A First Look at the Long-Tail of YouTube

Abstract

There has been considerable academic and non-academic interest in the recent Long-Tail phenomenon, which refers to the shape of the distribution curve that emerges when very large numbers of previously uneconomical offerings become viable for the first time. While demand for these niche offerings is relatively low, when aggregated, they can potentially rival the popular but relatively limited number of mainstream offerings. Several forces shape the distribution of this Long-Tail, and can either lengthen, fatten or flatten the distribution. We conduct an empirical investigation on part of the Long-Tail of YouTube.com to study the impact of these three forces. We find evidence that a longer tail leads to a greater number of hits, which fattens the head of the Long-Tail. This fattening of the head in turn leads to a lengthening of the tail. We also find that a widespread distribution of eWOM can flatten the head of the Long-Tail.

Keywords: Long Tail, e-marketing, Youtube, eWOM, electronic word of mouth

1. Introduction and Literature Review

The term Long-Tail was coined by Chris Anderson in an article in Wired Magazine[1] in 2004 (Anderson 2004). Anderson noted that the ease and reduced costs with which consumers could find and purchase niche products meant that previously unviable products became viable for the first time. This has lead to the consumption curve to follow a long-tailed distribution, where a large number of niche products have few but viable sales. While demand for these niche offerings is relatively low, when aggregated, they can potentially rival the popular but relatively limited number of mainstream offerings. A benchmark of the level of inequality between mainstream (hits) and niche offerings is the Pareto Principle or 80/20 Rule, which postulates that 20% of offerings account for 80% of revenue and vice versa. Violations of this informal rule in favor of the niches are generally regarded as evidence of a Long-Tail distribution. There has been some evidence to this, although the violation of the Pareto Principle has been of a limited magnitude (Brynjolfsson et al. 2007). Apart from studying the emergence of the Long-Tail phenomenon, attention has now started to focus on the structure of the Long-Tail distribution as well (Anderson 2006; Dellarocas et al. 2007). The Long-Tail distribution shows the existence of a head and a tail[2]. The head consists of the most popular videos or hits, which benefit greatly from the herding effect, which is often caused by information cascades (Duan et al. 2005). Another cause of the herding effect is the network effect (Duan et al. 2005). The tail consists of all the non-hits or niche offerings, which are much less popular than the hits, but have strength of numbers. Anderson (2006) identified three important forces that act on the Long Tail distribution. These three forces *lengthen*, *fatten*, and *flatten* the distribution (Anderson 2006). From an academic viewpoint, researchers have classified these forces as supply side drivers and demand side drivers (Brynjolfsson et al. 2006).

[1] While the Long-Tail has been used in the literature earlier, it is the first article that referred to the phenomenon in ecommerce demand for niche products using this term and popularize it, and is thus widely credited to Anderson.

[2] The term Long-Tail refers to the shape of the distribution and includes both the head and the tail.

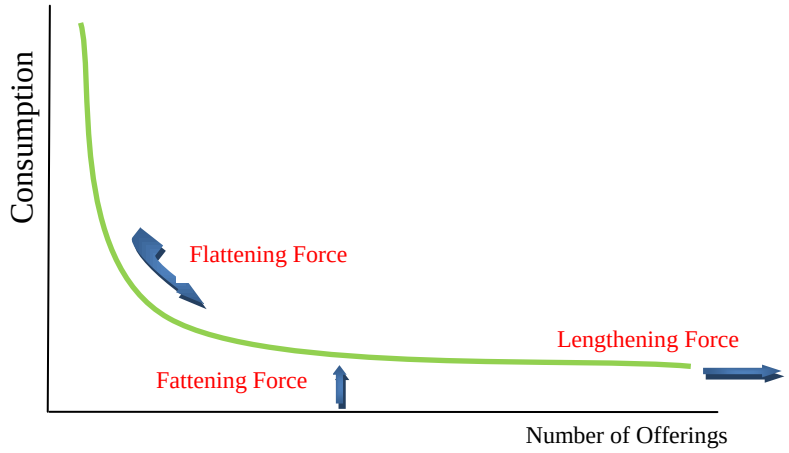

Figure 1 Three Forces acting upon the Long-Tail: Adapted from Anderson (2006)

1.1. Supply Side Forces

Supply side forces can be broken down into technological and economic drivers impacting production costs, inventory costs, and distribution costs (Brynjolfsson et al. 2006).

1.1.1. Production Costs.
Technological advances in the arena of ecommerce have dramatically reduced the cost of production, greatly expanding the domain of both offerings and producers (Bakos et al. 2000). As the cost of production decreases, the length of the tail increases as more producers are able to generate offerings, while at the same time the number of previously unviable offerings that are now viable increases as well (Anderson 2006; Brynjolfsson et al. 2006).

1.1.2. Inventory Costs
For brick-and-mortar stores, the stocking decisions are determined by the cost and limitation of shelf-space, and the geographical reach of the store (Brynjolfsson et al. 2006). However, the costs of maintaining large inventories are extremely low on the internet, especially since inventories of major Internet retailers are mostly virtual (Anderson 2006). In the case of digital content, the inventory consists of just one copy of each offering, which is then replicated on demand. The cost of cataloging is even more trivial, consisting of as little as a single entry in a database (Brynjolfsson et al. 2006).

1.1.3. Distribution Costs
Distribution costs are also much lower on the Internet, especially for digital content (Bakos 1998; Brynjolfsson et al. 2006). The disintermediation of middlemen online has also reduced distribution costs (Gallaugher 2002). This allows Internet firms to offer digital content to meet lower levels of demand which would have previously been uneconomical. For example, an Internet store that sold physical goods may not find it profitable to ship a single music track or short video clip individually due to the cost of distribution, but an Internet store selling digital content would face much lower distribution costs since a megabyte of bandwidth costs the same, whether it is utilized in discrete intervals or in a bundle. Thus the lowered distribution cost would make short video clips, single tracks and short e-books viable, thereby increasing the supply.

1.1.4. Impact on the Long-Tail

This increase in offerings has the effect of *lengthening* the Long-Tail distribution (Anderson 2006). This is particularly true with digital content, where inexpensive and off-the-shelf electronic devices and broadband connections have allowed nearly anyone to become a producer of content, although the impact is also evident on non-digital offerings such as physical books. Amazon.com, the leader in online book sales, has about 3 million books available for purchase, while Apple's iTunes, the leader in legitimate online music downloads, has over 6 million songs in its online catalog available for immediate purchase and download. However, both these online retailers are limited by the availability of offerings since they focus primarily on professionally created offerings. Anderson (2006) predicts that "democratizing production" and allowing amateurs to produce and offer content will have an explosive impact on the Long-Tail. Indeed, YouTube.com, the subject of this study and the most popular video-sharing site on the Internet, has a staggering 96 million videos available for instant viewing. YouTube has more videos offered under "Autos" than there are books offered by Amazon.com, and the "Sports" category has more offerings than all the tracks available for download on iTunes. Clearly YouTube has by far the longest tail on the Internet, thanks to a nearly unlimited supply of professional and amateur content.

1.2. Demand Side Forces

Reduced transaction and search costs for consumers are primarily responsible for generating ad enhancing Long-Tail demand (Brynjolfsson et al. 2006).

1.2.1. Reduced transaction costs

The Internet has reduced transaction costs significantly (Bakos 1998; Porter 2001), enabling consumers to access offerings that were previously priced out of range due to the high cost of obtaining the offering. The cost could be in terms of bandwidth expenses or the opportunity costs of waiting for a download. Cheap and fast broadband access has reduced transaction costs for consumers to the point of insignificance. This has in turn increased demand for digital offerings. Where once a consumer would carefully decide on the limited selection that he or she would consume due to the high costs of transaction, the consumer now selects a much larger and much more diverse set of offerings to consume. Many of these offerings are niche offerings that would not have been selected previously when the transaction cost was higher.

1.2.2. Reduced search costs and other benefits

By far the greatest impact on online demand and the emergence of the Long-Tail has been reductions in search costs. Even with a very large number of offerings, the prohibitive cost of search meant that most consumers only selected a handful of offerings – the mainstream ones. The large reductions in search costs with the advent of the Internet and its variety of search tools means that consumers can obtain information about previously invisible offerings. This in turn generates demand for these niche offerings (Anderson 2006). Indeed, research has shown that consumers derive a greater benefit online from the ability to consume a far greater choice set than they can via offline stores than just cost savings (Brynjolfsson et al. 2003). Search costs are reduced through effective information search tools. These tools can be "active search" tools or "passive search" tools (Brynjolfsson et al. 2006).

Active search tools are those that focus on consumer initiated search as well as sampling of offerings (Brynjolfsson et al. 2006). With these tools, consumers actively seek out offerings of

their choice and sample offerings before completing the transaction. These active search tools reduce search costs and increase demand by connecting consumers with offerings that they desire (Brynjolfsson et al. 2006). Examples of active search tools include Google Search, and previewing tools such as Amazon's "Search Inside" feature and iTune's 30 second music samples. Passive search tools are those that quietly seek out and direct consumers to offerings that are likely to match their taste. Passive search tools include recommender systems that use the habits of customers with similar tastes to recommend offerings, or use past preferences of the customer to recommend new offerings (Brynjolfsson et al. 2006). Most recommendation engines, including those used by Amazon, YouTube and NetFlix are passive search tools. Electronic Word-of-Mouth (eWOM) can be both an active and passive tool. eWOM is active in that it provides consumers with evaluative information about an offering before consumption (Amblee et al. 2007b), thereby reducing search costs. eWOM can also lead consumers to other offerings through links, comparisons and discussions of related offerings (Amblee et al. 2007b).

1.2.3. Impact on the Long-Tail

Reduced search costs make consumption of previously unviable offerings possible for consumers, and therefore increases demand in the Long-Tail. Reduced search costs will increase demand throughout the Long-Tail distribution, but will have the most impact on demand for niche offerings which may see consumption for the first time. The increased demand will lead to a *fattening* effect on the Long-Tail as demand rises (Anderson 2006). In addition to the fattening effect, there is another effect caused by search tools. Passive search tools such as recommender systems direct consumers to offerings that they might otherwise never consider or even be aware of. This moves consumers away from mainstream offerings and into exploring niche offerings (Anderson 2006). This *flattening* effect moves demand away from the mainstream into the niches, leading to a more even distribution (Anderson 2006; Brynjolfsson et al. 2007). In a recent study, Brynjolfsson, Hu and Simester (2007) examined a women's clothing retailer that sold the same set of items through a catalog as well as through a website. They found that the online sales distribution had a longer tail than the offline sales distribution, and concluded that the lower search costs on a website were responsible for the flatter tail online. In a newer study, Tucker and Zhang (2008) conducted an experiment with a website that lists wedding services vendors and found that popularity information – a quality signal, has a more positive impact for niche vendors than mainstream vendors (Tucker et al. 2008). Elberse and Oberholzer-Gee (2008) examined online movie rentals from 2000 to 2005, and found that movies that rarely rented offline rented twice as often online, although their findings were mixed on the impact of mainstream movie rentals (Elberse et al. 2008). In another study, Gal Oestreicher-Singer and Arun Sundararajan found that network structures affect demand online, and flatten the demand curve (Oestreicher-Singer et al. 2006). Another study related to network structures by Fleder and Hosanagar (2008) concluded that while recommender systems do lead individual consumers to new offerings that they may not have otherwise considered, they also lead most consumers to the same set of new offerings, which although likely to be new for each individual customer, are essentially the same across all customers (Fleder et al. 2008). This would have the net effect of reducing the flattening effect of network structures.

eWOM can both fatten and flatten the Long-Tail, since it is both an active and passive search tool. Prior research has shown that eWOM has an awareness effect, where the volume of eWOM, in the form of ratings and comments, can make customers aware about offerings (Liu 2006). This awareness reduces search costs which in turn increases demand. eWOM also helps connect

customers with niche offerings by discussing these offerings alongside the existing offering, which moves demand down the tail, thereby flattening the tail. While the volume of eWOM in the form of ratings and comments has been shown to be positively correlated with consumption (Amblee et al. 2007a; Amblee et al. 2007b), the density of eWOM is relatively low, with only a small fraction of consumers engaging in eWOM (Dellarocas et al. 2006). A study by Dellarocas and Narayan (2007) found that the distribution of eWOM is even more skewed towards hits than the distribution of consumption (Dellarocas et al. 2007). Thus eWOM is quite limited and likely to trail off much faster than the Long-Tail of consumption. However, eWOM reduces search costs, and this reduction in search costs is likely to benefit niche offerings more than mainstream offerings, even though mainstream offerings will have a greater volume of eWOM (Dellarocas et al. 2007). Thus we argue that if eWOM is more equitably distributed, the level of inequality in the Long-Tail should be lower. In other words, the Long-Tail will be flatter. Therefore we propose:

Hypothesis 1: A lower level of inequality in the distribution of eWOM will lead to a flatter distribution.

1.3. Interaction between the three forces

The Lengthening, Fattening and Flattening forces[3] constantly act upon one another to shape the Long-Tail. The lengthening of the tail due to increased supply creates more choices for consumers. This increased choice in turn increases the probability of having a "hit" offering, which moves into the mainstream (Anderson 2006). A greater number of hits will flatten the head of the Long-Tail distribution. A greater number of hits will also fatten the long tail due to increased consumption. This increased popularity will in turn increase supply, and the cycle will repeat itself in a virtuous loop.

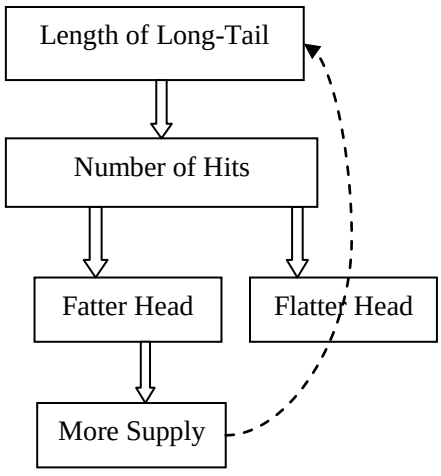

Figure 2. Interaction between the three forces on Long-Tail

[3] The terms lengthening, fattening and flattening were coined by Anderson (2006).

Based on the above model we propose:

Hypothesis 2: A greater supply of offerings will lead to a greater number of hits.

Hypothesis 3: A greater number of hits will lead to a fatter head for the Long-Tail distribution.

Hypothesis 4: A greater number of hits will lead to a flatter head for the Long-Tail distribution.

Hypothesis 5: A fatter head will lead to more supply into the Long-Tail.

2. An Empirical Study of YouTube

To test the proposed hypotheses, we conducted an empirical study of YouTube.com, the most popular video-sharing platform on the Internet. Despite it being a cultural phenomenon, there is no academic research that examines YouTube in any detail. We intend for this paper to be a first academic look into the most popular digital content-sharing platform of all time. YouTube also offers many distinct advantages in analyzing the structure of the Long-Tail. First, YouTube does not charge for viewing any of the content on the site. Therefore the price for the offerings is zero, which removes the impact of price and most of the risk from the decision-making process of consumers. This means that consumers will make viewing decisions based only on interest and curiosity, which will provide a better understanding of consumers' real tastes. Second, YouTube does not charge producers to list and disseminate content, which allows for producers to develop content without having to consider future costs into account. Third, YouTube allows both amateurs and professionals to upload content, and this allows for the emergence of the "true" Long-Tail distribution for the industry, where supply factors are only determined by production, inventory and distribution costs. Fourth, YouTube absorbs the cost of distribution by disseminating content for free. Fifth, YouTube is the most popular content-sharing platform on the Internet, and thus the research findings will be highly relevant.

2.1. Data Collection

We collected data from YouTube.com using the Google Data API, which provides controlled but vast access to YouTube's database. We used the GData PHP wrapper in the Zend Framework to write several PHP programs to retrieve the data in May 2008. YouTube restricts results to 1000 videos per search. Of the 1000 video details retrieved from each search, about 50 were dropped since they were duplicates. We collected data on the Top 1000 most viewed videos for each of the 14 categories[4] on YouTube. We also collected data on the Top 1000 most viewed videos of all time on YouTube. For each video, we collected the video's unique ID (assigned by YouTube), the number of views (view count), number of ratings, number of comments, and the average rating. We also collected the total number of videos in each category. Each video on YouTube can only belong to one category. Finally, we collected information on the 1000 most recent videos uploaded to YouTube.

2.2. Measures

[4] The category "Nonprofits" had only 215 videos in total.

2.2.1. Measure of Length of the Long-Tail

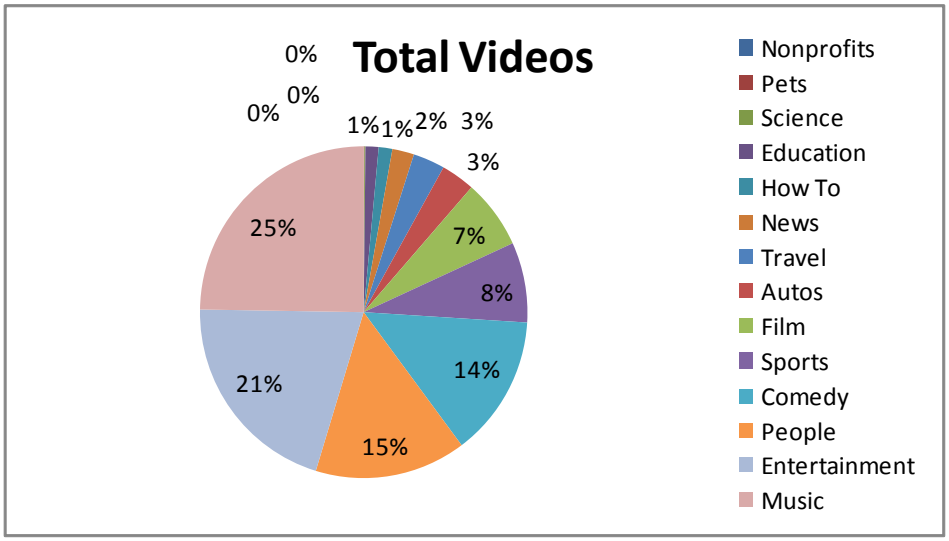

Figure 3. Distribution of available videos by category

We use the total number of videos in each category as a measure of length of the Long-Tail for each category on YouTube. We summate the number of videos in each category to calculate the total length of the YouTube Long-Tail distribution.

Table 1. YouTube categories and the corresponding number of available videos and total view counts

Category	Total Videos	Total Views
Nonprofits	215	96,001
Pets	63,226	192,290,435
Science	108,551	149,846,704
Education	1,232,727	186,100,228
How To	1,283,124	977,368,379
News	2,067,705	1,075,079,146
Travel	2,996,880	462,660,199
Autos	3,199,224	1,043,569,051
Film	6,452,652	1,863,744,247
Sports	7,475,302	1,567,681,088
Comedy	13,221,019	3,519,665,792
People	14,287,260	2,148,919,309
Entertainment	19,658,434	3,675,614,066
Music	23,686,032	7,580,421,814
Total Length / Top 1000	95,732,351	10,080,587,464

2.2.2. Measure of the Size (Fatness) of the head of the Long-Tail

We use the aggregate number of views of the top 1000 videos in each category as a measure of the size of the head of the Long-Tail. We do the same for the top 1000 most popular YouTube videos as well. We refer to the "top" of the Long-Tail distribution as the head, since it consists of the most popular videos.

2.2.3. Measure of the Inequality (Flatness) of the Long-Tail

We use the Gini coefficient as a measure of inequality in the Long-Tail. A flatter distribution will correspond to a lower Gini coefficient[5], and vice versa. The Gini coefficient has been used previously in academic research to measure the level of inequality in the Long-Tail (Brynjolfsson et al. 2007; Dellarocas et al. 2007; Oestreicher-Singer et al. 2006). The Gini coefficient varies from zero to one, with zero implying a perfectly equal distribution and one implying a perfectly unequal distribution. Therefore a higher Gini coefficient implies greater inequality of distribution, and a lower Gini coefficient indicates a flatter Long-Tail distribution. In this study, we focus on the head of the Long-Tail distribution. We calculated the Gini co-efficient for the Top 1000 videos for each category as well as for the Top 1000[6] most viewed videos across YouTube (see Table 2). We also calculated the Gini coefficient for the eWOM distribution for each category as well as the Top 1000 (the most viewed) videos on YouTube.

Table 2. Gini coefficients for View Counts, Ratings and Comments by category

Category	Gini View Counts	Gini Rating	Gini Comments
Autos	0.347	0.507	0.541
Comedy	0.339	0.570	0.566
Education	0.511	0.723	0.779
Entertainment	0.340	0.551	0.639
Film	0.383	0.599	0.621
How To	0.418	0.603	0.722
Music	0.342	0.463	0.542
News	0.374	0.623	0.767
Nonprofits	0.890	0.908	0.941
People	0.401	0.665	0.725
Pets	0.676	0.734	0.760
Science	0.541	0.690	0.790
Sports	0.313	0.479	0.540
Travel	0.426	0.649	0.683
Top 1000	0.301	0.467	0.539

For the view counts, we also calculated the Gini coefficient with 50 video intervals, to see the impact on the level of inequality as we move down the Long-Tail distribution starting at the top.

[5] Multiplying the Gini coefficient by a factor of one hundred gives the more familiar Gini Index.
[6] In actuality, only 950 videos were used.

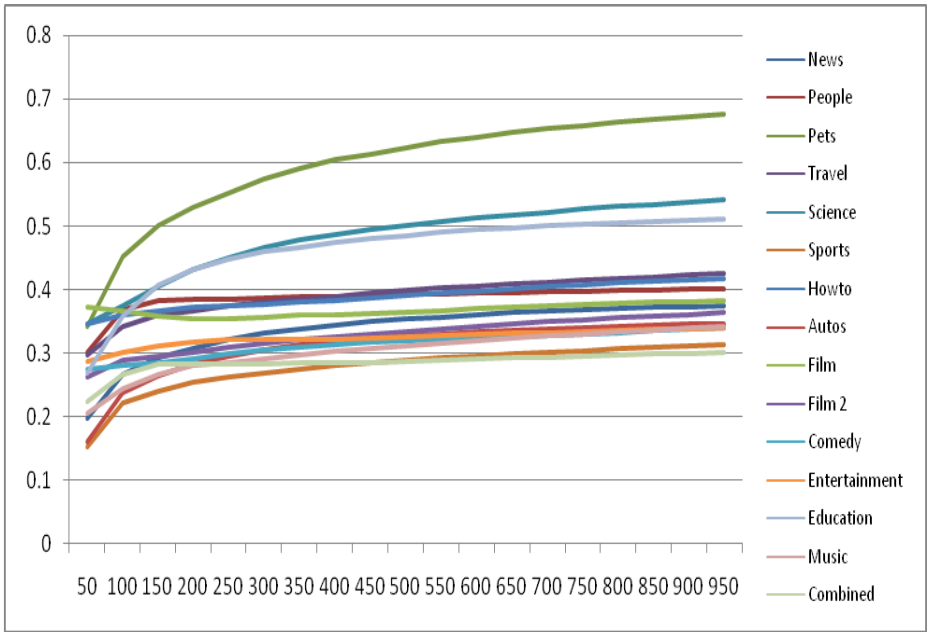

Figure 4. Gini coefficients moving down the Long-Tail

The above chart shows the Gini coefficient as we move across the distribution of the Top 1000 videos in each category as well as the combined Top 1000 regardless of category. The Gini coefficients for the top 50 videos are very low for each category, implying that the level of imbalance within the Top 50 is quite low. The Gini coefficients for the top 100 videos are also quite low, although higher than for the Top 50. The Gini coefficient then continues to climb gradually at a steady rate. This means that although there is significant herding at the top, the level of imbalance among the most popular videos is quite low. For example, the Dissimilarity Index (not shown) for the top 100 videos in the News category is 0.16, which means that only about 16% of the aggregate view counts will need to be redistributed in order to have an equal distribution.

The only category to buck this trend is Films, which shows the Gini starting high and then declining, before gradually increasing again. However, a closer inspection showed that the reason for this anomaly is the existence of two video clips in a non-English language that have extremely high view counts, but are excluded from being listed on any of YouTube's top lists. Removing these two clips shows a much more gradual Gini curve (See Film 2 in the above chart). While the Gini coefficients for the top 1000 videos are not very large, the values increase with the number of videos, implying the existing of a drop-off. This drop-off is the Long-Tail. The change in the Gini coefficient gets smaller as N increases, implying that the tail declines very gradually. This can be interpreted as evidence for a very long tail.

2.2.4. Measures of eWOM

We use the Gini coefficient again to measure the inequality in the distribution of eWOM in the Top 1000 videos in each category and the Top 1000 videos of all time (see Table 2). We calculate the Gini coefficient for the distribution of ratings as well as the distribution of comments. We aggregate the total number of ratings and comments in the Top 1000 lists to calculate the magnitude of eWOM in the head of the Long-Tail distribution.

3. Results

3.1. Summary Statistics

Table 3. Summary Statistics

Variable	N	Mean	Std Dev	Sum	Min	Max
logTotalViews	14	20.10	2.75	281.4	11.47	22.75
logTotalVideos	14	14.18	3.13	198.6	5.25	16.98
logHits	11	3.47	1.45	38.2	1.10	6.25
logNewVideos	14	3.57	1.43	49.9	0	5.30
logAvgView	14	5.92	0.89	82.8	5.01	8.02
logGiniViews	14	-0.84	0.30	-11.8	-1.16	-0.12
logGiniRatings	14	-0.48	0.18	-6.8	-0.77	-0.10
logGiniComments	14	-0.39	0.17	-5.5	-0.62	-0.06

Table 4. Legend

LogTotalViewCount	Log of sum of all view counts of the top 1000 videos in a category
LogTotalVideos	Log of count of total videos uploaded to a particular category
LogHits	Log of the number of videos from a category in the Top 1000 most popular videos
LogNewVideos	Log of number of videos in a category in the 1000 most recent videos uploaded
LogGiniViewCount	Log of the Gini coefficient for the distribution of view counts in the top 1000 videos in a category
LogGiniComments	Log of the Gini coefficient for the distribution of comments in the top 1000 videos in a category
LogGiniRatings	Log of the Gini coefficient for the distribution of ratings in the top 1000 videos in a category

Table 5. Pearson Correlation Coefficients

	Log Total View Count	Log Total Videos	Log Hits	Log New Videos	Log Gini View Count	Log Gini Ratings
LogTotal Videos	0.96 ***					
LogHits	0.85 ***	0.62 **				
LogNew Videos	0.85 ***	0.87 ***	0.76 ***			
LogGini ViewCount	-0.89 ***	-0.94 ***	-0.41	-0.71 ***		
LogGini Ratings	-0.81 ***	-0.81 ***	-0.50	-0.64 **	0.91 ***	
LogGini Comments	-0.75 ***	-0.77 ***	-0.47	-0.70 ***	0.84 ***	0.94 ***

* p<.10 **p<.05 ***p<.01

3.2. Top 1000 Videos by popularity on YouTube

Figure 5 shows the distribution for the most popular (Top 1000) videos on YouTube.com. Clearly it follows a Long-Tail distribution. After about the first 100 videos, the head tapers off into a long tail. While the distribution is very smooth, it is the aggregation of what Anderson (2006) referred to as many "tails within tails" (Anderson 2006) or alternatively, "heads within heads".

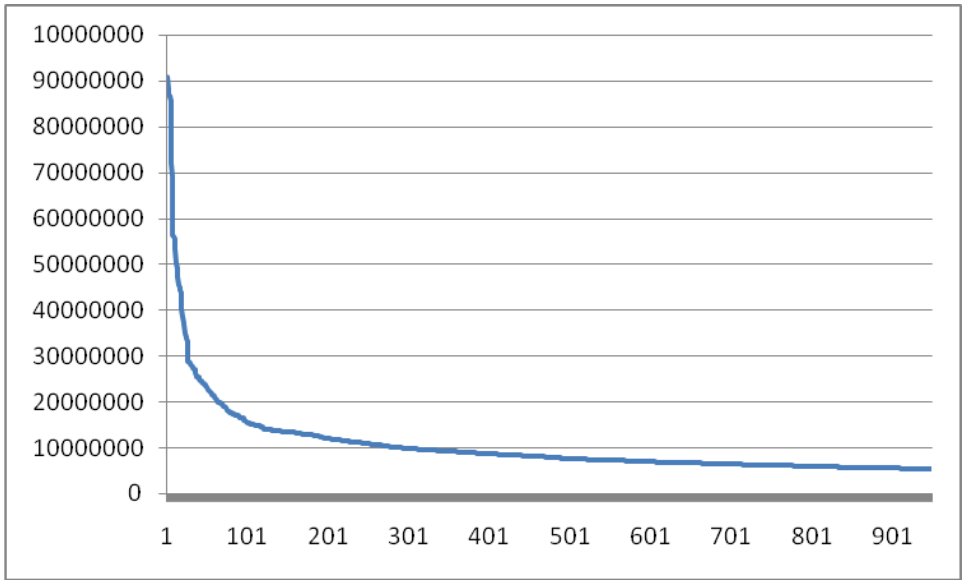

Figure 5. Long Tail of 1000 most popular videos

Figure 6 shows the sub-distributions within the head (Top 1000 videos), separated by category. Each category follows it own Long-Tail distribution. The dominance of the Music category is clearly evident. It is quite an interested result that the most popular video-sharing site of all time is in many ways predominantly a music-sharing site. The trend of downloading music has spread to videos as well. This phenomenon, studied separately, would lead to a better understanding of the online music industry.The pie-charts show the breakdown of the top 1000 most viewed YouTube videos. Music videos accounted for more than half of all the top videos on the site. Comedy videos are a distance second followed closely by Entertainment videos. These three categories account for over 80% of all views on the site[7]. Three other categories did not contribute any videos to the Top 1000 list. We also aggregated the total number of views in the top 1000 videos by category. We find that the results are almost identical to the breakdown by number of videos. This means that although a few categories dominate and the top 1000 videos show a Long-Tail distribution, no single category dominates in terms of average views per video in the top 1000.

[7] Interesting, this follows the 80/20 Pareto Principle, since approximately 20% (3 out of 14 categories) accounts for 80% of all the top views.

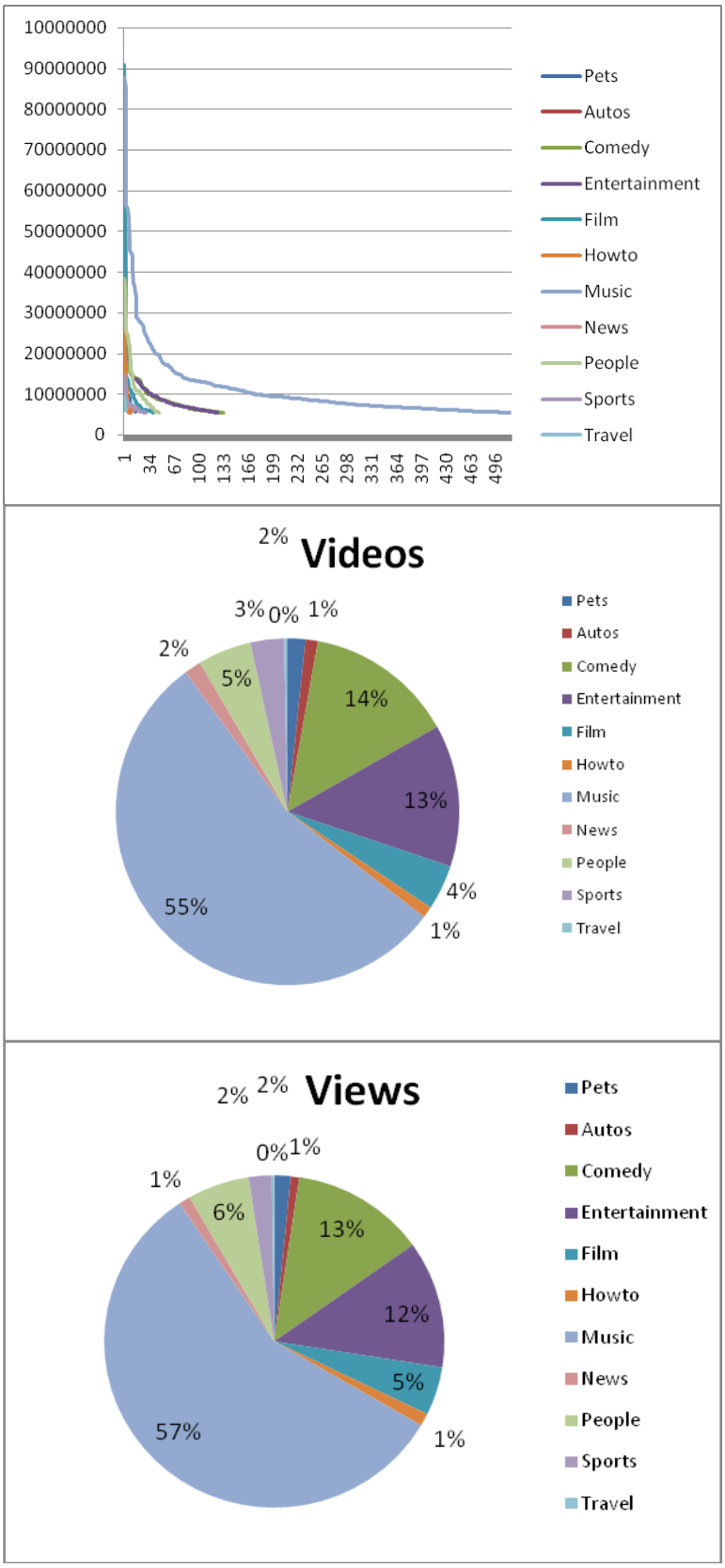

Figure 6. Distribution of View Counts for Top 1000 Videos

3.3. Hypotheses Testing

Hypothesis 1: A lower level of inequality in the distribution of eWOM will lead to a flatter distribution.

We use the Gini coefficient to measure the level of inequality (inversely related to flatness) for the distribution of both consumption (view counts) and eWOM (number of ratings and comments). Since prior research has shown a relatively strong correlation between consumption and the volume of eWOM (Amblee et al. 2007a; Amblee et al. 2007b), we control for the total volume of eWOM in order to remove any spurious correlation between the level of inequality in the distribution of consumption and eWOM. To do this we look at the partial correlation between the Gini of consumption and the Gini of eWOM (both for comments and ratings), controlling for the total volume of eWOM. The zero order correlation between LogGiniViews and LogGiniComments is 0.844 ($p<0.01$). The first order partial correlation, controlling for LogTotalViews is 0.70 ($p<0.01$). Therefore we detect a partial explanation effect. We run a regression to estimate the coefficients.

$$\text{LogGiniViewCount} = \alpha + \beta \cdot \text{LogGiniComments} + \beta \cdot \text{LogTotalComments} + \varepsilon$$

The results (See Table 6) show that the total volume of comments moderates the relationship between the degree of inequality in the view counts and the level of inequality in the distribution of comments. This can be interpreted as saying that "shifting" forces such as eWOM do in fact shift demand down the tail, but are more effective when these shifting forces are more evenly distributed.

We repeat the above tests for the number of ratings. The zero order correlation between LogGiniViews and LogGinRatings is -0.91 ($p<0.01$). The first order partial correlation, controlling for LogTotalRatings is -0.73 ($p<0.01$). Therefore we again detect an explanation effect. We run a regression to estimate the coefficients (See Table 6 for regression results).

$$\text{LogGiniViewCount} = \alpha + \beta \cdot \text{LogGiniRatings} + \beta \cdot \text{LogTotalRatings} + \varepsilon$$

Once again, we find that there is strong positive relationship between inequality in the distribution of view counts and the distribution of ratings, and that this relationship is moderated by the total volume of ratings. This can be interpreted as stating that a larger volume of ratings helps flatten the head of the distribution. We looked at the standardized coefficients for both regressions, and find that the flattening impact of the total volume of comments is stronger than the flattening impact of the total volume of ratings on the Long-Tail distribution. This result makes intuitive sense in that comments provide more information about an offering than just the aggregate number of ratings. It offers some contrast to previous findings that eWOM can sometimes increase inequality in the Long-Tail (Dellarocas et al. 2007). A t-test showed that there is a difference between the means of the two groups, with the distribution of comments being more skewed than the distribution of ratings. Furthermore, the distribution of both comments and ratings are more skewed than the distribution of view counts, which is in line with previous findings (Dellarocas et al. 2007).

Table 6. Regression Results

Y/X	Log Gini View Count	Log Gini View Count	Log Gini View Count
Constant	0.279 (0.186)	0.387* (0.190)	-0.742*** (0.163)
Log Gini Ratings	0.963*** (0.275)		
Log Gini Comments		0.815*** (0.253)	
Log Total Ratings	-0.0478** (0.020)		
Log Total Comments		-0.067*** (0.018)	
Log Top 1000 Hits			-0.059 (0.044)
N	14	14	11
F-value	42.46	38.78	1.84
R-square	0.87	0.85	0.077

* p<.10 **p<.05 ***p<.01

3.3.2. Length of Tail and Volume of Hits

Hypothesis 2: A greater supply of offerings will lead to a greater number of hits.

We use the total number of videos in each category as a measure of supply, and use the number of videos of a particular category in the Top 1000 most popular videos as a measure of the number of hits. Thus we derive the following models:

$$\text{Log}\,(\text{Top1000Hits}_{ci}) = \alpha + \beta \cdot \text{Log}\,(\text{TotalVideos}_{ci}) + \varepsilon$$

$$\text{Top1000Hits}_{ci} = \alpha + \beta \cdot \text{TotalVideos}_{ci} + \varepsilon$$

We use the natural log of the number of hits and total videos to obtain a linear relationship between the two variables. In the second model, we include categories that had zero hits in the top 1000, for which the log transformation cannot be applied. We tested the two models and find that there is a positive and significant relationship between the number of hits from a category and the length of the Long-Tail distribution of that category. The results (see Table 7) provide evidence that there is indeed a relationship between the length of a Long-Tail distribution and the number of hits it generates. This is an endogenous relationship since a greater number of hits will encourage producers to develop more content for that category, thereby lengthening the tail further.

Table 7. Regression Results

Y/X	Log Top 1000 Hits	Top 1000 Hits	Log Total ViewCount	Log New Videos
Constant	-4.49 (3.47)	-28.92 (30.09)	14.23*** (0.725)	-5.31*** (1.62)
Log Total Videos	0.529** (0.226)			
Total Videos		0.0000142*** (0.0000030)		
Log Total View Count			0.376*** (0.055)	0.442*** (0.080)
Log Top 1000 Hits			0.325*** (0.064)	
N	11	14	11	14
F-value	5.48	22.92	93.45	30.60
R-square	0.31	0.63	0.95	0.70

* p<.10 **p<.05 ***p<.01

3.3.3. Fatness of Head and Volume of Hits

Hypothesis 3: A greater number of hits will lead to a fatter head for the Long-Tail distribution.

We use the total number of view counts in the head (Top 1000 videos in each category) as a measure of the fatness of the head. We use the natural log transformation to derive a linear relationship. As per the model we developed, we intend to test if there is a positive relationship between the length of the Long-Tail distribution and the "fatness" of the distribution. We also propose that the number of hits is an intervening variable. We measure the partial correlation between the length of the Long-Tail (LogTotalVideos) and the fatness of the Long-Tail (LogTotalViewCount), controlling for the number of hits (LogTop1000Hits). We find a small partial explanation effect, since the zero order correlation drops from 0.96 ($p<0.01$) to 0.92 ($p<0.01$). We then test the following regression model:

$$\text{Log}\,(\text{TotalViewCount}_{ci}) = \alpha + \beta \cdot \text{Log}\,(\text{Top1000Hits}_{ci}) + \beta 2 \cdot \text{Log}\,(\text{TotalVideos}_{ci}) + \varepsilon$$

The results show a strong and statistically significant relationship between the length of the Long-Tail distribution, the number of hits in the Top 1000 videos and the "fatness" of the head of the distribution. Thus both the length of the Long-Tail and the number of hits positively impact the fatness of the Long-Tail.

3.3.4. Flatness of Head and Number of Hits

Hypothesis 4: A greater number of hits will lead to a flatter head for the Long-Tail distribution.

We use the Gini of the Long-Tail of view counts of the top 1000 videos in each category as the measure of the flatness[8] of the head of the Long-Tail distribution. As per our model, we expect

[8] The relationship between the Gini coefficient and level of flatness is an inverse one.

an intervening effect from the number of hits on the relationship between the length of the Long-Tail and the flatness of the head of the distribution. The zero order correlation between LogGiniViewCount and LogTotalVideos is -0.94 ($p<0.01$), and the first order partial correlation controlling for LogTop1000Hits is -0.86 ($p<0.01$). We thus test the following regression model:

$$\text{Log}\ (\ \text{GiniViewCount}_{ci}) = \alpha + \beta 1 \cdot \text{Log}\ (\text{Top1000Hits}_{ci}) + \beta 2 \cdot \text{Log}\ (\text{TotalVideos}_{ci}) + \varepsilon$$

While there is a strong inverse relationship between the length of the Long-Tail and the flatness of the head of the distribution, we do not find a statistically significant impact from the number of hits in the Top 1000 list and the flatness of the Long-Tail distribution. One explanation is that while a greater number of hits is likely to spread attention throughout the head, thereby reducing inequality, it is also likely to increase the inequality in the head by drawing attention to these particular hits. This opposing effect may not show up in our model.

3.3.5. Fatness of Head and Length of Tail

Hypothesis 5: A fatter head will lead to more supply into the Long-Tail.

We use the number of new videos in each category (1000 most recent videos) to measure the relative increase in supply for each category. To measure this growth in the length of the Long-Tail, we retrieved 1000 of the most recent videos uploaded to YouTube and broke it down by category.

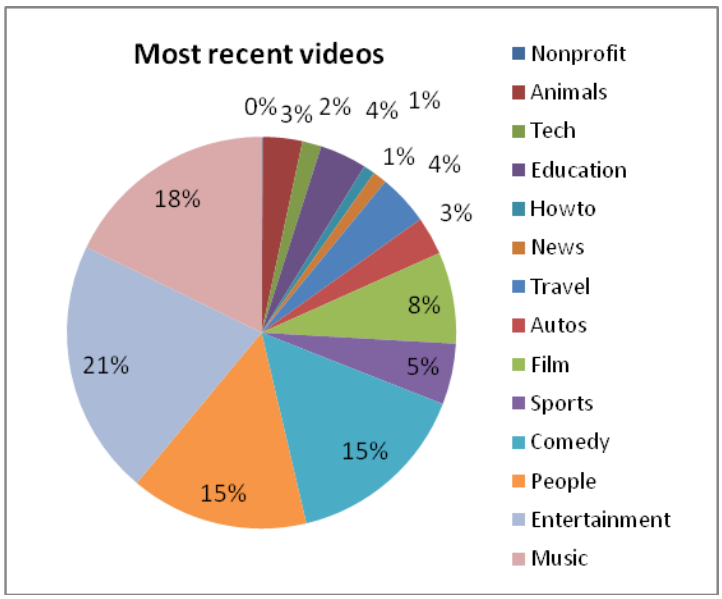

Figure 7. Top 1000 most recent videos

We again use the natural log transformation to derive a linear relationship. As per our model, we expect an intervening effect from the fatness of the head of the distribution on the relationship between the number of hits and the introduction of new videos for a particular category. The zero order correlation between LogTop1000Hits and LogNewVideos is 0.76 ($p<0.01$), while the first order partial correlation controlling for LogTotalViewCount is 0.50 ($p>0.10$). Thus we find evidence of explanation or control effect. We test the following regression model:

$$\text{Log}(\text{NewVideos}_{ci}) = \alpha + \beta \cdot \text{Log}(\text{TotalViewCount}_{ci}) + \varepsilon$$

We find a positive and statistically significant relationship between the "fatness" of the Long-Tail distribution of a category and the number of new videos (increased supply/length) in that category (See Table 7). This means that more popular categories receive a greater supply, thereby lengthening the Long-Tail distribution.

4. Discussion

Our first academic look at the Long-Tail of YouTube has provided valuable insight into the consumption patterns at a leading digital content site. We find that the widespread distribution of eWOM can help reduce inequality in the head of the Long-Tail. While the impact of eWOM on the structure of the Long-Tail was measured, we did not include eWOM in our interactive Three-Forces model due to the complex endogenous nature of its impact. eWOM is likely to not only impact the flatness of the distribution, but the fatness and the number of hits as well.
Our research confirms Anderson's findings (2006) that each Long-Tail consists of many sub-tails, each of which follows a Long-Tail distribution. Our initial results show the existence of a system that is likely never at equilibrium, with several forces acting simultaneously on the Long-Tail to constantly shape its distribution. The length of the Long-Tail is modeled as the "instigator" of this system, which in turn throws up more hits. These hits in turn fatten the head and both increase demand with that category and shift demand down the tail. The increased demand in turn adds supply, lengthening the tail further. We intend to follow up with a model sophisticated model and test it with structural equation modeling techniques.

5. Conclusion: Limitations and Future Research

Since we present this research as a first look at the Long-Tail of YouTube and the dynamic forces shaping the distribution, we refrain from making any recommendations. Instead we use the preliminary findings of this paper to chart a future course of research to better understand the dynamic nature of the Long-Tail. First, we intend to map out the entire Long-Tail distribution of a popular digital content provider such as YouTube in order to look beyond just the head of the Long-Tail. Second we intend to study the changes to the structure of the Long-Tail over time. Third, we intend to study the impact of eWOM and its interaction with the other forces more extensively. Fourth, we intend to conduct experiments to control for two of the three forces while studying the impact of the third. Fifth, we intend to conduct focus group sessions with viewers to better understand their motives while visiting popular digital content aggregators[9] such as YouTube.

[9] Anderson (2006) refers to sites that bring large numbers of offerings together as aggregators.

6. References

Amblee, N., and Bui, T. "The Impact of Additional Electronic Word-of-Mouth on Sales of Digital Micro-products over Time: A Longitudinal Analysis of Amazon Shorts," 40th Annual Hawaii International Conference on System Sciences (HICSS), Waikoloa, HI, 2007a.

Amblee, N., and Bui, T. "The impact of electronic word-of-mouth on digital microproducts: An empirical investigation of Amazon Shorts," 15th European Conference on Information Systems, St. Gallen, Switzerland, 2007b.

Anderson, C. "The Long Tail," *Wired Magazine* (12:10) 2004, pp 170-177.

Anderson, C. "The Long Tail: How Endless Choice Is Creating Unlimited Demand," Random House Business Books, New York, NY, 2006.

Bakos, Y. "The emerging role of electronic marketplaces on the Internet," *Communications of the ACM* (41:8) 1998, pp 35-42.

Bakos, Y., and Brynjolfsson, E. "Bundling and Competition on the Internet," *Marketing Science* (19:1) 2000.

Brynjolfsson, E., Hu, Y., and Smith, M.D. "From Niches to Riches: Anatomy of the Long Tail," *Sloan Management Review* (47:4) 2006, pp 67-71.

Brynjolfsson, E., Hu, Y.U.J., and Simester, D. "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales,") 2007.

Brynjolfsson, E., Yu, H., and Smith, M.D. "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers.," *Management Science* (49:11) 2003, pp 1580-1596.

Dellarocas, C., and Narayan, R. "A Statistical Measure of a Population's Propensity to Engage in Post-Purchase Online Word-of-Mouth," *Statistical Science* (21:2) 2006, pp 277-285.

Dellarocas, C., and Narayan, R. "Tall Heads Vs. Long Tails: Do Consumer Reviews Increase the Informational Inequality Between Hit and Niche Products?,") 2007.

Duan, W., Gu, B., and Whinston, A.B. "Analysis of Herding on the Internet–An Empirical Investigation of Online Software Download," *Proceedings of the Eleventh Americas Conference on Information Systems, Omaha, NE, USA August 11th-14th*) 2005.

Elberse, A., and Oberholzer-Gee, F. "Superstars and Underdogs: An Examination of the Long Tail Phenomenon in Video Sales," Harvard Business School Working Paper.

Fleder, D., and Hosanagar, K. "Blockbuster culture's next rise or fall: The effect of recommender systems on sales diversity Daniel Fleder Kartik Hosanagar,") 2008.

Gallaugher, J.M. "E-commerce and the undulating distribution channel," *Communications of the ACM* (45:7) 2002, pp 89-95.

Liu, Y. "Word of Mouth for Movies: Its Dynamics and Impact on Box Office Revenue," *Journal of Marketing* (70:3) 2006, p 74.

Oestreicher-Singer, G., and Sundararajan, A. "Network Structure and the Long Tail of Electronic Commerce,") 2006.

Porter, M.E. "Strategy and the Internet," *Harvard Business Review* (79:3) 2001, pp 63-78.

Tucker, C., and Zhang, J. "Long Tail or Steep Tail? A Field Investigation into How Online Popularity Information Affects the Distribution of Customer Choices,") 2008.