# A Mixed Bundling Pricing Model for News Websites (Working Paper)

Albert Sun
University of Pennsylvania,
*Senior Thesis in Economics,*
albert@albertsun.info

*Advisor:* Professor Kenneth Wolpin
University of Pennsylvania

April 2010

**Abstract**

This paper outlines a method for finding revenue maximizing mixed bundling prices for news websites. This can help better understand paid content strategies for online news content. Drawing on work in the field of bundling information goods, I apply a two-parameter model of consumer preferences to web site traffic data and a roughly estimated willingness-to-pay curve. We can then calculate revenues for different price points and find the optimal one for any given site. This method is applied to a sample of ten sites. At revenue maximizing prices, the majority of paid revenue for these sites comes from the sale of individual articles, rather than subscriptions. Site traffic showing highly loyal consumers is found to correlate with higher subscription prices. This model suggests that while it is possible for overall revenue to be higher with a paid content plan, total traffic will certainly fall.

# 1 Introduction

A wide array of challenges and opportunities confront the news industry as it struggles to adapt to a networked, digital world that has either removed completely or radically lowered nearly all barriers to entry. Among them is the opportunity to develop new revenue streams by charging consumers directly for the consumption of content.

This, of course, is not a new concept, but for a variety of reasons most news sites today do not charge consumers for access to their content. When news sites first began creating websites for themselves, many of them did try to charge for access to the sites, thinking of them as simple extensions of their paid print circulation. As the Internet experienced tremendous growth in audience and advertising spend online however, this led many sites unable to attract visitors or advertisers. In response, most dropped their paywalls to tap into this growth in advertising spending. News publishers are in the middle of a two-sided market, and so find it more profitable to subsidize readers and reap extra profits from advertisers. (Eisenmann et al., 2006) Demand for news in print is relatively inelastic (Lewis, 1995) compared to elasticity of demand for news online, (Chyi, 2005) meaning that to take advantage of the same two-sided market, publishers will give readers even greater subsidies.

However, as the Internet matured this two-sided market began to break down as a glut of online advertising inventory drove prices down and the same consumers could be reached by advertising messages on more and more sites.

Over the past year, there has been a rash of new discussion of paid content strategies online as publishers steel themselves to try again. In 2009 and early 2010 several publishers have either rolled out pay systems on their sites or announced plans to do so.

While it's by no means a settled question that news sites *should* charge for content, in this paper I will instead examine *how* they could implement one particular strategy.

# 2 Literature Review

## 2.1 Characteristics of Information Goods

Normal market goods are expected to be excludable, rival and transparent. DeLong and Froomkin (2000) explain how information goods have none of these three characteristics.

Information goods are non-rival. That is, one consumers consumption of an information good does not adversely affect consumption by others. Two people can read the same news article, but they cannot eat the same sandwich.

Information goods are non-excludable. It is difficult to impossible to bar someone from consuming the good because of how cheap and easy it is to copy digital data. The music and movie industries losing battle with peer-to-peer file sharers shows the futility of attempting to enforce this constraint on information goods.

They are also non-transparent. When making a purchasing decision, to have complete information about the good is to have the good itself, and so it is not possible to make perfectly informed decision because a consumers valuation of a good is known to no one ahead

of time. Additionally, the types of digital goods and services sold have become increasingly difficult to value as the goods and services themselves become more complex. (DeLong and Froomkin, 2000)

Finally, and obviously, information goods are characterized by a cost structure that is an extreme case of one that some normal goods have. They have an extremely high fixed cost of production and near zero marginal cost for producing and distributing additional units. Because of this value-based pricing, and in particular differential pricing, is essential. (Varian, 2000)

All these characteristics can be both good and bad for the firms that produce them and consequently a variety of methods of exploiting or minimizing their effects have been developed.

## 2.2 Solutions for News Content Producers

Since the information good producer most threatened is the online news publisher, Mings and White (2000) discussed possible business models and paths to profitability for online news publishers. They examined four different possible revenue models: subscription, advertising, transactional and bundled. In the time since, the market has had some chance to test them.

The subscription model is a close cousin of the print subscription model, charging consumers a periodic fee for content and other services on the publisher's website. Prominent examples of this model include The Wall Street Journal and The Financial Times. Many other publishers including the Philadelphia Inquirer, the Los Angeles Times and Slate.com have experimented with online subscriptions but ultimately abandoned payment plans in favor of trying to grow traffic. With the exception of advertising, all the models described in Mings and White (2000) have so far seen little success.

The problem for producers of content remains how to maximize revenues. How to maximize revenues with marginal costs constant at zero and high fixed cost of production. In a market with perfect substitutes for information goods and no collusion, a ruinous price war will inevitably ensue due to Bertrand competition. Each firm undercuts the other to capture the whole market and earn greater profit until price is driven down to marginal cost.

Fortunately, no producer of information goods faces competition with perfect substitutes. In a competitive context with imperfect substitutes and two companies each producing a single good that is an imperfect substitute for the other company's good it is possible to reach a strong equilibrium with non-zero prices if the two publishers pursue different bundling strategies. (Fishburn et al., 2000) One charges a fixed fee for its production, and the other charges a metered rate. In this situation it is occasionally possible to find strong equilibria with non-zero prices. However, at other times a price war ensues, and under nearly all conditions total producer profits are lower. The competitive scenario is one that will require future study.

If, however, publishers can differentiate their products enough to maintain pricing power they can do much better. Varian (1995) reviews two of the major areas of study related to information goods: price discrimination and bundling. He finds that in the case where a

producer can differentiate their good enough from the competition that they have pricing power, both strategies are necessary to maximize revenue.

The field is appealing as it touches on the real world decisions that many companies are facing today. For example in producing a suite of software products like Microsoft Office or as previously mentioned, Adobe's Creative Suite. A publisher producing a wide variety of news articles is also a multi-good monopolist as each article can be thought of as a separate good. How should such a publisher sell their products? Would aggregating them and selling them as a bundle increase profits?

Many experiments in this field are beginning as news publishers begin to tentatively roll out paid sections of their websites. A 2009 survey of newspaper executives reported that while only 10% had any sort of paid site, 58% were considering implementing a paid site. (ITZBelden, 2009) Studies conducted in 2007 and 2010 show some evolution of paid content plans among newspapers. Herbert and Thurman (2007) reviews the paid plans offered by British newspapers and finds that while almost all had some paid offerings, they were generally not for the core news content. Bruce (2010) finds that many more site features and core content have begun to move under the umbrella of a paywall in both British and American newspapers. However, very few sites allow readers to access nothing without paying.

## 2.3 Bundling and Unbundling

News articles whether printed or on screen are an information good. Printed news articles have historically been sold as a bundled product, combined into a newspaper instead of sold separately. The cost profile of an article in a printed newspaper fits squarely in the region well suited for pure bundling. Under different sets of assumptions, the authors also argue generally for bundling as a strategy for the distribution of information goods.(Bakos and Brynjolfsson, 1998, 1999)

Bundling of news articles in a printed product makes sense because of the cost side of the equation. But online, where the marginal cost of every aspect of production and distribution is quickly dropping to zero, it makes less sense to base pricing or bundling decisions on costs.

Bundling's profit maximizing power as presented both by Bakos and Brynjolfsson and Varian (1995) comes from the ability of bundling to average consumer valuations for goods and reshape the demand curve.

Bakos and Brynjolfsson (2000) propose aggregation of goods into bundles as a way to smooth consumer valuations and maximize revenue. Supposing that consumers valuations for a set of goods varies and that their valuation for a bundle is the sum of these separate valuations, the bundle of goods will be more likely to have a valuation close to some average. In their model, they provide a proof that if valuations for goods in the bundle are independent, then no matter what the distribution of those valuations is, bundling will increase profits.[1]

[1] Pure bundling of zero marginal cost goods with independent and identically-distributed valuations dominates pure unbundling.

In Chuang and Sirbu (1999) however, they conclude that while bundling can increase profits, mixed bundling is actually even more profitable and will always be a dominant strategy over both pure bundling and pure unbundled sales. By providing appropriately priced options to users it is possible to capture consumer surplus both from individuals with high valuations for specific pieces of content and zero valuation for others, and from individuals with low valuations for individual items in the bundle but a high value placed on the entire bundle.

Both these papers draw on the rich bundling literature studying the bundling of two goods in comparison to single good selling. Adams and Yellen (1976) had already concluded that in most cases mixed bundling would be the dominant strategy.

They proposed three benchmarks for assessing the profitability of a pricing strategy: complete extraction, exclusion, and inclusion. Complete extraction is the extraction of all consumer surplus, exclusion means excluding any consumers from consuming a good if their reservation price is below the cost of producing that good, and inclusion meaning every individual whose reservation price for a good exceeds its cost does consume the good. Because inclusion and exclusion are based on cost, they are much less useful for assessing information good bundling. Additionally, they also identified the correlation of an individual's valuation for the goods in the bundle to be an important determining factor in which pricing scheme is optimal.

Schmalensee (1984) discussed how different distributions of consumers affect the profitability of bundling. In particular, changes not only in the mean, but also in the standard deviation of willingness-to-pay can have a significant impact on profitability.

Chae (1992) studies the case of bundling subscription TV channels in the case of two channels and finds that when the consumer's reservation prices for individual channels are uniformly distributed, the welfare maximizing solution is mixed bundling.

Armstrong (1996) generalizes the problem of bundling to any number of products and parameters determining valuation, proving propositions showing that some low-valuation customers will always be excluded from the market. As will be shown later, a mixed bundling strategy will result in prices keeping a large percentage of customers from consuming.

# 3 Economic Model

We use a model of consumer choice inspired by a paper by Chuang and Sirbu (1999). In our model, a website selling access to online content presents consumers with a choice between purchasing a subscription for a fixed period of time or paying for each piece of content to be viewed separately.[2] They use data provided by King and Griffiths (1995) from a survey of readership of printed academic journals and the assumption that willingness-to-pay is uniformly distributed between 0 and 1.

A consumer's utility is dependent on the following five factors:

[2]For an excellent fuller description of mixed bundling and this economic model, really do read the Chuang and Sirbu paper.

- $F$: Price of a subscription.
- $f$: Price of a single article.
- $N$: Number of articles a consumer would consume in a single period for which the consumer has valuation greater than zero.
- $n$: Number of articles a consumer would purchase individually. The valuation for that article is greater than $f$.
- $WTP$: Consumer's willingness-to-pay for their most valued article in a period. Each additional article is assumed to be linearly decreasing in willingness-to-pay. So if $wtp_1 = WTP$, then the second most valued article has $wtp_2 = \frac{N-1}{N}WTP$ and so on. The $Nth$ article has $wtp_N = \frac{1}{N}WTP$.

Then, each consumer's utility is determined by the following equation.

$$U_{N,WTP}(f, F) = max\{\frac{N+1}{2}WTP - F, \sum_{i=0}^{n-1} \frac{N-i}{N}WTP - nf\} \tag{1}$$

To maximize utility, a consumer either chooses the sum over all positively valued articles, times $WTP$ and minus $F$, or the sum over all articles for which $wtp \geq f$ minus $n \cdot f$. The site's revenue from that consumer is then either $F$ or $n \cdot f$.

Referring to the equation 1 for a consumer's choice of purchase we then proceed to have a site determine a pair of prices for a subscription and a single article. At any pair of prices, by summing over the decisions of all consumers we can calculate the site's revenue in that period. By calculating this revenue across a range of values for $f$ and $F$ we find the revenue maximizing prices.

This relatively simple model requires some broad assumptions. The site must have a monopoly on the sufficiently differentiated product it is selling that consumers can't switch to some alternative. We ignore the effect of any competition. Also assumed is that a consumer's $WTP$ and $N$ are independent and that the consumer knows their own $N$ and $WTP$ for all articles in advance and so can make rational purchase decisions.

# 4 Data Source

## 4.1 Site Visitor Data

Our dataset consists of visit depth data for a single month from ten different newspaper websites of varying size. Visit depth is the number of pages on the website visited in each session on the website. As all sites in the sample are currently completely free, it makes sense to use each visitor's current number of pages visited as the number of pages they have positive valuation for. Visit depth is each consumer's $N$.

Of the ten sites, three are from small communities, three from medium size communities, three from large metropolitan areas and one is of national size. These sites will be denoted as $S1$, $S2$ and $S3$ for the small sites, $M1$, $M2$ and $M3$ for medium, $L1$, $L2$ and $L3$ for large, and $A1$ for national. The sites visitor traffic has descriptive statistics as follows.

Table 1: Summary statistics for site data

| | Pageviews per visit | | | |
|---|---|---|---|---|
| **Site** | Mean | 1st Quartile | Median | 3rd Quartile |
| **L1** | 2.81 | 1 | 1 | 3 |
| **L2** | 3.31 | 1 | 2 | 4 |
| **L3** | 3.28 | 1 | 2 | 4 |
| **M1** | 3.06 | 1 | 2 | 4 |
| **M2** | 3.21 | 1 | 2 | 4 |
| **M3** | 3.88 | 1 | 2 | 5 |
| **S1** | 3.38 | 1 | 2 | 4 |
| **S2** | 3.15 | 1 | 2 | 4 |
| **S3** | 4.39 | 1 | 2 | 6 |
| **A1** | 2.23 | 1 | 1 | 2 |

As Table 1 shows, the vast majority of visits to these sites only see one page, or as we are interpreting it, one article. There are several possible issues with this interpretation of pageviews. Many pages on a website are not articles, for example the homepage, section pages, author profiles, etc. Many articles on websites are broken up over two or more pages that a reader must click through. For visits with only a single pageview, it's possible that the visitor saw only the homepage and didn't read any articles, but more likely that they followed a link from an outside site directly to an article. There is little we can do about these kinds of measurement error.

There is some difficulty involved in accurately measuring a website's visitors due to the nature of the web technologies used to do so. The dataset used in this study of visitors, visits and pageviews are measured by the web server and is measured using tracking cookies set on visitor's web browsers.

Tracking cookies, however, are not a truly effective way of measuring the people visiting a website as they do not uniquely identify a person, but rather a single web browser and computer. If one person accesses the same site from multiple browsers or multiple computers they will be counted multiple times. Similarly if more than one person visits a site from the same web browser-computer pair they will be counted as a single visitor. Cookies are also periodically cleared from computers, depending on the privacy settings of the user.

A full discussion of the issues involved in tracking website visitors is beyond the scope of this paper,[3] but the net result is likely to be an over count of the number of people visiting a site, and an undercount of the number of pages per person.

## 4.2 Willingness-to-pay Data

In papers studying bundling more generally, researchers commonly assume that consumers' reservation prices fit some simple probability distribution. Schmalensee (1984) assumes Gaussian demand when he studies bundling. Chuang and Sirbu (1999) assume reservation prices are uniformly distributed. In other literature, assumptions about consumers' reservation price are also unsupported by empirical data.

There is no particularly good dataset for how much consumers are willing to pay for online content. Part of this has to do with the sparsity of data on the topic, with Chyi (2005) one of the few papers on the subject to take an empirical numbers based approach. In Chyi's study of news consumers in Hong Kong, she writes that consumer's response to fee-based services was ultimately "unenthusiastic". However, the analysis is far from inconclusive due to an extremely low $R^2$, likely indicating that several important predictors were not included.

In a survey of readers of Arabic electronic newspapers, AlShehri and Gunter (2002) report that 63% would be unwilling to pay to read newspapers online, many due to the easy availability of free alternatives.

Survey data that has been gathered comes from asking consumers if and how much they would be willing to pay for news online. Recently conducted surveys have produced widely varying results. (Boston Consulting Group, 2009; Harris Interactive, 2010) Simple questionnaire type surveys are highly dependent on the survey methodology and wording of questions asked. Answers to these surveys are likely to reflect how much people *want to pay,* rather than how much they are *willing* to pay.

The data used for this study comes from Boston Consulting Group (2009).

These numbers are used to represent each consumer's willingness-to-pay. An identical distribution was constructed for each segment of users modeled such that their willingness to pay for each individual article fit the above data.

Because the magnitude of the prices and revenue amounts are ultimately unimportant, we use survey results asking about willingness-to-pay for an entire month to represent willingness-to-pay for single articles under the assumption that these two curves should have the same shape.

The values from Table 2 are represented in the model code by cumulative distribution function constructed by interpolation of the lower ends of the intervals so as to conservatively

[3]There is an abundance of literature and discussion on the topic of web audience measurement as well as heated competition and acrimony between a variety of companies striving to be crowned champion of the field. The two leading methods of measuring traffic are the server-side method and the panel method. In the server-side method each page load "beacons" or "pings" a server with a message registering the page load. In the panel method, a representative panel of all Internet users voluntarily install tracking software on their computers to record every page they visit. The activities of this representative sample are then extrapolated to the entire population.

| Amount ($USD) | Percent of answers (%) |
|---|---|
| Nothing | 52 |
| 1 - 3 | 17 |
| 4 - 6 | 12 |
| 7 - 10 | 11 |
| 11 - 15 | 4 |
| More than 15 | 4 |

Table 2: Willingness to pay for online news
*How much per month are you willing to spend to get online news?*

estimate willingness-to-pay. 52% of the distribution is at 0. 17% of the distribution is between 0 and 1. 12% is between 1 and 4.

Figures 1 and 2 are an example of how consumers are modeled. Based on site visitor data a series of 20 bins are created for each site with $N$ ranging from 1 to 20. Each of these bins is assumed to have an identical distribution of willingness-to-pay. For each bin, based on $N$, $f$ and $F$ thresholds are calculated for the purchase decision of all consumers in that bin so that everyone with $WTP$ between some two numbers makes the same purchase decision. These thresholds are described in further detail in the following section.

# 5 Calculation Method

Instead of using analytic methods, numeric results were calculated using the Python programming language with NumPy and SciPy modules.[4] The code used can be seen in Appendix A. Calculations were done numerically by determining the different $WTP$ thresholds at which consumers will make different purchase decisions. These thresholds depend on a consumers $N$ value, which ranges from one to 20, and on the prices $f$ and $F$. In the dataset, all values of $N$ greater than 20 were lumped together and the number of such values were generally small, representing at most 3% of the dataset for one site, and less than 1% for the majority of sites.

Thresholds were calculated by solving the consumer's utility equation 1 for $WTP$. For each marginal purchase of another individual article, the individual's utility from that article must be greater than $f$.

$$\sum_{i=0}^{n-1} \frac{N-i}{N} WTP > nf \tag{2}$$

$$\sum_{i=0}^{n-1} \frac{N-i}{N} WTP - nf > \sum_{i=0}^{n-2} \frac{N-i}{N} WTP - (n-1)f \tag{3}$$

[4]Version 2.6.4 of Python was used. Version 1.3.0 of NumPy was used, and Version 0.7.1 of SciPy.

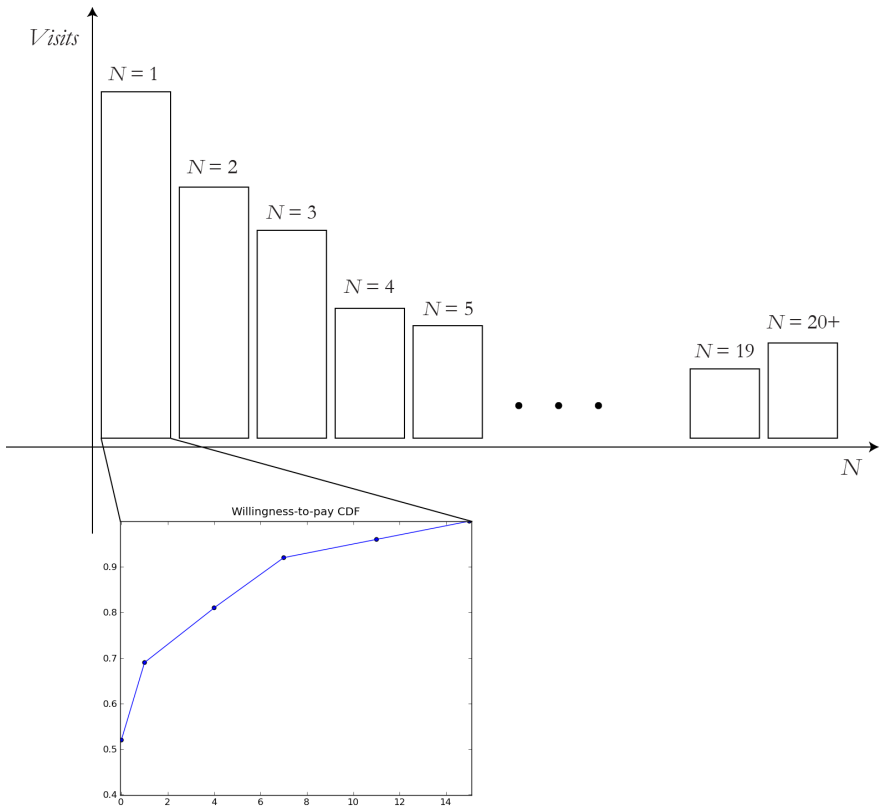



Figure 1: Readers for a site are distributed in bins ranging from $N = 1$ to $N = 20$. They each have the same CDF.

These two equations state that the consumer's utility must be greater than the price they are paying, and that consumer's surplus must be greater than if they were to buy one less article. The largest integer $n$ for which this holds true is the number of individual articles that would be purchased, so we compare $WTP$ against the following values.

$$\{0, f, \frac{Nf}{N-1}, \frac{Nf}{N-2}, \cdots, Nf\} \tag{4}$$

For $0 < WTP < f$, $n = 0$ and revenue is 0. For $f < WTP < \frac{Nf}{N-1}$, $n = 1$ and revenue is $f$. For $\frac{Nf}{N-1} < WTP < \frac{Nf}{N-2}$, $n = 2$ and revenue is $2f$, and so on.

The threshold for purchasing a subscription requires that it provides more utility than purchasing individual copies. The following equations must hold true. The first states that utility from consuming all articles is greater than the subscription price. The second states that the utility from consuming articles $n$ to $N$ is greater than $F - nf$. If both do, revenue is then $F$.

$$WTP > \frac{2F}{N+1} \tag{5}$$

$$WTP > \frac{F - nf}{\frac{N+1}{2} - \sum_{i=0}^{n-1} \frac{N-i}{N}} \tag{6}$$

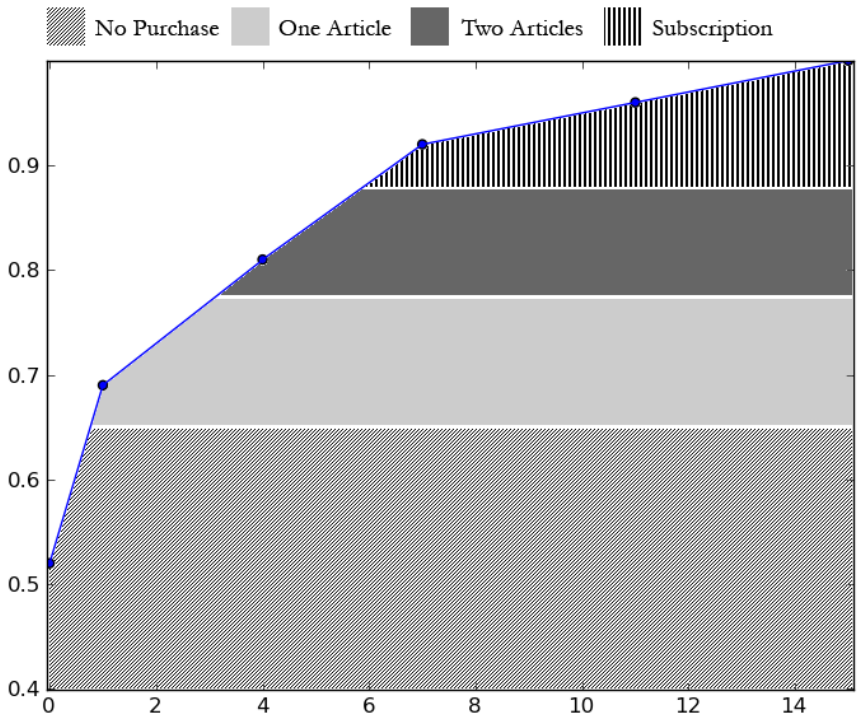



Figure 2: Within each bin, different regions of the CDF correspond to different purchase decisions.

Using these thresholds, the weighted revenue per consumer for each segment is calculated, this is then multiplied by the total number of visits in each segment to determine total revenue from that segment. This is summed over all 20 segments to determine total revenue for a site. Segments with low $N$ will typically choose to buy some number of individual articles, whereas segments with high $N$ will choose to buy either a subscription or nothing at all. Only in a small range of prices and at a middle $N$ will a segment have thresholds for both individual purchases and a subscription purchase. To see how this is implemented in code, see Appendix A section 3.

# 6 Results

## 6.1 Optimal Prices

After computing revenue across a wide range of possible prices, we find the following values maximize revenue for each site with $f$ accurate to within 0.01 and $F$ accurate to within 0.05.

Here in table 3 we see that values of f are scattered between 3.8 and 4.2. Values of F range between 21 and 35 and the ratios between prices range roughly from 5 to 8. Sites with a lower mean pageviews per visit like L1 and A1 have lower values of F and higher values of f as expected, and sites like S3 with higher mean pageviews per visit have a much higher F.

Table 3: Calculated optimal prices

| **Site** | f | F | F/f | Subscriptions (%) |
|---|---|---|---|---|
| **L1** | 4.11 | 21.35 | 5.19 | 30.18 |
| **L2** | 3.82 | 30.65 | 8.02 | 23.99 |
| **L3** | 3.86 | 27.85 | 7.21 | 25.77 |
| **M1** | 3.87 | 28.00 | 7.24 | 23.63 |
| **M2** | 3.85 | 26.30 | 6.83 | 25.26 |
| **M3** | 3.81 | 27.45 | 7.20 | 31.01 |
| **S1** | 3.86 | 26.35 | 6.83 | 28.09 |
| **S2** | 3.87 | 26.40 | 6.82 | 25.87 |
| **S3** | 3.85 | 31.10 | 8.08 | 38.52 |
| **A1** | 4.26 | 21.85 | 5.13 | 23.79 |
| **Average** | 3.92 | 26.73 | 6.82 | |

As can be easily seen in figure 5, sites A1 and L1 stand apart from the others. These two are characterized by the lowest mean pageviews per visit and the lowest 3rd Quartile, with visits at the 75th percentile generating only 2 and 3 pageviews respectively.

These relationships are shown as simple linear regressions in figures 3 and 4 showing that as average pageviews increases, F increases and f decreases.

The relationship between prices, site traffic and the percentage of revenue from subscriptions at the maximum is not as clear. Figure 6 shows that there is some evidence for a positive relationship between average pageviews and a higher percentage of revenue from subscriptions. However, L1 is well above the regression line, despite having a much lower $F$ than other sites.

At these prices, well over half of visitors will choose not to purchase anything.

Figure 7 shows large graphs of possible revenue at various levels of $f$ and $F$ for the site L1. In the far zoomed out view, you can see the two regions running parallel to each axis showing the revenue from either a pure bundling or pure unbundling plan. Even as the price of either $f$ or $F$ rise to extremely high levels, overall revenue declines only slightly. Along either axis, the site can make almost as much money by increasing the percentage of revenue from subscriptions or single article sales.

Charts of sites are shown smaller. Comparison of the charts in figures 8 and 9 show that across the entire set of sites the revenue space looks very similar with only slight shifts in the slopes and locations of maxima.

## 6.2 Discussion

How can these numbers and figures be applied to the real world problem? First, any conclusions drawn from this analysis must be taken warily. Merely programming these numbers into

Figure 3: Optimal $F$ plotted against average pageviews for each site

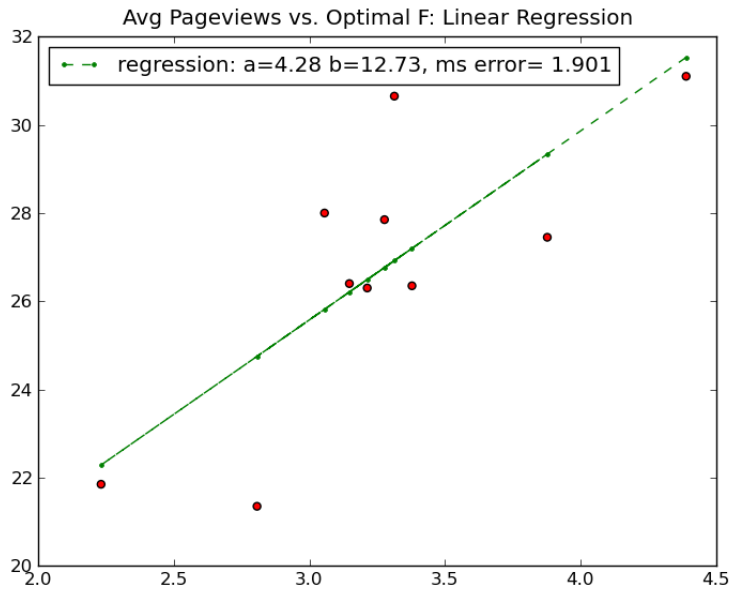



Figure 4: Optimal $f$ plotted against average pageviews for each site

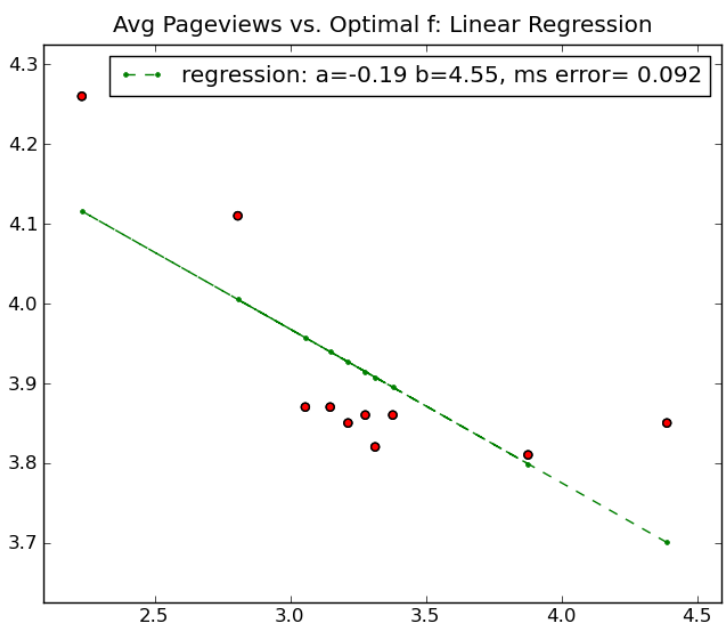


Figure 5: Locations of maximum revenue: $f$ on x-axis and $F$ on y-axis

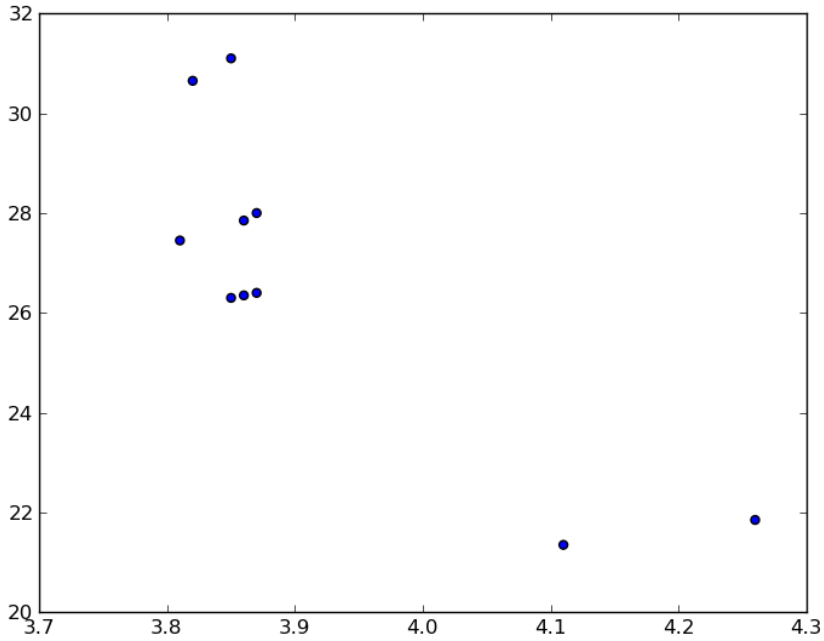
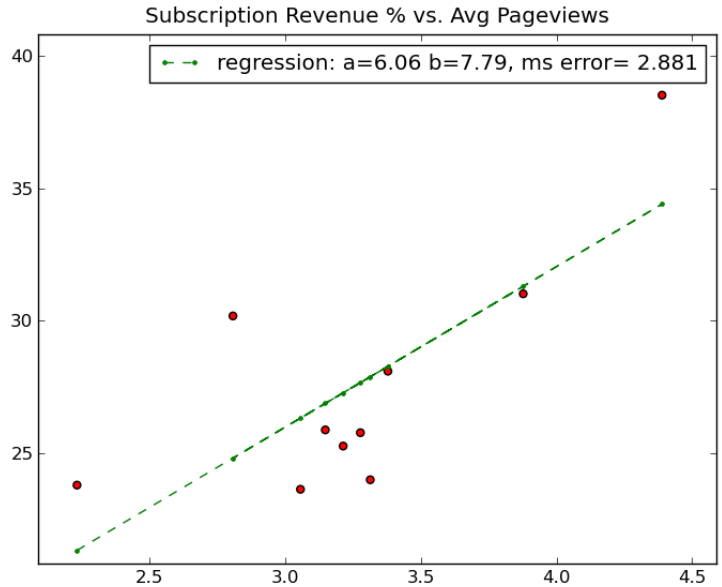



Figure 6: Percentage revenue from subscriptions plotted against average pageviews



a price gun would almost certainly result in disaster as the decision is much more complex than has been modeled.

Total revenue figures for each site from the paywall were not included as results of this study as they are very likely to be misleading. The calculated optimal prices for $f$ are significantly higher than what any newspaper publisher currently charges for an entire day's print edition at news stand rates. The survey data used was applied to $WTP$ for a single article, even though the survey question did not ask about the purchase of a single article. This was justified under the assumption that while the scale of the distribution would be wrong, the shape would still be right. Therefore, the nominal values of $f$ and $F$ are not particularly meaningful. Their relationship with each other and with a site's characteristics are.

If one is committed to interpreting the results for $f$ and $F$ in an absolute sense, it might be a good idea to apply some scaling factor to all the results. Whatever the factor, if it is applied to the estimate of consumer's willingness-to-pay, $f$ and $F$ equally, it will not impact the end result.

Another glaring factor left so far unmentioned is the trade-off with advertising and the exclusion of readers from the market. We are assuming an absolutist model where if a reader does not pay they cannot view the site at all. Since 52% of survey respondents indicated that they were not willing to pay anything for news content online, all those viewers' pageviews are lost along with the corresponding advertising revenue. Of course, this absolutist position is also not realistic. Even a site with the most stringent of paywalls would leave the homepage headlines and teasers for stories visible so as to attract new readers. That translates into pageviews without full article reads.

Table 4: Paid Content Revenues

| **Site** | Revenue/Visit | Revenue/Pageview |
|---|---|---|
| **L1** | 1.34 | 0.48 |
| **L2** | 1.50 | 0.45 |
| **L3** | 1.49 | 0.45 |
| **M1** | 1.42 | 0.46 |
| **M2** | 1.47 | 0.46 |
| **M3** | 1.69 | 0.44 |
| **S1** | 1.52 | 0.45 |
| **S2** | 1.45 | 0.46 |
| **S3** | 1.89 | 0.43 |
| **A1** | 1.15 | 0.52 |

We can make some attempt at estimating the trade-off between paywall revenue and advertising, though we now stray into much murkier territory. Stray (2010) makes a rough

estimation that NYTimes.com receives advertising revenue of $31 CPM (cost per thousand impressions) per page.[5] Let's make the generous estimate that all sites can monetize traffic at that same rate. Based on the calculated optimal prices for each site, we have revenues per visit and revenues per pageview as shown in table 4.

For any site in the sample, 1000 pageviews will generate more revenue from paying readers than from advertising site. Taking $(0.45 \times 1000)/31 \approx 14.5$ suggests that a small base of highly paying readers can support the same site that a much larger advertising audience would.

Many of the results of this study would seem to strain the imagination. It's important to keep in mind that they are the result of the data used. With better measured and more fine grained willingness-to-pay data, for example, results might correspond more closely to examples of successful paid content plans in place now.

## 6.3 Directions for Further Research

There are many possible directions for future research, most of them to improve upon inadequacies in this paper.

To begin with, only mixed bundling is studied. While mixed bundling is superior to either pure bundling or pure unbundling, it's unknown whether it is a preferable to a metered model, customized bundling, bundles of certain types of articles, or some other way of charging for content. No economic study has been conducted of the relative merits of these different models. Even within mixed bundling, the length of the subscription period or offering multiple options for subscription periods at different prices will lead to different results.

An obvious improvement would be to replace the dataset of pageviews per visit with one showing pageviews per unique visitor. This would better account for visitors making multiple visits to a site within a subscription period and would likely result in subscriptions making up a much larger proportion of revenue.

Another area that demands further examination is in estimating the demand curve for individual news articles. This might best be done on a news site with an existing paywall so that it can be derived from actual purchase decisions instead of survey responses. Failing that, a more rigorous survey methodology and a contingent valuation study could also provide better data on demand. This would allow analysis without assuming that $N$ and $WTP$ are independent.

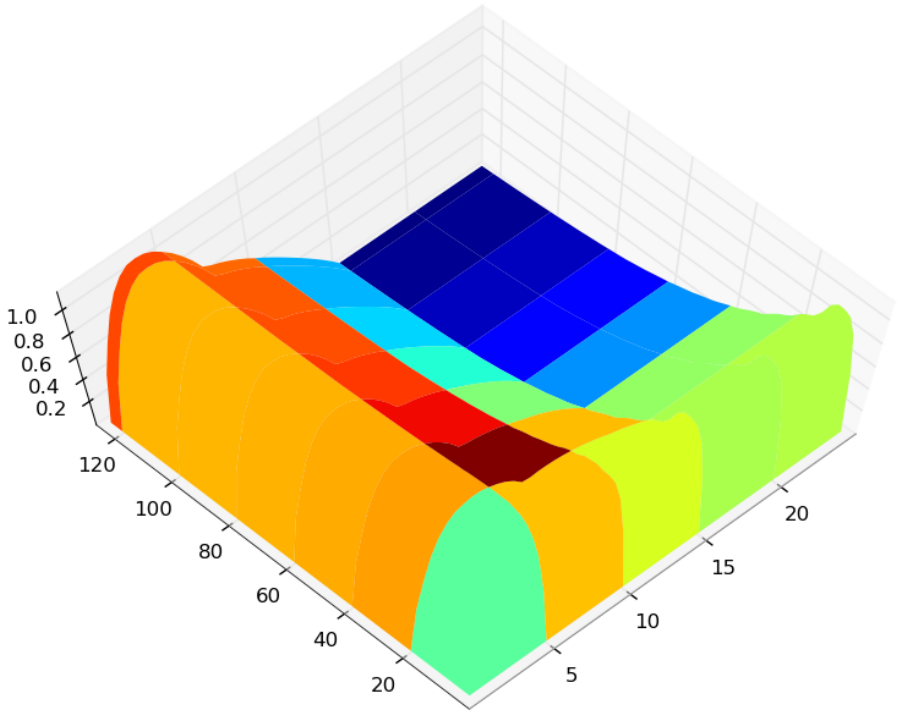
[5] CPMs are an industry standard way of measuring online advertising revenue and refers to the aggregate cost of purchasing 1000 displays of an advertisement on a site. Each page can also have multiple advertisements. The $31 CPM per page means that for every 1000 times a page is loaded, the site earns a total of $31. Stray estimates this number by taking a published figure of $100 million in yearly revenue and dividing that by an estimated 267 million pageviews in one month.

# 7 Conclusion

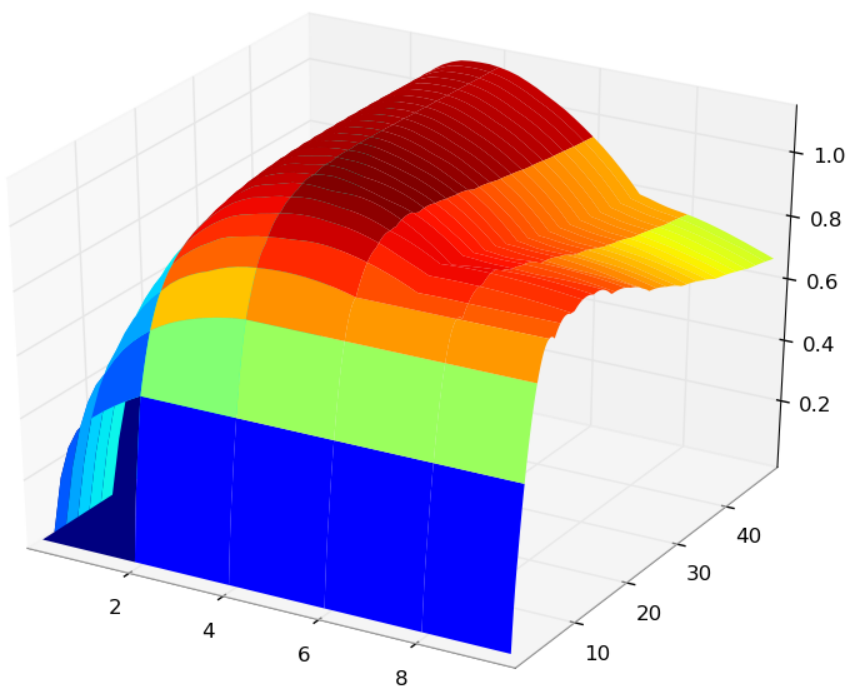
A mixed bundling pricing strategy has the potential to generate significant revenues for a news site. We've seen that when consumers are offered a choice between purchasing a subscription or some number of individual articles, the majority of revenue will come from the individual sale of articles. As expected, when the number of articles a consumer wants to read in each subscription period rises, both the optimal subscription price, and the percentage of revenue from subscriptions also rises. However, without better data it's hard to draw meaningful conclusions beyond these.

This study has proposed a method and framework which, with refinement, could be used as a guide for the publishers of news sites to create a significant second revenue stream. However, this would require the sacrifice of a large portion of the site's audience and could conflict with a publisher's desire to do public service journalism or command the attention of the public.

The real test of this or any other model will always be in the marketplace. Creating and selling news or other information goods online is too new a field and there are still too many unknowns for economic models or theory to explain or predict with much accuracy at all. Concrete answers will only come from jumping in and taking risks.

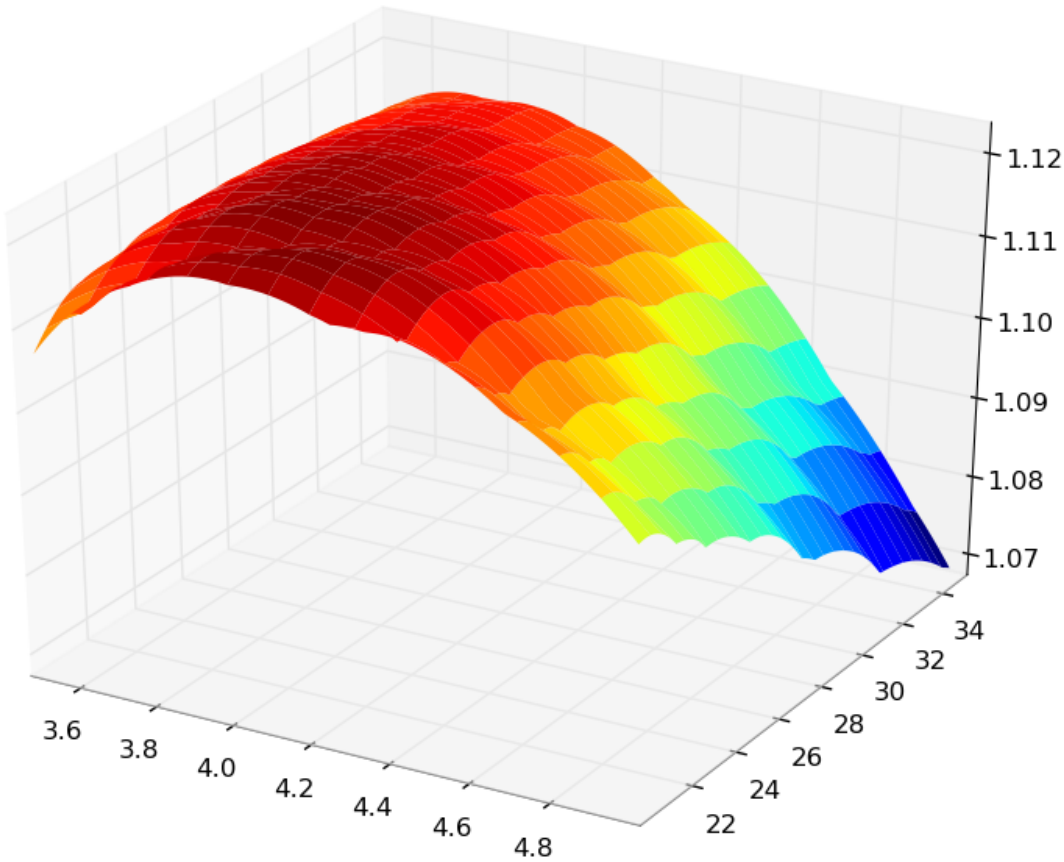



(a) $f$ from 0 to 25 and $F$ from 0 to 125 at intervals of 0.5 and 2




(b) $f$ from 0 to 10 and $F$ from 0 to 50 at intervals of 0.2

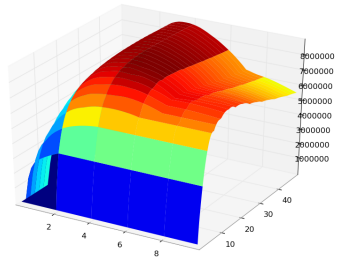
Figure 7: L1: First two graphs of $f$, $F$, and revenue ($)



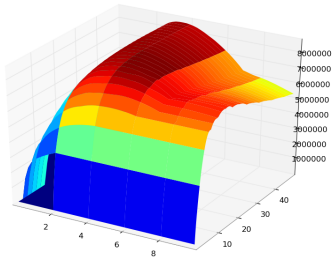

(c) $f$ from 3.50 to 5.00 and $F$ from 20 to 35 at intervals of 0.01 and 0.05

Figure 7: L1: Last graph of $f$, $F$, and revenue ($)

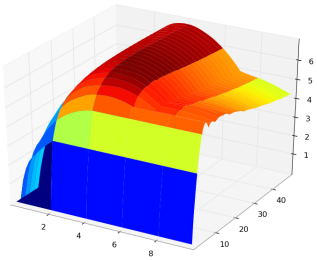
Figure 8: Other Sites: Graphs of $f$ from 0 to 10 $F$ from 0 to 50.



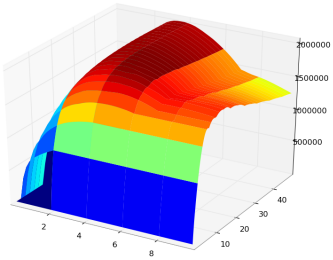
(a) L2



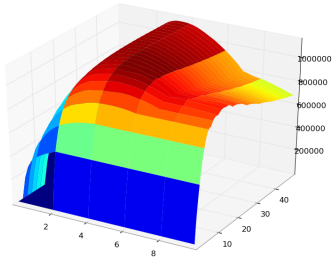
(b) L3



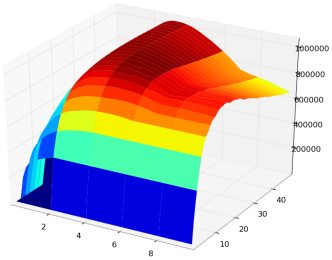
(c) A1



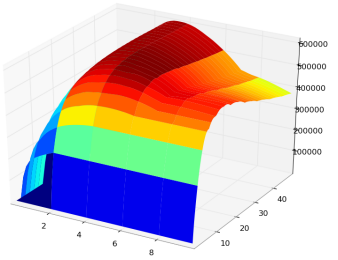
(d) M1



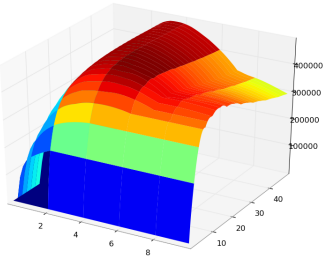
(e) M2



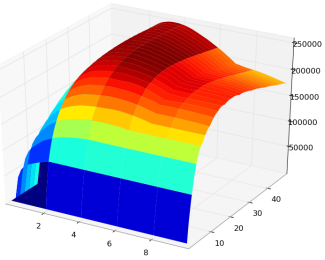
(f) M3



(g) S1

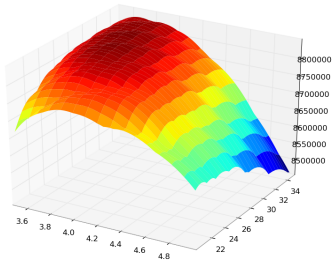


(h) S2

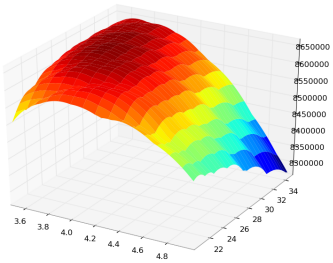


(i) S3

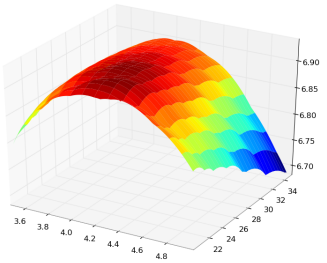
Figure 9: Other Sites: Graphs of $f$ from 3.50 to 5.00 and $F$ from 20 to 35.



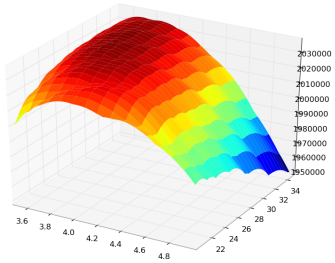

(a) L2



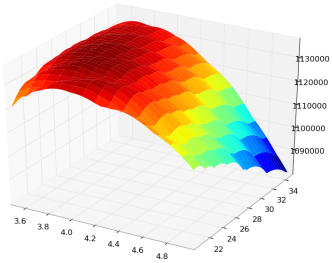

(b) L3



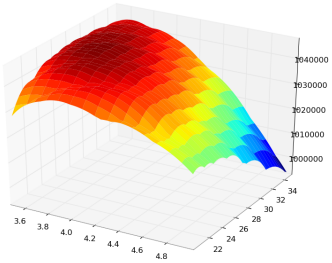

(c) A1



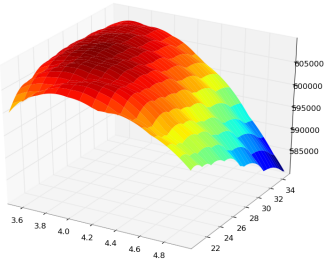

(d) M1



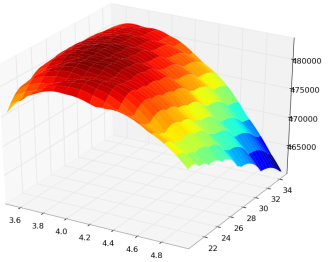

(e) M2



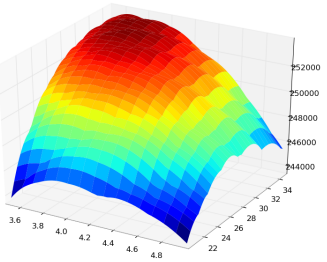

(f) M3




(g) S1




(h) S2




(i) S3

# A Source Code

Code Listing 1: Core calculation method

```
from pylab import *
from scipy import *
from scipy.interpolate import interp1d

from data_arrays import *
from model_v2_helpers import *
from segment import Segment
from results import Results

@print_timing
def calculate_points(name,site,wtp_cdf,
  f_start=0,F_start=0,f_range=10,F_range=30,f_step=.2,F_step=.2, format='list'):
    """
    The heart of the program.
    Calculates revenues for a site accross a range of f's and F's.
    """
    X = np.arange(f_start,f_start+f_range,f_step)
    Y = np.arange(F_start,F_start+F_range,F_step)
    if (format == 'array'):
        X, Y = np.meshgrid(X, Y)
        R = array([0.0]*X.size).reshape(X.shape) #total revenue
        P = array([0.0]*X.size).reshape(X.shape) #percentage individual income

        max = 0.0
        for i in xrange(len(X)):
            for j in xrange(len(X[0])):
                segments = build_segments(site, X[i][j], Y[i][j])
                ztup = integrate_stuff(segments,wtp_cdf)
                R[i][j] = ztup[0]
                try:
                    P[i][j] = 100.*float(ztup[1])/ztup[0]
                except ZeroDivisionError:
                    P[i][j] = 0.0
                if (ztup[0] > max):
                    max = ztup[0]
                    print "new max of $"+str(ztup[0])+" at f="+str(X[i][j])+" and
                      F="+str(Y[i][j])+", with ind%="+str(P[i][j])
        return {'X':X,'Y':Y,'Z':R,'P':P}
    elif (format == 'list'):
        r = Results(name,site,wtp_cdf,X,Y)
        print "+++++ calculating for "+str(r)+" ++++++"
        for x in X:
            for y in Y:
                segments = build_segments(site,x,y)
                z = integrate_stuff(segments,wtp_cdf)
                if (z[0] > r.max[2][0]):
                    r.max = (x,y,z)
                r.points.append((x,y,z))
        print "----max of $"+str(r.max[2])+" at f="+str(r.max[0])+" and F="+str(r.max[1])+"
          ----"
        return r

def build_segments(visitor_distribution,f,F):
    """Constructs a list of the 20 segments of a visitor distribution."""
    segments = []
    for v in visitor_distribution:
        s = Segment(v[1],v[0],f,F)
        segments.append(s)
    return segments
```

```python
def integrate_stuff(segments,cdf):
    rev = 0.0
    ind = 0.0
    sub = 0.0
    for s in segments:
        srev = s.integrate(cdf)
        rev += (srev[0]*s.visitors)
        ind += (srev[1]*s.visitors)
        sub += (srev[2]*s.visitors)
    return (rev,ind,sub)

"""A function that interpolates values from our WTP data."""
bcg = interp1d([0,1,4,7,11,15,1000],[.52,.69,.81,.92,.96,1.0,1.0])
```

Code Listing 2: Wrapper class for results

```python
class Results:
    """Wrapper class to hold results of calculate_points"""

    def __init__(self,name,site,cdf,f_range,F_range):
        self.name = name
        self.site = site
        self.cdf = cdf
        self.f_range = f_range
        self.F_range = F_range
        self.points = []
        self.max = (0.0,0.0,(0.0,0.0,0.0))

    def __repr__(self):
        return self.name

    def __getstate__(self):
        return {'name':self.name, 'points':self.points, 'max':self.max, 'site':self.site,
          'cdf':(self.cdf.x,self.cdf.y), 'f_range':self.f_range, 'F_range':self.F_range}

    def __setstate__(self,state):
        self.name = state['name']
        self.site = state['site']
        self.cdf = state['cdf']
        self.f_range = state['f_range']
        self.F_range = state['F_range']
        self.points = state['points']
        self.max = state['max']
```

Code Listing 3: A Segment of consumers

```
from pylab import *

class Segment:
    """
    A segment of consumers, defined at a pair of prices.
    Each segment calculates the range of WTP a consumer in that segment will make each
      purchase decision.
    """
    def __init__(self, N, visitors, f, F):
        assert(type(N) is int)
        assert(type(visitors) is int)

        self.N = N
        self.visitors = visitors
        self.f = float(f)
        self.F = float(F)
        self.steps = [(0,0)]
        for i in xrange(self.N):
            self.steps.append((self.N*self.f/(self.N-i),self.f*(i+1)))

        self.subscription_threshold = False
        if ((self.N*self.f) > self.F):
            self.subscription_threshold = self.s_threshold((0,self.N*self.f))

    def buys_subscription(self, wtp):
        """This just tests for any given wtp. Want a function that calculates lowest such
          wtp."""
        u_sub = self.utility_subscription(wtp, self.F)
        if (u_sub >= self.utility_individual(wtp, self.f)) and (u_sub > 0):
            return True
        else:
            return False

    def s_threshold(self, (start, end), acc=5):
        if (acc==0):
            return (start, end)
        else:
            X = arange(start,end,(end-start)/10.)
            prev = X[0]
            for x in X:
                if (self.buys_subscription(x)):
                    return self.s_threshold((prev,x), acc-1)
                else:
                    prev = x
            return self.s_threshold((prev,end), acc-1)

    def n_value(self, n, wtp):
        """Value of consuming some n articles"""
        val = 0.0
        for i in xrange(int(n)):
            val += float(wtp*(float(self.N-i)/self.N))
        return val

    def utility_subscription(self, wtp, F):
        """F: subscription price"""
        out = self.n_value(self.N, wtp) - F
        return out

    def num_individual_purchases(self, wtp, f):
        n = self.N-floor(float(f)/(wtp/float(self.N)))
        if (n<=0):
            return 0
        return n
```

```
def utility_individual(self, wtp, f):
    n = self.num_individual_purchases(wtp,f)
    out = self.n_value(n,wtp)-(float(n)*f)
    return out

def integrate(self,cdf):
    """
    Integrates over each step of purchase choice, weighted with the cdf function for wtp
    cdf: x->probability wtp is less than x. operates on array-like objects
    """
    if (self.subscription_threshold != False):
        ranges = [x for x in self.steps if x[0] < self.subscription_threshold[0]]
        ranges += [(self.subscription_threshold[0],self.F)]
    else:
        ranges = [x for x in self.steps]
    try:
        ranges = [(cdf(x[0]),x[1]) for x in ranges]
        ranges += [(1,)]
    except ValueError:
        print ranges
        print self.steps
        print self.subscription_threshold
        raise ValueError

    val = 0
    for i in xrange(len(ranges)-1):
        val += (float(ranges[i+1][0])-ranges[i][0])*float(ranges[i][1])
    if (self.subscription_threshold != False):
        """if a subscription is bought, return a tuple of three values. (total value,
          value from individual sales, value from subscription sales)
        subscription sales are the last segment"""
        sub_value = (float(ranges[i+1][0])-ranges[i][0])*float(ranges[i][1])
        return (val, val-sub_value, sub_value)
    else:
        return (val, val, 0.0)

def __repr__(self):
    return "Segment (depth:"+str(self.N)+", f:"+str(self.f)+", F:"+str(self.F)+")"
```

Code Listing 4: Helper functions

```
import time
import cPickle as pickle
from pylab import *
from mpl_toolkits.mplot3d import Axes3D
from matplotlib import cm

def show_graph(d,Z='Z'):
    """d a dictionary with X,Y,Z matrices"""
    fig=plt.figure()
    ax = Axes3D(fig)
    ax.plot_surface(d['X'],d['Y'],d[Z],cmap=cm.jet)
    plt.show()

def get_surface(results):
    X = array([x[0] for x in results.points])
    Y = array([x[1] for x in results.points])
    Z = array([x[2][0] for x in results.points])
    shape = (len(results.f_range),len(results.F_range))
    X = X.reshape(shape)
    Y = Y.reshape(shape)
    Z = Z.reshape(shape)
    return {'X':X,'Y':Y,'Z':Z}

def show_scatter(list):
    fig = plt.figure()
    ax = Axes3D(fig)
    xs = [x[0] for x in list]
    ys = [x[1] for x in list]
    zs = [x[2] for x in list]
    ax.scatter(xs,ys,zs)
    ax.set_xlabel('f')
    ax.set_ylabel('F')
    ax.set_zlabel('revenue')
    plt.show()

def print_timing(func):
    def wrapper(*arg, **kwargs):
        t1 = time.time()
        res = func(*arg, **kwargs)
        t2 = time.time()
        print '%s took %0.3f ms' % (func.func_name, (t2-t1)*1000.0)
        return res
    return wrapper

def save(object, filename):
    fout = open(filename,'wb')
    pickle.dump(object,fout)
    fout.close()

def load(filename):
    fin = open(filename,'rb')
    obj = pickle.load(fin)
    fin.close()
    return obj

def loadall(endname,base='pickles/',keys=['A1', 'L2', 'L3', 'M3', 'L1', 'S3', 'M1', 'S1',
  'M2', 'S2']):
    objs = []
    for k in keys:
        o = load(base+k+'/'+endname)
        objs.append(o)
    return objs
```

# References


William James Adams and Janet L Yellen. Commodity bundling and the burden of monopoly. *The Quarterly Journal of Economics*, 90(3):475–98, August 1976. URL http://ideas.repec.org/a/tpr/qjecon/v90y1976i3p475-98.html.

Fayez AlShehri and Barrie Gunter. Willingness to Pay for Online News: An Empirical Study on the Viability of the Subscription Model. *Aslib Proceedings*, 54(1):56–70, 2002. doi: 10.1108/00012530210426293. URL http://www.emeraldinsight.com/10.1108/00012530210426293.

Mark Armstrong. Multiproduct nonlinear pricing. *Econometrica*, 64(1):51–75, January 1996. URL http://ideas.repec.org/a/ecm/emetrp/v64y1996i1p51-75.html.

Yannis Bakos and Erik Brynjolfsson. Bundling Information Goods: Pricing, Profits and Efficiency. *SSRN eLibrary*, 1998. doi: 10.2139/ssrn.11488. URL http://ssrn.com/paper=11488.

Yannis Bakos and Erik Brynjolfsson. Bundling and Competition on the Internet. *SSRN eLibrary*, 1999. doi: 10.2139/ssrn.231598. URL http://ssrn.com/paper=231598.

Yannis Bakos and Erik Brynjolfsson. Aggregation and disaggregation of information goods: Implications for bundling, site licensing and micropayment systems. In Brian Kahin and Hal R. Varian, editors, *Internet Publishing and Beyond*, pages 114–137. The MIT Press, 2000.

The Boston Consulting Group. Willingness to pay for news online: Key findings from an international survey. Press release, November 2009. URL http://www.bcg.com/media/PressReleaseDetails.aspx?id=tcm:12-35297. Survey results distributed as press-release.

Alastair Bruce. Charging for content: How publishers are charging for online content or consumption and implementing paywalls and subscription services. Posted online, February 2010. URL http://www.slideshare.net/ajbruce/charging-for-content.

Suchan Chae. Bundling subscription tv channels : A case of natural bundling. *International Journal of Industrial Organization*, 10(2):213 – 230, 1992. ISSN 0167-7187. doi: 10.1016/0167-7187(92)90016-R. URL http://www.sciencedirect.com/science/article/B6V8P-45CX0N8-1B/2/537b5a66332964ba08bd5f2ede4dde40.

John Chung-I Chuang and Marvin A. Sirbu. Optimal bundling strategy for digital information goods: network delivery of articles and subscriptions. *Information Economics and Policy*, 11(2):147 – 176, 1999. ISSN 0167-6245. doi: DOI:10.1016/S0167-6245(99)00008-6. URL http://www.sciencedirect.com/science/article/B6V8J-3X11072-2/2/e28190889fd02b9be87f35e839107b85.

Hsiang Iris Chyi. Willingness to Pay for Online News: An Empirical Study on the Viability of the Subscription Model. *Journal of Media Economics*, 18(2):131–142, April 2005. doi: 10.1207/s15327736me1802_4. URL http://www.informaworld.com/10.1207/s15327736me1802_4.

J. Bradford DeLong and Michael A. Froomkin. Speculative microeconomics for tomorrow's economy. In Brian Kahin and Hal R. Varian, editors, *Internet Publishing and Beyond*, pages 6–44. The MIT Press, 2000.

Thomas Eisenmann, Geoffrey Parker, and Marhall W. Van Alstyne. Strategies for two-sided markets. *Harvard Business Review*, October 2006.

Peter C. Fishburn, Andrew M. Odlyzko, and Ryan C. Siders. Fixed-fee versus unit pricing for information goods: Competition, equilibria, and price wars. In Brian Kahin and Hal R. Varian, editors, *Internet Publishing and Beyond*, pages 167–189. The MIT Press, 2000.

Harris Interactive. Just two in five americans read a newspaper almost every day. Press release, January 2010. URL http://news.harrisinteractive.com/profiles/investor/ResLibraryView.asp?ResLibraryID=35425&GoTopage=5&Category=1777&BzID=1963&t=30. Survey results distributed as press-release.

Jack Herbert and Neil J. Thurman. Paid content strategies for news websites: An empirical study of british newspapers' online business models. *Journalism Practice*, 1(2), 2007.

ITZBelden. Revenue initiatives 2009. Presented at Newsmedia Economic Action Plan Conference, Sept. 13-15, 2009, September 2009. URL http://americanpressinstitute.org/09/Newsmedia/.

Donald W. King and Jose-Marie Griffiths. Economic issues concerning electronic publishing and distribution of scholarly articles. *Library Trends*, 43(4):713 – 740, 1995.

Regina Lewis. Relation between newspaper subscription price and circulation, 1971-1992. *Journal of Media Economics*, 8(1):25–41, 1995. ISSN 08997764. URL http://www.informaworld.com/10.1207/s15327736me0801_3.

Susan M. Mings and Peter B. White. Profiting from online news: The search for viable business models. In Brian Kahin and Hal R. Varian, editors, *Internet Publishing and Beyond*, pages 62–96. The MIT Press, 2000.

Richard Schmalensee. Gaussian demand and commodity bundling. *Journal of Business*, 57(1):S211–30, January 1984. URL http://ideas.repec.org/a/ucp/jnlbus/v57y1984i1ps211-30.html.

Jonathan Stray. Play paywall!, the new web game sweeping the newspaper industry. Nieman Journalism Lab, January 2010. URL http://www.niemanlab.org/2010/01/play-paywall-the-new-web-game-sweeping-the-newspaper-industry/. Blog post on www.niemanlab.org.

Hal R. Varian. Pricing information goods. In *Scholarship in the New Information Environment*. Research Libraries Group Symposium, May 1995.

Hal R. Varian. Versioning information goods. In Brian Kahin and Hal R. Varian, editors, *Internet Publishing and Beyond*, pages 190–202. The MIT Press, 2000.