Website Ad Quantities:
An Empirical Analysis of Traffic, Competition, and Business Model

Laura J. Kornish
Leeds School of Business, University of Colorado at Boulder
kornish@colorado.edu

Jameson Watts
Eller College of Management, University of Arizona
jamesonw@email.arizona.edu

April 24, 2012

Abstract

In running a website, a firm balances two potential streams of revenue: sales of goods, services, or information content to visitors; and sales of advertising space to other organizations. Web-based businesses thus operate in "two-sided markets," selling something of value to visitors and selling visitors' attention, and advertising choices affect both sides of those markets. In this paper, we empirically investigate the determinants of ad quantity on websites. Recent theoretical literature on this topic, especially the work of Katona and Sarvary (2008) [KS]; Godes, Ofek, and Sarvary (2009) [GOS]; and Kind, Nilssen, and Sorgard (2009) [KNS], has examined the relationship of site traffic, competition, and ad quantity. However, those papers yield some contradictory predictions. We focus on three issues. Do sites that have more competition devote more space to advertising? GOS say no and KNS say perhaps yes. Do sites that have more traffic devote more space to advertising? KS say no, KNS say yes, and GOS say it depends. And how do the answers to those questions change depending on the business model of the site, i.e., whether it is an advertising-only business model vs. a hybrid of the two streams? We study these questions with a large sample of websites taken from the top sites as ranked by quantcast.com.

Keywords: Internet advertising; media economics

Acknowledgments: We would like to thank Avi Goldfarb, Ken Koput, Gary McClelland, and Ken Wilbur for their help.

1. Introduction

Millions of new websites are launched each year. Many of these sites hope to make money with an advertising business model: launch the site, create good content, and sell ads that bring in revenue based on visitors. However, there are questions about how to best run such a business. Should a site use advertising sparingly, thereby sacrificing advertising revenue but possibly increasing the attractiveness—and hopefully the traffic—of the site? Or should the site use advertising full bore, hoping that consumers aren't too distracted or bothered by the ads?

Marketing scholars have developed analytical models to study the question of the optimal level of advertising, and the answer to the question of how much advertising is, of course, "it depends." Two recent papers look at one factor it depends on, namely, competition, and the papers have somewhat conflicting predictions. Godes et al. (2009) [GOS] argue that in a competitive setting, a site can ill-afford to bother visitors with too many ads. Conversely, Kind et al. (2009) [KNS] argue that more competitive categories need to rely more on ads because of cutthroat competition in content sales. The literature has also considered another predictor of ad quantity: site traffic. Again, the results in different papers don't agree. Katona and Sarvary (2008) [KS] find that more popular (higher traffic) sites restrict advertising displays, giving visitors fewer opportunities to click away from the site. Conversely, KNS find that competition leads to "favorable terms" for consumers, which results in both more advertising and more traffic.

In this paper, we analyze the empirical evidence related to these competing theories. We examine the effects of competition on ad quantity and the correlational relationship between traffic and ad quantity. We study how those effects are different for sites in general compared to sites that have an advertising-only business model. This analysis provides a snapshot of how the tension in selecting ad quantity is being resolved in the market today. Further, it can inform decisions in practice about how to make trade-offs inherent in selecting ad quantity.

We focus on competition and site traffic, given their prominence in the literature and their amenability to measurement. We also discuss another factor that plays a role in determining ad quantity, the nuisance cost or information value associated with ads, even though it is hard to measure that directly.

We develop a novel data set to study these questions. We use an importance sample from a list of top websites, ranked by traffic by web metrics company Quantcast. We take measures of ad quantity with a customized spider that crawls the home pages of sites in the sample and looks for calls to ad servers, noting the total size of ads on the page. Consistent with the recently published theoretical models, we develop multiple measures of competition, including number of competitors and a concentration index. We also use measures of an individual site's competitiveness: market share based on traffic and a measure of competitive intensity in the ad market. To compute the number, concentration, and share, we identify a site's category using a large web directory, dmoz. To determine concentration and market share, we use estimates of site traffic from Quantcast. For competitive intensity in the ad market, we use Google's AdWords keyword competition metrics. For most of our measures, we have four observations, each spaced two months apart.

Our empirical findings are as follows. First, we see that more intense competition is associated with fewer ads. This finding supports the logic in the literature that ads degrade a visitor's experience, and more competition forces companies to offer a better experience. It contradicts the logic that asserts that competition drives sites toward ads, away from content. Second, we see that higher-traffic sites devote more space to advertising than lower-traffic ones. This result holds for both a general sample of sites and for advertising-only sites specifically. Less popular sites are more conservative with their advertising space than those more popular, which contradicts existing theory that predicts that lower traffic sites sell more ads than higher traffic ones.

2. Theoretical framework

There is a vast body of work about advertising in the marketing and economics literatures (see Bagwell, 2007). Our focus is on the media carrying the advertising, concentrating on the question of how much advertising they display (i.e., sell). Our work is therefore related to work about media markets (Dukes 2004, 2006; Gal-Or and Dukes 2003; Godes et al. 2009; Kaiser and Wright 2006; Katona and Sarvary 2008; Kind et al. 2009; Rochet and Tirole 2006; Wilbur 2008), two-sided markets that sell two things, content to viewers and advertising space to advertisers. Synthesizing themes from that literature, we examine the determinants of ad quantity on websites, focusing on the roles of competition, site traffic, and business model.

2.1 Competition

Two recent papers address the relationship between the level of competition and quantity of advertising on websites. GOS capture three notions of competition: the number of competitors (monopoly vs. duopoly), the "competitive intensity in the content market" (denoted □), and the "competitive intensity in the ad market" (denoted h). KNS examine two dimensions of competition: the closeness of site substitutability (essentially the degree of differentiation, denoted s, which can be seen as similar to GOS's □ and/or h) and second, the number of sites that could contain similar advertising (denoted m, related to GOS's consideration of monopoly vs. duopoly).

The results from the two papers are somewhat at odds. GOS describe competitive mechanisms that either lower ad quantity or leave it unaffected. First, they find that the rivalry of duopoly "causes each media firm to lower the number of ads it runs" (p. 25) to offer a better deal to consumers. Second, they find that optimal ad quantity is unaffected by competitive intensity in the content market (p. 27). Third, they find that higher competitive intensity in the ad market leads to lower ad quantity on each site. In contrast, KNS describes a competitive mechanism which raises ad quantity. They find that the less differentiated the media products are (i.e., the closer the competition), the more ads a site runs. (See Lemma 3 on p. 1118.) This reduction occurs because with less differentiation, the less able sites are to sell their content, and the more they resort to advertising revenues. The paper does not contain a prediction about the relationship between number of competitors and ad quantity. However, it does show that advertising revenues for each site decrease with the number of competitive sites.

Both papers treat the multiple dimensions of competition as distinct to understand the separate effect of each dimension. In reality, we would expect that competitive dimensions are related. For example, the more competitive sites there are (m in KNS), the closer the substitutes are likely to be (s in KNS).

Summarizing the above discussion, the theory supports the following hypotheses:

H1: Sites that face more competitors have less advertising.

H2a: Sites that face greater competitive intensity have more advertising.
H2b: Sites that face greater competitive intensity have less advertising.
H2c: Competitive intensity does not affect ad quantity.

The analyses in KNS and GOS represent what KS call the "commercial web"—sites that have *at least some* non-advertising revenue sources. In KS, this non-advertising revenue source is called "content" and is modeled as the "income from consumers," e.g., sales of products or services from the website.[1] Consistent with KS's notion of content value, both KNS and GOS focus on interior solutions when solving for the quantity of advertising and the quantity of content sales, meaning that there is activity on both sides of the traditional two-sided media markets. We note that the KS definition of commercial web does not encompass all web activity, or even all activity that could be deemed as "commercial." Sites that use an advertising-only business model can also be commercial, but operating on only one side of the media market.

In an appendix, GOS also study these advertising-only sites. Setting content price to zero is an empirically common corner solution in the optimization of ad and content revenues. Looking at those advertising-only sites, GOS show that the effect of competition on ad quantity is no longer a straightforward comparison. Whether competition raises or lowers ad quantity depends in a complex way on the model parameters, including strength of competition, value of the site, effectiveness of advertising, etc. The logic determining ad quantity is clearly different in this zero-price-content case.

2.2 Traffic

The recent literature also has much to say, again with little agreement, about the relationship between site traffic and ad quantity. The two studies discussed above, KNS and GOS, which

[1] Interpreted more broadly, the "income" to the firm could be other value from consumers, for example, increased brand awareness that leads to sales through another channel. That interpretation would be like the firm using the website space as advertising for its own products. A key feature of the KS commercial web is that there is some value to the firm from visitors, beyond advertising sales.

focus on competition, also investigate the relationship between site traffic[2] and ad quantity. In addition, KS examine the relationship between site traffic and ad quantity explicitly. All three papers include predictions about ad quantity on sites that have at least some non-advertising revenue sources (KS's commercial web). The three predictions are all different. KS find a negative relationship between traffic and ad quantity, i.e., lower traffic sites feature more ads than higher traffic sites. GOS find no relationship. KNS find a positive relationship.

In the GOS appendix in which content prices are constrained to zero, the logic changes again. For those sites, they find a positive relationship between site quality and ad quantity, implying that higher traffic sites (i.e., the higher quality ones in their model) have greater ad quantity.

KS analyze a network model. In their model, sites buy and sell advertising links to one another. They derive an equilibrium in which the "higher content" sites sell fewer ads and buy more ads than "lower content" sites. (Recall that the content level of a site is modeled as the content income per visitor to a site.) In equilibrium, higher content sites have more traffic. One implication is that higher traffic sites will have (sell) fewer ads than lower traffic sites. They show that their result holds even with higher advertising rates on the higher traffic sites in the model. Higher content sites have more to lose by visitors navigating away from the site by clicking on an ad, so they sell fewer of them.

GOS and KNS both consider "media" firms, which provide content and have the ability to sell ads. The setting applies to both web-based and traditional media. In contrast to KS, GOS find that there is no relationship between the quantity of ads and traffic (or content sales volume). The null relationship is the result of two opposite forces. In their model, the higher volume sites are higher "quality." And the higher the quality, the higher the price the site can charge for content and the higher the demand for that content. Because the content price is endogenous, it effectively insulates the decision about ad quantity from the quality (and therefore the traffic) level.

[2] Both GOS and KNS use a "representative consumer" model of demand for content. KNS describe how the content sales volume (C_i in their model) can represent both "time that each viewer spends watching channel *i* and … the number of viewers of channel *i*" (p. 1114). GOS use the language of "units of content" rather than "viewers" or "visitors."

In the GOS appendix that restricts content prices to zero, they find that higher quality sites will have more ads than lower quality sites. With a content price constrained to zero, the price is no longer endogenous, and higher quality sites, which also have more traffic, optimally increase revenues by selling more ads.

KNS do not impose the zero-price-on-content restriction, and yet their prediction is also for a positive relationship between traffic and advertising (see p. 1113, middle of column 2). KNS explain that media firms that have a larger emphasis on advertising revenues compared to content sales will have relatively more traffic: the ad business model offers a low price to consumers, so ad-business-model firms will have higher traffic.

Summarizing the above discussion, the theory supports the following hypotheses:

H3a: Sites that have more traffic have more advertising.

H3b: Sites that have more traffic have less advertising.

H3c: There is no relationship between site traffic and ad quantity.

For advertising-only sites, we test the GOS prediction.

H4: For advertising-only sites, higher traffic sites have more advertising.

We cannot observe all the elements of the models of KS, GOS, and KNS. For example, we cannot observe the profit per visitor for a site nor the prices for individual ads on different sites. However, there are reasonable estimates available for site traffic (like r in KS), and it is challenging yet possible to gather a measure of ad quantity displayed on a site (like d_{out} in KS).

2.3 Another issue: disutility from ads

We have discussed one corner solution, in which the site sets the content price to zero and runs an advertising-only business model. Another corner solution is possible: the site runs no ads. This type of corner solution is actually highly prevalent.

GOS, KNS, and KS do not explicitly discuss the no-ads solution. However, all three papers include an element that can explain it: consumers' disutility from ads.[3] If the disutility of ads were high enough, that would lead sites to optimally have (sell) no ads. While not addressed in the papers, sites could be heterogeneous in ad disutility, resulting in the no-ads solution for some sites.

We mention disutility of ads because it is a plausible explanation to a feature in the data set, the prevalence of sites with no ads. However, we do not focus on it in our analysis for two reasons. First, we did not see conflicting theory in the literature on this point. Second, because heterogeneous disutility for ads is difficult to measure.

3. Overview of Data

Our data are derived from a list of the top two-hundred-thousand websites, as determined by traffic, based on estimates from the web metrics company Quantcast. Using that population, we drew an importance sample, with probability of selection proportional to traffic, of ten thousand sites. For a subset of the sites in the samples, we have the site category, measures of ad quantity, measures of competition, and a demographic characterization of visitors. We provide the specifics of these elements below. We collected data at four points in time: August, October, and December 2011 and February 2012. The multiple observations help us control for unobserved heterogeneity in the sites.

3.1 Traffic

Quantcast ranks the top million sites by traffic and lists them at quantcast.com (Quantcast, 2012). We built a customized spider to crawl the top 20% of that list. The spider took about five days to run, which is why we extracted only part of the list. The list includes the site web address (URL), traffic rank, and an estimate of traffic in terms of U.S. visitors per month. The top-two-hundred-thousand list and the associated traffic were gathered at all data collection points.

[3] In GOS and KNS, disutility from ads is an essential part of the demand model; in KS, disutility from ads is treated in an appendix.

Although we have *ranks* for the top two hundred thousand sites, Quantcast doesn't display a *traffic estimate* for every single one of them; they listed estimates for 81.6% of the sites. For sites missing traffic estimates, we interpolate the traffic level based on rank.

3.2 Site categories

We used dmoz (http://www.dmoz.org/), an Internet directory (the "open directory project"), to categorize the sites in the top two-hundred thousand. Dmoz uses human raters to classify sites into categories and subcategories. There are fifteen top level categories, including arts, business, computers, reference, sports, shopping and news. There are a few hundred subcategories at the next level of the hierarchy, and thousands at the lower levels. Being human-rated, dmoz is far from a comprehensive directory of Internet sites, but it does have a large database. We built a customized spider to look up all the sites on our initial top-two-hundred-thousand list. Of those, 47,170 sites--just under one-fourth--were present in dmoz.

Sites can be listed in multiple categories in dmoz. Many sites have only a single classification, but some have hundreds. To keep the data collection time reasonable (under a week), we examined up to the first 300 categories listed for each site. We considered only the categories for the root domain, e.g., www.site.com or site.com, and not subdomain.site.com or www.site.com/subdirectory/. We also removed the top level categories of World and Regional because those categories often refer to the geographic location of the company headquarters rather than the category of business. If there were still multiple categories (e.g., reuters.com is listed under both Business/News_and_Media and News/Media), we randomly selected a category. For each site, we used the top and second level categories.

We collected the site categories at the time of the first observation. We did not collect it again because the category isn't a characteristic that fluctuates, the way a traffic or ad measurement would. We collected the categories for the whole top 200,000 list, not just the importance sample, because we use this information to derive some of our measures of competition (described below).

3.3 Ad quantity

We created a customized crawler to collect a measure of ad quantity for the home pages of the sites in our importance sample. We measure ad quantity by finding the total count of pixels devoted to ads on the home page. To make the counts for so many sites, our crawler looks for calls to a list of approximately 2,700 ad servers. We compiled the list by starting with a publicly available list of ad servers and adding ones we discovered as the spider crawled the pages. In this data collection, we distinguish between “inbound” and “outbound” ads. Inbound ads link to another part of the given site; outbound ads link to a different site. Our analysis is based on outbound ads. The pixel count was calculated from the dimensions specified for each displayed ad. We tallied both text ads and display ads; including static banners, those with rich media, those that include Flash and those that are videos embedded in Flash. We did not include pop-ups, except for pop-ups that collapse to a display ad. Expandable ads counted as the size they appear when the site loads. Approximately 19% of the sites in the sample show ads.

We also attempted to collect another measure, the fraction of the page covered in ads. However, we were unable to reliably measure the “size” of the page itself, so we could not use that measure. Our measure of ad quantity was gathered at all data collection points.

3.4 Competition

Consistent with the theoretical work we draw on, we collected multiple measures of competition for each of the sites in the importance sample.

First, we collected the number of competitors, as defined by the number of other sites in the dmoz category at the second level of the tree (top-level category plus subcategory) that also appear in the list of top two-hundred thousand sites.

Second, we use a measure of competitive intensity in the content market (like the □ parameter in GOS and partially like the *s* parameter in KNS). The measure of competitive intensity is a “concentration” measure, the Herfindahl–Hirschman Index (HHI), based on the sites in the category and their traffic level. The HHI is the sum of the squared share of the sites. It is a measure between 0 and 1: an HHI of 1 indicates no competition, a perfect monopoly; HHI close to 0 indicates great competition—a huge number of competitors and equally split shares.

Third, we use a measure of market share within a category for each site, also based on the dmoz categories and traffic. We measure market share by calculating the fraction of traffic in the category that a given site has. Our first two measures, number of competitors and competitive intensity, both characterize the entire category. In other words, for a given category, there is a single value of each measure. All the sites in the category are treated symmetrically, with regards to these measures. In KNS and GOS, indeed, the competitive sites are modeled as homogeneous, and they ask questions about what happens when the general level of competition increases. We recognize, however, that sites are not homogeneous, and our market share measure captures this aspect of competitive heterogeneity. We consider market share as a site-level measure of competition: the greater the market share, the better the competitive position.

Fourth and finally, we use a measure of competitive intensity in the ad market (like the parameter *h* in GOS). We used a competition metric from Google, part of a Keyword Suggestion utility for AdWords, their search engine advertising program. Keyword Suggestion, suggests, as the name implies, relevant words and phrases for a site, based on an examination of the site content. For each keyword, there is an estimated measure of competition for the phrase, a normalized metric between 0 and 1. Appendix A contains an example of keywords and competition metrics for the site petfinder.com. For each site, we measure competition by using the average competition score for the top suggested keywords. If there were more than ten suggested words, we just used the top ten. Google provides an application programming interface (API) to access the data, which can be used for a small fee per query. We were able to get this measure of competition for the vast majority (99.7%) of sites in our sample. We hoped to collect this data at each of the data collection points, but Google did not renew our license to use the API after the first collection, so we have only a single observation of this element of data.

Table 1 shows correlations as well as summary statistics for our variables.

Table 1: Descriptive Statistics of the Data Samples

	Ln(Ad Pixels)	Ln(Traffic)	Ln(Market Share)	Ln(HHI)	# Competitors	Comp Intensity in Ads
Ln(Ad Pixels)	1					
Ln(Traffic)	0.1397*	1				
Ln(Market Share)	0.0920*	0.7088*	1			
Ln(HHI)	0.0609*	0.1563*	0.4388*	1		
# Competitors	-0.0569*	-0.0526*	-0.4827*	-0.7631*	1	
Comp Intensity in Ads	-0.1024*	0.0882*	-0.0088	-0.0013	0.0111	1
Mean	2.09	12.41	-5.13	-2.97	644.79	0.462
Standard Deviation	4.43	1.99	2.24	1.14	665.97	0.264

4. Analysis

Our overall goal in this research is to understand the relationship between traffic, competition, and ad quantity (our dependent variable), and to understand how the relationship changes based on the site's business model. In our analyses, we take steps to control for unobserved heterogeneity of the sites and to tackle questions of causality.

4.1 Model

Our initial model uses a linear form relating ad quantity and the characteristics of the sites. This linear model can be solved with generalized least squares (GLS), specifically, we use random effects GLS with robust errors. We also report results from the Tobit model, which more explicitly accounts for the sites with no ads. Our initial model includes both time-invariant and time-varying components:

$$\mathrm{AQ}_{it} = \alpha + \tau_t + \beta_1 \mathrm{Tr}_{it} + \beta_2 \mathrm{K}_{it} + \beta_3 \Lambda_i + \mu_i + \varepsilon_{it} \qquad (1)$$

The dependent variable AQ is the observation of ad quantity, and the subscripts i and t indicate site i at time t. We measure ad quantity with ad pixels. The distribution on ad pixels spans many orders of magnitude, and shows a skewed shape, naturally bounded below by zero and with a long

right tail. To handle that shape of distribution, we use the natural log of one plus the pixel count. We add the one to handle the zero-pixel case. We don't want to exclude the zero-pixel sites from the analysis, as the same considerations drive both the choice of whether or not to sell ads and the quantity if ads are sold.
The α is a constant, the τ is a control for which observation period (out of four), the Tr is traffic, and the K is competition (HHI and market share, which change each period due to new traffic estimates). The Λ captures "observed" heterogeneity, i.e., observations that differ by site but not by period. With our data, the Λ captures the AdWords competition measure, which we were only able to gather in one period, and the number of competitors.

The μ_i is the error component which captures the unobserved heterogeneity. This is a random-effects model: the μ_i is random, but it applies to each site i in the same way in each time period. Finally, ε_{it} is a standard error term for each time period. Both μ_i and ε_{it} are assumed normal and independently distributed. Further, we assume that the unobserved site-stable heterogeneity, the μ_i, is independent from the explanatory variables (Greene 2003). We believe this is a reasonable assumption: if there are unobserved characteristics driving ad quantity, they are unlikely to be the same characteristics across such a diverse set of sites. Moreover, we test this correlation post-estimation.[4]

The model in Eq. (1) is structured to analyze the correlational relationship between ad quantity and competition and/or traffic. What about causation? Edwards and Bagozzi (2000) list the four conditions commonly used to argue for causation in the social sciences: that the cause and effect are distinct entities, that the cause and effect co-vary, that the cause precedes the effect temporally, and that the rival explanations are eliminated. Here we describe our approach to analyzing causation.

First, we look to the theoretical papers for guidance. In KNS and GOS, there are clearly causal mechanisms at play with competition: they derive effects of changes in (exogenous) parameters

[4] For the models without a lagged dependent variable all predictors have correlations under 0.025 with μ_i. Not surprisingly, for models that include a lagged dependent variable, correlations between predictors and μ_i are virtually indistinguishable from 0. Further, we see that in the lagged versions, σ_μ=0, suggesting that none of the variance in ad quantity can be explained by intraclass correlation (Greene 2003).

representing competition on ad quantity (as we summarize in Section 2). The papers contain analytical models, and they are silent about the possibility that there may be reverse causation, i.e., the level of ad quantity could affect the level of competition. Empirically, we can and do consider that reverse causation. In contrast to competition, the theoretical predictions about traffic are *not* based on causal arguments. KS derive a network equilibrium: in the equilibrium, higher traffic sites have lower ad quantity, but KS do not argue that higher traffic "causes" lower ad quantity. Traffic and ad quantity are endogenous and jointly determined. Similarly, in the models of KNS and GOS, both traffic and ad quantity are endogenous; there is no implication that one causes the other.

Second, we look at lagged relationships in the data to help make the case for causation. If more intense competition precedes higher ad quantity (and not vice versa) that is suggestive that competition leads to changes in ad quantity (and not vice versa). We look at the effect of one of the variables, lagged, on the other (e.g., competition lagged on ad quantity) and then the reverse. If only one direction is significant, that supports the view that one variable causes another. If both directions are significant, the interpretation is that there is a feedback loop between the variables. Finally, if neither is significant, that does not support causation in either direction. Thus, we use variations on Eq. (1) that include lagged terms and reverse the dependent and independent variables.

4.2 Results on Competition

To test the hypotheses related to competition, we perform analyses relating our measures of competition to our measure of advertising quantity. Recall that H1 posits more competitors, less advertising; and H2a, b, and c, posit greater competitive intensity has more, less, and no effect, respectively, on advertising.

The results shown in Table 2 use our four different measures of competition. We run concentration (HHI), market share, and number of competitors in three different models because the measures of competition are moderately to highly correlated. (See Table 1.) The theory in GOS and KNS looks at the effect of competition on a market: competition is a market-level (or, in our case, a category-level) variable that is felt symmetrically by all firms. For example, in KNS, a higher *s* parameter means that all firms face a greater degree of substitutability of their product with others. In our analysis, we have two of those category level measures, but we also have two site-specific measures. Site specific measures speak to the site's level of competitiveness or market power relative to the other sites. This type of effect is not explicitly in the analytical models, but the logic is similar.

Table 2: Regression results using natural log of ad pixels as the dependent variable and various measures of competition as the independent variables.

VARIABLES	(1)	(2)	(3)
Ln(HHI)	0.212***		
	(0.0558)		
Ln(Market Share)		0.164***	
		(0.0282)	
# Competitors			-0.128**
			(0.0531)
Comp Intensity Ads	-1.688***	-1.704***	-1.650***
	(0.263)	(0.263)	(0.264)
Period	-0.158***	-0.140***	-0.147***
	(0.0231)	(0.0240)	(0.0230)
Constant	3.838***	4.040***	3.904***
	(0.244)	(0.225)	(0.346)
Observations	11,571	11,113	11,571
N	2,928	2,928	2,928
R^2	0.0158	0.0204	0.0135

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In aggregate, these results all support H1 and H2b, the two hypotheses that suggest that ad quantity and competition are negatively related. That negative relationship is consistent with the GOS logic that the more competition there is, the more pressure there is for the firms to offer a "better" experience to visitors, namely, with fewer ads.

Model (3) directly tests Hypothesis 1, and we find support for it. The coefficient on number of competitors is negative and significant: the more competitors, the lower the ad quantity.

Models (1) - (3) all contain tests of the three variations of Hypothesis 2, using different three measures of competition. Model (1) shows significant effects of both concentration (HHI) and competitive intensity in the ad market. The coefficient on HHI is positive and significant: the greater the concentration, the lower the competitive intensity, and larger the quantity of ads. The coefficient on competitive intensity in the ad market is highly significant and negative. The greater the competitive intensity, the lower the ad quantity. Model (2) shows the same negative, significant effect. In addition, it shows a positive and significant effect of market share. Sites with a stronger competitive position (as measured by share of traffic within the dmoz category) have more advertising. Model (3) again shows the same effect of competitive intensity in the ad market, and the support for H1 already mentioned.

Turning our attention to the temporal sequence of events to understand causal possibilities, we show the two lagged versions of the regressions.

Table 3: Regression results using natural log of ad pixels as the dependent variable and lagged measures of competition and other variables as the independent variables.

VARIABLES	(1)	(2)	(3)
Lagged Ln(Ad Pixels)	0.651***	0.646***	0.652***
	(0.0138)	(0.0140)	(0.0138)
Lagged Ln(HHI)	0.0904***		
	(0.0266)		
Lagged Ln(Market Share)		0.0562***	
		(0.0141)	
# Competitors			-0.000108***
			(3.80e-05)
Comp Intensity Ads	-0.523***	-0.541***	-0.519***
	(0.117)	(0.121)	(0.117)
Period	0.0773**	0.0953**	0.0821**
	(0.0392)	(0.0405)	(0.0392)
Constant	0.866***	0.866***	0.647***
	(0.167)	(0.166)	(0.140)
Observations	8,605	8,341	8,605
N	2,927	2,927	2,927
R^2	0.452	0.447	0.452

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The coefficients on the lagged competition metrics in Table 3—HHI in Model (1), market share in Model (2), and number of competitors in Model (3) —show the exact same patterns as in the non-lagged versions in Table 2. Competitive intensity in the ad market continues to be negative and significant in all the models.

Perhaps not surprisingly, lagged pixels predicts pixels: if a site had more ads in the last period, we predict more in the next.

Reversing the temporal sequence, we regress the measures of competition on lagged ad quantity. Although the theory in GOS and KNS doesn't address the possibility of reverse causation, i.e., that ad quantity affects competition, that direction of causation would also explain the pattern in Table 2. Table 4 shows the results from the appropriate lagged analysis.

Table 4: Regression results using competition measures as the dependent variable and lagged measures of ad quantity and other variables as the independent variables.

VARIABLES	(1) HHI	(2) Market Share
Lagged Ln(Ad Pixels)	0.000106	-0.000499
	(0.000379)	(0.00105)
Lagged DV	0.990***	0.985***
	(0.00154)	(0.00273)
Comp Intensity Ads	-0.0199***	0.0148
	(0.00599)	(0.0178)
Period	-0.00259	-0.00797
	(0.00222)	(0.00636)
Constant	0.0388***	-0.0914***
	(0.00898)	(0.0256)
Observations	8,734	8,170
N	2,928	2,852
R^2	0.978	0.953

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The lagged ad quantity is not a significant predictor of competition, for either concentration (HHI) or market share. Again, not surprisingly, the lagged DVs *are* significant: the past level of competition predicts the future level.

Our overall interpretation of the two lagged models is that the results support the direction studied by KNS and GOS (what is the effect of competition on ad quantity?), rather than the reverse direction (what is the effect of ad quantity on competition?). Further, we see support for H1, as predicted by GOS (and not inconsistent with KNS), and for H2b, as predicted by GOS (and somewhat at odds with KNS).

Finally, we also examined the relationship between competition and ad quantity in advertising-only sites by using a subsample comprised of sites in the News category in dmoz. (Of course some News sites have content for sale, but the News category is known for its reliance on advertising revenues.) None of the independent variables are significant in the regression, consistent with the GOS finding that competition and ad quantity are no longer simply related under advertising-only business models.

4.3 Results on Traffic

Following the theory, our hypotheses on traffic are split between sites in general for the versions of H3 and advertising-only sites for H4. The results for H3 are shown in Table 5.

Table 5: Regression results using natural log of ad pixels as the dependent variable and traffic as an independent variable.

VARIABLES	(1) Contemporaneous	(2) Lagged
Ln(Traffic)	0.211***	
	(0.0154)	
Lagged Ln(Ad Pixels)		0.659***
		(0.00913)
Lagged Ln(Traffic)		0.0819***
		(0.00831)
Period	-0.104***	0.0225
	(0.0117)	(0.0193)
Constant	-0.827***	-0.653***
	(0.177)	(0.109)
Observations	34,161	26,016
N	9,501	9,474
R^2	0.0182	0.474

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

This analysis supports H3a, that higher traffic sites have more advertising, both Models (1) and (2), the contemporaneous and the lagged versions of the model, respectively. This result is consistent with the prediction from KNS and is the opposite of the prediction from KS.

We also analyzed the reverse lagged version: traffic on lagged pixels, which is also significant and positive. (We omit the table of coefficients.) In this analysis, we are not surprised to see a bi-directional effect. None of the three theory papers argue for causation (i.e., they do not say that a traffic level determines an ad level), but rather traffic and ad quantity are jointly endogenous outcomes.

We further note that the *positive* relationship between lagged pixels and traffic does not suggest a directly causal story. Higher ad quantity results in more traffic? That seems unlikely. However, the KNS model helps explain the missing pieces needed to understand that result. KNS explain that when sites give consumers a better deal on the content (e.g., higher quality content and/or a lower price for it), the sites may raise the ad quantity in the bargain. It is the higher quality that drives both the higher ad quantity and the higher traffic.

For advertising-only sites, we test the GOS prediction with a similar analysis as in Table 5, except using only the News sites from the dmoz categorization. The results are shown in Table 6.

Table 6: Regression results for sites in the News category, using natural log of ad pixels as the dependent variable and traffic as an independent variable.

VARIABLES	(1) Contemporaneous	(2) Lagged
Ln(Traffic)	0.488***	
	(0.171)	
Lagged Ln(Ad Pixels)		0.611***
		(0.0451)
Lagged Ln(Traffic)		0.245**
		(0.0983)
Period	-0.109	0.0184
	(0.138)	(0.242)
Constant	-0.0850	-0.992
	(2.251)	(1.506)
Observations	670	497
N	172	172
R^2	0.0234	0.392

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In Table 6, again we do see that traffic is a significant positive predictor of ad quantity, supporting Hypothesis 4, based on the theory from GOS. The result holds in both Models (1) and (2), the contemporaneous and the lagged versions of the model, respectively. Again, we analyzed the other lagged (reversed) version: traffic on lagged pixels. In that direction, the relationship is not significant. (We omit the table of coefficients.) For the general sample, we explained that the significant positive relationship between lagged pixels and traffic can arise due to a better price for content; if the site is an advertising-only business model, it would not be possible to give a better price for content.

4.4 Summary of Results

We summarize the results of the hypothesis tests in Table 7.

Table 7: Summary of Hypothesis Tests.

Hypothesis (DV, IV)	**Finding**
H1 (ad quantity, number of competitors)	H1 supported: negative relationship between ad quantity and number of competitors.
H2: (ad quantity, intensity of competition)	H2b supported: negative relationship between ad quantity and intensity of competition.
H3: (ad quantity, traffic)	H3a supported: positive relationship between ad quantity and traffic
H4: (ad quantity, traffic) for News sites	H4 supported: positive relationship between ad quantity and traffic for News sites.

4.5 Further Analysis of the KS Relationships

A central message of KS is a negative relationship between traffic and ads, but we don't see that effect in the data. Here we investigate which parts of their mechanism do show in the data.

As we explained above, KS describe sites with a "content level," which represents non-ad revenue per visitor. Content can include physical goods for sale, information goods like music or news for sale, or a subscription to a service. In their results, content is positively related to traffic and to cost per click of ads sold and negatively related to ad quantity sold. In our data, we have direct measures of only two of these constructs, traffic and ad quantity. Although we cannot measure content or cost per click directly, we do have a proxy. Our measure of competitive intensity in the ad market, the Google AdWords competition metric, reasonably represents cost per click: sites featuring high-competition keywords can charge more for a click on an ad. We further infer that sites with lucrative ads have a potentially profitable target audience. As KS find, high rates for cost per click positively correspond to the "content level."

We see the following patterns in our data, as predicted by the equilibrium analysis in KS:

1) Sites that have higher competitive intensity in the ad market—which we interpret both as receiving higher cost per click and profiting more from sales of content to the visitors—have a lower ad quantity. (Tables 2 and 3, Models (1)-(3) in both tables.)
2) Those same sites also have higher traffic. (The correlation between competitive intensity in the ad market and traffic is positive and significant $\rho=0.09$, $p < 0.05$.)

The piece of the KS logic that is *inconsistent* with our data is their prediction that those high profitability, high traffic sites have fewer ads. In our data, we see that the high profitability sites *do* have fewer ads, but the higher traffic sites have *more* ads (H3a).

Overall, these patterns create a story more complex than any single one of the three models we study. On the one hand, we see some support for elements of each (the two points listed above for KS, H2b and H4 for GOS and H3a for KNS). On the other hand, we see contradictions to each. Even though competition drives down ad quantity (H2b), and traffic and ad quantity are positively related (H3a), there is a positive relationship between competitive intensity in the ad market and traffic. Sites that can make money from visitors (either from "content" or from ads) make a concerted effort to generate traffic (e.g., by offering "favorable terms" à la GOS or KNS or by buying ads à la KS). But contrary to the logic in KS, in which those sites restrict ads to reduce visitors' opportunities to click away, those sites choose to monetize the valuable traffic by showing ads.

4.6 Panel Tobit Model

In addition to our main GLS analysis, we also estimate a Tobit model to account for the left-censoring on the ad pixels measure (i.e., it is not possible to have negative pixels).

A Tobit model can capture censoring, typically used, as in our case, when the dependent variable is restricted to be positive. For us, that is useful because of the sites that have no advertising: there is a mass of sites with zero pixels. The dependent variable in our analysis is the natural log of (ad pixels + 1) to maintain the zero value even when taking the log.

In the results shown below, we include both competition and traffic measures as dependent variables. We lag the independent variables, and the dependent variable is ad quantity, consistent with the direction of causality assumed in KNS and GOS and with our results above.

Table 8: Tobit results using natural log of ad pixels as the dependent variable and traffic and competition measures as the independent variables.

VARIABLES	(1)	(2)	(3)
Lagged Ln(Ad Pixels)	1.908***	1.911***	1.907***
	(0.0459)	(0.0459)	(0.0459)
Lagged Ln(Traffic)	0.754***	0.959***	0.782***
	(0.106)	(0.149)	(0.105)
Lagged ln(HHI)	0.393**		
	(0.179)		
Lagged Ln(Market Share)		-0.201	
		(0.127)	
# Competitors			-0.000764**
			(0.000329)
Comp Intensity Ads	-3.626***	-3.741***	-3.626***
	(0.747)	(0.749)	(0.747)
Period	0.468*	0.488**	0.489**
	(0.240)	(0.240)	(0.240)
Constant	-24.51***	-29.26***	-25.59***
	(1.804)	(2.545)	(1.658)
σ_{μ}	0	0	0
	(0.551)	(0.549)	(0.555)
σ_{e}	11.13***	11.13***	11.13***
	(0.247)	(0.247)	(0.247)
Observations	8,341	8,341	8,341
N	2,927	2,927	2,927

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The results from this analysis shown in Table 8 are largely consistent with the GLS analysis. First, competition and ad quantity are negatively related: Model (3) supports H1; and all three models support H2b—Model (1) using HHI as a measure of competitive intensity, Model (3) using number of competitors, and all three models using competitive intensity in the ad market. (Model (2) shows that we do not replicate the result with market share as a measure of competitiveness.) Second, all three models show a positive relationship between traffic and ad quantity, supporting H3a. These consistent patterns reveal that the results are robust to different forms of the regression.

5. Conclusion

Our empirical analysis contributes to the debate in the theoretical literature about website advertising. In particular, we find support for two main results. First, that competition is a force that drives down ad quantity. Second, that there is a positive association between website traffic and advertising: the more popular sites have more ads. In this final section, we explore the limitations of our analysis and discuss the results.

5.1 Limitations

We have attempted to create empirical measures as close as possible to the concepts in the relevant theoretical models, but inevitably, we have made compromises in collecting and analyzing the data. For example, we crawl only the home page of the sites in our sample, not all the pages in the sites, due to the technical challenges and time intensiveness of a more extensive crawl. The decision to crawl only the home pages reduces the measurements of ad quantities on all the sites. Further, we acknowledge that the linear form of Eq. (1) is not ideal for estimating a dependent variable with a mass at zero. But it is simple, easy to interpret, runs in a reasonable amount of time on panel data, and symmetric in our temporal treatment of reverse causality.

Another limitation is the reliance on dmoz. Dmoz is human coded, and is a less comprehensive index of the web than the search engines. We used it because it was an available, extensive, and relatively objective source of categorizing websites. However, there are clearly biases in selection into dmoz: larger traffic and higher ad quantity websites are disproportionately represented. We recognize our choice for home-page crawling and the use of dmoz as limitations, but in each case, we proceeded only when we felt reasonably sure that there weren't systematic biases with direct bearing on our research questions.

Another limitation is the reliance on the temporal arguments as evidence of causation. Of course a temporal relationship is only a necessary condition for establishing causation, not a sufficient one.

We investigated an instrumental variables approach to augment the temporal analysis, but we didn't find guidance for appropriate instruments in the literature.

5.2 Discussion

Our result on competition confirms the results of GOS, that sites recognize ads as a nuisance to consumers, and that they reduce their prevalence in the face of competition. That result contradicts the alternative hypothesis, the more subtle logic proposed by KNS, that competition forces sites to offer a better deal to consumers on content, and therefore pushes up the ad quantity.

Even though we don't see support for that positive main effect between competition and ad quantity, we do see support for elements of the KNS story. In particular, we see support for the KNS prediction that higher traffic sites tend to have higher ad quantity. KNS explain that

> [O]ur model predicts that media products that are mainly advertising financed have relatively large audiences. … This is not because the media firms seek a broad audience as such, but because the competitive pressure forces them to behave so that they attract a larger audience (p. 1113).

They argue, in other words, that optimal content provision (overall value, considering price and quality) drives both ad quantity decisions and traffic outcomes. It is the very situations in which the site will decide to have high ad quantity that they will take actions that encourage high traffic. We see some evidence that their proposed mechanism is at play, namely, that lagged ad quantity predicts traffic. Surely greater ad quantity is not *causing* more traffic, but their story about the role of content choices explains that pattern. Therefore, even though their proposed effect of competition on ad quantity is empirically dominated by the "compete to lower ads" logic of GOS, we do see its fingerprints.

In finding support for the aforementioned KNS story, our result of a positive relationship between traffic and ad quantity directly contradicts a major result of KS. They summarize their results:

> [S]ites…with low content, specialize in selling links (i.e., traffic), whereas sites with high content tend to buy links (advertise) in order to benefit from content (product) sales (p. 770).

In fact, what we find is that sites with higher traffic also tend to have more advertising. When the target audience for a site is especially lucrative, the site can convert that potential both by selling its "wares" (KS's "content") *and* by selling to advertisers who will pay a premium to reach the coveted audience. In that case, the site has a doubly strong incentive to drive traffic.

Why might the KS prediction not hold? The KS surfing model has web visitors randomly start at sites across the web and then follow links, and in an extension they consider search engine use. However, the model doesn't emphasize direct traffic, in which visitors directly type in the web address (or, equivalently, use a bookmark) to arrive at the page. Direct traffic is possibly an important source of visitors to many highly popular primarily-advertising supported sites (like the ones mentioned above, e.g., Facebook). Sites that are deliberate destinations could sell a lot of advertising and not much make much money from content.

In addition, the random walk model might also be an insufficient description of web surfing behavior. They model a web surfer who is equally likely to take any of the available paths, treating clicking on an available link on par with staying on the page. If there are nine advertising links on a page, the visitor has a 10% chance— 1/(9+1)—of staying on the page. Perhaps this is not a reasonable assumption; perhaps it is that the higher the "content" of the site, the more likely the person is to stay on the page, or return to the page, regardless of the advertisements.

One way to further test the KS mechanism would be to look at whether sites that spend more on advertising have higher traffic, as they predict. Currently, we do not have the data on ad spend to test that idea.

In conclusion, compared to other forms of media such as printed magazines or newspapers, acquiring advertising on websites is relatively easy. Sites (the "publishers") can join advertising networks or programs such as Google's AdSense, which handle many of the details of matching advertisers and publishers. The online advertising industry is developing new tools, including Demand Side Platforms (DSPs) and Supply Side Platforms (SSPs) to make it even easier for advertisers and publishers to connect. From a logistical perspective, a site could run as much advertising as it wants. Of course sites recognize that there is a trade-off in gaining more revenue from ads and degrading the visitor's experience on the site. What we have shown is that higher traffic, lower competition sites tend to contain more advertising. This suggests that trade-off is made systematically differently depending on the site's circumstances (namely traffic and competition levels), not just on how bothersome ads will be to a site's specific audience.

Our result on traffic is unwelcome news for fledgling web-based businesses that hope to find success through an advertising business model: the more popular sites that are able to sustain more advertising, not the less popular ones. A less popular site already earns less revenue by having fewer visitors and therefore earning fewer impressions of the ads; our results show that the disadvantage is made worse because less popular sites also have fewer ads.

Appendix A: Example Competition Data for http://www.petfinder.com

Phrase	Competition
adopt a pet	0.577990
pet for adoption	0.741832
lost pets	0.642122
humane society of south brevard	0.073914
senior dog rescue of oregon	0.146667
humane society of west louisiana	0.104353
precious pets adoption league	0.293182
save the animals rescue team ii	0.090043
shiba rescue gta	0.051147
fresh start bird rescue	0.164119

References

Bagwell, Kyle. 2007. "The Economic Analysis of Advertising" in *Handbook of Industrial Organization*, Volume 3.

Dukes, Anthony. 2004. "The Advertising Market in a Product Oligopoly," *Journal of Industrial Economics*. 52(3):327-348.

Dukes, Anthony. 2006. "Media Concentration and Consumer Product Prices," *Economic Inquiry*. 44(1): 128-141.

Edwards, Jeffrey R. and Richard P. Bagozzi. 2000. "On the Nature and Direction of Relationships Between Constructs and Measures," *Psychological Methods*. 5(2) 155-174.

Gal-Or, Esther and Anthony Dukes. 2003. "Negotiations and Exclusivity Contracts for Advertising," *Marketing Science.* 22(2):222-245.

Godes, David, Elie Ofek, and Miklos Sarvary. 2009. "Content vs. Advertising: The Impact of Competition on Media Firm Strategy," *Marketing Science*. 28(1):20-35.

Greene, W. 2003. *Econometric Analysis.* 5th ed. New York: MacMillan.

Kaiser, Ulrich and Julian Wright. 2006. "Price Structure in Two-Sided Markets: Evidence from the Magazine Industry," *International Journal of Industrial Organization*. 24(1):1-28.

Katona, Zsolt and Miklos Sarvary. 2008. "Network Formation and the Structure of the Commercial World Wide Web," *Marketing Science.* 27(5):764-778.

Kind, Hans Jarle, Tore Nilssen, and Lars Sorgard. 2009. "Business Models for Media Firms: Does Competition Matter for How They Raise Revenue?" *Marketing Science*. 28(6):1112-1128.

Quantcast. 2012. http://www.quantcast.com/top-sites-1 . Last accessed February 20, 2012.

Rochet, Jean-Charles and Jean Tirole. 2006. "Two-Sided Markets: A Progress Report," *RAND Journal of Economics*, 37(3):645-67.

Wilbur, Kenneth C. 2008. "A Two-Sided, Empirical Model of Television Advertising and Viewing Markets," *Marketing Science*. 27(3):356-78.