social Networks", IEEE ICC, Oxford, UK, 2015.
- **R. Farahbakhsh**, N. Crespi, A. Cuevas, Neetya Shrestha, M. Mani, Poompat Saengudomlert, "Improved P2P Content Discovery by Exploiting User Social Patterns", ICNC, San-Diego, USA 2013.
- **R. Farahbakhsh**, N. Crespi, A. Cuevas, S. Adhikari, M. Mani, T. Sanguankotchakorn, "socP2P: P2P Content Discovery Enhancement by considering Social Networks Characteristics", IEEE ISCC, Cappadocia, Turkey, July 2012.
- **Reza Farahbakhsh**, Angel Cuevas, Noel Crespi, "Characterization of cross-posting activity for professional users across Facebook, Twitter and Google+", submitted to IEEE/ACM ASONAM 2015.
- R. Gonzales, Reza Motamedi, **Reza Farahbakhsh**, Angel Cuevas, Ruben Cuevas, Reza Rajae, "Head-to-Head Comparison of Major OSNs", under submission for ACM CoSN 2015.
- Roberto Gonzalez, **Reza Farahbakhsh**, Reza Motamedi, Angel Cuevas, Ruben Cuevas and Reza Rejaie, "Characterization of information propagation in Google+ and its Comparison with Twitter", under submission for ACM CoSN 2015.

1.4 Structure of the Thesis

This contributions of this thesis can be divided in two parts: i) Social networks and ii) P2P systems. To follow the contributions, this manuscripts is also organized in two main sections to overview separately the works that has been done in each domain.

First of all an overall overview of the related work and state of the art is presented in section 2. Section 3 presents two of the studies regarding to the user behavior analysis in P2P systems. The first one explores evolution of Internet multimedia content in the past few years presented in subsection 3.2 and the second study, which is presented in subsection 3.3, is an investigation about the effects of two major antipiracy actions on the P2P content publishers.

The second part of the thesis, section 4, presents three studies related to social networks. It starts with a study presented in subsection 4.2, in which the publicly disclosed information of Facebook profile users has been analyzed. Second and third studies are related to professional users' behavior characterization over Cross OSNs which are presented in subsections 4.3 and 4.4 respectively. Those studies includes characterizing cross-posting activity of professional users across Facebook, Twitter and Google+ which is presented in 4.3 and characterization of Professional Users' Strategies in those major OSNs at subsection 4.4.

Finally section 5 concludes this report by describing some of the ongoing and future works.

Chapter 2

State of the Art

Contents

2.1 Summary

Related work and state of the art to this thesis can be divided in two groups. Studies that characterize some of the large Internet applications and works that implemented large scale measurement tools to collect datasets from popular Internet applications. In this chapter we overview major studies on the two mentioned groups, specifically in those ones measuring Online Social Networks and Peer to Peer systems. The literature review presented in this chapter overview the general studies relevant to the two mentioned topic and later for each presented study in this thesis, a separate and detail overview of the related work will be presented.

2.2 Internet applications Characterizations

Characterization of the popular and large scale Internet applications have attracted the attention of the research communities in the last decade. One of the main reason is that this type of studies are crucial and useful for different players: i) Internet service providers with the goal of evaluating their popular applications that is implemented, ii) for the network expansion plan to the companies that are running this type of application such as Facebook or Akamai iii) or companies that aim to create a successful large scale application. To this end, next we overview the recent studies which aim to characterise different aspect of Social networks and P2P systems.

2.2.1 Social Networks

The research community has dedicated a fair amount of work to characterize OSNs in the last years. The conducted studies can be classified into three broad classes:

Connectivity properties & social graph: The connectivity properties of the social graph for Facebook [2–4], Twitter [5, 6], Google+ [7–9] and other less popular OSNs [10] have been carefully analyzed by the referred works. The results presented in those studies along with the results in our studies depict a complete comparison study of the activity and connectivity of these OSNs.

In addition, we can find works analyzing the graph of other social players different than the main OSNs like [10] that analyze the graph properties for static snapshots of four social systems Orkut, Flickr, LiveJournal, and YouTube. In particular, Magno et. al in [7] perform an early analysis on G+ and identify the main similarities and differences with other OSNs like FB and TW. However, their comparison only focuses on the social graph level, and does not cover user activities or reactions, which is the actual scope of our

conducted studies. Finally in [9], authors compare the connectivity properties of the social graph of FB, TW and G+.

Temporal Evolution of OSN properties: Previous works have studied the evolution of the relative size of the network elements for G+ [9] or Flickr and Yahoo [11]. Furthermore, other works have analyzed the evolution of the social graph properties [9,12–17], the evolution of the interactions between users [18] and the evolution of users' availability over time [19]. In addition in [17] the authors have analyzed the evolution of users' activity in MySpace and Twitter.

Information Disclosure in Social Networks. There are several studies that investigate the level of information disclosure in social networks focusing a group of users from a specific country [20,21] or city [22] or users from a university [23] but just few studies are available that look on a random sample of users [24]. Conceptually similar to our efforts, Quercia et al. (2012) [20] found a correlation between the degree of openness and gender, using a dataset of 1323 profiles from the United States. Gross et al. in [23] studied the patterns of information revelation in Facebook. They analyzed around 4K Carnegie Mellon University students' profiles, specifically those that joined a popular social networking site catering to college students. In other work, Chang et al. [21] studied the privacy attitudes of U.S. Facebook users of different ethnicities. Another U.S.-based study [25] used a questionnaire and with considering 1,710 students' profiles shows that women are more likely to maintain a higher degree of profile privacy than men; and that having a private profile is associated with a higher level of online activity. Authors in [26] examined disclosure in Facebook profiles looking at only 400 Facebook profiles. In a similar work to the previous one, authors in [27] employed surveys and interviews to study the factors that influence university students to disclose personal information on Facebook. In addition, we also study the amount of disclosed information on Facebook profiles on a dataset including half million users [24].

Some other studies provide methodologies which use available Facebook users' profile attributes to do different type of estimation or prediction such as estimating the birth year [22], predicting the friendship [28,29] or predicting the attributes of another user [30].

Users' Behavior in Online Social Networks: Users' behavior needs to be characterized from real data collected from OSNs. In particular, previous works have used two different strategies: Passive measurements [31,32] vs. Active measurements [9,33,34]. The former captures traces of traffic or click streams that allow the reconstruction of the behavior of users whereas the latter uses crawling techniques similar to those described in our studies.

Many studies has been conducted to characterize the behavior of users based on the

real data collections. Gyarmati et al. [34], in accordance to our studies, used active measurements to characterize users' activity in few different OSNs. Gyarmati et al. analyzed less popular OSNs such as Bebo, MySpace, Netlog, and Tagged and they defined activity as the time a user stays on the system but they do not characterize users' reactions which, as we have demonstrated, are key features. Authors in [35] investigate emotional contagion of facebook users which occurs outside of in-person interaction between individuals. In another work, author conducted a large scale experiment over 61-million facebook users and study the social influence and political mobilization [36].

2.2.2 Peer-to-Peer

Popularity Evolution of P2P Applications: P2P networks are already widely used around the Internet, mainly for file sharing. The massive sizes of some P2P networks contain huge numbers of all kinds of content. There are several papers that look at the evolution of P2P traffic along the time *e.g.*, [37, 38]. The most recent one [38] studies the Inter-AS traffic associated to several ISPs across the Internet. The authors suggest that P2P traffic is becoming less representative and mention the migration process discussed in this paper as a possible cause. Furthermore, [39] studies the impact of BitTorrent in the Internet traffic over a period of two years between Nov 2008 and Nov 2010. The authors briefly mention a reduction of 10% in the number of peers that partially validates our observations. They argue that this reduction may be due to a drop in the system popularity and at the same time acknowledge the difficulty of validating this hypothesis so that they do not explore it. Our study which is presented in section 3.2, is different in nature than the previous works in the literature since we do not analyze the network footprint of BitTorrent, instead we perform a comprehensive analysis of the evolution of BitTorrent popularity at aggregate and local level across both publishers and consumers. In addition, we face the difficult task of finding the root causes for the discovered trends that to the best of our knowledge has not been addressed before.

Socio-economic Studies in BitTorrent: The popularity of BitTorrent attracted the attention of the research community to examine various aspects of swarming mechanism in BitTorrent [40–43] and propose different techniques to improve its performance [44, 45]. Furthermore, other aspects of BitTorrent such as demographics of its ecosystem [46–48] along with security [49] and privacy issues [50, 51] have also been studied. However (to the best of the author's knowledge) despite of its importance, little work has been conducted on the understanding of socio-economic aspects of P2P applications in general and BitTorrent ecosystem in particular [47, 50, 52]. Authors in [52] studied the incentives that drive users to publish content in BitTorrent.

2.3 Large scale measurement tools implementation

2.3.1 Social Networks

Considering the ongoing researches that aim to understands the phenomena of social media, a necessary first step is to collect good enough data from various available OSNs and other types of social media. To this end, proper data collection tools and crawlers are being developed to gather data from different sources. Large-scale data collection from OSN services mainly depends on the functionality provided by the analysed system. Possible solutions to collect data include the use of a systems available APIs, as in Google+ and Twitter [9, 53], and the integration of their own applications to attract people and to provide access to their profile information, as in Facebook [54]. Alternatively, web crawling techniques have been used to analyse Facebook [4,55,56] and other platforms, e.g., Myspace [57], Flickr [12, 58], and YouTube [59]. Usually, web crawling is applied in cases where the required data cannot be accessed via an available API, or when the revealed data is insufficient for subsequent analysis. The research community uses the collected data to analyse social media from different perspective [7, 22, 24, 60–62].

2.3.2 Peer-to-Peer

In recent years, several studies measure the P2P ecosystem from various perspective and different measurement methodologies for implementing large scale measurement tools. Authors in [63] conduct a complete survey on different methods of measurement and simulation in BitTorrent. In summary two popular way to collect data from P2P systems is observing the trackers log [64] or using mirror script such as HTML script [65]. Authors in [66] surveys the existing measurement studies and also collected BitTorrent traffic at four major European ISPs and investigate how is the BitTorrent traffic pattern from ISPs perspective. There are several other studies such as [67, 68] which investigate different aspects of P2P system by implementing large scale data collection tools.

2.4 Conclusion

This section provided a general overview over the major previous efforts relevant to this thesis. In summary it provided some key studies on two research lines “characterizing a large Internet application” and “Implementation of large scale measurement tools” for OSNs and P2P systems. In addition to this section, for each of the presented studies in this thesis, a separate related work will be provided which focuses on the main relevant studies to that specific study.

Chapter 3

Profiling regular and professional users in BitTorrent

Contents

3.1 Summary

The main focus of this chapter is to characterize professional and regular users behavior on P2P systems. To this end, two studies are presented that look to some aspects of the users characterizations such as their content consumptions trend on a major BitTorrent portal namely "ThePirateBay" as well as how the professional and regular users react to antipiracy actions in terms of their activity and publishing behavior.

More specifically, subsection 3.2 presents a measurement study which show us how the evolution of multimedia contents in BitTorrent has been changed over 3 years and following to that, subsection 3.3 includes a study about the reaction of professional BitTorrent Content Publishers to two major Antipiracy Actions (shutdown of Megaupload and Hadopi law).

Keywords

Multimedia Content, P2P, BitTorrent, Content Availability, Content Popularity, Content Size, Piracy, law, Cyberlocker, Megaupload, P2P, BitTorrent, Hadopi.

3.2 Evolution of multimedia content in the Internet through BitTorrent glasses

3.2.1 Introduction

In the last years Internet traffic has been mostly dominated by multimedia content [69]. This has led to the development of new technologies to distribute this content: (*i*) P2P technologies that allow end-users to share content without the necessity of a dedicated infrastructure, (*ii*) Cyberlockers that are web-based portals that allow users to both upload and download content, (*iii*) multimedia content distribution platforms such as YouTube (video), Netflix (TV shows and Movies) or Spotify (music). In addition, in order to reduce the cost and improve the efficiency of the content distribution a new network infrastructure namely Content Delivery Network (CDN) was proposed [70]. A CDN uses caching and prefetching strategies in order to store the content close to those users that are likely to consume it so that the amount of data crossing long paths in the Internet is reduced and the user's experience is enhanced. Some of the main players in the business of CDNs are Akamai and Limelight that provide their service worldwide to a large number of customers. Furthermore, some Content Providers such as Google have also deployed their own CDNs. Finally, network operators have to continuously adapt their network infrastructures in order to efficiently serve the large demand of multimedia content. For instance, some of the major operators have recently started to develop their own CDN.

The described scenario, along with the expected steady growth of the traffic associated to multimedia content in the near future [71], makes it interesting to study the evolution of such content. Understanding this evolution will help the aforementioned players to adapt their algorithms, infrastructures and resources to meet the needs of their clients and, at the same time, increment their revenues.

In this study, we present a first step to study the evolution of different types of content in the Internet using BitTorrent as reference system. We believe that BitTorrent is the most appropriate platform to conduct our study due to the following reasons: (*i*) BitTorrent is the application that aggregately contributed more Internet traffic in the last decade [72] [73] and it is still among the three that generate more traffic [69]; (*ii*) The most popular (and recent) content (e.g. last Hollywood movies) are typically available in BitTorrent; (*iii*) Other successful platforms such as Neftflix, YouTube or Spotify are specialized in a single type of content. Instead, BitTorrent offers a broader catalogue of content types (e.g, video, audio, games, etc). Therefore, it allows to perform a comparative study across different types of multimedia content.

Our study is based on a large scale dataset collected from the most popular BiTorrent portal, namely The Pirate Bay (TPB), over a period longer than two years between Nov.

2009 and Feb. 2012. Note that TPB received more than twice daily visits compared to the second most popular BitTorrent portal, according to Alexa ranking [1]. We have collected 4 different snapshots over this time window that collectively account for more than 160K content that attracted more than 185M download sessions. This dataset constitutes a solid ground to provide meaningful insights regarding the content evolution in BitTorrent and by extension in the Internet. In particular, our study address three concrete but very relevant issues: (*i*) We analyze the evolution of the content availability and popularity associated to different content types over the considered period; (*ii*) We study the evolution of the content size for all aggregate content and its division into the different categories; and, (*iii*) We quantify the end-users' feedback activity by means of the number of comments that each content receives.

Our main insights are:

- Video of different types (Movies, TV Shows, Porn) represents 40-50% of the overall content and attracts 80% of the download sessions.
- The median size of the available content has doubled in a two years period.
- High-resolution content has multiplied by 5 its availability and popularity to represent 10% of the multimedia content and downloads in Feb. 2012.
- Finally, we have observed that end-users' feedback is typically very reduced.

3.2.2 Background

There are two separated processes in BitTorrent functionality. On the one hand, we find the process in which a user (publisher) makes a content available, or *publishing phase*. On the other hand, once the content is available end-users (consumers) download it in the *downloading phase*.

In the publishing phase, the publisher generates a .torrent file associated to that content and uploads it in a BitTorrent portal such as The Pirate Bay (TPB). In addition, the publisher registers the content in one (or more) Tracker(s), which is a server that manages and monitors the swarm (the set of peers sharing a content) associated with a given content. As part of its services, the Tracker keeps track of all the peers (*i.e.* IP addresses) that share the content and classifies them either as seeders (which have the full content) or leechers (which have only some pieces of the content). The .torrent file includes (among other information): the IP address of the Tracker (and optionally a list of other backup trackers) that manages the swarm associated to the content, the content size and its name. In addition, major torrent portals like TPB provide a web page for every uploaded content

that includes information such as size, category, number of leechers and seeders, content description, users' comments, etc.

In the downloading phase, a BitTorrent client gets the .torrent file associated to the desired content from a BitTorrent Portal (e.g. TPB). That client subsequently sends a request to the Tracker included in the .torrent file. The Tracker replies with: (*i*) the number of seeders and leechers that are currently connected to the swarm, and (*ii*) N (typically 50 with a limit of 200) random IP addresses of peers participating in the swarm. Next, the BitTorrent client connects to those peers in order to start receiving pieces of the content (and after getting some pieces serves them to other peers). From time to time, during the downloading process, the BitTorrent client may contact the Tracker to obtain more peers.

3.2.3 Related Work

The demonstrated weight of BitTorrent in the Internet has attracted the attention of many computer scientists, who have made in depth studies of the functionality of the BitTorrent ecosystem [47] [48], have generated models that capture its behaviour [40], have provided new algorithms to improve its performance [45] [44], have analyzed and proposed mechanisms regarding its security [49] [51], and have evaluated socio-economic reasons that motivate users to upload and consume BitTorrent content [52]. Therefore, the technical and socio-economical aspects of BitTorrent have been thoroughly studied. However, to the best of our knowledge, except [72] [73] which are technical reports that provide some general insights to the evolution of P2P protocols in different countries, there is no study that analyzes the evolution of the BitTorrent content in a long term.

3.2.4 Measurement Methodology

The goal of our measurement process is to collect a large number of contents and the following information for each one of them: (*i*) the content Category/Subcategory as defined by TPB, (*ii*) the number of download sessions, (*iii*) the content size, (*iv*) the number of comments provided by end-users.

Towards this end, we leverage the RSS feed of TPB to detect the availability of any new .torrent file. When a new torrent is detected, in addition to gather its size (from the .torrent file) and Category/Subcategory from TPB, our crawler tool periodically queries the tracker in order to obtain the IP addresses of the participants in the content swarm and always solicits the maximum number of IP addresses (*i.e.* 200) from the Tracker. To avoid being blacklisted by the Tracker, we issue our queries at the maximum rate that is allowed by the tracker (*i.e.* 1 query every 10 to 15 minutes depending on the tracker load). Given this constraint, we query the tracker from several geographically-distributed

Table 3.1: Datasets Description

	pb09	pb10	pb11	pb12
Crawling Period	11/28/09 - 12/18/09	04/09/10 - 05/05/10	10/21/11 - 12/13/11	01/28/12 - 02/12/12
Duration (days)	21	27	54	16
Torrents	15.8K	38.2K	72.0K	21.0K
Downloads	-	95.6M	79.0M	11.1M

machines so that the aggregated information by all these machines provides an adequate high resolution view of the participating peers (*i.e.* number of download sessions). We continue to monitor a target swarm until we receive 10 consecutive empty replies from the Tracker. This allows us to capture for each new content its size, Category/Subcategory and the number of associated download sessions.

Finally, in order to gather the number of comments for a given content, we crawled TPB page of all collected content in June 2012. It must be noted that at that time some of the contents collected by our crawling tool had been removed from TPB, and thus we could not gather their number of comments.

Using the described methodology we have collected four snapshots of TPB content between Nov. 2009 and Feb. 2012. We refer to them as pb09, pb10, pb11 and pb12 based on the year in which each dataset was collected. Table 3.5 summarizes the main characteristics of these datasets (as it is shown in the table we do not have the number of download sessions for pb09). All the snapshots together contribute more than 160K torrents (*i.e.* contents) and 185M download sessions. These numbers allow us to perform a comprehensive analysis on how the content (and its division into different categories) has evolved over the two years period that separates the four datasets.

3.2.5 Content Evolution Analysis

In this subsection we investigate how the relative weight (in %) of the different content categories evolves in the period under study. For that, we first classify all the collected contents following the Category/Subcategory schema defined by TPB. Next, we analyze each of them from an availability (portion of content available in each category) and a popularity (portion of downloads for each category) perspective.

3.2.5.1 Content Availability Evolution

Table 3.2 shows the portion of content available in each Category/Subcategory for pb09, pb10, pb11 and pb12 snapshots.

VIDEO is the dominant category and doubles, in all the snapshots, the number of contents available in any other category. The VIDEO category shows a very slight increment

Table 3.2: Distribution of content availability (proportion of available content) by categories/subcategories and datasets (pb09, pb10, pb11 and pb12)

Category	pb09 (%)	pb10 (%)	pb11 (%)	pb12 (%)
AUDIO	**15.958**	**15.208**	**12.535**	**13.884**
Music	10.118	10.796	7.984	8.414
Audio Books	0.376	0.728	0.579	0.608
Sound Clips	0.162	0.076	0.095	0.120
FLAC	1.757	1.218	1.894	1.910
Other	3.546	2.390	1.984	2.833
VIDEO	**39.234**	**41.266**	**52.260**	**46.272**
Movies	23.004	20.084	20.623	19.924
Movies DVDR	-	1.625	1.448	2.029
Music Videos	1.646	2.340	1.151	1.608
Movie Clips	-	0.433	0.237	0.493
TV shows	11.913	14.216	21.996	15.435
Handhled	0.207	0.258	0.353	0.110
Highres - Movies	1.348	0.644	1.842	1.728
Highres - TV shows	-	0.603	3.690	4.039
3D	-	-	0.072	0.014
Other	1.115	1.062	0.849	0.890
APPLICATIONS	**16.788**	**9.922**	**3.986**	**5.006**
Windows	13.514	9.283	3.371	3.647
Mac	0.726	0.258	0.238	0.345
UNIX	0.071	0.089	0.136	0.235
Handheld	0.292	0.133	0.031	0.014
IOS(Ipad/Iphone)	-	-	0.051	0.302
Android	-	-	0.097	0.349
Other OS	2.184	0.159	0.061	0.115
GAMES	**4.997**	**3.253**	**3.084**	**4.236**
PC	3.636	2.599	2.642	3.039
Mac	0.039	0.037	0.043	0.072
PSx	0.181	0.063	0.088	0.254
XBOX360	0.201	0.099	0.070	0.148
Wii	0.389	0.198	0.141	0.168
Handheld	0.551	0.258	0.102	0.053
IOS(Ipad/Iphone)	-	-	0.026	0.211
Android	-	-	0.232	0.177
Other	0.402	0.279	0.092	0.115
PORN	**8.264**	**21.553**	**21.140**	**23.007**
Movies	5.950	10.767	9.097	10.386
Movies DVDR	-	0.532	0.014	0.057
Pictures	1.232	1.688	0.971	1.206
Games	0.091	0.026	0.015	0.077
Highres - Movies	0.201	0.511	1.878	2.422
Movie Clips	-	7.308	8.670	8.313
Other	0.791	0.720	0.494	0.546
OTHER	**14.759**	**8.798**	**6.994**	**7.595**
E-books	5.185	4.352	3.865	5.068
Comics	0.421	1.059	1.316	1.278
Pictures	2.930	2.173	1.227	1.163
Covers	0.058	0.016	0.021	0.005
Physibles	-	-	-	0.005
Other	6.164	1.198	0.565	0.077

in its presence between pb09 and pb10 from 39% to 41%. It keeps a stable growth to reach 52% (*i.e.* at this point there was more video content than the sum of all other categories) of the overall content in pb11, and then it surprisingly shows a considerable drop of 6 percentage points (to 46%) in the two months separating pb11 and pb12.

We now turn our attention to the PORN category that shows an important increment in its representativeness during the five months between pb09 and pb10. This increase allows PORN to scale from the 5^{th} category in terms of availability in pb09 (8%) up to the 2^{nd} position in pb10 accounting for 21% of the total content. From this moment on, it remained in the 2^{nd} position and maintained its weight, 21% in pb11 and 23% in pb12.

The remaining categories (AUDIO, APPLICATIONS, GAMES and OTHER) follow a

similar trend over time. They steadily reduce their weight between pb09 and pb11 and change this slope in pb12. Although the trend is similar we can find a much more marked representativeness loss in the APPLICATIONS and OTHER categories. The APPLICATIONS category almost halves its presence between pb09 (16.8%) and pb10 (10%), and maintains that descendent line to only accounts for 4% of the contents in pb11, followed by a small increase up to 5% in pb12. The OTHER category shows a strong decrement of its weight between pb09 (15%) and pb10 (8.7%) to later slows down the slope of this loss to end up in 7% of the total content in pb11 and slightly increases this value (7.5%) in pb12. Contrary to these cases, GAMES and AUDIO categories present a smoother contribution reduction between pb09 and pb11 of 3 percentage points for AUDIO and 2 percentage points for GAMES, to later increase 1 percentage point in pb12.

After analyzing the evolution of each category we can present three interest insights:

- Movies and TV Shows (in the VIDEO category) are the most available contents. Both subcategories together always sum up more than 34% of the total content, and they reach a peak of presence in pb11 when both together surpassed 40%. Furthermore, if we add the PORN-Movies subcategory, we end up with a range between 40%-50% for Movies and TV Shows.

- There is a relevant increment of the High Resolution content. While that type of content only represented about 1.5% in pb09 and pb10 (summing up Highres-Movies from PORN and VIDEO and Highres-TV Shows from VIDEO), it grew to 7.4% and 8.2% in pb11 and pb12, respectively.

- The presence of Windows related content has dramatically decreased. It represented 13% of the total available content in pb09, while in the most recent snapshots its presence is reduced to a mere 3%.

3.2.5.2 Content Popularity Evolution

The previous subsection has analyzed the content availability in TPB. We now study the popularity of the different Categories/Subcategories over time based on the number of download sessions associated with each content in our snapshots.

Table 3.3 shows the portion of download sessions in each Category/Subcategory for pb10, pb11 and pb12 snapshots. As we mentioned earlier, we did not collect download information for pb09.

VIDEO is the most popular category by attracting more than 3/5 of the downloads in all the snapshots. However, it shows a relevant drop in its popularity over the time. VIDEO represented 71% of the downloads in pb10 and steadily decreased after that, to 64% and 59% in pb11 and pb12 respectively.

Table 3.3: Distribution of content popularity (proportion of download sessions) by categories/subcategories and datasets (pb09, pb10, pb11 and pb12)

Categories	pb10 (%)	pb11 (%)	pb12 (%)
AUDIO	**4.671**	**5.574**	**4.972**
Music	3.814	3.977	1.036
Audio Books	0.119	0.213	0.093
Sound Clips	0.011	0.065	0.053
FLAC	0.208	0.297	0.292
Other	0.518	1.021	3.498
VIDEO	**71.299**	**64.080**	**58.925**
Movies	41.394	29.874	22.667
Movies DVDR	0.937	1.027	0.943
Music Videos	0.443	0.245	0.284
Movie Clips	0.066	0.037	0.097
TV shows	26.448	27.010	28.349
Handhled	0.127	0.040	0.014
Highres - Movies	0.766	3.533	3.702
Highres - TV shows	0.723	2.205	2.826
3D	-	0.025	0.000
Other	0.396	0.086	0.043
APPLICATIONS	**2.117**	**0.996**	**0.810**
Windows	2.041	0.934	0.725
Mac	0.050	0.041	0.027
UNIX	0.002	0.002	0.000
Handheld	0.018	0.001	0.000
IOS(Ipad/Iphone)	-	0.003	0.002
Android	-	0.012	0.054
Other OS	0.006	0.001	0.001
GAMES	**1.274**	**2.182**	**1.013**
PC	0.790	1.747	0.756
Mac	0.003	0.003	0.000
PSx	0.018	0.023	0.006
XBOX360	0.027	0.119	0.165
Wii	0.144	0.102	0.019
Handheld	0.216	0.022	0.001
IOS(Ipad/Iphone)	-	0.005	0.006
Android	-	0.154	0.056
Other	0.075	0.007	0.004
PORN	**17.256**	**24.300**	**31.012**
Movies	11.259	13.209	17.685
Movies DVDR	0.034	0.014	0.025
Pictures	0.740	0.255	0.598
Games	0.007	0.004	0.009
Highres - Movies	0.385	1.727	3.089
Movie Clips	4.559	8.827	8.388
Other	0.272	0.264	1.218
OTHER	**3.383**	**2.868**	**3.268**
E-books	1.337	2.099	2.604
Comics	0.326	0.225	0.115
Pictures	1.307	0.266	0.258
Covers	0.003	0.000	0.000
Physibles	-	-	0.000
Other	0.410	0.278	0.291

PORN appears as the second most popular category among BitTorrent users. Contrary to VIDEO, PORN presents a steady increase in its weight since it accounts for 17% of the download sessions in pb10, 24% in pb11 and 31% in pb12. The growth in the PORN's share (14 percentage points) almost matches the VIDEO category drop (12 percentage points). Finally, it is very important to notice that the sum of these two categories represents about 90% of the total downloads for the three snapshots. More interestingly, by zooming in our analysis into the subcategories, we realize that out of that 90%, 80% belongs to the subcategories: VIDEO/Movies, VIDEO/TV Shows, VIDEO/Highres-Movie, VIDEO/Highres-TV Shows, PORN/Movies, PORN/Highres-Movies.

The popularity of High-resolution PORN and VIDEO content follows the increasing

availability of this type of content. While it only attracted 1.87% of the downloads in pb10, it has increased its popularity 5 times by receiving 9.62% of the downloads in pb12.

If we analyze the remaining categories: (i) we find that AUDIO contributes 5% of the downloads (with variations smaller than 1 percentage point over the trhee snapshots). (ii) APPLICATIONS goes from 2% in pb10 to less than 1% in pb11 and pb12. It is worth noting that APPLICATIONS category contribution is mainly due to Windows applications. (iii) GAMES starts at 1.2% in pb10, gains 1 percentage point in pb11, and loses it again in pb12. (iv) Finally, the OTHER category remains stable around a 3% with variations smaller than 0.5 percentage points.

In a nutshell, PORN is compensating for loss in VIDEO, which in the worst case attracts 3/5 of the downloads. Both categories together account for 90% of the downloads. Furthermore, we observe a significant increase in the High-resolution content. Finally, the rest of the categories remains steady over the time with very small variations showing a small but stable interest from BitTorrent consumers in each one of them.

3.2.5.3 Content Availability Vs Content Popularity Discussion

The most relevant content in BitTorrent (according to its major portal, TPB) in terms of availability and popularity are Movies (including porn ones) and TV Shows. Although this type of content represents only 1/2 of the available content, it accounts for 4/5 of the downloads.

In the case of PORN content we perceive a stable presence in the available content (a bit higher than 20%), but an increment of its popularity based in the portion of downloads, from 17% to 31%. In particular, PORN is taking up the popularity reduction suffered by the VIDEO category. Similarly to VIDEO, the proportion of available content for PORN is lower than its weight in number of downloads (except in pb10).

For the rest of the categories the portion of available content exceeds the portion of downloads. The AUDIO category represents between 12%-15% of the available content but only attracts 5% of the downloads. In the case of the GAMES category, it contributes between 3%-5% of the content to get 1%-2% of the downloads. The OTHER category feeds 7%-9% of the content (without considering pb09) and only captures 3% of the downloads. Finally, the APPLICATIONS category contributes 10%, 4% and 5% of the content in pb10, pb11 and pb12, to attract 2%, 1% and 0.8% of the downloads, respectively.

Therefore, we can conclude that if TPB removes all the categories except VIDEO and PORN, although it would lose half of its available content, it would not suffer a significant reduction in the downloading activity. In addition, the results suggests that High-resolution content is rapidly increasing its availability and popularity.

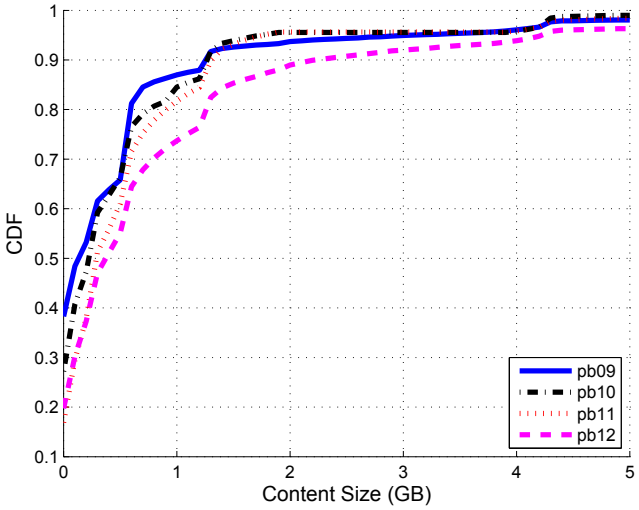


Figure 3.1: Torrents Size CDF

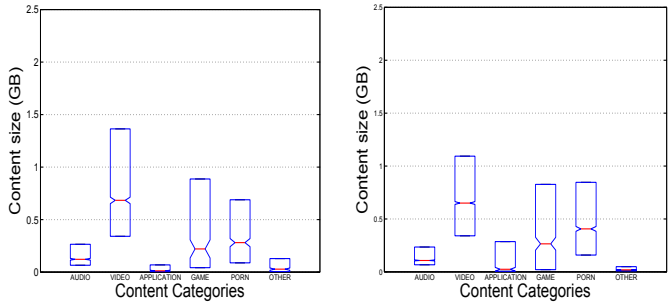
3.2.6 BitTorrent's Content Size Analysis

In this subsection we characterize the evolution of the content size across the four snapshots. This allows us to understand whether the size of BitTorrent content is increasing linked to the presence of everyday larger content in the multimedia arena (e.g increment of High-resolution content presence). To perform this discussion, we will first look at the aggregate content size distribution across all the snapshots to later narrow down our analysis to individual categories.

3.2.6.1 Aggregate Content Size Distribution

Figure 3.1 depicts the CDF of content size for our four snapshots. For a better understanding, the graph only shows the CDF for content up to 5 GB (that includes the DVD standard size of 4.7GB), which accounts for more than 96% of the content within our dataset. The graph shows a steady increase of the content size over the 2-years period under study. The median value of the content size in pb09 was 223MB and increased by 53% (to 341MB) in the next five months (pb10), and it kept growing up to 370MB and 458MB in pb11 and pb12 respectively. The conclusion is that BitTorrent content has doubled its size (in median) in a period of 2 years.

We also want to highlight that the content larger than a standard DVD of 4.7GB (not included in the graph) increases its representativeness by almost 2 percentage points from 2.06% in pb09 to 3.85% in pb12. In addition, it is interesting to notice that all snapshots contain some content of huge size. In all the cases (except pb10 in which the largest content



(a) pb09 (b) pb10

(c) pb11 (d) pb12

Figure 3.2: Box plot of content size per category for pb09, pb10, pb11 and pb12 datasets. For each category we show the 25th, 50th (median) and 75th percentiles represented by the bottom horizontal blue line, the middle horizontal red line and the top horizontal blue line, respectively.

is 66GB) we can find contents above 150GB. By manual inspection we have discovered that they are actually collections of files (e.g. several seasons of a TV series) that all together reach that size.

3.2.6.2 Content Size per Category

Having depicted the overall picture for the content size, we devote our efforts on dissecting our analysis for the different categories. Figure 3.2 shows the box plot (which includes 25th-percentile, 50th-percentile or median, and 75th-percentile) of the content size for every category in each one of the four snapshots. The obtained results allow us to divide the categories into two groups: (*i*) low-size categories composed by AUDIO, APPLICATIONS and OTHER, and (*ii*) large-size categories formed by VIDEO, GAMES and PORN.

Among the low-size categories, AUDIO is the one presenting a larger size and a very stable distribution over the 2-years period under study. The median size for AUDIO is around 120MB with a small variance. Its 75th-percentile doubles the median and stays close to

250MB, except in pb11 that goes above 300MB. The APPLICATIONS and OTHER categories show very low median values below 25MB in all the datasets. The only remarkable issue for these two categories happens for APPLICATIONS in pb10, which shows a much higher 75th-percentile (285MB) than in the other cases.

In the large-size categories, VIDEO is the one with the largest median content size over time. It is interesting to observe that it presents a quite stable median value with only a slight variation between 650MB and 700MB. While the 25th-percentile is also stable (∼340MB) in the four snapshots, the 75th-percentile presents a moderate increment of 31% from its lowest value of 1.1GB in pb10 to the highest one of 1.44GB in pb12. PORN ranks as the category with the second largest median content size. In the case of PORN all percentiles grow over time. For instance, the median size evolves as follows: 280MB (pb09), 405MB (pb10), 434MB (pb11) and 558MB (pb12). This demonstrates that PORN content has doubled its median size in a period of only two years. Finally, GAMES is the category that presents a major variance between the different percentiles. The 25th-percentile and the median are always lower than the same parameter in the VIDEO and PORN categories. However, the 75th-percentile becomes the largest one in pb11 and pb12. These large variability between the different percentile thresholds occurs because we can find a large set of games with a very small size (e.g for smartphones, portable videoconsoles, etc), and at the same time a large set of games of very big size (e.g. DVD, Blue-ray, etc). In particular, the extreme variability shown in pb11 and pb12 responds to the recent appearance of video-game consoles and PCs that are Blue-ray capable and the increasing presence of Blue-ray games in the market.

3.2.6.3 Content Size Increment Discussion and Implications

First of all, it must be noted that those categories that massively contribute to BitTorrent (*i.e.* VIDEO and PORN) happen to be the ones presenting a larger size, while those categories with a minor presence contain content of small size. The only exception is the GAMES category that shows an extreme variability in the size of its content, especially in pb11 and pb12 snapshots.

The increment of the content size can be explained by 3 main factors: (*i*) The important evolution in the availability of High-resolution content (large size) which in pb12 already represents 8.2% of the content. (*ii*) PORN that represents more than 20% of the content is doubling its size (in median); and, (*iii*) 75th-percentile for VIDEO content size (which represents 40%-50% of the available content) has increased a 31%, probably due to the major presence of High-resolution Movies and TV Shows.

The fact that BitTorrent content (and by extension multimedia content in the Internet) has doubled its size in the last two years is something that major Internet players (ISPs,

Table 3.4: Percentage of contents with comments in different categories (contents with at least 1 comment — contents with three or more comments)

Category	**pb09** (%) ≥1 | ≥3	**pb10** (%) ≥1 | ≥3	**pb11** (%) ≥1 | ≥3	**pb12** (%) ≥1 | ≥3
AUDIO	39.21 | 10.79	37.29 | 11.43	34.99 | 9.46	26.91 | 5.75
VIDEO	41.42 | 15.06	46.32 | 15.56	33.00 | 9.74	32.83 | 10.12
APP.	47.93 | 19.28	72.03 | 31.36	61.75 | 25.28	53.44 | 20.84
GAMES	63.83 | 34.89	66.06 | 30.30	67.33 | 29.45	54.11 | 23.19
PORN	33.50 | 6.32	31.62 | 6.53	15.41 | 2.56	12.36 | 2.08
OTHER	37.24 | 11.99	33.28 | 7.89	44.98 | 13.76	31.41 | 9.29
Aggregate	41.25 | 14.01	42.66 | 13.74	31.80 | 9.30	28.35 | 8.29

multimedia content providers, CDN operators, etc) need to take into account in order to update their infrastructures and resources. For instance, if the content growth speed depicted by our results remains stable over time, CDN operators will need to provision their servers with larger caches, content service providers (e.g. Cyberlockers) will need to double its storage capacity every two years, etc.

3.2.7 User Comments on BitTorrent Contents

An interesting aspect related to the BitTorrent content analysis is to study the users' interaction and feedback. In order to measure such activity we have crawled the TPB page of each content in our dataset (unless they had been removed from TPB) to capture the number of comments that BitTorrent users wrote. With this data we are able to study how the number of comments have evolved over the period under study.

Table 3.4 shows the percentage of content aggregated and per category that received at least one comment on their TPB page as well as the portion of content that collected 3 or more comments.

First of all, if we look at the aggregate content results we conclude that the social activity around BitTorrent content is quite reduced and the users are just focused on accessing the content without sharing much about its experience. The portion of content receiving three or more contents is in the best case 14% (pb09) . In addition, we did not find any content with more than nine comments. Furthermore, the number of comments per content decreases over the time. For instance the portion of content that presents at least one comment goes down from more than 40% in pb09 and pb10 to 32% and 29% for pb11 and pb12, respectively. This happens because the time is an important variable that increases the likelihood that a content receives one or more comments, the longer the content is exposed the more likely is that a user comments on it.

The GAMES and APPLICATIONS categories are the ones containing a larger portion of content with comments. Although in pb09 GAMES is largely leading this ranking, in the other snapshots both categories present similar results alternating in the first position. We

could roughly account between 55%-65% of the content with at least one comment and 20%-30% with three or more comments. GAMES and APPLICATIONS content usually requires some particular knowledge to manage them (e.g. what movement each button generates in a video game, how to find different options in an application, etc). In addition, in the case of applications the installation process could be challenging for non-skilled users. These two factors increase the need for BitTorrent consumers to interact with the content publisher (or with other consumers of the same content) in order to solve some issue and manage the downloaded game or application.

After the two leading categories, we can find a second group that includes AUDIO, VIDEO and OTHER categories. We can establish rough intervals of 30%-40% and 5%-15% for the portion of content in those categories presenting at least one comment and 3 or more comments in their TPB page, respectively. This makes a relevant difference of 15 percentage points from the two previous categories.

Finally, the PORN category attracts the fewest comments from end-users. Roughly 15%-30% of the PORN content shows at least one comment, and only 2%-6% has three or more comments. PORN is not only the category with the smallest portion of content attracting comments, but it is also the one experiencing the largest reduction of this parameter over the time. It loses 20 percentage points from 33% (in pb09) to 12% (in pb12) during the time period under study. It is obvious that PORN is a very controversial content that is still considered immoral in much of the world, and even forbidden in many countries. Therefore, although it is massively consumed (2nd category in content availability and popularity), consumers prefer not to comment on it.

In a nutshell, this subsection demonstrates the low interest of BitTorrent users in commenting on their downloaded content.

3.2.8 Conclusion

This study has presented a thorough analysis on the evolution of multimedia content available in the Internet based on the content available in the most popular BitTorrent portal in a two years period between Nov. 2009 and Feb. 2012. Our results predict a steady and important increment of the multimedia content traffic, which already represents the major part of the Internet traffic, sustained in three main findings: (*i*) Multimedia content has doubled its size in a period of only 2 years, (*ii*) the major part (80%) of the consumed multimedia content corresponds to TV Shows and Movies (including porn) that belong to those categories with a largest size , and (*iii*) High-resolution content, which has very large size, is increasing its presence and it already represented 8% of the available content and 10% of the downloads in our most recent snapshot dated at the beginning of 2012. These findings are useful to those Internet players (*i.e.* ISPs, CDN operators) involved in

the content distribution business in order to update their infrastructures, resources and algorithms to efficiently distribute and serve multimedia content. Furthermore, if the size of the multimedia content keep growing as in the last two years CDNs and content providers (e.g. Cyberlockers, multimedia content distribution platforms, etc) should properly dimension the storage capacity of their caches and servers to cope with the distribution of larger content.

3.3 Reaction of BitTorrent Content Publishers to Antipiracy Actions

3.3.1 Introduction

During the past decade, the Internet has witnessed an increasing level of online piracy of copyrighted content. In particular, Peer-to-Peer content distribution applications (*e.g.*, BitTorrent, Gnutella) and Cyberlocker services (*e.g.*, Megaupload) have facilitated illegal sharing of copyrighted content. At the same time, the availability of copyrighted content by these systems at no cost, has led to an explosion in their popularity and therefore to their contribution in overall Internet traffic. While legal actions were taken against few major and many minor violators who illegally published, consumed or facilitated the distribution of copyrighted content, online piracy appears to become even more widespread in different countries. In recent years, these trends have prompted copyright holders to demand the legislation and implementation of more effective online antipiracy laws in several countries. However, such an effort has faced strong opposition by various stake holders in several countries. In fact, we are only aware of a small number of countries that have legislated and implemented an online antipiracy law. Given the difficulty to put in place an online antipiracy law, an interesting question is *"whether or not and to what extent an antipiracy law and its associated enforcement actions can affect the behavior of violating users?"*

This intriguing question is very difficult to answer for at least three reasons as follows: First, the effect of an antipiracy event (*e.g.*, publicizing relevant laws or enforcement actions) can be assessed on different groups of users including those who publish or consume copyrighted content, users with different levels of involvement (as publisher or consumer), or users for a specific system or in a particular country. Clearly, the impact of an antipiracy event could vary significantly across different groups. Second, there could be other (potentially some unknown) co-existing social, economical, and technical factors that have a dominant and possibly opposing effect on piracy behavior among users. More importantly, it is very challenging to identify and capture all the relevant major factors, and assess their level of impact on piracy behavior among users. For example, the drop in the number of online pirating for movies in the the US could be due to a combination of antipiracy actions against a few users and/or due to user access to cheap and legal content via Netflix. Furthermore, the effect of an antipiracy action could be short- or long-lived. Third, there is no ground-truth to reliably validate any finding about user reactions to antipiracy events. A survey of users can be conducted to obtain a more accurate view of the behavior for a relatively small group of users (*e.g.*, few thousands). However, only a small fraction of surveyed users may be involved in antipiracy and those users may not indicate their intention because of any concern for legal action against them.

Despite these challenges, a few recent studies have examined the effect of specific antipiracy actions on the behavior of a particular group of users (i.e. consumers) in a single country using measurement [74], or survey of users [75] or businesses [76]. All these studies presented a collection of evidences to illustrate that the enforcement of local antipiracy laws succeeded in reducing the downloading activity of copyrighted content among their target group of consumers. To our knowledge, the effect of antipiracy actions on content publishers have not been examined, and it is essential because they feed the ecosystem of online piracy and in some cases gain substantial profit [52].

In this study, we investigate the effect of antipiracy actions on the publishers of copyrighted content. To cope with the challenges in tackling such a broad question, we limit the scope of our study in two ways as follows: First, we only examine the effect of two major antipiracy actions: *(i)* the closure of Megaupload was a sudden event that was publicized worldwide, and *(ii)* the French antipiracy law (*Hadopi* law) that was debated, legislated and fully implemented over a two year period. We intuitively expect these antipiracy actions to have a dominating impact on the behavior of their corresponding group of users. Therefore, any potential error in our analysis due to potentially unknown factors should be relatively small. Second, we only consider the effect of these two antipiracy actions on the content publishers in the largest BitTorrent portal, namely the *The Pirate Bay* (TPB). Since a significant majority of BitTorrent publishers upload copyrighted material [77] [78], they provide a large population of publishers that are actively engaged in online piracy and therefore their reactions offer relevant and meaningful insights for this study.

One key contribution of this study is our methodology to leverage the reaction of BitTorrent publishers for assessing the effect of selected antipiracy actions. Toward this end, we capture snapshots of all BitTorrent publishers along with their uploaded (and downloaded) files through TPB. The timing of our snapshots are properly aligned to the target antipiracy actions to increase the likelihood of detecting any measurable effect even if its impact is short-lived. We use the changes in the daily number of relevant BitTorrent publishers and their contribution over the proper time frame as our basic metrics to assess the effect of each antipiracy action. We show that this basic metric does no always paint a clear picture of publisher behavior. Therefore, we deepen our analysis by grouping publishers based on different criteria to identify the most likely cause of the observed changes in publishers' behaviour. These criteria include: (*i*) level of activity (*e.g.*, active vs casual publishers), (*ii*) publishers' business profile (*e.g.*, profit-driven vs altruistic) or (*iii*) monitoring policies of their hosting facilities (soft vs strict). Finally, we corroborate our findings with a few independent sources including Google trends and other reports to gain more confidence. While there is not ground truth to validate our findings, we believe that the number of discovered evidences and their temporal alignment offer a very convincing explanation for

how these selected antipiracy actions have influenced the behavior of BitTorrent publishers.

The second key contribution of this study is to demonstrate some of the subtleties in identifying the potential effect of an antipiracy action, and properly relating them to their cause. These findings of our "detective work" are summarized as follows:

The closure of Megaupload: Many publishers joined BitTorrent most likely from Megaupload (and other Cyberlockers) right after its closure. This resulted in an increase in the overall number of TPB publishers but, surprisingly, had no impact on their overall publishing rate. This is due to the fact that major BitTorrent publishers that maintain a private BitTorrent portal (*i.e.* a similar business to a Cyberlocker) reduced their publishing rate in reaction to this event.

French antipiracy Law: The French population have followed the legislation and implementation of the 3 strike law that targets both consumers and publishers on any copyrighted content through P2P applications. We show that the first two steps of the Hadopi law have been very effective in decreasing the number of casual publishers that as we demonstrate, are indeed active consumers. However, the number of active publishers (*i.e.* uploading more than one content per day on average) remained stable and they considerably increased their publishing rate. This reaction is surprising given the reduction of French consumers for copyrighted content through P2P applications as reported in [75], [79]. Our closer examination revealed that most of top French publishers do not publish any French content. In fact, the concentration of these publishers in a particular hosting facility in France appears to be motivated by the absence of a strict policy for avoiding the use of BitTorrent on its servers. These professional publishers are legally savvy and realize that the opportunity to freely operate simply outweighs any unlikely antipiracy action as a result of the Hadopi law.

3.3.2 Related Work

There has been several studies by behavioral scientists on the motivation of users to engage in online piracy [80–83]. These studies typically rely on the collected data from a small-scale survey (a couple of hundreds user). We are only aware of two prior measurement studies on the effect of antipiracy events on illegal file sharing among Internet users.

First, Alcock et al. [74] recently analyzed the impact of the New Zealand antipiracy law on different applications. They monitored the traffic of DSL connections for 4000 users at three different time periods in 2011 and 2012. Their study demonstrates that the consumption of copyrighted material has decreased among users and concludes that this is the effect of the local antipiracy law. Similar to this study, our work relies on a collection of evidences to draw a conclusion about user behavior. However, we focus on the behavior of publishers (rather than consumers) who are clearly engaged in online piracy.

	pb10	pb11	pb12
Crawling Period	04/09/10 - 05/05/10	10/21/11 - 12/13/11	01/28/12 - 02/12/12
Duration (days)	27	54	16
Pubishers (username)	7.1K	6.9K	3.3K
Torrents	38.2K	72.0K	21.0K
Consumers	27.3M	25.6M	5.1M
Downloads	95.6M	79.0M	11.1M

Table 3.5: Dataset Description

Second, Lauinger et al. [84] examined the contents of a large number of uploaded files in eight Cyberlockers to measure the impact of Megaupload closure on the availability and lifetime of copyrighted files in other Cyberlockers. They demonstrate that after Megaupload closure, other Cyberlockers proactively increased the filtering of copyrighted material from their servers, as had been reported in the press [85]. In addition, in the same study the authors present a qualitative discussion of the potential impact that the SOPA (US) law [86] could had achieved if it had been implemented. While this work is in spirit similar to ours, we examine the behavior of thousands of BitTorrent publishers (uploading tens of thousands of files).

We are also aware of two prior reports on the effect of the Hadopi law on French Internet consumers. Preliminary results of a longitudinal survey of 2K users carried out by the Hadopi commission indicated a decrease in the download of copyrighted material [75]. In particular, 72% of the respondents to this survey who had received a warning, declared that they had decreased or stopped their activity, and 50% of the respondent indicated that they have increased the consumption of legal copyrighted content. The second study [76] analyzed the data from iTunes record sales by four major labels and reported 25% increase in the purchase of iTunes music among French users after the enforcement of the Hadopi law while the increase in a neighbor country such as Spain was negligible. Based on these two pieces of evidence, they concluded that the Hadopi law has been successful in deterring P2P downloads of copyrighted material.

To the best of our knowledge, our study is the first investigation of the effect of antipiracy actions on the behavior of publishers of copyrighted content that is based on measuring a large number of such publishers. More importantly, our investigation goes beyond obvious metrics and reveals the impact of other social, economical, and technical factors through data-driven analysis.

3.3.3 Data Collection and Datasets

Our objective is to capture multiple snapshots of the BitTorrent ecosystem over time in order to characterize longitudinal trends in the population and activity of publishers. We use these trends to assess the impact of antipiracy events on content publishers. Towards

this end, we leverage active measurement over The Pirate Bay (TPB) portal using the methodology and tools that were developed in a previous study [52]. We focus on TPB in this study since it is the most popular BitTorrent portal as reported by scientific studies [47] and Alexa ranking [1]. In particular, TPB is one of the top-100 most popular websites in the Internet and receives at least twice (and in most cases significantly larger) daily visit than any other BitTorrent portal based on Alexa information. Furthermore, all the indexed content on TPB portal are explicitly uploaded by a publisher in contrast to the other major portals (*e.g.*, Torrentz or IsoHunt) that use crawling techniques to identify their indexed content. These features make TPB a suitable venue to capture snapshots of the BitTorrent ecosystem and conduct our analysis. This subsection describes a brief overview on BitTorrent and our measurement methodology as well as the main characteristics of our collected datasets.

3.3.3.1 Background on The Pirate Bay

TPB is simply a rendezvous point between content publishers and consumers. When a publisher wishes to make a content available within the BitTorrent ecosystem, its first step is to generate a unique id known as the *infohash* and register the content with one (or multiple) tracker(s). A tracker keeps track of the IP addresses for a group of peers that concurrently participate in the delivery of a content (*i.e.* form a swarm). A participating peer can be of two types: peers with a complete copy of a content are known as *seeders* while other peers are *leechers*. Therefore the content publisher is the first seeder in a swarm. The second step is to advertise the content by generating a torrent file that provides meta-information for consumers including the IP address of the associated tracker(s). The publisher uploads the .torrent file to TPB and possibly other BitTorrent portals. In the case of TPB, the publisher needs to be registered with the portal and uses her account (with a specific username) to advertise a content. TPB creates a separate webpage for each registered user in which all its published content along with publishing times are listed. Finally, TPB offers an RSS ("Really Simple Syndication") service where consumers can subscribe and receive a notification as soon as a new content becomes available. To download a content, a consumer typically retrieves the .torrent file from a portal, extracts the IP address of the tracker and connects to it. The tracker provides a list of IP addresses for a random subset of participating peers in the swarm to the new peer so that the new peer can connect to them and join the swarm.

3.3.3.2 An Overview of the Measurement Methodology

Our measurement tool can capture a rather complete snapshot of all active publishers, their published files and associated consumers within a window of time. To achieve this

goal, our tool subscribes to TPB's RSS service to get a notification for any new content that is published on the portal[1]. The RSS feed provides the .torrent file along with the username of the content publisher. Our tool retrieves the IP address of the tracker from the .torrent file (or the magnet link) and immediately connects to it. By connecting to the tracker immediately after the content is published, we are able to identify the IP address of the initial seeder (*i.e.* the publisher's location) in many torrents. Our tool periodically connects to the tracker to retrieve the IP addresses for (typically) 200 randomly-selected participating peers (*i.e.* consumers) while respecting the reconnection time imposed by the tracker in order to avoid being banned. To cope with this limitation, our tool probes a tracker from eight geographically-distributed nodes in parallel and captures the IP address of a majority of consumers. We use MaxMind [87], an IP-to-geo mapping database, to determine the location of discovered publishers and consumers. In summary, our captured snapshots contain the following information for each published torrent on TPB portal: *(i)* publisher's username and IP address, *(ii)* list of IP addresses for associated consumers. Further details for a similar measurement methodology can be found at [52].

3.3.3.3 Datasets

Using our measurement tool, we have collected three snapshots of TPB system in April 2010, Nov 2011 and Jan 2012. Table 3.5 summarizes the main characteristics of each snapshot including: crawling period, the number of unique publishers, consumers, torrents (*i.e.* published files) and downloads for the three datasets labeled as pb10, pb11 and pb12. Each dataset was collected over a sufficiently long time such that any daily or even weekly variations among users and their activities are captured. The pb11 and pb12 snapshots are captured shortly before and after the closure of Megaupload site [88]. Therefore, we use these two snapshots to examine the impact of Megaupload closure. Moreover, our pb10 and pb11 snapshots were collected 18 months apart and are used to investigate the effect of French antipiracy law on French users.

3.3.4 Effect of a Global Antipiracy Event

In this subsection, we investigate how BitTorrent publishers reacted to a major antipiracy action, the closure of Megaupload [89]. We focus on this antipiracy action against Megaupload because it was a major player in illegal sharing of copyrighted content. Megaupload was the most popular Cyberlocker website. Cyberlockers provide storage service to end users that enables them to share their online stored content with other users through a

[1]Note that since Feb. 2012, TPB only indexes magnet links instead of .torrent files. We have accordingly updated our tool to properly operate with this new indexing strategy.

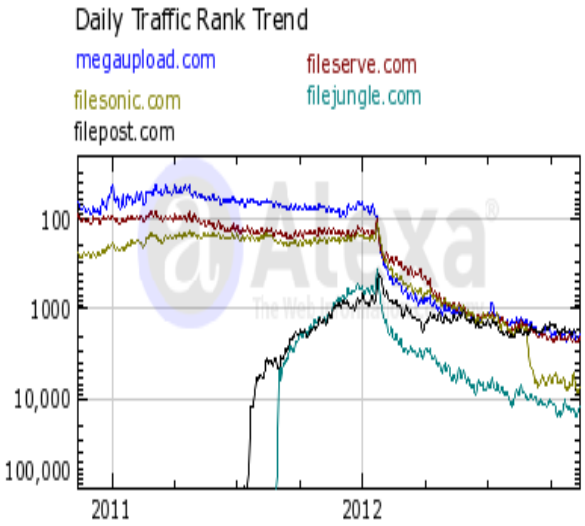


Figure 3.3: Evolution of Alexa ranking for five popular Cyberlockers over the last two years (source Alexa).

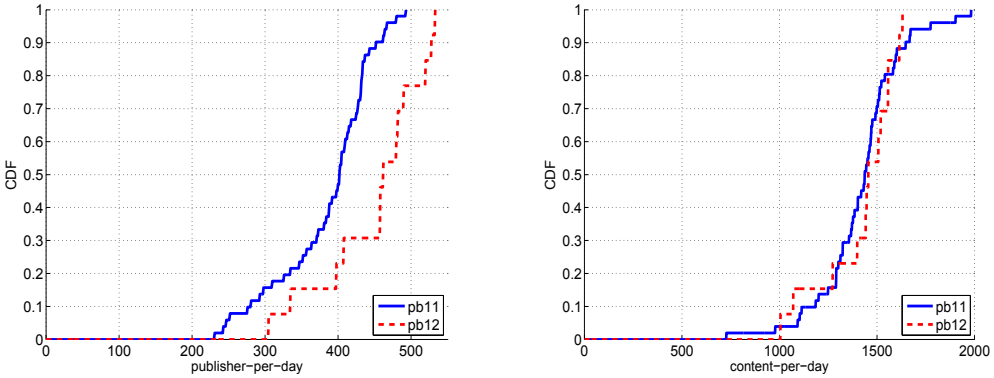
URL. They have quickly become very popular among users for sharing copyrighted material (*e.g.*, movies, TV shows, music, etc) through their websites [90]. These websites became very profitable through posting from ads and selling premium subscriptions that provide end users with a better experience (*e.g.*, higher download rate). Moreover, they even encouraged users to publish interesting content by offering some income to publishers whose content became popular [91]. To illustarte the popularity of Megaupload, we note that Megaupload had 180M registered users and 50M daily visitors, and stored 12 billions unique files with the aggregate size of 25 petabytes [89]. We first provide some info about Megaupload closure and then examine its impact on BitTorrent publishers.

3.3.4.1 The Closure of Megaupload

On January 19th 2012, the FBI (in coordination with other agencies across multiple countries) shut down Megaupload website and arrested their owners on charges of worldwide online piracy that produced $175M unlawful income and caused $500M loss for the copyright owners [92]. This antipiracy event had a worldwide coverage. To demonstrate the overall effect of this well publicized event on the Cyberlockers' ecosystem, Figure 3.3 presents the evolution of the Alexa ranking [1] for five popular Cyberlockers over the past two years. This figure shows two points: *(i)* Before Megaupload closure, all of these cyberlockers were either already among the top-200 websites in Alexa ranking or their ranking was rapidly improving until the closure of Megaupload. *(ii)* After the closure of Megaupload, the ranking of all Cyberlockers (and thus their popularity) were rapidly and consistently dropping.

	pb11	pb12
Avg. daily publishers	367	420 (+14.4%)
Avg. daily contribution	1334	1314 (-1.5%)

Table 3.6: Aggregate results for publishing activity in BitTorrent. The table shows the average daily publishers and the average daily uploaded content in pb11 and pb12. The value in parenthesis indicates the relative difference between pb11 and pb12.



(a) CDF daily publishers (b) CDF daily content

Figure 3.4: CDF for the number of daily publishers and daily contribution in pb11 and pb12 datasets.

This effect could be due to the adoption of new strategies by Cyberlockers to actively remove all copyrighted content as reported by a recent study [84] and in press [85]. In summary, these evidences confirm that the closure of Megaupload had a significant impact on all Cyberlockers and possibly all systems that facilitated illegal sharing of copyrighted content.

3.3.4.2 Effect on BitTorrent Publishers Activity

We rely on our pb11 and pb12 BitTorrent snapshots that were collected shortly before and after the closure of Megaupload. Given the short time between both snapshots and our target event, and the fact that (to the best of our knowledge) no other major relevant event occurred during this period, we are confident that any change in the behavior of BitTorrent publishers is most likely triggered by the closure of Megaupload. We use two metrics to measure the effect of Megaupload closure on BitTorrent publishers as follows: *(i)* the average daily number of active BitTorrent publishers, and *(ii)* the average daily number of discovered uploaded content. Using these daily average values enables us to compare these characteristics of publishers across different datasets despite the differences

n=Avg. content/day	Active (n≥10)	Regular (1≤n<10)	Casual (n<1)
Avg. daily publishers pb11	13.6	92.4	261
Avg. daily publishers pb12	14.7	113.8	291.5
Avg. daily publishers difference	+1.1 (+8.1%)	+21.4 (+23.1%)	+30.5 (+11.3%)
% new publishers in pb12 after Megauplaod	6.25%	15.23%	42%
Avg. daily contribution pb11	471	423	440
Avg. daily contribution pb12	374	510	431
Avg. daily contribution difference	-97 (-21%)	+87 (+17%)	-9 (-2%)

Table 3.7: Number of publishers and daily contribution for next groups of publishers classified based on their contribution to the system. Active publishers (n≥10 content/day, Regular publishers (1≤n<10 content/day), and Casual publishers (n<1 content/day)

in dataset durations[2].

Table 3.6 presents the average daily number of publishers and uploaded files for snapshots pb11 and pb12. A more detailed view of these characteristics is provided in Figures 3.4(a) and 3.4(b) that depict the distribution of daily number of publishers and uploaded files for both snapshots, respectively. These statistics reveal that the average number of publishers increased by 14% over 1.5 months whereas their activity remained roughly unchanged. The observed increase in the number of publisher (over such a short time) is surprising and is very likely caused by the migration of publishers from Megaupload (and other Cyberlockers) to BitTorrent after the closure since we are not aware of any other event during this period that can explain such increment. To validate this observation, we take a closer look at the timing of published files by individual publishers in the pb12 dataset. To obtain this information, we have crawled the TPB page of all publishers in our pb12 dataset and captured the number of files they uploaded in each day during the 75 day window between two snapshots (12/01/2011 to 02/12/2012). We observe that 42% of the publishers in pb12 snapshot published their first file after the closure of Megaupload which suggests that they most likely joined BitTorrent after this event. Using this information, we have also determined the aggregate number of active pb12 publishers in each day during this period as shown in Figure 3.5. This figure demonstrates that the number of daily publishers is relatively stable around 200 until the date of Megaupload closure and then it rapidly doubles in a few days once the implications of the event becomes clear to publishers. *These evidences collectively suggest that our observed changes in the publishers' demographics and activity between pb11 and pb12 must be due to the closure of Megaupload.*

Active vs Casual Publishers The lack of increase in the daily number of uploads despite the clear growth in the daily number of publishers after the Megaupload closure is

[2]We evaluated both metrics for different time windows in pb11 (54 days) and they remain the same independently of the used window.

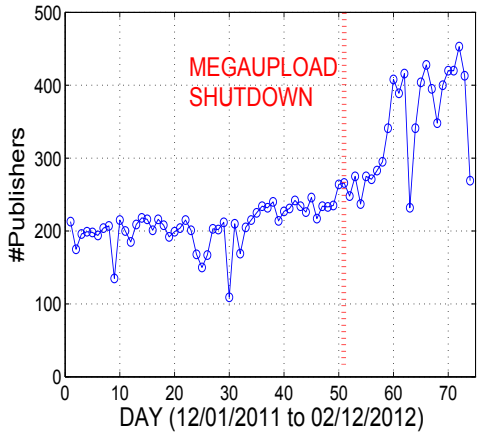


Figure 3.5: Daily number of publishers among those ones collected in our pb12 snapshot during the period 12/01/2012 to 02/12/2012.

username	pb11 day cont.	pb11 rank	pb12 day cont.	pb12 rank	Business	URL
scenebalance	107.17	1	31.06 (-71%)	4	BT Private Portal	www.scenetime.com
TvTeam	80.9	2	65.94 (-18%)	1	BT Private Portal	www.torrentday.com
exmnova	58.38	3	35.38 (-39%)	3	BT Private Portal	www.69bits.com
sceneline	53	4	29.93 (-43%)	5	BT Private Portal	www.speed.cd
chkm8te	33.96	5	59.5 (+75%)	2	Promoting Website	www.4ufrom.me
UltraTorrents	24	6	6.12 (-74%)	26	BT Private Portal	www.ultratorrents.com
FluxXxu	16.46	7	19.6 (+19%)	7	Promoting Website	www.starpix.us
RockSaltS	15.13	8	6.68 (-55%)	24	Promoting Website	http://jolypic.com/
adultvideotorrents	13.06	9	2.375 (-82%)	92	BT Private Portal	www.adultvideotorrents.com
.BONE.	12.11	10	5.25 (-56%)	35	Altruistic	
Black1000	11.98	11	OUT	OUT	Altruistic	
MirrorRu	11,65	12	OUT	OUT	Fake	
bigbluesea	11.63	13	OUT	OUT	Altruistic	
eztv	11.39	14	8 (-29%)	18	BT Private Portal	http://eztv.it/
scene4all	10.29	15	10.56 (+2.56%)	13	Altruistic	

Table 3.8: A summary of main characteristics (daily contribution rate and business profile) of the 15 active BitTorrent publishers in pb11 and the changes in their level of publishing between pb11 and pb12.

counter-intuitive. To explain this finding, we divide the publishers in both datasets based on their average daily contribution into the following three classes: *Active* publishers that upload more than 10 contents per day on average, *Regular* publishers that upload between one and 10 contents per day, and *Casual* publishers that contribute less than one content per day. Table 3.7 shows the average number of publishers per day from each class and their aggregate daily contributions in snapshots pb11 and pb12. We first analyze the results for pb11 snapshot as the starting point and later discuss the evolution of each class between pb11 and pb12.

It is interesting to notice that while the number of casual publishers in pb11 is roughly three and 20 times larger than the number of regular and active publishers, respectively, the overall daily contribution of all three groups is roughly the same (between 420 to 470

files a day) before the Megaupload closure. Between pb11 and pb12, the number of casual publishers has increased by 11%, while their contributions remain unchanged (less than 2%) during this period. The number of regular publishers has increased by 23% and this has led to a roughly proportional increase (17%) in their daily contribution. Finally, the number of active publishers grew by roughly 8% but their contribution dropped by 21%. In addition, 42% of casual, 15% of regular and 6% of active publishers in pb12 are newcomers who joined BitTorrent during this period.

In summary, most of the newly arriving BitTorrent publishers after Megaupload closure are casual or regular publishers. The overall contribution of three groups were rather balanced before the Megaupload closure. However, after the closure, the increase in the contribution of regular publishers is roughly the same as the decrease in the contribution of active publishers which led to the unchanged overall rate between two snapshots. This raises the question that "why the relatively small number of very active publishers have dropped their publishing rate after the Megaupload closure?" We tackle this question in the next subsection.

Business Profile of Active Publishers The active publishers that upload more than 10 files a day are in most of the cases professional publishers behind profitable websites as it is demonstrated in [52]. Therefore, their wide spread reaction to (measurably) lower their contributions right after the Megaupload closure must be related to this event. To explore this issue, we employ a similar methodology to the one used in [52] to determine the business profile of all 15 active publishers in pb11. The basic idea in this methodology is to download a few published files by a publisher and manually inspect whether, where and how a consumer might be redirected to another web site associated with the publisher. This methodology broadly divides publisher profiles into the following categories: *(i) BitTorrent Private Portals* are associated with private trackers that offer a better experience to BitTorrent users for a seeding ratio or a fee. *(ii) Promoting Websites* basically publish content for the sole purpose of attracting users to their web sites that are often hosting image services. *(iii) Fake Publishers* are either antipiracy agencies or malicious users that inject fake (non-existent) content in order to warn users of downloading copyrighted material or infect their computers, respectively. *Altruistic Publishers* simply publish content to share with others without any expectation of direct gain.

Table 3.8 shows the following information for the 15 Active publishers in pb11 in different columns: their usernames, daily publishing rate, rank in pb11 and pb12, their business profile and the URL to their web site[3] (if applicable). Note that the Top five publishers

[3]Our goal in providing the identity of these publisher is to demonstrate the fact that many of these publishers are indeed real companies.

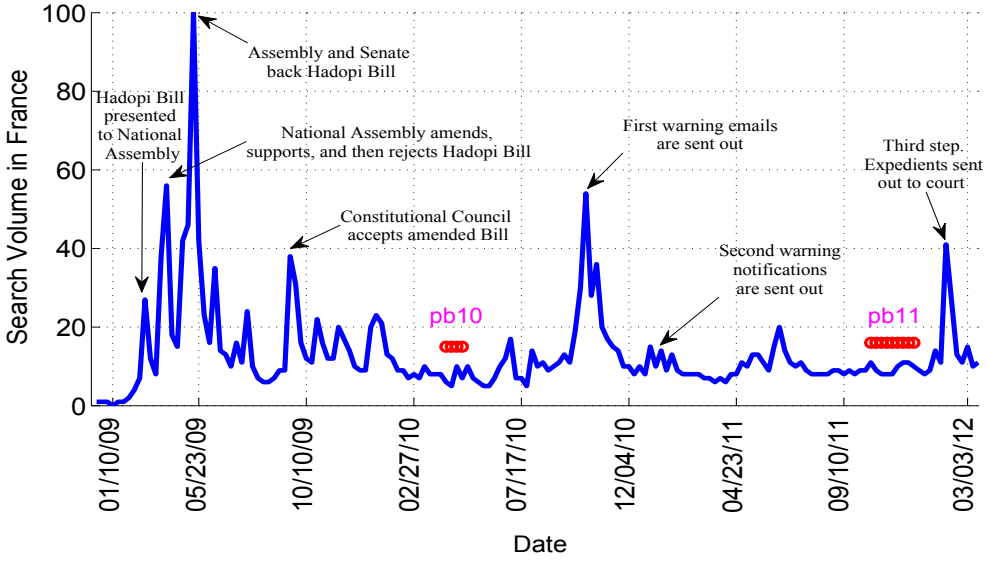


Figure 3.6: Volume search in France of the keyword Hadopi during the period Jan. 2009 to Feb. 2012 according to Google Trends. In addition, we describe the events associated with major searching peaks.

in pb11 and pb12 are the same. Among Active publishers, seven of them are private BitTorrent portals, three of them are promoting web sites, four are altruistic publishers, and 1 is fake. The fake accounts are quickly removed by TPB and two of the altruistic users have also removed their accounts from TPB. Interestingly, all seven private portals have significantly reduced their publishing rate (*i.e.* their aggregate publishing rate dropped to half from 347 to 178 files per day) whereas other groups of publishers show mixed reactions.

A plausible explanation for the consistent reaction among publishers who manage a BitTorrent private portal is as follows: since the main business model of private portals is very similar to Megaupload (i.e. facilitating access to copyrighted material), they decreased their visibility (i.e. footprint) in the BitTorrent ecosystem to not be viewed as a major player in order to reduce the likelihood of any antipiracy action against them. This reaction is actually similar to the one observed in several Cyberlockers that tried to reduce the availability of copyrighted content in their portals [85]. Such a behavior seems to be aligned with the theory in Economics that punishing a player who performs a non-legitimate activity generates negative incentive for other players involved in similar activities [93, 94]. Finally, it is interesting to note that provided disclaimers in some of the active publishers' website confirm that they are clearly aware of copyright infringement and use the disclaimers to

protect themselves against any potential legal action [4].

3.3.5 Effect of a Local Antipiracy Law

In this subsection, we investigate the effect of a local antipiracy law in a single country, namely France, on content publishers that illegally share copyrighted material through BitTorrent. Toward this end, first we briefly justify our focus on France and provide the required background on the French antipiracy law, called the *Hadopi* law [95]. Afterwards, we examine the longitudinal trend among publishers as the law was legislated, approved and implemented.

To limit the number of unknown variables on our investigation, we focus on a country that has a publicized and properly enforced an antipiracy law. We note that several western countries have had unsuccessful legislative efforts to pass a major antipiracy law. For example, the SOPA law in the US triggered the largest Internet "strike" and was tabled [86, 96]. The Digital Economy Act in the UK [97] has also been delayed till 2014 after the appeal by major ISPs such as British Telecom [98]. The Sinde law in Spain [99] is going to be ineffective even if it is implemented due to its bureaucratic process for suing a potential copyright infringing website [100]. In contrast, there are few countries such as France, New Zealand [101] [74], Korea [102] or Japan [103] that have passed and implemented antipiracy laws that have been reported to be (at least partially) successful. Any of these countries offer a good example for investigating the effect of such a law. However, we focus on France primarily because French publishers have a large contribution in pb10, namely 10% of the uploaded content, in the BitTorrent ecosystem while the contribution of publishers from New Zealand, Japan or Korea is significantly smaller (<1%). Finally, we are neither aware of any popular competing technology for legal and cheap delivery of copyrighted material to users in France (such as Netflix [104] in the US), nor other antipiracy event happening in France that could affect the outcome of our analysis.

Operation of the Hadopi Law: The Hadopi law targets users that share copyrighted content (*i.e.* both consumers and publishers) in Peer-to-Peer (P2P) applications among which BitTorrent is the most popular one. It is a 3-strikes law that is implemented as follows: *(i)* P2P users sharing copyrighted material are identified by their ISPs and receive a warning email to stop their illegal activity. *(ii)* The ISP of the notified users continues to monitor their activity and if they repeat their violation during the next 6 months, they will receive a 2^{nd} warning email together with a certified letter. *(iii)* The ISPs continue to

[4] An example of such disclaimers is the following: "None of the files shown here are actually hosted on this server. The links are provided solely by this site's users. The administrator of this site (www.69bits.net) cannot be held responsible for what its users post, or any other actions of its users. You may not use this site to distribute or download any material when you do not have the legal rights to do so. It is your own responsibility to adhere to these terms".

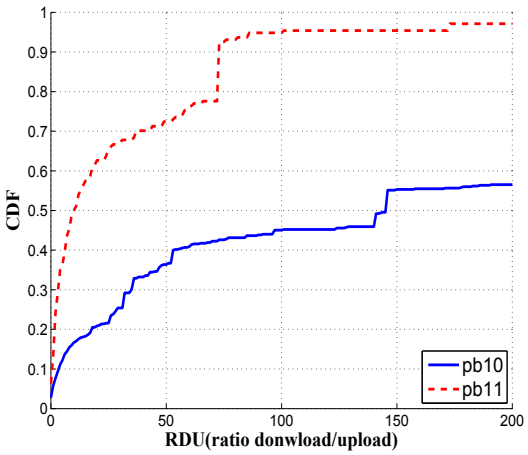


Figure 3.7: Ratio download/upload (RDU) for casual publishers in France.

monitor the notified users for one more year and if they repeat their violation, the Hadopi commission may send the violating users to the court. At this stage, a judge will determine the proper sanction that can be a fine up to 1500 euros and/or the shutdown of their Internet connection for a period no longer than one month. Further details of the law can be found in [75] [76].

History of the Hadopi Law: The Hadopi bill was first presented to the French Senate on June 18th 2008. After a long discussion in the French Assembly and Senate, the law was amended and passed on June 2009. The last legal step took place on October 22^{nd} 2009 when the Constitutional Council finally approved the law. At the end of December 2009, a committee of experts was nominated to implement the law. This process took a long time till October 2010 when the first set of warning emails were sent out. The second round of notifications occurred at the end of February 2011 (six months after the first warning). Finally in February 2012, some expedients were sent to the court as the third strike. It has been reported that since October of 2010 the number of users that have received the first and second warning have been 1.15M and 100K while only 340 expedients have been identified in the third phase, and 14 have been sent to the court. In September 2012, the first condemnatory sentence condemned a user to pay a fine of 150 euros [79]. Figure 3.6 depicts a temporal diagram of the volume of web searches (originated in France) provided by Google Trends for the keyword "Hadopi" over which we have specified the time of the above major events as well as the collection time of pb10 and pb11 snapshots. The temporal alignment of the pronounced peaks in the search volume for Hadopi with the time of major events is a clear indicator that the French population follows this antipiracy law.

The exponential reduction in the number of warnings sent in the consecutive rounds indicates that the first two rounds were the most important since only 340 out of the

	pb10	pb11
Avg. daily publishers-All	487	367 (-25%)
Avg. daily publishers-FR	93	51 (-46%)
Avg. daily contribution-All	1.4K	1.3K (-6%)
Avg. daily contribution-FR	156	184 (+18%)

Table 3.9: Comparison of the number of daily publishers and uploaded content for the entire BitTorrent (BT) ecosystem and in France between pb10 and pb11. The value in parenthesis indicates the normalized difference for each metric.

1.15M identified violating users in the 1^{st} round reached the 3^{rd} strike. The first two rounds took place in October 2010 and February 2011, respectively. Our pb10 snapshot was collected around April 2010 when the law was passed but still not implemented and no warning had been sent out whereas the pb11 snapshot was collected around November 2011, a few months after the 2^{nd} strike. Therefore the pb10 and pb11 datasets are suitable to examine the effects of the main two rounds of the Hadopi on publishers' behavior. Finally, it is important to notice that $\geq 99\%$ of BitTorrent users are consumers (see Table 3.5 for number of consumers vs. publishers). This suggests that a vast majority of the 1M warnings were actually received by BitTorrent consumers who must have reacted by stopping their downloading activity (as demonstrated by the exponential reduction in the number of warnings sent out in the subsequent rounds). However, it is uncertain how French BitTorrent publishers[5], and in particular professional publishers, reacted to the "Hadopi" law.

3.3.5.1 Effect on Publishers Activity

In this subsection, we investigate whether the Hadopi antipiracy law has prompted French BitTorrent (*i.e.* P2P) publishers to reduce or stop their activity. To tackle this issue, we examine the average daily number of publishers and uploaded files among French publishers and compare them with all BitTorrent publishers (as a reference) in snapshots pb10 and pb11. The results are summarized in Table 3.9. The drop in the number of French publishers is roughly twice the drop among all BitTorrent publishers. However, the daily average number of uploads for the entire system dropped by 6% while that measure has increased by 18% among French publishers. This significant increase in the activity (*i.e.* uploads) by French publishers despite the large drop in their number is indeed surprising.

To further explore this issue, we divide the publishers into three classes of casual, regular and active publisher based on their daily average number of uploaded content as we defined in the previous subsection. Table 3.10 summarizes the average daily number of French publishers from each class as well as their average daily number of uploads in pb10

[5]Those BitTorrent publishers whose location of IP address is in France.

n= Avg. content/day	Active (n≥10)	Regular (1≤n<10)	Casual (n<1)
Avg. daily publishers pb10	1.96	19.3	72.6
Avg. daily publishers pb11	1.96	17.8	31.2
Avg. daily publishers difference	0 (0%)	-1.5 (-8%)	-41.4 (-57%)
Avg. daily contribution pb10	34	66	56
Avg. daily contribution pb11	83	77	24
Avg. daily contribution difference	+49 (+144%)	+11 (+17%)	-32 (-57%)

Table 3.10: Number of publishers and daily contribution for next groups of publishers classified based on their contribution to the system in France for pb10 and pb11 snapshots. Active publishers (n≥10 content/day, Regular publishers (1≤n<10 content/day), and Casual publishers (n<1 content/day)

and pb11. Table 3.10 shows that *(i)* the average daily number of casual publishers and their contributions have both dropped by 57%, *(ii)* the average number of regular publishers has dropped by 8% but their contributions have increased by 17%, *(iii)* finally, there are roughly two daily active publishers in both snapshots but their contribution increased by 144%.

We take a closer look at each one of these trends to identify their underlying causes. First, our hypothesis is that those casual publishers (57%) leaving the system are indeed active consumers who altruistically publish very few content. Therefore, the Hadopi law has motivated them to stop their downloading activity which in turn has led to the drop in their publishing rate as well. We verify this hypothesis by examining the distribution of the ratio of the number of downloads to uploads (RDU) for casual publishers in pb10 and pb11 that is shown in Figure 3.7. This figure illustrates that the RDU ratio among casual publishers in pb11 is roughly an order of magnitude lower than in pb10. This confirms our observation that most of the departing publishers between our two snapshots are indeed active consumers with a significant drop in their publishing activity. Note, that these active consumers are likely to be regular Internet users for whom the sanctions associated with the Hadopi law is considered too costly (*e.g.*, a fine up to 1500 euros).

Second, to uncover the factors that led to the significant increase in the publishing rate of active users, we take a closer look at top French publishers. We noticed that one of these publishers is *scenebalance* that is the most active one among all BitTorrent publishers worldwide in pb11 (as shown in Table 3.8). Scenebalance is a professional worldwide publisher injecting more than 100 contents per day into the BitTorrent ecosystem, most of them from France[6]. This raises a couple of interesting issues as follows:

(i) Since the number of French consumers has rapidly dropped, these active publishers must be targeting consumers that are outside France. To explore this issue, we have checked

[6]In [52], authors demonstrate that usually active publishers upload their content from different IP addresses that in many cases are located in different countries.

the published content by top five French publishers in both pb10 and pb11 snapshots. For pb10 snapshot, we found that two of the top publishers upload porn content in English, another one publishes TV series and shows in English, and the two remaining ones upload only Spanish content. In the case of pb11, we discovered a similar situation where three of the top five publishers upload only content in English, and the remaining two publish Spanish content. We extended this probe to top 20 French publishers in pb10 and pb11 and could identify only one publisher who is clearly uploading content for French consumers (*e.g.*, French content or content with French subtitles). This investigation confirmed that major French publishers primarily target worldwide consumers.

(ii) It is then intriguing why these professional publishers operate from France while their consumers must be mostly outside France, and there is an enforced antipiracy law that could affect them. Closer examination of the top 20 French publishers revealed that more than 80% of them in both pb10 and pb11 snapshots are located at a particular hosting facility. In fact, 29 of worldwide top 100 BitTorrent publishers from pb10 and 25 of them from pb11 were hosted at that particular hosting facility. This hosting facility provides professional publishers with powerful servers to perform their intensive activity of uploading and serving (*i.e.* seeding) the large amount of content they make available through BitTorrent. We contacted that hosting facility to gain some insight into its popularity among the professional BitTorrent publishers and learned that the hosting facility does not proactively monitor the activities of its customers unless a violation is reported by a third party and the customer does not cease its "improper" activity. Such a passive monitoring strategy is unusual as most of the hosting providers in recent years (*e.g.*, Server Intellects [105]) have adopted strict monitoring policies to prevent the distribution of copyrighted material from their servers through P2P applications. These evidences collectively suggest that the "BitTorrent-friendly" policy that hosting facility is much more valuable for publishers than the cost of any potential antipiracy action against publishers in France. It is important to note that professional publishers have major financial interest in publishing copyrighted material [52]. Therefore, they carefully examine any law that might affect them, take advantage of existing loopholes, and weigh the likelihood as well as the implications of any legal action against them. This suggests that even if the Hadopi law intends to targets publishers, it is much more difficult to deter at least professional publishers compare to consumers. In a nutshell, many professional publishers operate from France simply because of a major hosting facility passive monitoring policy accommodates their illegal activities.

In summary our results reveal that French antipiracy law has been quite effective on reducing the number of casual publishers in BitTorrent who were primarily consumers and the potential of receiving a fine or temporal loss of Internet connection as a result of the

Hadopi law is considered costly and thus has a deterrent effect. However, the law has not succeeded in reducing the publishing rate of copyrighted material by professional publishers. These publishers seem to have most of their servers in a particular hosting provider in France primarily due to the "BitTorrent-friendly" policy of this provider. The benefits of having access to a facility that does not monitor the sharing of copyrighted material allows these businesses to comfortably operate which is clearly more valuable than the potential risk of any fine that is negligible compare to their profits.

3.3.6 Conclusions

This study presents a detailed study on how two major antipiracy actions affect the behavior of publishers in the largest BitTorrent portal who primarily publish copyrighted content. In our first case study, we focused on the impact of the Megaupload closure as a worldwide antipiracy event on BitTorrent publishers. We showed that the Megaupload closure triggered an immediate drop in the activity of professional BitTorrent publishers that are running their own private BitTorrent portals. Furthermore, a group of casual publishers also migrated to BitTorrent most likely from Megaupload and other Cyberlockers. Our second case study revealed that the French Hadopi law was effective in reducing the number of casual BitTorrent publishers that are actually consumers. However, it did not have any impact on the activity of professional publishers from France. The concentration of very active publishers in a particular hosting facility in France suggests the popularity of this facility among BitTorrent publishers that appears to be due to its passive monitoring for copyright infringement activity. Therefore, legally savvy publishers are willing to take the chance and operate from France and are not concerned about a potentially small fine. Our findings provide a valuable insight about the effect of antipiracy actions on publishers who are engaged in online piracy and also reveal the complexity of identifying the affected group of publishers. While it is impossible to validate our findings, the collection of all supporting evidences, their temporal alignment and the dominance of target events suggest that the observed behavior among publishers are most likely driven by the corresponding antipiracy events.

Chapter 4

Profiling regular and professional in major Online Social Networks

Contents

4.1 Summary

This chapter focus is on characterizing professional and regular users behavior on Social networks. It is worth mentioning that in contrast to other OSNs (e.g., Twitter) there are different type of accounts for regular and professional users. On the one hand, regular users are connected to other users in the network by means of bidirectional connections and this is if a User A is connected to a user B, then the user B automatically has a link with user A. We refer to A and B as friends. A regular FB user account is limited to only 5000 friends. On the other hand, we can find Facebook Pages that are usually created by popular users (e.g., politician, musician, celebrities, sportsmen, etc.) or a large variety of commercial brands (e.g., coke, BMW, BBC, Zara, etc.). FB Pages do not create bidirectional relationship, but instead unidirectional links in which regular users, referred as fans, subscribe to the page. Different to the regular profiles, FB pages are not limited in the number of fans that can subscribe to them. Therefore, FB presents two clear different types of users from very early registration process, regular users and Pages, that require a different treatment and analysis. Finally, it is important to notice that our approaches for crawling information for regular users and Pages is totally different, and it required designing and implementing two different crawling tools.

In this chapter three relevant studies are presented that start with a study on regular users and shows what amount of information is publicly disclosed on users Facebook profile (sebsection 4.2). The two other studies which are presented in this section evaluate the professional users behavior across three major OSNs namely Facebook, Google+ and Twitter. The second study present a novel methodology to identify Cross-posting activity of a professional users across different accounts of her in three mentioned OSNs and quantify the volume of this type of activity (Section 4.3). And finally the third study provides an innovative model to find the strategies of professional users per sector in OSNs in term of their publishing activities and involvement and also evaluate the success level of their followed strategies (Section 4.4).

Keywords

Online Social Networks, Facebook, Privacy, Information Disclosure, Professional Users, Strategies, Cross Activity.

4.2 Analysis of publicly disclosed information in Facebook profiles

4.2.1 Introduction

Facebook is the most popular On-line Social Network (OSNs) with more than one billion subscribers. Users mainly utilize Facebook to share their opinions, interests, personal content like pictures with users who are connected to them. An important element that Facebook incorporates is the possibility of defining a detailed profile where users provide information about themselves. In Facebook we find more than 20 different attributes that can be utilized in a user profile. Those attributes include potentially sensitive information such as contact info, birth date, current city, home town, employers, college, high school, etc. Furthermore, together with that personal details, Facebook users can complete their profiles by expressing their interests in different categories such as music, movies, books, television series, games, teams, sports, athletes, activities and inspirational figures, which in many cases facilitates deriving sensitive information from a user (e.g. personality characteristics, political leanings). Depending on the person, their status and this information's social context, publicly disclosing this sort of information could lead to some serious privacy issues. To avoid or at least mitigate these problems, Facebook allows each user to define a degree of privacy for different attributes in the profile. That is, for each attribute, a Facebook user can decide among several privacy options: (*i*) leaving an attribute blank so that no one will get access to that information; (*ii*) filling out an attribute and defining its privacy level as *"only me"* meaning only the user has access to that information; (*iii*) defining the attribute privacy level as *"friends"* which allows a user's Facebook contacts to access the information; (*iv*) defining the attribute privacy level as *"friends of friends"* that makes the information available not only to the user's contacts but also to the Facebook friends of those contacts; (*v*) defining the attribute privacy level as *"custom"* in which the user can define one by one which users can access the attribute information (e.g. just some part of her friends); and, (*vi*) defining the privacy level as *"public"* so that any user can access that information. Based on the Facebook strategies by default most of the attributes are publicly available except the birthday, Political views, Religion and Contact Info that are in the level of "only Friends". For these attributes users can change the privacy level to public or more private.

The information included in the profile of Facebook users is precious for external users/entities and these have very divergent objectives, from non-lucrative activities such as research to lucrative ones, including marketing campaigns. Given the privacy management provided by Facebook, external entities can only access attributes that has been defined as "public" by users. Therefore, an important question to answer is what is the amount

of public information that an external user/entity can find in Facebook profiles. In other words, what is the portion of Facebook users that publicly disclose (i.e. indicate privacy level "public") each of the profile attributes. By answering this question for each attributes we will be able to understand which type of information is considered more sensitive by Facebook users, and to the contrary, what are the attributes experiencing major public exposure.

Toward this end we have collected the public profiles of 479K randomly-selected Facebook users, and analyze 19 of the profile's attributes by computing the portion of the collected users that publicly disclose each attribute in their profiles. We divide the analyzed attributes into two groups: *personal* and *interest-based* attributes. The former category refers to attributes that contain personal life information about the user (e.g. location, education, work history, etc). Interest-based attributes, on the other hand reflect the tastes of Facebook users, revealed by their preferences (e.g. in music, television, sport teams, etc). The results will let us determine the attributes that users consider more sensitive. Furthermore, we explore the correlation degree among the different personal attributes. That is, determining if a user disclosing a personal attribute A has some relation to that user also publicly sharing a different attribute B. In order to get a meaningful answer, in this study we correlate 9 personal attributes pairwise.

Our attribute-based analysis tells us how much information can be retrieved for a particular attribute, but it does not contribute anything regarding the expected amount of information that we can extract from a typical Facebook user. Therefore, we seek to understand the public exposure habits of Facebook users themselves. To that end we have defined a very simple yet meaningful metric that accounts for the number of attributes that are publicly disclosed in a Facebook profile, and refer to it as the *Degree of Public Exposure (DPE)*. The DPE ranges from 0 for user profiles that do not have any attribute publicly available, to 19 when a user has made all the analyzed attributes available, including personal and interest-based attributes. Hence, we can assign each of the 479K users in our dataset a DPE value. Using this metric and our dataset we are able to identify what type(s) of users present a higher degree of public exposure.

Finally, in the last part of this study, we define three simple use cases to illustrate how some external entities can utilize the information that is publicly accessible in Facebook. First, we perform a gender-based division of different personal attributes to discover whether men or women show a significant predisposition to publicly disclose particular type of information. Second, we depict the distribution of the ages of our 479K Facebook users based on those users that publicly share their ages. Third, we check the accuracy that could be achieved by using Facebook users as an estimator for the distribution of the world wide population in cities, e.g. to estimate the portion of human-beings living in cities with

a population 5 million or more.

The main observations extracted from the study are:
(*i*) Friend-list is the attribute with the largest public exposure with almost 63% of users publicly sharing their contacts, whereas a users' age (i.e. Birth date attribute) rate as having the highest privacy value for from Facebook users, since only 3% disclose this information.
(*ii*) There are strong correlations between *Current City* and *Home Town* attributes. This may be because both attributes provide a type of "location" information, and users revealing one tend to also share the other. In addition, we found a second high correlation between education (i.e. *College* and *HighSchool*) and professional experience (i.e. *Employers*) attributes. Typically, education and professional experience complement each other and are closely related. A clear example of this is the case of CVs where we always include both, education and professional information.
(*iii*) The average Facebook user makes more than four attributes publicly available in their profiles.
(*iv*) Men show a larger public exposure then women for all personal attributes except *birth date*. This exception is very surprising given the widespread assumption that women tend to hide their real age more than men, which is not the case in Facebook.
(*v*) The age range most-represented, based on the publicly available information, is 18-25. That age range accounts for 1/2 of the users among those making their birth date publicly available.
(*vi*) We show that Facebook data very accurately estimates the portion of people that live in cities of more than 5 million (according to a recent United Nation report [106]). It also provides an accurate estimation for the proportion of people living in cities ranging between 500K-1M inhabitants, whereas it has a 10% deviation for cities of less than 500K and for cities with between 1M and 5M citizens.

4.2.2 Related work

We explore the prior efforts regarding to user privacy in online social networks that establish the basis for our work. In a concept similar to our study, Quercia et al. (2012) [20] found a correlation with the degree of openness and gender, using a dataset of 1323 profiles from the United States. Our work has many distinctions from this study. Firstly, our dataset is much larger and broader (479K profiles widely distributed throughout the world compared to a little more than 1K profiles exclusively from U.S.). Secondly, our data was gathered directly from Facebook profiles, while Quercia et al. used a form of questionnaire administered by a specific Facebook application. Lastly, we study most of the available attributes in the FB profiles, and for some of them we deeply investigated the correlation between the attribute

type and profile characteristics. They also concluded that men tend to make their profile information more publicly available. In another work by these authors [107], they study the personality characteristics of popular Facebook users.

Gross et al. in [23] studied the patterns of information revelation in Facebook. They analyzed just around 4K Carnegie Mellon University students' profiles, specifically those that joined a popular social networking site catering to college students. Gross et al. evaluate the amount of information students disclose and their usage of the site's privacy settings. Also these authors in [108] study the evolution of the profiles privacy of around 5k of their collected profiles in past years.

In other work, Chang et al. [21] studied the privacy attitudes of U.S. Facebook users of different ethnicities. Another U.S.-based study [25] used a questionnaire and with considering 1,710 students' profiles shows that women are more likely to maintain a higher degree of profile privacy than men; and that having a private profile is associated with a higher level of online activity. The authors in [26] examined disclosure in Facebook profiles looking at only 400 Facebook profiles. In a similar work to the previous one, authors in [27] employed surveys and interviews to study the factors that influence university students to disclose personal information on Facebook.

In a study of the Facebook users' profile attributes, authors in [22] present a method to estimate the birth year of 1M Facebook users in New York City, based on the information available on their profiles, such as their friends. Authors in [30] examined the possibility of using the attributes of users, in combination with their social network graph, to predict the attributes of another user in the network. Other similar work [28] presents a study of Facebook profile attributes by analyzing a dataset of 30,773 Facebook profiles. They were able to determine which profile attributes are most likely to predict friendship links and discuss the theoretical and design implications of their findings. They explore how profile attributes relate to the #Friends of a user's profile. An investigation of Facebook users' privacy evolution in a dataset of a large sample of New York City (NYC) Facebook users, was presented in [109]. That study shows how the close/disclose status of profiles attributes changed over time.

Apart of the above-mentioned works as well as many other similar studies, some surveys on the literature of security and privacy in online social networks has been done to formulate related concerns, such as those in [110] and [111]. Authors in [112] discuss the design issues involved in the security and privacy of OSNs. Another work [113] investigated the privacy and security of users in online social networking sites such as: Facebook, Google+, and Twitter. Authors in [114] explored the negative impacts of social networking sites on its users.

By considering the previous work, the study presented here is a new effort in the arena

of social networks; one that by uses a large dataset of Facebook profiles to analyze the profile information disclosure patterns.

4.2.3 Data Collection and Attributes Definition

In contrast to other OSNs (e.g., Twitter) Facebook subscribers are able to create different type of account. Regular users are connected to other users in the network by means of bidirectional connections but a regular FB user account is limited to only 5000 friends. The account for professional users, Facebook Pages, do not create bidirectional relationship, but instead unidirectional links in which regular users, referred as fans, subscribe to the page. FB pages are not limited in the number of fans that can subscribe to them.

We have implemented an HTML crawler that is able to collect publicly-available information from a Facebook user's profile. The crawler collects up to 19 attributes from each profile. It must be noted that our tool respects the privacy of users since we only collect information that users themselves decide to share publicly. Base on this we can not differentiate that one attribute is blank or it is closed to public.

Our goal was to capture the publicly available information from a random sample of Facebook profiles. We run our crawler between March to June 2012 and captured the profile of 479k Facebook users randomly selected throughout the world. For each user we store up to 19 different attributes (only those publicly available). We classify those attributes into two categories: personal and interest-based. The first category refers to information related to an individual's life, while the second includes information regarding user's "likings". The 19 attributes are listed below in their respective category.
Personal attributes: Friend-list, Current City, Hometown, Gender, Birthday, Employers, College and HighSchool.
Interest-based attributes: Music, Movie, Book, Television, Games, Team, Sports, Athletes, Activities, Interests and Inspired people.

The meaning of the personal attributes present are obvious and self-contained. It worth mentioning that some of them such as *Employers, College,* or *HighSchool* could include more than one item. For instance, a user can include their current employer as well as previous ones or, in the case of college, a user could list several names if she obtained degrees in different universities or other post-secondary schools. In the case of Interest-based attributes, all of them can contain more than one item. Facebook users use these attributes express their likings for the categories referred to by the attribute. For instance, in the music category we can find singers, music bands, music styles (e.g. jazz, rock, etc.), music albums, etc. We need to note that in our analysis we insert an "artificial" interest-based attribute, called *Aggregate-Interests* which is a binary attribute, i.e. it is 1 if the user publicly shares at least one item among all the interest-based attributes, and 0 otherwise.

The *Aggregate-Interests* attribute lets us know if a user shares any interests without taking into account the separate categories.

Finally, in order to perform personal attribute correlations, and to gain further insights into some of them, we have divided our main dataset into several attribute-based groups. Basically, a given group A includes all the users in our main dataset that publicly disclose attribute A. For instance, from this point onwards in the study, when we mention the *Gender* group we are referring to the group that includes all the users in our dataset that make their gender available in their Facebook profile.

4.2.4 Public exposure of Facebook profile attributes

In this subsection we define the degree of publicly disclosed information in Facebook[1]. We first perform an attribute-based analysis to study the portion of Facebook users that disclose each attribute. Next, we study the correlation among pairs of personal attributes. That is, of those users that make an attribute A public, what is the portion of users that also disclose attribute B. Towards this end, we create one group of users for each personal attribute, so that the group for attribute A includes all the users in our dataset that make that attribute available, and later correlates those users with the remaining attributes. This analysis will provide useful insights on whether some attributes are correlated and we will discuss some potential reasons for such correlation.

4.2.4.1 Degree of attributes disclosure

We provide some global numbers that paint a global picture of the amount of information (i.e. attributes) that Facebook users make publicly available. To this end first of all we study the default status of the attributes in Facebook. The study shows that out of the 479k analyzed users, only 11.62% do not share any attribute, 19.26% disclose a single attribute, while the remaining users, 69.12%, have two or more attributes in their profile that are publicly accessible. These values give a first reference point to help understand that external users/entities can retrieve an enormous amount of information from Facebook profiles.

Our goal is to determine the level of privacy awareness that Facebook users present with respect to the different attributes. Table 4.1 shows the portion of users in our main dataset that publicly disclose each of the studied attributes. We first focus on personal attributes and then discuss interest-based attributes.

[1]We clarify that, for better readability, in the rest of the study when we mention that a user discloses, shares or makes available an attribute we are explicitly saying that this attribute was assigned a private level of "public" and so any other user has access to it.

Table 4.1: Portion of users with publicly disclosed personal and interest-based attributes in Facebook profiles.

	Attribute	% Profiles accessible
Personal attributes	Friend-list	62.7
	CurrentCity	36.1
	Hometown	34.6
	Gender	53.5
	Birthday	2.9
	Employers	22.5
	College	16.8
	HighSchool	13.2
Interest-based attributes	Aggregate-Interest	48.4
	Music	41.0
	Movie	28.3
	Book	16.7
	Television	31.8
	Games	9.4
	Team	8.5
	Sports	2.3
	Athletes	10.7
	Activities	20.5
	Interests	10.9
	Inspire	1.9

Table 4.2: Attributes correlation. Each value in the table refers to the portion of users belonging to the group indicated in the column that disclose the attribute indicated by the row.

Attribute	All	Friend-list	CurrentCity	Hometown	Gender	Age (Birthday)	Job (Employers)	College	HighSchool
Friend-list	62.7	100	79.6	79.3	64.8	72.5	82.8	83	87
CurrentCity	36.1	45.9	100	**74**	42	56.2	55.4	59.4	57
Hometown	34.6	43.7	**71**	100	35.7	58.2	55.3	54.2	50.8
Gender	53.5	55.3	61.7	55.2	100	58.8	55.7	79.9	86
Birthday	2.9	3.4	4.6	4.9	3.2	100	5	4.9	4.2
Employers	22.5	29.7	34.5	35.9	23.4	38	100	**59**	**53**
College	16.8	22.2	27.6	26.3	25	28	**43.8**	100	64.6
HighSchool	13.2	18.3	20.8	19.3	21.2	18.7	31.1	**50.7**	100

Personal attributes The friend-list appears as the attribute with the greatest public exposure. Table 4.1 shows that almost 63% of the users make their friend-list available. This clearly indicates that FB users do not consider that exposing their connections could lead to any privacy issue. At the other extreme, the attribute with the lowest exposure is Birthday. Less than 3% of the users reveal their age, which means that users regard this attribute as highly private. Here it worth to mention again that Birthday attribute is in the privacy level of "only Friends" by default in Facebook and this 3 % of users they changed this level to publicly available. Also, 1/2 of the users share their gender. A bit less, around 35% of users, make their current city and their home town available publicly. This implies that users consider personal location information to be more sensitive than the information related to their contacts, but much less sensitive than their age. In addition, users seem to

be more concerned about privacy issues linked to disclosing their job information since a little less than 1/4 of them publicly list their employers. We close the analysis of personal attributes by evaluating those related to education, where 17% and 13% of users publicly share their college and High School. Education-related attributes are thus the next-most private attributes after age. In summary, we can list the attributes in terms of public exposure (from more to less exposure) as follows: friend-list, gender, job, education, and age.

Interest-based attributes Table 4.1 shows that almost 1/2 the users share at least one interest within the interests-based attributes, which means that Facebook users are not very concerned about the potential privacy implications that could be derived from sharing their interests. These attributes are initially less sensitive than personal attributes in terms of privacy. However, in some cases a particular interest of a user regarding some controversial issue could potentially lead to privacy issues. Looking at the results in the table we observe that the more popular categories are music (41%), Television (32%) and Movies (28.3%). It is interesting that almost all users that share an interest (48%) are actually sharing Music (41%). In contrast very few users share information in relation to their sports interest and as to what inspires them, just 2.3% and 1.9% respectively. The remaining interest-based attributes are made available by 10%-20% users. Finally, it is worth to mention that personal attributes such as Friendlist, CurrentCity, Hometown or Gender are more accessible than users' interests.

4.2.4.2 Correlation of Facebook Attributes

We now turn our attention to the different groups that include all the users that disclosed a particular attribute (e.g. CurrentCity), and how they correlate with the remaining personal attributes. Table 4.2 shows the portion of users from a given group (columns) that share one of the remaining attributes (rows). For instance, the value crossing Current City column with Friend-list row means that 79.6% of the users in the *CurrentCity* group (i.e. those users from our dataset with their CurrentCity attribute available) also disclose their Friend-list. In addition, table 4.2 includes the results obtained from our main dataset, referred to as *All group* (the first column in the table), for comparison purposes.

First of all we observe that all the analyzed groups present a larger percentage for their available attributes than in *All* group, which implies that users that share one personal attribute will likely share some other attributes. This assumption is supported by the observation that 2/3 of Facebook users disclose more than one attribute, as previously reported in this subsection. It is especially noteworthy that most of the users (71%) disclosing their *CurrentCity* also make public their *Hometown*, and close to 74% of users that share

their *Hometown* attribute also disclose their *CurrentCity*. This indicate that Facebook users relate these two attributes together, and in case they share the place where they currently live, they also disclose the place where they were born. In fact, these two parameters are the only ones that directly provides a physical location (i.e. College or employers can provide location information but in an indirect manner), and it is clear that most Facebook users providing location information tend to share both of these indicators. Therefore, we can conclude that *CurrentCity* and *Hometown* attributes are highly correlated since 3/4 of users disclosing one of these attributes will also share the other one.

We also find a significant correlation when we relate the employment and the education attributes. The users composing the *Employment* group tend to also share some educational information. In particular, 44% of the users that make their job information available also show their College, and 31% identify their High School. This is also validated in the other direction as 59% and 53% of users in the *College* and *HighSchool* groups, respectively, made their employer available. In addition, as we would expect, the two education attributes are highly correlated with each other. In contrast, user groups that are not related to education or employment information show a much lower correlation to these attributes, always below 38%, 28% and 21% for Employment, College and HighSchool, respectively. This observation reveals that a large portion of Facebook users understand that employment and education attributes complement each other, which is certainly obvious to anyone who prepares their CV where professional experience and education are always included, and often complement each other. Furthermore, the high number of users (44%) disclosing their College within the *Employers* is significant even though 44% reflects less than half of all, that figure is quite high given that a large number of users in Facebook that cannot share their college because they simply never attended (or did not graduate). Then, that 44% is actually a very relevant number that roughly demonstrates that whoever indicates their employer (or employment status) in Facebook and has obtained a University degree wants to make it public. This hypothesis is validated by the fact that only 31% users in the *Employers* group share its HighSchool, and obviously there are more users in the *Employers* group who went to the High School than the ones who went to the University. Previous statement is validated by the fact that 65% of users in *HighSchool* group also report their College, whereas this portion is reduced to 50% for those users in *College* group that also report their *HighSchool* information. Therefore, we can extract two main conclusions from the correlation analysis between education and employment: (*i*) These two attributes are clearly correlated in Facebook, and (*ii*) an important fraction of users in Facebook understand that disclosing the University they attended does not imply any privacy issue, instead they seem to believe it provides them with a good reputation.

In the *Gender* group we do not find any strong correlations, only very weak correlations

with *CurrentCity* (42%), *Hometown* (36%) and *Employers* (23%) compared to the correlations of the rest of the groups with these attributes. This would suggest that users sharing their gender have strong privacy concerns with respect to their location and employment information. Although we do not yet have any supportive argument to present with these results, we think it is worth mentioning.

Finally, we cannot find any relevant correlation between the *Friend-list* and *Birthday* groups and the other attributes. In a nutshell, we have found strong correlations between: (*i*) the *CurrentCity* and *Hometown* attributes, and (*ii*) the education attributes, *College and HighSchool*, between each other and with the *Employment* attribute. We believe that the first correlation is because that users roughly perceive both parameters as location information, so if they do not have privacy concerns with one, they also do not have an issue with the other. Our hypothesis for the second correlation is that users, with anyone preparing a resume, find that education and employment attributes complement each other.

4.2.5 Public Exposure of Facebook Users

To this point, we have performed an attribute-based analysis that has allowed us to understand which attributes are more privacy-sensitive for Facebook users, and to identify the correlation that exists (or not) among the different attributes. However, this analysis did not account for the public exposure of Facebook users. Towards this end we need to perform a user-based analysis. Instead of taking one attribute and counting how many users share it, we now need to look at individual users and determine how many attributes (among all those possible ones) she is disclosing. For that we take into account all 19 attributes collected with our tool from a Facebook profile (Personal + Interest-based attributes). We define a simple but functional metric named as Degree of Public Exposure (DPE), which ranges from 0 to 19. Basically, we go through the 19 parameters and whenever one can be accessed we sum +1 to the DPE value for that user. By defining this metric we are able to easily compare the level of profile's attribute openness without considering any kind of difference between the attributes.

Table 4.3 shows the median and average value of the DPE metric for our main dataset, as well as each of the previous attribute-based groups, while Figure 4.1 provides further details of the DPE distribution for the different groups by means of a box plot graph that shows the 25th, 50th (median) and 75th percentiles. If we first consider the results for *All* group, we extract that a typical Facebook user presents an average DPE of 4.27. The remaining groups (except for *Friend-list* and *Gender*) show an average DPE higher than 7. This means that users in these groups publicly disclose more than seven attributes. It is worth noting that the users with a higher public exposure are those ones that share their education information, i.e users in *College* and *HighSchool* groups, which present an

Table 4.3: Median and Mean of DPE metric

Attribute	Median	Mean
All	3	4.27
Friend-list	5	5.61
Likes-list	7	7.11
CurrentCity	7	7.18
Hometown	7	7.35
Gender	4	5.26
Birthday	7	7.60
Employers	7	7.26
College	8	7.95
HighSchool	8	8.02

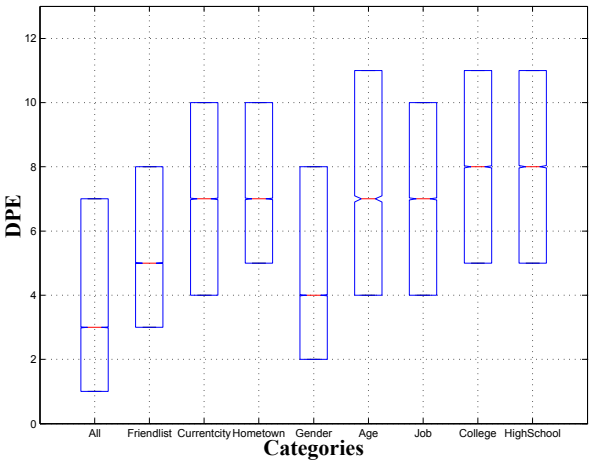


Figure 4.1: Box plot of DPE for categories

average DPE of 7.95 and 8.02, respectively. If we analyze the results shown in Figure 4.1, we can observe that all the groups except *All*, *Friend-list* and *Gender* present a DPE 75th percentile $\geq$10. This means that there are a relevant portion of users that disclose more than 10 attributes. Therefore, those users may be very attractive for external entities since they have a quite complete information regarding them.

In a nutshell, our results demonstrate that anyone can find substantial personal information from Facebook profiles since it is publicly available. In particular, our results suggest that if an entity wants to maximize the amount of information (i.e. attributes) retrieved from Facebook profiles, she should target users disclosing their education information.

4.2.6 Examples of Public Facebook Information Usage

In this subsection we show three examples of how the information available in Facebook can be used for different purposes. First we present a gender analysis to understand whether men or women show a major predisposition to disclose personal attributes. Next, we use

Table 4.4: Gender analysis per categories of attributes

attributes' categories	%Male	%Female
All	51.33	48.67
Friend-list	53.99	46.01
CurrentCity	52.81	47.19
Hometown	54.05	45.95
Gender	51.33	48.67
Birthday	49.23	50.77
Employers	55.23	44.77
College	53.30	46.70
HighSchool	55.89	44.11

Table 4.5: Age of users with disclosed birthday

Age category	% of users inside Birthday group
Teenagers ($\leq$ 18)	0.85
Post-Teenagers (19 - 25)	48.29
Young (26 - 30)	27.22
Mature (31 - 50)	19.71
Senior ($>$ 50)	3.93

the age information available in our dataset to depict the distribution of the users' ages in among Facebook. Finally, we use *CurrentCity* information to estimate the distribution of worldwide population across cities according to their size, and crosscheck the result with the data provided by the United Nations.

4.2.6.1 Gender attribute: Men vs. Women public exposure

In each attribute-based group we found users that provide their gender information and study which portion of them are males and which portion females. Table 4.4 shows the percentage of users for each gender and group. Male is the dominant gender for all the attributes except Birthday. This seems to indicate that generally men are less concerned about privacy issues than women, however the difference for most of the parameters is small, and never goes above 11 percentage points. The higher differences occur for Employers and HighSchool attributes. Finally, it is somewhat surprising that women share their age information slightly more frequently than men, which contradicts the "cultural" assumption that women tend to hide their age more often than men.

We can find many reports that explore gender differences in different disciplines like sociology and psychology such as [115], etc, which in many cases has a large diffusion even reaching general media. This example demonstrates that the publicly available information in Facebook is a potential source of information for these types of studies.

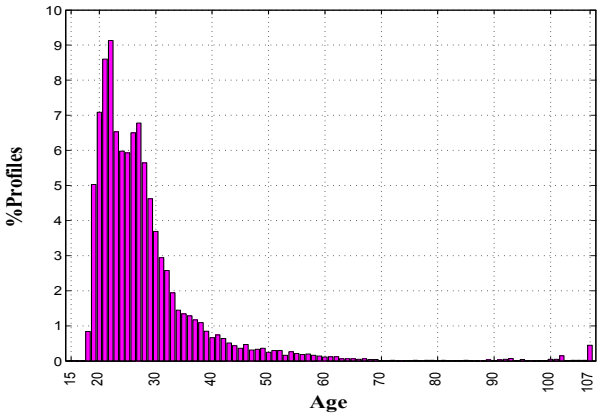


Figure 4.2: %Profiles in different age range

Table 4.6: Gender distribution in different categories of age

Age category	% Female	% Male
Teenagers ($\leq$ 18)	62.67	37.33
Post-Teenagers (19 - 25)	54.83	45.17
Young (26 - 30)	48.10	51.90
Mature (31 - 50)	45.37	54.63
Senior ($>$ 50)	37.54	62.46

4.2.6.2 Age distribution analysis

Analyzing the distribution of ages among those few users (i.e. 2.9% of the 479K) that publicly share their birth date reveals some unexpected results. Figure 4.2 shows the portion of users in our dataset belonging to each age from 13 to 107 (Facebook does not allow accounts to be opened for users younger than 13). Surprisingly, we found very few users $\leq$18 years old, and we did not find any Facebook rule that penalizes the disclosure of birthdays for users less than 18 years old. The ages in the interval of 19-28 contain more than 50% of the users revealing their age, with 21 and 22 the most represented ages containing more than 8% of the users each one. From 28 years upwards we found an exponential decrement, in some few cases reaching ages above 100. Particularly, we observe that almost 0.5% users report an age of 107 (indicating1905 as their birthday year, which at the time we collected the data was the oldest year allowed by Facebook). It is very likely that these are fake ages introduced by users who do not want to provide their real age.

In order to provide aggregate numbers we have classified users into 5 different ages groups : Teenagers ($\leq$18), Post-Teenagers ($18-25$), Young ($26-30$), Mature ($31-50$) and Senior ($>$50). Table 4.5 reports the portion of users included in each of these categories. The results reveal that Teenagers very rarely (0.85%) disclose publicly their age, which was totally unexpected statistic. In contrast, as the results confirmed our expectation that

Senior users, which are the ones less representative in OSNs like Facebook, would present a low weight in the *Birthday* group. Therefore, the big majority of users sharing their age belongs to he interval 18-50. In particular, Post-Teenagers between 18 and 25 years old represent 1/2 of the users sharing their birthday, followed by Young that accounts 1/4 of the users, and Mature group comprising 1/5 of the users from *Birthday* group.

The results in the previous subsection revealed that women share their age a bit more often than men, and we want to check whether this is constant across different age categories. Table 4.6 shows for each age category the portion of users whose gender is male or female. In the case of Teenage women expose their age much more than men. In the case of post-teenagers we find 10% more women than men among the users disclosing their age. The observed tendency changes for young people between 26-30 years old where we find slightly more men sharing their age. This change of tendency is confirmed in the Mature and Senior categories where there are 10% and 25% more men with open ages as compared to women, respectively. In summary, we can conclude that there is a clear trend, the younger the age group the larger the portion of women disclosing their age is as compared to men, and the other way around, the older the age group the more the portion of men disclosing their age.

As it happened for the gender analysis, there are other disciplines that use age groups to perform different types of analysis. We have demonstrated that Facebook allows researchers to easily identify users of particular ages who also have other personal and interest-based attributes accessible.

4.2.6.3 CurrentCity population analysis

In this subsection we aim to validate the accuracy of a small sample of Facebook to compute the distribution of worldwide population across cities according to their size. For this use case we need to perform a more complex analysis than in the previous use cases where the results were directly derived from our database.

We found 8.473 different cities in the *CurrentCity* attribute inside our dataset. We used debepedia [116] (a crowdsourcing effort to extract structured information for Wikipedia) in order to retrieve the population associated to those cities. We were able to identify population for 1.840 cities that aggregately include 173,026 profile out of the users with open *CurrentCity* attribute our database. We classify these cities into six categories according to their population: <1K, 1K-10K, 10K-100K, 100K-500K, 500K-1M, 1M-5M, and >5M. For each category we have extracted the portion of FB profiles (corresponding to those cities) belonging to each class. Furthermore, we've used official statistics reported by the United Nations (UN) in its 2011 World Urbanization Prospects report [106] (see page 25). Unfortunately, this report only includes granularity for cities with more than 500k citizens.

Table 4.7: Population distribution of Facebook (#Profiles) and world (#Inhabitants) in different city size class

City Size Class (#Inhabitants)	%Profiles (FB)	%Inhabitants (UN) [106]
< 1K	0.14	
1K - 10K	3.60	
10K - 100K	18.50	
100K - 500K	18.39	50.9 (<500k)
500K - 1M	8.78	10.10
1M -5M	33.04	21.30
> 5M	17.55	17.07

Table 4.7 collects the results for the FB and the UN report.

Facebook results reveal that less than 0.2% of users live in small villages with less than 1K inhabitants. Our hypothesis for this result is that people living in such small villages is usually senior people (>50), which, as demonstrated in subsection 4.2.6.2, is very low population in Facebook. Therefore, we believe this data may not reflect the reality. Larger villages up to 10K citizens are reported by 3.6% of users. Again we think this data is biased by the same reason explained before. Small towns going from 10K to 100K citizens and big towns between 100k-500K inhabitants show the same portion of profiles, roughly 18.5% each of them, so 37% both categories together. We found almost 9% of Facebook users in cities from 0.5M to 1M citizens. Finally, cities above 1M users include more than 1/2 users, which are divided as follows. One-third of Facebook users report that they live in cities with a population between 1M and 5M, and 17.5% of the users live in very big cities with more than 5M.

Here we compare the Facebook results to the UN data in order to check the accuracy of a small Facebook sample (i.e. users in our dataset belonging to those 1,840 cities for which we were able to identify their population) to estimate the worldwide population distribution across cities according to their size. First of all, our data is able to very accurately estimate the portion of worldwide population in cities with more than 5M citizens. Furthermore, we also found a quite accurate estimation of the population in cities whose population ranges between 500k and 1M, since there is a discrepancy a bit higher than 1%. In contrast, we found an important discrepancy for the case of cities between 1M-5M citizens and towns whose population is less than 500k. In the former case our data assign 33% of Facebook users to those cities, while UN data only reports 21%, a 12 percentage point difference. This is aligned to the 11 percentage point difference for cities with less than 500k citizens, since our data predicts 40% and UN data 51%. We believe that part of this deviation is due to the small amount of users our data reports for villages below 10K users (less than 4%), since probably this portion is considerably larger in reality, but we believe people on those villages shows a much lower penetration in the use of technology (including OSNs)

and thus Facebook results are biased.

4.2.7 Conclusion

In this study with the goal of understanding the degree of Facebook profile's informartion disclosure, we study the privacy status of Facebook profiles by analyzing the profile's attributes disclosure degree in a dataset including 479K Facebook profiles publicly available information that we have crawled from March to June 2012. The analysis of this data reveals the following main insights about the disclosed information in Facebook profiles. (*i*) Friend-list is the attribute with the largest public exposure, whereas Birthday attribute is the one showing major privacy concerns from Facebook users. (*ii*) We find strong correlations between *Current City* and *Home Town* attributes as well as (i.e. *College* and *HighSchool*) and professional (i.e. *Employers*) attributes. (*iii*) In average Facebook users make more than 4 attributes publicly available in their profiles. (*iv*) Men show a larger public exposure as compared to women for all personal attributes except *birthday*. (*v*) The more representative age range based on the public available information is 18-25 that accounts for 1/2 of the users among those ones making its Birthday publicly available. (*vi*) We show that Facebook accurately estimates the portion of people living in different class of cities.

4.3 Cross-posting activity of professional users across Facebook, Twitter and Google+

4.3.1 Introduction

Online Social Networks (OSNs) have become one of the most popular services in the Internet attracting billions of subscribers and millions of daily active users. This tremendous success has created a very profitable market in which major OSN players have acquired an important role on the current Internet. We can find three dominant OSNs according to their number of subscribers: Facebook (FB), Twitter (TW) and Google+ (G+). While these systems have been demonstrated to be very attractive to regular users that perform a wide variety of social interactions on them, they also present a golden opportunity to professional players (i.e. brands, politicians, celebrities, etc.) to interact with a huge amount of potential customers/voters/fans to increase their reputation and popularity, to run marketing campaigns, to attract voters, etc. In a nutshell, we can find a number of professional players that are using OSNs with a similarly professional goal.

Most professional users do not limit their activity to a single OSN, but usually they have accounts in multiple OSNs, including the most popular ones such as FB, TW and G+. Then an interesting question is whether professional players use all OSNs in the same way, or actually they use each OSN for different purposes. In other words, when a professional user wants to advertise or notify some update, does she publish that information in several OSNs?, or contrary, she publishes it in a single OSN depending on the type of information (e.g., if it is a personal update she publishes a post in one OSN, but in case it is a commercial update she selects another OSN). We refer to the information that a professional player publishes in multiple OSNs as *cross-posting activity*. Therefore, if a professional user publishes a post in FB and a post TW that contain the same information we consider them as a cross-post.

To the best of our knowledge, although there are other works that have analyzed the behaviour of regular users across two OSNs [117, 118], this study presents the first large scale study on cross-posting activity of professional users across the three major OSNs, i.e., FB, TW and G+. We analyze the activity of 616 (popular) professional users with active accounts in the three referred OSNs. Among these users we can find big companies, politicians, athletes, artists, celebrities, public institutions, etc. To perform the study we have analyzed more than 2M posts distributed across the 616 users in TW, FB and G+.

The first contribution of this study is a simple yet efficient methodology that is able to precisely determine whether two posts contain the same information, and thus classify them as a cross-post. This methodology relies in a hierarchical algorithm implemented in two steps. The first step applies NTLK Fuzzy logic [119] to compare a pair of posts, and

provides a binary decision on whether they actually represent a cross-post. Those pairs of posts obtaining a positive comparison are already classified as cross-post at this stage, while the pairs failing in this comparison go to the second step of the algorithm. The second step of the algorithm uses two metrics, cosine similarity [120] and string similarity [121], which provides a similarity value ranging between 0 and 1 for each pair of posts under comparison. Then the closer the similarity value is to 1 the more similar the posts are. We classify as cross-post any pair of posts obtaining a similarity value ≥ 0.5 for both metrics, cosine similarity and string similarity. The validation of our methodology shows an accuracy of 99% for the classification of cross-posts.

Based on this methodology, the first goal of the study is to characterize the cross-posting activity of professional OSN users across FB, TW and G+. In order to achieve this objective we perform a data analysis that allows us to shed light to three key aspects of the cross-posting activity. (*i*) The first immediate question is whether the cross-posting phenomenon actually exists, and if it exists what fraction of the activity from a professional user is associated to cross-posting. (*ii*) In case the cross-posting activity is relevant, we aim at understanding between which OSNs it is more frequent. This means, can we find more cross-posts between FB-TW, FB-G+, or TW-G+? (*iii*) Finally, we measure what is the benefit, if any, that professional users obtain from the cross posting activity in terms of engagement.

Once we have characterized the cross-posting behaviour, we study which is the preferred OSN of professional users as initial source to inject information. Indeed, when a professional user decides to publish her updates first in an OSN than other, she is privileging the first OSN that somehow is showing the "breaking news" for that user.

Finally, our last effort defines cross-posting behavioural patterns for users with some representative characteristic that determines their profile. First, we characterize the behaviour of professional users with a strong preference for initiating their cross-posts in a particular OSN using three metrics: (*i*) similarity of their cross-posts, (*ii*) type of content associated to the cross-post they publish, and (*iii*) sites more frequently linked by the urls contained in their cross-posts. In addition, we repeat the analysis but classifying the users based on their median inter-posting interval, which refers to the time gap between the moment they publish the cross-post in the first OSN and the instant it is uploaded in the second OSN.

Following, we list the main findings of our research:

(1) Cross-posting is a frequent practice across professional users. In median a professional user share in other OSN 25% of the posts published in FB and G+, and only 3% of the tweets. However, we must note that professional users are much more active in TW than FB and G+, hence, in absolute terms, the TW account of professional users generate a

larger volume of cross-posts than G+ accounts and similar volume to FB accounts.
(2) The cross-posting phenomenon mainly happens between FB and TW, but it is also relevant between FB and G+. However, it is surprising that is more likely to find a cross-post published in FB, TW and G+, than only in TW and G+. Therefore, professional users do not find any benefit on sharing information between their TW and G+ accounts.
(3) Professional users obtain a substantial benefit in FB and TW when they publish cross posts since they attract 30% and 100% more engagement as compared to non-cross-post. However, in the case of G+ non-cross-posts attract 2× more engagement than cross-posts.
(4) Among the 616 analyzed users 50% prefer FB as most frequent option to initially upload their cross-posts, 45% prefer TW, and only 5% give priority to G+.
(5) Professional users with a strong preference for TW publish cross-posts that: (*i*) are very similar across the different OSNs, (*ii*) mostly includes textual content, and (*iii*) mostly include links to websites different than OSNs sites.
(6) Professional users with a strong preference for FB publish cross-posts that: (*i*) mostly includes audiovisual content, and (*ii*) mostly include links to content stored in major OSNs sites.
(7) As the inter-posting interval decreases: (i) the similarity of cross-posts increases, (ii) the portion of audiovisual content attached to cross-posts decreases, (iii) and a larger portion of urls included in cross-posts refers to major OSNs sites.

4.3.2 Related Work

There exist several works that have studied the graph and connectivity properties of Facebook, [2–4], Twitter [5, 6], and Google+ [7, 8]. In addition, there are other works in the literature that compare two or more OSNs based on their graph properties [10, 17]. However, these works do not consider the same users in the different OSNs for their analysis since their goal is to characterize OSNs at a macroscopic level.

There are only few works that try to characterize the behaviour of the same user or group of users across different OSNs. The main reason is that it is not an easy task to identify and collect the information of the same users across different system and, in addition, it requires to have one data collection tool for each system. There are some few tools and platforms available in the market [122, 123] that provide some few information (for free) of a given user across different OSN. However, that information is usually limited to the number of followers, the number of published posts, aggregated engagement and/or popularity trends. Therefore, these tools do not provide enough detail on the activity of a user to perform a comprehensive analysis of its behaviour in different OSNs.

Nevertheless, some few studies in the literature have analyzed the behaviour of the same users across different OSNs. Authors in [118] compare 195 users from the archival

community and study their activity pattern in TW and FB. This is a small-scale study based on 2,926 links to external documents. In [124], we find again a comparative analysis for users having accounts in FB and TW. This work studies the behaviour of 300 users from a psychological perspective and the results reveal a correlation between end-users personality and their use of FB and TW. Finally, the most similar work to our study is a very recent study [117] that compares the behaviour of 30,000 regular users across TW and Pinterest. Although this study similar in spirit to our work, we differ from [117] since we are focusing in professional OSN players instead of regular users, and we are comparing TW, FB and G+ instead of TW and Pinterest.

4.3.3 Data Collection Methodology

This section briefly explains our data collection methodology to construct the required dataset to achieve the objectives addressed in this study.

Our first challenge was to identify a numerous group of relevant professional users having active and popular accounts across FB, TW and G+. To this end, we rely on a large dataset that includes thousands of professional and regular users with an account in the three OSNs collected for a previous work [125]. From these users we were interested in those ones that meet two requirements: (i) have an active account in FB, TW and G+; (ii) present a high popularity in at least two of the systems. We found 616 professional users that satisfy the popularity requirement. We validated that the selected users were actually relevant in all the three OSNs by means of an external source [122] ranks professional users in each system in terms of popularity. Subsequently we briefly introduce the crawlers developed to retrieve the activity of the professional users from each OSN. For more details on these crawlers we refer the reader to [9, 125]:

FB crawler We have implemented a Crawler Facebook fan pages based on the FB API[2] which is able to collect different information for a user including its popularity, activities and reactions. The crawler receives a user ID (or username) as input and uses the FB API to collect the posts published by the user in her FB account. The API provides quite a lot information from a post from which the most relevant for our study is: (*i*) the description of the post that refers to the text included by the user in that post, (*ii*) the timestamp associated to the exact publication time of the post, and (*iii*) the type of content associated to the post, which could be photo, video, link (when the post includes an url) and status (that refers to the post that only include text). It must be noted that FB API imposes a maximum threshold of 600 queries every 10 minutes. Hence, in order to speed up our data collection process, we used multiple instances of the crawler working in parallel. In

[2]https://developers.facebook.com/tools/explorer/

summary our developed crawler is able to gather two category of information: i) user level information as shown in table 4.8 and ii) post level information as shows in table 4.9.

Table 4.8: FB Users attributes collected by the crawler

Category	**Information**
Page Information Crawler (per Page)	Page Name Page ID Page #fans Page #people talking about Page Category

Table 4.9: FB Posts' Attributes collected by the crawler

Category	**Information**
Post information (all this information is per Post)	Post ID Post type Post Description Post Created Time Post Updated Time Post total #Likes Post total #Comments Post total #Shares
User's Like Info. (per Post)	user id who put Like in the post user name who put Like in the post
User's Comment Info. (per Post)	user id who put comment user name comment ID (one unique ID per comment) user comment's text (the context of the comment) User comment's Created Time User comment's Like Count (#likes that other users put for the comment)

It should be noted that the Facebook API has several limitations in terms of number of queries and the amount of information that can be retrieved from it. Firstly, for gathering posts information, the crawler needs to access the API using an access token that has its own limitations. The API queries are limited to 600 queries per 600 seconds for each access token that we use to connect to API. As all our crawlers are based on the API, this limitation applies to all of them. This is mitigated by using different machines (with different IP addresses) to make it in parallel. Secondly, FB API only provides the identity of the last 5K users that clicked on the button like for a post, and the last 1K users commenting on the posts. This means that for posts that are very popular and attract more than 5K likes and/or 1K comments we are just able to gather the identity of the last users that reacted to the posts.

TW crawler In collaboration with ONRG team[3] in University of Oregon, we had access to their Twitter crawler and collected required datasets for this research. The crawler receives as input a user identifier that can be either the user's id or the user's screen name and queries the Twitter API to obtain the user's profile attributes, the total number of

[3]Oregon Network Research Group, `http://mirage.cs.uoregon.edu/index.html`

Table 4.10: Dataset description

OSN	total posts	avg. posts per user
FB	422 K	685
G+	173 K	280
TW	1.64 M	2664

published tweets, and the last 3,200 tweets posted by the user along with the number of reactions associated with each one of the user's tweets, except the responses (i.e., comments) for a tweet. Consequentially if a user has published more than 3,200 tweets we can only retrieve the last 3,200. Twitter imposes a limit of 150 requests per hour per IP address. To overcome this limitation, we use PlanetLab [126] infrastructure to parallelize our data collection process. Specifically, our crawler sends requests to TW API using approximately 450 PlanetLab machines as proxies, so that we can multiply the speed of our data collection in proportion to the number of used proxies.

G+ Crawler In collaboration with people from NETCOM group[4] in Universidad Carlos III de Madrid, we had access to their G+ crawler and we have collected required datasets for this research. This crawler is composed by two modules. The first one collects the public profile information as well as the connectivity information of all the users in the largest connected component (LCC) of G+. This module is a web-crawler that parses the web page of G+ users to collect the previous information. The second module uses the G+ API to collect all the public posts as well as their associated reactions. Google limits the number of queries to the G+ API to 10K per hour per access token. In order to overcome this limitation we have created several hundred accounts with their correspondent access tokens and leverage the proxies infrastructure in PlanetLab explained above to speed up our crawling data collection.

Table 4.18 summarizes the datasets used in this study. In total, we analyze more than 2M posts published across 616 professional publishers in FB, TW and G+. Finally, it must be noted that the collection campaign finished on May 2013, thus our dataset may not include novel features released by any of the analyzed OSNs after that period.

4.3.3.1 Methodology to Identify Cross-posts

In order to being able to compare cross-posting activity of professional users we need to have an accurate mechanism that detects when two posts are actually containing the same information. For instance, a given user could upload a post in FB and TW which refers exactly to a recent event, but in the case of FB she uploads a picture and in TW she adds a link to the picture. In other words, a tool that only detects as cross-posts

[4]NETCOM Research Group, `http://netcom.it.uc3m.es/`

those posts that are exactly the same in two OSNs is inaccurate for our research. Hence, we have implemented a hierarchical classification algorithm that determines whether two posts can be considered as cross-posts in two steps. Then, given the description (i.e. the text associated to a post) of two posts, P_1 retrieved from the account of user U in OSN_A and P_2 published by U in her account of OSN_B, our algorithm proceeds as follows:

(1) We compare P_1 and P_2 using NTLK Fuzzy Match [119] that provides a binary decision based on the similarity of the compared texts. NTLK Fuzzy Match generates a positive answer (i.e., same text) when both texts are very similar and only differ in some few characters. Therefore, in the context of cross-posting analysis if NTLK Fuzzy Match determines that $P1$ and $P2$ are similar, we can safely classify them as cross-post. However, in the case that the output is negative we cannot guarantee that P_1 and P_2 are not referring to the same information, thus we cannot classify them as non-cross-post. In summary, all the pairs of posts receiving a positive classification are labelled as cross-posts while the remaining pairs need to go through the second step of our algorithm.

(2) We compare P_1 and P_2 using two similarity metrics: cosine similarity [120] and string similarity [121]. These two metrics provide as output a value ranging between 0 and 1, so that the closer is the output to 1 the more similar P_1 and P_2 are. Based on the obtained results, we classify P_1 and P_2 as cross-post if both metrics, cosine similarity and string similarity, are ≥ 0.5. Later in this section we validate our methodology and demonstrate why we have selected the 0.5 threshold.

The previous algorithm serves to classify any pair of posts as cross or non-cross based on their description. In addition, we must note that our algorithm is not bound to any particular alphabet, so it can be applied in multiple languages. However, the use of the hierarchical algorithm is not enough for the purpose of this research. Following we describe two more elements we had to integrate in our methodology to ensure the accuracy of the results obtained in the study.

First, we had to define which pairs of posts should be compared together. A straightforward solution had been to compare, for a given user, each post in FB to all posts in TW and all posts in G+. However, that option would be inaccurate because we have observed that some users utilize repetitive patterns over time. For instance, we found a user that publishes frequently posts with the content "love you my fans", thus following an all to all comparison approach would lead to a wrong classification for quite a lot cross-posts. In order to be accurate and efficient we applied the following methodology. Given a post P_{FB} published by a user U in her FB account at the timestamp t_{FB}, we compare P_{FB} with all the posts that user U published in her TW and G+ accounts in a time window starting one week before and finishing one week after t_{FB}. In other words, we compare each post in a time window of two weeks around the date that post was published.

Table 4.11: Methodology validation, false positive (FP) and false negative (FN) rates of different similarity threshold (ST) in our cross-posting identification methodology.

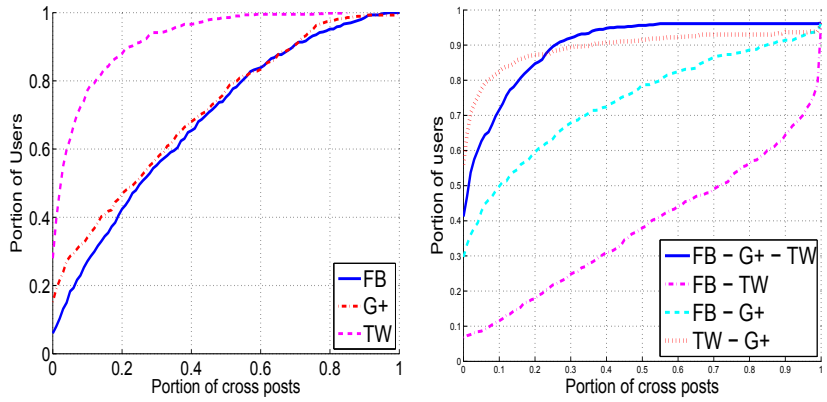
ST>0.3 similarity		ST>0.5 similarity		ST>0.7 similarity	
FP	FN	FP	FN	FP	FN
15.006	0.194	0.140	1.117	0.016	4.593

Second, TW API limits the number of retrieved posts for any user to the last 3,200 posts she published, while FB and G+ do not have that limitation and provide all the posts published by the user since she registered in the system. Hence, it may happen that for a given user we only have 6 months of posts for TW, but several years for FB and G+. Therefore, in this case it only makes sense to analyze that user for the last 6 months because we would not be able to determine if the information associated to a post published in FB or G+ one year ago was also available in TW at that time. Hence, in order to perform an accurate study, we have restricted our cross-post analysis to the time window imposed by the limitation of TW API for each user in our dataset. It must be noted that the number of posts depicted in table 4.18 already consider this limitation.

We applied the described methodology to the selected 616 OSN professional users and we found 176K cross-posts across their OSNs accounts.

Methodology Validation In order to ensure the accuracy of the proposed methodology 3 persons manually classified 12.8K random posts as cross-posts or non-cross-posts. In order to have a meaningful validation set we ensured that half of the posts had been labelled as cross-post and half as non-cross-posts by our classification tool. Then, given two posts published by a user in two different OSNs we classify them as a cross-post if at least 2 out of the 3 persons performing the manual inspection indicate that both posts contain the same information. This allows us to obtain a ground truth set to determine the false positive and false negative rate of our methodology. A false positive occurs when our tool classifies as cross-post two posts (published by the same user in two different OSNs) that are actually referring to a different piece of information. A false negative happens when our tool classifies as non-cross-post two posts that actually contains the same information.

Based on the ground truth set we compute the false negative and false positive rate for our methodology using three different thresholds for the second step of the algorithm: 0.3, 0.5 and 0.7. Basically, a lower threshold requires less similarity between the compared posts to classify them as cross-post. Table 4.11 shows the false positive and false negative rate for our algorithm when it uses each of the evaluated thresholds. The results clearly determine that 0.5 is a very good threshold since it presents a very low rate for false positives (0.14%) and false negatives (1.11%). However, on the one hand, a threshold of 0.3 imposes a very low similarity to classify two posts as cross-post and thus it presents an unacceptable false



(a) (b)

Figure 4.3: (a) CDF for the portion of cross-posts per user in FB, G+ and TW. (b) CDF for the portion of cross-posts and in each possible cross-posting pattern (FB-TW, FB-G+, TW-G+ or FB-TW-G+).

Table 4.12: median and average values for absolute number (and percentage) of Cross posts across users in FB, G+ and TW

	#Cross Posts		Portion of cross posts (%)	
OSNs	Median	Average	Median	Average
FB	114	231	26.42	32.14
G+	20	83	24.50	29.31
TW	85	196	3.34	7.36

positive rate (15%). On the other hand, a threshold of 0.7 is too strict and it skips some pairs that actually contains the same information and classify them as non-cross-posts, thus presenting a poor false negative rate (4.5%). Therefore, based on this experimental validation of our methodology, we decided to establish a threshold of 0.5 in our algorithm.

4.3.4 Cross-Posting Characterization

The first question we aim to answer in this section is whether the cross-posting phenomenon exists in the activity of professional users, and what is its weight in FB, TW and G+. Subsequently, we look at how this cross-posting occurs among the three OSNs under analysis. To this end, we quantify what is the volume of cross-posting happening between FB-G+, FB-TW, TW-G+ and FB-TW-G+, in order to determine what pair of OSNs is actually sharing more common information. Finally, we also want to characterize the impact of cross-posting in the attracted engagement measured in terms of likes comments, and shares.

4.3.4.1 Quantification of cross-posting activity

The goal is to quantify the cross-posting phenomenon for professional users in FB, TW and G+. Towards this end, we compute for each user and each OSN the portion of cross posts with respect to all the posts each user has published. For instance, given a user U and her FB account we compute how many posts published in that account also appear in TW, G+ or both. We quantify the same parameter for the TW and G+ accounts of user U[5].

Figure 4.3(a) shows the CDF for the portion of cross posts across the 616 users analyzed in the three OSNs. The x axis refers to the portion of posts and the y axis to the portion of users. For instance, the point {x=0.2, y=0.4} in the line associated to FB indicates that 40% of the users have $\leq$ 20% of cross-posts in their FB accounts.

The first immediate conclusion extracted from the graph is that most of the professional users have published some cross-post. In particular, when we consider FB accounts we find that only 6% of the users do not have any cross-post, which means for those users the information published in FB cannot be found neither in TW nor in G+. This number grows up to 15% and 28% for G+ and TW, respectively. Therefore, a vast majority of professional users published some cross-post at some point. Hence, the first conclusion is that in general professional users find some value on the cross-posting activity.

If we compare the results obtained for the three OSNs, we clearly observe that, in relative terms, the cross-posting activity is more frequent for those posts published in FB and G+ than in TW. The results for TW show that most of the tweets are not replicated neither in FB nor in G+. The median value, which shows the typical portion of cross-posts for a user in each OSN, shows that for a typical professional user around 1/4 of the posts that appear in FB and 1/4 of the posts that appear in G+ are also available in at least one more OSN. However, in the case of TW, out of 100 tweets only 3 of them are replicated in other OSNs. Finally, we can find quite a lot professional users with an intensive cross posting activity. In particular, 25%, 23% and 1.5% of the analyzed users, in FB, G+ and TW, respectively, have published more cross posts (i.e., $\geq$ 50%) than posts appearing exclusively in a single OSN. We refer to these posts as *non-cross-posts*.

The previous analysis refers to the cross-posting activity in relative terms. However, it is important to notice that, according to the overall activity of the professional users in our dataset, the publishing rate of professional users in TW is 4$\times$ higher than in FB and G+. Table 4.12 presents the median and average values for the absolute number and portion of cross-posts per user in each OSN. The results reveal that although TW presents a much lower cross-posting activity in relative terms, it actually has a larger number of cross-posts

[5] It must be noted that for this analysis we do not take into account where the post appears first, but only consider whether it is unique in an OSN or it appears in 2 or 3 of them.

than G+, and it is much closer to FB in absolute cross-posts number.

4.3.4.2 Inter-OSN Cross-Posting

Once we have demonstrated that cross-posting is a common practice among professional users in FB, TW and G+, we analyze how cross-posting happens among them. Then, our goal is to find whether professional users prefer to share things in FB and TW, or rather it is more frequent finding common posts in FB and G+, or actually there are lots of cross-posts published in TW and G+.

In order to perform this analysis we proceed as follows. For a given user U we get all her cross-posts in FB (independently whether the first appearance happened in that OSN or another one) and compute which portion of them also appears in TW, which portion in G+ and which portion in both TW and G+. We repeat the same process for the TW and G+ accounts of user U. Therefore, for each user we know what is the cross-posting level for the following relations: $FB-TW$, $FB-G+$, $TW-G+$ and $FB-TW-G+$.

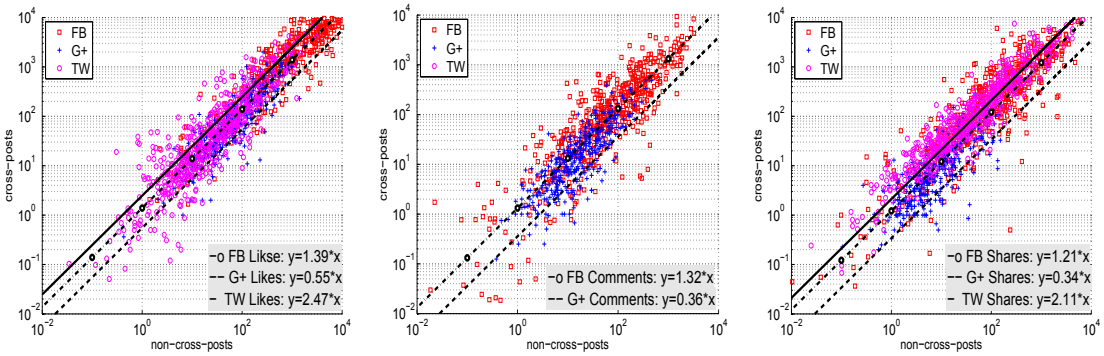
Figure 4.3(b) shows the CDF for the portion of cross-posts that occurs for the four referred relations across the users in our dataset. Again in this figure the x axis refers to portion of posts and the y axis shows the portion of users. Then for instance the point x=0.4, y=0.3 in the $FB-TW$ line indicates that 30% of the users publish $\leq$40% of their cross posts in FB and TW.

The results in the figure demonstrate that professional users perform much more cross-posting between FB and TW than in any other combination of OSNs. This claim is supported by the fact that in median a professional user publishes 70% of their cross-posts in FB and TW. In addition, we can only find 8% of the users that never shared a post between their FB and TW accounts, while this value grows to 30% between FB and G+, to 40% for the case in which the three OSNs are involved, and to 55% when we consider TW and G+. Therefore, this last result surprisingly states that is more likely that a user publishes the same posts in the three OSNs than only in TW and G+.

In a nutshell, we can find more cross-posting between FB and TW (in either direction) than with G+, while the specific cross-posting between TW and G+ (in either direction) appears as the least preferred option, since users prefer to publish the information in all the 3 OSNs than only in TW and G+.

4.3.4.3 Engagement Analysis

A plausible reason of why professional OSN users publish the same information across different OSNs is to try to increase the coverage in order to engage as many end-user as possible within their accounts. Therefore, in this subsection we want to conclude whether cross-posts achieve more engagement than non-cross-posts in FB, TW and G+. In order



(a) Likes (b) Comments (c) Shares

Figure 4.4: Users' average attracted engagement per post, for cross posts initiated in each OSNs vs. non-cross posts.

to measure the engagement we use standard reaction mechanisms available for end users in OSNs: likes, comments and shares[6]. As we acknowledge in Section 4.4.3, our TW collection tool could not retrieve comments. These reaction mechanisms allow professional users to interact with end-users through its OSN account and obtain a very valuable first-hand feedback from them. Therefore, engaging as many end-users as possible is an important goal for professional OSNs users.

In order to measure the efficiency of cross-posts to attract engagement in one OSN we measure, for a given user U, the average engagement for $U's$ non-cross-posts and $U's$ cross-posts initiated in that OSN in terms of likes, comments and shares. We apply this methodology to all the users for their FB, G+ and TW accounts.Figure 4.4 shows a scatter plot for FB, G+ and TW for each of the engagement type: likes (Figure 4.4(a)), comments (Figure 4.4(b)) and shares (Figure 4.4(c)). Each point in the graphs represents a user with an x coordinate referring to the average engagement for non-cross-posts and y coordinate referring to the average engagement for cross-posts initiated by that user in that OSN. In addition, all the figures include three lines (one per OSN) showing the linear regression for the cloud of points represented by an equation[7] of type $y = ax$. When the slope of the linear regression, represented by the value of a, is greater than 1, it means that for that OSN cross-posting is worthy since cross-posts attract more engagement than non-cross-posts in average.

The results demonstrate that cross-posts in FB and TW allows professional users to attract more engagement than non-cross-posts. However in the case of G+ cross-posts

[6]This is the nomenclature employed in FB. A like is associated to a +1 in G+ and to a favourite in TW. A share is associated to reshare in G+ and a retweet in TW.

[7]Usually a linear regression is represented as y=ax+b, but in the figure we just use y=ax, since we are interested in the slope, but not in the offset

Table 4.13: Cross-Posts initiated in FB, TW and G+.

OSN	#Posts	%Posts
FB	74355	42.17
G+	12002	6.81
TW	80497	45.66
other	9451	5.36

receive considerably less attention than non-cross-posts. In more detail, a FB user attracts 39% more likes, 32% more comments and 21% more shares in FB when she uses cross-posts instead of non-cross-posts. In the case of TW cross-posting provides even more benefit. This is, a cross-post initiated in the TW account of a professional user attracts 2.47× and 2.1× more likes (i.e., favourites) and shares (i.e., retweets) than a non-cross-posts. Finally, in the case of G+ a cross-post roughly achieves 1/2 of the likes (i.e, +1's), 1/3 of the comments and 1/3 of the shares compared to non-cross-posts. Therefore, cross-posting seems to be a bad strategy if the goal of a professional user is to attract as many reactions as possible in G+.

In summary, cross-posting exists and it is a frequent practice across professional users in FB, TW and G+. It mostly happens between the FB and TW accounts of professional users, and it very rarely occurs between TW and G+. Finally, in terms of attracted engagement, cross-posting is beneficial in FB and TW, but not in G+.

4.3.5 Preference of Professional Publishers

Professional users utilize OSNs to interact with their followers and share with them more or less relevant information. In previous section we have demonstrated that quite frequently an end-user can find the same information in two (or more) OSNs. Based on this finding, in this section we tackle two interesting questions. First, we want to know in overall which OSN is used more frequently as first option to publish fresh information that later will be republished in other OSNs. Second, we want to understand what is the OSN that professional users prefer to publish first the information. Answering the first question will determine which OSN is used more times as source of cross-OSN information, while the response to the second question will roughly determine what is the OSN that professional users value more to publish first their fresh updates.

4.3.5.1 OSN-based Analysis

Table 4.13 shows the number and portion of cross-posts in our dataset that were initiated in FB, TW and G+. The results demonstrate that TW appears as initial source of information for 45% of the cross-posts closely followed by FB with 42%, while G+ is rarely chosen as first option. Finally, we find a very interesting result associated to the category *"other"* that represents those cross-posts that could not be assigned to a particular OSN since they

Table 4.14: Portion of cross-posts published for first time in FB, TW or G+ for different cross-posting patterns: $FB-TW-G+$, $FB-TW$. $FB-G+$, $TW-G+$. The table also includes the portion of posts that are published in at least two OSNs at the same time (i.e., exact timestamps)

cross-posting pattern	#Posts	%Posts	%FB (1st)	%G+ (1st)	%TW (1st)	%posts with same publishing time
FB - G+ - TW	18619	10.56	34.93	12.32	49.80	2.95
FB - G+	34337	19.48	73.68	24.07	-	2.26
FB - TW	117276	66.52	36.28	-	56.80	6.92
G+ - TW	6073	3.44	-	23.79	75.96	0.25

Table 4.15: Preferred OSN per user

OSN	#Users	%Users
FB	307	50
G+	30	5
TW	275	45

were published exactly at the same time (i.e., same timestamp) in at least two OSNs. It is surprising that almost 10K cross-posts, which represent 5.3% of all the cross-posts in our dataset, experienced this parallel publication. This reflects the use of automatic publishing tools that upload in parallel some information to two or more OSNs.

As we determined in the previous section, most of the posts are not published in all the three OSNs, but just two of them. Therefore, it is interesting to analyze for each particular publishing pattern which OSN appears more frequently as initial source of information. Table 4.14 shows the results for all the possible cross-post patterns: $FB-TW-G+$, $FB-TW$, $FB-G+$ and $TW-G+$. First of all, the results confirm the conclusion obtained in the previous section since 2/3 of the cross-posts appear exclusively in FB and TW, 1/5 belong to the category $FB-G+$, and as we already stated it is more likely finding cross-posts across the three OSNs (10%) than only across G+ and TW (3.4%). In the most popular category, i.e., $FB-TW$, TW appears as first option for 57% of the posts while FB is chosen in first place only 36% of the times. When G+ competes individually either with FB or TW, it is source of information only 1/4 of the times. For those posts published in the three OSNs, 1/2 of them appear first in TW, 1/3 in FB and 1/10 in G+.

Finally, we want to highlight that all the categories include some portion of posts that where published in parallel at the same exact time in two OSNs. This phenomenon is especially relevant for cross-posts between $FB-TW$.

In summary, the OSN-based analysis demonstrates that Twitter is the OSN selected as initial source of information more frequently. FB appears as the second option close to Twitter. Finally, G+ is the least preferred option.

Table 4.16: Users classification based on different OSN preference criteria: (*i*) users initiating 100% of their cross-posts from one OSN; (*ii*) users initiating ≥80% of their cross posts from one OSN; (*iii*) users starting <50% of their posts from all three OSNs.

Criteria	#User	#FB	%FB	#G+	%G+	#TW	%TW
100%	32	11	1.79	2	0.32	19	3.08
≥80%	182	75	12.18	5	0.81	102	16.56
<50%	95	-	-	-	-	-	-

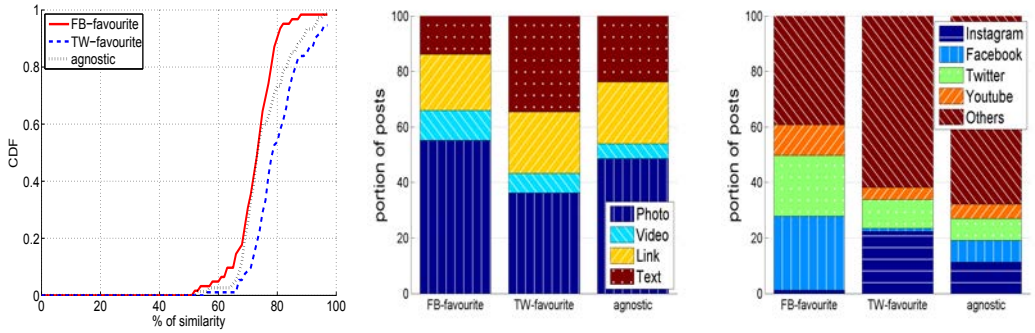
4.3.5.2 User-based Analysis

The OSN-based analysis revealed that Twitter is chosen as first option for a larger number of cross-posts. However, we cannot extract from that analysis that TW is the preferred OSN for most of the users, since it may happen that very active users contributing a large number of posts prefer TW but less active users prefer FB or G+. Therefore, in this section we analyze which is the preferred OSN for professional users. For a given user its preferred OSN is the one she selected in first place for a major number of posts. For instance, if a user has generated 20 cross-posts from which 10 were first published in FB, 6 in G+ and 4 in TW, we define FB as the preferred OSN for that user. Table 4.15 shows the number and portion of users in our dataset that prefer each OSN. The results reveal that half of the professional users prefer FB, closely followed by 45% of the users that prefer TW, while only 5% of the users chooses G+ as initial OSN for publishing their post. Therefore, FB and TW has exchanged their positions as compared to the OSN-based results. As we indicated above, the difference between the post-based and user-based results comes from the fact that users tend to be more active in TW.

Once we have classified professional users' preference, a subsequent question is, can we find users that shows a strong preference for a particular OSN? In other words, are there users that utilize as source of information one single OSN for most of their cross-posts?

Table 4.16 shows the number and portion of professional users in our dataset that choose either FB, TW or G+ to initiate 100% or 80% of their cross-posts showing a clear strong preference. In addition, we also quantify the number and portion of users that publish in first place less than 50% of their posts in all three OSNs and thus do not show any strong preference. We can find 19, 11 and 2 users that always choose TW, FB and G+ as initial source for their cross-posting activity, respectively. If we move down the threshold to 80% the number of users showing a clear evidence of which OSN they prefer grows a lot for FB and TW, but not for G+ that only accounts for 5 users. There are 75 (12.18%) users with a preference for FB and 102 (16.56%) with a noticeable preference for TW. In contrast to these users showing a clear OSN preference, we can find 95 (15.4%) users that are not biased towards any OSN, even though they make use of cross-posts.

In summary, professional users are (more or less) equally divided into those that prefer



(a) CDF of cross-post similarity (b) Type of post (c) Website associated to urls

Figure 4.5: Cross-posting behaviour characterization based on professional users preference.

TW and those that prefer FB, and very few cases that show a preference for G+.

4.3.6 Cross-posting Behavioural Patterns

We have fully characterized the cross-posting phenomenon as well as what is the preferred OSN for professional users in the context of cross-posting. To finalize this study we want to explore the presence of explicit differences in the cross-posting activity for groups of users presenting different but well defined profiles according to a given characteristic. We will focus on two characteristics: (*i*) OSN preference and (*ii*) inter-posting interval. First, the goal is to determine whether there are significant behavioural differences in the cross-posting pattern for professional users showing a strong preference for TW, professional users showing a strong preference for FB and agnostic users. Second, we separately analyze whether professional users publishing their cross-posts in two OSNs in a short time window show some behavioural differences compared to users that delay a lot the publication of the cross-posts in the second OSN. We refer to the time window between the publication in the first OSN and the second OSN as *inter-posting interval*.

We characterize the cross-posting behaviour using three parameters that will help to determine the difference among the profiles we are comparing. These parameters are: (*i*) the cross-post similarity value obtained from the methodology described in Section 4.3.3.1, (*ii*) the type of content associated to the cross-posts according to the category assigned by the FB API to the posts, (*iii*) the website associated to the urls contained within TW version of the cross-posts (i.e., tweets).

Due to lack of space we will perform this analysis for the cross-posts shared between FB and TW that, as Table 4.14 depicts, represent 66% of the total cross-posts, which increases to more than 75% if we also consider the cross-posts that appear in the 3 OSNs (thus also in FB and TW).

4.3.6.1 Cross-Posting behaviour based on Preference

We create three groups of users according to the results obtained in the previous section (see Table 4.16). The first group, referred to as *TW-favourite*, is formed by the 102 users that show a strong preference for TW. The second group, referred to as *FB-favourite* is formed by the 75 users showing a clear preference for FB. Finally, the last group is formed by the 95 "agnostic" users that do not show any strong preference, and we refer to it as *Agnostic*. Next we characterize the cross-posting pattern for these four groups based on the three characteristics introduced at the beginning of this section.

Figure 4.5(a) shows the CDF for the cross-post similarity across the users in the three groups. The results show that the users with a strong preference for TW publish more similar cross-posts between FB and TW than the users that prefer FB. In median *TW-favourite* group achieves an average similarity close to 80%, while *FB-favourite* and *Agnostic* groups reach a median similarity a bit higher than 70%.

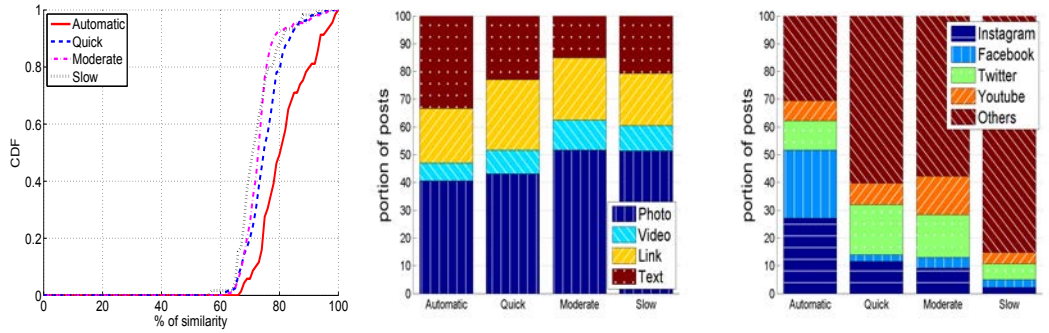
In order to classify the type of content embedded in the posts we rely on the content type assigned by the Facebook API to each post that can be: text, link, video or photo. It must be noted that posts classified as photo or video by FB API may not include the video or photo in TW, but a link to them. Figure 4.5(b) shows a bar plot presenting the average portion of each type of content published within each group. We observe a substantial difference between *FB-favourite* and *TW-favourite* groups. Users that prefer FB attach photos to half of the cross-posts. Even more, users in this group are the ones that publish a larger portion of videos. In contrast, *TW-favourite* group includes users that publish much less photos and videos (36% and 7% in average, respectively), but much more posts including only text (35% in average). The agnostic users ranges in between *FB-favourite* and *TW-favourite*.

Finally, we want to find what are the sites more frequently linked from the cross-posts. For this we rely on the urls included in the TW version of the cross-posts (i.e., tweets)[8]

We have found that the most popular websites linked from cross-posts are actually OSNs. In particular, the most linked sites are Facebook, Twitter, Youtube and Instagram. It must be noted that a link to those websites refers in most of the cases to some content (e.g. photo, video, etc) stored in that OSN. Based on these results we analyze the portion of urls linking to those four sites and we group together the remaining urls in a category referred to as *Other*.

Figure 4.5(c) shows a bar plot depicting the average portion of posts including a url linking to Facebook, Twitter, Youtube, Instagram, and Other. Again the results show

[8]TW usually employs short urls. Hence, to obtain the website behind the short urls we had to reverse the process and obtain the original urls from the short urls using "Expand url portal" (http://expandurl.appspot.com/expand?url=http)



(a) CDF of cross-post similarity (b) Type of post (c) Website associated to urls

Figure 4.6: Cross-posting behaviour characterization based on inter-posting interval.

different patterns for users preferring FB and users preferring TW. For users in the former group 60% of their urls point to one of the four OSNs, with a clear preference for FB (26%) and TW (22%) and a negligible presence of Instagram. In contrast, for users in *TW-favourite* group more than 60% of the urls link to websites different than the four main OSNs. However, among the OSNs, Instagram is the most popular one (22% of the urls) while the number of urls for FB is negligible. Agnostic users are the users including a larger portion of urls to "Other" websites (almost 70%).

In a nutshell, the cross-posting profile of a TW-favourite user is as follows: (i) she has a higher similarity for the cross-posts, (ii) she publishes more textual content than audiovisual content, and (iii) she links more frequently websites different than OSNs, but across OSNs it mostly link content in Instagram. In contrast, the profile of a FB-favourite user is as follows: (i) she mostly publishes audio-visual content, and (ii) she mostly contains ursl linking content stored on major OSNs, especially stored in FB and TW. Finally agnostic users show an intermediate behaviour between the TW-favourite profile and the FB-favourite profile.

4.3.6.2 Cross-Posting behaviour based on Inter-Posting Interval

We have shown that professional users present different cross-posting pattern depending on the preferred OSN. Similarly, in this section we explore whether the inter-posting interval time reveals different cross-posting profiles as well. Towards this end we group the professional users in our dataset based on the inter-posting interval between FB and TW (independently of the direction). In order to create the groups we apply the K-means clustering algorithm [127] using as parameter the median inter-posting interval of each user. We use this mechanism to discover all the groups, except one that we form manually.

The reason for creating a manual group responds to the fact that more than 5% of the

Table 4.17: Average and maximum cross-posting interval within *Automatic*, *Quick*, *Moderate* and *Slow* groups.

Groups	#Users	Avg(inter-posting interval)	Max.
Automatic	69	2.3 Minutes	68 Minutes
Quick	159	8.3 Minutes	87 Minutes
Moderate	109	2.35 Hours	5.1 Hours
Slow	59	13.5 Hours	6.4 days

posts in our data set are published at the same time in two OSNs, and this portion grows to 7% if we only consider the cross-posts between FB and TW. Even more, 30% of the cross-posts between FB and TW are posted in both OSNs in a timer interval lower than 10 seconds. Therefore, we thought that there might be a relevant number of users that publish most of their $FB-TW$ cross-posts (in either direction) in less than 10 seconds. We make the assumption that for a human-being is unlikely to manually publish a post both in FB and TW in so short gap. Therefore, we consider that if a post is published in both OSNs in a time interval lower than 10 seconds, it means that the user (e.g., the community manager managing the OSNs accounts of the user) is utilizing some automatic tool to upload the cross-post. Actually, there is quite a few tools that provide this feature [Argyle Social[9], dlvr.it[10], bufferapp[11]]. Based on this discussion, we have manually formed a group that includes those users that have published more than 1/2 of their cross-posts in less than 10 seconds both in FB and TW. The group is formed by 69 professional users from our dataset. We refer to this group as *Automatic* since the users forming it are making an intensive use of automatic tools to perform its cross-posting activity.

After creating this group, we run the K-means algorithm [127] to classify the remaining users according to their median inter-posting interval time. We have found 3 clusters whose characteristics (number of users, average and maximum inter-posting interval) along with the characteristics of the *Automatic* group are depicted in Table 4.17. We refer to these groups as *Quick*, *Moderate* and *Slow*. In a nutshell: (*i*) users in the *Quick* group publish their cross-posts in both OSNs in the order of minutes, (*ii*) users in the *Moderate* group take more than 2 hours, and (*iii*) users in the *Slow* group takes a gap of more than 13 hours between the publication in the first OSN and the time they republish the post in the second OSN.

Following, we analyze the behaviour of the users in terms of similarity, type of content, and links to websites.

Figure 4.6(a) shows the CDF for the cross-post similarity across the users in the different groups for $FB-TW$. We find a very interesting pattern that correlates the similarity to the inter-posting interval. Basically, the shorter the inter-posting interval the higher the

[9] `http://www.argylesocial.com/`
[10] `https://www.dlvr.it/`
[11] `https://www.bufferapp.com/`

similarity across the cross-posts. This actually is reasonable since if you publish the same information in two OSNs within a very short of time (e.g. $< 10s$) the posts are going to look very similar. However, if you post something in TW (FB) and you republish the same information in FB (TW) after some few hours it is more likely that you introduce some change. Finally, we do not observe any relevant difference between the *Moderate* and the *Slow* since the capacity of modifying the post is the same for the associated intervals to these groups.

Figure 4.6(b) show the portion of posts belonging to each content category according to the type assigned by FB API. Again we observe an interesting correlation between inter-posting time and type of published content. As the inter-posting interval increases the portion of cross-posts associated to audiovisual content (i.e., photos and videos) also increases, while the combination of more textual content (i.e., pure text + links) decreases. In particular, the quickest category, i.e., *Automatic* group, publishes around 45% of audiovisual posts, which increases to 50% for *Quick* group and to 60% for *Moderate* and *Slow* groups. Therefore, it seems that automatic tools are employed less frequently to upload videos and photos.

Figure 4.6(c) shows a bar plot depicting the average portion of posts including a cross-post linking to Facebook, Twitter, Youtube, Instagram, and Other. We observe a large divergence for the results across the different groups. First, almost 70% of the urls included in cross-posts of users belonging to *Automatic* group are linking to one of the four main OSNs, with a strong presence for FB (25%) and Instagram (27%). All the remaining groups contains more urls linking "Other" websites than urls linking the four OSNs. It is particularly interesting the very marked pattern of the users in the *Slow* group for which only 15% of the links go to OSNs.

In a nutshell, as the inter-posting interval decreases: (*i*) *the similarity of cross-posts increases,* (*ii*) *the portion of audiovisual content attached to cross-posts decreases,* (*iii*) *and a larger portion of urls included in cross-posts refers to major OSNs sites.*

4.3.7 Conclusions

This study presents the first large-scale measurement-based characterization of the cross-posting activity for OSN professional users across FB, TW and G+. We have used a simple yet efficient methodology that is able to determine with an accuracy of 99% whether two posts, even from different OSNs, contains the same information, and if so classify them as cross-post. We have used that methodology to classify more than 2M posts published for 616 professional publishers with active accounts in FB, TW and G+. Following we list the main outcomes of the study. First, we have demonstrated that professional users frequently publish the same information in at least two OSNs, especially in the case of FB

and G+. Although professional users in TW present a low portion of cross-posts, the fact that they are very active implies that in absolute terms we can find quite a lot cross-posts in their TW accounts. Second, a professional user publishes (in median) 70% of her cross-posts exclusively in FB and TW, and around 15% in FB and G+. Furthermore, we demonstrated that the cross-posting activity between TW and G+ is negligible. Third, professional users benefit of cross-posting in their TW and FB accounts since they attract 2× and 30% more engagement with cross-posts than non-cross-posts, respectively. However, cross-posts in G+ leads to halve the engagement as compared to non-cross-posts. Fourth, professional users equally prefer FB and TW as initial source of information, but they rarely choose G+. Fifth, users with a strong preference for TW present cross-post with a higher similarity (across different OSNs), publish more textual content than photos and videos, and use to include links to websites different than major OSNs. In contrast, users preferring FB publish mainly audiovisual content and a major portion of urls in their cross-posts refer to OSN content. Finally, as the user inter-posting interval time decreases: (i) the similarity of her cross-posts increases, (ii) the portion of audiovisual content attached to her cross-posts decreases as wall, (iii) and a larger portion of urls included in her cross-posts refers to major OSNs sites.

4.4 Characterization of Professional Users' Strategies in major OSNs

4.4.1 Introduction

The tremendous success of Online Social Networks (OSNs) has created a golden opportunity to professional players (i.e. big industry brands, politicians, celebrities, etc.) in order to: interact with a huge amount of potential customers/voters/fans, improve their reputation and popularity, run marketing campaigns, etc. The presence and interest of professional users in OSNs as well as their concern to engage more people [128] with their OSNs accounts is becoming so relevant that we can even find an award ceremony to best professionals users in social media [129].

In this context there is an increasing research interest, especially in the area of management and marketing, to study what are the strategies[12] that professional users apply in their use of OSNs [130–132]. It seems that understanding the factors that allow professional users to engage more people with their OSN activity will have a tremendous value in the future for marketing purposes. To the best of our knowledge most of the studies available in the literature only focus on a limited number of users and extract very particular conclusions for those users that cannot be generalized. Furthermore, all previous studies are either based on manual inspection of OSNs accounts [133] or interviews [134] that cover very few aspects that again lead to not generalizable conclusions. Therefore, we believe that a large-scale data-driven approach based on the actual activity of a large number of professional users across major OSNs will help to shed light into the challenging problem of devising the way professional users utilize OSNs. Towards this end in this study we rely on a dataset formed by 616 very popular users with active accounts in FB, TW and G+. For each user we capture her activity (i.e., published posts) in the three systems over a long-term time window (almost all their activities in OSNs from the time they initiated their accounts) that overall generates a corpus of 2M posts.

In contrast to previous studies we do not aim at studying the strategy of individual users. Instead, our main goal is to make a global analysis to characterize the strategy of a particular sector/category (e.g., Cars Industry, Politician, Athletes, News Media, etc) in OSNs. This analysis can be only conducted for those sectors that fulfil the following hypothesis: *professional users that belong to a particular sector present a similar strategy in OSNs.* Therefore, the first objective of this study is to determine whether this hypothesis is true for some sector. For this we classify the 616 users in our dataset into 62 categories according to the sector reflected by their FB account. Out of these 62 groups only 16

[12]In this study, we will use indistinguishably the terms strategy and behaviour to refer the way a professional utilizes an OSN.

had enough users to perform a meaningful validation of the hypothesis. We apply the methodology proposed in [135] that determines whether the behaviour of the users within a category is significantly similar and, in addition, differs from the behaviour of the users outside that category. The results reveal 8 categories whose users present a common behaviour. These categories are: Athletes, Cars, Media News, Movie, Musician-Band, News Website, Politician, and Sports Teams. After discovering 8 sector fulfilling the baseline hypothesis, we devote our effort to derive the behavioural elements that characterize their use of OSNs.

We base our analysis in a set of meaningful behavioural elements that allow us to discriminate the strategy of each sector. These elements include: activity rate, preference among FB, TW and G+, popularity and type of content published. Using these behavioural elements we are able to describe the strategy and highlight the differential characteristics of each category. There is a last element that, to the best of our knowledge, has never been used to analyze the strategy of professional users across multiple OSNs, which is referred to as *cross-posting activity* and was already discussed in section 4.3. This element captures the volume of common information that a user publishes in more than one OSN. This means, when a professional user wants to post some information she can decide to publish it in a single OSN, or in multiple OSNs. Even more, when she decides to post it in multiple OSNs, there are several combinations of OSNs she could use (e.g., FB-TW or FB-G+ or TW-G+, or the three OSN in our work). Hence, we believe that the *cross-posting activity* of a user is an important behavioural element that for instance reveals whether a user utilizes each OSN for different purposes or not. In a previous study [136] which is presented in section 4.3, we characterize the *cross-posting phenomenon* across professional users.

Finally, to conclude this research we address the very challenging question of whether the strategies implemented by each category are successful or not. To the best of our knowledge there is no standard mechanism in the literature that allows measuring the success of a strategy in OSNs. Therefore, in this study we propose a simple methodology to quantitatively measure such success. The rationale of this methodology is to estimate the number of reactions per post a category should attract based on its popularity, and compare that estimation to the actual number of reactions received by the category. We provide an estimation of the success of each category for eight types of reaction: FB Likes, FB comments, FB shares, G+ +1s, G+ reshares, TW favourite and TW retweets.

The key insights of this study can be summarized as follows:

(1) We demonstrated that for some sectors professional users present a common behaviour. The sectors we found in the study that fulfil this statement are: Athletes, Cars, Media News, Movie, Musician-Band, News Website, Politician, and Sports Teams.

(2) Each of the categories listed above present differential elements in their use of OSNs.

For instance, Athletes activity and preference is biased to TW; categories related to news are extremely active in the three OSNs; Cars is the category with major interest in G+, and Movie shows a low activity and a clear preference for FB.
(3) The categories listed above can be further clustered into three significant groups based on the similarities in their strategies: *individual users* (Athletes, Musician-Band, and Politician), *commercial brands* (Cars and Sport Teams) and *news* (Media News and News Website).
(4) We demonstrate that the level of engagement of a professional user is linearly correlated to her popularity, which allows us to define a model that estimates the number of reactions per post a category should obtain according to its popularity.
(5) The only categories with a successful strategy in FB are Movie (which is successful in all OSNs) and Politician, which is the only category that do not cover the engagement expectation in G+. Similarly, the only two categories that fail in attracting the expected number of reactions in TW are Media News and News Website.

4.4.2 Related Work

There are a number of books [131,132] and reports [137] that propose general guidelines to companies to enhance their marketing strategies in social media. However, most of these guidelines are based on qualitative elements rather than quantitative metrics. Following this line, authors in [138] manually look to the publishing activity of 11 brands from 6 different categories, and provide some general guidelines for the manager of those brands on how to enhance the engagement of their followers in social media. Another study [139] aims at studying the importance of brands' Fans and the Fans' friend as a key factor in the strategy of three Facebook accounts. However, their study is limited to just three brands and they only considered one metric, the number of fans for each brand. Therefore, the last two references only derive ad-hoc conclusion for very few users that cannot be generalized. We found some larger scale works like [133] where authors manually look to the type of activity of 275 non-profit organization profiles in FB. However, they just look at two elements: how the users disseminate their messages and what type of posts they are considering in their strategies. This work differentiates from our study in three main aspects: they only look at FB, they do not look into professional users, and they only use type of content to evaluate the behaviour of the FB users. In addition, authors in [130] explore the strategic use of social media for 250 of U.S. based companies on Facebook, Twitter, and YouTube. Although this work is more similar to our study due to the analysis of multiple OSNs it presents considerable differences in the methodology and the analyzed behavioural elements. First of all the authors in this study rely on manual inspection of the accounts that is a much more subjective method than a data-driven approach. In addition,

Table 4.18: Dataset description

OSN	#posts	avg(posts)	%cross posts	#like	#comments	#shares
FB	423K	695	33.63	2.9B	98M	235M
G+	175K	304	29.36	27M	5M	3M
TW	1.7M	2648	7.17	274M	-	491M

Table 4.19: Categories in the Dataset with more than 10 users.

#	**category**	**#user**	#	**category**	**#user**
1	Musician_band	134	9	Food_beverages	18
2	Tv_show	40	10	Website	16
3	Public_figure	32	11	Cars	15
4	Media_news_publishing	29	12	Clothing	13
5	Actor_director	28	13	Movie	12
6	Athlete	24	14	News_media_website	12
7	Sports_team	23	15	Tv_network	12
8	Product_service	20	16	Politician	6

they use a number of social metrics (adoption, integration, code of conduct, human voice, dialogic loop, activity and stakeholder willingness) that are not linked to the actual activity of a user in OSN, and again are subjective. In contrast to these previous works relying on manual inspection, we have found a number of works that uses surveys or interview community managers to analyze the strategy of some few brands. The most relevant work is [134] in which the authors interview nine community managers of NBA teams. This study just focus in a single sector and perform a qualitative analysis based on the replies of the community managers. Finally, we also find a couple of studies that attach the success of a social media brand to the popularity [139] and to the number of reactions [138]. However, none of them take into account that both parameters are related and that considering success using them isolated may lead to wrong conclusions.

In summary, the main novelties of our work compared to the previous studies are: (*i*) it is the first data-driven approach over a large number of professional users. (*ii*) It aims at understanding the strategies from a global point of view per sectors. (*iii*) It is a longitudinal study across the three major OSNs: FB, TW and G+. (*iv*) It proposes a quantitative estimation of the success of OSNs strategies.

4.4.3 Dataset

This part briefly describes the dataset of this study, and introduce the way we classify the users into categories. The detail of the data collection methodology and selected users has been presented in section of this manuscript.

Table 4.18 summarizes the datasets used in this study. In total, we analyze more than 2M posts published by 616 professional publishers in FB, TW and G+.

In order to address the main goal of the study we need to assign the 616 users to the categories they are representing. Towards this end, we have used a straightforward approach based on the category each professional user selects when they register their FB page. Therefore we assign each user to the category they have selected in FB. Overall, the 616 users are classified into 62 different categories. The goal of this study is to find whether users in some category present a common behaviour on their utilization of OSNs, describe the strategy in that category and determine its degree of success in FB, TW and G+. We can only perform that analysis for those categories in our dataset that includes enough users. Then, we have decided to study categories represented by at least 10 users in our dataset. Table 4.19 shows the number of users associated to the 15 categories[13] that meet that requirement[14]. We have made an exception for the category *Politician*, which is formed by only 6 users. Although we acknowledge that 6 users must not be enough to generalize the strategy of politicians, we believe it is worthy to study such an interesting category. We believe the 16 categories we are going to analyze present a quite interesting heterogeneity of sectors (e.g., popular individuals, big industrial companies, news agencies, TV or the Internet) that address different audiences.

4.4.4 Detection of Common Strategies by Sectors

The goal of this section is to verify the baseline hypothesis of whether the users of a particular sector present a similar behaviour in their use of OSNs. Then we first introduce the behavioural metrics used to describe the strategy of a user, and later apply the methodology proposed in [135] to discriminate which categories follows our hypothesis.

4.4.4.1 Metrics to Capture the Behaviour

The strategy of a user is defined by the decisions that she takes when posting information across several OSNs. Therefore, the elements we use to define the activity are behavioural metrics directly related to those decisions. Each behavioural metric is captured with one (or more) values in each OSN as it is detailed below. Overall each user is represented with a behavioural vector of 33 values that defines her strategy across FB, TW and G+. We wanted to provide the same weight to all the parameters, hence all the values range between 0 and 1 in the vector. This has led us to normalize one of the metrics, the activity rate. We have performed the normalization using the 90th-percentile[15] of that parameter considering all the users in our dataset. All the users with a value above the

[13]We use News Website instead of News_media_website and Media News instead of Media_news_publishing from now on in the study.

[14]The reader can find the name of the users in each category in [140].

[15]We did not use the maximum since we have checked that usually for the parameters we had to normalize the maximum value was actually an outlier.

90th-percentile was assigned a value equal to 1 in the normalization. Note that we perform the normalization individually for each OSN.

Activity rate: We measure the average posts/day published by the user. As it is reported in [125], OSN users are intrinsically much more active in TW than in FB and G+. Therefore, we are interested on knowing how active is a user in a particular OSN with respect to the activity of other users in that OSN. With the proposed normalization for this metric we achieve that goal. This metric generates 3 values in the behavioural vector, one per OSN.

Fraction of Cross-Posting: We use as metric the portion of cross-posts in each OSN per user (3 values in the vector).

Cross-Posting pattern: This metric captures the volume of common information that a user publishes in more than one OSN. We defined and characterized this phenomenon in a previous study. For more information please check [136]. We use as metric the portion of cross-posts happening in each possible OSN combination, i.e., FB-TW, FB-G+, TW-G+ or FB-TW-G+ (4 values in the vector).

Preference: This element is measured using the portion of cross-posts initiated in each OSN. This metric allows us to establish what is the preference of a user among the evaluated OSNs (3 values in the vector).

Type of content in regular-posts: This metric measures the portion of posts assigned to different type of content from the regular posts published by the user. In the case of FB and G+ the options are: photos, videos, links and text. In the case of TW only text or link. This metric generates 4 values in the vector for FB, one per type of content, 4 values in G+ and 2 Values in TW (10 values in total in the vector).

Type of content in cross-posts: This metric is similar to the previous one but in this case it only considers cross-posts (10 values in the vector).

4.4.4.2 Identifying Categories Whose Users Present a Similar Strategy

We compare the similarity in the strategy of two different users by computing the Euclidean distance between their vectors. Hence, the lower the Euclidean distance the closer the strategies of the two users are. We can apply this process to compute what we refer to as intra-category and inter-category similarity. The former refers to the Euclidean distance between each pair of users within the category, while the latter is represented by the Euclidean distance of each user in the category to all the user outside that category.

We now apply the methodology proposed in [135] to find what are the categories whose users present a similar strategy across FB, TW and G+. First, we measure the intra-category and inter-category cohesion of each category using a Kernel Density Estimation

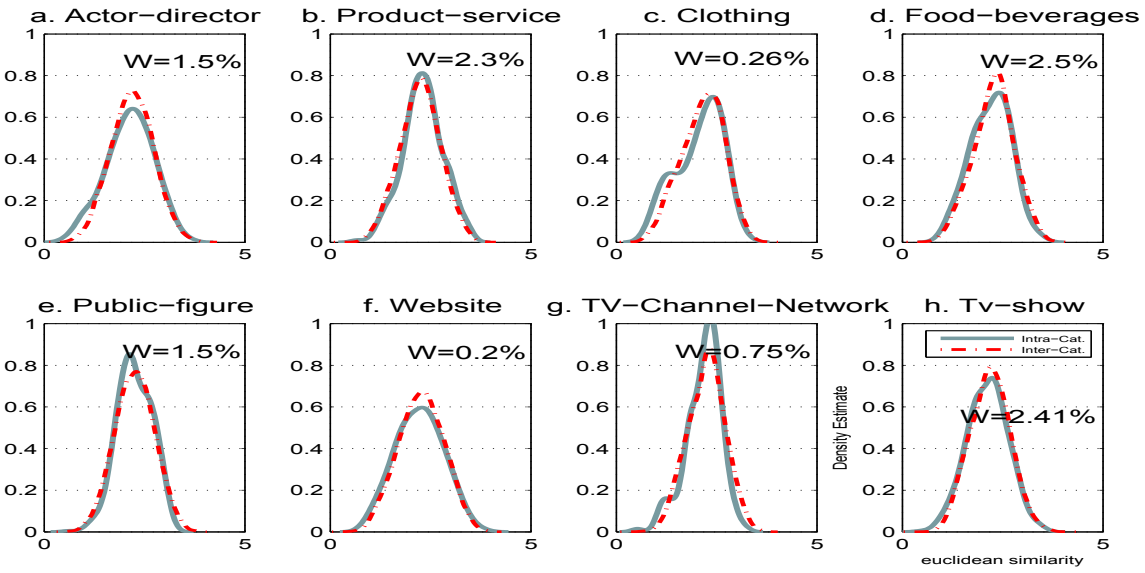


Figure 4.7: Kernel Density Estimation of the intra-category and inter-category Euclidean distance for those categories whose users do not present a common strategy.

(KDE) [141] method which is computed as follows:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)$$

where K is the gaussian kernel and h ¿ 0 is a smoothing parameter called the bandwidth. The cohesion is measured based on the Euclidean distance which is computed for two series of p and q as follows.

$$d(p,q) = \sqrt{\sum_{i=1}^{n}(q_i - p_i)^2}$$

In addition, for each category, we run the Wilcoxon rank-sum test [142] on the distributions of the intra-category and inter-category Euclidean distance. This is a non-parametric test of the null hypothesis that two populations are the same. The Wilcoxon test also provides the parameter W that measures the distance between the median of both distributions. In our analysis $W = Median\_inter - Median\_intra$, thus the larger W is the stronger is the intra-category cohesion. We note that we compute the parameter W as the difference of the medians in percentage (instead of absolute term) that provides more clear insights.

Figure 4.7 shows the KDE results for those categories in which the Euclidean distance among the users inside the category is very similar to the Euclidean distance with external

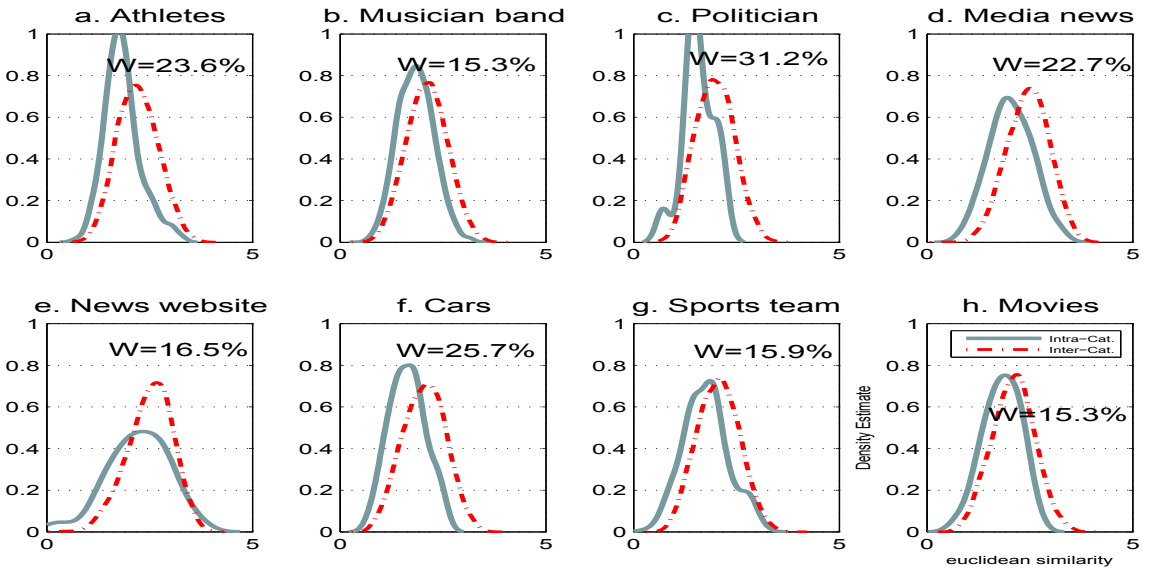


Figure 4.8: Kernel Density Estimation of the intra-category and inter-category Euclidean distance for those categories whose users present a common strategy.

users. This can be easily observed since the distributions are overlapped. Aligned to this result, the Wilcoxon test validates the null-hypothesis in all the cases (i.e., the distributions are the same), and W is below 2.5% in all the cases. Therefore, we conclude that the users in those eight categories do not present a common behaviour.

Contrary, Figure 4.8 depicts the KDE for those categories with a major intra-category cohesion. In this case, the Wilcoxon test rejects the null-hypothesis in all cases. This means that the intra-category and inter-category distributions are statistically different ($p\text{-}value<0.001$) for these eight categories. This statement is supported by the fact that for these categories W ranges between 15% and 30%. Therefore, these results uncover eight categories whose members present common behavioural elements (i.e., strategy) that globally differs from the strategy of the users outside that category. These eight categories are: Athletes, Cars, Media News, Movie, Musician-Band, News Website, Politician and Sport Team.

We note that from now on in the study the strategy of each category will be represented by the centroid[16] of the category.

[16]Each of the 33 values characterizing the centroid corresponds to the median of each metric across the users in the category.

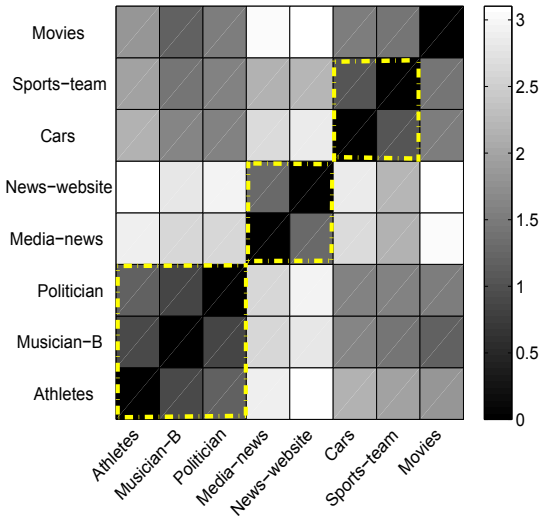


Figure 4.9: Colormap that represents the Euclidean distance between the behaviour of the eight categories with a similar strategy. The closer the strategy of two categories is the darker the cell representing their Euclidean distance. We find three relevant clusters among the analyzed users that are highlighted using a yellow dotted line.

4.4.4.3 Similarity Between Categories' Behaviour

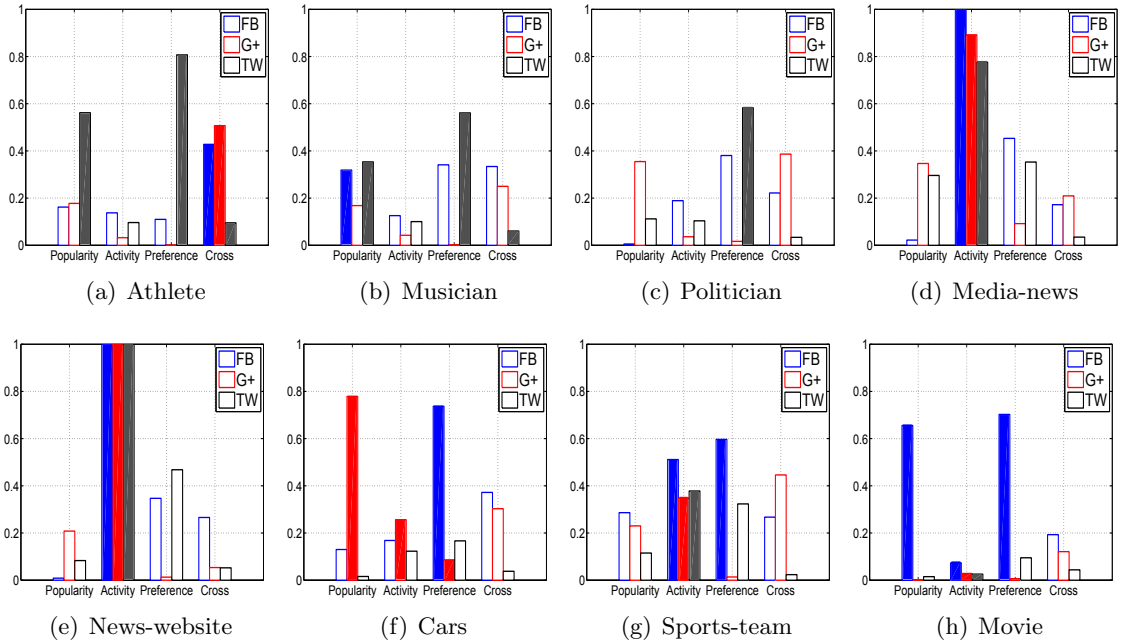
We have demonstrated that there are 8 categories whose users present a similar use of OSNs. However, the previous analysis neither says how close are the strategies of these categories nor defines the main elements of each strategy. In this subsection we address the first point, while the second question is covered in the next section.

To compare the strategies between two categories we calculate the Euclidean distance between their centroid. Figure 4.9[17] shows a colormap in which each cell unveils the Euclidean distance between the centroid of two categories. Visually, the closer the strategy of two categories is the darker the cell is[18].

The results reveals three interesting clusters. First, Media News and News Website have very different strategies to any other category, while they present some commonalties in their use of OSNs. Second, the categories that represent individual users, i.e., Athletes, Music-Band and Politician, present a more similar strategy among them than to other categories. Third, Cars and Sport Teams, the two categories representing companies, present a major similarity to each other than to any other category. Finally, Movie present a strategy that is neither far away nor close to any other category except the two categories referring

[17]We advise the reader to visualize all the figures from this point in the computer to get a better color resolution.

[18]Note that in Figure 4.9 the black diagonal act as a mirror. The results are the same in the upper and lower part of the diagonal since the Euclidean distance between two categories is a bidirectional parameter.



(a) Athlete (b) Musician (c) Politician (d) Media-news

(e) News-website (f) Cars (g) Sports-team (h) Movie

Figure 4.10: Bar plot that shows the value of the following metric for each category and OSN: *popularity*, *activity rate*, *preference* and *fraction of cross-posts*.

to news.

It is important to highlight that the fact that two categories present a higher similarity in their strategy does not mean they present exactly the same behaviour (i.e., the same values in the metrics). Instead, the correct interpretation is that those two categories will present some commonalities in some behavioural elements that make their strategies closer with respect to other categories.

4.4.5 Unveiling Strategies

In this Section we reveal and discuss what are the most significant elements in the strategy of the 8 categories under analysis. Towards this end we use all the behavioural elements introduced in Section 4.4.4.1 except *Cross-Posting pattern* because it is only relevant in the strategy of Cars. The other categories closely follow the general results reported in Section 4.3.4.2 for this metric. In addition to the behavioural parameters, we use the popularity (i.e., number of followers) of each category in each OSN in the analysis. The reason is that although the popularity is not a behavioural element itself, it can influence the decisions of a user. As we did for the activity rate, we have normalized the popularity using the 90th-percentile in each OSN.

Figure 4.10 shows one bar plot per category in which each bar shows the value of

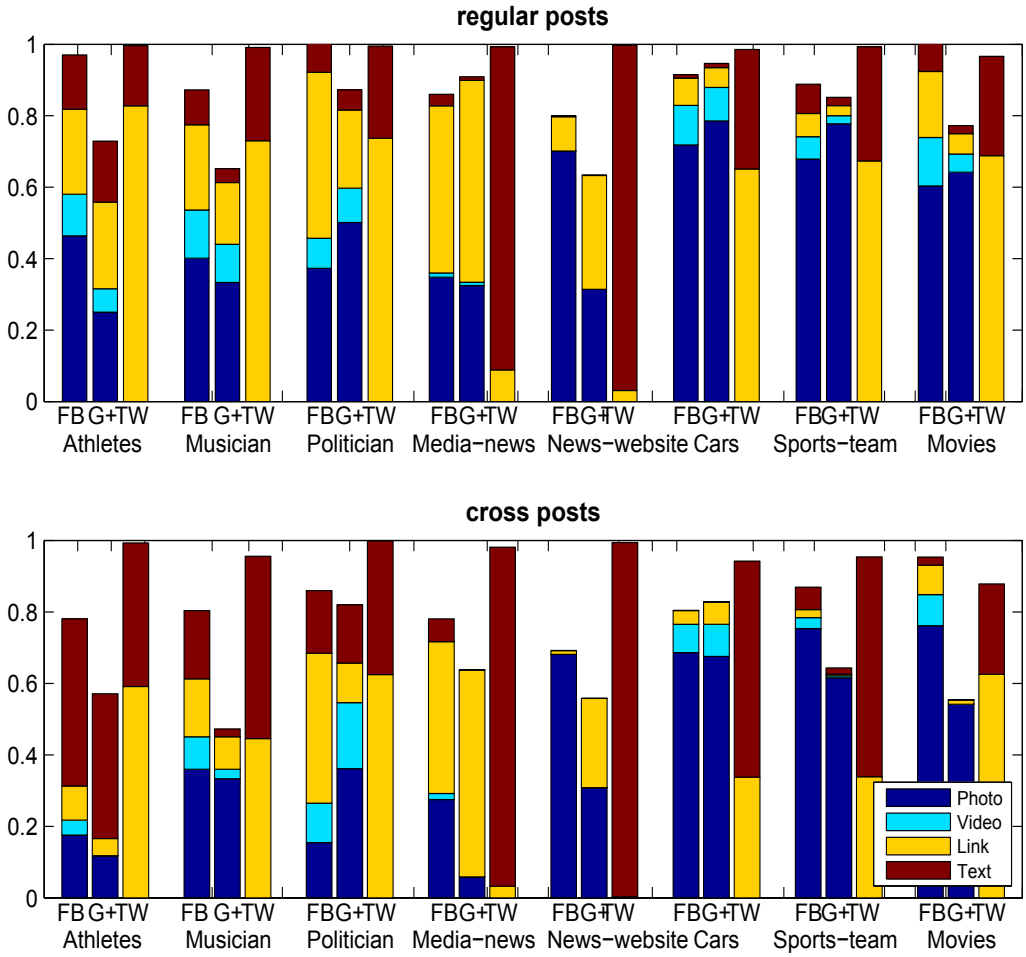


Figure 4.11: Bar plot that shows the type of content published in each category per OSN.

popularity, *activity rate*, *preference* and *fraction of cross-posts* in each OSN, respectively. We have highlighted in full color the bars that represent the most significant elements of the behaviour of each category. In addition, Figure 4.11 shows *type of content in regular-posts* and *type of content in cross-posts* for each category and OSN, respectively. Following, we describe the strategy of each category:

Athlete: It is the category with the strongest preference for TW and with the most intense cross-posting activity in the three OSNs. It presents a low activity in all OSNs compared to other categories. Regular posts are mostly photos and links in FB and G+, however cross-posts are dominated by text in these two OSNs. This is explained because most of the cross-posts are initiated by TW (as shown by the strong TW preference) and replicated in FB and G+ as text. Finally, it is the most popular category in TW, which may explain its strong preference for this OSN.

Musician-Band: This category presents a clear preference for TW and an important level of cross-posting in this OSN (only surpassed by Athletes). The posts published in FB and G+ are mostly audiovisual content, both in cross-posts and regular-posts. The activity rate is low in the three OSNs. Finally, in terms of popularity, Musician-Band is the second most popular category in FB and TW behind Movie and Athlete, respectively.

Politician: Similar to Athlete and Musician-Band this category presents a preference for TW as well as a low activity in all 3 OSNs. The most interesting behavioural element of Politician is that it uses different content in FB and G+. Politician publishes more links in FB than in G+, where it mostly publishes audiovisual content. They also opt for using links in most of the tweets.

Media News: The differential strategy of this category is clearly a very high activity rate in the three OSNs. Actually, this seems reasonable since the users in this category are news agencies, portals, etc that are continuously publishing recent news. In addition, a second particularity of Media News is that the most common type of content in FB and G+ is link. However, it very rarely uses links in TW. In addition, together with News Website, is the category with a more balanced preference between FB and TW.

News Website: As the previous category, the differential behavioural element of News Website is its extraordinary high activity rate in all OSNs. In addition, News Media Website also shows a quite balanced preference between FB and TW. Contrary to Media News, in this case posts in FB are mostly photos, while in G+ they are balanced between photos and links.

Cars: Cars is the category with a major interest in G+, which may be due to its high popularity in that OSN. The behavioural elements that shows that interest are: (*i*) it is the only category in which the selection of G+ as initial source of information is relevant (it happens in almost 10% of the cross-posts), (*ii*) Cars is the only category in which its (relative) activity rate is higher in G+ than in TW and FB, and (*iii*) Cars is the only category in which the cross-posting activity between TW and G+ is not negligible since this pattern appears in 15% of the cross-posts. Apart from its interest in G+, Cars is clearly biased to FB in terms of preference and mostly uses audiovisual content in its posts. This seems reasonable since the business of Cars companies has a lot to do with presenting an attractive view of their cars and this requires the use of audiovisual material.

Sports Team: There are three elements that denote the behaviour of Sport Teams. First, a clear preference for FB. Second, an intense use of photos in its posts. Three, a considerably high activity in the three OSNs compared to the other categories (with the exception of the two categories related to news).

Movie: The behaviour of this category is defined by a strong preference of FB, the use of photos in most of its FB and G+ posts, and the lowest activity rate in the three OSN

among the categories under analysis. This happens because the OSN accounts associated to movies are only active in a short period of time around their release and later they just keep a residual activity. Finally, there is a big contrast in its popularity since it is the most popular category in FB, but the least popular in TW and G+.

We conclude our analysis by enumerating the common behavioural aspects for the three clusters identified in Section 4.4.4.3. (1) All the *individual* users present a preference for TW and a relatively low activity in all OSNs compared to other categories. (2) Cars and Sports Teams, which represent *commercial companies*, shows a clear preference for FB and mostly post audiovisual content in FB and G+. (3) The categories related to news reporting coincides in having a very high activity rate.

4.4.5.1 Evaluation of Strategies Success

To conclude this study we want to assess the success of the strategies adopted by the analyzed categories. To the best of our knowledge it does not exist any standard metric or methodology to evaluate the success of an strategy in OSNs. Our approach is based on the conviction that the number of reactions that a user attracts in her posts is the only objective available metric to capture the interest/engagement of end-users in the activity of a professional user. Therefore, in this study we propose to measure the success of the strategy of a category as a function of the average number of reactions that the category attracts per post. We believe that the proposed methodology is a useful tool to estimate the success of a particular strategy in the context of this study. However, we do not pretend to present it as a reference methodology to globally evaluate success in OSNs. Following, we first introduce our methodology and later we discuss the results extracted from applying it.

Table 4.20: Pearson coefficient, p-value, and Regression Coefficient of the correlation between popularity and reactions.

Reaction	PPMC	p-value	Regression Coefficient
FB likes	0.97	6e-5	1.78e-3
FB comments	0.94	4e-4	4.92e-5
FB shares	0.94	4e-4	1.14e-4
G+ +1s	0.76	0.03	7.02e-5
G+ comments	0.14	0.73	-
G+ reshares	0.94	5e-4	8.11e-6
TW favourite	0.78	0.02	2.07e-5
TW retweet	0.71	0.049	5.04e-5

4.4.5.2 Methodology to Measure the Success Degree of Strategies

Our methodology proposes to compute the success of the strategy of a category as the difference between the expected number of reactions per post that category should receive

and the actual number of reactions it receives. Therefore, our goal is to propose a model that estimates the expected volume of reactions per post for the eight categories under discussion.

Our intuition is that the number of reactions that a user attracts in a post in an OSN is strongly correlated to her popularity in that OSN. Therefore, our first step is to validate this hypothesis that would allow us to formulate the expected number of reactions as a function of the popularity.

We calculate the Pearson Product-Moment Correlation Coefficient (PPMCC) between the popularity and all the reaction types separately. The PPMCC measures the degree of linear dependence between two variables and calculates as follows

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

which becomes higher as the PPMCC moves to 1.

Table 4.20 shows the PPMCC and p-value for the correlation associated to each reaction type. The results reveal a very strong linear positive correlation between popularity and volume of reactions per post for all type of reaction in all OSNs (PPMCC>0.7 and p-value<0.05). There is only one exception, G+ comments (p-value>0.05), which are omitted from our analysis in the rest of the section.

Based on these results, we propose a simple linear model that estimates the number of reactions a category should receive based on its popularity. Hence, we perform a linear regression to obtain the regression coefficient, listed in Table 4.20, associated to each type of reaction. In a nutshell, we estimate the number of reactions per post for a particular type of reaction in a category multiplying the popularity of that category by the regression coefficient for that reaction type.

Once we have the model to estimate the expected number of reactions we are able to evaluate the success of the different strategies. Figure 4.12 shows a colormap that represents the level of success of each category for each type of reaction. The colormap shows a positive (associated to green color) and negative (associated to red color) scale. For instance, a value of +2 implies that the category under analysis is obtaining 2× more reaction per post than what our model suggests. In contrast, a value of -2 indicates that the category is attracting 1/2 of the expected reactions per post. Note that the darker is the green color in a cell the higher is the success. Similarly, the darker is the red color in a cell the less efficient the strategy is. Each row corresponds to one category and presents a visual overview of the success of its strategy across the different OSNs and types of reactions.

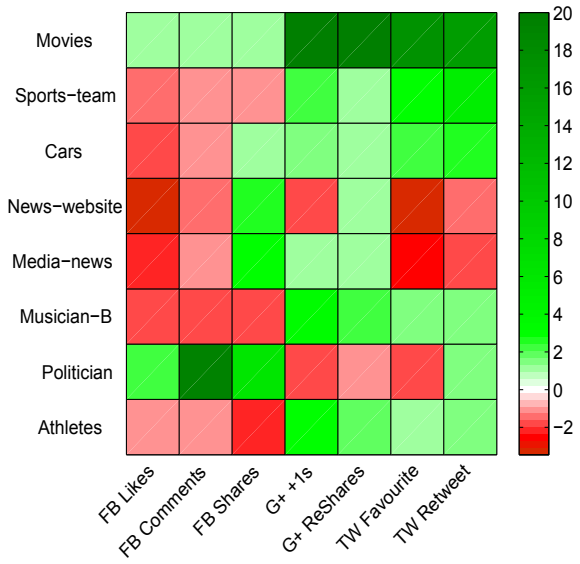


Figure 4.12: Colormap that represents the success of the strategy of each category across different types of reaction. The green color represents success and the red color represents failure.

4.4.5.3 Discussion of Strategies' Success

Movies is the only category with a successful strategy in all OSNs according to the volume of reactions it receives per post. This is an indicator that the adopted strategy is well adapted to the requirements of its audience in each OSN.

Athletes and Musician-Band are successful in TW and G+, but they fail in FB. Based on their clear preference for TW, it seems its strategy is adequate to cover their main objective, however they should modify their behaviour in FB in order to increase the engagement of end-users.

Politician has a successful strategy in FB, especially on attracting comments, but it fails in G+. In the case of TW it manages to get more retweets than expected, but does not cover the expectation in number of favourites. Its strategy is fair enough in FB to cover the expected reactions. In the case of TW, if its major interest focuses on spreading tweets its strategy is also adequate.

It seems that the interest of Cars in G+ is obtaining its reward since it manages to attract more reactions than the estimation of our model. In contrast, it seems Cars should revise their behaviour in FB since it only succeeds on the number of shares, even though it has a strong preference for this OSN.

Sports Team fails in FB, but is successful in TW and G+. Therefore, it should change some behavioural aspects to increase their engagement in FB.

Finally, Media News and News Website categories present a quite similar success pattern with the exception of G+ likes. We believe the most important type of reaction for news agencies and portals is share, reshare and retweet, since their goal is to spread the reported news as much as possible. For these reactions they present an almost identical result that reflects a success in FB and G+, but a failure in TW. This is a quite negative outcome since TW is considered a very relevant communication channel to disseminate news nowadays.

4.4.6 Conclusions

This study advances the state of the regarding the strategy used by professional users in OSNs in three main elements: (*i*) To the best of our knowledge this is the first study that follows a data-driven approach to analyze the strategy of professional users in OSNs. (*ii*) We evaluate the global strategy of some professional sectors in the three major OSNs, namely FB, TW and G+. In contrast, most previous work focuses in the analysis of individual users and obtain adhoc conclusions. (*iii*) To the best of our knowledge, this study is the first study that proposes a quantitative estimation of the success of a strategy. In order to be able to make an analysis per sector, our first step has been to demonstrate that there are sectors whose users present similar behavioural elements that define a common strategy in OSNs. In particular, we have found eight sectors with a common strategy: Athletes, Cars, Media News, Movie, Musician-Band, News Website, Politician, and Sports Teams. The more interesting findings for the analyzed sectors are: (*i*) the two categories related to news show an extremely intense activity in the three OSNs; (*ii*) Athlete shows a strong preference for TW that directly impacts the information published in FB and G+; (*iii*) Cars gives a high value to G+ where they have a much stronger presence than any other category, and, (*iv*) Movie is very active around the release of the film but later the activity becomes residual. Finally, we estimate the success of each strategy. The success is measured as the difference between the actual volume of engagement (i.e., reactions per post) and the expected volume of engagement based on the popularity of the category. Movie is the only category that overpasses the engagement expectation in all OSNs. Politician is the only category, in addition to Movie, with a clear success in FB, but it is the only category that does not reach the expectation in G+. Finally, the news-related categories are the only ones that do not reach the expected engagement in TW, neither in retweets nor in favourites. In addition to all the previous findings, this work presents an aside contribution that characterizes the cross-posting phenomenon for professional users across FB, TW and G+. We have demonstrated that this phenomenon exists and is relevant. The dominant cross-posting pattern is FB-TW, while it is very rare finding information shared between TW and G+.

Chapter 5

Conclusion and Future Work

Contents

5.1 Conclusion

5.1.1 Summary of Contributions

Appearance of new and very popular Internet services had a huge impact of people life specially the way that they communicate with each other. Online Social Networks (OSNs) and Peer-to-Peer (P2P) systems are two examples of this type of services which have engaged a huge number of different type of customers both regular and professional users. In this thesis we have characterized different aspects of regular and professional users in term of their publishing strategies and content consumptions in 5 different studies.

- **Disclosed Information on Facebook:** The main results of this study answer the question that how much information is available on facebook users profile. The main insights from this study are as follow: (*i*) Friend-list is the attribute with the largest public exposure, whereas Birthday attribute is the one showing major privacy concerns from Facebook users. (*ii*) There is a strong correlations between *Current City* and *Home Town* attributes as well as (i.e. *College* and *HighSchool*) and professional (i.e. *Employers*) attributes. (*iii*) In average we found that Facebook users has more than 4 attributes publicly available in their profiles. (*iv*) In general Men has larger public exposure as compared to women for all personal attributes except *birthday* in their profiles. (*v*) The age range of 18-25 (half of the dataset) is the group of users with the most public available information in their profile. (*vi*) Our study shows that, Facebook data can be utilized for estimating the portion of people living in different class of cities.

- **Cross Posting Activity:** The main outcomes of the study are as follow: (*i*) we have demonstrated that professional users frequently publish the same information in at least two OSNs, especially in the case of FB and G+. Although professional users in TW present a low portion of cross-posts, the fact that they are very active implies that in absolute terms we can find quite a lot cross-posts in their TW accounts. (*ii*) We saw that a professional user publishes (in median) 70% of her cross-posts exclusively in FB and TW, and around 15% in FB and G+. Furthermore, we demonstrated that the cross-posting activity between TW and G+ is negligible. (*iii*) Cross-posting in TW and FB accounts has benefit for professional users in term of attracting people engagement. They attract 2× and 30% more engagement with cross-posts than non-cross-posts, respectively. However, cross-posts in G+ leads to halve the engagement as compared to non-cross-posts. (*iv*) As initial source of information, professional users equally prefer FB and TW, but they rarely choose G+. (*v*) Users with a strong preference for TW present cross-post with a higher similarity (across different OSNs),

publish more textual content than photos and videos, and use to include links to websites different than major OSNs. In contrast, users preferring FB publish mainly audiovisual content and a major portion of urls in their cross-posts refer to OSN content. (*vi*) As the user inter-posting interval time decreases: (i) the similarity of her cross-posts increases, (ii) the portion of audiovisual content attached to her cross-posts decreases as wall, (iii) and a larger portion of urls included in her cross-posts refers to major OSNs sites.

- **Professional Users' Strategies in OSNs:** The main findings of this study on different categories of professional users is: (*i*) The two categories related to news show an extremely intense activity in the three OSNs; (*ii*) Athlete category shows a strong preference for TW that directly impacts the information published in FB and G+; (*iii*) Cars category gives a high value to G+ where they have a much stronger presence than any other category, and, (*iv*) Movie category is very active around the release of the film but later the activity becomes residual.

- **Multimedia Evolution on P2P:** The main outcome of this study on the evolution of multimedia content can be summarized in three main findings: (*i*) Multimedia content has doubled its size in a period of only 2 years. (*ii*) The major part (80%) of the consumed multimedia content corresponds to TV Shows and Movies (including porn) that belong to those categories with a largest size. (*iii*) High-resolution content, which has very large size, is increasing its presence and it already represented 8% of the available content and 10% of the downloads in our most recent snapshot dated at the beginning of 2012.

- **Reaction to Antipiracy Actions:** Two main finding on this study on the reaction of professional publishers on two major antipiracy actions is as follow: (*i*) the Megaupload closure triggered an immediate drop in the activity of professional BitTorrent publishers that are running their own private BitTorrent portals. Furthermore, a group of casual publishers also migrated to BitTorrent most likely from Megaupload and other Cyberlockers. (*ii*) The French Hadopi law had an effect on the number of casual BitTorrent publishers and reduces it. However, it did not have any impact on the activity of professional publishers from France due to a particular hosting facility with passive monitoring policy for copyright infringement activity.

In summary the main contributions of this thesis can be summarized in four parts as follow:

- **outcomes of the studies**, Five studies has been presented in this thesis which

characterize some behavioral aspects of regular and professional users over major OSNs and P2P systems.

- **Measurement Methodologies**, We proposed four measurement methodologies to evaluate the behavior of regular and professional users in OSNs and BitTorrent.
- **Implemented Measurement Tools**, Seven advanced data collection tools have been implemented during this thesis to collect users information from Facebook and BitTorrent.
- **Available Datasets**, Four of the collected datasets are available for further research collaborations which include Facebook professional and regular users information.

More details about the contributions of this thesis is available in Section 1.2.

5.1.2 Impact on the Community

Some of our conducted researches had media impacts and several news articles were published about our results in some major online news agencies. Table 5.1 summarize some of the published articles.

5.2 Ongoing and future work

This section summarize some of the ongoing and future research directions on this research.

5.2.1 Ongoing and future work on Social Network

5.2.1.1 Characterizing Group-Level User Behavior in Major Online Social Networks

In this study, we conduct a detailed measurement study to characterize and compare the group-level behavior of users in Facebook, Twitter and Google+. We focus on Popular, Cross (with account in three OSNs) and Random group of users in each OSN since they oer complementary views. We capture user behavior with the following metrics: user connectivity, user activity and user reactions. Our group level methodology enables us to capture major trends in the behavior of small but important groups of users, and to conduct inter- and intra-OSN comparison of user behavior. Furthermore, we conduct temporal analysis on different aspects of user behavior for all groups over a two-year period. Our analysis leads to a set of useful insights including: (i) The more likely reaction by Facebook and Google+ users is to express their opinion whereas TW users tend to relay a received post to other users and thus facilitate its propagation. Despite the culture of reshare among

Table 5.1: Published articles about our studies

Source	Article name/link
Original press	actu-des-tic.telecom-sudparis.eu/2014/01/les-editeurs-professionnels-alimentent-les-systemes-p2p-mondiaux-depuis-la-france/ www.mines-telecom.fr/es/estudio-demuestra-que-editores-profesionales-alimentan-los-sistemas-p2p-mundiales-desde-francia/ www.mines-telecom.fr/wp-content/uploads/2014/01/20141301_CPPubliIEEE_P2P_ES.pdf www.mines-telecom.fr/une-etude-montre-que-des-editeurs-professionnels-alimentent-les-systemes-p2p-mondiaux-depuis-la-france/ www.mines-telecom.fr/en/a-study-has-revealed-that-professional-publishers-feed-global-p2p-systems-from-within-france/ www.mines-telecom.fr/wp-content/uploads/2014/01/20141301_CPPubliIEEE_P2P_janv2014.pdf
PC INpact	www.pcinpact.com/news/85319-selon-etude-hadopi-a-fait-diminuer-nombre-d-uploadeurs-en-france.htm ?utm_source=PCi_RSS_Feed&utm_medium=news&utm_campaign=pcinpact
ZDNet	www.zdnet.fr/actualites/p2p-la-france-telecharge-moins-mais-alimente-le-monde-39797025.htm
Info Utiles	www.info-utiles.fr/modules/news/article.php?storyid=9754
Tropgeek	www.tropgeek.com/news/Un+bilan+de+l%E2%80%99Hadopi,+le+nombre+de+contenus+ill %C3%A9gaux+mis+en+ligne+depuis+la+France+a+augment%C3%A9+de+18+%25
Numerama	www.numerama.com/magazine/28045-une-etude-sur-hadopi-pointe-du-doigt-ovh-et-son-laxisme.html
Ginjfo	www.ginjfo.com/actualites/politique-et-economie/un-bilan-de-lhadopi-le-nombre-de-contenus- illegaux-mis-en-ligne-depuis-la-france-augmente-de-18-20140110
Forumdupirate Softonic	forum.journaldupirate.com/viewtopic.php?f=55&t=8053 actualites.softonic.fr/2014-01-14-p2p-la-loi-hadopi-est-elle-efficace
Wingwit	fr.wingwit.com/?p=760
IT Channel	www.itchannel.info/index.php/articles/145867/hadopi-riposte-graduee-mais-incomplete.html
ITR Mobiles	www.itrmobiles.com/index.php/articles/145867/hadopi-riposte-graduee-mais-incomplete.html
ITR Manager	www.itrmanager.com/articles/145867/hadopi-riposte-graduee-mais-incomplete.html
ITR News	www.itrnews.com/articles/145867/hadopi-riposte-graduee-mais-incomplete.html
La Vie Numerique	www.lavienumerique.com/articles/145867/hadopi-riposte-graduee-mais-incomplete.html
Actualite24h	www.actualite24h.com/actualites/p2p-la-france-telecharge-moins-alimente-le-monde
Scoop.it!	www.scoop.it/t/free-mobile-orange-sfr-et-bouygues-telecom/p/4014115874/2014/01/13/p2p-la-france-telecharge-moins-mais-alimente-le-monde
Le Journal du Geek	www.journaldugeek.com/2014/01/19/une-etude-montre-que-des-editeurs-professionnels-alimentent-les-systemes-p2p-mondiaux-depuis-la-france/
WebRadar	webradar.me/68827862
Direct Matin - Lille plus	www.directlille.com/pdf/lil_lilleplus_04_02_14.pdf
OVH	www.ovh.com/fr/a1326.reponse-ovh-etude-hadopi
Themediashaker	www.themediashaker.com/en/list-of-articles/more-pirated-files-but-fewer-pirates/
UC3M	netcom.it.uc3m.es/news/20140128-News-Publishers_Feed_P2P_Systems-EN
Madrimasd	www.madrimasd.org/blogs/sociedadinformacion/2014/01/28/132210
IMDEA	www.networks.imdea.org/whats-new/news/2014/study-has-revealed-professional-publishers-feed-global-p2p-systems-within-france

Twitter users, a post by a Popular Facebook user receives more Reshares than a post by a Popular Twitter user. (ii) Added features in an OSN can significantly boost the rate of action and reaction among its users.

5.2.1.2 Facebook Network Architecture Analysis, How Far is Facebook from Me!

In this work we are going to investigate the architecture of Facebook in term of its physical server location around the world including Akamai servers (is a CDN network that serve most of Facebook services). To this end we study the accessibility of around 50 service (server url) from 500 Planetlab node distributed around the world by sending ping and tracerout commands six times per day. By this results we can see how is the reachability to the Facebook services from different location across the world. Here we are going to do study in the country and continent level to see what category of countries have better access to what type of the services.

5.2.1.3 Community Similarity Degree: Community Selection for Community Recommendation

In this study we present the *Community Similarity Degree (CSD)* that is a metric to compute the degree of similarity among the users within a community. To evaluate the utility of our metric we rely on a dataset that includes more than 200K Facebook users. Using this dataset we define four different types of community: Friend-based (group of friends of a user), Interest-based (group of users sharing a common interest), Location-based (group of users from a city) and Random-based (group of users selected at random). We use the CSD to quantify users' similarity based on the interests they share within a community for five well-defined Facebook profile attributes: television, books, music, movies and games. Surprisingly, our results reveal that Interest-based communities are the ones showing a larger similarity degree, with a CSD between 1.5× and 4.5× larger than Friend-based communities. We use this outcome to demonstrate that communities with a larger similarity degree increase the efficiency of recommendation systems. We have emulated an OSN recommendation system in which an Interest-based recommendation strategy outperforms in 52% the efficiency shown by a Friend-based recommendation approach. This result demonstrates the practical usefulness of the proposed CSD metric.

5.2.1.4 Real life Change effect on the Facebook Profile Attributes information

In this work we aim at studying the effect of major changes in real life of Facebook users on their profile attributes information. For example if a user changes her residence city or job in her real life what is the change effect in the profile attributes such as number of friends, number of interests or even the privacy policy of attributes. for this work we have collected around 73K users profile information in two time window with 8 month duration. By this data we can compare the changes in the profile of each user to see how the attributes evolute in different aspects.

5.2.1.5 Popularity Trend Analysis of Professional Users in Facebook

In this study, we are analysing the popularity evolution of professional users in Facebook. To this end we collected the popularity metrics (#Fans and #Talking) of 10k top Facebook Fan pages 6 times per day for a duration of 18 months. The goal is to understand how the popularity is changing over time in overall, category based and user based. The expected results will show us different cluster of users that has a similar pattern in term of their popularity trends. As an example, we already found a group of professional users that have a clear jump in their popularity due to some impact of their OSN involvement strategies. In the other hand we found users that have a substantial decrease in their popularity which

shows they are loosing their followers in time. We are trying to come up with a model to predict the popularity trend of a user based on her historical pattern and activities. This model also can provide users some hints to improve their strategies on attracting new fans.

5.2.2 Ongoing and future work on P2P networks

5.2.2.1 Film Factory Losses: Is BitTorrent a Major Responsible?

This work is actually a subset of a bigger work that targets the activity monitoring of around 1 million torrents in 18 main trackers. Here for each tracker we collect 4 times per day its torrents information in scrape mode. Later by using TorrentZ portal we identifies the categories of torrent as shown in table 5.2. Later just for movie category that includes 241K torrents we study deeply and identified the movies by using IMDB portal. Here we study movies' losses by investigating the number of download and businesses information of movies like their budget and incomes to see what is the effect of bittorrent on this industry.

Table 5.2: Torrents Category (based on torrentZ portal categories)

Category	**# Torrents**	**%**
Audio	92 321	9,97
Ebooks	27 747	3
Games	30 238	3,26
Movie	241 129	26,03
None	240 652	25,98
Pictures	8 931	0,96
Porn	1 078	0,12
Software	43 369	4,68
Video	240 956	26,01
Total	**926 421**	-

References

[1] "Alexa." [Online]. Available: http://www.alexa.com

[2] J. Ugander, B. Karrer, L. Backstrom, and C. Marlow, "The Anatomy of the Facebook Social Graph," *CoRR*, vol. abs/1111.4503, 2011.

[3] L. Backstrom, P. Boldi, M. Rosa, J. Ugander, and S. Vigna, "Four Degrees of Separation," *CoRR*, vol. abs/1111.4570, 2011.

[4] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, "Walking in facebook: A case study of unbiased sampling of osns," in *IEEE INFOCOM*, 2010.

[5] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?" in *WWW*, 2010.

[6] M. Cha, H. Haddadi, F. Benevenuto, and K. Gummadi, "Measuring user influence in twitter: The million follower fallacy," in *AAAI ICWSM*, 2010.

[7] G. Magno, G. Comarela, D. Saez-Trumper, M. Cha, and V. Almeida, "New kid on the block: Exploring the google+ social graph," in *ACM IMC*, 2012.

[8] D. Schiöberg, F. Schneider, H. Schiöberg, S. Schmid, S. Uhlig, and A. Feldmann, "Tracing the birth of an osn: Social graph and profile analysis in google," in *ACM WebSci*, 2012.

[9] R. Gonzalez, R. Cuevas, R. Motamedi, R. Rejaie, and A. Cuevas, "Google+ or google-?: dissecting the evolution of the new osn in its first year," in *WWW*, 2013.

[10] A. Mislove, M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee, "Measurement and Analysis of Online Social Networks," in *ACM IMC*, 2007.

[11] R. Kumar, J. Novak, and A. Andtomkins, "Structure and evolution of online social networks," in *ACM KDD*, 2006.

[12] A. Mislove, H. Koppula, K. Gummadi, P. Druschel, and B. Bhattacharjee, "Growth of the flickr social network," in *WOSN*, 2008.

[13] Y. Ahn, S. Han, H. Kwak, S. Moon, and H. Jeong, "Analysis of topological characteristics of huge online social networking services," in *WWW*, 2007.

[14] X. Zhao, A. Sala, C. Wilson, X. Wang, S. Gaito, H. Zheng, and B. Zhao, "Multi-scale dynamics in a massive online social network," in *ACM IMC*, 2012.

[15] S. Gaito, M. Zignani, G. Rossi, A. Sala, X. Wang, H. Zheng, and B. Zhao, "On the bursty evolution of online social networks," in *ACM KDD HotSocial Workshop*, 2012.

[16] S. Garg, T. Gupta, N. Carlsson, and A. Mahanti, "Evolution of an online social aggregation network: an empirical study," in *ACM IMC*, 2009.

[17] R. Rejaie, M. Torkjazi, M. Valafar, and W. Willinger, "Sizing Up Online Social Networks," *IEEE Network*, 2010.

[18] J. Jiang, C. Wilson, X. Wang, P. Huang, W. Sha, Y. Dai, and B. Y. Zhao, "Understanding Latent Interactions in Online Social Networks," in *ACM IMC*, 2010.

[19] A. Boutet, A. Kermarrec, E. Le Merrer, and A. Van Kempen, "On the impact of users availability in osns," in *ACM SNS*, 2012.

[20] D. Quercia, D. B. L. Casas, J. P. Pesce, D. Stillwell, M. Kosinski, V. Almeida, and J. Crowcroft, "Facebook and privacy: The balancing act of personality, gender, and relationship currency," in *ICWSM*, 2012.

[21] J. Chang, I. Rosenn, and L. Backstrom, "epluribus: Ethnicity on social networks," in *ICWSM*, 2010.

[22] R. Dey, C. Tang, K. Ross, and N. Saxena, "Estimating age privacy leakage in online social networks," in *INFOCOM, IEEE*, march 2012.

[23] 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*, ser. WPES '05, 2005, pp. 71–80.

[24] R. Farahbakhsh, X. Han, A. Cuevas, and N. Crespi, "Analysis of publicly disclosed information in facebook profiles," in *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ser. ASONAM '13, 2013.

[25] K. Lewis, J. Kaufman, and N. Christakis, "The taste for privacy: An analysis of college student privacy settings in an online social network," in *Journal of Computer-Mediated Communication*, 2008.

[26] A. Nosko, E. Wood, and S. Molema, "All about me: Disclosure in online social networking profiles: The case of facebook," *Computers in Human Behavior*, vol. 26, no. 3, pp. 406–418, 2010.

[27] A. L. Young and A. Quan-Haase, "Information revelation and internet privacy concerns on social network sites: a case study of facebook," in *Proceedings of the 4th international conference on Communities and technologies*, 2009, pp. 265–274.

[28] N. Lampe Cliff A.C., Ellison and C. Steinfield, "A familiar face(book): profile elements as signals in an online social network," in *SIGCHI, Conference on Human Factors in Computing Systems*, pp. 435–444.

[29] X. Han, L. Wang, S. Park, A. Cuevas, and N. Crespi, "Alike people, alike interests? a large-scale study on interest similarity in social networks," in *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*, Aug 2014.

[30] A. Mislove, B. Viswanath, K. P. Gummadi, and P. Druschel, "You are who you know: inferring user profiles in online social networks," in *WSDM '10*, 2010, pp. 251–260.

[31] F. Benevenuto, T. Rodrigues, M. Cha, and V. Almeida, "Characterizing user behavior in online social networks," in *ACM IMC*, 2009.

[32] F. Schneider, A. Feldmann, B. Krishnamurthy, and W. Willinger, "Understanding online social network usage from a network perspective," in *ACM IMC*, 2009.

[33] Z. Xu, Y. Zhang, Y. Wu, and Q. Yang, "Modeling user posting behavior on social media," in *ACM SIGIR*, 2012.

[34] L. Gyarmati and T. Trinh, "Measuring user behavior in online social networks," *Network, IEEE*, 2010.

[35] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, no. 24, pp. 8788–8790, 2014. [Online]. Available: http://www.pnas.org/content/111/24/8788.abstract

[36] J. J. J. A. D. I. K. C. M. J. E. S. J. H. F. Robert M. Bond, Christopher J. Fariss, "A 61-million-person experiment in social influence and political mobilization," *Nature*, vol. 489, p. 295298.

[37] P. Borgnat, G. Dewaele, K. Fukuda, P. Abry, and K. Cho, "Seven years and one day: Sketching the evolution of internet traffic," in *INFOCOM*, 2009, pp. 711–719.

[38] C. Labovitz, S. Iekel-Johnson, D. McPherson, J. Oberheide, and F. Jahanian, "Internet inter-domain traffic," in *SIGCOMM '10*, 2010.

[39] J. S. Otto, M. A. Sánchez, D. R. Choffnes, F. E. Bustamante, and G. Siganos, "On blind mice and the elephant: understanding the network impact of a large distributed system," in *SIGCOMM '11 conference*, 2011.

[40] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang, "Measurements, analysis, and modeling of bittorrent-like systems." in *Proc. of ACM IMC'05*.

[41] A. Legout, N. Liogkas, E. Kohler, and L. Zhang, "Clustering and sharing incentives in bittorrent systems," in *Proc. of ACM SIGMETRICS '07*.

[42] S. Kaune, R. Cuevas, G. Tyson, A. Mauthe, C. Guerrero, and R. Steinmetz, "Unraveling BitTorrent's File Unavailability: Measurements, Analysis," in *IEEE P2P'10*, 2010.

[43] M. Kryczka, R. Cuevas, C. Guerrero, A. Azcorra, and A. Cuevas, "Measuring the bittorrent ecosystem: Techniques, tips, and tricks," *Communications Magazine, IEEE*, vol. 49, no. 9, pp. 144–152, September 2011.

[44] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani, "Do incentives build robustness in BitTorrent?" in *Proc. of NSDI'07*, 2007.

[45] N. Laoutaris, D. Carra, and P. Michiardi, "Uplink allocation beyond choke/unchoke or how to divide and conquer best," in *Proc. of ACM CoNEXT'08*.

[46] J. Pouwelse, P. Garbacki, D. Epema, and H. Sips, "The BitTorrent P2P file-sharing system: Measurements and analysis," in *Proc. of IPTPS'05*.

[47] C. Zhang, P. Dhungel, D. Wu, and K. Ross, "Unraveling the bittorrent ecosystem," *IEEE Transactions on Parallel and Distributed Systems*, 2010.

[48] C. Zhang, P. Dhungel, D. Wu, Z. Liu, and K. W. Ross, "Bittorrent darknets," in *Proceedings of the 29th conference on Information communications*, ser. INFOCOM'10, 2010.

[49] M. Sirivianos, J. Han, P. Rex, and C. Yang, "Free-riding in bittorrent networks with the large view exploit," in *In IPTPS '07*, 2007.

[50] S. L. Blond, A. Legout, F. Lefessant, W. Dabbous, and M. A. Kaafar, "Spying the world from your laptop," *LEET'10*, 2010.

[51] D. R. Choffnes, J. Duch, D. Malmgren, R. Guimera, F. E. Bustamante, and L. Amaral, "Strange bedfellows: Communities in bittorrent," *IPTPS'10*, 2010.

[52] R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, and R. Rejaie, "Is content publishing in bittorrent altruistic or profit-driven?" in *CoNEXT '10*, 2010.

[53] R. Dey, Y. Ding, and K. W. Ross, "Profiling high-school students with facebook: How online privacy laws can actually increase minors' risk," in *Proceedings of the 2013 Conference on Internet Measurement Conference*, ser. IMC '13, 2013.

[54] A. Nazir, S. Raza, and C.-N. Chuah, "Unveiling facebook: A measurement study of social network based applications," in *Proceedings of the 8th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '08.

[55] M. Gjoka, M. Kurant, C. Butts, and A. Markopoulou, "Practical recommendations on crawling online social networks," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 9, pp. 1872–1892, oct. 2011.

[56] C. Wilson, B. Boe, A. Sala, K. Puttaswamy, and B. Zhao, "User interactions in social networks and their implications," in *ACM Eurosys*, 2009.

[57] M. Torkjazi, R. Rejaie, and W. Willinger, "Hot today, gone tomorrow: On the migration of myspace users," in *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, ser. WOSN '09, 2009.

[58] M. Valafar, R. Rejaie, and W. Willinger, "Beyond friendship graphs: A study of user interactions in flickr," in *Proceedings of the 2Nd ACM Workshop on Online Social Networks*, ser. WOSN '09, 2009.

[59] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I tube, you tube, everybody tubes: Analyzing the world's largest user generated content video system," in *Proceedings of the 7th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '07, 2007.

[60] I. Foudalis, K. Jain, C. Papadimitriou, and M. Sideri, "Modeling social networks through user background and behavior," in *Proceedings of the 8th International Conference on Algorithms and Models for the Web Graph*, ser. WAW'11, 2011.

[61] K. Punera and S. Merugu, "The anatomy of a click: Modeling user behavior on web information systems," in *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10, 2010.

[62] N. Archak, V. S. Mirrokni, and S. Muthukrishnan, "Mining advertiser-specific user behavior using adfactors," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10, 2010.

[63] R. Xia and J. Muppala, "A survey of bittorrent performance," *Communications Surveys Tutorials, IEEE*, vol. 12, no. 2, pp. 140–158, Second 2010.

[64] M. Kryczka, R. Cuevas, C. Guerrero, and A. Azcorra, "Unrevealing the structure of live bittorrent swarms: Methodology and analysis," in *Peer-to-Peer Computing (P2P), 2011 IEEE International Conference on*, Aug 2011, pp. 230–239.

[65] M. Yoshida and A. Nakao, "Measuring bittorrent swarms beyond reach," in *Peer-to-Peer Computing (P2P), 2011 IEEE International Conference on*, Aug 2011, pp. 220–229.

[66] M. Varvello, M. Steiner, and K. Laevens, "Understanding bittorrent: A reality check from the isps perspective," *Computer Networks*, vol. 56, no. 3, pp. 1054 – 1065, 2012, (1) Complex Dynamic Networks (2) {P2P} Network Measurement.

[67] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang, "Measurements, analysis, and modeling of bittorrent-like systems," in *Proceedings of the 5th ACM SIGCOMM Conference on Internet Measurement*, ser. IMC '05, 2005.

[68] A. Mahanti, C. Williamson, N. Carlsson, M. Arlitt, and A. Mahanti, "Characterizing the file hosting ecosystem: A view from the edge," *Perform. Eval.*, vol. 68, no. 11, pp. 1085–1102, Nov. 2011. [Online]. Available: http://dx.doi.org/10.1016/j.peva.2011.07.016

[69] Sandivine, "Sandvine global internet phenomena complete'," Spring 2011, Fall 2011, Spring 2012.

[70] R. Buyya, M. Pathan, and A. Vakali, *Content Delivery Networks*, 1st ed. Springer Publishing Company, Incorporated, 2008.

[71] Cisco, "Visual Networking Index: Forecast and Methodology, 2011-2016," May 2012.

[72] Ipoque, "Internet study 2007." [Online]. Available: http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2007.pdf

[73] Ipoque., "Internet study 2008-2009." [Online]. Available: http://www.ipoque.com/sites/default/files/mediafiles/documents/internet-study-2008-2009.pdf

[74] S. Alcock and R. Nelson, "Measuring the impact of the copyright amendment act on new zealand residential dsl users." in *Proc. of ACM IMC'12.*

[75] T. Meyer, "Graduated response in france: The clash of copyright and the internet," *Journal of Information Policy*, vol. 2, no. 0, 2012. [Online]. Available: http://jip.vmhost.psu.edu/ojs/index.php/jip/article/view/71

[76] B. Danaher, M. D. Smith, R. Telang, and S. Chen, "The effect of graduated response anti-piracy laws on music sales: Evidence from an event study in france," in *Social Science Research Network (SSRN).*, 2012. [Online]. Available: http://ssrn.com/abstract=1989240

[77] R. Layton and P. Watters, "Investigation into the extent of infringing content on bittorrent networks," *Technical Report*, 2010.

[78] E. Felten, "Census of files available via bittorrent," 2010, https://freedom-to-tinker.com/blog/felten/census-files-available-bittorrent/.

[79] Le Figaro, "Hadopi : un premier internaute condamné." [Online]. Available: http://www.lefigaro.fr/hightech/2012/09/13/01007-20120913ARTFIG00599-hadopi-un-premier-internaute-condamne.php

[80] T. Ramayah, N. H. Ahmad, L. G. Chin, and L. May-Chiun, "Testing a causal model of internet piracy behavior among university students," *European Journal of Scientific Research*, 2009.

[81] S. Hinduja, "Trends and patterns among online software pirates," *Ethics and Inf. Technol.*, vol. 5, no. 1, June 2003.

[82] W. D. Gunter, G. E. Higgins, and R. E. Gealt, "Pirating youth: Examining the correlates of digital music piracy among adolescents," *International Journal of Cyber Criminology*, vol. 4, no. 1&2, 2010.

[83] M. W. Kampmann, "Online piracy and consumer affect : to pay or not to pay," July 2010. [Online]. Available: http://essay.utwente.nl/60470/

[84] T. Lauinger, M. Szydlowski, W. G. Onarlioglu, K., E. Kirda, and K. C., "Clickonomics: Determining the effect of anti-piracy measures for one-click hosting," in *NDSS' 13*, 2 2013.

[85] PCMAG.com, "After megaupload, storage sites shutter services." [Online]. Available: http://www.pcmag.com/article2/0,2817,2399238,00.asp

[86] A. Bridy, "Copyright policymaking as procedural democratic process: A discourse-theoretic perspective on acta, sopa, and pipa," in *Social Science Research Network (SSRN).*, 2012. [Online]. Available: http://ssrn.com/abstract=2042787

[87] "MaxMind- GeoIP." [Online]. Available: http://www.maxmind.com/app/ip-location

[88] BBC, "Megaupload file-sharing site shut down." [Online]. Available: http://www.bbc.co.uk/news/technology-16642369

[89] Wikipedia, http://en.wikipedia.org/wiki/Megaupload.

[90] Envisional, "An estimate of infringing use of the internet." [Online]. Available: http://documents.envisional.com/docs/Envisional-Internet%5FUsage-Jan2011.pdf

[91] Z. Jelveh and K. W. Ross, "Profiting from filesharing: A measurement study of economic incentives in cyberlockers," in *P2P*, 2012, pp. 57–62.

[92] US Department of Justice., "Justice department charges leaders of megaupload with widespread online copyright infringement." [Online]. Available: http://www.fbi.gov/news/pressrel/press-releases/justice-department-charges-leaders-of-megaupload-with-widespread-online-copyright-infringement

[93] G. S. Becker, "Crime and punishment: An economic approach," *Journal of Political Economy*, vol. 76, 1968.

[94] I. Ehrlich, "Crime, punishment, and the market for offenses," *Journal of Economic Perspectives*, vol. 10, no. 1, 1996.

[95] French Senate, "Haute autoritÈ pour la diffusion des uvres et la protection des droits sur internet," June' 2009. [Online]. Available: http://www.senat.fr/dossier-legislatif/pjl07-405.html

[96] SOPA, http://www.sopastrike.com.

[97] UK Government, "Digital economy act 2010," 2010. [Online]. Available: http://www.legislation.gov.uk/ukpga/2010/24/section/1?view=plain

[98] TechWeekEurope, "Britain's anti-piracy act delayed by cost dispute," April 2012. [Online]. Available: http://www.techweekeurope.co.uk/news/copyright-anti-piracy-act-delayed-75259

[99] Spanish Government, "Ley 2/2011,de economÌa sostenible."

[100] Intereconomia. [Online]. Available: http://www.intereconomia.com/noticias-gaceta/cultura/sgae-%E2%80%9C-valoracion-todos-los-sectores-culturales-negativa%E2%80%9D-20120724

[101] Parliamentary Council Office of New Zeland, "Copyright (infringing file sharing) amendment act 2011," April 2011. [Online]. Available: http://www.legislation.govt.nz/act/public/2011/0011/latest/DLM2764312.html

[102] Global Censorship Chokepoints, 2011, http://globalchokepoints.org/countries/south-korea.

[103] Internet Initiative Japan Inc, "Traffic shifting away from p2p file sharing to web services," *Internet Infrastructure Review*, vol. 8, no. 0, pp. 25–30, August 2010.

[104] http://www.netflix.com.

[105] "Server Intellect Use Policy," http://www.serverintellect.com/terms/aup.aspx.

[106] United Nations, "World urbanization prospects, 2011 revision." [Online]. Available: http://esa.un.org/unup/pdf/WUP2011_Highlights.pdf

[107] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J. Crowcroft, "The personality of popular facebook users," in *CSCW*, 2012.

[108] Stutzman, Fred; Gross, Ralph; and Acquisti, Alessandro , "Silent listeners: The evolution of privacy and disclosure on facebook," in *Journal of Privacy and Confidentiality: Vol. 4: Iss. 2, Article 2*, 2012.

[109] R. Dey, Z. Jelveh, and K. W. Ross, "Facebook users have become much more private: A large-scale study," in *PerCom Workshops*, 2012.

[110] P. Joshi and C.-C. Kuo, "Security and privacy in online social networks: A survey," in *ICME, IEEE*, july 2011, pp. 1–6.

[111] H. Gao, J. Hu, T. Huang, J. Wang, and Y. Chen, "Security issues in online social networks," *Internet Computing, IEEE*, vol. 15, no. 4, pp. 56–63, july 2011.

[112] C. Zhang, J. Sun, X. Zhu, and Y. Fang, "Privacy and security for online social networks: challenges and opportunities," *Network, IEEE*, vol. 24, no. 4, pp. 13–18, july 2010.

[113] H. O. Abdullahi, A. said, J. B. Ibrahim, "An investigation into privacy and security in online social networking sites among iium students," in *WCSIT*, ser. vol. 2, no. 2, 2012.

[114] B. D. Biswajit Das, "Social networking sites a critical analysis of its impact on personal and social life," *International Journal of Business and Social Science*, vol. 2, no. 14, 2011.

[115] Joiner R, Gavin J, Brosnan M, Cromby J, et al, "Gender, internet experience, internet identification, and internet anxiety: a ten-year followup," in *Cyberpsychol Behav Soc Netw.*, 2012.

[116] Depepedia Portal, "Depepedia: structured information of wikipedia." [Online]. Available: http://dbpedia.org/page/Paris

[117] J. P. W. M. C. W. A. M. R. Ottoni, D. Casas and V. Almeida, "Of Pins and Tweets: Investigating How Users Behave Across Image- and Text-Based Social Networks," in *ICWSM*, 2014.

[118] A. Crymble, "An Analysis of Twitter and Facebook use by the Archival Community," in *The Journal of the Association of Canadian Archivists*, 2010.

[119] NLTK, *NLTK madules for similarity*, 2014, http://www.nltk.org/api/nltk.metrics.html.

[120] A. Singhal, "Modern Information Retrieval: A Brief Overview," *IEEE Data(base) Engineering Bulletin*, vol. 24, pp. 35–43, 2001.

[121] Q. X. Yang, S. Y. Sung, L. Chun, L. Zhao, and S. Peng, "Faster algorithm of string comparison," *Pattern Analysis and Applications*, 2003.

[122] Socialbakers, *Socialbakers Portal*, 2014, http://www.socialbakers.com/.

[123] PageData, *Pagedata Portal*, 2014, http://www.pagedatapro.com/.

[124] M. B. A. L. David John Hughesa, Moss Rowea, "A tale of two sites: Twitter vs. Facebook and the personality predictors of social media usage," in *Journal of the Association of Canadian Archivists*, 2011.

[125] R. Motamedi, R. Gonzalez, R. Farahbakhsh, A. Cuevas, R. Cuevas, and R. Rejaie, "Characterizing group-level user behavior in major online social networks, Technical Report available at: http://mirage.cs.uoregon.edu/pub/CIS-TR-2013-09.pdf," 2014.

[126] B. Chun, D. Culler, T. Roscoe, A. Bavier, L. Peterson, M. Wawrzoniak, and M. Bowman, "Planetlab: an overlay testbed for broad-coverage services," *ACM SIGCOMM Computer Communication Review*, 2003.

[127] J. MacQueen, "Some methods for classification and analysis of multivariate observations," *Berkeley Symposium on Mathematical Statistics and Probability*, 1967.

[128] *WWE top 100 million facebook fans at wrestlemania*, http://corporate.wwe.com/news/2013/2013_04_10.jsp.

[129] *Annual Shorty Awards Winners*, http://shortyawards.com/.

[130] M. W. DiStaso and T. McCorkindale, "A benchmark analysis of the strategic use of social media for fortunes most admired u.s. companies on facebook, twitter, and youtube." *Public Relations Journal, 7(1), 1-33*, 2013.

[131] L. Evans, in *Social Media Marketing: Strategies for Engaging in Facebook, Twitter & Other Social Media*. Pearson Education, 2010.

[132] M. Barlow and D. B. Thomas, in *The Executive's Guide to Enterprise Social Media Strategy: How Social Networks Are Radically Transforming Your Business*. John Wiley and Sons, 2010.

[133] A. L. Richard D. Watersa, Emily Burnettb and J. Lucasb, "Engaging stakeholders through social networking: How nonprofit organizations are using facebook," *Public Relations Review, Elsevier*, 2009.

[134] M. Wysocki, "The Role of Social Media in Sports Communication: An Analysis of NBA Teams Strategy," Tech. Rep., 2012, http://www.american.edu/soc/communication/upload/Capstone-Wysocki.pdf.

[135] E. Ferrara, O. Varol, F. Menczer, and A. Flammini, "Traveling trends: Social butterflies or frequent fliers?" ser. COSN '13. ACM, 2013.

[136] R. Farahbakhsh, A. Cuevas, and N. Crespi, "Characterization of cross-posting activity for professional users across facebook, twitter and google+, Technical Report available at: http://www.it.uc3m.es/acrumin/papers/CrossPosts_TR.pdf," 2014.

[137] A. Martin and R. van Bavel, "Assessing the benefits of social networks for organizations," Tech. Rep., 2013, http://ftp.jrc.es/EURdoc/JRC78641.pdf.

[138] S. G. Lisette de Vriesa and P. Leeflang, "Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing," *Journal of Interactive Marketing, 2012, Vol.26(2), pp.83-91.*

[139] G. M. Andrew Lipsman and M. Rich, "The power of like: How brands reach and influence fans through social media marketing," *Journal of advertising research*, 2012. [Online]. Available: https://hospitalityandtravel.files.wordpress.com/2012/09/73177656.pdf

[140] R. Farahbakhsh, A. Cuevas, R. Cuevas, and N. Crespi, "Characterization of professional users strategies in major osns, Technical Report available at: http://www.it.uc3m.es/acrumin/papers/WWW_TR_strategy.pdf," 2014.

[141] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning*. Springer, 2006.

[142] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics bulletin*, pp. 80–83, 1945.

List of figures

List of tables