Income Inequality in the Attention Economy[*]

Kevin S. McCurley
Google Research

ABSTRACT

The World Wide Web may be viewed as a gigantic market for information. In this market there are producers (authors) and consumers (readers) and the currency for information is *attention*. In this paper we examine the distribution of attention across the World Wide Web. Through study of the habits of web users, we conclude that the currency of attention is highly concentrated on a relatively small number of web resources, and that the rich appear to be getting slightly richer over time. We also study the effect of search engines on the distribution of attention, and conclude that search engines produce a more uniform distribution of attention than generic surfing habits. Finally, we show that the observed distribution of attention is in substantial disagreement with the distribution that is suggested by the random surfer model embodied in the PageRank algorithm.

Categories and Subject Descriptors

J.4 [**Computer Applications**]: Social and Behavioral Sciences—*Economics*

General Terms

Welfare Economics, Attention Economy, Distribution of Traffic, PageRank

1. INTRODUCTION

As we all know, economics is concerned with the study of allocation of resources in markets, with particular focus on scarcity and competitiveness. When economists think of markets, they generally think of exchanges of goods and services for monetary value, and some information markets fall under this classical view. The World Wide Web has changed that somewhat however, since most information is available "for free". The World Wide Web may still be viewed as a market for information, in which producers (authors) and consumers (readers) participate in transactions. The scarce resource in this market is not the information itself, nor is it money to buy the information. The scarce resource that plays the role of currency in this market is *attention*. The recognition that attention is the scarce resource in information markets seems to have originated with Herbert Simon [29] in 1971, and the concept has been popularized recently (e.g., see [14, 11]).

Authors compete for attention because attention has value. The essence of advertising is to steer the attention of consumers toward products offered by a seller. When advertising succeeds, it is because it completes a transaction that turns attention into monetary value from sale of goods or services. Attention can often be converted into other forms of value, such as reputation. In some sense, reputation is to attention as wealth is to income, because reputation and wealth represent stored value of their respective currencies.

A prerequisite for monetization of web resources is to garner attention. In order to understand the forces that shape information markets, it is therefore important that we understand the dynamics and characteristics of how attention is distributed across the web.

In this paper we study the allocation of resources in information markets from the point of view of welfare economics. This subfield of economics is primarily concerned with two concepts, namely the efficiency of markets and the distribution of income. In the context of information markets such as the World Wide Web in which the primary currency is attention, the distribution of income corresponds to the distribution of attention among information resources. In what follows we shall examine this distribution of attention among users of the World Wide Web.

The two primary questions about income distribution are typically the following.

- to what degree is the distribution of income unequal or inequitable?
- to what extent does the distribution of income affect social welfare, including growth and market efficiency?

In this paper I will address both questions, though not in complete generality. First, inequality and inequity are two different things. I will restrict my attention to the inequality of attention among resources, and let others form conclusions about the social impact of this inequality. Moreover, market efficiency is something that is very difficult to quantify in attention markets, so I will confine my attention to the question of growth. The approach taken in this paper is largely empirical, but I will include a discussion of mathematical modeling issues along the way.

Economists have long known that information is a fundamental ingredient in how markets are shaped. In particular, the 2001 Nobel prize in economics was awarded for work on

[*]Working Draft, November 14, 2007. The current version may be found at mccurley.org/papers/effective/

markets with asymmetric information. The premise is that buyers who have access to more and better information are able to make more rational decisions, which results in more efficient markets. By making information freely and widely available, it is reasonable to believe that the World Wide Web has had a profound effect on a number of markets, resulting in better efficiency and more growth. It is therefore imperative that we understand the characteristics of how information is consumed on the Web.

The World Wide Web has differing effects on different markets, and it is impossible to address every effect on every market. Among other influences, the World Wide Web has been observed to enabled many niche markets to flourish, and fueled a great deal of interest among economists in the study of "long tail markets". Music and books present good examples of markets of this type, since it might be possible to find products of very specific and narrow genres. In order for these markets to grow, they need to attract enough attention. If consumers are not aware that markets exist, they may not seek them out.

Distribution of income in the attention economy

In studying the distribution of income in the United States, it has been observed that the top 20% of households in the U.S. had 49.7% of the income [18]. Moreover, it is generally agreed that the percentage of total income going to the top 20% in the U.S. has been increasing since the 1960s [30]. In this paper I show that a similar phenomenon appears to be happening in the attention market on the World Wide Web. The startling observation is that, while the number of web sites and URLs has grown tremendously in recent years, the distribution of attention seems to be showing greater inequality, with attention concentrated on a relatively small number of web sites. This confirms a statement by Benkler [5, pp. 214]:

> ...the Internet is, in fact, exhibiting concentration: Both infrastructure and, more fundamentally, patterns of attention are much less distributed than we thought. As a consequence, the Internet diverges from the mass media much less than we thought in the 1990s and significantly less than we might hope.

There is no obvious explanation for this concentration of attention, though several explanations seem possible. There is also evidence of whether this trend will continue. New mechanisms for distribution of information have emerged at a dizzying pace in the last decade, and web properties such as myspace, facebook, and youtube continue to emerge and reshape the distribution of attention. It may also be that the long tail markets have yet to emerge, and that we have not yet seen the biggest economic impact from the World Wide Web. Forces that focus attention on any particular part of the web can have a tremendous impact on the growth of these niche markets.

It should be noted that the study [18] has been subjected to a great deal of criticism, since it fails to take into account non-cash income such as capital gains, and it omits the effects of taxation. The same kind of pitfalls and disputes arise in studying the distribution of attention as a commodity, since it is not clear what to count as attention, not all attention can be easily measured, and attention varies in quality. It is also difficult to determine what impact attention has. The conversion of attention into monetary value depends crucially on these factors.

The influence of search

The emergence of search engines has had a profound effect on the structure of the web, by making information more accessible for people with known information needs, and by directing people to parts of the web that they might not otherwise discover by browsing. Browsing still plays a strong role however, particularly when it comes to entertainment media. It has been argued [6, 7] that search engines serve to concentrate attention on a narrow class of pages, and that newer pages have a hard time garnering attention. One of the goals of this paper is to examine the differences in attention distribution induced by the activities of browsing and search.

An outline of this paper is as follows. In Section 2, I will describe the measures of inequality that are commonly used in welfare economics. In Section 6, I will consider the relationship to the concept of PageRank, and models of growth of information markets. In Section 3, I will describe the data used for this project, and give some summary observations. The major empirical observations will be summarized in Section ??, where we will cover the distribution of attention across web resources. In Section 4, I will examine the relative effects of search vs. browsing. In Section 7, I will examine the implications of different models for distribution on the growth of the World Wide Web.

2. MEASURES OF INEQUALITY

There is a voluminous literature in economics regarding statistical measures of inequality. For surveys on the subject, see [3, 8, 9, 27, 28]. The goal in designing a measure of inequality has been to quantify the degree of inequality in a distribution with a single numeric value, with 0 representing a completely uniform distribution, and larger values representing a more unequal distribution. A further goal is to devise a measure that can be used to break down a population and expose the determining factors for inequality.

2.1 The Gini Index

Probably the most commonly used statistic among economists is the Gini statistic. This is defined in terms of the Lorenz curve of a distribution, which may be defined for any probability distribution on $[0,\infty)$ with finite mean (discrete or continuous). For a cumulative distribution function F, the Lorenz function $L : [0,1] \to [0,1]$ is defined as

$$L(F(x)) = \frac{\int_0^x t dF(t)}{\int_0^\infty t dF(t)}.$$

Thus $L(t)$ represents the fraction of mass that is occupied by values that are less than or equal to a fraction t of the population. In other words, each point (x, y) on the Lorenz curve represents a generalized Pareto principle statement of the form "$100x$% of the population has $100y$% of the mass". Given a set of observations from the distribution, the piecewise linear approximation is generally used as an estimator for the Lorenz curve, but since the curve is convex this is always an overestimate. An approximate Lorenz curve computed this way is shown in Figure ??. Goldie [15] has proved that the piecewise linear approximation is a consistent estimator for the true Lorenz curve, in the sense that the former converges to the latter.

The Gini index of a distribution is defined quite simply as twice the area of the region between the line $y = x$ and the Lorenz curve of the distribution, or

$$G(F) = 1 - 2\int_0^1 L(t)dt.$$

If a distribution is uniformly distributed, then the Gini index is 0, and if the distribution is concentrated on a single value then the Gini index is 1. For an empirical distribution $y_1, \leq \ldots \leq y_n$, the Gini index may be estimated easily as

$$G = \frac{2}{\bar{y}n^2}\sum_{i=1}^{n} iy_i - \frac{n+1}{n}.$$

This estimate is consistent but not unbiased [10]. For the size of populations that we study, this bias is inconsequential.

2.2 The Theil Index

The most commonly used measure of inequality in mathematics and computer science is that of entropy. It is common in economics to use a variant called the Theil index [31]. For a cumulative probability distribution F with finite mean $\mu(F)$, the Theil index is defined by the equation

$$T = \int \frac{x}{\mu(F)} \log\left(\frac{x}{\mu(F)}\right) dF(x).$$

For observations $y_1, \ldots, y_n$ we can use the estimator

$$\begin{aligned} \bar{T} &= \log n + \sum_{i=1}^{n} \frac{y_i}{S}\log(\frac{y_i}{S}) & (1) \\ &= \sum_{i=1}^{n} \frac{y_i}{\bar{y}}\log(\frac{y_i}{\bar{y}}) & (2) \end{aligned}$$

where $S = \sum_{i=1}^{n} y_i$ and $\bar{y} = S/n$. Computer scientists will recognize this as being closely related to the observed entropy for the distribution (subtracted from the maximum possible entropy). In the case where the distribution is uniform, the T measure is zero, and when the distribution becomes unequal, the T measure grows to a maximum possible value of $\log n$.

A word should be said about numerical methods in the calculation of inequality indices, because the empirical distributions that are being considered in this paper consist of billions of tiny observations. In particular the expression in equation 1 should not be used directly, because accumulation of many values of $x \log x$ for x near 0 will produce dramatic roundoff errors if performed naively. By contast the expression in equation 2 is more amenable to direct calculation since it avoids this problem. Other tricks may also be employed to group values with the same value of y_i and thereby reduce the number of summands, but in general care should be employed in the calculation. For small sample sizes, bootstrap methods are preferable for estimating the Gini index [20].

2.3 The Axiomatic Approach

Statistical measures of inequality are typically desired to satisfy a set of axioms [8, 9]. For convenience of understanding, we state them in terms of income:

The Pigou-Dalton Transfer Principle A transfer from a poorer person to a richer person should not cause a decline in the measure of inequality.

Scale Independence Inequality measures should be unaffected if there is a uniform proportional change in scale among values.

Dalton's principle of population If an entire population is replicated, then the overall inequality measure of the population should be unchanged.

Symmetry principle The inequality should depend only on the values, and not on other characteristics (sometimes called anonymity).

Decomposability The overall inequality should be consistently related to subsets of the population. For example, if the population is partitioned into two subgroups A and B, then if inequality rises within the two groups, it should also rise within the entire population. Refinements of this axiom state that there should be a formula relating the inequality within A and within B and the inequality between A and B.

All of the measures discussed here satisfy the first four axioms. The decomposability axiom is by far the most complicated, but is motivated by the desire to break down a population into constituent parts so that the determining factors of inequality may be understood. Cowell has proved that the only measure that satisfies all of these axioms in complete generality is a generalized class of entropy measures [9], given by

$$GE[\beta] = \frac{1}{n\beta(\beta-1)}\sum_{i=1}^{n}\left[\left(\frac{y_i}{\bar{y}}\right)^{\beta} - 1\right]$$

Here the entropy measure corresponds to the limiting case of $\beta = 1$. By choosing different values of β it is possible to give greater weight to different parts of the scale.

Decomposability implies that it is possible to break down the population into subgroups and measure the contribution to inequality that arises from the individual groups among themselves and between the groups. The Theil index is particularly simple in this way, since the relationship is a simple sum $T(p) = T_w + T_b$ consisting of the intra-group inequality and the between-group inequality.

2.4 The choice of statistics

Each of these indices suffer from inadequacies, but they complement each other well. The Gini index fails to satisfy the strongest version of the decomposability axiom, since it mixes between-group distribution and across-group distribution. It still satisfies a weaker form of the decomposability axiom [26]. It is interesting to note that the variance fails to satisfy the scale independence and scalability axioms. The Theil index makes it difficult to compare populations of different sizes, since the only absolute upper bound depends on the size of the population. For more information on this subject, see [8, 9].

There has been a great deal of debate about which of these measures is most appropriate for analyzing income data, where social judgements have enormous political consequences. It is impossible for a single statistic to capture every property of the inequality of a distribution, and numerous other measures of inequality have been proposed. Examples include the Hoover index (also known as the Robin Hood index), the Atkinson index, the mean logarithmic deviation of income, the Dalton index, and the Herfindahl index. In addition, one may compare different distributions using the concept of stochastic or Lorenz dominance [9]. In the interests of brevity we confine ourselves to the Gini and Theil index, which are the two most common indices.

One potential objection that might be raised in the context of attention distribution is the fact that there are very few obstacles to the creation of new web content, and when we include all pages in the discussion we skew the distributions by adding pages that are very unlikely to contribute value to humans. This is in fact one of the effects that forms the motivation for this paper, but in some of the analysis that follows I will confine my attention to only the k most popular resources as measured by traffic. Changes to these distributions are less susceptible to distortion that would be induced by counting all pages and sites, but still exhibit interesting characteristics.

2.5 Parametric methods

In the web community it has been popular to hypothesize a mathematical model for an observed distribution (e.g., indegree), and then fit a parameterized curve to the data. This approach was pioneered by the economist Pareto [25] in the 19th century with his early study of income distribution. For example, we might hypothesize that the distribution of attention on web pages is distributed as a Zipfian distribution, as $p(k) \sim k^{-\beta}$ for the $k - th$ most popular web page. If we fit such a curve to the data, then β becomes a measure of the inequality of distribution. Unfortunately, such a statistic often provides relatively little insight into the underlying data, whereas the nonparametric measures provide clear intuitive understanding.

Moreover, this approach depends on assumptions that are largely untested and subject to dispute. In cases where we have high confidence about the determinants of the distribution, the parametric approach may be justified. The forces that drive web traffic are anything but simple however, since they reflect almost every aspect of society including seasonal effects, monetary effects, effects from world events, effects from other forms of mass media, technological shifts, linguistic effects, political effects, etc. I would claim that in this case it is premature to hypothesize a comprehensive mathematical model for how users consume information, and in this paper I shall mostly confine myself to nonparametric methods and empirical observations. It will remain an extremely interesting area of research to devise mathematical models that describe the characteristics that can be seen in such a dynamic information market as the World Wide Web.

3. DATA FROM BROWSING

In trying to understand the distribution of attention across web resources, we should let the data speak for itself. The primary data source used in this study comes from users of the Google Toolbar. This piece of software has been installed on a very large number of machines, and if the user elects to allow it, the URLs that people surf to are reported to Google in order to retrieve the PageRank and other information about the page. This provides a large stream of data for analysis of browsing habits, though some precautions must be given regarding inference from this data.

- Google toolbar users tend to be technologically savvy, and either capable of installing their own software or else acquired their software from partnership arrangements. This distinguishes them from average web users.

- Due to the obvious and serious privacy implications concerns about this data, users must opt in for advanced features of the toolbar in order for their data to be reported. This skews the selection of users, probably reducing the number who surf to socially sensitive sites (e.g., porn), as well as those whose primary information consumption has business implications or government activity (e.g, corporate and government employee web users).

- URLs reported from the toolbar are canonicalized to remove potentially sensitive data (again due to privacy concerns, since some web sites encode user-specific data in URL arguments). By design, this sometimes obscures other non-sensitive URLs and results in our seeing somewhat fewer pages than are actually viewed.

- Identical web resources (or nearly identical) often can be fetched by many different URLs. No duplicate elimination has been done on the underlying data, so we may actually overestimate the amount of information that is being viewed.

- Google toolbar users tend to use the Google search engine, so our data cannot be expected to accurately reflect traffic to competing sites.

- Toolbar data is not authenticated, so is unreliable as a measure of true traffic.

In spite of these caveats, the data gathered in this way provides an expansive view of web browsing habits by many millions of users and many billions of individual clicks. There are currently very few sources of available data for such a broad range of web users across so many cultures and languages.

The data used for some of this investigation was compiled from toolbar usage for a one-month period overlapping September and October in 2005, 2006, and 2007. The choice of this time period was to eliminate periodic fluctuations with small period (e.g., weekly and daily). Unfortunately, a fair amount of web traffic exhibits yearly periodicity (e.g., sports, holidays, weather, travel, etc). Using only a month of data means that we see only a subset of the data that is consumed during the course of a year. The selection of three successive years during the same month allows us to examine what is happening to the distribution over a longer period of time.

In counting web resources, there is some doubt about exactly what granularity is sensible for examining data. In this study I worked with three different levels of granularity, namely URLs, hostnames, and sitenames. Individual URLs are perhaps the most natural, but they can be skewed by the fact that the same resource can have so many distinct URLs, and the raw size of the data makes calculations difficult. In order to overcome this, I worked with a random sample of data, sampled randomly according to the hash of the URL.

Aggregation by hostnames has the advantage that the URLs on a hostname often represent a coordinated information source with a common theme and coordinated strategy for attracting and holding attention. In order to remove some redundancy, I stripped leading `www.` strings from hostnames. Under this level of aggregation, visits to a URL on the same hostname are counted as visits to the same resource.

Unfortunately grouping by hostnames suffers from the fact that it considers all of sites such as `geocities` as a single information unit, when in fact it consists of a large number of individually authored subunits. It also fails to recognize that `research.google.com` and `www.google.com` are related in that they both correspond to the same organizational unit. In order to address this, I also used a unit of aggregation called a site. The definition of a site uses a few heuristic rules to determine the domain associated with a URL, but also keeps a few large sites such as `geocities` as distinct. Thus each of `geocities.com/comp_go`, `google.com`, and `cam.ac.uk` are counted as "sites".

The effort to consider different aggregations made substantially little difference in the overall results however. URLs tended to exhibit more inequality as might be expected, since top-level URLs accumulate a substantial fraction of the traffic on a site during navigation. Aggregating by hostname or sites influenced which sites came out on top, but had little effect on the overal shape of the distributions. Part of the reason for this is undoubtedly due to the stripping of URL parameters as reported in the toolbar data.

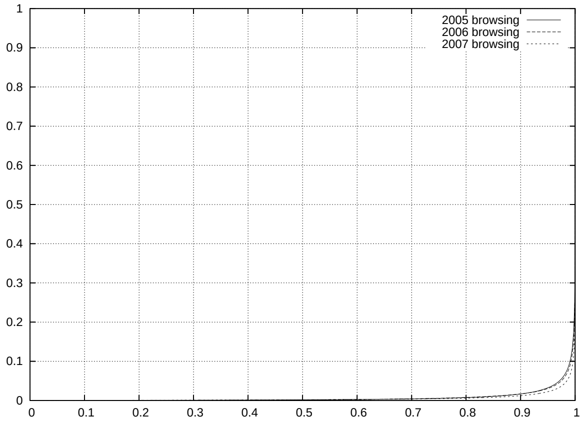

Figure 1: Lorenz curves for the distribution of attention in browsing over all hostnames over a one-month period in three successive years. Note that 50% of the attention is focused on approximately 0.2% of the hostnames, and 80% of the attention is focused on 5% of hostnames. Though barely perceptible in this graph, each year's curve is dominated by the previous year's curve, which indicates that attention is being focused on a smaller number of the top million hostnames.

4. THE IMPACT OF SEARCH ENGINES

Given the level of concentration of attention on such a small number of hostnames, it is natural to ask what forces have caused this. It is important to observe that web browsing falls into four broad categories:

Communication and community This includes email, social networking, and reading blogs of friends.

Education and discovery This is a broad cateogory, and covers the human's basic desire to understand the world around them. In cases where a user is uncertain where to go for information, search plays a very strong role here.

Commerce Web usage for online shopping overlaps somewhat with education, since buying decisions are often influenced by their search for information about products and services.

Entertainment a good example of this is most of the traffic to youtube.

In pursuit of each of these, users tend to employ a mixture of hypertext navigation and search. In the last decade, search engines have come to occupy a central role in determining which pages are seen by people. Prior to the emergence of search engines, researchers in hypertext often spoke of the problem of being "lost in hypertext" (see [22]). Users who have an information need that is at least partially defined now routinely start with a search engine.

The previous sections have focused on the habits of toolbar users as they surf the web. Typical user behavior observed in this way is a mixture of browsing and search. Users will often type a search, examine the results on the search engine page, and potentially surf to a few of these pages. They may also use the search engine as a way to find a place to start their browsing, surfing off through many pages without returning to the search engine. By interacting with a search engine, users are presented with a huge quantity of pages to potentially browse to, but due to the limitations of a search engine interface, they are usually directed toward only ten results per query.

An immediate question that springs to mind is whether the use of search increases or decreases the inequality of distribution of pages that are viewed. In this section we examine the distribution of attention across pages that are search results clicks, with an eye to whether this distribution exhibits more or less inequality in the distribution.

In order to answer this question, I took data from a month of user result clicks, and compared the distribution against the distribution of browsing as exhibited by toolbar users. Since this is aggregated across all users and queries, it reflects the aggregate influence of using search. The results are shown in Figure 2.

5. INEQUALITY BETWEEN AND ACROSS GROUPS

In trying to understand the nature of income inequality, economists strive to break down the inequality into different groups and draw inferences on the cause of inequality in order to guide social policy. While there seems to be general agreement that income inequality is increasing in the United States, there is less agreement about the causes of it, or whether this is a global phenomenon. Some economists have argued that globalization is contributing to increasing income inequality [12, 21, 1] in developed countries, though at least one Nobel-prize winning economist has argued otherwise [4]. A recent IMF report [19] has suggested that

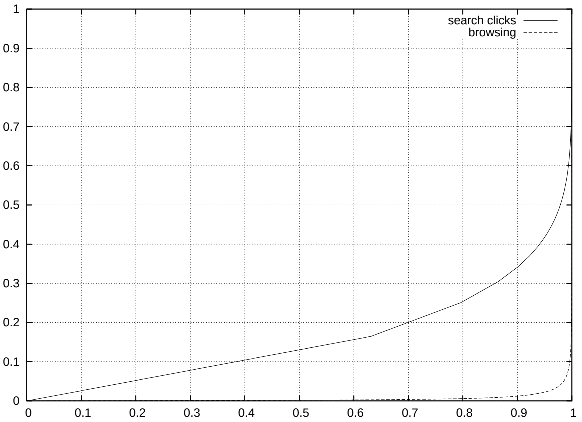

Figure 2: The Lorentz curves for the distribution of browsing of result pages clicked on by Google users in a one-month period, and browsing of all pages. The curves are dramatically different, indicating that search produces a much more equal distribution of attention across pages on the web.

technological progress has had a greater impact on inequality within countries. O'Rourke has argued that the primary source of inequality over the last 200 years has been due to a rise in between-country inequality [21]. All of these analyses seek to explain the origin of inequality in income.

We can apply the same approach in analyzing the distribution of attention as a currency, by breaking down by various groups to see if attention inequality exists within web sites, or whether it arises from inequality between web sites. We may also examine how inequality is distributed among different languages, geography, or different types of content (e.g., news, entertainment, etc). In this section I will show that while there is some variation in the inequality between different groups, the trend toward highly skewed distributions appears to be ubiquitous.

5.1 Commercial influence in the web

One might suspect that inequality arises in part from competitive economic pressures, and the increasing commercialization of the web. Figure 3 shows the differences between the distribution of attention for URLs in the `co.jp` domains vs. URLs in the `ac.jp` domains. While both distributions have fairly high Gini indices (0.992 and 0.921, respectively), there is a noticeable difference between the two distributions. This suggests that most of the inequality comes from commercial parts of the web. Similar characteristics were exhibited for the `co.uk` and `ac.uk` domains, though the difference was less pronounced due to the presence of a large number of special-purpose research sites in the `ac.uk` domain.

By contrast, if we investigate the distribution of attention in the `.org` and `.com` domains, there is a less noticable difference due to the presence of a few domains in `.org` with a huge amount of attention (e.g., wikipedia.org). The Gini indices for the distributions among domains in `.org` is 0.985 vs. a Gini index of 0.994 for browsing over all domains.

5.2 The Influence of Language

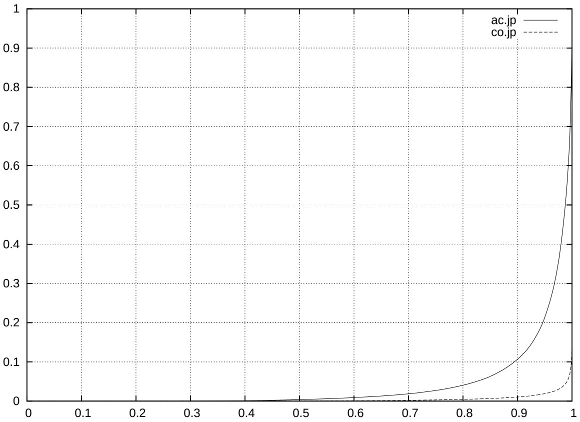

Figure 3: Lorenz curves for the distribution of attention in the `ac.jp` and `co.jp` domains. The commercial domain exhibits considerably more inequality.

In this section I describe the influence of language in determining the inequality of the attention distribution. Information consumers tend to segregate themselves by language, since most people can understand documents written in only a few languages. Google users are allowed to set their primary language in their preferences, and the data in this section shows the distribution of attention broken down by this setting.

The Lorenz curves for the different attention distributions are shown in Figure 4. The curves exhibit a fair amount of variation in behavior, with English exhibiting a significantly higher concentration of attention on a relatively small number of sites.

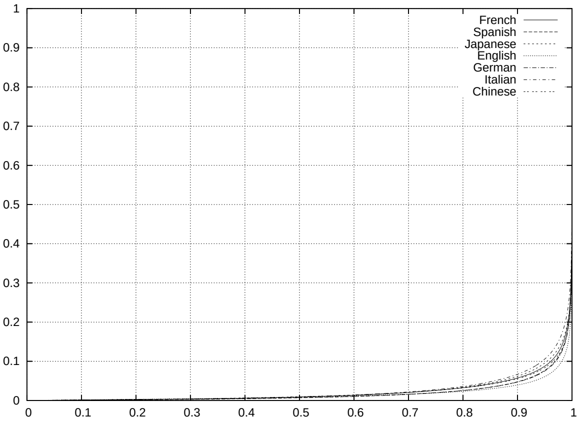

Figure 4: Lorenz curves for the distribution of attention when URLs are partitioned by the preferred language of the user.

While there is some noticable variation in inequality among different slices of browsing behavior, none of the differences are even close to being as different as the difference between browsing behavior and that induced by web search.

6. PAGERANK AS A MODEL OF ATTENTION DISTRIBUTION

> It is a capital mistake to theorize before you have all the evidence. It biases the judgements.
>
> - A. C. Doyle, 1888

The original motivation of the PageRank algorithm [23] was to model a web user as a finite state machine, in which at any given time they perform one of two choices. With probability α, they choose an outlink from the page they are currently viewing, and follow that to a new page. Alternatively, with probability $1 - \alpha$ they jump to a page chosen according to some other criterion. The random jump process is often referred to as "teleportation", and the original paper suggested several possible choices for a teleportation model. In one model they suggest that the user will jump to a page selected uniformly at random. In another model, they might jump to a small set of "trusted" pages, from which they would once again begin their random walk process. More generally, we can model the teleportation process as jumping to a page u with a probability distribution $T(u)$. Some have suggested that this distribution should also depend on the user [17].

Using this model, we can think of the web user as being modeled by a Markov process with transition probability matrix

$$A = (\alpha L + (1 - \alpha)T)$$

where L is the incidence matrix of the hyperlink graph, and T represents the teleportation process. In the case of uniform teleportation, it is easy to prove that the underlying Markov chain is irreducible and there exists a unique stationary distribution p that represents the probability that a user will encounter the page.

Several papers [24, 32] have argued that the PageRank distribution has a power-law distribution, but this requires some explanation. In [24], they consider the limiting case of uniform teleportation when $\alpha \to 1$. In [32], they also restrict their attention to the uniform teleportation model, and observe that PageRank scales with $1/n$ due to the teleportation contribution. In order to normalize things, they consider the scale-free version of PageRank namely $R(u) = nPR(u)$. They then build a mathematical model of the scale-free PageRank in which they argue that $P(R > X) \sim cx^{-\beta}$ for some constant β that depends upon α. Hence the *values* themselves of PageRank are not distributed as a power law, but rather the number of pages with PageRank larger than a given scaled value is distributed as a power law.

In the case of uniform teleportation, the actual values of PageRank are bounded below by a constant fraction of the total mass, namely $\frac{1-\alpha}{n}$. Using the language of welfare econonomics, this corresponds to a minimum poverty level for a page. Before we push this analogy for welfare economics too far, we should remember that web pages are *not* conscious beings who deserve a minimum standard of living, nor is there any apparent justification for a "tax" to be placed on links in order to donate traffic to random web pages. In fact, the experience of search engines combatting web spam has shown just the opposite to be true. In reality, there are many web pages being created with no inherent human value other than to the creator, namely to artifically enhance the rank of other pages for a commercial interest of the producer. Moreover, in this market there is essentially no scarcity of resources that inhibits the production of new pages for this purpose. If we apply the principles of welfare economics to the web, there is no intrinsic justification for using uniform teleportation, and in fact a strong case can be made against it.

In fact, the major motivation for using uniform teleportation to randomly selected pages did not arise from any fairness considerations. Instead, it was motivated by a desire to guarantee that the underlying Markov chain is irreducible, that a unique stationary exists, and that the power algorithm will converge in a reasonable amount of time because the second largest eigenvalue of the incidence matrix corresponds to the probability of taking a random jump [16]. Anyone who actually tailored their browsing habits to use random jumps with probability 0.15 at each page would quickly grow tired of the process and give up. A more natural model is that users conceive of an information need (possibly very broadly, such as to be entertained), and then either use a search engine to point them in the right direction or else draw upon another source for a page that satisfied their need. Models of web browsing should probably use a teleportation probability distribution that more accurately reflects information needs of users.

All of this says essentially nothing about the use of PageRank in search, but rather it argues that as a model of traffic, PageRank suffers from a few weaknesses. In reality PageRank remains a very useful signal for deciding which pages best match a user's information need among the many that match a query.

If we consider the Lorenz curve of a distribution whose values are bounded below by $\frac{1-\alpha}{n}$, it is bounded below by a line of slope $1-\alpha$, so that the Gini index is bounded above by α. In order for PageRank to represent an accurate model of information consumers, we need to understand their actual or ideal behavior. This highlights one of the problems with using uniform teleportation in the calculation of PageRank, since the Gini index of the actual attention distribution as witnessed on real traffic will be shown in section 3 to be significantly larger than the publicly discussed values for α.

In [2], the authors showed that the distribution of PageRank in the uniform teleportation model appears to be highly concentrated in the OUT section of the web graph when the teleportation parameter is chosen as $\alpha = 0.85$. In order to compensate for this perceived inequity, they suggest choosing α close to 0.5. It is interesting to note that this would result in a distribution that is much more uniformly distributed, which makes PageRank even less predictive as a model of attention distribution.

In the original paper [23] that describe the PageRank algorithm, the possibility of modifying the teleportation probabilities to concentrate on a few "trusted pages" was suggested as a possibility. Unfortunately, while this is very helpful in controlling manipulation of PageRank, it results in a very different probability distribution, and unless it is exercised with care it can have troubling social consequences for the web. This was observed previously in [13], where we showed the stark difference between the distribution of PageRank computed using uniform teleportation and PageRank using only a few trusted teleportation destinations. Figure 5 of [13] clearly shows a distribution that decays exponentially with the distance from the teleportation points, and this is intuitively clear since the PageRank decays as α^D for a page at distance D from the teleportation points.

In the context of this paper, PageRank has been considered only as a model of attention distribution, and nothing has been said about the relative rankings on pages that are induced on by computing PageRank with this choice of teleportation. It also says nothing about how to incorporate a different distribution into an overall scoring system for search results, or the overall effects on search quality. Given the pronounced inequality that has been observed in surfing habits of users, it may be the case that teleportation to a small number of trusted pages leads to a more realistic distribution of attention. It may also be the case that a teleportation distribution that is itself heavy-tailed may be a better choice for modeling attention distribution. From the point of view of economics, this would correspond to a progressive tax rather than a flat tax on the income of attention.

7. THE EFFECTIVE SIZE OF THE WEB

The overall size of the World Wide Web continues to grow, but it has become increasingly clear that just counting individual URLs is not an adequate measure of growth because many of the pages are not created with consumers in mind. It is therefore reasonable to ask what the *effective* size of the web is, and how it will evolve over time.

One approach to quantifying the effective size of the web arises from information theory. In the field of data compression, we often speak of the effective size of the data in terms of the amount by which it can be compressed, which is in turn related to the entropy of the underlying distribution. For a probability distribution defined on n possible values, we can speak of the effective range of the random variable as $E(X) = 2^{H(X)}$. For a uniform probability distribution on n values, the entropy would be $H(X) = \log n$, and the effective size of the range would be n.

Thus the attention probability distribution provides us with a way to measure the effective size of the World Wide Web. Moreover, any mathematical models of the web may be analyzed in these terms to see what they say about the rate of growth of the web.

Consider for example a Zipfian attention distribution, where the probability of the k most popular page is $\sim ck^{-\beta}$ for some constants c and $\beta > 1$. In this case the entropy of the probability distribution would be

$$H(X) \sim \beta \sum_{k=1}^{n} \frac{\log k}{k^{\beta}}.$$

The startling observation about this distribution is that as the number of pages added to the web grows to infinity, the effective size remains bounded! By contrast, consider the case of a distribution induced by PageRank with uniform teleportation. In this case,

$$H(X) > \sum_{k=1}^{n} \frac{(1-\alpha)\log k}{k}$$

which clearly tends to infinity as $n \to \infty$. Hence a web modeled with uniform teleportation grows without bound.

The observed probability distributions from toolbar data and search result clicks both tend toward a distribution that is at least as skewed as a Zipfian distribution, leading to the conjecture that the effective size of the web will remain bounded as time goes on. In order to track this, I used the attention distribution predicted by browsing with the toolbar, and calculated the entropy of the underlying probability distributions. During the three years of 2005, 2006, and 2007 the number of hostnames encountered by toolbar users increased each year, with a 40% jump from 2005 to 2006, and more than doubling from 2006 to 2007. In order to compare populations of the same size, I restricted attention to the top one million hostnames from each year, and computed the empirical entropy upon this set. In addition, I computed the Gini index of each distribution. The results are shown in Table 1. The data clearly shows that the attention distribution became more unequal with each successive year. In fact, the Lorenz curves (shown in Figure 1) from one year to the next were completely dominated. This shows that at *all* points in the rank, the attention was shifted to more popular hostnames.

Year	Gini	entropy	effective size
2005	0.98514	14.97	32092
2006	0.98592	14.76	27744
2007	0.98732	14.27	19823

Table 1: The effective size of the World Wide Web as measured by the entropy of the observed surfing distribution from toolbar data on the top one million hostnames. Since the entropy declines each year and the Gini coefficient rises, this indicates that the trend is toward more attention being concentrated on fewer sites. The effective number of sites according to this distribution is fewer than 20,000 sites among the million with the most attention.

8. CONCLUSIONS

Observations of web browsing behavior described in this paper lead to the overall conclusion that attention is being focused on a fairly small number of web pages and web sites. Moreover, there is slight evidence that this trend has accelerated in the last three years. This suggests that the rate of growth for top-rated sites is outpacing the rate of growth for the “long tail markets”. These observations apply whether the data is derived from visits to URLs, hostnames, or sites.

The cause of the inequality in distribution of attention among web users is open to speculation. Due to the structure of most web sites, the links available to a user in an ordinary browsing session tend to be very redundant. Empirical evidence seems to suggest that the increasing commercialism of the web also has resulted in a tendency to have fewer offsite links.

We have also investigated the attention distribution induced by users who click on search results. This distribution exhibits less inequality than distributions observed by ordinary browsing. This is to be expected, since search engines are able to cast users into completely unknown territory that might be many clicks away from their bookmarks or browsing history.

The World Wide Web continues to evolve rapidly, with huge numbers of web sites appearing each year, and a few sites that gain tremendous traffic in a relatively short time. In spite of this growth, this study suggests that user browsing habits are being concentrated on a shrinking portion of the Web. We can only speculate as to the reasons why this is happening. One possible explanation is that the recently developed sites have stimulated so much of a following that they have cannibalized attention from older sites. Another possibility is that the inter-site link structure of the web is losing its importance. The increasing commercialization of the web, and the introduction of many sites and pages of poor quality may be inhibiting page authors from linking to sites beyond their control. Whatever the reasons are, it holds potentially serious consequences for the monetization of web content, since attention will continue to be a prerequisite for monentization.

9. REFERENCES

[1] *The Inequality Predicament.* Report on the World Social Situation. United Nations Department of Economic and Social Affairs, 2005.

[2] Konstantin Avrachenkov, Nelly Litvak, and Kim Son Pham. Distribution of PageRank mass among principle components of the web. In *Proc. 5th Workshop On Algorithms And Models For The Web Graph*, Lecture Notes in Computer Science, San Diego, 2007. Springer. to appear.

[3] R. L. Basmann, K. J. Hayes, and D. J. Slottje. *Some New Methods for Measuring and Describing Economic Inequality.* JAI Press, 1993.

[4] Gary S. Becker and Richard Posner. World inequality, December 2006. Blog entry.

[5] Yochai Benkler. *The Wealth of Networks.* Yale University Press, 2006.

[6] Junghoo Cho and Sourashis Roy. Impact of search engines on page popularity. In *Proceedings of the World-Wide Web Conference*, 2004.

[7] Junghoo Cho, Sourashis Roy, and Robert E. Adams. Page quality: In search of an unbiased web ranking. In *Proc. ACM International Conference on Management of Data (SIGMOD)*, 2005.

[8] Frank A. Cowell. *Measurement of Inequality*, volume 1 of *Handbook of Income Distribution*, chapter 2, pages 87–166. 2000.

[9] Frank A. Cowell. Measuring inequality. Oxford University Press, third edition, 2000.

[10] Partha Dasgupta, Amartya Sen, and David Starrett. Notes on the measurement of income inequality. *Journal of Economic Theory*, 6:180–187, 1973.

[11] Thomas H. Davenport and John C. Beck. *The Attention Economy.* Harvard Business School Press, 2001.

[12] Axel Dreher and Noel Gaston. Has globalisation increased inequality?. Technical report, Swiss Federal Institute of Technology, Zurich, June 2006.

[13] Nadav Eiron, Kevin S. McCurley, and John A. Tomlin. Ranking the web frontier. In *Proc. 13th International World Wide Web Conference*, pages 309–318, 2004.

[14] Michael H. Goldhaber. The attention economy. *First Monday*, 2(4), 1997. Online at firstmonday.org/issues/issue2_4/goldhaber/.

[15] Charles M. Goldie. Convergence theorems for empirical Lorenz curves and their inverses. *Advances in Applied Probability*, 9:765–791, 1977.

[16] Taher Haveliwala and Sepandar Kamvar. The second eigenvalue of the Google matrix. Technical report, Stanford University, 2003.

[17] Taher Haveliwala, Sepandar Kamvar, and Glen Jeh. An analytical comparison of approaches to personalizing PageRank. Technical report, Stanford University, 2003.

[18] Arthur F. Jones, Jr. and Daniel H. Weinberg. The changing shape of the nation's income distribution. Technical Report P60-204, U. S. Department of Commerce, Economics and Statistics Administration, 2000.

[19] Subir Lall, Florence Jaumotte, Chris Papageorgiou, Petia Topolova, Stephanie Denis, and Patrick Hettinger. Globalization and inequality, chapter 4. World Economic Outlook. International Monetary Fund, October 2007.

[20] Jeffrey A. Mills and Sourushe Zandvakili. Statistical inference via bootstrapping for measures of inequality. *Journal of Applied Econometrics*, 12, no. 2:133–150, 1997.

[21] Kevin H. O'Rourke. Globalization and inequality: Historical trends. In *World Bank Conference on Development Economics*, May 2001.

[22] M. Otter and H. Johnson. Lost in hyperspace: metrics and mental models. *Interacting with Computers*, 13, no. 1:1–40, 2000.

[23] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. Technical report, Stanford Digital Library Technologies Project, 1998.

[24] Gopal Pandurangan, Prabhakar Raghavan, and Eli Upfal. Using PageRank to characterize web structure. *Internet Mathematics*, 3, no. 1, 2006.

[25] V. Pareto. *Cours d'Economie Politique.* Droz, Geneva, 1896.

[26] Graham Pyatt. On the interpretation and disaggregation of Gini coefficients. *The Economic Journal*, 86, no. 342:243–255, 1976.

[27] Amartya Sen and James Foster. *On Economic Inequality.* Oxford University Press, 1997.

[28] Jacques Silber. *The Handbook of Income Inequality Measurement.* Springer, 1999.

[29] H. A. Simon. *Designing Organizations for an Information-Rich World*, pages 37–72. Computers, communications, and the public interest. Johns Hopkins Press, 1971.

[30] Michael Strudler, Tom Petska, Lori Hentz, and Ryan Petska. Analysis of the distributions of income, taxes, and payroll taxes via cross section and panel data, 1979-2004. Technical report, United States Internal Revenue Service, 2006.

[31] Henri Theil. *Economics and Information Theory.* Rand McNally, 1967.

[32] Yana Volkovich, Nelly Litvak, and Debora Donato. Determining factors behind the PageRank log-log plot. In *Proc. 5th Workshop On Algorithms And Models For The Web Graph*, Lecture Notes in Computer Science, San Diego, 2007. Springer. to appear.