Study of the long tail formation within an eWOM community. The case of Ciao UK

PhD Thesis

in fulfilment of the requirements for the Doctorate Degree in “Strategic Management and International Business”

(Programa de Doctorado en Gestión Estratégica y Negocios Internacionales)

Submitted by:

María Olmedilla Fernández

born on May 12th 1984 in Burgos, Spain

Academic supervisor: Prof. Dr. Mª del Rocío Martínez Torres

Line of Research: Information Systems and Knowledge Management

Department: Business Administration and Market Research (Marketing)

Faculty: Economics and Business Sciences

Study of the long tail formation within an eWOM community. The case of Ciao UK

Author: María Olmedilla Fernández

Academic supervisor: Prof. Dr. Mª del Rocío Martínez Torres

PhD Program: Strategic Management and International Business *(Gestión Estratégica y Negocios Internacionales)*

Line of Research: Information Systems and Knowledge Management

Department: Business Administration and Market Research (Marketing)

Faculty: Economics and Business Sciences

University of Seville

Seville, December 2016

Sgd.: Mª del Rocío Martínez Torres

Sgd.: María Olmedilla Fernández

A mi madre, el ángel que me dio sus alas para poder volar

Acknowledgements

I would like to express my sincere gratitude my academic supervisor Professor Dr. María del Rocío Martínez Torres, who provided me an opportunity to join the research group “Big Data and Business Intelligence in Social Media” (SEJ548). I thank her for the support of my Ph.D. study and related research and for her motivation. My sincere thanks also go to Professor Dr. Sergio Toral Marín for his great academic support and for sharing his expertise so willingly.

This appreciation is extended to my research group and also to the Department of Business Administration, Marketing and Market Research, where the knowledge acquired have opened new horizons to me.

I would also like to thank all of my friends who supported me in writing, and encouraged me to strive towards my goal.

Finally, special thanks to all my family, in particular to my mother. Words cannot express how grateful I am for all of the sacrifices that she has made on my behalf. She is the most important person in my world and I dedicate this thesis to her.

Table of contents

List of figures

List of tables

Introduction

"Such is the power of the Long Tail. Its time has come."

Chris Anderson - Wired Magazine 2004

Motivation

Prior to Internet era, economic scale gave advantage to products oriented to a big amount of customers –such as books– instead of those aimed at niche markets (Anderson 2004). Nowadays, the inexpensive online medium and the reduced distribution costs have lowered the barriers of entrance (Kumar, Norris and Sun 2009, Martínez-Torres 2014). Besides and as stated by Anderson (2008a, p.4-5) *"increasingly, the mass market is turning into a mass of niche (…) and as the cost of reaching it falls—consumers finding niche products, and niche products finding consumers—it's suddenly becoming a cultural and economic force to be reckoned with."* In this regard, electronic word-of-mouth (eWOM) has widened customers' choices for assembling impartial product information from other customers, giving them the opportunity to offer their own consumption-related recommendation (Jansen et al. 2009). Continuous communication among people and ubiquitous online access are fundamental characteristics of online eWOM communities that are facilitating the distribution of a broad range of products and services. Thus, through eWOM communities, a great audience of users is able to acquire knowledge from reviews concerning products and services that are less popular to the majority. In that respect, the distribution of product sales is changing due to the increment of product information available to consumers (Brynjolfsson, Hu and Smith 2010). Actually, the long tail phenomenon is a manifestation of such transformations (Anderson 2004).

Challenges

Although many previous authors give a good understanding of the main idea behind long tail within sales distributions in product markets such as Amazon (Brynjolfsson, Hu and Smith 2003, Brynjolfsson, Hu and Smith 2010), this Thesis applies new

methodologies –elbow criterion– and extends others –power-law distribution– by Clauset, Shalizi and Newman (2009) to mathematically measure the long tail in other environments, such as the eWOM community Ciao.

Whereas most eWOM studies focus just on the potential of eWOM facilitating the long tail effect to find rare or niche products (Hennig-Thurau, et al. 2004, Khammash and Griffiths 2011) and how eWOM is enabling zero-cost dissemination of information about products (Odić, et al. 2013) and so forth, not many noticed that for each product type enclosed in the tail of the sales distribution there might be different impacts. For instance, the findings of this Thesis could be resembled to the interpretations of Lee, Lee and Shin (2011). Actually, the results within this Thesis might indicate that vendors could adopt alternative product strategies depending on with which niche product type (search or experience good) the tails of sales distribution would be formed. More specifically, this Thesis proposes an approach for detecting whether there is a long tail for each product type and thus, cases should be differentiated when niche products represent a significant portion of overall product sales.

Likewise, given the volume of the user-generated content in the web and its speed of change this Thesis also presents two important highlights in this regard. First, the implementation of an effective web crawler that can gather and identify big amounts of user-generated content. Second, the stages followed on this crawling process, which are the identification and collection of important data, and the maintenance of the gathered data. Consequently, social science needs to develop adequate methodologies to deal with huge amounts of data, such as the one outlined within this Thesis and overcome the distance between technology and social sciences.

Methodology

Despite a growing research literature about the niche products across the long tail in Internet, there is still a lack papers focusing on analytical models within online recommender systems. Hence, within this Thesis the chosen approach in the methodology has been to avoid the worst pitfalls from previous research by triangulating the method of power-law distribution of data gathered with other method, the elbow

criterion in order to identify the long tail. That is, to compare the all the type of products among the eWOM Ciao UK, the probability power-law distribution function was represented as a tool to measure the long tail. Besides, to extra validate such method the elbow criterion was also used to identify where was located the optimal cut-off point that distinguishes the products characterized by the long tail.

Furthermore, this Thesis outlines an architectural framework and methodology to gather user-generated data the eWOM community Ciao UK. To that end, a new methodology describes the implementation of a web crawler from other disciplinary perspective: the computing science discipline. In fact, due the large amount of user-generated data retrieved, this methodology reveals really interesting things about human behaviour in general (e.g. user interactions, participation analysis, content analysis, analysis of topics, sentiment analysis and the studied long-tailed phenomenon).

Contributions

The present thesis aims to contribute, by contrasting three hypotheses (see Figure 6) to the study of the long tail phenomenon in an eWOM community and what product types are enclosed there. In this regard theoretical and practical implications of this paper contributed with the development of two methodologies for identifying the long tail: through the power-law distribution method and by using the elbow criterion for corroboration on the examined data. Also a new product type categorization is proposed to differentiate type of products from eWOM by the degree of objectivity and by the degree of experiencing. Likewise, this Thesis has contributed with a methodology, which is not only understood from the perspective of the social science discipline but also includes practices on data accessing and computing. Such methodology comprehends the development of a web crawler for gathering user-generated data from eWOM communities, in which traditional data retrieval methods are challenged. Thanks to data collection, a better understanding of the insights found along the data set gathered from 28 main product categories of the portal Ciao UK has been achieved.

Last but not least, publications of refereed journals papers (indexed in JCR/JSCR) as well as conference papers related to the main topic of this Thesis have resulted from this Thesis as well.

Structure of the Thesis

This Thesis is organized as follows (see following Figure 1). The Part I discusses the theoretical framework of this Thesis by conducting a literature review to address the three hypotheses in the end. Accordingly, a review about eWOM communities, their current classification and a product type categorization within eWOM communities as been addressed. Also the literature on the phenomenon of the long tail within eWOM communities and the two methods to identify such phenomenon (power-law distribution and elbow criterion) has been addressed. Likewise, a theoretical background about the gathering of online user-generated data in the context of social science has been provided. Along the Part II the design of the research and the methodology describes the application of the two methods for identifying the long tail based on the literature reviewed. Besides, it is drawn the framework of the methodology to gather user-generated data from an eWOM community. Afterwards, Part III details the case study and data collection from the eWOM portal Ciao UK. The analysis and discussion of results are next presented in Part IV. The aim of this part is connecting the theoretical framework to the methodology and the case of study by way of the research hypotheses. Finally, Part V concludes the Thesis with the conclusions and a discussion of the implications and limitations of this study as well as plans for future research.

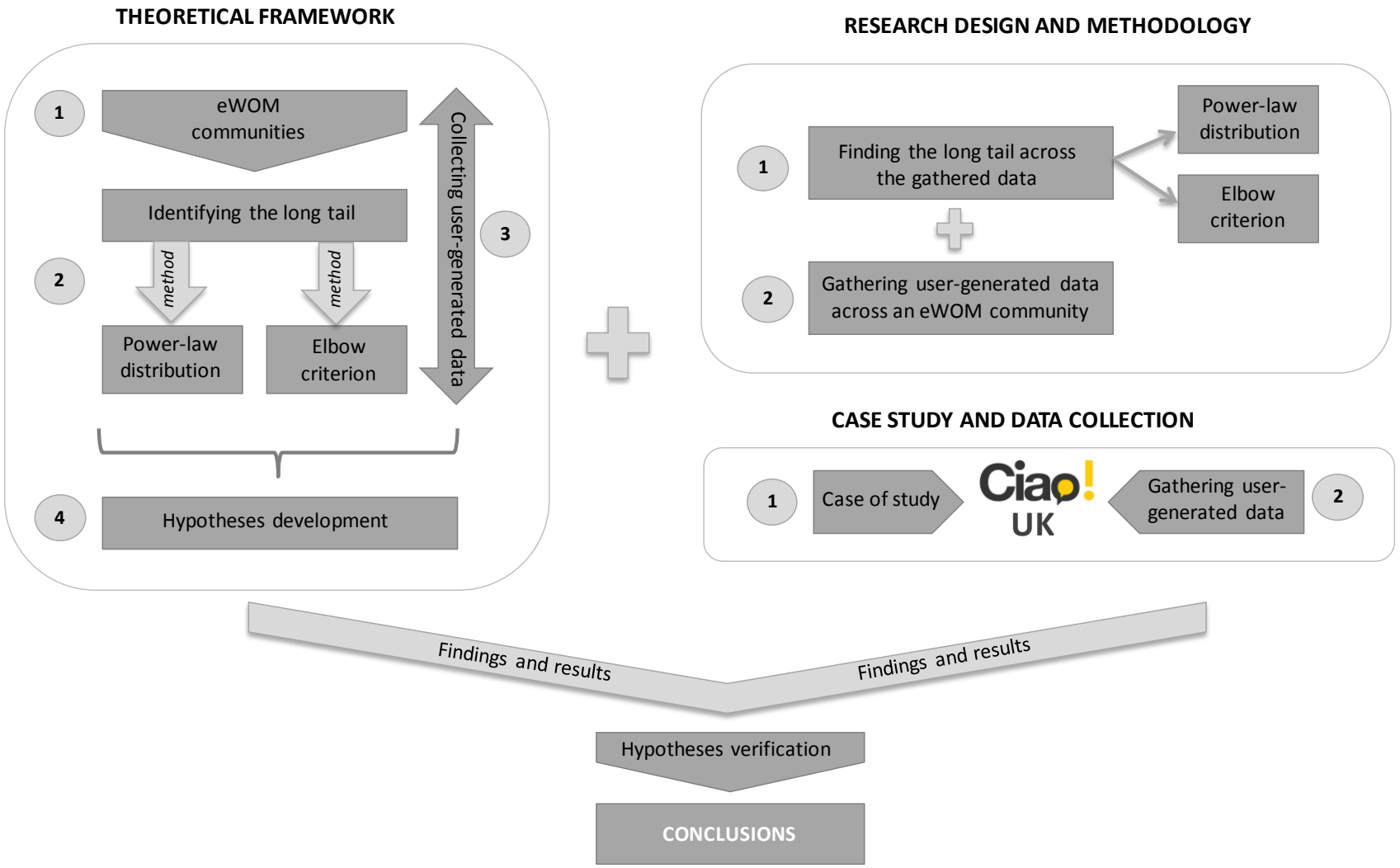


Figure 1 – Schematic structure of the thesis

Part I: Theoretical Framework

Along this first part of the dissertation, the literature is reviewed so as to address the subsequent hypotheses. On the one hand, the first two sections focus on the field of study, which is the phenomenon of the long tail across eWOM communities. To that end, section 1 assesses the definition of eWOM communities from their origin and illustrates an overview of their current classification as well as a product type categorization within eWOM communities. Section 2 assesses the phenomenon of long tail through the literature, its importance within social science, the way in which is linked with eWOM communities and two methods to identify such phenomenon: (1) power-law distribution and (2) elbow criterion. Afterward, as eWOM communities provide massive quantities of information produced by and about people, section 3 provides a theoretical background about the gathering of online user generated data in the context of social science. This first part ends setting the hypotheses of the Thesis.

1 eWOM Communities

This chapter begins by describing the concept of electronic word-of-mouth (eWOM) communities defined in the literature, outlining and explaining it from its origin. Then, an overview of what an eWOM community is and how it emerged is followed. Thereafter a structured classification of all kind of eWOM communities is displayed. Finally, a product type categorization of eWOM communities according to its attributes and evaluation standards is presented.

1.1 Classic WOM shifting towards eWOM

Word-of-mouth (WOM) has been recognized for several years as a great influence on what people know, feel and do. The power of word-of-mouth and its influence on consumer decision-making are well proven in academic literature (Engel, Kegerreis and Blackwell 1969, Herr, Kardes and Kim 1991, Bone 1995, Goldsmith and Horowitz 2006). Besides, those authors generally support that WOM is more influential on behaviour than other marketer-controlled sources. In that regard Katz and Lazarsfeld (1955) conducted one of the first formal studies where the authors found that WOM was the most important source of influence when making a purchase decision concerning of household products and food. Their study revealed that WOM was twice effective in influencing consumers' behaviour as radio advertising, four times as effective as personal selling, and seven times as effective as newspapers and magazines. Subsequently, WOM as a phenomenon has been researched and defined during the 1960s and its WOM definitions have evolved over time. For instance, Engel, Kegerreis, and Blackwell (1969) found that about 60 per cent of the consumers mentioned WOM as the most influential source regarding their adoption of an automotive diagnostic centre. Similarly, Arndt (1967, p. 3) was one of the earliest researchers into the influence of WOM on consumer behaviour. He defined WOM as: *Oral person-to-person communication between a receiver and a communicator whom the person perceives as non-commercial, regarding brand, product or a service.* Later, Westbrook (1987, p. 261)

characterized WOM more broadly including: *"all informal communications directed at other consumers about the ownership, usage, or characteristics of particular goods and services or their sellers".*

While WOM has traditionally been studied from the perspective of face-to-face communication involving sharing product or brand information (Schindler and Bickart 2005), it is also becoming prevalent in online shopping environments (Dellarocas 2003). Besides, the emergence and impact of user-generated content through the Web has made classic WOM to shift towards eWOM (Hennig-Thurau, et al. 2004), which has been enabled by Internet and online communication (Moe and Trusov 2011). Regardless of the form of WOM, the focus of the communication is the sharing of information regarding individuals' experiences with various products and services (Steffes and Burgee 2009). However, it is important to understand the relevant differences between electronic and traditional WOM.

In the online context, there is typically no familiarity between senders and receivers of eWOM (Gupta and Harris 2005) and, in contrast with traditional WOM, conversations are visible to the rest of the consumers (Gonzalez-Rodriguez, Martinez-Torres and Toral 2014). Statements made by customers within eWOM can be exchanged through a variety of online communities such as discussion forums, electronic bulletin board systems, newsgroups, blogs, review sites, and social networking sites (Goldsmith and Horowitz 2006). Furthermore, by using search engines customers can seek out the opinions of strangers, given that overall information search and dissemination costs are lower online than offline (Kumar, Norris and Sun 2009). This rarely occurs in traditional WOM contexts where opinion providers belong to the inner circle of individuals and well-known people are considered more credible (Sun, *et al.* 2006). Thus, the strength of the relationship between a communicator and a receiver is one of the most distinctive differences between WOM and eWOM (Chatterjee 2001). Unlike traditional WOM, there is a lack of familiarity between eWOM receivers and senders, since consumers exchange opinions and experiences outside their personal social network (Cheung and Thadani 2012), which may heighten the potential for the posting and use of fraudulent techniques. Conversely, according to Dellarocas (2003) Internet allows a powerful

social force to be precisely measured and controlled through proper engineering of the information systems that mediate online feedback communities. Besides, the author argues that such online feedback mechanisms or the so-called eWOM differentiate from traditional WOM because of "*(1) their unprecedented scale, achieved through the exploitation of the Internet's low-cost, bi-directional communication capabilities, (2) the ability of their designers to precisely control and monitor their operation through the introduction of automated feedback mediators, and (3) new challenges introduced by the unique properties of online interaction such as the volatile nature of online identities and the almost complete absence of contextual cues that would facilitate the interpretation of what is, essentially, subjective information".* Moreover, much of the persuasive nature of eWOM is associated to the assumption that sometimes consumers trust more communications from consumers rather than from marketers or advertisements (Lee and Youn 2009). Basically, the notion of eWOM has caught significant consideration in both business and academic communities (Steffes and Burgee 2009) and the scope of published studies is large and fragmented. Furthermore, there is an increasing assortment of definitions and descriptions on the concept eWOM in the current literature tending to deal with two thought streams, eWOM as an aspect of e-commerce (Alanah and Khazanchi 2008, Goldsmith 2006) or as interpersonal influence (Zhang, Craciun and Shin 2010, Lis 2013). For instance, according to Hennig-Thurau et al. (2004, p. 39) the term eWOM refers to *"any positive or negative statement made by potential, actual, or former customers about a product or company, which is made available to a multitude of people and institutions via the Internet".* As can be observed, even though the authors give a good understanding of the main idea behind eWOM, this definition is quite general. In that regard, the authors Kietzmann and Canhoto (2013, pp. 147-148) developed an improved a clearly defined framework of the concept eWOM as: *"any statement based on positive, neutral, or negative experiences made by potential, actual, or former consumers about a product, service, brand, or company, which is made available to a multitude of people and institutions via the Internet (through web sites, social networks, instant messages, news feeds…)".*

1.2 The concept and emergence of eWOM communities

Nowadays, consumers are no longer restricted to the traditional one-way seller-to-buyer communications (Moe and Trusov 2011). Consumers use eWOM as they search for online information in order to reduce this perceived risk that accompanies high-involvement products (Prendergast, Ko and Siu Yin 2005, Schindler and Bickart 2005). For this reason, several of the principal online retailers, such as Office Depot, Amazon, Home Depot and Macy's, facilitate eWOM by allowing online reviews on the products they offer through discussion boards and other online communication tools (Gupta and Harris 2005). However, there are websites or online communities that are exclusively dedicated to exchange eWOM. Those are the eWOM communities, which have emerged to influence customers directly and create interest with efficacy and flexibility in spite of geographic boundaries (Duan, Gu and Whinston 2008). They provide rich and objective product information that is influencing customers' decision making (Gu, Tang and Whinston 2013, Kim and Gupta 2009, Zacharia, Moukas and Maes 2000), due to the credibility, empathy and relevance they offer to customers as opposed to the information provided by marketer-designed websites (Bickhart and Schindler 2001, Pan, MacLaurin and Crotts 2007). Additionally, a further significant attribute of an eWOM community is its speed, accessibility, one-to-many reach, and its lack of face-to-face human pressure (Phelps, et al. 2004). Besides, through eWOM communities, users are able to freely post their reviews about any product or service, and share those reviews with other users in order to better understand a product (Hennig-Thurau, et al. 2004). Similarly, eWOM communities enable the interaction among users, as they can share their experiences and also comment or rate other users' reviews (Arenas Márquez, Martínez-Torres and Toral 2014). In that regard, eWOM communities provide an aggregated rating for each product given by the reviewers' scores, showing a quick general impression of the product (Qiu, Pang and Lim 2012).

Hence, consumers have become progressively more online active in sharing product and service experiences and as aforementioned eWOM is widely recognized as an influential information source on consumer decision-making, which is well established in academic literature (Jiménez and Mendoza 2013, Park, Lee and Han 2007). In this

respect, also the authors Cheung and Thadani (2012) have conducted a literature analysis based on prior eWOM studies, where they built an integrative framework explaining the impact of eWOM communication on consumer behaviour. Conversely, some concerns related to the use of an eWOM community have been also addressed in the literature. For instance, the authors Cheng and Zhou (2010, p. 4) suggested that *"consumers have to put more attention to evaluating the content of eWOM to determine whether it is credible or not".* Nevertheless, the author Yang (2013) uses a scenario simulation experiment method to show that eWOM information quality and credibility are positive to information, which in turn depends on factors such as source's popularity or community status. In this line, research has shown that by participating in an eWOM community, customers derive both social and economic value (Sridhar Balasubramanian 2001) and consequently may have different motivations in using or producing eWOM (Hennig-Thurau, et al. 2004).

Although all the previous authors show different levels of specificity but with common aspects to give an understanding of the concept eWOM community, it is also important to illustrate an appropriate working definition for the sake of clearness. Hence, an eWOM community can be described as:

> *A community of users who freely share their experiences, thoughts and opinions about a wide variety of products and services through online computer-mediated communication. Such users also receive the feedback of the rest of the community in the form of comments and usefulness score, which is key to rely on their shared opinions. Accordingly, the consumer information shared is usually more effective than marketer-generated one since it is more convincing and relevant to the consumers.*

1.3 Classification of eWOM communities

According to Bickhart and Schindler (2001), individuals who sought for product information from corporate websites were less interested in making a purchase decision

than individuals who sought for product information from eWOM communities, such as discussion forums. Furthermore, Cheong and Morrison (2008) also notice that within online discussion boards, when reading posts, participants rarely evaluate the source of the content and trust that only because it is generated by peers. In general, consumer-generated communication is more useful than marketer-generated communication (Thorson and Rodgers 2006). This indicates that the place in which eWOM emerges affects and maximizes the persuasive effectiveness of the eWOM messages (Sussan, Gould and Weisfeld-Spolter 2006). Hence, it is important to understand types of eWOM communities, which in fact could be of interest to marketers.

The online world offers a variety of appropriate platforms for eWOM by which consumers can exchange information such as blogs, microblogs, discussion forums, review websites, shopping websites, virtual consumer communities and social media websites (Cheung and Thadani 2012, Goldsmith and Horowitz 2006, Strutton, Taylor and Thompson 2011). While certain literature has addressed those different types of eWOM and their differences, there has not been a consolidated conceptualization of such differences. To reach this objective, a grounded classification will serve as a distinction of eWOM communities' features and as identification of their commonalities and differences. The classification proposed within this thesis is delimited to eWOM communities and it is also part of the one provided by Cheung and Thadani (2012). The study enumerates five different types of eWOM according to their literature analysis of prior eWOM studies. However, the authors do not give a detailed description of what each type encompasses. Additionally and within this line, Kozinets et al. (2010) show an understating about the forces characterizing eWOM. As stated by the authors, those forces work together to change the nature of the eWOM message by transforming it from a commercial promotion to communally valuable information. Within their study they explain that eWOM is influenced by the subsequent four important factors or forces: (1) character type, meaning different archetypal patterns in how people offer perspectives that unfold over time; (2) blog forum, where communications take place; (3) communal norms, which rule the manifestation, transmission, and reception of a message and its meanings; and (4) the promotional characteristics of the marketing campaign. Accordingly, following these two studies and other gathered literature about

eWOM and a variety of platforms to post user-generated content, the next Table 1 shows the eWOM communities classification grounded on the level of user interactivity and participation, the typology of eWOM platform or the communication strategy.

Table 1 – eWOM communities' classification
Source: Own elaboration based on Cheung and Thadani (2012)

Type of eWOM community	Description	Communication strategies	User interactivity
Social Media sites *(e.g. Facebook, Twitter)*	Truly appropriate platforms for eWOM (Canhoto and Clark 2013, Trusov, Bucklin and Pauwels 2009) where opinions, experiences and promotion about products or services are exchanged with friends and acquaintances, which might influence users (Kozinets, et al. 2010, Chu and Kim 2011).	Users create and promote profiles relating to products and services of brands (Chen, et al. 2014). Besides, users share their comments through written texts, pictures, videos or even applications (Erkan and Evans 2016).	Users display their preferences to their network, such as becoming a fan of brands, interacting with brands posts through liking and commenting, or posting a brand included content (Erkan and Evans 2016). This allows consumers to gain trustworthy information from others' positive and negative opinions (Jansen, et al. 2009)
Blogs *(e.g. blogger, wordpress)*	Online personal journals in which an individual publish his or her opinions, experiences, and stories about particular topics and visitors are able to comment and to interact with the publisher (Thorson and Rodgers 2006, Schmallegger and Carson 2008). Accordingly, consumers gain useful information from blogs to facilitate their purchase decisions. (Schmallegger and Carson 2008).	Bloggers are expected to conform communal norms such as the length of posts, or the use of photos and to post messages appropriate to their forums (Kozinets, et al. 2010). Besides, when increasing the hit rate of blogs also more emotional topics are created in order to generate consumer opinions (Yang, Huang and Lin 2009).	Blogs are not only about sharing information but also they build trust, friendship, and alliances (Rettberg 2008). Thus, through advertising and promotions influential bloggers on the basis of their lifestyle relevance can influence visitors and consumers (Kozinets, et al. 2010). Besides, blogs *"are second only to newspapers as a trusted information source"* (Lee, Brown and Broderick 2007, p. 16).
Online discussion forums *(e.g. travellerspoint, rankers)*	An online discussion forum is a general concept of a platform, a virtual avenue, for members with common interests in certain products to exchange knowledge, social bonds, and enjoyment that is relevant to those products (Schindler and Bickart 2005, Chan and Li 2010, Cheung, et al. 2009, Bickart and Schindler 2001).	Members rarely make their identities known to others, thus the role of influencers is bigger (Cheong and Morrison 2008). This anonymity makes viewers focusing more on information helpfulness, consumer's posting history and the feedback from others to analyse the information quality (Xun and Reynolds 2010, Burton and Khammash 2010)	Posted messages are considered trustworthy sources since consumers do not gain financial benefit from product companies (Bickart and Schindler 2001). In this line, administrators should design message-rating systems that allow users to evaluate messages in several attributes, such as argument strength, understandability or objectivity, instead of just offering a general evaluation score (Cheung, et al. 2009).

Type of eWOM community	Description	Communication strategies	User interactivity
Brand communities *(e.g. community smartthings, oracle community)*	*"A specialized, non-geographically bound community, based on a structured set of social relations among admirers of a brand*" (Muniz and O'guinn 2001, p. 412). It is likely to be formed around brands with strong image and a rich lengthy history (Muniz and O'guinn 2001). This community might have either evolved on its own or could have been enabled by the brand owner (Raj Devasagayam, et al. 2010).	Members of a brand community establish the agenda and specific community activities by the inter-relationships among members who possess the same brand by exchanging information and meanings about the brand (Muniz and Schau 2005). Besides, the community devises rituals and mores of its own, beyond the control of the brand owner (Raj Devasagayam, et al. 2010).	There is a space for committed members to express their brand improvement ideas and/or complaints about bad experiences with the brand. Commitment refers to each member's attitude toward the community and it is used as a predictor of members' actual behaviours, such as participating in community activities, offering help to the community, and solving problems for others (Hur, Ahn and Kim 2011).
Online consumer review sites *(e.g. Ciao, Epinions)*	Users write reviews, experiences, evaluations, and opinions about products or services on sellers' or third party's websites and other members rate the usefulness of such reviews. (Hennig-Thurau and Walsh 2003, Dellarocas 2003). This usefulness is based on the authorship of the review as well as the content of the review (Li, et al. 2013).	Users are also able to post their comments online in the form of an account of their own experience with a product (Hennig-Thurau and Walsh 2003). However, these communities are not only a place for sharing, but also have great potential to significantly influence readers who intend to use on-line recommendations for purchase decisions (Cheung, et al. 2009).	A reviewer reputation system is used to express credibility information about the reviewer. Users rate multiple aspects of reviewed items. Thus, users can directly invest trust in other members based on their posting histories. The reviewer's reputation could also be conferred based on past contributions and posting records (Cheung, et al. 2009, Dellarocas 2003).

1.4 Product type categorization along eWOM: the role of product attributes and evaluation standards

eWOM communities often have a strong impact on product judgements because the information received is more accessible. Reviews are published on a great variety of products and services, and have become part of the decision making process for consumers (Mudambi and Schuff 2010). Besides, a product review is focused on a specific product category, and the effect of eWOM depends on the product itself (Park and Lee 2009). Depending on the nature of their specific attributes, products can be generally classified as either search products or experience products (Cui, Lui and Guo 2012). On that subject, Nelson (1970, 1974) was the first author who classified products into search and experience according to consumers' capacity to acquire product quality information before purchase. Search products are described as goods, which quality can be evaluated by consumers through objective and specific attributes before purchase. Contrariwise, experience products are typically evaluated by affective attributes, which come from the consumers' experiences with the product. Thus, those are more difficult to be described using specific attributes (Nelson 1970, Nelson 1974, Weathers, Sharma and Wood 2007, Mudambi and Schuff 2010, Cui, Lui and Guo 2012). Despite the fact that several products comprise a combination of search and experiences attributes, this categorization of search and experience goods is still relevant and commonly accepted (Huang, Lurie and Mitra 2009).

Nelson (1970, 1974) argues that consumers conduct minimal pre-purchase information search for experience goods, but achieve extensive information for search goods. Also in this regard, Park and Lee (2009, p. 63) agree that *"finding search attributes on the Internet is easier than finding experience attributes, consumers are likely to possess a much more detailed, rich cognitive structure for search goods than for experience goods".* However, existing studies of eWOM have principally deal with information on products such as books (Brynjolfsson, Hu and Smith 2003, Chevalier and Mayzlin 2006), movies (Yang, Huang and Lin 2009, Yeap, Ignatius and Ramayah 2014), or music (Morales-Arroyo and Pandey 2010). These products can be classified as experience products. For this reason, online reviews within eWOMs are very important

when consumers are choosing products they do not have first-hand experience with. Besides, the results from the study by Park and Lee (2009) ensure that the eWOM effect is greater for experience goods than for search goods. Their article describes how eWOM information's direction and website's reputation contribute to the eWOM effect focusing on the moderating role of the product type (search vs. experience). Following this line of thinking, Huang, Lurie and Mitra (2009) investigate the differences in consumer's search patterns between search and experience goods in the online context. Their study reveals that the interaction between products before buying through consumers' product reviews has a higher effect on online consumer's search and purchase behaviour for experience than for search goods (Huang, Lurie and Mitra 2009). Similarly, Senecal and Nantel (2004) conducted a study across multiple product categories and found that consumers trust online recommendations sources for experience products much more than for other types of products. The analysis made by Ku, Wei and Hsiao (2012) also shows that product type has an important effect on members' reputation, in accordance with the results from those previous studies, which indicate that the effect of eWOM depends on product type.

Nelson's (1970, 1974) original classification is still relevant in the online context since it it provides insights into online consumer behaviour. Nevertheless, it might be argued that there are other approaches to product classification —such as Lee, Lee and Shin (2011) classification based on product evaluation standards. The authors' study develops a new approach to product categorization and shows that the impact of eWOM on sales distribution is different across product types. According to Lee, Lee and Shin (2011), for product evaluation, customers examine various attributes using two types of evaluation standards: objective versus subjective standards. Some attributes associated with objective evaluation standards are capacity, warranty, class, and power, whereas attributes associated with subjective evaluation standards are colour, design, style, and genre. The authors evaluate products through different combinations of the objective/subjective attributes and the major objective or subjective attributes is chosen. Similarly, Ghose and Ipeirotis (2011) describe a technique that identifies which information within online reviews is objective and which is subjective. They classify reviews with objective information as those "*listing the characteristics of the product,*

and giving an alternative product description that confirms (or rejects) the description given by the supplier". On the contrary, the authors classify reviews with subjective information as those *"in which the reviewers give a very personal description of the product, and give information that typically does not appear in the official description of the product"* (Ghose and Ipeirotis 2011, p. 1502).

Both studies (Lee, Lee and Shin 2011, Ghose and Ipeirotis 2011) attempt to explain that the type of product differentiation according to its evaluation standards can also determine the position of a product on the classification made by Nelson (1970, 1974). Although several products include a mix of search and experience attributes, the studies concur that search goods include those products that are evaluated through more objective standards, such as feature-based products, where their quality[1] is evaluated before purchase. On the contrary, for experience goods, the product is evaluated through more subjective standards that describe those aspects that are not captured by the product description and their quality is not evaluated before purchase.

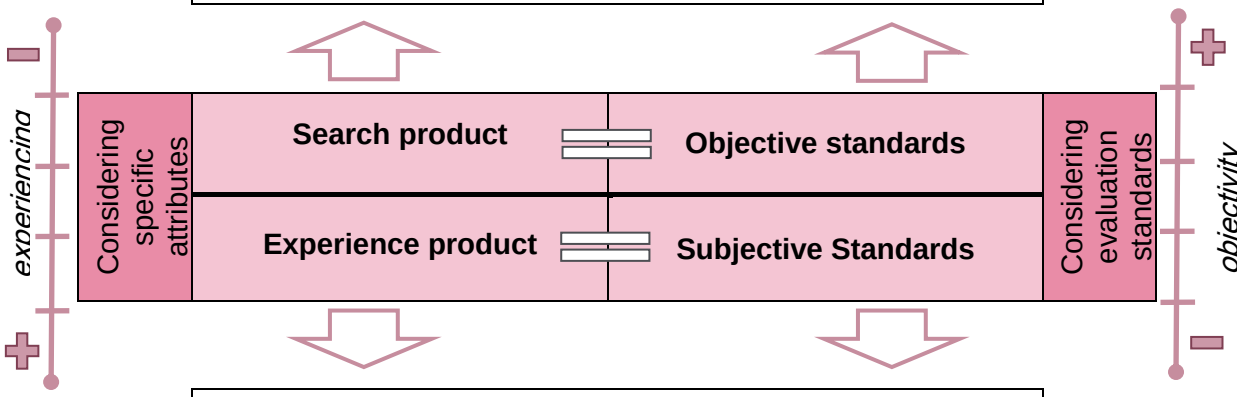
According to the results from both articles, the following Figure 2 provides a product category scheme that differentiates type of products from eWOM. The degree of *objectivity,* considering product evaluation standards, is given by the significance of the objective attributes in the total product assessment. Likewise, the degree of *experiencing*, considering specific attributes, is determined by the affective/feeling evaluative cues in the total product assessment.

[1] The main information in eWOM about a product is quality, in other words, whether this product is good or bad (Forman, Ghose and Wiesenfeld 2008, Li and Hitt 2008).

Description:
Products distinguished and ordered by quality, which availability of quality is checked before purchasing (Lee, Lee and Shin 2011).

Attributes:
Length, weight, class, warrantee, capacity, power, etc. (Lee, Lee and Shin 2011)

Examples:
Computers, home and garden (Ku , Wei and Hsiao 2012), cameras (Nelson 1970), consumer electronics (Cui, Lui and Guo 2012), natural supplement pills (Weathers, Sharma and Wood 2007), clothing and jewellery (Nelson 1974).



Description:
Products that cannot be distinguished and ordered through quality. Besides, quality will not be evaluated before purchase.

Attributes:
Colour, shape, design, style, genre, brand, etc. (Lee, Lee and Shin 2011).

Examples:
Cars and motorsports, restaurants and gourmet (Ku , Wei and Hsiao 2012), wine (Klein 1998), movies, books (Mudambi and Schuff 2010), video games (Cui, Lui and Guo 2012), DVDs (Ghose and Ipeirotis 2011), music (Nelson 1970, Bhattacharjee, Gopal and Lertwachara 2006), health and beauty products (Nelson 1974), Household equipment and appliances (Nelson 1970).

Figure 2 – Product type categorization depending on specific attributes or evaluation standards

2 Identifying the Long Tail among eWOM communities: A comparison of two methods

This section begins by addressing the concept of long tail, its main features and the importance of studying such phenomenon in the discipline of Social Science. Furthermore, this chapter reviews the literature based on the two disciplines or methods to identify the long tail. The first method provides the discussion of the power-law distribution and how it is behaved among the eWOM communities in order to identify the long tail. The second method embraces the elbow criterion and how it is used to categorize the long tail as well.

2.1 Anatomy of the long tail

From a theoretical point of view, the term long tail refers to the principle of statistical distribution that emphasizes the considerable weight of infrequent or minor events compared to frequent or major events (Benghozi and Benhamou 2010). In a normal bell curve, the highest frequency events appear at the centre of the distribution, and then they are gradually narrowed down at the upper and lower ends. Those narrowed ends refer to the long tail. Long tail events rarely occur, but sometimes the total aggregate number of low-frequency events that are found in the tail can be superior to the group of high-frequency events in the short head (Lew 2008). Accordingly, the term long tail is applied to a distribution curve, which often form power-laws and are thus long-tailed distributions in the statistical sense (Brynjolfsson, Hu and Smith 2003, Lee, Lee and Shin 2011). From an economical point of view, the long tail phenomenon denotes the behaviour of economic areas that provide products in reasonably low volume, but are able to make profit by providing a bigger diversity of products in aggregate. Contrariwise, the short head sectors profit is based on a narrower assortment of products selling in much higher volume (Lew 2008). Furthermore, the concept long tail describes the structure and success of Internet-based activities by representing a new approach to the marketing and selling of products that did not exist prior to the advent of

the Internet (Lew 2008). In this line, an early mention of the long tail in the context of the Internet was first outlined by Anderson (2004), who defined this new trend in the field of economics. He argued that niche products would become more prevalent since online retailers sell more products that are less popular than traditional retailers do. His view held that online recommendations networks based on user-generated content such as Amazon guide consumers into the long tail of personalization by covering that percentage of the market of lower-selling niche products. Moreover, with the proliferation of online channels, consumers will find it easier to search for and discover those obscure products since Internet has removed many of the communication barriers between geographically distant locations (Brynjolfsson, Hu and Smith 2003, Brynjolfsson, Hu and Smith 2006). In this regard, Anderson (2008a) outlined two different but related ideas. The first one is that sales of products below a certain volume increase proportionately to those of products above a certain volume because physical and cost constraints on selection disappear. The products do not have to be displayed on stores due to the elimination of the communication barriers between geographically distant locations through Internet. The second idea is that online channels are changing the shape of the demand curve since people are interested in exploring those niche products, which are more likely to be geared to their particular interests than in those oriented to the mass market. Those products can collectively comprise a market share that competes or surpasses current bestsellers, but only if the distribution channel is large enough (Odić, et al. 2013). Thus, the Internet frees many businesses from traditional location factors since it provides an inexpensive medium for individuals and businesses to reach audiences and potential customers (Lew 2008). At the same time it enables the long tail as more rare products are made available to consumers (Anderson 2008a). Furthermore and according to Lew (2008), whether storage and distribution costs are high, only the most profitable products can be sold at a competitive price. Thus, the Internet is probably the only inexpensive approach to economic success in the long tail market-place.

In order to have a better understanding of the anatomy of the long tail, the following Figure 3 illustrates an example of standard demand curve that could be applied to any sales of products.

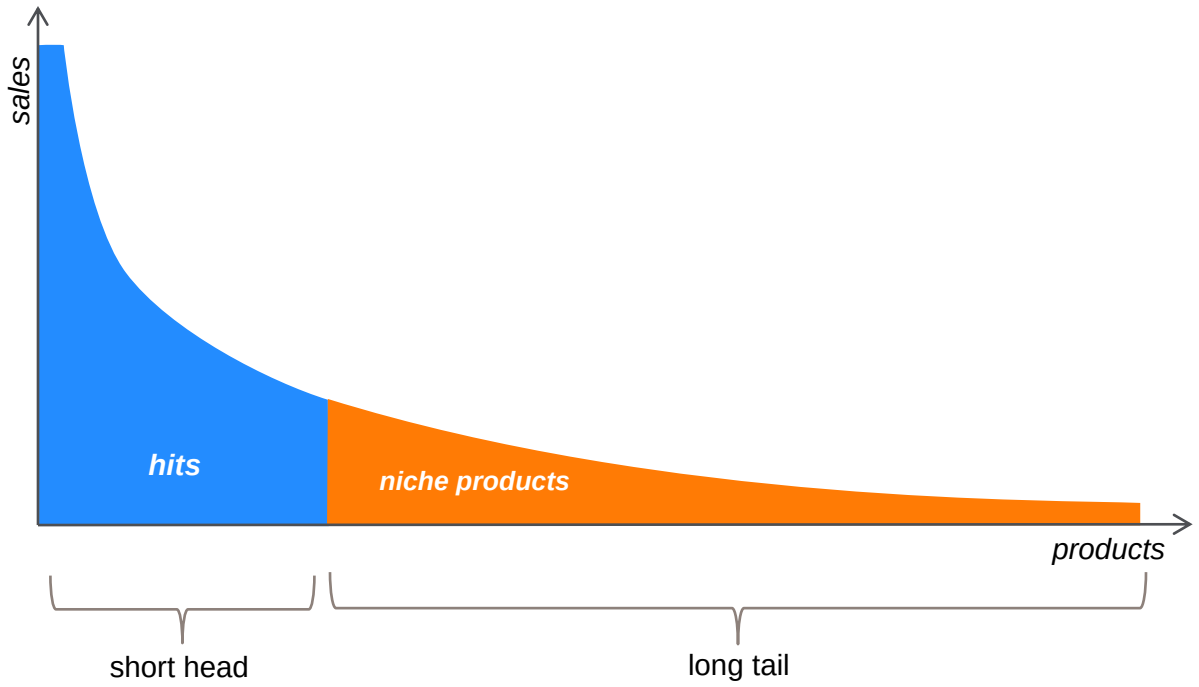


Figure 3 – Example chart representing the long tail

Here, all possible offerings in an imagined product sector are ranked by their sales volume. The vertical axis represents sales, while the horizontal represents the products. The blue part of the curve depicts the high volume sales or super star products, which have been dominating brick-and-mortar markets. The orange part is the lower-selling or niche products, which reveals a previously untapped demand.

2.2 Importance of the long tail in Social Science

The majority of data generated and collected in science is important to the scientific process of theory development and evaluation. According to Heidorn (2008), the data across the long tail is a significant resource in social science. In his article he discusses how the data from long tail economics can be brought to light and made more useful to the scientific community. In this line, the author argues that data in the tail tends to be more poorly curated and less visible to other scientists and gives potential solutions for proper management of this class of data.

Understanding the long tail in social science may suggest, for instance, where researchers could study how Internet retailers could use the availability of niche products as a strategic tool in its competition with other Internet retailers (Brynjolfsson, Hu and Smith 2010). Even though it would be interesting to study whether an Internet retailer can adopt a long tail strategy to move consumers from shopping high-volume sales products towards niche products that are more likely to fit their tastes (Brynjolfsson, Hu and Smith 2010). Consequently, such data along the long tail can also offer new ways for academics and practitioners to research consumer behaviour.

Following this idea, several explanations for the study of long tail in Internet can be traced to several authors. Brynjolfsson, Hu and Simester (2011) draw attention to Internet's long tail by analysing data collected from a clothing retailer. They discover that consumers' usage of Internet search and discovery tools, for instance recommendation engines, are related with an increase in the share of niche products. Elberse (2008) analyses the Anderson's Long Tail idea and also studies how the tail of the sales distribution is getting longer and fatter by looking at Rhapsody Music or Quickflix among others. Elberse and Oberholzer-Gee (2007) show evidence that Internet retailing has shifted demand toward niche video products over time and argue that exploiting the tail could be unprofitable if many titles do not sell at all. Tucker and Zhang (2011, p. 828) examine data from a web site that lists wedding service vendors and suggest the following: *"popularity information may benefit niche products with narrow appeal disproportionately, because the same level of popularity implies higher quality for narrow-appeal products than for broad-appeal products".* Moreover, according to Huberman and Wu (2008), the same logic can be applied to peer-to-peer file-sharing networks that also offer content diversity. Since digital products can be offered at practically no additional cost, it is a viable strategy for online retailers to offer products that sell only in small quantities, which guarantees the real time provision of any content regardless of its popularity.

As can be observed the long tail is of utmost importance across the social science, since the field of marketing and product development focuses more on these small, minority niche market segments. This is because they enhance consumer preferences

by innovating to meet niche and individual demands and interests. Hence, long tail supports and grows diversity in terms of new innovations and in terms of conserving traditions that might otherwise be gone (Anderson 2004, Lew 2008).

2.3 Linking the long tail and eWOM communities

In the case of online shopping, having product information is essential since consumers purchase products without physical examination. Therefore and to avoid the risk of buying undesirable products, consumers gather all type of information presented on the Internet before making their buying decisions (Lee, Lee and Shin 2011). In that respect, for online customers, eWOM communities are not only a primary source of product information, but they also assign value to informational content (Dwyer 2007). As a consequence, several products that are not in the list of bestselling now become visible as they bring together users of similar interests, similar standards of judgment, or geographic location (Elberse 2008). Besides, thanks to lower distribution costs and new ways of connecting demand and supply at a world scale, a shift in demand from the most popular products to niche products might occur facilitating again the long tail phenomenon (Peltier and Moreau 2012).

Nevertheless, there are two sets of studies suggesting opposite directions. Proponents of the long tail idea maintain that increasing the assortment variety of products offered through online channels (e.g. eWOM communities) will in turn intensify sales of lower-selling niche products (Brynjolfsson, Hu and Simester 2011, Elberse 2008, Anderson 2004, Brynjolfsson , Hu and Smith 2009); whereas others defend that these online channels will promote the sales of popular products with even high ratings (Standifird 2001). Paradoxically, unlike these studies, according to Lee, Lee and Shin (2011), the roles of eWOM in both cases are the same; it helps consumers to find the product they are looking for. When consumers look for a less popular product, eWOM helps them and causes the long tail to appear. But if customers look for a popular product, eWOM also helps them to find it; nonetheless, in this case the tail of the sales distribution is shortened.

2.4 Method 1: Identifying the long tail through power-law distribution

2.4.1 *The concept of power-law distribution*

Along the literature numerous empirical analysis of diverse real phenomena such as the population of the cities (Blank and Solomon 2000), earthquakes (Gutenberg and Richter 1956), the annual income of the people (Okuyama, Takayasu and Takayasu 1999, Drăgulescu and Yakovenko 2001), the viscosity of cooled liquids (Taborek, Kleiman and Bishop 1986), the magnetic field power spectra (Fedi, Quarta and De Santis 1997), the World Wide Web (Albert, Jeong and Barabási 1999), etc. show power-law behaviour in the upper tail of their distributions over different scales. Besides, data collected to measure the parameters of such distributions often extend over a large-scale range, where the majority of values in the distribution may be beyond that of the scale range of observation (Pickering, Bull and Sanderson 1995).

As can be observed, statistical patterns are uncovered in nature and society and their distribution might correspond to mathematical models. Furthermore and according to Clauset, Shalizi and Newman (2009), many of the things that scientists measure have a typical size or 'scale'—a typical value around which individual measurements are centred. An example made by the authors would be the heights of adults pulled randomly from a human population. In this regard, those many measured processes in nature have density functions following a Gaussian distribution. Moreover, the Central Limit Theorem claims that the sum of random variables with a finite mean and a finite variance converge to a Gaussian distribution (Rosenblatt 1956). The mean corresponds to the average value of the random variable and the variance corresponds to the measure of individuals differing from that average (Goodman 1963). Hence, a Gaussian distribution occurs as a result of universal statistical processes and not similar causes. Nevertheless, not all measured processes are peaked around a typical value. Some vary over a huge dynamic range, sometimes many orders of magnitude (Newman 2005). For instance, when the variance and/or mean is not finite anymore, subsequently the Central Limit Theorem does not predict Gaussians but it predicts a variety of sum-stable distributions, which all look like power-laws. Besides, large events are extremely

rare within Gaussian distributions. On the contrary, such events in the tail of the distribution —the part of the distribution representing large but rare events— are more likely to happen in a power law distribution (Clauset, Shalizi and Newman 2009). Consequently, when the probability of measuring a particular value of some quantity varies contrariwise as a power of that value, the quantity is said to follow a power-law— also known as Zipf's law or Pareto distribution (Newman 2005).

According to Virkar and Clauset (2014), a quantity x obeys a power-law if it is drawn from a probability distribution with a density of the form

$$p(x) = \alpha\, x^{-\alpha} \qquad (1)$$

where $\alpha > 1$ is the exponent or scaling parameter and $x > x_{min} > 0$. Thus, if the function $p(x)$ describes the probability of being above a quantity x_{min} it is said that the tail of the distribution follows a power-law. Since power-laws usually describe systems where the larger events are rarer than smaller events α remains positive. This ensures that the power-law is a monotonically decreasing function. In order to have a better understanding of the power-law, the Figure 4 illustrates four random examples of two different distributions (Gaussian and power-law) in which the probability, p, of a particular value, x, occurring is plotted against that value. (1) The graph on the left (pink) displays a bell-curve, which is symmetrical with the highest centre at the mean value. That is to say, more sample data is presented around the mean value, which presents a single peak with tails on the high and low ends of the curve. On the right graph (green) a power-law distribution is depicted, which has more sample data with extreme value than a Gaussian distribution, showing a curve with a long tail dropping as the value increases. (2) In these graphs logarithmic axes are used. The example shows the Gaussian distribution (pink) that has a mean value 0 and variance 1, while the power-law distribution (green) has an exponent –0.6. As can be observed, when data are plotted on a log-log scale, the power-law is a straight line sloping down from left to right.

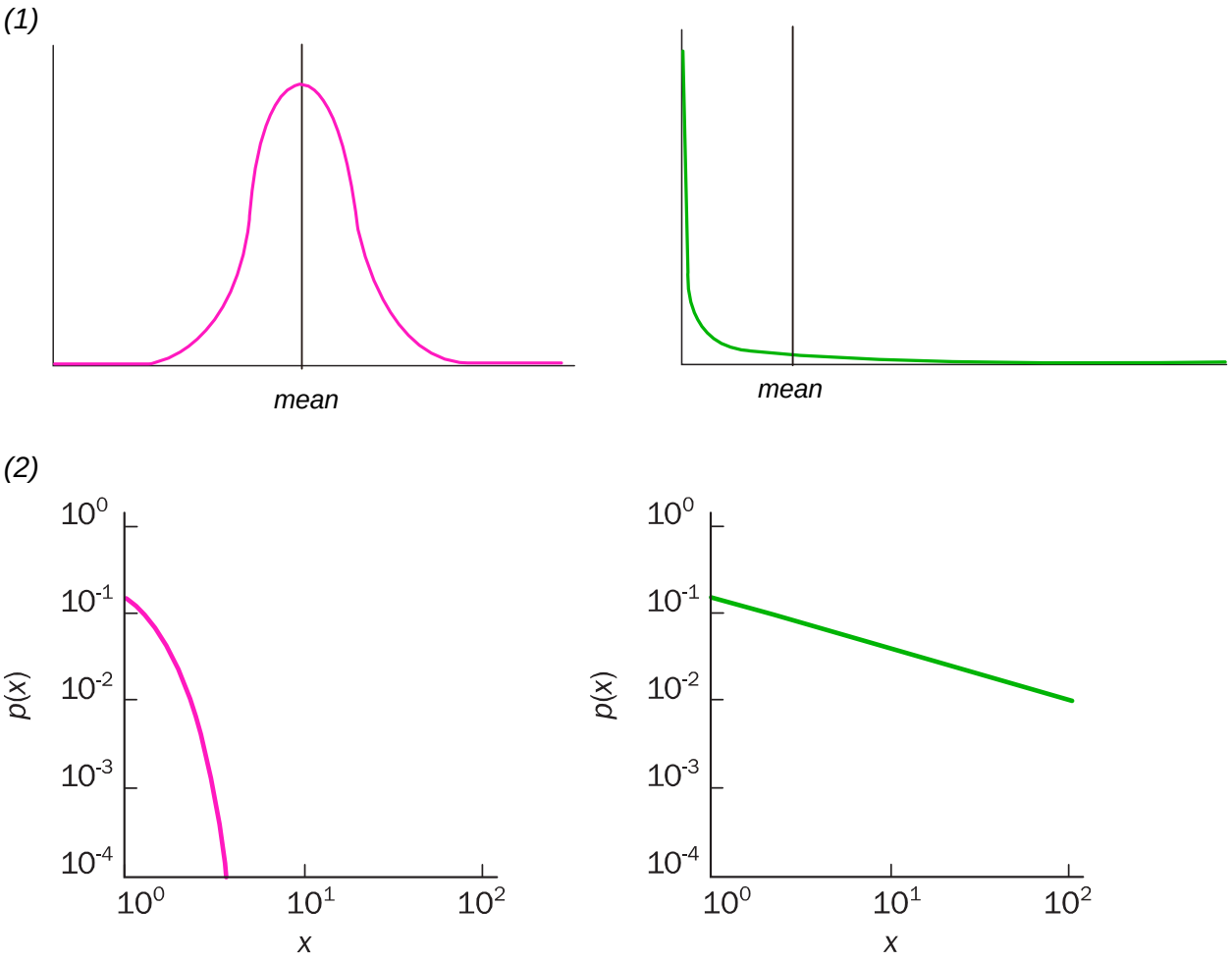


Figure 4 – Comparison of graphs representing Gaussian distribution and power-law distribution

2.4.2 *The long tail revealed as power-law relationships across eWOM communities*

A previous study of the Internet's long tail phenomenon made by Brynjolfsson, Hu and Smith (2003) suggested that a power-law can be used to describe the relationship between Amazon sales and Amazon sales rank and to estimate which niche books are not typically stocked in brick-and-mortar bookstore. Hence, from a mathematical point of view, the long tail is a manifestation of power-law relationships (Mahanti, et al. 2013).

Also, according to Anderson, customer demand across a product space takes the form of a power-law (Anderson 2008a). Furthermore, according to a series of breakthroughs in network theory by authors such as Huberman (2003), among others, it is known that power-law distributions tend to arise in social systems where several people express their preferences among numerous choices. Such pattern has also been identified within web objects access frequencies (Breslau, et al. 1999). Further examples of such phenomena are the World Wide Web (Adamic and Huberman 2000), Internet TV (Cha, et al. 2008), paper citations distribution (Redner 1998) or web pages visibility (Martínez-Torres y Díaz-Fernández 2013). In the same way, other widely reported examples are YouTube video popularity (Phillipa, et al. 2007) and popularity of TV channels in Internet TV workloads (Cha, et al. 2008), where they appear to follow a Zipf-like distribution, which belongs to the family of discrete power-law probability distributions. On the Internet, power-law seems to be the rule rather than the exception. Nevertheless, detection and characterization of power-laws remains still complex, especially due to the large fluctuations that arise in the tail of the distribution and also because of the difficulties of identifying the range over which power-law behaviour is supported (Clauset, Shalizi and Newman 2009).

In short, as stated by the aforementioned literature large events happen more often in systems that exhibit power-law distributions. This means that the events in the tail of the distribution are more likely to occur in a power-law distribution. That is why power-laws are also called heavy-tailed (Gomes, et al. 2000). Thus, the extreme tails are important because they give an idea of how often, on average, the largest events might occur. Within this line, it is important to emphasize that the literature reviewed suggests a relevant association between eWOM communities, power-law distribution and the long tail phenomenon. Firstly, eWOM communities allow customers to obtain information related to products from a massive, geographically dispersed group of people, who have experience with niche products. Secondly, such exchange of information facilitates a long tail effect, as more customers are able to access low-volume products. Consequently, power-law distribution is considered a valuable tool to measure these information uncertainties since Gaussian distributions cannot handle when happening at a certain probability.

2.5 Method 2: Identifying the long tail through the elbow criterion

In order to manage big quantities of data, it is convenient to use data analysis techniques that permit extracting its best information. Clustering is one of these techniques for statistical data analysis, and it is used in many fields, including data mining (Ng and Han 1994), pattern recognition (Duda and Hart 1973) or image analysis (Celenk 1990) among others to find relationships in the data and to concisely model the data distribution (Palmer and Faloutsos 2000). More accurately, according to Kuri-Morales and Rodríguez-Erazo (2009, p. 58), clustering is the data mining procedure that consists of *"processing a large volume of data to obtain groups where the elements of each group exhibit quantifiably (under some measure) small differences between them and, contrariwise, large dissimilarities between elements of different groups".* There are many algorithms and methods proposed in the literature for clustering. The *k-means* clustering algorithm is the most frequently used due to its simplicity (Kodinariya and Makwana 2013) and also one of the most popular heuristics for solving the *k-means* problem. It is based on a simple interactive scheme for finding a locally minimal solution (Hartigan and Wong 1979). Likewise, there are several approaches in the literature to determine the number of clusters for *k-means* clustering algorithm such as cross-validation, variance based method, the elbow criterion and so on (Kodinariya and Makwana 2013). Additionally, when there is no a clear idea of the number of clusters a *k-means* cluster analysis using the "elbow criterion" is an appropriate method to estimate the optimal number of clusters.

The elbow criterion is a visual method (Kodinariya and Makwana 2013) to determine what number of clusters should be chosen for *k-means* clustering. This method is based on the plot of the ratio of the between-group variance to the total variance as a function of cluster number. If the number of clusters increases then, the total variance explained by the data increases monotonically until an additional increase in the number of clusters reaches a plateau (Krahe, et al. 2011). More specifically, when graphing the percentage of variance explained by the clusters against the number of clusters, the first clusters might add much information, however at some point the marginal gain will drop, giving an angle in the graph (Madhulatha 2012). Accordingly, the elbow criterion

proposes an intuitive selection of the optimal number of clusters such that the decrease in the value of the objective function is not substantial enough (Twarakavi, Šimůnek and Schaap 2010).

Figure 5 illustrates the elbow criterion for the selection of the appropriated number of clusters. This figure plots the total sum of squared deviations *F* versus the number of clusters *K*.

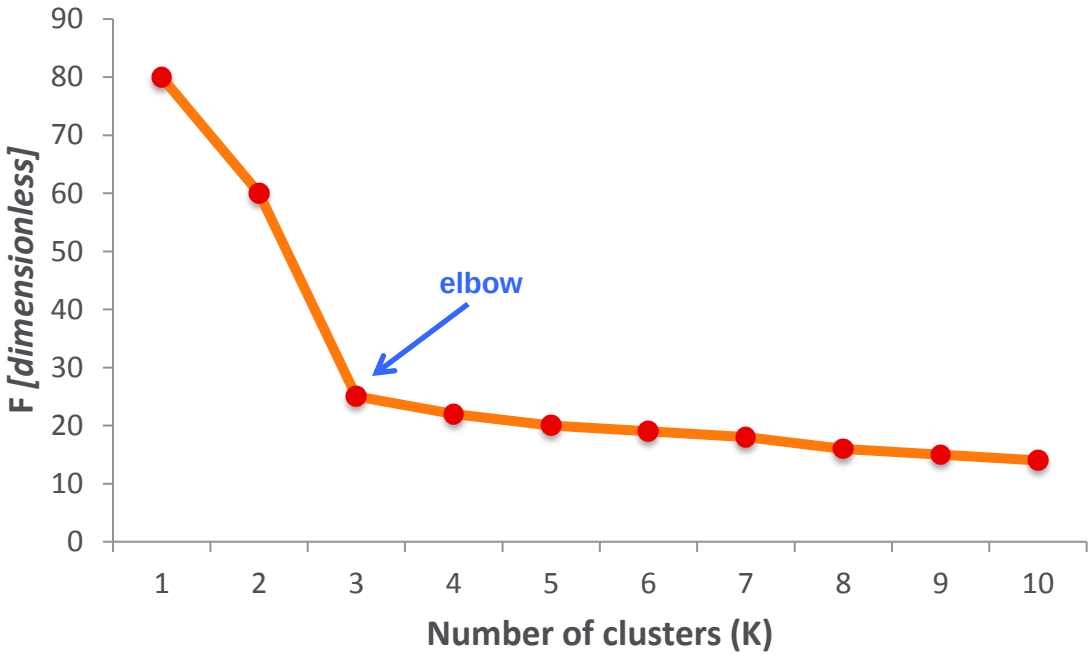


Figure 5 – Location of the elbow to delineate the appropriate number of clusters within a given dataset

The idea is that at some value for K the value (the one required) falls noticeably and after that it reaches a plateau when increasing it further. In the example of Figure 5, this breakpoint happens at *K=3*. The *F* goes down quickly with *K* increasing from 1 to 2, and from 2 to 3, and then reach an elbow at *K=3*, afterwards the *F* value goes down gradually. Therefore, the number of clusters needed for this data set is 3, since that corresponds to the elbow of this curve. This indicates that there is a high probability that the optimal number of clusters is located at that point.

The rationale is that the elbow criterion is a common rule that permits determining what number of clusters should be chosen. In this regard and as appreciated in the Figure 5, an inflection in the curve (the elbow) reflects the number of clusters that account for a substantial fraction of the variance. Beyond such point, an addition of further clusters yields no more sufficient information and do not noticeably increase the amount of variance accounted for (Krahe, et al. 2011). Additionally, it should be distinguished that the lack of an elbow might be a signal that there is no cluster structure in the data (Krahe, et al. 2011).

3 Collecting online user-generated data

There is a vast amount of information shared among eWOM communities. Besides, users are becoming more powerful and deeply involved as content generators in such online environments. Therefore, it is important to understand that all these Big Data within social science is not only about the created content nor its consumption. Actually, it is also about the capture, search, discovery, and analysis tools that help gaining insights from unstructured data. In that regard, Big Data is gaining meaning into social science from quantitative datasets side, which differs from traditional social science where collecting data has always been hard, time consuming, and resource intensive. The emergent field of computational social science is broadening researchers' perspectives. However, it also requires a multidisciplinary approach involving several and different knowledge areas. Consequently, this chapter is focused on illustrating the importance of online user-generated data, its massive collection within social science and the significant role played by the social scientists with such data.

3.1 Collecting Big Data in Social Science

A better access to information is powering the interest in Big Data (Anderson 2008b). Over the next years, it is expected that the increasing volume of data created and collected in Internet persists (Kaisler, et al. 2013). Besides, with the increased

automation of data collection and analysis, handling the emergence of an Era of Big Data is critical (Boyd and Crawford 2012). However, most of the Big Data still remains wild and unstructured. Likewise, selecting the content of interest from the huge and constantly expanding universe of user-generated data exhibits one of the most fundamental challenge for applications for data collection: to explore large volumes of data and extract useful information or knowledge for future actions (Leskovec, Rajaraman and Ullman 2014). When using appropriate instrumentation for data collection, it is possible to take advantage of the information that comes from user-generated content such as clickstreams, tweets, user opinions, auction bids, consumer choices or social network exchanges (Chang, Kauffman and Kwon 2014). In numerous situations, the knowledge extraction process has to be very efficient and close to real time because storing all observed data is nearly unfeasible. In that regard, advanced computational techniques are exploiting the potential of technology to capture and analyse such big amounts of data from the Internet in increasingly powerful ways (Eynon 2013). This is offering the humanistic and social science disciplines the possibility of making many social spaces quantifiable, so they can be studied following a quantitative approach (Boyd and Crawford 2012). Actually, the evolution in computer aided research methods is changing the way in which social science research and data processing is done (Demchenko, et al. 2013).

Hence, for an intelligent system to handle such acquisition of Big Data the essential key is to provide a processing framework, which includes considerations on data accessing and computing, as well as algorithms that can extract knowledge. In addition to providing a variety of data analysis methods, such knowledge discovery must supply a means of storing and processing the data at all stages of the pipeline, meaning from initial ingest to serving results (Begoli and Horey 2012). To achieve such goal, an overview of crawlers (or spiders, robots, wanderers, etc.) for collecting and indexing all accessible web documents will be here introduced. Nevertheless, it has to be emphasized that, all this process of gathering Big Data cannot be effectively understood from the unique disciplinary perspective of social science. Convergence among several disciplines to deal with the emergence of Big Data should be taken into account (Chang, Kauffman and Kwon 2014). Furthermore, according to McCloskey (1985), what gives

accuracy to social scientists' work is not only rooted in all the way to data analysis and interpretation of the results but also in their systematic approach to data collection. To that end, a researcher can retrieve the data stored in the web through APIs provided by most social media services and largest media online retailers, which are not complicated to use. For example, the public API provided by Twitter to request specific information on the social network (Teutle 2010). However, in many cases they do not provide all the data required by researchers. Besides, in order to archive all tweets from Twitter pipeline payment is required, since normally twitter only makes a part of its pipeline available to public. For instance, some additional features of users can be necessary to perform data cleaning and filtering operations, such as previous experience of users or their popularity or reputation. Such specific information is not usually available using APIs, and more computational specialized techniques are then necessary (Manovich 2011). Therefore, collecting Big Data is a skill set generally restricted to those with a computational background. They use methods from the discipline of computer science such as web crawlers in order to capture the full potential of Big Data without any restriction.

The rapid growth of the web poses unprecedented scaling challenges for web crawlers, which seek out pages in order to obtain data. According to Najork (2009, p. 3462), a web crawler is a *"program that, given one or more seed URLs, downloads the web pages associated with these URLs, extracts any hyperlinks contained in them, and recursively continues to download the web pages identified by these hyperlinks"*. Several crawling systems and architecture have been described in the literature. For instance, Chakrabarti, Van den Berg and Dom (1999) and Seyfi, Patel and Júnior (2016) describe in their papers a focused crawler and briefly outline its basic process, which seeks, acquires, indexes, and maintains pages that represent a narrow segment of the web rather than crawling the entire web. Equally, well established is the principle of operation of web crawlers stated by Cothey (2004). The author presents an experiment that examines the reliability of web crawling as a data collection technique. Prior to these authors, Pinkerton (1994) describes the architecture of the web crawler and some of the trade-offs made in its design. The author specifies three actions

performed by a crawler: (1) marking the document as retrieved, (2) deciphering any outbound links and (3) indexing the content of the document.

Additionally, it is important to highlight that given space limitations in dealing with extremely large datasets extracted from crawling the web – especially when working with a very large and diverse information collection – there seems to be fundamental to create a database system in order to have an organized collection of data. What is more, due to such large size of data it becomes very difficult to scale it or processing it in an efficient way as well as to extract value from it (Katal, Wazid and Goudar 2013). In this context, MapReduce-based systems for large-scale data processing and storing such as Hadoop are being used in both industry and academia. This is due to the ease-of-use, scalability, and failover properties (Dittrich and Quiané-Ruiz 2012, Chen, Alspaugh and Katz 2012). Nevertheless, one of the main performance problems with Hadoop MapReduce is its physical data organization including data layouts and indexes (Dittrich and Quiané-Ruiz 2012). Besides, in the context of data gathered form eWOM communities it is not necessary such systems or enormous databases to handle large-scale data processing needs, but a relational system. This is because data through eWOM communities has plenty useful information and its structured data is organized in a way so that it can be managed easily. Thus, a well integration with a relational database is preferred.

3.2 User-generated data in Internet

Social science has been traditionally handling collection of data in passive observation or active experiments, which aim to verify one or another scientific hypothesis (Demchenko, et al. 2013). On that subject, it is still common practice in social science to develop further survey models to collect data sets directly from the users. Contrariwise, the public is increasingly choosing not to respond to surveys (Curtin, Presser and Singer 2005, De Leeuw and De Heer 2002). Besides, advances in data collecting technologies and data storage make it possible to obtain and preserve massive data generated directly or indirectly by users in Internet to generate valuable new insights

(Michael and Miller 2013). In the same way, with the emerging capabilities to collect data sets from diverse real world contexts, Internet has become the researcher's new behavioural research lab (Chang, Kauffman and Kwon 2014). Especially during the last years the rapid expansion of social networking applications, such as Facebook or Twitter have allowed users to generate content freely and amplify the already massive web volume of data (Fan and Bifet 2013). In that regard, among the current literature there are several studies and projects in which user-generated online content have been used to carry out analysis in social sciences. For instance, Antenucci et al. (2014) from the University of Michigan used Twitter data to create three job-related indexes for the US economy: job loss, job search and job posting. Similarly Llorente et al. (2015) focused on tweets about unemployment to demonstrate how social media activity relates to the socio-economic situation across Spanish regions. In the financial field, also a growing number of papers are investigating whether the data coming from online social networks can help to improve the prediction of financial variables such as the study conducted by Nardo, Petracco and Naltsidis (2016).

User-generated content is a significant means through which consumers express themselves and communicate with others online. It is what is produced in the moment of being social and the object around which sociality happens (Boyd and Ellison 2010, Smith , Fischer and Yongjian 2012). As the consumption, creation, and distribution of user-generated content continues to evolve, online communities are becoming more usable and accessible to consumers, facilitating the creation of manageable information space that is both customized and relevant (Daugherty, Eastin and Bright 2008). In this regard, customers' reviews fall into the general category of user-generated content, and they may well be the leading form of user-generated content (Liu, Karahanna and Watson 2011). Besides, according to Cheong and Morrison (2008) user-generated content is related to, but not equal with eWOM. While user-generated content is wider than eWOM, the two overlap noticeably. The authors emphasize that the main difference between them depends on whether the content is generated by users or the content is conveyed by users. For instance, videos posted on YouTube, which are generated and posted by users are considered user-generated content. Nonetheless, a user sending friends a link to a YouTube site is engaging in eWOM. Besides, as stated

by the authors, whether the content conveyed has been generated by users, it can be both user-generated content and eWOM. Consequently, *"to be successful, eWOM depends on the dissemination of content, and user-generated content has less influence without eWOM"* (Cheong and Morrison 2008, p. 39).

Significant challenges are present in the search to capture the full potential of user-generated content. Whereas several data collection and analysis tools have become accessible on the web (e.g. APIs and crawling methods from Twitter or Facebook) during the last few years, still, they are quite limited (Manovich 2011). In that respect, Jagadish et al. (2014) emphasizes that a good resource to avoid those tools limitations is the development of data cleaning techniques during the gathering. Hence, as Chang, Kauffman and Kwon (2014) observe, researchers should deliberate how the tools participate toward gathering and extracting maximal value from data. Web crawlers might facilitate this process. While some researchers relay on APIs and other tools to retrieve user-generated content, web crawlers do a better an exhaustive harvesting as well as they are more topic-specific (Pant, Srinivasan and Menczer 2004). This is because web crawlers can also extract specific content information while browsing the target website.

3.3 The role of Social Scientists within Big Data

The phenomenon of Big Data is closely bound to the appearance of *data science* or the so-called *computational organization science*, a discipline that combines mathematics, programming and scientific instinct. Such discipline has widened researchers' perspectives on social systems, by embracing computational models that combine social science and computer science (Reips and Garaizar 2011, Jagadish, et al. 2014, Syed, Gillela and Venugopal 2013). Besides, according to Manovich (2011) this combination of data abundance and the appearance of computational data analysis have shaped three kinds of divisions among people, which are the so-called *"new data-classes of the big data society"*: people creating data, people with skills collecting data, and people with expertise analysing data. On this line of thinking, Davenport and Patil

(2012) present the role of the data scientist, who basically comprehends the skills of the three aforementioned new data-classes defined by Manovich (2011). The authors emphasize that not only are important the technologies for taming Big Data, but also the people with the skills to put those technologies to good use and to retrieve meaning from the unstructured gathered information. In this regard, recently, a far wider range of social scientists have become more involved about the potential of Big Data, which is creating challenges and opportunities for interdisciplinary researchers (Chang, Kauffman and Kwon 2014). For instance, in his article in Wired magazine, Anderson (2008b) suggested that research methodologies in social science should not only be based on building theoretical models but also on having better data and using better analytical tools. The beginning of digital convergence in the social sciences is accelerating the way phenomena are studied (King 2014). Furthermore, the recent advancements in Big Data technologies such as software tools to gather the content of interest from user-generated data facilitate the paradigm change in the so-called modern e-Science (Boyd and Crawford 2012).

In general, most of the researches focus just on the analysis or modelling step of the Big Data pipeline. While that step is essential, the other phases such as data gathering are at least as important (Jagadish, et al. 2014). Nevertheless, not only is an important aspect of the data scientist the ability to write code together, but also the ability of analysing large quantities of data. In this regard, Boyd and Crawford (2012) state that, although computational scientists have started engaging in acts of social science, it does not mean that methodological issues are no longer relevant when dealing with Big Data. As said by the authors *"understanding sample, for example, is more important now than ever"*.

4 Hypotheses development

Following from the above theoretical framework, several explanations for the study of long tail in Internet can be traced to several authors. In fact, there are mixed results from previous studies that have found differing evidence regarding the shift in the sales distribution for different types of products. On the one hand the authors Gu, Tang and Whinston (2013) examine the long tail phenomenon through the informative effect in the context of eWOM. They show that positive reviews in Amazon improve the sales of popular products more than the sales of niche products. Therefore, they suggest that online WOM restrains the formation of the long tail. On the other hand, using also data from Amazon a previous study by Chevalier and Mayzlin (2006) find the opposite, that online reviews influence book sales. The authors argue that the total number of books sold at Amazon is higher than it would have been absent the provision of customer review features. This argument is also consistent with Zhu and Zhang (2010), whose study advocates that online consumer reviews affect the diffusion and adoption of less popular products. Also in this line, Elberse and Oberholzer-Gee (2007) found that the number of non-selling video titles increased in the time period 2002 to 2005.

This leads to the question of whether or not the eWOM communities affect product types differently. Hence, in order to understand the shifts in sales distributions for different products, it is also necessary to understand the product types and how eWOMs affects them. eWOM introduces less popular products to customers and offers detailed information of the products so that it informs customers, who would not have bought the products without eWOM, about those products (Hennig-Thurau, et al. 2004, Arenas Márquez, Martínez-Torres and Toral 2014). In this regard, Huang, Luri and Mitra (2009) argue that the Internet serves as an important information source for the two product categories: experience and search goods. The authors emphasize that the online context favours experience goods since it provides a useful channel to propagate quality information and “experience” the product before buying it. Also in this line, findings by Lee, Lee and Shin (2011) reveal thinner heads and longer tails in the sales distribution of products with less objective evaluation standards. Based on this line of

thinking and to complement the above literatures, the following hypotheses are postulated:

H_1: The experience products from the distribution of product categories within an eWOM are more likely to exhibit a long tail.

H_2: The search products from the distribution of product categories within an eWOM are less likely to exhibit a long tail.

According to Standifird (2001) eWOM may inhibit the long tail phenomenon by promoting the sales of popular products with high ratings. Thus, the head part of the sales distribution becomes thicker generating high-frequency events in the short tail. Customers who search for a product that is well rated by other customers often find the one by sorting through popular products (Chevalier and Mayzlin 2006, Lee, Lee and Shin 2011). Paradoxically, in both cases (less popular product and super-hits) the eWOM helps customers to find the product they are looking for. In this context, whether customers try to obtain different information about the products, the way they search and make choices is different for the two types of products – search and experience products (Huang, Lurie and Mitra 2009). Therefore, considering product type is important because it determines customers' product-search patterns and might result afterwards in different sales distributions. For instance, *"if customers buy books, they may look for the ones that are conforming to their personal taste but are not necessarily popular products. In contrast, if customers buy television sets, they may look for best-selling products whose quality is verified by other customers"* (Lee, Lee and Shin 2011, p. 467).

The following hypothesis follows from the preceding discussion:

H_3: The distribution of product categories within an eWOM that have high frequency events or super-hits in the short head are not particularly associated with search or experience products.

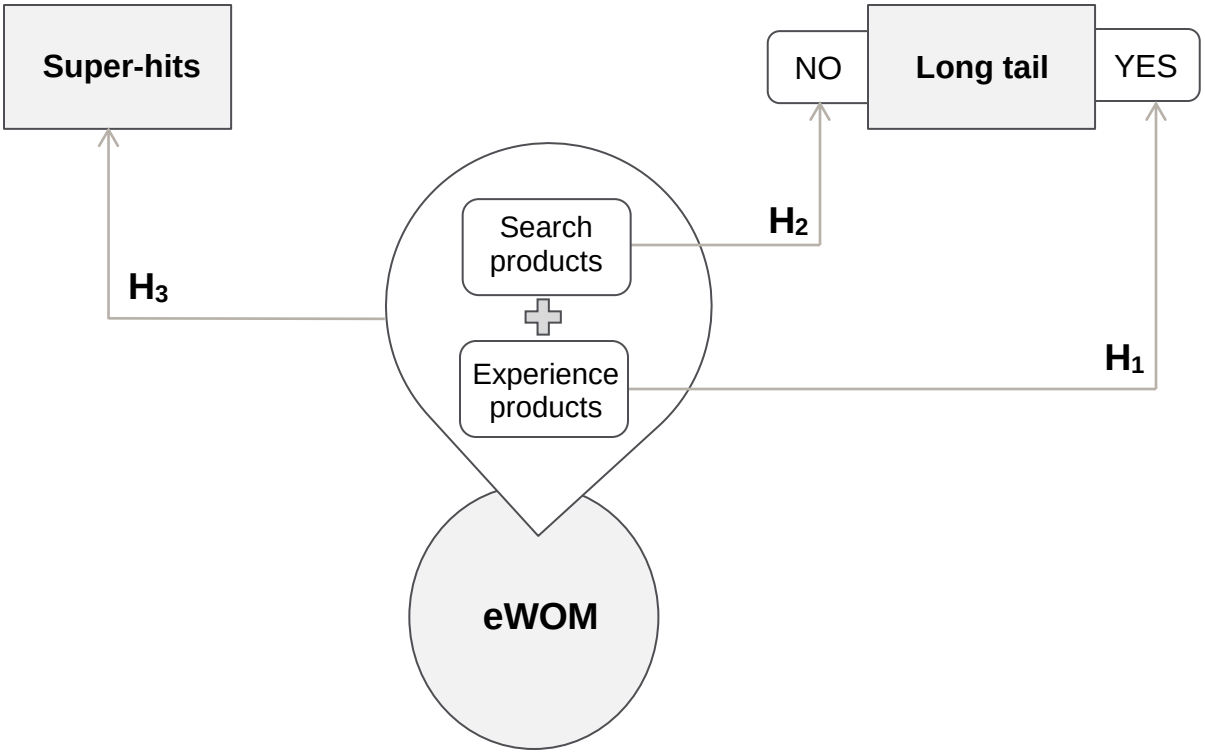


Figure 6 – Hypotheses development model based on theoretical framework

Part II: Research Design and Methodology

Since the literature about products across the long tail still does not discern or quantify the long tail of the sales distribution, one of the contributions of this Thesis is focusing on methods, such analytical models, to identify niche products within an eWOM community, providing a deeper understanding of the long tail phenomena. To this end, on the one hand, the first section of this second part explains the application of the aforementioned two methods for identifying the long tail based on the above literature reviewed. On the other hand, section no. 2 outlines an architectural framework and methodology to collect Big Data from an eWOM community containing user-generated content.

1 Finding the long tail across the data

1.1 Fitting the power-law to gathered data

Brynjolfsson, Hu and Smith (2010) made in their paper a review of the literature about the long tail. They found that the long tail could be defined and measured in three different ways: (1) Absolute long tail, (2) Relative long tail and (3) exponent based, *"because the ordinal rank to cardinal sales relationship often follows a power-law distribution, the exponent provides an indication of the relative importance of the head versus the tail of the sales distribution".* Particularly, the third one has been taken into account within this Thesis.

According to this line of thought and to McKelvey and Andriani (2005), power-law distributions appear under two conditions, when tension increases and when the cost of connections decreases. Since in the global economy the costs of connections are quickly decreasing and IT infrastructures dramatically reduce the cost of information transmission, power-law distributions are becoming even more prevalent (Aarstad 2014). This exploits the fact that many empirical quantities gather around a typical value, even the larger deviations are still only near a factor of two from the mean in either direction and thus the distribution can be satisfactory characterised by quoting just its mean and standard deviation (Clauset, Shalizi and Newman 2009). However, when values or events are independent, interactive, or even both not all distributions fit this pattern, such as the power-law distribution.

The mathematics of power-law cumulative distributions imply a power-law form when the distribution function is defined as:

$$P(x) = Cx^{-\propto} \tag{1}$$

Where $P(x)$ is the probability (frequency) that the variable takes the value x; $\propto$ the exponent of the distribution (with $\propto > 0$, otherwise when normalizing the function does not converge); x the variable to be analysed; and C a constant that depends on the type of event.

Taking logarithms on both sides of (1), it is observed that for a power-law

$$ln(x) = ln\, C - \propto lnx \ , \tag{2}$$

which means that in a graph with logarithmic scale, the relationship between $ln(x)$ and lnx is described by a straight line whose negative slope is $\propto$.

In practice, identifying power-law behaviour is difficult due to the large number of fluctuations that occur in the tail of the distribution. However, it is possible to infer it. In this regard, in many cases it is convenient to use the complementary cumulative distribution function (CDF) of a power-law distributed variable, which is denoted as

$$P(x) = \int_x^{\infty} p(x)dx = \left(\frac{x}{x_{min}}\right)^{-\propto+1} \tag{3}$$

Actually, there must be some lowest value x_{min} at which the power law is obeyed, and the statistics are only considered for x values greater than this value. Besides, the equation indicates that the probability of large events is very small and the probability of small events is high.

Within this Thesis, the method defined in the article written by Clauset, Shalizi and Newman (2010) will be considered to decide whether or not the data set follows a power-law distribution. These are their proposed steps:

1. Estimate $\propto$ using the maximum likelihood estimator for the $\propto$ scaling exponent

2. Find x_{min}using the goodness of fit value to estimate where the scaling region begins. The curve can follow a power-law on the right or upper tail, so above a given threshold x_{min}.

3. Calculate the goodness of the model using the goodness of fit given by the Kolmogorov-Smirnov statistic equation

$$D = max_{x \geq x_{min}} |S(x) - P(x)| \ , \tag{4}$$

   where $S(x)$ is the cumulative distribution function (CDF) of the data to be fitted

with $x \geq x_{min}$, and $P(x)$ is the CDF for the power-law model that best fits the data in the region $x \geq x_{min}$. The estimation of x_{min} is actually the value of x_{min} that minimizes the distance D. The distance D is calculated for the observed data set and the best-fit power law distribution computed as described by the authors. A *p-value* can be calculated to determine whether the value of D is too high. Besides, the *p-value* quantifies if the data sets are consistent with a power-law distribution based on goodness of fit. That is to say, if the *p-value* is smaller than 0.05, the power-law model can be firmly discarded. Nevertheless, if the *p-value* is near 0.05, then the power-law could be a plausible fit to the data.

1.2 Finding the optimal x_{min} based on the elbow criterion

Along this section a mathematical method to adjust the elbow of any curve with an exponential or power-law distribution is illustrated. In the proposed methodology, finding an optimum *"k"* value (x_{min}) is performed by the elbow criterion. Here, the elbow criterion plays a central role in the indication of the cut-off x_{min}. This cut-off happens on the graph of the within-cluster dispersion, as a function of the number of clusters, such that it generally diminishes as quantity of clusters increases. The decrease becomes sharper whenever the correct number of clusters is obtained. In this regard, the elbow criterion is based on choosing the values where such sharp decrease of the slope of the function was observed. Thus, a technique for getting such cut-off from all parts of data set –that implies a power-law distribution– is to calculate the tangent to the power-law curve at that point. Then, the slope of the function $P(x) = C\, x^{-\propto}$ at the point x_{min} means the slope of the tangent at the point x_{min}, where the sharp "elbow" is clearly visible in the graph.

The tangent line to the graph of a function $y = f(x)$ at the point $(x_0, f(x_0))$ on the graph is the unique straight line that passes through that point and has the same slope as the graph at that point. In other words, the slope of the tangent line at the point $(x_0, f(x_0))$ on the graph is defined as the value $f'(x_0)$ of the derivative at the point x_0.

The derivative of the function $y = f(x)$ at the point $x = x_0$, (denoted $f'(x_0)$,) is given by the limit

$$f'(x_0) = \lim_{x \to x_0} \frac{f(x) - f(x_0)}{x - x_0} \quad , \tag{4}$$

if the limit exists. In this case it is said that $f(x)$ is differentiable at x_0.

The idea is to find an "elbow" in the graph of the data set, that is to say, a point from which the eigenvalues are approximately equal. The criterion is to stay with a number of components that exclude those associated to small values and those of the approximately same size. When the values are practically equal means that there is barely variation. In this regard, the derivative gives the slope of the variation. Thus, when there is barely variation is the same as when the derivative is close to 0 (slope of a horizontal line). To capture more precisely the idea, the following Figure 7 illustrates the selection of the elbow criterion based on choosing the values where a sharp decrease in the slope of the tangent.

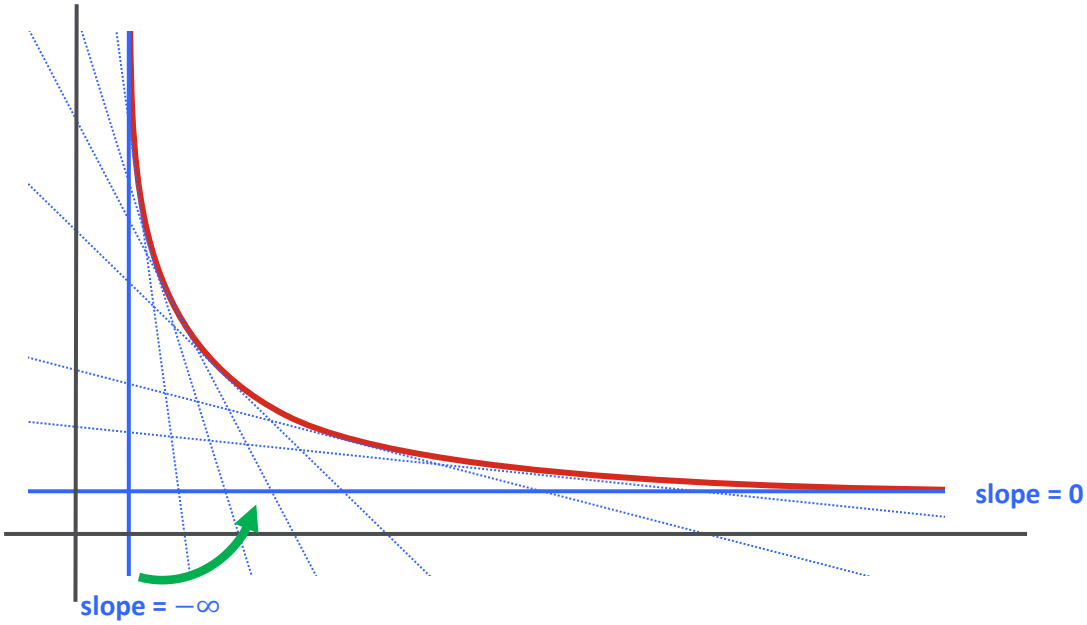


Figure 7 – Obtaining the elbow point (x_{min}) by solving combinations of the slope of the tangent

The green arrow in the graph indicates the evolution of the tangent along the curve. Moreover, the blue lines represent the initial and the final tangent. Then, the blue dotted lines are representing the variation of the tangent, meaning how the tangent begins and how it ends (the domain within which the tangent will be moving). The graph varies from $-\infty$ to 0 and the midpoint of this change is -1 (the slope of tangent – 45º), which is what it is sought to find the elbow. Besides, as the number of clusters increases, the eigenvalues decrease. In this sense, the point at which the drop in eigenvalues markedly decreases is the elbow point, which is the cut-off x_{min} point. Thus, the x_{min} is the point of average slope between the two asymptotes of the graph, which is supposed to be the point from which it starts decreasing very fast to decrease very slowly. Nevertheless, in some cases the elbow cannot always be unambiguously identified since either there is no elbow, or several elbows as shown (Kodinariya and Makwana 2013).

2 Gathering user-generated data across an eWOM community

Following the theoretical framework on data gathering it is clear that user-generated data can be obtained with those APIs specialized on web crawling provided by most social media services and largest media online retailers such as YouTube, Flickr or even Amazon (Manovich 2011). Those APIs provide an easy technique to obtain or scrape data. For example, through Amazon Web Services, developers can access product catalogue, customer reviews, site ranking and historical pricing. Nonetheless, they offer very poor functions such as the ones for search and acquisition since the content is produced without directly involving a person (Boyd and Crawford 2012). In some cases, it is interesting to obtain more information than the one provided by APIs, for instance to perform some filtering over the collected data of interest. For example, there are accounts that are actually bots, or even many users that are not active and can compromise the validity and reliability of the subsequent analysis. Those users cannot be removed unless some additional information is collected such as reputation, previous experience of users, etc. Moreover, in some cases APIs only provide a fraction

of all the available information worsening one of the benefits of social Big Data, which is the possibility of collecting the whole data instead of just a sample. This might be the case of Twitter APIs, which only makes available to typical researchers a roughly 10 per cent of public tweets (Boyd and Crawford 2012). Besides, it is important to mention that not all websites offer an API. Consequently, web scraping is a great alternative to grabbing the required data. So, within this section, a set of commands or methods are explained in order to retrieve all the data stored in an eWOM community that contains user-generated content.

The methods applied to data collection have involved two different steps: (1) data crawling from a web with user-generated content and (2) data storage in a Data Base.

Firstly, in order to crawl data from the web *Python* was used because it is a dynamic, portable and performing language combined with an open source web crawler framework called *Scrapy*. Although there are simpler *Python* alternatives and other open source scrapers in *Java*, *Ruby*, and *PHP*, *Scrapy* is a much better alternative since it is the most popular tool for web crawling written in *Python*, as well as it is simple and powerful, with plenty of features and possible extensions (Wang and Guo 2012). The scraping cycle went through the definition of several *items*, which are containers defined to contain the data to be collected from the page. Then, to crawl or scrape information several classes named *spiders* where programmed. *Spiders* define how a certain site will be crawled, including how to perform the crawl and how to extract structured data from their pages. To that end, the *spiders* define an initial list of URLs to download, how to follow links, and how to parse (analyse) the contents of pages to extract *items*. Obviously there is a huge amount of data in webs with user-generated content and the *spiders* provide access to useful and relevant information with the goal of browsing as many web pages as possible. Thus, the basic algorithm programmed here was fetching the web that contained all indexed pages that link all the information to gather. To that end, the method *parse_start_url()* was called. Such method contained a list of *URLs* where the spider began to crawl from. So, the first pages downloaded were those listed there and the subsequent *URLs* were generated successively from data contained in the start *URLs*. Then, different programmed *parse()* methods where in

charge of processing the response and returning scraped data and more URLs to follow. Those methods returned *item* objects. The main advantage of web crawlers is that they can also extract specific information while browsing the site. This can be done using *XPath* language, which can be easily integrated within *Scrapy.* The following Figure 8 illustrates some examples of how were programmed the methods *parse()* and also how the language *Xpath* was used.

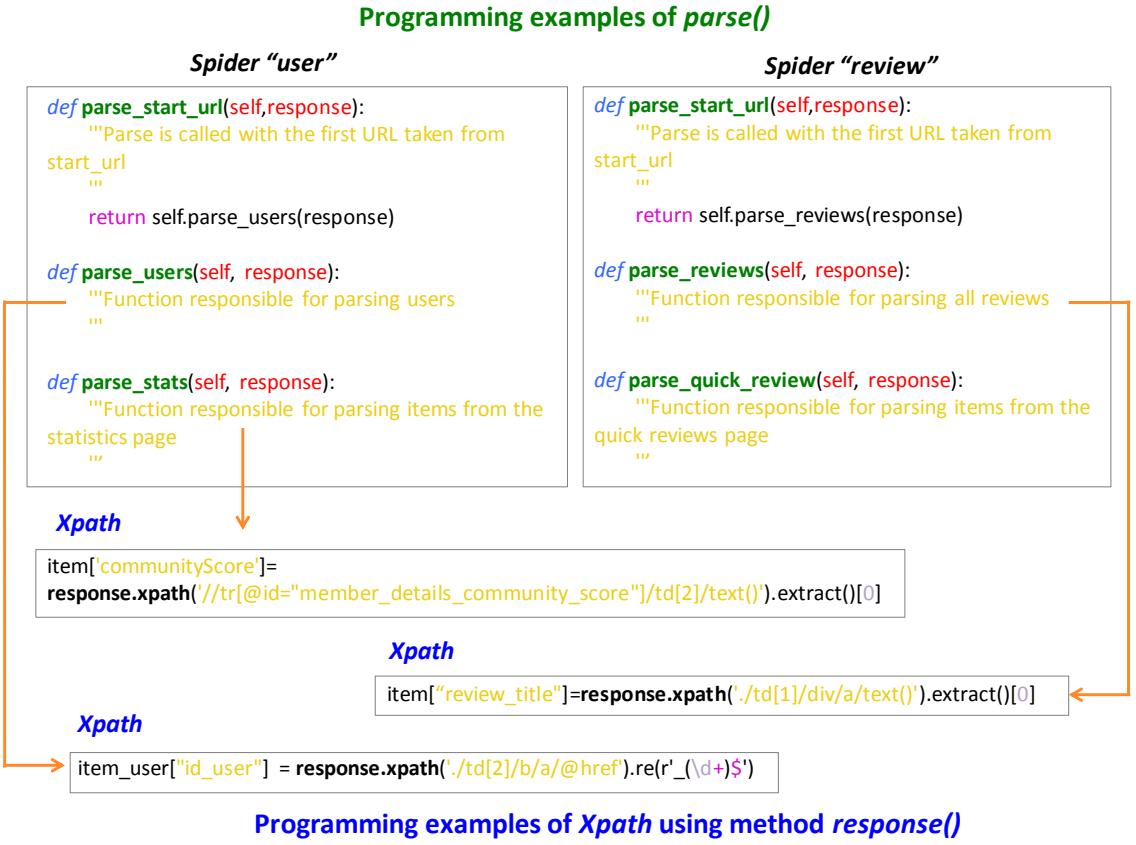


Figure 8 – Programming examples of methods parse() and response.xpath()

XPath is a language created for doing queries in *XML* content and it is used to turn an *XML* document into a hierarchical form to better organize information into a tree structure (Gottlob, Koch and Pichler 2003). Those selectors are applied to part of the source code of the web and perform data extractions from the *HTML* source using expressions to navigate a document and extracting information using the library *Lxml*. Using *XPath*, researchers can select whatever content they consider meaningful in the

context of their on-going research, without the limitations of APIs, restricted to the information decided by the API provider.

Secondly, all of these steps have involved storing information in a database. Hence, the items retrieved from the spiders were persisted to a relational database due to the data-intensive storage and in order to have an organized collection of data. A relational database consists of one or more tables that have relationships, or links, between them, either in a one-to-one or a one-to-many relationship. The relational database systems are generally efficient since different tables from which information has to be linked and extracted can be easily manipulated by querying data in the form in which it is desired. The database was designed in *MySQL* because it is efficient, ubiquitous and has an open-source engine available for all major platforms (Vicknair, et al. 2010). Several tables were created containing all dataset consisting of meta-information about user-generated data. Besides and in order to look at the data and analyse it in different ways, it is possible to apply the *SQL querying language*.

Finally, because the web is constantly changing and indexing is done periodically, proper data extraction also requires solid data validation and error recovery to handle data extraction failures as well as exceptions from the data storage. Thus, crawl monitoring and diagnostics dealing with exceptions were also created.

To conclude the following diagram in Figure 9 illustrates the afore-described process and the steps performed to collect and to store the data for this Thesis.

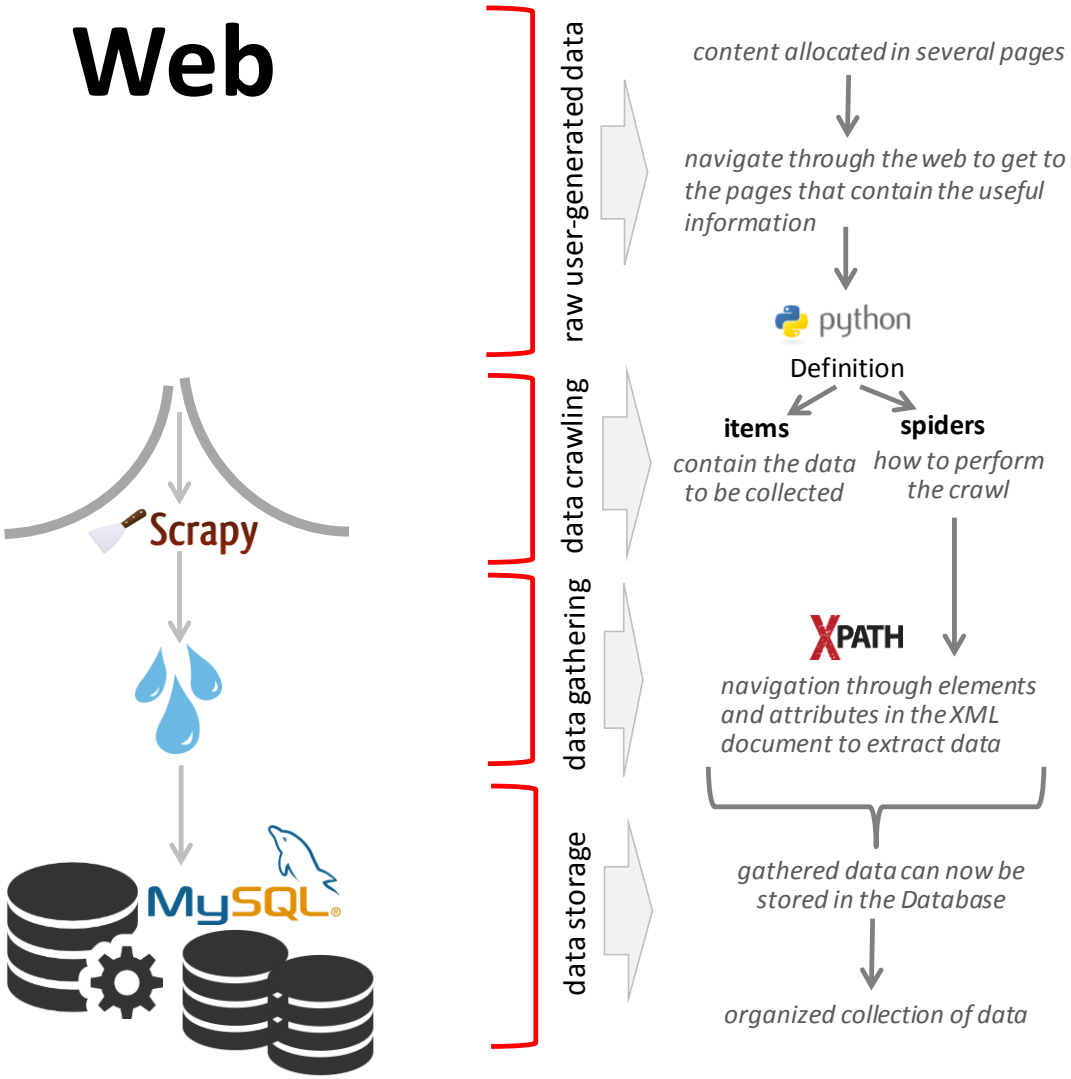


Figure 9 – Process of collecting user-generated data from a web

Part III: Case Study and Data Collection

This third part describes in depth in the first section a particular eWOM community that has user-generated content, Ciao UK. Then, along the second section it is outlined the process of the implementation of a web crawler to gather data from the portal Ciao UK using the web crawling approach from both perspectives the social science and the computing science discipline. Some experimental results in terms of time, size and database design are also included.

1 Case of Study: Ciao UK

Data collection has involved accessing data from the website Ciao, which is a mass eWOM community where registered users can make reviews about any product or service. Ciao is one of the largest eWOM communities in Europe available in local-language versions, with more than 1.3 million members that have written more than 7 million reviews on 1.4 million of products (Arenas Márquez, Martínez-Torres and Toral 2014, Olmedilla, Martinez-Torres and Toral 2016). Basically, the website Ciao is structured in three main sections: *reviews*, *shopping* and *"My Ciao"* together with their corresponding subsections as illustrated in the following *Table 2*. The sections *reviews* and *shopping* are organized through categories of products and services. Principally, there are 28 main categories established by Ciao as well as subcategories created by registered users whenever they post and share reviews about any product. The section *reviews* also contains all the reviews, video-reviews posted by the users and the questions with the concerning a user have about a specific review. The section *shopping* has the top-10 list of more sell and rated products and the top-seller charts. The web also contains a member webpage section named *"My Ciao"* for each registered user that assembles all of the information that is relevant to Ciao members. It not only contains community and system announcements alerting members to new features and possible scheduled site downtimes, but also many useful guidelines documents offering advice on how to write reviews and comments, give ratings and use the "circle of trust".

Table 2 – 3-section structure of web Ciao UK

Reviews	Shopping	My Ciao
Product main category	Product main category	User's webpage
Product subcategory	Product subcategory	Member Centre
Latest reviews	Top 10 list of products	User's announcements
Latest questions	Ciao top-seller charts	User's guestbook
Latest videos	Recent products	User's statistics

Ciao is available free of charge to users, who can register on the web. In order to create an account the users have to provide some data about themselves (although it is not mandatory) such as first name, gender, age or country. Likewise, they have several

scopes of activity as illustrated in Figure 10: (1) create a review, (2) rate a review, a product or another user for the benefit of other consumers and (3) trust other registered users.

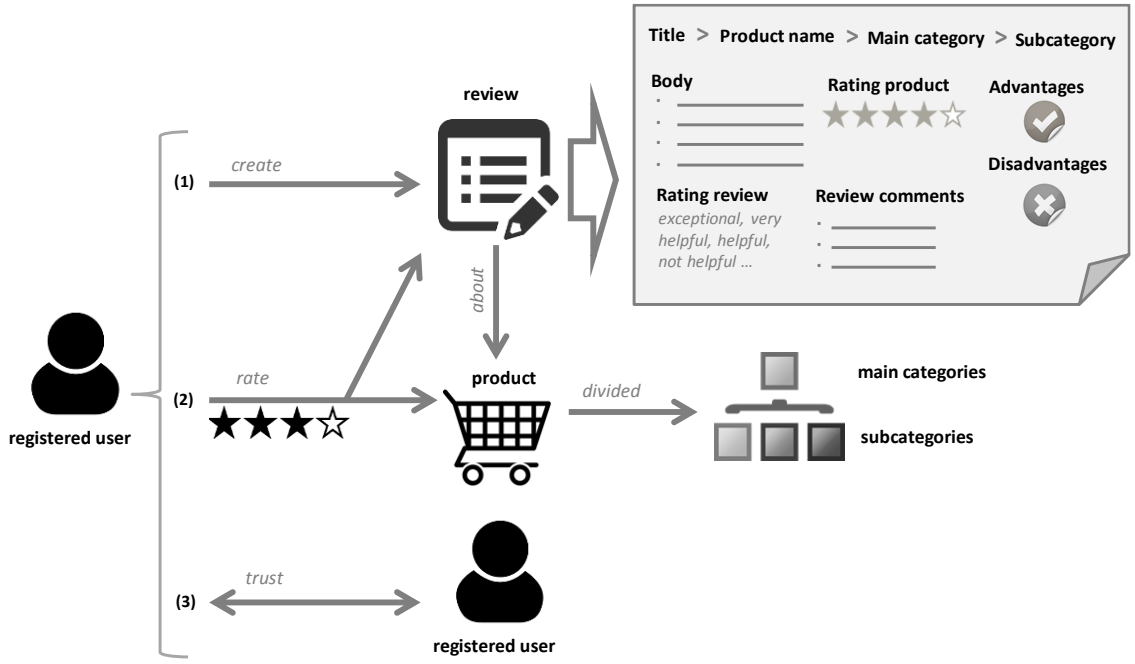


Figure 10 – Scopes of activity within Ciao for a user

Firstly, when creating a review some fields to fill are required such as the title, the name of the product the review belongs to, the body with the user's opinion about the product or the advantages and disadvantages about the product reviewed. The standard review must be at least 120 words long and should aim to give readers an idea of what the product in question is like. Moreover, the users can only write reviews of products that are listed in the Ciao product categories. Secondly, a user can rate either a review, or a product or another user. To rate a review, the user has to choose one of the six options listed beneath the review to describe how useful he or she found it: exceptional, very helpful, helpful, somewhat unhelpful, not helpful and off topic. Besides, the user can score products using qualitative ratings by giving it from 1 to 5 stars depending on how satisfied he or she is with it. Thirdly, the users are able to join their own *"circle of trust"* whenever they consider another registered user's reviews are consistently interesting and helpful. A user can invite up to 100 users to join his or her *circle of trust* but an

unlimited number of users can choose to trust him or her. Any user who earns another user's trust is awarded community points for having done so, which influences the weighting given to his or her review ratings, and determines how visible the user is on the website. Moreover, the registered user's activity is traced including a record of all of the actions that he or she has performed as shown in the following Table 3. Some examples, among others, are the date of his/her first/last review, how many reviews has he/she received from other users, the users included in his/her circle of trust, etc.

Table 3 – Record of all the actions a user has performed

	Information about performed activities
Registered user	Status (online/offline)
	Date of since when is a member
	Date of the first review
	Date of the last review
	Reviews written
	Comments written of a review
	Comments received of a review
	Ratings given
	Ratings received
	Members who trust the user
	Members the user trusts
	Community score

2 Gathering user-generated data

Considering the content of the web and taking into account the data of interest with a focus on social science research the basic algorithm programmed to crawl the web was fetching the web that contained all indexed pages that link all users, reviews, products, ratings and circle of trust belonging to a category within Ciao. For such purpose data collection has required accessing to the *Member Centre* page of Ciao, where all the information about registered users is presented. The following Figure 11 structures an example of the subsequent-detailed process of data gathering.

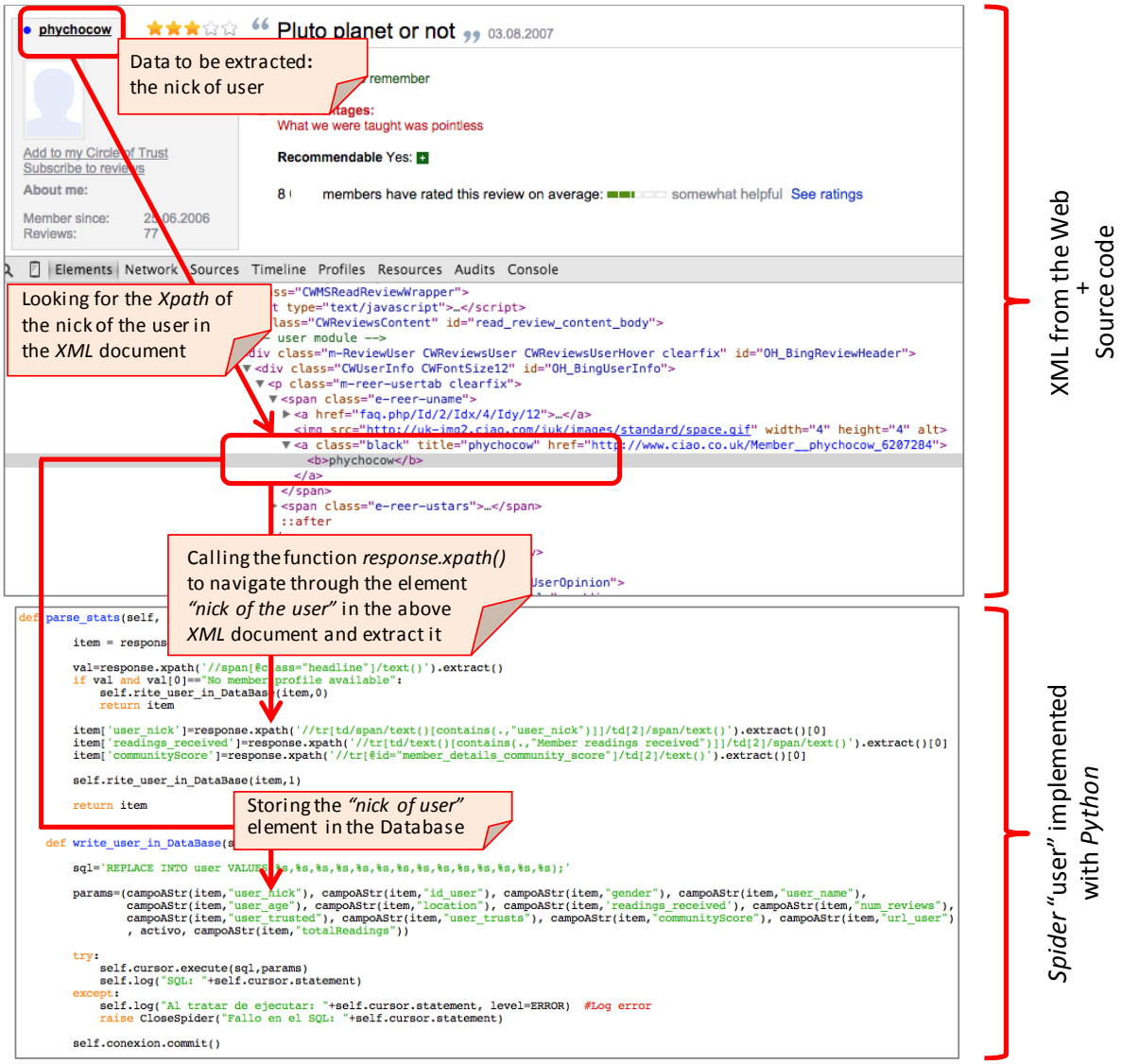


Figure 11 – Structure of the process of data gathering of "nick of user"

Firstly, a list of users was collected in order to obtain the more important pages rapidly. To that end, a *spider* or crawler that follows the hyperlink structure of the users' webpages has been developed using *Scrapy* with *Python*. Principally, the crawler browses the website of the user, storing a list of users containing their nick, id, name, gender, location or *URLs* among other data. This has provided a fast way of maintaining an index of the web through the id and *URLs* of the users that can be queried for updates. Secondly, a link was made from the list of users in order to collect the rest of the data (reviews, products, ratings, etc.). For this purpose several spiders were

programmed using *XPath* language calling the function *response.xpath()* from the *Scrapy* base library to navigate through elements and attributes in the *XML* document of each webpage (see Figure 8). This method was used for selecting nested data and to extract all the textual data the selector *extract()* was called or the selector *re()* for extracting data using regular expressions. In this case most of programmed *XPath* selectors have selected the link that contained the text *'Next Page'* since the majority of the pages to extract were linked to other pages through this link. Finally and once all the spiders were programed, two functions for performance evaluation measures were created. The first function was used to capture errors and when a specific *URL* generated an exception, it was stored to an error table in the database. The other function was a list of *URLs* corresponding to the pages that were already downloaded with the spider without any error.

Gathering big data from a web site can be a time-consuming task, so programmed algorithms should be fast enough to save time. For instance, within Ciao there is a huge number of information sources as well as different levels of accessibility to its user generated-content, which presents a complex information gathering control problem. Furthermore, the scale of the dataset is very large – there are about 45 thousand registered users in Ciao UK – which has meant the crawling procedure has taken relatively long time. Nevertheless, in spite of such a time consuming and complex process the website was completely crawled. During this process, the downloading speed has fluctuated due to exceptions and power failures as illustrated in Table 4, which shows the duration of active data downloading of each spider represented in the first column. For instance, capturing the data from the ratings of a review has been done in three steps due to the appearance of exceptions and errors from things like validating the extracted data, removing duplicated items, storing in a database, etc. which has slowed down the gathering process. Another reason of such delay was that extracting data from the ratings required accessing content from not only a webpage but from six different pages (the six options of rating: *exceptional, very helpful, helpful, somewhat helpful, not helpful, off topic*) and implement their six link extractors in order to get to the pages that contain the useful information. Conversely, the amount of time that has taken the data gathering of the user's circle of trust was insignificant in comparison with the

rest of the extractions. It was because there was only one link extractor for the spider, not all the users in Ciao have a circle of trust and storing only two fields in the database was fast.

Table 4 – Time spent on data gathering

Data gathered	Start date	End date	Duration	% of time
users	6/4/15 - 14:15	6/4/15 - 20:42	6,45 hours	1,08 %
reviews, products and categories	7/4/15 - 20:20	8/4/15 - 1:04	4,73 hours	0,79 %
user's circle of trust	8/4/15 - 8:17	8/4/15 - 8:29	0,20 hours	0,03 %
review ratings 1	22/4/15 - 17:36	16/5/15 - 23:09	24 days and 5,55 hours	97,66 %
review ratings 2	17/5/15 - 21:37	18/5/15 - 0:02	2,42 hours	0,41 %
review ratings 3	18/5/15 - 10:05	18/5/15 - 10:13	0,13 hours	0,02 %

Ciao provides a good example in the context of a statistical analysis due to the variety of information and knowledge that contains. This data of course has to be processed, stored, analysed and visualized to have any meaning. Actually, the key of translation between gathered data and posterior structured data suitable for analytics lies on well-defined data characterisations (often in tables) stored in relational databases. Therefore and as aforementioned, a relational database was designed in this paper. It is composed of several tables and ordering schemes, which gives the possibility of tracking almost everything stored inside. As can be observed, the following relational model depicted in Figure 12 illustrates the database model based on all the gathered data with a representation in terms of tuples, grouped into relations.

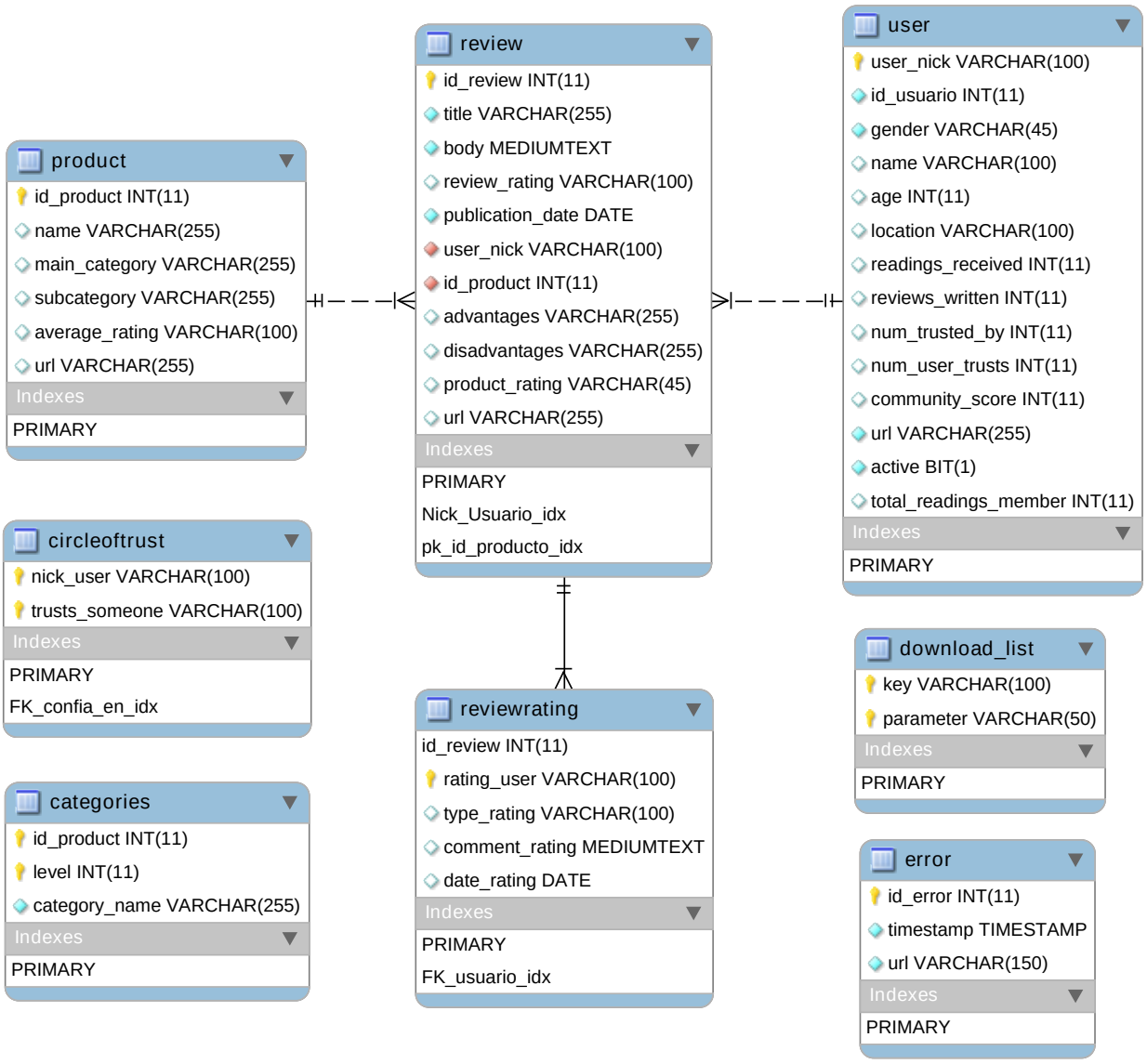


Figure 12 – Relational model of the database

This model organizes data into eight tables representing each item: type of users, products, reviews, circle of trust among the users, ratings of the reviews, categories of products and two tables that store errors and exceptions resulting form the data gathering. The represented relations among the tables contain a unique key for each row. For example, the table that describes a *product* with columns for id of product *(unique key)*, name, category, rating, and so forth. Another table describes a *review*: id, title, body, rating of review, date, user nick, id of product *(foreign key)*, advantages, disadvantages, and so forth. Because each row in the table *product* has its own unique

key (id of product), rows in the table *product* can be linked to rows in the table *review* by storing the unique key (id of product) of the row to which it should be linked, where such unique key is known as a "foreign key". Furthermore, with the *SQL querying language* applied to the database it is possible to retrieve the data based on specific criteria. For instance, if anybody would like to know how many users have rated another user the following query should be written:

```
select r.user_nick, rr.*
from review r join review rating rr on rr.id_review=r.id_review;
```

Additionally, the next Table 5 indicates the size of the complete database, which are 760 megabytes as well as the size in rows of all the data comprised in each table of the database. As can be observed the table *review rating* is the one with the higher number of rows since for each review all the users can rate up to 6 types of evaluations. Otherwise the table circle of trust has only 8.356 rows, which corresponds with the same number of users. This means that not every user has to have “trusters” thus a circle of trust.

Table 5 – Size of data comprised in the database

Stored data	Size	
Table “users”	44.352 rows	760 Megabytes
Table “reviews”	105.918 rows	
Table “products”	68.650 rows	
Table “categories”	283.240 rows	
Table “review ratings”	3.444.316 rows	
Table “circle of trust”	8.356 rows	

Part IV: Analysis of Results and Discussion

The aim of this fourth part is concerned with presenting the analysis of the data. Descriptive results are reported on section 1 based upon the methodology applied to gather online data. All the results state the findings of this Thesis arranged in a logical sequence without bias or interpretation. Then, along section 2 the results are used to validate the postulated Hypotheses. Finally, section 3 illustrates the discussion to interpret and describe the significance of the results in light of what was already known about the research problem being investigated and to explain any new insight about the problem after the results have taken into consideration. To this end the discussion is connected to the Theoretical Framework by way of the research hypotheses and the literature reviewed.

1 Descriptive results

A list of categories and subcategories of the posted reviews has been extracted for each registered user using the above-described crawler (see Part III - Section 2). The gathered result is shown in Figure 13, which illustrates the distribution of subcategories created by the users who have posted reviews over the 28 main categories distinguished by Ciao.

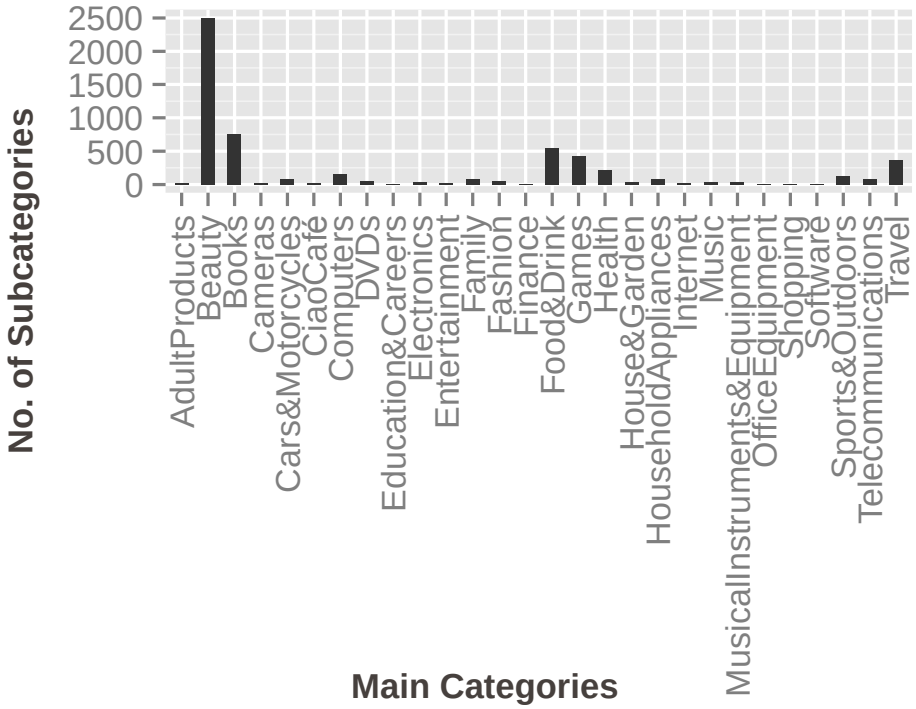


Figure 13 – Distribution of posted reviews for the 28 Main Categories

As it can be observed the number of reviews is not evenly distributed over categories. The category Beauty is the one gathering almost half of the total number reviews (about 2500), whereas other categories such as Cameras or Software only have a small number of reviews.

On the one hand, Table 6 describes in detail the long tail parameters of the 28 main categories contained in Ciao UK according to the power-law adjustment. The first and the second column show the x_{min} and the $\propto$ respective values of the fitted power-law distribution. The third column corresponds to the goodness of fit, the *p-value*, given by

the distance D displayed in the fourth column. The fifth column shows the total length of the distribution, and the sixth column shows the length of the long tail. Such length has been calculated as the number of subcategories having a number of reviews below the x_{min}, that is, the number of subcategories that are not part of the fitted power-law distribution. The five lasts columns represent the areas and their corresponding percentage calculated from the x_{min}. The values in x_{min} illustrate the frequency calculated to fit the power-law distribution, that is, the number of reviews from which number of reviews the fitting power-law distribution is made. Correspondingly, the categories below the x_{min} refer to the products characterized by the long tail. Likewise, by observing the $\propto$ exponent it is possible to discern whether a category follows a power-law distribution. Besides, the fitting has an associated *p-value*. For each category were the calculated *p-value* is considerably lower than 0.05, the null hypothesis is rejected, which means that such category does not to follow a power-law distribution. But if the resulting *p-value* is greater than 0.05, that means that the null hypothesis cannot be rejected and hence the distribution is likely to follow a power-law distribution. The area 1column denotes the area contained to the left of the cut-off point of the x_{min}, which are the dominating products or best sellers. The area 2 column embodies the long tail, that is, the region to the right of the cut-off point of the $\mathrm{x_{min}}$.

On the other hand, Table 7 describes the long tail parameters of the 28 main categories according to the elbow method. The first column displays the x_{min} values, which correspond to the turning point, that is, the point from which the function goes from decreasing very fast to decreasing very slowly. Accordingly, the function is traversed from the right to the left, from x_{max} to -1, until the first point/value that meets the searching condition. This is the point of average slope between the two asymptotes of the graph. The second column shows the total length of the distribution. The third column is the length of the long tail, which has been calculated as did with the power-law adjustment. Also in this case the five last columns represent the areas and their corresponding percentage calculated from the x_{min} in the same way as clarified before. Similarly, the Area 1 column is calculated to the left of the cut-off point of the x_{min}, and the Area 2 column to the right of the cut-off point of the x_{min}.

Table 6 – Long tail parameters of the 28 main categories of Ciao UK according to the power-law method

MAIN CATEGORIES	x_{min}	$\propto$	*p-value*	*D*	*Total length*	*Length tail*	*Area 1*	*Area 2*	*Total area*	*% Area 1*	*% Area 2*
Adult Products	3	2,97	0,602	0,0928	16	11	14,5	17	31,5	46,03	53,97
Beauty	3	2,02	0	0,0479	2505	1651	7799,5	2695	10494,5	74,32	25,68
Books	7	2,20	0,011	0,0634	751	571	3261	1267	4528	72,02	27,98
Cameras	4	1,63	0,231	0,1444	19	2	524	10,5	534,5	98,04	1,96
Cars & Motorcycles	3	1,84	0,047	0,1017	80	24	648	73	721	89,88	10,12
Ciao Café	18	1,62	0,085	0,2271	15	3	1388,5	24,5	1413	98,27	1,73
Computers	5	1,74	0,002	0,1059	155	86	1812,5	173	1985,5	91,29	8,71
DVDs	5	1,78	0,435	0,1018	46	24	555	54	609	91,13	8,87
Education & Careers	8	1,53	0,098	0,4191	4	0	234,5	0	234,5	100	0
Electronics	20	2,21	0,683	0,1094	45	25	1099	139,5	1238,5	88,74	11,26
Entertainment	18	2,72	0,940	0,1096	23	13	350	82,5	432,5	80,92	19,08
Family	57	3,50	0,520	0,1335	86	73	1047	809	1856	56,41	43,59
Fashion	2	2,09	0,065	0,1046	56	20	157	44,5	201,5	77,92	22,08
Finance	118	3,24	0,590	0,3145	6	3	384,5	80	464,5	82,78	17,22
Food & Drink	2	1,81	0,086	0,0351	551	262	3174	442,5	3616,5	87,76	12,24
Games	9	2,13	0,171	0,0601	421	339	2213	794	3007	73,59	26,41
Health	17	2,39	0,064	0,0778	218	189	1069	519	1588	67,32	32,68
House & Garden	19	2,01	0,683	0,0993	39	14	1675	91,5	1766,5	94,82	5,18
Household Appliances	71	3,50	0,703	0,1174	75	61	1354	1219	2573	52,62	47,38
Internet	13	1,50	0,048	0,2164	17	2	3595	20	3615	99,45	0,55
Music	8	3,50	0,179	0,1637	38	27	100,5	66,5	167	60,18	39,82
Musical Instruments & Equipment	8	2,70	0,711	0,1171	30	25	57,5	69,5	127	45,28	54,72
Office Equipment	4	1,52	0,161	0,2942	6	0	207,5	0	207,5	100	0
Shopping	11	1,50	0,458	0,3526	3	0	408,5	0	408,5	100	0
Software	7	1,52	0,538	0,2773	5	1	236,5	4,5	241	98,13	1,87
Sports & Outdoors	5	2,21	0,036	0,0999	130	89	468	177	645	72,56	27,44
Telecommunications	71	2,22	0,020	0,1532	80	72	1322,5	250	1572,5	84,10	15,90
Travel	3	1,98	0,157	0,0484	372	198	1856	377	2233	83,12	16,88

Table 7 – Long tail parameters of the 28 main categories of Ciao UK according to the elbow method

MAIN CATEGORIES	x_{min}	Total length	Length tail	Area 1	Area 2	Total area	% Area 1	% Area 2
Adult Products	2	16	9	20	11,5	31,5	63,49	36,51
Beauty	2	2505	1265	8459	2035,5	10494,5	80,60	19,40
Books	5	751	522	3497	1031	4528	77,23	22,77
Cameras	4	19	2	524	10,5	534,5	98,04	1,96
Cars & Motorcycles	2	80	15	689,5	31,5	721	95,63	4,37
Ciao Café*	8	15	2	1401,5	11,5	1413	99,19	0,81
Computers	4	155	78	1837	148,5	1985,5	92,52	7,48
DVDs	4	46	23	569,5	39,5	609	93,51	6,49
Education & Careers	8	4	0	234,5	0	234,5	100,00	0,00
Electronics	24	45	28	1009	229,5	1238,5	81,47	18,53
Entertainment	23	24	17	269,5	163	432,5	62,31	37,69
Family	65	86	75	860	996	1856	46,34	53,66
Fashion	2	56	20	157	44,5	201,5	77,92	22,08
Finance	149	6	4	251	213,5	464,5	54,04	45,96
Food & Drink	2	262	551	3174	442,5	3616,5	87,76	12,24
Games	11	421	352	2078	929	3007	69,11	30,89
Health	23	218	197	949	639	1588	59,76	40,24
House & Garden	10	39	12	1689,5	77	1766,5	95,64	4,36
Household Appliances	82	75	64	1127,5	1445,5	2573	43,82	56,18
Internet	10	17	1	3606,5	8,5	3615	99,76	0,24
Music	5	38	26	115	52	167	68,86	31,14
Musical Instruments & Equipment	9	30	26	49	78	127	38,58	61,42
Office Equipment	8	6	1	201,5	6	207,5	97,11	2,89
Shopping	107	3	1	349,5	59	408,5	85,56	14,44
Software	7	5	1	236,5	4,5	241	98,13	1,87
Sports & Outdoors	7	130	101	412	233	645	63,88	36,12
Telecommunications	83	80	73	1245,5	327	1572,5	79,21	20,79
Travel	2	372	137	1975,5	257,5	2233	88,47	11,53

In order to evaluate whether among the above gathered data distributions appears a long tail the choice of a decision-making rule is needed. In decision theory, a decision-making rule is a function, which maps an observation to an appropriate action (Peterson 2009). In statistics, a decision-making rule is a formal rule that states, based on the obtained data, when to reject the null hypothesis H_0. Usually, it specifies a set of values based on the collected data, which are contradictory to the null hypothesis H_0.

To construct the decision-making rule the joint-ratio 80:20 on the above-presented data has been applied. The null hypothesis H_0 "there is a long tail among the data distribution according to its percentage of area" has been settled. Firstly, a cut-off value – the most extreme value – that marks the starting point of the set of values that comprise the rejection region has been chosen. To this end, from both methods (power-law and elbow) the maximum (61.41%) and the minimum (0%) sample within the column *Area 2 %* (see Table 6 and Table 7) has been selected to formulate the rejection region and corresponding decision rule. A change of scale (from 0-100 to 0-61.41) was required in order to calculate the rejection region. The equation for the line through the points is used for this change of scale. For the next null hypothesis H_0 "there is a long tail among the data distribution according to its tail length", a new cut-off value has been chosen. Also from both methods the maximum (91.25%) and the minimum (0%) sample within the column *Length tail* (see Table 6 and Table 7) has been selected. Likewise, a change of scale (from 0-100 to 0-91.25) was required. After determining the rejection region, the decision-making rule compares the data in Table 6 and Table 7 as indicated in the Figure 14.

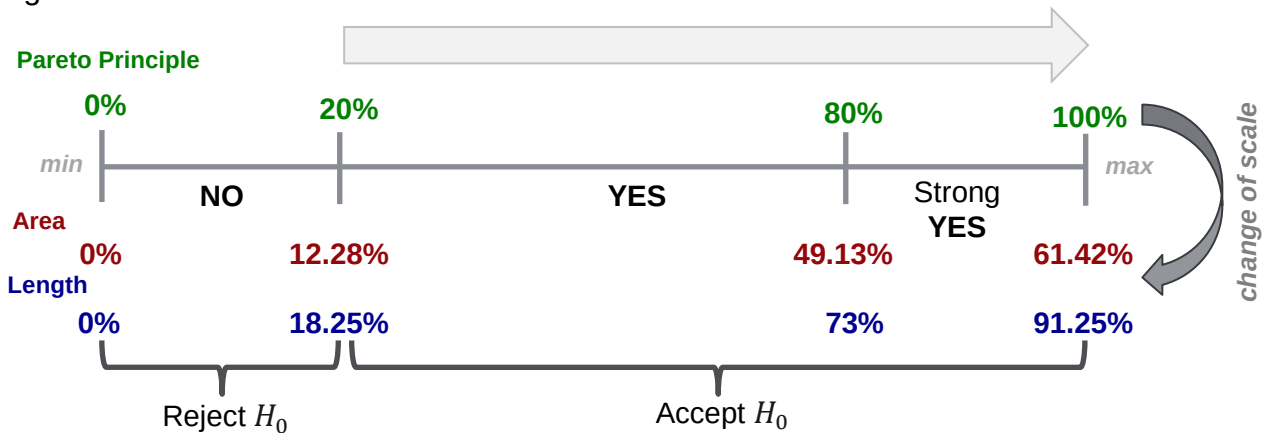


Figure 14 – Construction of the decision rule and calculation of the rejection region

Table 8 – Validity of the long tail presence through the consistency of the decision rules

MAIN CATEGORIES	Areas		Long tail	Length of tail		Long tail	COMPARISON
	Power-law	Elbow		Power-law	Elbow		Areas vs. Tail
Adult Products	Strong Yes	Yes	YES	Yes	Yes	YES	**True**
Beauty	Yes	Yes	YES	Yes	Yes	YES	**True**
Books	Yes	Yes	YES	Strong Yes	Yes	YES	**True**
Cameras	No	No	NO	No	No	NO	True
Cars & Motorcycles	No	No	NO	Yes	Yes	YES	False
Ciao Café	No	No	NO	Yes	No	Uncertain	**Uncertain***
Computers	No	No	NO	Yes	Yes	YES	False
DVDs	No	No	NO	Yes	Yes	YES	False
Education & Careers	No	No	NO	No	No	NO	True
Electronics	No	Yes	Uncertain	Yes	Yes	YES	**Uncertain***
Entertainment	Yes	Yes	YES	Yes	Yes	YES	**True**
Family	Yes	Strong Yes	YES	Strong Yes	Strong Yes	YES	**True**
Fashion	Yes	Yes	YES	Yes	Yes	YES	**True**
Finance	Yes	Yes	YES	Yes	Yes	YES	**True**
Food & Drink	Yes	Yes	YES	Yes	Yes	YES	**True**
Games	Yes	Yes	YES	Strong Yes	Strong Yes	YES	**True**
Health	Yes	Yes	YES	Strong Yes	Strong Yes	YES	**True**
House & Garden	No	No	NO	Yes	Yes	YES	False
Household Appliances	Yes	Strong Yes	YES	Strong Yes	Strong Yes	YES	**True**
Internet	No	No	YES	No	No	NO	False
Music	Yes	Yes	Yes	Yes	Yes	YES	**True**
Musical Instruments & Equipment	Strong Yes	Strong Yes	YES	Strong Yes	Strong Yes	YES	**True**
Office Equipment	No	No	NO	No	No	NO	True
Shopping	No	Yes	Uncertain	No	Yes	Uncertain	**Uncertain**
Software	No	No	NO	Yes	Yes	YES	False
Sports & Outdoors	Yes	Yes	YES	Yes	Strong Yes	YES	**True**
Telecommunications	Yes	Yes	YES	Yes	Strong Yes	YES	**True**
Travel	Yes	No	Uncertain	Yes	Yes	YES	**Uncertain***

The above Table 8 summarizes the validity and consistency of the decision-making rules introduced under the perspectives of the area and the length of tail. It can be observed that there are some cases denoted as “Uncertain”. Those cases are the pair of values from both methods (power-law and elbow) which position corresponds to both extremes of the region: accepting H_0 and rejecting H_0. Among the “Uncertain” cases, there are also some highlighted with an asterisk (*). This means that the value either is positioned more closely to the direction of the rejection region in the case of the category Ciao Café or more closely to the direction of the accepting region in the cases of Electronics and Travel.

Once the set of values for which the null hypothesis H_0 is rejected or accepted is known, this information is considered to obtain the outcome through a comparison of values. Accordingly, whether H_0 is accepted for both methods (power-law and elbow) and for both perspectives (area and length of tail) the corresponding category is very likely to be characterized as a long tail. This can be appreciated in 15 of the 28 main categories (*Adult Products, Beauty, Books, Entertainment, Family, Fashion, Finance, Food & Drink, Games, Health, Household & Appliances, Music, Musical Instruments, Sports & Outdoors, Telecommunications*), which are depicted on the following graphs.

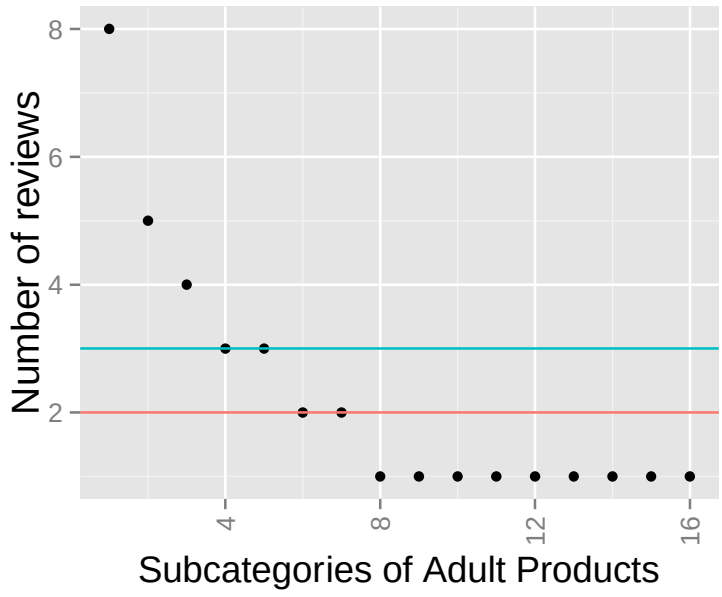


Figure 15 – Distribution of reviews for the main category “Adult Products”

Figure 16 – Distribution of reviews for the main category "Beauty"

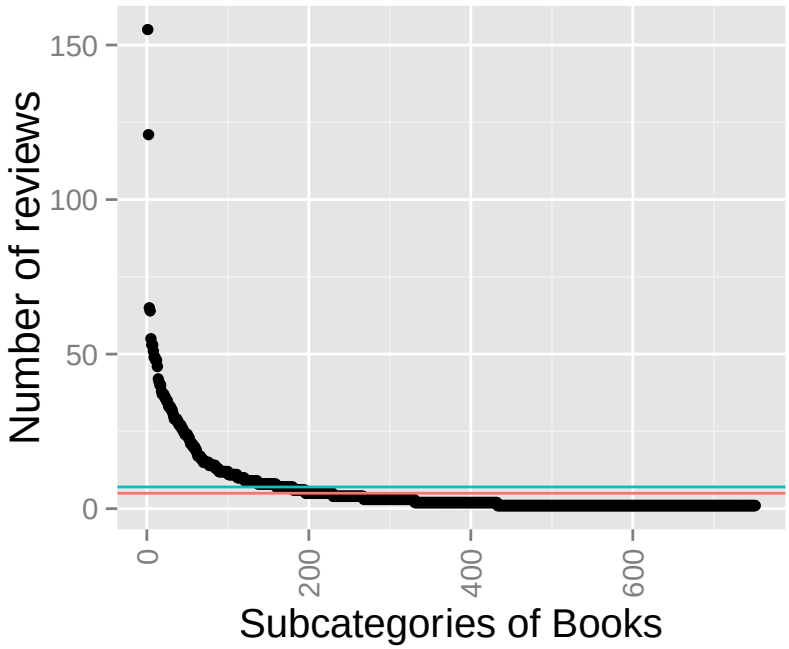


Figure 17 – Distribution of reviews for the main category "Books"

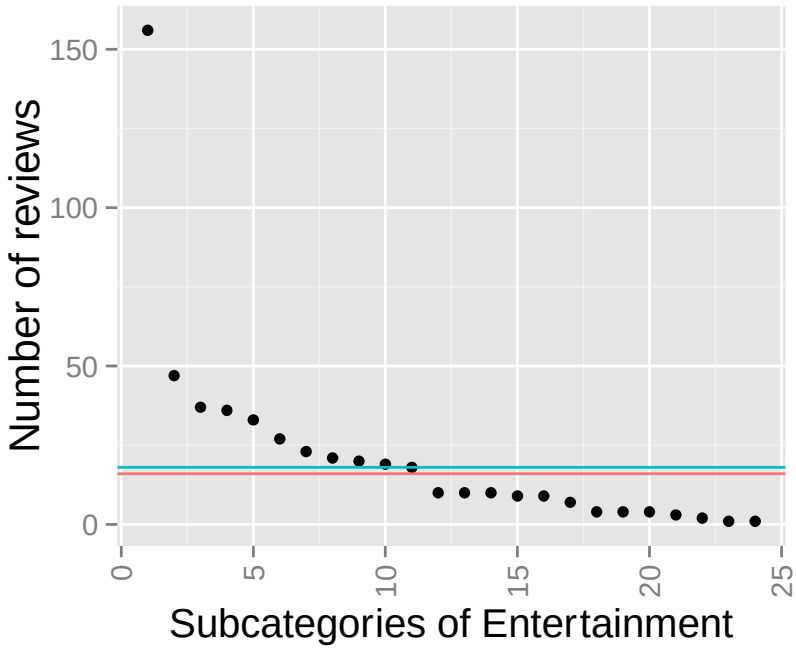


Figure 18 – Distribution of reviews for the main category "Entertainment"

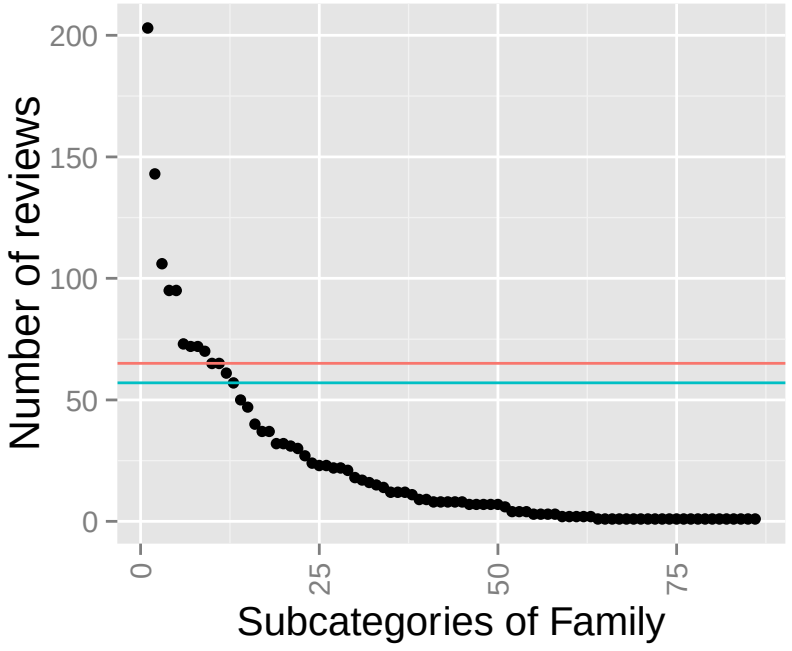


Figure 19 – Distribution of reviews for the main category "Family"

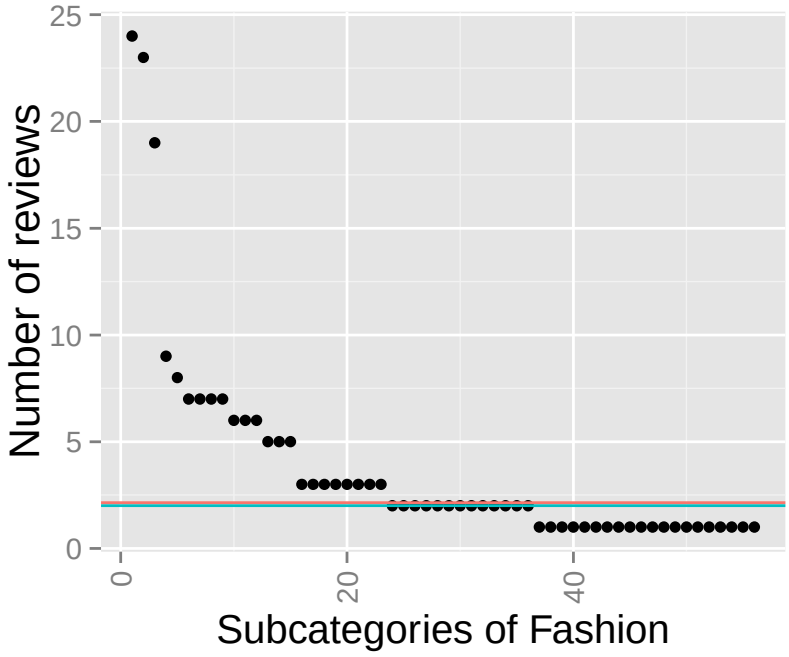


Figure 20 – Distribution of reviews for the main category "Fashion"

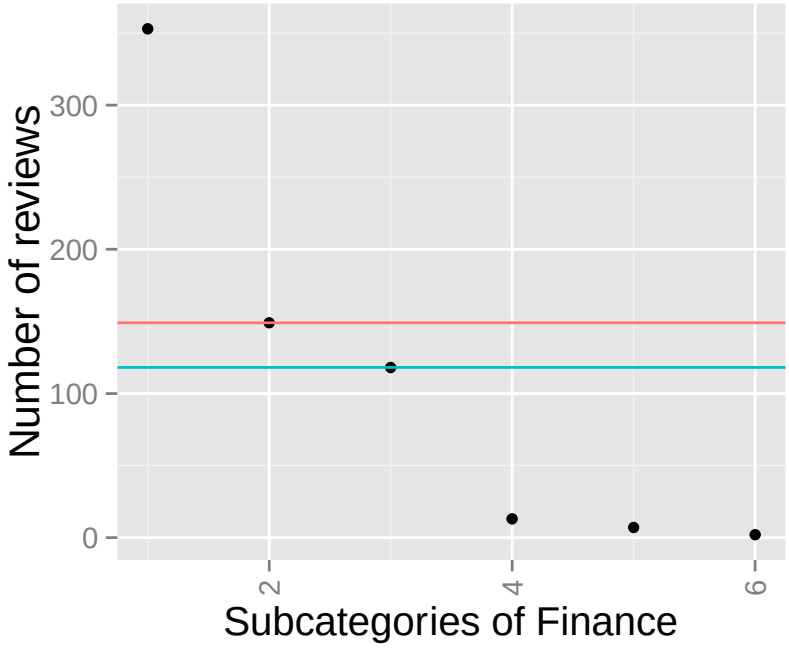


Figure 21 – Distribution of reviews for the main category "Finance"

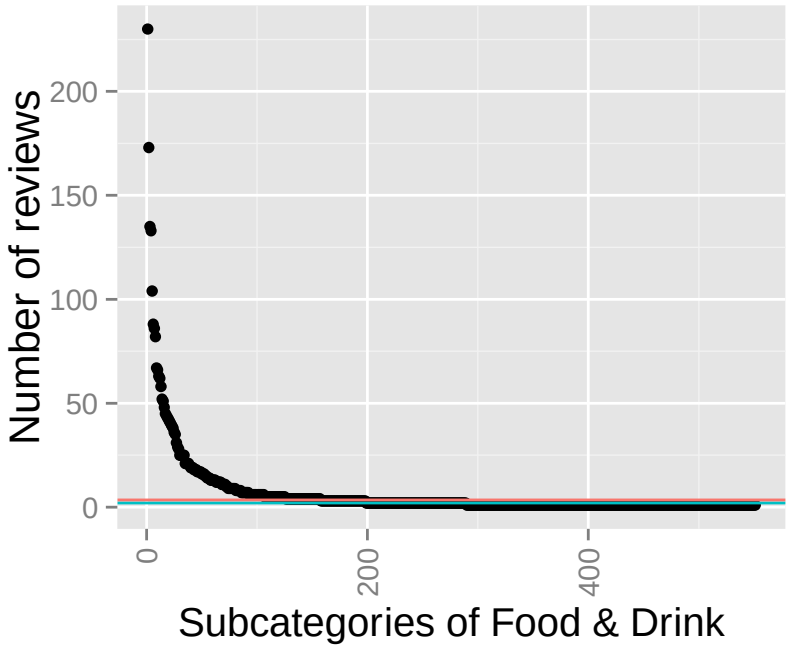


Figure 22 – Distribution of reviews for the main category "Food & Drink"

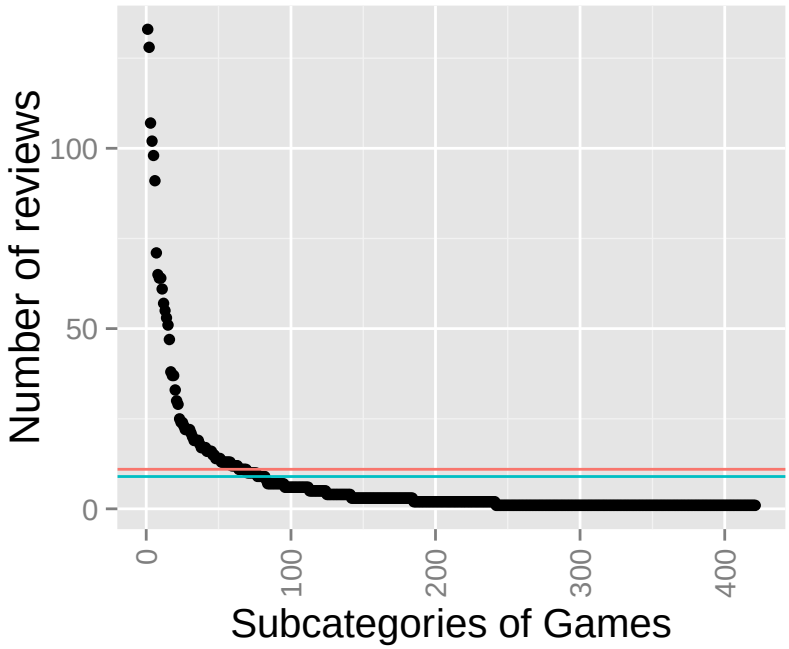


Figure 23 – Distribution of reviews for the main category "Games"

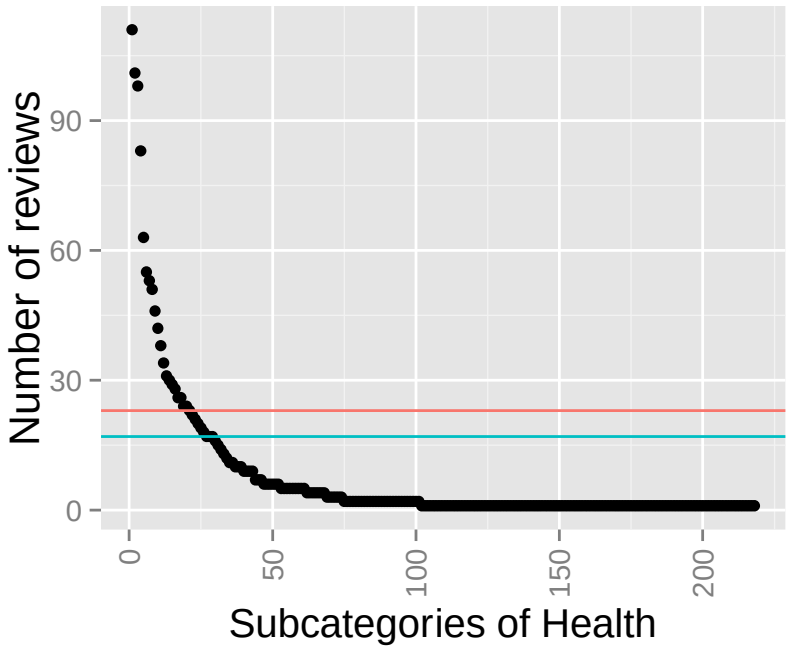


Figure 24 – Distribution of reviews for the main category "Health"

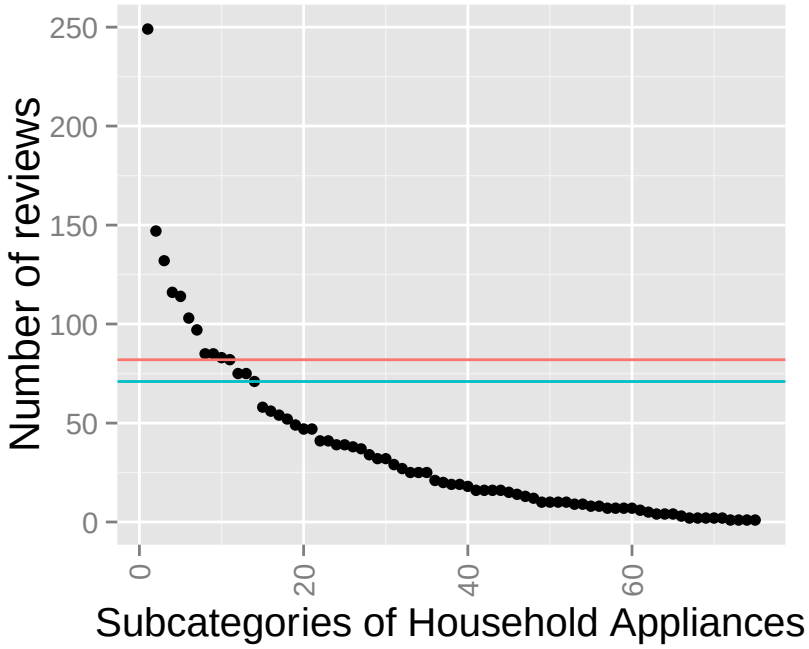


Figure 25 – Distribution of reviews for the main category "Household Appliances"

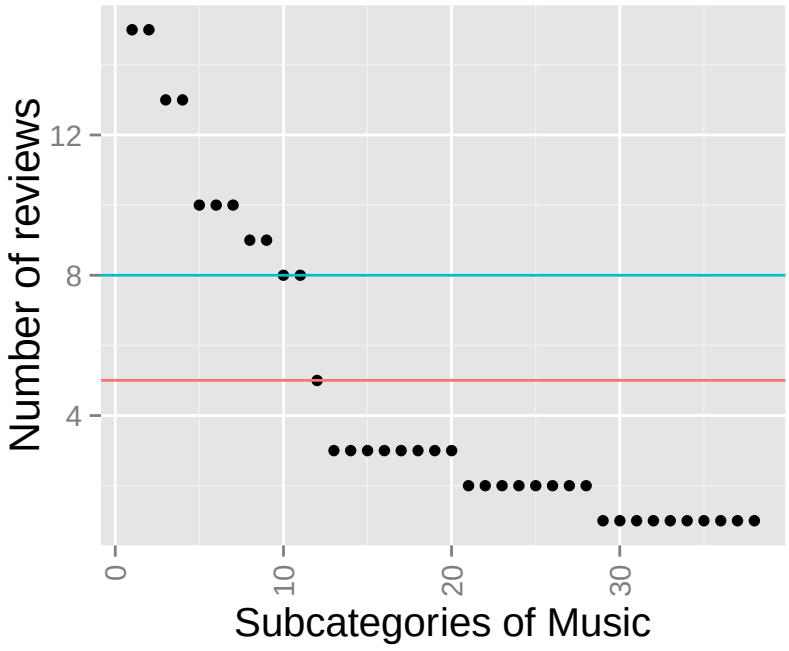


Figure 26 – Distribution of reviews for the main category "Music"

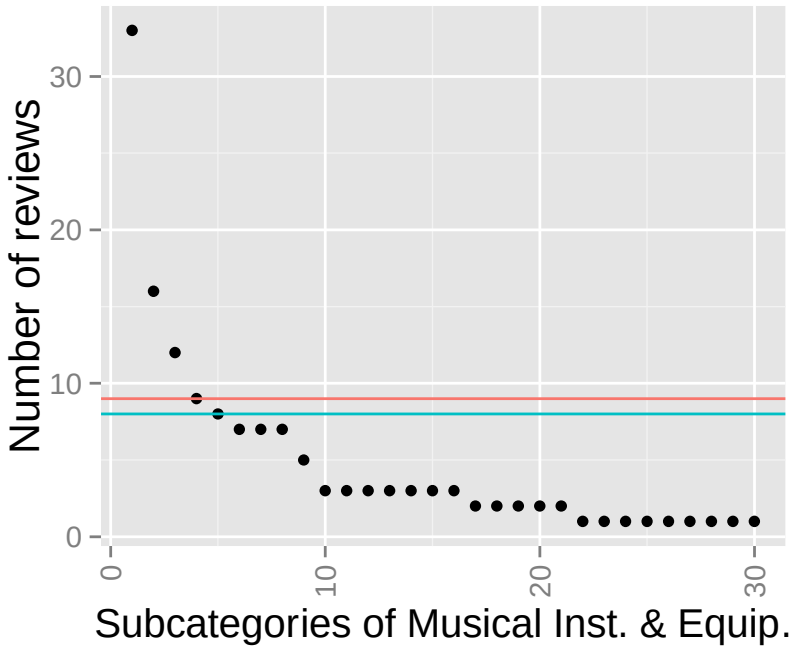


Figure 27 – Distribution of reviews for the main category "Musical Instruments & Equipment"

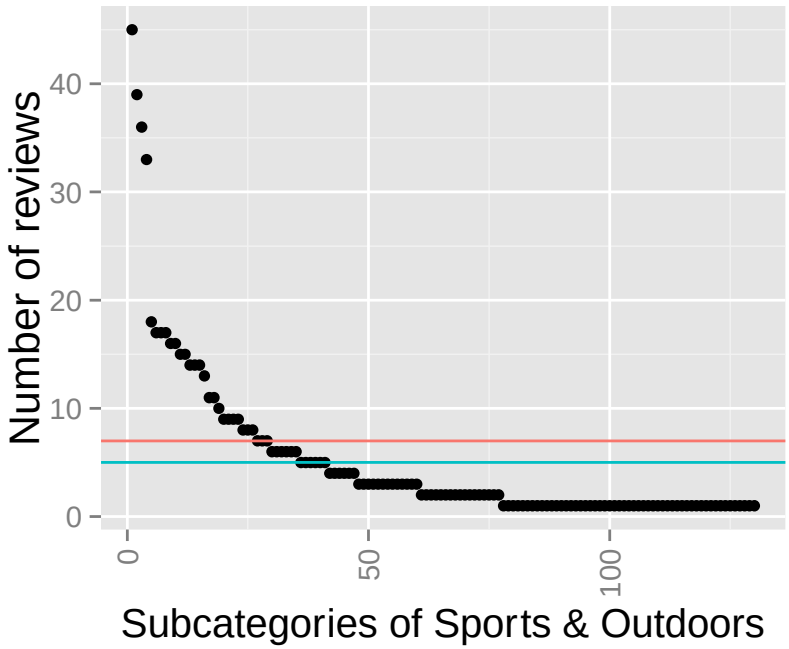


Figure 28 – Distribution of reviews for the main category "Sports & Outdoors"

Due to the lack of number of subcategories, some depicted categories, such as *Finance* (see Figure 21) may not have the visual appearance of an example of a power-law graph showing a representation of the classic "long tail" style curve. Graphs are better defined when there is a higher number of subcategories, so more information can be used to apply both methods. This could be the ideal case. However, all these 15 categories finally exhibit a long tail according to both perspectives (area and length of tail).

The above Table 6 show results from the fitting of a power-law form to each of the gathered categories using the methods described in Part II- Section 1.1. In particular, when examining all categories, the *p-values* reveal that 7 out of the 28 categories among the data sets (*Beauty, Books, Cars & Motorcycles, Computers, Internet, Sports & Outdoors, Telecommunications*) should be firmly discarded. All have *p-values* small enough that the power-law model can be firmly discarded, according to their goodness of fit. However, when examining all these categories and its graphical representation in the following graphs, it can be appreciated that some of them might fit the power-law

distribution on the long peak of the function, at least graphically (e.g. Figure 29 or Figure 30).

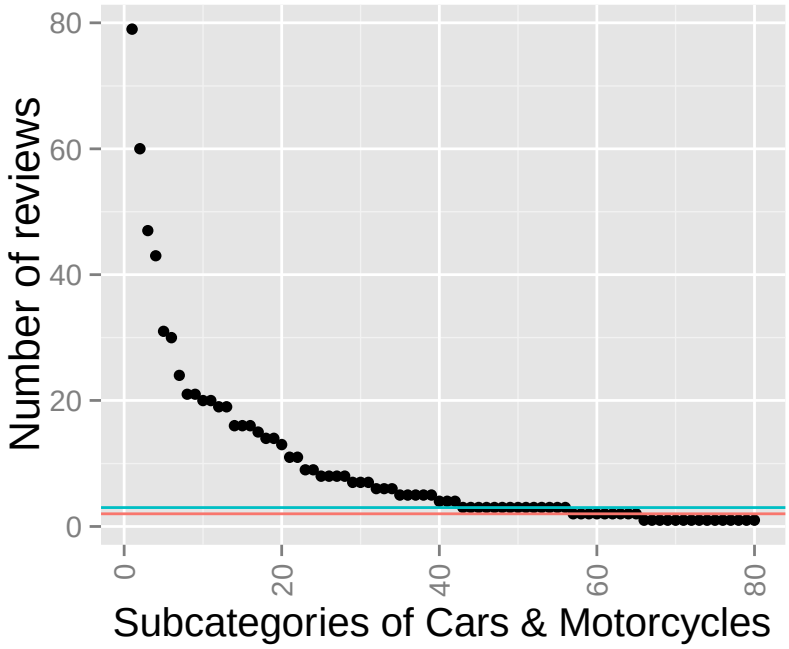


Figure 29 – Distribution of reviews for the main category "Cars & Motorcycles"

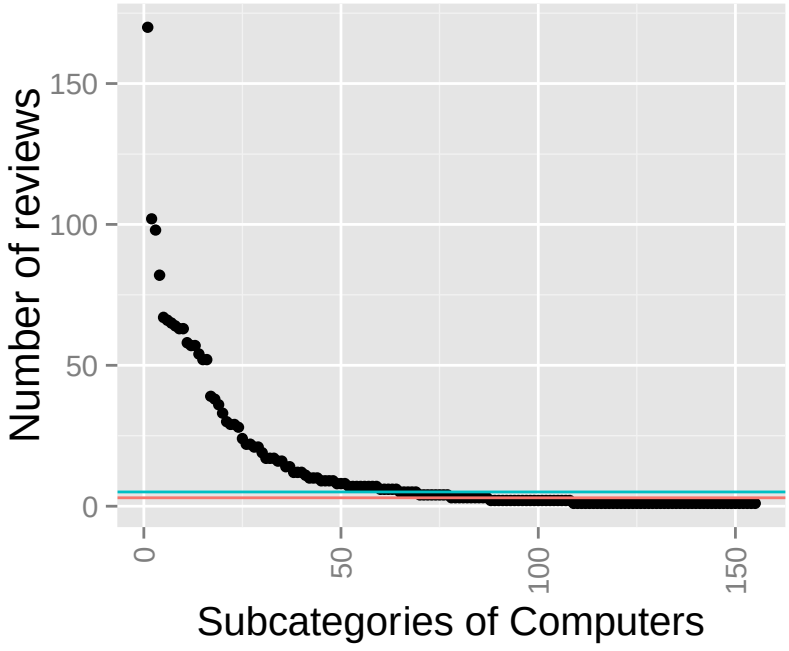


Figure 30 – Distribution of reviews for the main category "Computers"

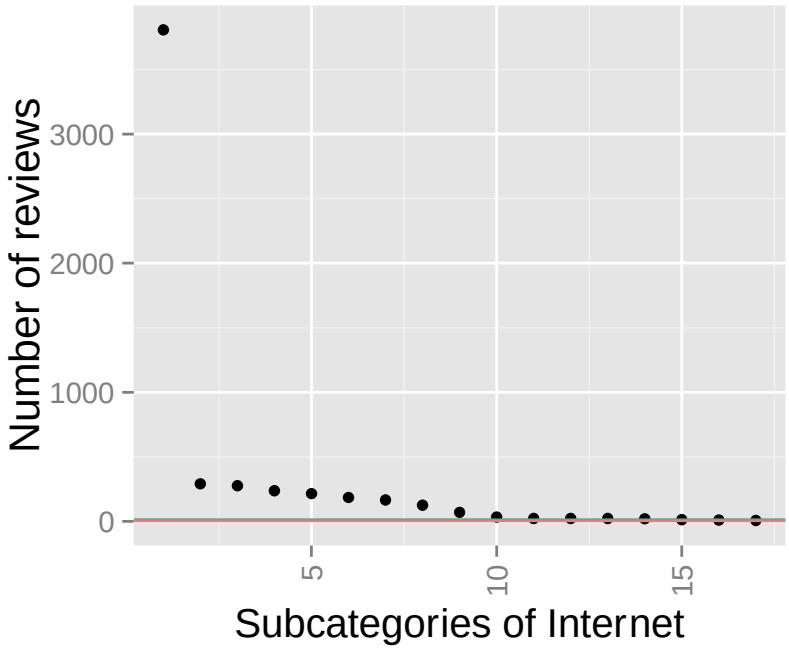


Figure 31 – Distribution of reviews for the main category "Internet"

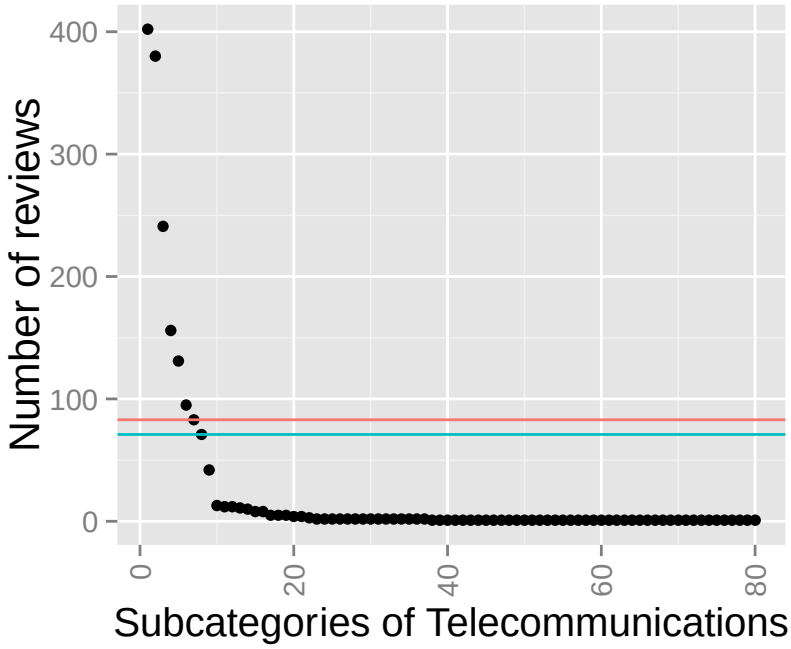


Figure 32 – Distribution of reviews for the main category "Telecommunications"

Additionally and as clarified before, the categories *Beauty, Books, Sports & Outdoors* and *Telecommunications* are some of the cases where the calculated *p-value* is considerably lower than 0.05 (see Table 6). Thus, null hypothesis is rejected, which indicates that such categories do not to follow a power-law distribution. Unlikely, and as revealed by the validity and consistency of the decision-making rules, under the perspectives of the area and the length of the tail, these categories exhibit a long tail.

On the other hand and also according to their *p-values,* 21 out of the 28 data sets might be consistent with a power-law distribution. Consequently and as explained along the Theoretical Framework part of the Thesis, when there is a power-law distribution the events in the tail of the distribution are more likely to happen. Moreover, these categories, according to their $\propto$ exponent are following a power-law distribution with a long peak filled with popular products. Nevertheless, among all those 21 cases there are some that do not exhibit a long tail *(DVDs, Education & Careers Cameras, House & Garden, Office Equipment)*, others –as illustrated above– that do exhibit a long tail according to the validity and consistency of the decision-making rules *(Adult Products, Entertainment, Family, Finance, Food & Drink, Games, Health, Household Appliances, Music, Musical Instruments & Equipment, Fashion)*, and the rest there is uncertainty as to exhibit a long tail *(Ciao Café, Shopping, Travel, Electronics)*. Furthermore, when observing the graphic representation of those categories consistent with a power-law distribution according to their *p-values* –plotted on the following figures– it can be observed that they cannot plausibly be considered to follow a power-law distribution. This is because there are not enough products highlighted as bestsellers, which could be related – in some of the cases – to lack of information, that is, the minor number of subcategories posted (see following Figure 33, Figure 34, Figure 36, Figure 39, Figure 40, Figure 41). The probability of getting by chance a fit as poor as the one observed is very small in each of these cases. Therefore, it might be difficult to discern whether the number of products fit the power-law adjustment or is characterized by a long tail, which is below the x_{min} demarcated by a horizontal line.

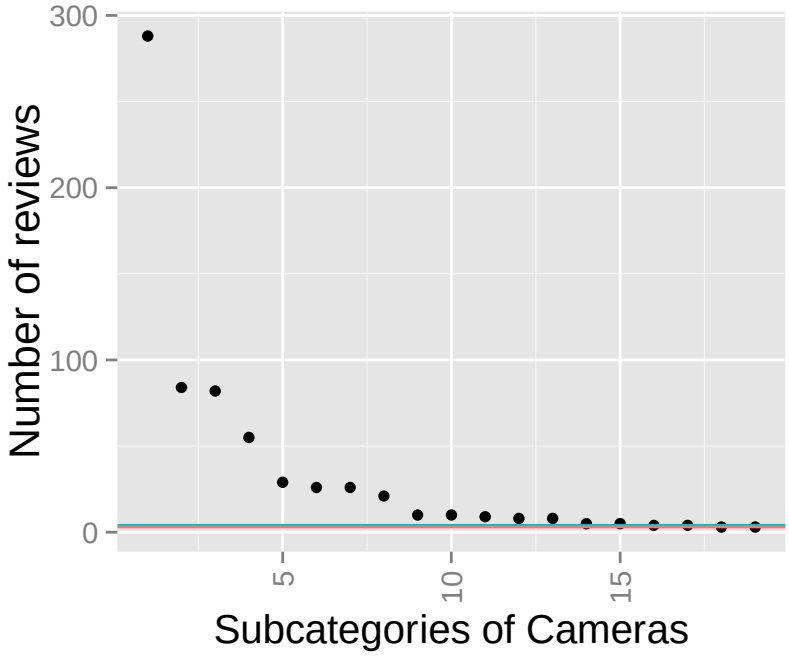


Figure 33 – Distribution of reviews for the main category "Cameras"

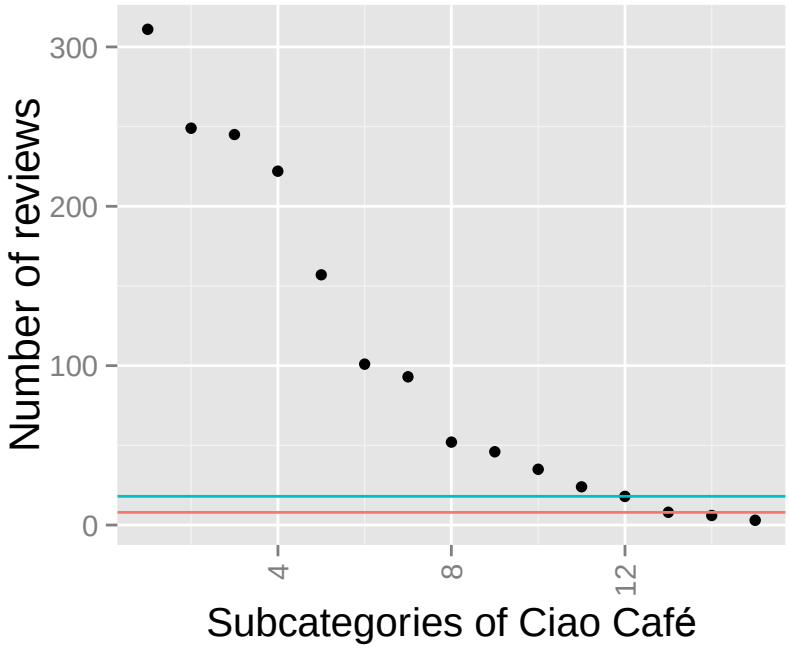


Figure 34 – Distribution of reviews for the main category "Ciao Café"

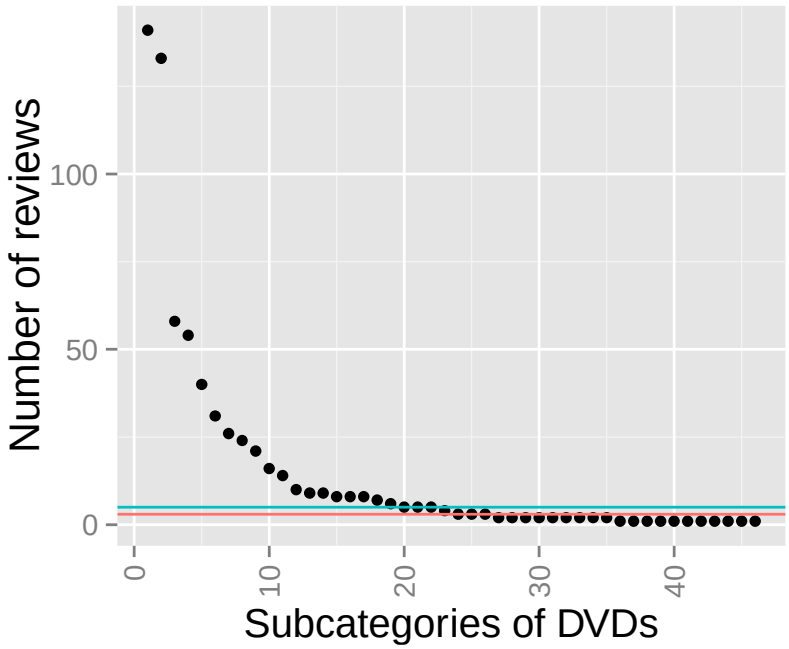


Figure 35 – Distribution of reviews for the main category "DVDs

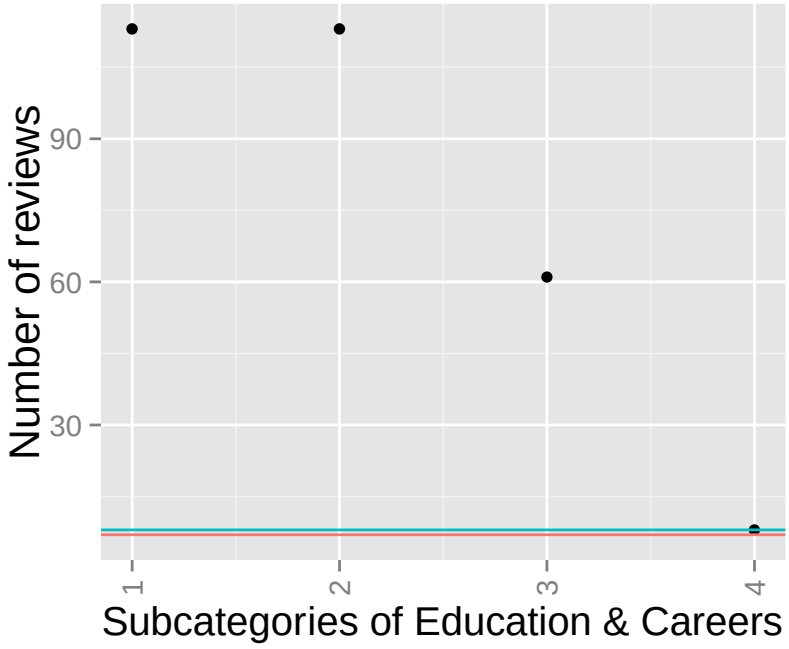


Figure 36 – Distribution of reviews for the main category "Education & Careers"

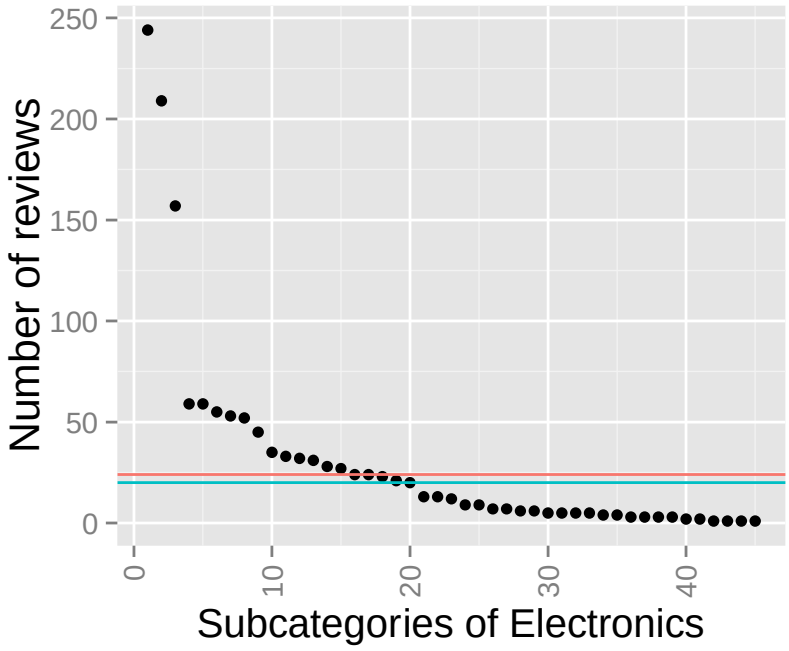


Figure 37 – Distribution of reviews for the main category "Electronics"

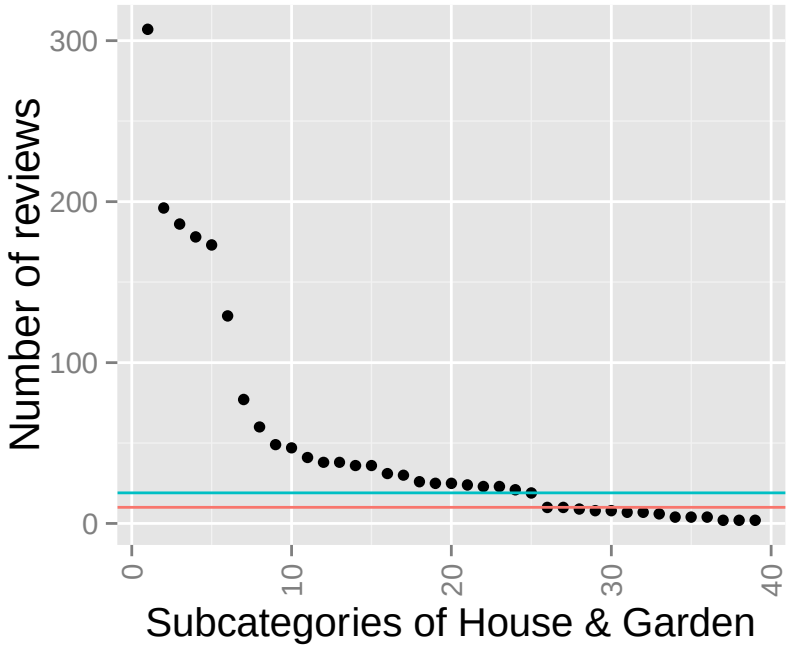


Figure 38 – Distribution of reviews for the main category "House & Garden"

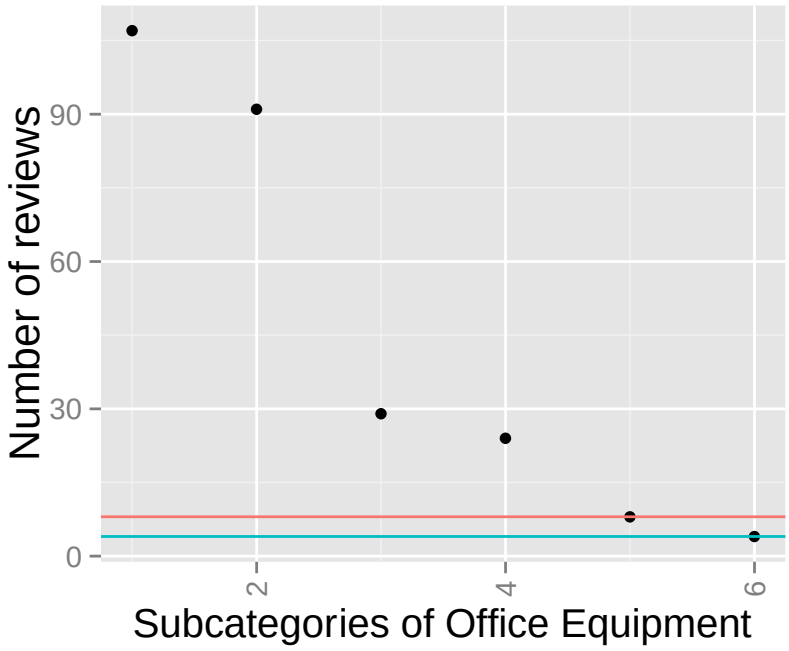


Figure 39 – Distribution of reviews for the main category "Office Equipment"

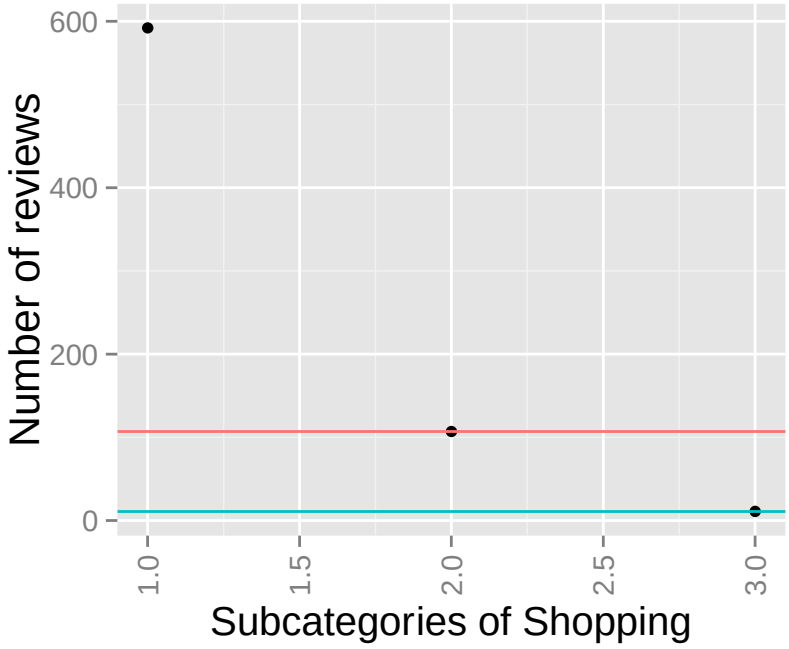


Figure 40 – Distribution of reviews for the main category "Shopping"

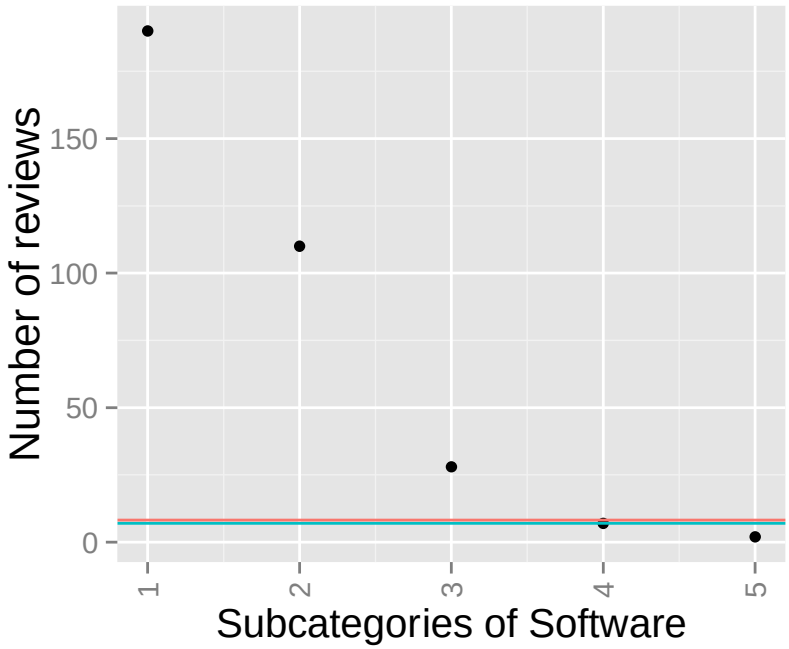


Figure 41 – Distribution of reviews for the main category "Software"

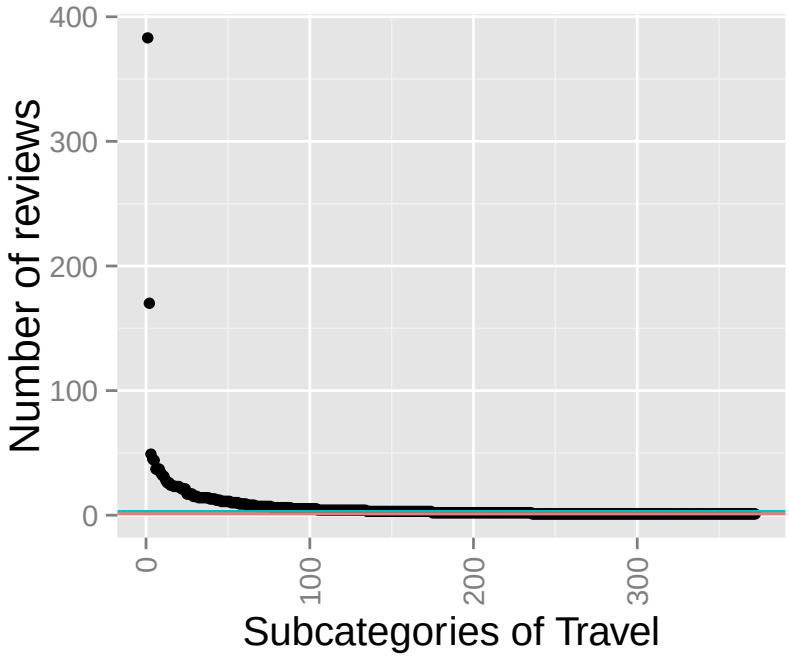


Figure 42 – Distribution of reviews for the main category "Travel"

It is relevant to emphasize that all the x_{min} calculated according to the power-law adjustment (see Table 6) and through the elbow method (see Table 7) appear to be close one to another. In the graphs the x_{min} is demarcated by a horizontal red line representing the one calculated by the elbow method, and by a horizontal blue line representing the one calculated by the power-law method. Furthermore, there are some x_{min}, which are special cases. An instance of this is the case of the main category Education and Careers, of which calculated area of the long tail is zero in both methods. In order to calculate the x_{min} according to the elbow method, the graph is traversed from the right to the left. When comparing pairs of values and the condition is not met (slope lower strict than -1), the first x_{min} is taken (marked in blue, see Figure 36). Perhaps it may not be the ideal in such cases like this, but the only option. In this case the following values are in the coordinate axis, where the subcategories are represented: 113, 113, 61, 8. As can be noticed, the pairs of points are compared from the left to the right and it is observed that none is greater than -1 until the last one, thus 8 is chosen as the x_{min}.

The next stage has to do fundamentally with characterizing products types. Hence, Table 9 shows a description of the main features of all the 28 gathered categories, which encompass diverse products and services reviewed by Ciao UK users as well as their product type categorization according to their characteristics.

Table 9 – Description of each of the 28 main categories of Ciao UK

MAIN CATEGORIES	Description	Product type
Adult Products	Selection of adult products from vibrators and dildos to adult games and underwear	Experience
Beauty	Information about hair care, make-up products, perfume brands for him or her and spa or beauty clinics	Experience
Books	Browse through all fiction categories to find a mysterious new world in the Fantasy section or a terrible crime in a familiar setting	Experience
Cameras	Digital, video, SLR or standard cameras. Everything by popular manufacturers such as Sony, Nikon, Panasonic and Canon. Likewise, Accessories needed for the cameras including: memory cards, batteries, lenses and cases	Search
Cars & Motorcycles	New cars, used cars, motorcycles, car accessories, car insurances or a car rental firms	Experience
Ciao Café	Stories and opinions of other members and views on current topics	Experience
Computers	Laptops, LCD-monitors, printers, tablets and computer accessories among others. Including laptops from the most popular brands such as Apple, Sony, HP, Samsung and Dell	Search
DVDs	DVDs encompassing action, horror, comedy, romance or family films	Experience
Education & Careers	Education and career resources to assist in current or future professional life. Resources on higher education, further education, career, placement, internship, gap year and alternative activities	Experience
Electronics	Electronics from iPods and 3D TV's to Projectors and Headphones from the best brands. There are also Blu-ray players from Samsung, Sony & Panasonic and eBook readers like the Amazon Kindle	Search
Entertainment	This section covers games, music, books, movies and magazines. Including also information on TV and radio programmes	Experience
Family	Product or information on breastfeeding, prams & pushchairs, bedding, toys, baby care and much more including many budget products	Experience
Fashion	Clothing for men, women and kids, as well as accessories of all types to compliment the outfits	Search
Finance	Banking and personal finance, pension, stockbroker, insurance and other legal financial services. Likewise, advice on renting, mortgages and loans.	Experience
Food & Drink	Coffee, beer, wine and tea plus much more. There are reviews about the food and drinks users have tried and tasted	Experience
Games	Reviews about new or old games. From PlayStation 1 and Super Nintendo to Wii-U and Xbox 36. From handheld devices to classic consoles. Likewise, controllers, cables and headsets	Experience

MAIN CATEGORIES	Description	Product type
Health	Products for eye & health care, family planning, as well as the best supplements for every diet and advice on all health concerns	Experience
House & Garden	Items from barbecues and lawn mowers to cacti and bird food	Search
Household Appliances	Products such as ovens, washing machines, fridges, cookers, freezers or one of many other appliances for the house	Experience
Internet	The best places for every task on the web, from dictionaries and email to holiday booking site and voucher offers	Experience
Music	From popular pop group to folk bands. Likewise, from those relaxing CD's to music for starting a night of partying. There is also a great mp3 player section	Experience
Musical Instruments & Equipment	Products such as guitars, keyboards, percussions, DJ turntables, as well as recording equipment and software	Experience
Office Equipment	Printers and calculators as well as the most ergonomic chairs and desks	Search
Shopping	Shopping, services and a shopping by city guide. Accessories, jewellery, books shop, home improvement, DIY, department store, postal service, telecommunication, travel service, customer service as well as the best shopping in the UK & Ireland, Europe and worldwide	Experience
Software	Products such as mobile apps, operating systems and computer programs	Experience
Sports & Outdoors	All things related to football, golf, swimming and fitness. Besides, it is found sport equipment for winter and extreme sport as well as information on sports resources on the Internet, sport locations	Experience
Telecommunications	Mobile phones, telephones or accessories among many others	Search
Travel	Users' travel experiences, such as which restaurants, flights and hotels they recommend. Reviews on cities and hotels include destinations such as London, America and France plus many others across all continents	Experience

In addition, according to the results gathered in Table 8 and Table 9, Figure 43 and Figure 44 show the categorization of the categories of products with and without long tail whose locations have been indexed by their major attributes (with respect to the experiencing level) and product-evaluation standards (with respect to the objectivity level of product) according to the above Theoretical Framework (see Part I-Section 1.4).

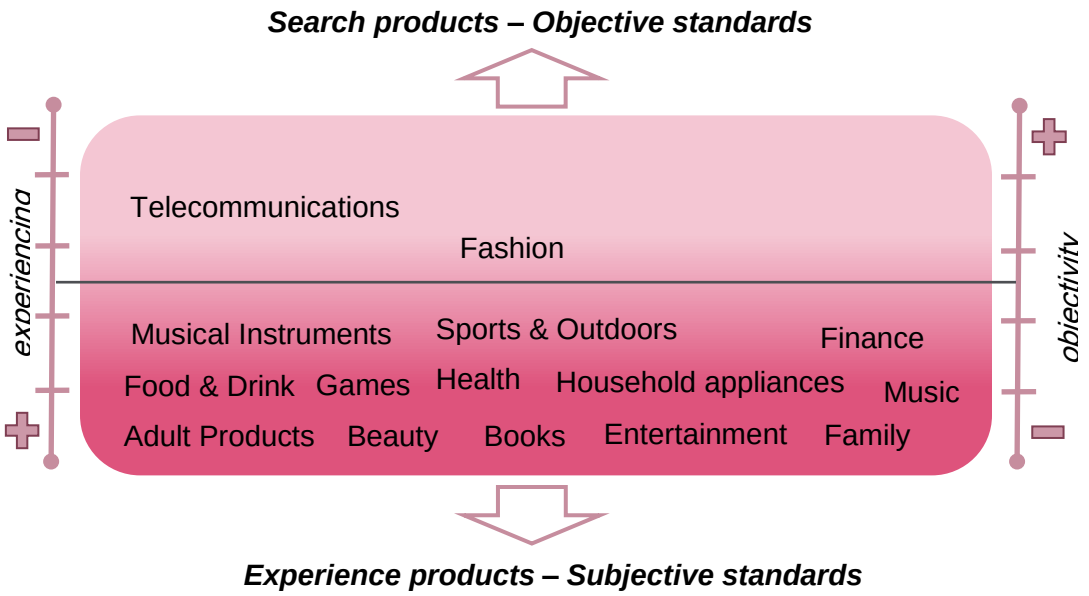


Figure 43 – Categorization of the 15 main categories of products that exhibit a long tail

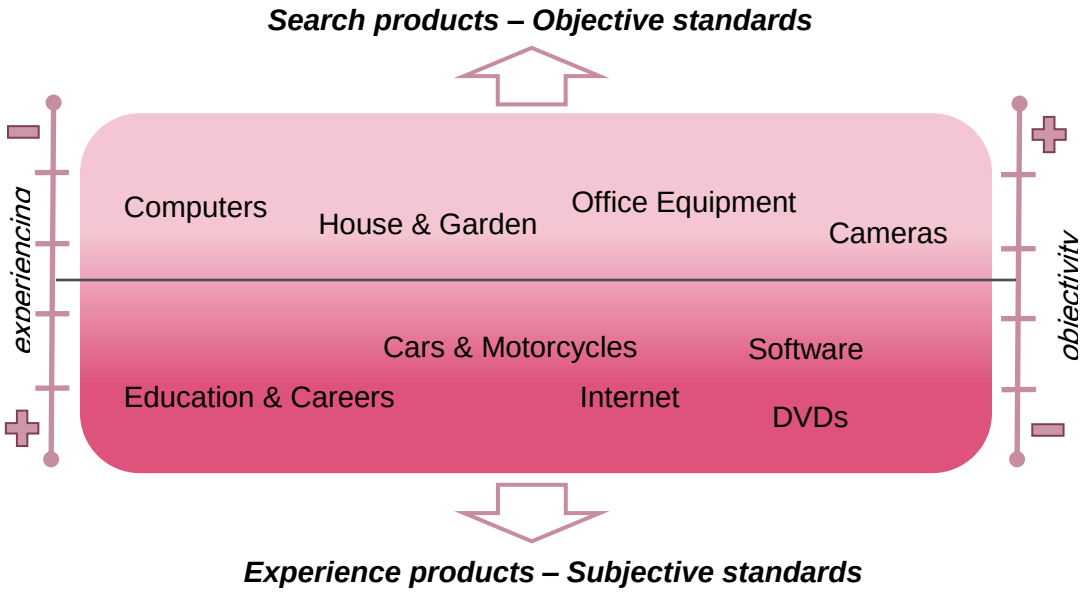


Figure 44 – Categorization of the 9 main categories of products that do not exhibit a long tail

2 Hypotheses testing

		Type of product	
		Experience	***Search***
Long tail	YES	13	2
	NO	5	4
	Uncertain	3	1

Hypothesis	Description	Result
H_1	The experience products from the distribution of product categories within an eWOM are more likely to exhibit a long tail	Supported
H_2	The search products from the distribution of product categories within an eWOM are less likely to exhibit a long tail	Supported

		Type of product	
		Experience	***Search***
Super-hits	YES	16	5
	NO	5	2

Hypothesis	Description	Result
H_3	The distribution of product categories within an eWOM that have high frequency events or super-hits in the short head are not particularly associated with search or experience products	Supported

Figure 45 – Summary of findings

Figure 45 shows how the above-presented hypotheses are proven to be feasible. In the case of the first hypothesis ($\boldsymbol{H_1}$) the analysis of the results (see Descriptive results section) reveals that the majority of categories of products that encourage the long tail phenomenon are classified as experience goods or those categories where reviews can be done through subjective evaluation standards. However, among all the categories of products with long tail, only two of them emerged to differ. These categories of products sustain a portion of more objective or impartial evaluation standards, and thus, are classified as search products. Furthermore, it was found that among all the gathered search products there are 66.7% of them that do not exhibit a long tail. Consequently,

the second hypothesis ($\boldsymbol{H_2}$) is supported since search products are barely interrelated with the long tail.

Additionally, for the third hypothesis ($\boldsymbol{H_3}$) validation in the Figure 45 it is appreciated that according to the gathered results, there are a mix of products that might be consistent with a power-law distribution. Thus, the long peak should be filled with super-hits. Interestingly, those products do not follow a specific pattern in regard to product type, but they are a mix of different experience and search products.

3 Discussion

Anchored in the results gathered form the analysis of the case study Ciao UK it is observed that not all the users' reviews across all categories distinguished within Ciao UK follow a power-law distribution. Moreover, results show that many distributions follow a power-law only in the tail and others concentrate most of the information in the head region. Nevertheless, not all long tails are power-law. In this regard, there is also literature that mathematically examines the power-law behaviour. For instance, Clauset, Shalizi and Newman (2009) present a statistical framework for discerning and quantifying power-law behaviour in empirical data occurring in the tail of the distribution. In this sense, in order to reveal the power-law form of the distribution as well as the long tail, this Thesis has focused on two analytical models within an online recommender system such as an eWOM community. The first one was the power-law method by Clauset, Shalizi and Newman (2009), which was used for analysing user-generated data gathered online. The second one was the elbow criterion, which is used in many fields, but it has not yet been applied to verify whether there is a long tail among data as it has been applied for this Thesis. Those lines of study find evidence to support whether there is power-law behaviour or a long tail on Internet data, however, it is also important to go beyond data. As many assortments of niche products are only sold online and their consumers are more prone to use online reviews as the primary source for information, it is crucial to find out what kind of products are enclosed there. To this end, as already explained, the classification drawn by Nelson (1970, 1974) to

distinguish between search goods (products dominated by product information attributes prior to purchase) and experience goods (products dominated by information attributes only appreciated after purchase and use) has been used.

Interestingly, prior research has found evidences regarding the shift in the sales distribution for different types of products within eWOM communities. Brynjolfsson, Hu and Smith (2009) considered books –classified as experience product– reviewed in Amazon and found a long tail in the sales distribution online. In another study, Elberse and Oberholzer-Gee (2007) also found a long-tail effect in the video titles –an experience product– that sell only a few copies every week. More specifically, research by Lee, Lee and Shin (2011) argued that eWOM is able to change the rule of the long tail theory by differentiating the cases across product types. Their categorization of product types has been based on the objectivity of product evaluation standards. In this regard, the authors' results indicated that the sales distribution within the eWOM community Amazon show a thin head and a long tail in products evaluated by subjective standards phenomenon. In this regard, the analysis of results within this Thesis– reinforced by the implementation of the two analytical models power-law method and elbow criterion– reveals that experience goods are those encouraging the long tail phenomenon. These results coincide with the results identified by Lee, Lee and Shin (2011) on their conclusions and are also supported by the first hypothesis ($\boldsymbol{H_1}$). Likewise, further results within this Thesis verified the second hypothesis ($\boldsymbol{H_2}$) revealing that search products are those that do not show a long tail coinciding again with the results by Lee, Lee and Shin (2011). The authors' results showed sales concentration with a short tail when the product was evaluated by objective standards. Notwithstanding that the categorization of search and experience products is relevant to distinguish those having a long tail, many products still involve a mix of search and experience attributes (Huang, Lurie and Mitra 2009). In this regard, some of the main categories might cover also subcategories that have not been categorized as the same type of product as the main category has been. Consequently, it is difficult to establish a standard format. In particular, the cases of the products category *Telecommunications* and *Fashion* that have been categorized as a search product having a long tail have enclosed subcategories that could be categorized as experience products. For example,

when observing the description in Table 9 the subcategories of *Telecommunications* encompass mobile phones, telephones and others referred as “accessories”. Such accessories have nothing to do with objective evaluation standards (e.g. class, capacity or power) as the other subcategories but with subjective (e.g. colour, design, style) evaluation standards. Also in this respect, Chatterjee (2001) stated that the content of online reviews might vary from subjective to objective.

More existing studies of eWOM have found evidence of the super-hits effect. For instance, Standifird (2001) supports that eWOM promotes the sales of popular products, so the head part of the sales distribution becomes thicker. Likewise, the authors Elberse and Oberholzer-Gee (2007) not only found in their study evidence of long tail effect in home video sales, but also they found evidence of super-hit effect. Such effect was more pronounced among best-selling titles, which accounted for the majority of sales. The authors dealt with information about video titles, which are categorized as experience products. In this regard, this Thesis goes further and has also found evidence to support a third hypothesis ($\boldsymbol{H_3}$). That is, when there is a super-hit effect within a power-law distribution of product categories within an eWOM, this is indicating that products contain either search or experience attributes.

Finally, it is important to emphasize that to obtain the findings the online data collection component of research has made it possible to acquire virtual experience directly by users and transform it into valued information. In this regard, Anderson (2008b) suggested that the field of social science should embrace the use of analytical tools to have better data. Besides, advances in data collecting technologies, such as the web crawler used for this Thesis has developed a new cross-cutting discipline: integration of computational models and social science.

Part V: Conclusions

The purpose of this fifth part is to conclude the main aspects of the findings in relation to the hypotheses and aim presented at the beginning of this Thesis. Hereafter, the practical and theoretical contributions are outlined, as well as lessons learned and suggestions for future research, which could further contribute to the field.

1 Conclusions

This Thesis analyses whether there is a long tail characterization in an eWOM community and what product types are enclosed across. To that end, the previous sections explore two different methodologies that allow finding power-law behaviour among data distribution as well as specifying in which cases occur the long tail. Likewise, the methodology of this Thesis also shed light on the importance of online user-generated data collection in the context of social science. In eWOM communities, the online reviews can be considered as user-generated content as they consist of comments published by users on the products and services they experienced. Besides, data collection has allowed a better understanding of the insights found along the data set gathered from 28 main categories of the portal Ciao UK. In particular, to better specify in which cases occur the long tail by discerning the cases through product types.

Due to the current importance of eWOM in businesses and economics, this Thesis shows what products of sales distribution form the long tail in an eWOM and proposes a type of product categorization that allows specifying the cases in detail. To do so, a product type categorization is proposed to differentiate type of products from eWOM by the degree of objectivity and by the degree of experiencing. Specifically, in order to validate the hypotheses and to differentiate the shape of sales distribution across the long tail a panel data collected from ciao.co.uk has been used for validation. Subsequently, to compare the entire product types in Ciao UK, the probability power-law distribution function was depicted as a tool to measure the long tail. For extra statistical support the elbow criterion method was also used to corroborate where was located the optimal cut-off point that distinguishes the products characterized by the long tail. The results supported all the three proposed hypotheses. In this sense, this Thesis presents important new findings. Firstly, it is evidenced that products having a long tail are those with subjective evaluation standards, which are classified as experience products. Secondly, it is also corroborated that search products, which have a high level of objective attributes in the total product assessment do not encourage the long tail phenomenon. Thirdly, there is a combination of products when there are super-hits in the short head of the distribution. Thus, those are not particularly associated with

search or experience products since they contain either objective or subjective evaluation standards. Finally, it is also remarkable to highlight that not all the categories fitting a power-law distribution are characterized by a long tail and on the contrary, some of those having a long tail do not fit a power-law.

In general, the findings also suggest the potentials of eWOM, which, in general, might generate a long tail effect, where a large number of small-volume vendors coexist with a few high-volume ones.

2 Research contributions

This Thesis has contributed to both theory and practice, essentially, in three different ways: (1) with a methodology of collection of online user-generated data in the context of social sciences; (2) with the development of two more accurate methods to identify niche products within an eWOM community, providing a deeper understanding of the long tail phenomena and the type of products; and (3) with publications of refereed journals papers (indexed in JCR/JSCR) as well as conference papers related to the main topic of this Thesis.

(1) Following from the above-described theoretical framework, several explanations for opportunities that Big Data offers to the field of social science can be traced to several authors. Boyd and Crawford (2012) outlined Big Data as a socio-technical phenomenon and explained how to handle it through the use of data collection and analysis tools. The authors described themselves as an example of social scientists working together with computer scientists and informatics experts, who focus on Big Data in social media context. Likewise, Chang Kauffman and Kwon (2014) presented a comparison of examples gathered from the literature of the importance the role Big Data plays in the so-called computational social science research. Furthermore, when referring to a methodology of crawling user-generated data from a web there is neither a specific approach nor a common agreement in the literature. However, there are

some authors in the literature describing the challenges and trade-offs inherent in web crawler design. Such as Reips and Garaizar (2011), who described in their paper web collectors of Big Data for Wikipedia and Twitter respectively. While those authors have contributed with good ideas related to the matter, the specific attention of this Thesis has resided on defining a practical framework and methodology to collect Big Data from an eWOM community that has user-generated content. To that end, a portrait of a new model and architecture for a web crawler in which traditional data retrieval methods are challenged has been provided. Additionally, this Thesis has contributed with a methodology that is not only understood from the perspective of the social science discipline but also includes practices on data accessing and computing.

(2) Theoretical and practical implications of this Thesis contribute with a new approach on analysing the information gathered from eWOMs to understand the context of the long tail phenomenon. In this regard, some authors have drawn attention to the Anderson's (2004) long tail by analysing data collected from eWOMs such as Amazon or sales portals such as Quickflix among others (Brynjolfsson, Hu and Smith 2006, Brynjolfsson, Hu and Smith 2010, Elberse 2008). Although those ideas continue to hold sway and this Thesis has absorbed all of them, the specific attention has resided on defining a new methodology for finding and distinguishing niche products focusing in the long tail from eWOM communities. To this end, this Thesis has explored different product types – experience and search products– characterized by the formation of the long tail among all the categories enclosed in product reviews of Ciao UK. Built on this foundation, two theoretical and practical methodologies have contributed with a new approach on examining data through the power-law distribution method and by using the elbow criterion for verification.

From the managerial perspective, this Thesis has also has presented a challenge to businesses by suggesting that they should embrace the opportunity of finding first-hand information on different niche products that exceed the geographic boundaries. Specifically, the long tail impact can be observed differently across

product types (search or experience product) that could affect information search. What is more, methods like those proposed here permit managers to identify and find niche products, as well as to leverage advantages into dominating global blockbusters. Thus, they should look for those markets where those niche products stand to be of great interest for customers outside their immediate social network, since that could have a significant impact on business' sales.

(3) The following listed publications are those resulting from this Thesis.

Journal papers:

- Martinez-Torres, Olmedilla, M. (2016). Identification of innovation solvers in open innovation communities using swarm intelligence. Technological Forecasting and Social Change, 109, pp. 15–24. DOI: 10.1016/j.techfore.2016.05.007

- Olmedilla, M., Martinez-Torres, M., Toral, S. (2016). Harvesting Big Data in Social Science: A methodological approach for collecting online user generated content. Computer Standards & Interfaces, 46, pp. 79-87. DOI: 10.1016/j.csi.2016.02.003

- Olmedilla, M., Martinez-Torres, M., Toral, S. (2015). Examining the power-law distribution among eWOM communities: A characterization approach of the long tail. Technology Analysis & Strategic Management, pp. 1-13. DOI: 10.1080/09537325.2015.1122187

Conference papers:

- Olmedilla, M., Arenas-Marquez, F. J, Martínez Torres, MR, Toral, S. (2016). Features of Reputed Users in eWOM Using Evolutionary Computation, 9th International Conference on Developments in e-Systems Engineering, DeSE2016

- Olmedilla, M., Arenas-Marquez, F. J, Martínez Torres, MR, Toral, S. (2016). Identification of Influencers in eWord-of-Mouth communities using their

Online Participation Features, Proceedings of the 1st international Conference on Advanced Research Methods and Analytics, CARMA2016, pp. 38-45

- Olmedilla, M., Martínez Torres, MR, Toral, S., Teso E. (2015). Examining Gender Discourse Differences in Shared Reviews about Books in eWOM, 8th Conference on Developments in e-Systems Engineering, DeSE2015

- Olmedilla, M., Martinez-Torres, M., Toral, S. (2015). A long tail study of eWOM communities. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, New York, USA, 9 (6), pp. 1279-1283

- Martinez-Torres, M., Toral, S., Olmedilla, M. (2015). A quantitative study of the evolution of Open Source Software Communities. World Academy of Science, Engineering and Technology, International Journal of Computer, Electrical, Automation, Control and Information Engineering, New York, USA, 9(6), pp. 1268-1273.

3 Limitations and future work

The principal limitation of this Thesis could be that the methodology has been implemented in just one eWOM community. Nevertheless, the applied gathering methodology can be adapted to other eWOM websites since it is possible to follow the programming code pattern. Actually, that would be achievable by rewriting something similar but for the structure of the new eWOM website. Additionally, it would be possible to extend the two methodologies (power-law and elbow criterion) to measure the long tail of other eWOM communities. In this regard, probable future research plans on gathering information to study other eWOM communities are focused on TripAdvisor. Such eWOM provides reviews of travel-related content and its structure has very similar characteristics to Ciao.

Other possible methodological limitation in the analysis of this paper would be the sample of the data set. That is, Ciao does not represent all the population but a particular sub-set, since members using Ciao are not representative of the global population. Besides, it is important to take into account that accounts and users within Ciao are not equal since users might have multiple accounts, whereas some accounts might be used by many people. Thus, some accounts might be 'bots' producing automatic content without the involvement of a person. Furthermore, not all information within Ciao is provided by the registered users. For instance, the 95.21 % of the users are sharing their age, unlike the 11,45 % who are sharing their real name or the 10.87 % who share their location. Nonetheless, is important to clarify that this limitation is just personal data from a registered user, a trace about a user's activity and interactions (reviews, score, circle of trust, etc.) is always registered.

Since data is continuing increasing in more abundance than before, further research and potential areas for future work in regard to data collection would be reducing the crawling speed as well as the response time to gather larger collections of data, which is also a major issue. The goal would be discover another crawling strategy, i.e., a strategy for determining which URLs to download next or to have a highly optimized system architecture that could download a large number of URLs per second while being robust against errors and programming exceptions. Consequently, a better data-processing, data-gathering and data-storing technology would be necessary.

Further research can also extend the findings by characterizing the long tail for each subcategory of the posted reviews in an eWOM portal, and defining a tool for extracting and representing all the niche products across the long tail. The goal would be discovering some common patterns among niche products through the creation of social network models, where nodes would represent types of products and edges would connect products that have received reviews from the same user. Likewise, base on the product type categorization this Thesis could be extended in several directions. First, future research could compare the influence of online reviews among various types of products within the same product categorization. Second, more research could

contribute to a more detailed evaluation of the way information is handled for experience versus search products.

Bibliography

Aarstad, J. “Possible suboptimal diffusions of technological innovations in clustered scale-free networks.” *Technology Analysis & Strategic Management* 26, no. 3 (2014): 267-277.

Adamic, Lada A., and Bernardo A. Huberman. “Power-law distribution of the world wide web.” *Science* 287, no. 5461 (2000): 2115-2115.

Alanah, Davis, and Deepak Khazanchi. “An Empirical Study of Online Word of Mouth as a Predictor for Multi-product Category e-Commerce Sales.” *Electronic Markets* 18, no. 2 (2008): 130-141.

Albert, Réka, Hawoong Jeong, and Albert-László Barabási. “Internet: Diameter of the world-wide web.” *Nature* 401, no. 6749 (1999): 130-131.

Anderson, Chris. “The long tail.” *Wired Magazine*, no. 12.10 (October 2004).

Anderson, Chris. *Long Tail: Why the Future of Business is Selling Less of More.* New York: Hyperion Books, 2008a.

Anderson. “The end of the theory: the data deluge makes the scientific method obsolete?” *Edge.* 2008b. http://www.edge.org/3rd_culture/anderson08/anderson08_index.html (accessed June 15, 2015).

Antenucci, D., M. Cafarella, M. Levenstein, C. Ré, and M.D. Shapiro. “Using social media to measure labor market flows.” *National Bureau of Economic Research* w20010 (2014).

Arenas Márquez, Francisco José, María Rocío Martínez-Torres, and Sergio L. Toral. “Electronic word-of-mouth communities from the perspective of social network analysis.” *Technology Analysis & Strategic Management* 26, no. 8 (2014): 927-942.

Arndt , J. “Role of product-related conversations in the diffusion of a new product,, 4 (1967), pp.” *Journal of Marketing Research* 4, no. 3 (1967): 291–295.

Burton, Jamie, and Marwan Khammash. “Why do people read reviews posted on consumer-opinion portals?” *Journal of Marketing Management* 26, no. 3-4 (2010): 230-255.

Begoli, E., and J. Horey. “Design principles for effective knowledge discovery from big data.” *Software Architecture (WICSA) and European Conference on Software Architecture (ECSA)* (IEEE), August 2012: 215-218.

Benghozi, P.J., and F. Benhamou. “The long tail: Myth or reality?” *International Journal of Arts Management* 12, no. 3 (2010): 43-53.

Bickhart, Barbara, and Robert M. Schindler. “Internet forums as influential sources of consumer information.” *Journal of interactive marketing,* 15, no. 3 (2001): 31-40.

Bhattacharjee, S., R.D. Gopal, and K. Lertwachara. “Consumer Search and Retailer Strategies in the Presence of Online Music Sharing.” *Journal of Management Information Systems* 23, no. 1 (2006): 129-159.

Blank, A., and S. Solomon. “Power laws in cities population, financial markets and internet sites (scaling in systems with a variable number of components).” *Physica A: Statistical Mechanics and its Applications* 287, no. 1 (2000): 279-288.

Boyd, D., and K. Crawford. “Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon.” *Information, communication & society* 15, no. 5 (2012): 662-679.

Boyd, D., and N. Ellison. “Social network sites: definition, history, and scholarship.” *IEEE Engineering Management Review* 3, no. 38 (2010): 16-31.

Bone, P.F. “Word of mouth effects on short-term and long-term product judgments.” *Journal of Business Research* 32, no. 3 (1995): 213-223.

Breslau, L., P. Cao, L. Fan, G. Phillips, and S. Shenker. “Web caching and Zipf-like distributions: Evidence and implications.” *INFOCOM'99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings.* New York: IEEE, (1999): 126-134.

Brynjolfsson, Erik, Yu Jeffrey Hu, and Duncan Simester. “Goodbye pareto principle, hello long tail: The effect of search costs on the concentration of product sales.” *Management Science* 57, no. 8 (2011): 1373-1386.

Brynjolfsson, Erik, Yu Jeffrey Hu, and Michael D Smith. “Consumer surplus in the digital economy: Estimating the value of increased product variety at online booksellers.” *Management Science* 49, no. 11 (2003): 1580-1596.

Brynjolfsson, Erik, Yu Jeffrey Hu, and Michael D Smith. “From niches to riches: The anatomy of the long tail.” *MIT Sloan Management Review* 47 (2006): 67-71.

Brynjolfsson , Erik, Yu Jeffrey Hu, and Michael D Smith. “A Longer Tail?: Estimating the Shape of Amazon's Sales Distribution Curve in 2008. In.” *Workshop on Information Systems and Economics (WISE)*, 2009.

Brynjolfsson, Erik, Yu Jeffrey Hu, and Michael D Smith. “Research commentary-long tails vs. superstars: The effect of information technology on product variety and sales concentration patterns.” *Information Systems Research* 21, no. 4 (2010): 736-747.

Canhoto, Ana Isabel, and Moira Clark. “Customer service 140 characters at a time: The users' perspective.” *Journal of marketing Management* 29, no. 5-6 (2013): 522-544.

Celenk, M. “A color clustering technique for image segmentation.” *Computer Vision, Graphics, and Image Processing* 52, no. 2 (1990): 145-170.

Cha, Meeyoung, P. Rodriguez, J. Crowcroft, S. Moon, and X. Amatriain. “Watching Television Over an IP Network.” *Proceedings of the 8th ACM SIGCOMM conference on Internet measurement.* New York: ACM, (2008): 71-84.

Chakrabarti, S., M. Van den Berg, and B. Dom. “Focused crawling: a new approach to topic-specific Web resource discovery.” *Computer Networks* 31, no. 11 (1999): 1623-1640.

Chan, Kimmy Wa, and Stella Yiyan Li. “Understanding consumer-to-consumer interactions in virtual communities: The salience of reciprocity.” *Journal of Business Research* 63, no. 9 (2010): 1033-1040.

Chang, R.M., R. J. Kauffman, and Y. Kwon. “Understanding the paradigm shift to computational social science in the presence of big data.” *Decision Support Systems* 63 (2014): 67-80.

Chatterjee, Patrali. “Online review: do consumers use them?” *Advances in Consumer Research* 28 (2001): 129–133.

Chen, Y., S. Alspaugh, and R. Katz. “Interactive analytical processing in big data systems: A cross-industry study of mapreduce workloads.” *Proceedings of the VLDB Endowment* 5, no. 12 (2012): 1802-1813.

Chen, Y.L., K. Tang, C.C. Wu, and R.Y. Jheng. “Predicting the influence of users’ posted information for eWOM advertising in social networks.” *Electronic Commerce Research and Applications* 13, no. 6 (2014): 431-439.

Cheong, Hyuk Jun, and Margaret A. Morrison. “Consumers’ reliance on product information and recommendations found in UGC.” *Journal of Interactive Advertising* 8, no. 2 (2008): 38-49.

Cheng, Xiufang, and Meihua Zhou. “Study on effect of ewom: A literature review and suggestions for future research.” *International Conference on Management and Service Science*, 2010: 1-4.

Cheung, C. M., and D. R. Thadani. “The impact of electronic word-of-mouth communication: A literature analysis and integrative model.” *Decision Support Systems* 54, no. 1 (2012): 461-470.

Cheung, M. Y., C. Luo, C.L. Sia, and H. Chen. “Credibility of Electronic Wordof-Mouth: Informational and Normative Determinants of On-line Consumer Recommendations.” (International Journal of Electronic Commerce) 13, no. 4 (2009): 9-38.

Chevalier, Judith A., and Dina Mayzlin. “The effect of word of mouth on sales: Online book reviews.” *Journal of Marketing Research* 43, no. 3 (2006): 345–354.

Chu, Shu-Chuan, and Yoojung Kim. “Determinants of consumer engagement in electronic word-of-mouth (eWOM) in social networking sites.” *International journal of Advertising* 30, no. 1 (2011): 47-75.

Cothey, V. “Web crawling reliability.” *Journal of the American Society for Information Science and Technology* 55, no. 14 (2004): 1228-1238.

Cui, G., H. K. Lui, and X. Guo. “The effect of online consumer reviews on new product sales.” *International Journal of Electronic Commerce* 17, no. 1 (2012): 39-58.

Curtin, R., S. Presser, and E. Singer. “Changes in telephone survey nonresponse over the past quarter century.” *Public opinion quarterly* 69, no. 1 (2005): 87-98.

Clauset, Aaron, Cosma Rohilla Shalizi, and Mark EJ Newman. “Power-law distributions in empirical data.” *SIAM review* 51, no. 4 (2009): 661-703.

Drăgulescu, A., and V.M. Yakovenko. “Exponential and power-law probability distributions of wealth and income in the United Kingdom and the United States.” *Physica A: Statistical Mechanics and its Applications* 299, no. 1 (2001): 213-221.

Eynon, Rebecca. “The rise of Big Data: what does it mean for education, technology, and media research?” *Learning, Media and Technology* 38, no. 3 (2013): 237-240.

Elberse, Anita. “Should you invest in the long tail?” *Harvard business review* 86, no. 7/8 (2008): 88-96.

Elberse, Anita, and Felix Oberholzer-Gee. “Superstars and underdogs: An examination of the long tail phenomenon in video sales.” *Hardvard Business School* (Division of Research, Harvard Business School) Working Paper Series (2007): 07-015.

Engel, J.F., R.J. Kegerreis, and R.D. Blackwell. “Word-of-mouth communication by the innovator.” *Journal of Marketing* 33, no. 3 (1969): 15-19.

Erkan, Ismail, and Chris Evans. “The influence of eWOM in social media on consumers’ purchase intentions: An extended approach to information adoption.” *Computers in Human Behavior* 61 (2016): 47-55.

Dwyer, Paul. “Measuring the value of electronic word of mouth and its impact in consumer communities.” *Journal of Interactive marketing* 21, no. 2 (2007): 63-79.

Duan, Wenjing, Bin Gu, and Andrew B. Whinston. “Do online reviews matter?—An empirical investigation of panel data.” *Decision Support Systems* 45, no. 4 (2008): 1007-1016.

Duda, R.O., and P.E. Hart. *Pattern classification and scene analysis.* Vol. 3. New York: Wiley, 1973.

Davenport, T.H., and D.J. Patil. “Data scientist.” *Harvard business review* 90 (2012): 70-76.

Daugherty, T., M.S Eastin , and L. Bright. “Exploring consumer motivations for creating user-generated content.” *Journal of Interactive Advertising* 8, no. 2 (2008): 16-25.

De Leeuw, E., and W. De Heer. “Trends in household survey nonresponse: A longitudinal and international comparison.” *Survey nonresponse*, 2002: 41-54.

Dellarocas, C. “The digitization of word of mouth: promise and challenges of online feedback mechanisms.” *Management Science* 49, no. 10 (2003): 1407–1424.

Demchenko, Y., P. Grosso, C. De Laat, and P. Membrey. “Addressing big data issues in scientific data infrastructure.” *2013 International Conference on Collaboration Technologies and Systems (CTS)* (IEEE), May 2013: 48-55.

Dittrich, J., and J.A. Quiané-Ruiz. “Efficient big data processing in Hadoop MapReduce.” *Proceedings of the VLDB Endowment* 5, no. 12 (2012): 2014-2015.

Fan, W., and A. Bifet. “Mining big data: current status, and forecast to the future.” *ACM sIGKDD Explorations Newsletter* 14, no. 2 (2013): 1-5.

Fedi, M., T. Quarta, and A. De Santis. “Inherent power-law behavior of magnetic field power spectra from a Spector and Grant ensemble.” *Geophysics* 62, no. 4 (1997): 1143-1150.

Forman, C., A. Ghose , and B. Wiesenfeld. “Examining the relationship between reviews and sales: the role of reviewer identity information in electronic markets.” *Information Systems Research* 19, no. 3 (2008): 291–313.

Gu, Bin, Qian Tang , and Andrew B. Whinston . “The influence of online word-of-mouth on long tail formation.” *Decision Support Systems* 56 (2013): 474–481.

Gupta, P., and J. Harris. “How e-WOM recommendations influence product consideration and quality of choice: a motivation to process information perspective.” *Journal of Business Research* 63, no. 9 (2005): 1041–1049.

Gutenberg, B., and C.F. Richter. “Earthquake magnitude, intensity, energy, and acceleration (second paper).” *Bulletin of the seismological society of America,* 46, no. 2 (1956): 105-145.

Ghose, A., and P.G. Ipeirotis. “Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics.” *Transactions on Knowledge and Data Engineering* (IEEE) 23, no. 10 (2011): 1498-1512.

Goldsmith, Ronald E. “Electronic word-of-mouth.” *Encyclopedia of e-commerce, e-government and mobile commerce*, 2006: 408-412.

Goldsmith, Ronald E., and David Horowitz . “Measuring motivations for online opinion seeking.” *Journal of interactive advertising* 6, no. 2 (2006): 2-14.

Gonzalez-Rodriguez, M. R., M. R. Martinez-Torres, and S. L. Toral. “Monitoring travel-related information on social media through sentiment analysis.” *Proceedings of the 2014 IEEE/ACM 7th International Conference on Utility and Cloud Computing.* IEEE Computer Society, 2014. 636-641.

Gomes, C.P., B. Selman, N. Crato, and H. Kautz. “Heavy-tailed phenomena in satisfiability and constraint satisfaction problems.” *Journal of automated reasoning* 24, no. 1-2 (2000): 67-100.

Goodman, N.R. “Statistical analysis based on a certain multivariate complex Gaussian distribution (an introduction).” *The Annals of mathematical statistics* 34, no. 1 (1963): 152-177.

Gottlob, G., C. Koch, and R. Pichler. “The complexity of XPath query evaluation.” *Proceedings of the twenty-second ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (ACM), June 2003: 179-190.

Hennig-Thurau, Thorsten, K. P. Gwinner, G. Walsh, and D.D. Gremler. “Electronic word-of-mouth via consumer-opinion platforms: What motivates consumers to articulate themselves on the Internet?” *Journal of interactive marketing* 18, no. 1 (2004): 38-52.

Hennig-Thurau, T., and G. Walsh. “Electronic word-of-mouth: Motives for and consequences of reading customer articulations on the Internet.” *International Journal of Electronic Commerce* 8, no. 2 (2003): 51-74.

Huang, P., N.H. Lurie, and S. Mitra. “Searching for experience on the Web: An empirical examination of consumer behavior for search and experience goods.” *Journal of Marketing* 73, no. 2 (2009): 55–69.

Huberman, B.A., and F. Wu. “Bootstrapping the long tail in peer to peer systems.” *Managing Complexity: Insights, Concepts, Applications* (Springer Berlin Heidelberg), 2008: 263-272.

Huberman, Bernardo A. *The laws of the Web: Patterns in the ecology of information.* Cambridge, Massachusetts: MIT Press, 2003.

Hur, W. M., K. H. Ahn, and M. Kim. “Building brand loyalty through managing brand community commitment.” *Management Decision* 49, no. 7 (2011): 1194-1213.

Hartigan, J.A., and M.A. Wong. “Algorithm AS 136: A k-means clustering algorithm.” *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, no. 1 (1979): 100-108.

Heidorn, P. Bryan. “Shedding light on the dark data in the long tail of science.” *Library Trends* 57, no. 2 (2008): 280-299.

Herr, P.M., F.R. Kardes, and J. Kim. “Effects of word-of-mouth and product-attribute information on persuasion: an accessibility-diagnosticity perspective.” *Journal of Consumer Research* 17, no. 4 (1991): 454-462.

Jagadish, H. V., Gehrke, J., Labrinidis, A., Papakonstantinou, Y., Patel, J. M., Ramakrishnan, R., & Shahabi, C. “Big data and its technical challenges.” *Communications of the ACM* 57, no. 7 (2014): 86-94.

Jansen, Bernard J., M. Zhang, K. Sobel, and A. Chowdury. “Twitter power: Tweets as electronic word of mouth.” *Journal of the American society for information science and technology* 60, no. 11 (2009): 2169-2188.

Jiménez, F.R., and N.A. Mendoza. “Too popular to ignore: The influence of online reviews on purchase intentions of search and experience products.” *Journal of Interactive Marketing* 27, no. 3 (2013): 226-235.

Kaisler, S., F. Armour, J.A. Espinosa, and W. Money. “Big data: issues and challenges moving forward.” *46th Hawaii International Conference on System Sciences (HICSS)* (IEEE), January 2013: 995-1004.

Katz , Elihu, and Paul F. Lazarsfeld. *Personal influence: The part played by people n the flow of mass communications.* Glencoe, Illinois: The Free Press, 1955.

Katal, A., M. Wazid, and R.H. Goudar. “Big data: issues, challenges, tools and good practices.” *Sixth International Conference on Contemporary Computing (IC3)* (IEEE), August 2013: 404-409.

Kietzmann, J., and A. Canhoto. “Bittersweet! Understanding and managing electronic word of mouth.” *Journal of Public Affairs* 13, no. 2 (2013): 146-159.

King, G. “Restructuring the Social Sciences: Reflections from Harvard's Institute for Quantitative Social Science.” *PS: Political Science & Politics* 47, no. 1 (2014): 165-172.

Kim, Hee Woong, and Sumeet Gupta. “A comparison of purchase decision calculus between potential and repeat customers of an online store.” *Decision Support Systems* 47, no. 4 (2009): 477-487.

Klein, L. “Evaluating the Potential of Interactive Media Through a New Lens: Search Versus Experience Goods.” *Journal of Business Research* 41, no. 3 (1998): 195-203.

Kozinets, R.V., K. De Valck, A.C. Wojnicki, and S.J. Wilner. “Networked narratives: Understanding word-of-mouth marketing in online communities.” *Journal of marketing* 74, no. 2 (2010): 71-89.

Kodinariya, T. M., and P. R. Makwana. “Review on determining number of Cluster in K-Means Clustering.” *International Journal* 1, no. 6 (2013): 90-95.

Krahe, T. E., R.N. El-Danaf, E.K. Dilger, S.C. Henderson, and W. Guido. “Morphologically distinct classes of relay cells exhibit regional preferences in the dorsal lateral geniculate nucleus of the mouse.” *The Journal of Neuroscience* 31, no. 48 (2011): 17437-17448.

Ku , Y.C., C.P. Wei, and H.W. Hsiao. “To whom should I listen? Finding reputable reviewers in opinion-sharing communities.” *Decision Support Systems* 53, no. 3 (2012): 534-542.

Kumar, Chetan, John B. Norris, and Yi Sun. “Location and time do matter: A long tail study of website requests.” *Decision Support Systems* 47, no. 4 (2009): 500-507.

Kuri-Morales, A., and F. Rodríguez-Erazo. “A search space reduction methodology for data mining in large databases.” *Engineering Applications of Artificial Intelligence* 22, no. 1 (2009): 57-65.

Lee, Jung, Jae-Nam Lee, and Hojung Shin. “The long tail or the short tail: The category-specific impact of eWOM on sales distribution.” *Decision Support Systems* 51, no. 3 (2011): 466-479.

Lee, N.J., J. Brown, and A.J. Broderick. “Extending social network theory to conceptualize on-line word-of-mouth communication.” *Journal of Interactive Marketing* 21, no. 3 (2007): 2-19.

Lee, Mira, and Seounmi Youn. “Electronic word of mouth (eWOM) How eWOM platforms influence consumer product judgement.” *International Journal of Advertising* 28, no. 3 (2009): 473-499.

Lee, T.M. “Information direction, website reputation and eWOM effect: A moderating role of product type.” *Journal of Business research* 62, no. 1 (2009): 61-67.

Leskovec, J., A. Rajaraman, and J.D. Ullman. *Mining of massive datasets.* United Kingdom: Cambridge University Press, 2014.

Lew, Alan A. “Long tail tourism: New geographies for marketing niche tourism products.” *Journal of Travel & Tourism Marketing* 25, no. 3-4 (2008): 409-419.

Li , X., and L.M. Hitt. “Self selection and information role of online product reviews.” *Information Systems Research* 19, no. 4 (2008): 456–474.

Li, M., L. Huang, C.H. Tan , and K. K. Wei. “Helpfulness of online product reviews as seen by consumers: Source and content features.” *International Journal of Electronic Commerce* 17, no. 4 (2013): 101-136.

Lis, Bettina. “In eWOM We Trust.” *Business & Information Systems Engineering* 5, no. 3 (2013): 129-140.

Liu, Q.B., E. Karahanna, and R.T. Watson. “Unveiling user-generated content: Designing websites to best present customer reviews.” *Business Horizons* 54, no. 3 (2011): 231-240.

Llorente, A., M. Garcia-Herranz, M. Cebrian, and M. Moro. “Social media fingerprints of unemployment.” *PloS one* 10, no. 5 (2015): e0128692.

Najork, M. “Web crawler architecture.” *Encyclopedia of Database Systems* (Springer US), 2009: 3462-3465.

Nardo , M., M. Petracco, and M. Naltsidis. “Walking down wall street with a tablet: a survey of stock market predictions using the web.” *Journal of Economic Surveys* 30, no. 2 (2016): 356-369.

Nelson, Phillip. “Advertising as information.” *Journal of political economy* 82, no. 4 (1974): 729-754.

Nelson, Phillip. “Information and Consumer Behavior.” *Journal of Political Economy* 78, no. 2 (1970): 311–329.

Ng, R.T., and J. Han. “Efficient and Effective Clustering Methods for Spatial Data Mining.” *Proc. 20th International Conference on Very Larga Data Bases*, September 1994: 144-155.

Newman, M.E. “Power laws, Pareto distributions and Zipf's law.” *Contemporary physics* 46, no. 5 (2005): 323-351.

Madhulatha , T. Soni. “An overview on clustering methods.” *Journal of Engineering IOSR* 2, no. 4 (April 2012): 719-725.

Mahanti, Aniket, N. Carlsson, M. Arlitt, and C. Williamson. “A tale of the tails: Power-laws in internet measurements.” *IEEE Network* 27, no. 1 (2013): 59-64.

Manovich, L. “Trending: The promises and the challenges of big social data.” *Debates in the digital humanities* 2 (2011): 460-475.

Martínez-Torres, M.R. “Analysis of open innovation communities from the perspective of Social Network Analysis.” *Technology Analysis & Strategic Management* 26, no. 4 (2014): 435-451.

Martínez-Torres, M.R., and M.C. Díaz-Fernández. “A study of global and local visibility as web indicators of research production.” *Research Evaluation* 22, no. 3 (2013): 157-168.

McCloskey, D.N. “From methodology to rhetoric.” *The Rhetoric of Economics* ((University of Wisconsin Press), 1985: 20–35.

McKelvey, Bill, and Pierpaolo Andriani. “Why Gaussian statistics are mostly wrong for strategic organization.” *Strategic Organization* 3, no. 2 (2005): 219-228.

Michael, K., and K. W. Miller. “Big data: New opportunities and new challenges [guest editors' introduction].” *Computer* 46, no. 6 (2013): 22-24.

Moe , Wendy W., and Michael Trusov. “The Value of Social Dynamics in Online Product Ratings Forums.” *Journal of Marketing Research* 48, no. 3 (2011): 444-456.

Morales-Arroyo, M., and T. Pandey. “Identification of critical eWOM dimensions for music albums.” *IEEE International Conference on Management of Innovation and Technology (ICMIT)* (IEEE), 2010: 1230-1235.

Mudambi, S.M., and D. Schuff. “What makes a helpful online review? A study of customer reviews on Amazon.com.” *MIS Quarterly* 34, no. 1 (2010): 185–200.

Muniz, A.M., and H.J. Schau. “Religiosity in the abandoned Apple Newton brand community.” *Journal of Consumer Research* 31, no. 4 (2005): 737-47.

Muniz, Albert M., and Thomas C. O'guinn. “Brand community.” *Journal of consumer research* 27, no. 4 (2001): 412-432.

Odić, Ante, M. Tkalčič, J.F. Tasič, and A. Košir. “Predicting and detecting the relevant contextual information in a movie-recommender system.” *Interacting with Computers* 25, no. 1 (2013): 74-90.

Okuyama, K., M. Takayasu, and H. Takayasu. “Zipf's law in income distribution of companies.” *Physica A: Statistical Mechanics and its Applications* 269, no. 1 (1999): 125-131.

Olmedilla, M., M.R. Martinez-Torres, and S. Toral. “Examining the power-law distribution among eWOM communities: a characterisation approach of the Long Tail.” *Technology Analysis & Strategic Management* 28, no. 5 (2016): 601-613.

Palmer, C.R., and C. Faloutsos. “Density biased sampling: an improved method for data mining and clustering.” *ACM* 29, no. 2 (2000): 82-92.

Pan, Bing, Tanya MacLaurin, and John C. Crotts. “Travel blogs and the implications for destination marketing.” *Journal of Travel Research* 46, no. 1 (2007): 35-45.

Pant, G., P. Srinivasan, and F. Menczer. “Crawling the web.” *Web Dynamics* (Springer Berlin Heidelberg), 2004: 153-177.

Park, C., and T.M. Lee. “Information direction, website reputation and eWOM effect: A moderating role of product type.” *Journal of Business research* 62, no. 1 (2009): 61-67.

Park, D. H., J. Lee, and I. Han. “The effect of on-line consumer reviews on consumer purchasing intention: the moderating role of involvement.” *International Journal of Electronic Commerce* 11, no. 4 (2007): 125–148.

Peltier, Stéphanie, and François Moreau. “Internet and the ‘Long Tail versus superstar effect’debate: evidence from the French book market.” *Applied Economics Letters* 19, no. 8 (2012): 711-715.

Peterson, Martin. *An introduction to decision theory.* Cambridge: Cambridge University Press, 2009.

Phelps, J.E., R. Lewis, L. Mobilio , and D. Perry. “Viral marketing or electronic word-of-mouth advertising: Examining consumer responses and motivations to pass along email.” *Journal of advertising research* 44, no. 4 (2004): 333-348.

Phillipa, Gill, M. Arlitt, Z. Li, and A. Mahanti. “Youtube traffic characterization: a view from the edge.” *Proceedings of the 7th ACM SIGCOMM conference on Internet measurement.* New York: ACM, 2007. 15-28.

Pickering, G., J.M. Bull, and D.J. Sanderson. “Sampling power-law distributions.” *Tectonophysics* 248, no. 1 (1995): 1-20.

Pinkerton, B. “Finding what people want: Experiences with the WebCrawler.” *Proceedings of the Second International World Wide Web Conference* 94 (October 1994): 17-20.

Prendergast, Gerard, David Ko, and V. Yuen Siu Yin. “Online word of mouth and consumer purchase intentions.” *International Journal of Advertising* 29, no. 5 (2005): 687-708.

Qiu, Lingyun, Jun Pang, and Kai H. Lim. “Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence.” *Decision Support Systems* 54, no. 1 (2012): 631-643.

Raj Devasagayam, P., C.L. Buff, T.W. Aurand, and K.M. Judson. “Building brand community membership within organizations: a viable internal branding alternative?” *Journal of Product & Brand Management* 19, no. 3 (2010): 210-217.

Redner, Sidney. “How popular is your paper? An empirical study of the citation distribution.” *The European Physical Journal B-Condensed Matter and Complex Systems* 4, no. 2 (1998): 131-134.

Reips, U. D., and P. Garaizar. “Mining twitter: A source for psychological wisdom of the crowds.” *Behavior research methods* 43, no. 3 (2011): 635-642.

Rettberg, Jill Walker. *Blogging.* Cambridge: Polity Press, 2008.

Rosenblatt, M. “A central limit theorem and a strong mixing condition.” *Proceedings of the National Academy of Sciences* 42, no. 1 (1956): 43-47.

Schindler, Robert M., and Barbara Bickart. “Published word of mouth: Referable, consumer-generated information on the Internet.” In *Online Consumer Psychology: Understanding and Influencing Consumer Behavior in the Virtual World*, by Karen A. Machleit, Richard Yalch Curtis P. Haugtvedt, 35-61. New Jersey: Psychology Press, 2005.

Schmallegger, Doris, and Dean Carson. “Blogs in tourism: Changing approaches to information exchange.” *Journal of vacation marketing* 14, no. 2 (2008): 99-110.

Senecal, S., and J. Nantel. “The influence of online product recommendations on consumers’ online choices.” *Journal of retailing* 80, no. 2 (2004): 159-169.

Seyfi, A., A. Patel, and J.C. Júnior. “Empirical evaluation of the link and content-based focused Treasure-Crawler.” *Computer Standards & Interfaces* 44 (2016): 54-62.

Smith , A.N., E. Fischer, and C. Yongjian. “How does brand-related user-generated content differ across YouTube, Facebook, and Twitter?” *Journal of Interactive Marketing* 26, no. 2 (2012): 102-113.

Sridhar Balasubramanian, Vijay Mahajan. “The economic leverage of the virtual community.” *International Journal of Electronic Commerce* 5, no. 3 (2001): 103-138.

Standifird, S.S. “Reputation and e-commerce: eBay auctions and the asymmetrical impact of positive and negative ratings.” *Journal of management* 27, no. 3 (2001): 279-295.

Steffes, E. M., and L. E. Burgee. “Social ties and online word of mouth.” *Internet Research* 19, no. 1 (2009): 42-59.

Strutton, David, David G. Taylor, and Kenneth Thompson. “Investigating generational differences in e-WOM behaviours: for advertising purposes, does X= Y?” *International Journal of Advertising* 30, no. 4 (2011): 559-586.

Sussan, Fiona, Stephen Gould, and Suri Weisfeld-Spolter. “Location, location, location: The relative roles of virtual location, online word-of-mouth (eWOM) and advertising in the new-product adoption process.” *Advances in Consumer Research* 33 (2006): 649.

Syed, A., K. Gillela, and C. Venugopal. “The future revolution on big data.” (Future) 2, no. 6 (2013): 2446-2451.

Taborek, P., R.N. Kleiman, and D.J. Bishop. “Power-law behavior in the viscosity of supercooled liquids.” *Physical Review B* 34, no. 3 (1986): 1835.

Teutle , A. R. M. “Twitter: Network properties analysis.” *20th International Conference on Electronics, Communications and Computer (CONIELECOMP)* (IEEE), February 2010: 180-186.

Thorson, Kjerstin S., and Shelly Rodgers. “Relationships between blogs as eWOM and interactivity, perceived interactivity, and parasocial interaction.” *Journal of Interactive Advertising* 6, no. 2 (2006): 5-44.

Trusov, Michael, Randolph E. Bucklin, and Koen Pauwels. “Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site.” *Journal of marketing* 73, no. 5 (2009): 90-102.

Tucker, C., and J. Zhang. “How does popularity information affect choices? A field experiment.” *Management Science* 57, no. 5 (2011): 828-842.

Twarakavi, N.K., J. Šimůnek, and M.G. Schaap. “Can texture- based classification optimally classify soils with respect to soil hydraulics?” *Water resources research* 46, no. 1 (2010).

Vicknair, C., M. Macias, Z. Zhao, X. Nan, Y. Chen, and D. Wilkins. “A comparison of a graph database and a relational database: a data provenance perspective.” *Proceedings of the 48th annual Southeast regional conference* (ACM) 42 (April 2010).

Virkar, Y., and A. Clauset. “Power-law distributions in binned empirical data.” *The Annals of Applied Statistics* 8, no. 1 (2014): 89-119.

Wang, J., and Y. Guo. “Scrapy-based crawling and user-behavior characteristics analysis on Taobao.” *International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)* (IEEE), October 2012: 44-52.

Weathers, Danny, Subhash Sharma, and Stacy L. Wood. “Effects of online communication practices on consumer perceptions of performance uncertainty for search and experience goods.” *Journal of Retailing* 83, no. 4 (2007): 393-401.

Westbrook, R.A. “Product/consumption-based affective responses and postpurchase processes.” *Journal of Marketing Research* 24, no. 3 (1987): 258-270.

Xun, Jiyao, and Jonathan Reynolds. “Applying netnography to market research: The case of the online forum.” *Journal of Targeting, Measurement and Analysis for Marketing* 18, no. 1 (2010): 17-31.

Yang, Wen, Yun Kuei Huang, and Yu Hsiang Lin. “Study of Comments on Official Movie Blogs.” *International Journal of Electronic Business Management* 34, no. 3 (2009): 201-210.

Yang, Shuang. “Effects of information quality and source credibility on EWOM adoption in context of virtual community.” *International Conference on Management Science and Engineering*, 2013: 194-200.

Yeap, J.A., J. Ignatius, and T. Ramayah. “Determining consumers’ most preferred eWOM platform for movie reviews: A fuzzy analytic hierarchy process approach.” *Computers in Human Behavior* 31 (2014): 250-258.

Zacharia, Giorgos, Alexandros Moukas, and Pattie Maes. “Collaborative reputation mechanisms for electronic marketplaces.” *Decision Support Systems* 29, no. 4 (2000): 371-388.

Zhu, F., and X. Zhang. “Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics.” *Journal of marketing* 74, no. 2 (2010): 133-148.

Zhang, Jason Q., Georgiana Craciun, and Dongwoo Shin. “When does electronic word-of-mouth matter? A study of consumer product reviews.” *Journal of Business Research* 63, no. 12 (2010): 1336-1341.

Statutory declaration of original authorship

I hereby declare that I have authored this doctoral thesis independently, that I have not used other than the declared sources, and I have explicitly marked all material, which has been quoted either literally or by content from the used sources. This doctoral thesis has not been previously submitted, in identical or similar form, to another academic institution, nor has it yet been published.

Seville, 06/12/2016

_____

Sgd.: María Olmedilla Fernández