A long tail study of eWOM communities

M. Olmedilla, M. R. Martinez-Torres, S. L. Toral

***Abstract*—**Electronic Word-Of-Mouth (eWOM) communities represent today an important source of information in which more and more customers base their purchasing decisions. They include thousands of reviews concerning very different products and services posted by many individuals geographically distributed all over the world. Due to their massive audience, eWOM communities can help users to find the product they are looking for even if they are less popular or rare. This is known as the long tail effect, which leads to a larger number of lower-selling niche products. This paper analyzes the long tail effect in a well-known eWOM community and defines a tool for finding niche products unavailable through conventional channels.

***Keywords*—**eWOM, Online user reviews, Long tail theory, Product categorization, Social Network Analysis.

I. Introduction

On line channels like eWOM communities are facilitating the distribution of a wide range of products and services. They provide abundant, and objective product information that is influencing customers' decision making [1]-[2]. Through eWOM, users can freely post their reviews about any product or service, and share this review with thousands of other or potential users all over the world. eWOM communities also facilitate the interaction among users, so they can share their experiences and also comment or rate other users' reviews [3]. Typically, eWOM systems provide an aggregated rating for each product based on all reviewers' scores to facilitate a quick and overall impression of the product [4]. Gaining consumers' trust is also a key factor in eWOM. Consumers need to be confident about what other users are posting. For this purpose, eWOM systems introduce some metrics about the reputation of users. For instance, the quality of the review can also be scored by other users, leading to an individual score of each user as a reviewer. Some eWOM reputation models are based on the idea of trust networks, which refer to a network of reviewers whose reviews and ratings are considered valuable by other community members [5]. This is the case of well-known eWOM communities such as Epinions or Ciao.

The importance of eWOMs as a new marketing channel has been well recognized in prior studies [6]. Internet is influencing consumer decisions and even changing consumption habits. As a difference to conventional channels, the potential audience of shared opinions is huge, and it is well known that Internet is a perfect platform for distribution and transaction of products within the "long tail" [7]. The long tail refers to the sales of non-hit or niche products that account for a significant portion of overall product sales. These products can collectively comprise a market share that rivals or exceeds the relatively few current bestsellers but only if the distribution channel is large enough [8]. eWOMs, the same than other information sharing mechanisms through the Web, enables the discovery of new and niche products and encourages the long tail formation.

This paper analyzes and mathematically measures the long tail within eWOM, and proposes a methodology for detecting niche products within the long tail.

The rest of the paper is organized as follows. Section II discusses the related work. Section III introduces the proposed methodology. Section IV describes the case study and the data collection process. Section V presents the empirical results and Section VI concludes the study.

II. Related Work

The long tail model was first coined by Chris Anderson [2], who used this model to explain the success of Amazon book and Netflix DVD rental recommendations system to promote obscure products. The proliferation of Internet and Web 2.0 has reduced the search and distributed costs in the new digital economy [9]. Before the Internet age, marketplaces needed to be located in geographically relevant places as they were only able to serve local clients. As a result, retailers were mainly focused on hits able to attract a large amount of customers rather than targeting niche products, with a much lower potential audience. However, the inexpensive online media and reduced distribution costs have lowered the barriers of entrance [10]. Initially, the Internet was seen as an opportunity for marketers to interact with customers [11]. But it became clear very quickly that customers were also able to interact with each other and share product information the same way they do offline [12]. Since then, the number of websites and opinions communities where users can read and write product reviews has increased year by year, changing consumption habits. As a result, consumers have access to a lot of information resources before making their buying decisions. But another consequence of the abundance of information is that many products or services that are not in the list of bestselling became now visible through eWOM. They are part of long tail and, although they do not have enough potential

M. Olmedilla is with the Business School, University of Seville, Seville, Spain (e-mail: mariaolmedilla@hotmail.com).

M. R. Martinez-Torres, is with the Business School, University of Seville, Seville, Spain (e-mail: rmtorres@us.es).

S. L. Toral is with the School of Engineering, University of Seville, Seville, Spain (corresponding author to provide phone: +34 955481293; fax: +34 9544487373; e-mail: storal@us.es).

customers when offered locally, the potential audience grows when offered at a world scale. Therefore, eWOM mechanisms increase the sales of the less popular products and facilitate the long tail phenomenon. Moreover, the tail will grow not only longer, as more obscure products are available, but also flatter, as consumers discover better suited products than those more popular [13].

However, some other studies have challenged this idea and claim that eWOM may inhibit the long tail phenomenon by promoting the sales of popular products with high ratings [14]. In [15], it is shown that the impact of eWOM on sales distribution is different across product types, leading to a long or a short tail depending on the case. For instance, on the one hand it has been found that the sales distribution of a retailer's online channel is less concentrated than that of its traditional channels, indicating that a reduction in search costs contributes to the long tail phenomenon [16]. But on the other hand other studies reveal the presence of both long tail and superstar phenomenon. That means that although unpopular products increase their sales, at the same time, an even smaller number of products account for the bulk of sales [17]. The superstar phenomenon is based on the idea that popular products are more likely to be recommended and purchased. It has been argued that the presence of product popularity information usually displayed in eWOMs significantly increases the sales of popular products, facilitating this superstar phenomenon [18]. This argument was tempered in [15], where authors conclude that the long tail appears with products with subjective evaluation standards. The common characteristic of such products is that people have various preferences and do not have any objective evaluation criteria for the products [19].

From a mathematical point of view, the long tail has been frequently modeled on the assumption that the data obeys a power-law distribution. It is widely accepted that customer demand across a product space takes the form of a power law [7], [20]. Many previous studies have identified such pattern related to the phrases the tourists use to describe destination image [21] or the web use of learning websites [22]. Zipf-like distribution, belonging to the family of discrete power law probability distributions has been also used to model visits to websites [23] and users' requests for online information [10].

This paper analyzes a well-known eWOM considering its internal categories of products and models the reviews posted for each one as a power law distribution. The goal is checking which categories exhibit a long tail behavior.

III. Methodology

The long tail is a manifestation of power-law relationships. A long tail exemplifies the statistical property that there are many more low-frequency events compared to a Gaussian distribution [24]. Mathematically, a power law probability distribution is given by (1):

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

where C and $\propto$ are positive constants. In practice, few empirical phenomena follow a power law distribution for all values of x. Usually, the power law applies only for values higher than some minimum value x_{min}. Then it is said that the tail of the distribution follows a power law. Fitting power laws to empirical data means estimating the scaling parameter $\propto$ the lower-bound on the scaling region x_{min}. However, the characterization of power laws is difficult due to the large number of fluctuations that occur in the tail of the distribution. Standard methods such as least-squares fitting are known to produce systematically biased estimates of the two parameters. In this paper we use the method defined in [25], based on maximum likelihood methods. The goodness of fit is given by the Kolmogorov-Smirnov statistic, (2):

$$D = \max_{x \geq x_{min}} |S(x) - P(x)| \tag{2}$$

Where $S(x)$ is the 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 in [25]. A p-value can be calculated to determine if the value of D is too high. The p-value quantifies the probability that the data were drawn from the power law distribution based on goodness of fit. If the p-value is much lower than one, the power law model can be ruled out. However, if the p-value is near 1, then the power law could be a plausible fit to the data.

IV. Case Study and Data Collection

Ciao.com is an eWOM website where registered users can critically review and rate products and services for the benefit of other consumers. It is one of the largest eWOM in Europe, with 1.3 million members that have written more than 7 million reviews. It is available free of charge, and registered users can write comments and score products using qualitative ratings that correspond to numerical values (currently, the website contains reviews on 1,4 million of products). Posted reviews can also be scored by other members.

Ciao website is structured through categories of products and services. Basically, there are 28 main categories, which in turn are subdivided in many subcategories. The 28 main categories are established by Ciao, while subcategories are created by users as long as they post and share reviews.

Data collection requires accessing the categories and subcategories in which registered users post their reviews. It is important to collect data aggregated by users in order to build the social network of the long tail, connecting those subcategories where a given user has posted his or her reviews. Ciao website contains a webpage for each registered user where some statistics about his or her past history are displayed. Obviously, registered users use an alias and no personal information is displayed. This member webpage contains, among other information, those categories and subcategories where each specific user has posted his or her reviews. The main limitation is that there is no index about registered users. As a result, data collection has been done as a

two-stage procedure. The first step consists in collecting members' webpages. For this purpose, a crawler program that follows the hyperlink structure of Ciao website has been developed in R. Basically, the crawler browses the website, storing those webpages corresponding to members and discarding the rest of them. The second step consists in extracting the list of categories and subcategories for each member stored in step 1. The function *readLines()* from the R base package, that reads data from a URL, was used to access each member webpage. However, webpages are formatted in HTML code, and accessed data contains both the webpage content and the HTML tags. Therefore, it is necessary to parse HTML file using the function *htmlParse()*, that generates an R structure representing the HTML tree. Once online webpages are available as an R structure, meaningful data (categories and subcategories for each member) can be easily identified using regular expressions that are also supported in R, for instance, in packages like XML.

V. Results

Data were collected from the website ciao.co.uk, which is the Ciao website in UK. Currently, there are about 45 thousand registered users in the UK. However, it is almost impossible to extract all of them by crawling the whole website, as this process will take an extraordinary amount of time. Therefore, the website has only be partially crawled by extracting a subset of 4574 registered users. For each user and using a program in R, it has been extracted the list of categories and subcategories of the posted reviews.

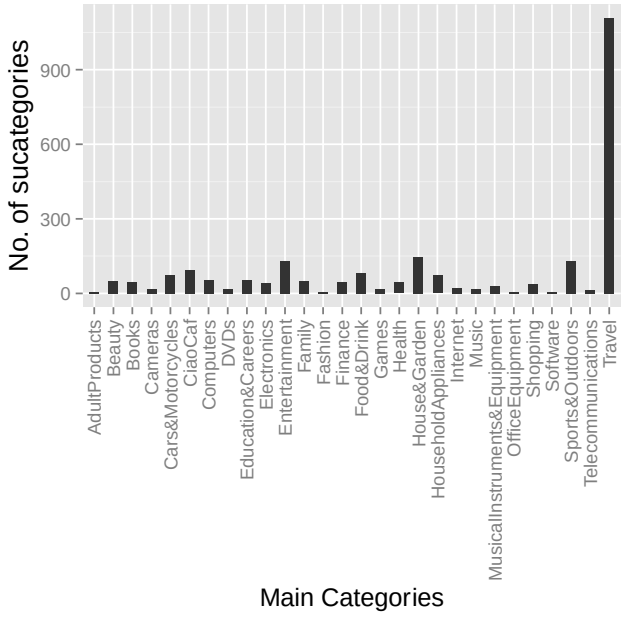

Fig. 1. Distribution of posted reviews per main categories.

The aggregated result is illustrated in Fig. *1*, which shows the distribution of posted reviews over the 28 main categories distinguished by Ciao. It can be observed that the number of reviews is not uniformly distributed over categories. On the contrary, the category Travel accumulates almost half of the total number reviews while some other categories only have a small number of reviews.

Table I details the power law adjustment and the long tail characterization for each of the 28 main categories of Ciao. The first column is the number of subcategories created by users posting reviews. The second and third column are the $\propto$ and x_{min} values of the fitted power law distribution. The fourth column is the goodness of fit, given by the distance *D* and the p-value. Finally, the last column is the length of the long tail calculated as the number of subcategories with a number of reviews below the value of x_{min}, that is, the number of subcategories that are not part of the fitted power law distribution.

TABLE I. POWER LAW AND LONG TAIL PARAMETERS OF THE 28 MAIN CATEGORIES OF CIAO.

Cat	SubCat	$\propto$	x_{min}	D (p)	Length
Adult Prod.	5	1.50	2	0.262 (0.28)	0
Beauty	50	1.71	91	0.115 (0.14)	11
Books	48	1.55	31	0.115 (0.12)	12
Cameras	19	1.60	6	0.126 (0.54)	2
Cars&…	75	2.13	13	0.125 (0.03)	42
CiaoCaf	94	1.73	20	0.089 (0.32)	60
Computers	53	2.36	55	0.082 (0.96)	32
DVDs	17	3.50	2319	0.200 (0.02)	12
Educ&…	55	1.69	2	0.112 (0.23)	21
Electronics	42	2.92	95	0.185 (0.25)	33
Entertainment	131	1.97	10	0.055 (0.89)	91
Family	50	1.68	37	0.139 (0.01)	15
Fashion	6	1.50	10	0.288 (0.52)	2
Finance	48	2.44	16	0.1159 (0.00)	30
Food&…	83	2.25	93	0.108 (0.16)	39
Games	19	3.50	580	0.241 (0.99)	12
Health	43	1.86	34	0.123 (0.11)	12
House&…	147	1.86	14	0.065 (0.39)	58
Household…	72	3.36	170	0.102 (0.89)	61
Internet	24	1.63	22	0.107 (0.80)	6
Music	16	1.71	78	0.179	5

				(0.43)	
MusicalInstr.	28	2.73	8	0.107	19
				(0.75)	
OfficeEquip.	7	1.54	9	0.194	1
				(0.70)	
Shopping	39	2.12	24	0.147	17
				(0.11)	
Software	6	1.50	6	0.170	1
				(0.92)	
Sports&…	132	2.88	17	0.060	102
				(0.86)	
Telecomm.	13	1.50	4	0.113	1
				(0.85)	
Travel	1108	2.04	6	0.028	690
				(0.30)	

Anchored in the statistical significance testing and results provided by the case, for each category if the calculated p-value is considerably lower than 0.1, then it is clear that such category does not to follow a power-law distribution. If on the other hand the resulting p-value is greater than 0.1, then in this case it might follow or not a power-law distribution.

Nevertheless, among all these categories gathered in TABLE *I* there are few where the power law distribution fits the long peak of the function, and it does not occur the long tail. Likewise, there are some categories, which do not fit a power law distribution; still those ensure a long tail formation.

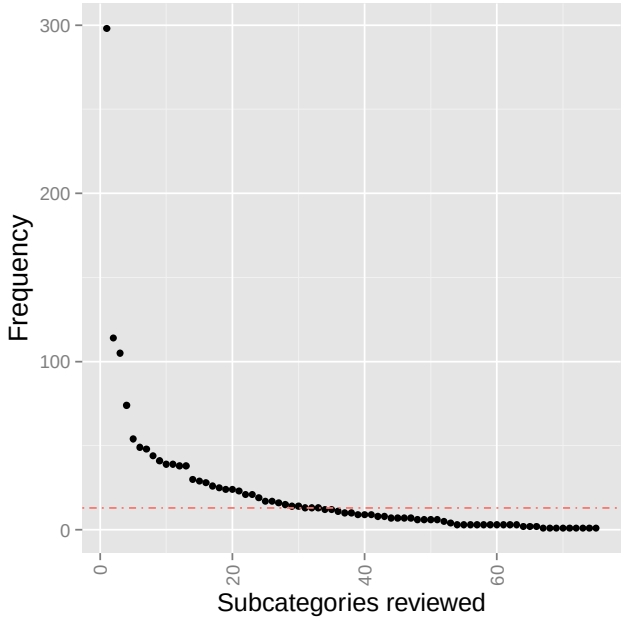

Fig. 2. Distribution of posted reviews per subcategories for main category Cars&Motorcycles.

For instance, in that regard, when observing the category Cars&Motorcycles from the TABLE *I*, according to its p-value such case does not fit a power law distribution, since there are not enough products highlighted as bestsellers. Conversely and as illustrated in Fig. *2*, it is possible to discern the number of products characterized by a long tail of length 42, which are below the x_{min} demarcated by a horizontal red line.

Additionally, the category Travel, which is depicted in *3*, constitutes a good example of the long tail phenomenon, since it obeys a power law distribution where a long tail appears and it is characterized by users' posted reviews having various and subjective preferences about their opinion on travelling.

Conversely, categories such as Telecommunications or Cameras, which are characterized by users' objective evaluation criteria of a product, obey a power law distribution with a long peak filled with popular products and a short tail.

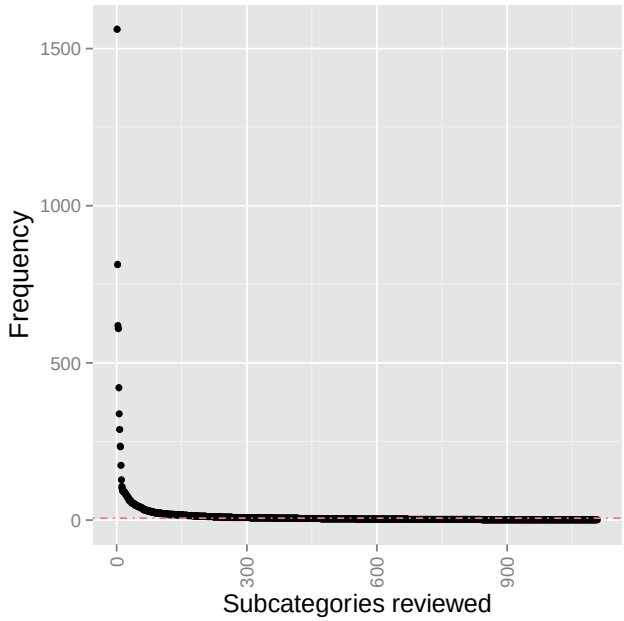

Fig. 3. Distribution of posted reviews per subcategories for main category Travel.

VI. CONCLUSION

This paper establishes an initial attempt to examine the power law distribution among all the categories enclosed in product reviews of an eWOM portal by analyzing whether there is a long tail characterization. In that respect, the described previous sections aim to explore if a distribution of 28 categories of data set gathered from the reviews in the eWOM portal Ciao UK fit a power law distribution and if so, to stipulate in which cases occur the long tail formation. The findings reveal that not all the categories fitting a power law distribution are characterized by a long tail, and alternatively some of those having a long tail do not fit a power law distribution. Furthermore, most of cases with subjective and personal evaluation standards encourage the long tail phenomenon, whereas those with more objective and impartial evaluation standards encourage the superstar phenomenon.

Further research can extend these findings by characterizing the long tail for each principal category of reviews in an eWOM portal, and defining a tool for identifying all the niche products across the long tail. The goal is to find some common patterns among niche products through the creation of social

network models, where nodes represent types of products and edges are connecting products that have received reviews from the same user.

ACKNOWLEDGMENT

This work was supported by the Consejería de Economía, Innovación, Ciencia y Empleo under the Research Project with reference P12-SEJ-328 and by the Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad under the Research Project with reference ECO2013-43856-R.

REFERENCES

[1] H.-W. Kim, S. Gupta, A comparison of purchase decision calculus between potential and repeat customers of an online store, *Decision Support Systems*, Vol. 47, Iss. 4, pp. 477–487, 2009.

[2] G. Zacharia, A. Moukas, P. Maes, Collaborative reputation mechanisms for electronic marketplaces, *Decision Support Systems*, Vol. 29, Iss. 4, pp. 371–388, 2000.

[3] F. J. Arenas-Marquez, M. R. Martinez-Torres, S. L. Toral, Electronic word of mouth communities from the perspective of Social Network Analysis, *Technology Analysis & Strategic Management*, Vol. 26, Iss. 8, pp. 927-942, 2014.

[4] L. Qiu, J. Pang, K. H. Lim, Effects of conflicting aggregated rating on eWOM review credibility and diagnosticity: The moderating role of review valence, *Decision Support Systems*, Vol. 54, Iss. 1, pp. 631–643, 2012.

[5] Y. C. Ku, C. P. Wei, H. W. Hsiao, To whom should I listen? Finding reputable reviewers in opinion-sharing communities, *Decision Support Systems*, Vol. 53, pp. 534–542, 2012.

[6] Y. Chen, J. Xie, Online consumer review: a new element of marketing communications mix, Management Science, Vol. 54, Iss. 3, pp. 477–491, 2008.

[7] C. Anderson, *Long Tail: Why the Future of Business is Selling Less of More*, Hyperion Books, New York, NY 2008.

[8] A. Odic, M. Tkalčič, J. F. Tasič, and A. Košir, Predicting and Detecting the Relevant Contextual Information in a Movie-Recommender System, Interacting with Computers, Vol. 25, no. 1, pp. 74-90, 2013.

[9] F. Zhu, X. Zhang, Impact of online consumer reviews on sales: the moderating role of product and consumer characteristics, *The Journal of Marketing*, Vol. 74, Iss. 2, pp. 133–148, 2010.

[10] C. Kumar, J. B. Norris, Y. Sun, Location and time do matter: A long tail study of website requests, *Decision Support Systems*, Vol. 47, pp. 500–507, 2009.

[11] F. Feather, Future consumer.com. Toronto: Warwick Publishing, 2000.

[12] M. Khammash, G. H. Griffiths, 'Arrivederci CIAO.com, Buongiorno Bing.com'—Electronic word-of-mouth (eWOM), antecedences and consequences, International Journal of Information Management, Vol. 31, pp. 82–87, 2011.

[13] A. Elberse, Should You Invest in the Long Tail?, *Harvard Business Review*, Vol. 86, no. 7/8, pp. 88-96, 2008.

[14] S. Standifird, Reputation and ecommerce: eBay auction and the asymmetrical impact of positive and negative ratings, *Journal of Management*, Vol. 27, Iss. 3, pp. 279–295, 2001.

[15] J. Lee, J. N. Lee, H. Shin, The long tail or the short tail: The category-specific impact of eWOM on sales distribution, *Decision Support Systems*, Vol. 51, pp. 466–479, 2011.

[16] E. Brynjolfsson, M.D. Smith, Y.J. Hu, Goodbye pareto principle, hello long tail: the effect of search costs on the concentration of product sales, in: MIT working paper, 2007.

[17] Elberse, F. Oberholzer-Gee, Superstars and underdogs: an examination of the long tail phenomenon in video sales, in: Harvard Business SchoolWorking Paper Series, 07–015, 2007.

[18] M. Sun, How does variance of product ratings matter? Management Science, Vol. 58, Iss. 4, pp. 696–707, 2012.

[19] B. Gu, Q. Tang, A. B. Whinston, The influence of online word-of-mouth on long tail formation, Decision Support Systems, Vol. 56, pp. 474–481, 2013.

[20] E. K. Clemons, G. Gao, Consumer informedness and diverse consumer purchasing behaviors: traditional mass-market, trading down, and trading out into the long tail, *Electronic Commerce Research and Applications*, Vol. 7, no. 1, pp. 3–17, 2008.

[21] B. Pan, X. R. Li, The long tail of destination image and online marketing, *Annals of Tourism Research*, Vol. 38, Iss. 1, pp. 132-152, 2011.

[22] X. Li, Y. Xu, Y. Zhang, J. Shi, Long Tail Distribution in the Web Usage of a Chinese Learning Website, 2012 International Symposium on Information Science and Engineering (ISISE), pp. 64-67, 2012.

[23] Q. Jiang, C.-H. Tan, C. W. Phang, J. Sutanto, K.-K. Wei, Understanding Chinese online users and their visits to websites: Application of Zipf's law, International Journal of Information Management, Vol. 33, Iss. 5, pp. 752-763, 2013.

[24] A. Mahanti, N. Carlsson, A. Mahanti, M. Arlitt, C. Williamson, A tale of the tails: Power-laws in internet measurements, *IEEE Network*, Vol. 27, Iss. 1, pp. 59-64, 2013.

[25] A. Clauset, C. R. Shalizi, M. E. J. Newman, Power-law distributions in empirical data, *SIAM Review*, Vol. 51, pp. 661-703, 2007