Manufacturing & Service Operations Management

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Cost-per-Click Pricing for Display Advertising

Sami Najafi-Asadolahi, Kristin Fridgeirsdottir

To cite this article:
Sami Najafi-Asadolahi, Kristin Fridgeirsdottir (2014) Cost-per-Click Pricing for Display Advertising. Manufacturing & Service Operations Management 16(4):482-497. http://dx.doi.org/10.1287/msom.2014.0491

M&SOM MANUFACTURING & SERVICE OPERATIONS MANAGEMENT
Vol. 16, No. 4, Fall 2014, pp. 482–497
ISSN 1523-4614 (print) | ISSN 1526-5498 (online)

informs

http://dx.doi.org/10.1287/msom.2014.0491

Cost-per-Click Pricing for Display Advertising

Sami Najafi-Asadolahi
Leavey School of Business, Santa Clara University, Santa Clara, California 95053, snajafi@scu.edu

Kristin Fridgeirsdottir
Management Science and Operations, London Business School, London NW1 4SA, United Kingdom, kfridgeirsdottir@london.edu

Display advertising is a $25 billion business with a promising upward revenue trend. In this paper, we consider an online display advertising setting in which a web publisher posts display ads on its website and charges based on the cost-per-click pricing scheme while promising to deliver a certain number of clicks to the ads posted. The publisher is faced with uncertain demand for advertising slots and uncertain traffic to its website as well as uncertain click behavior of visitors. We formulate the problem as a novel queueing system, where the slots correspond to service channels with the service rate of each server inversely related to the number of active servers. We obtain the closed-form solution for the steady-state probabilities of the number of ads in the publisher's system. We determine the publisher's optimal price to charge per click and show that it can increase in the number of advertising slots and the number of promised clicks. We show that the common heuristic used by many web publishers to convert between the cost-per-click and cost-per-impression pricing schemes using the so-called click-through-rate can be misleading because it may incur substantial revenue loss to web publishers. We provide an alternative explanation for the phenomenon observed by several publishers that the click-through-rate tends to drop when they switch from the cost-per-click to cost-per-impression pricing scheme.

Keywords: queueing systems; online advertising; pricing; Markov chains; cost-per-click
History: Received: November 28, 2011; accepted: March 21, 2014. Published online in *Articles in Advance* August 1, 2014.

1. Introduction

Display advertising is currently a $25 billion business (Anandan 2012), which is expected to reach $200 billion in "a few short years," according to Google. As a result, Google is now investing heavily in display advertising (Peterson 2011) and not focusing only on its sponsored search advertising, where textual ads are displayed along with search results. In addition, display advertising is expected to continue to grow at a faster pace and overtake sponsored search advertising by 2015 (Fredricksen 2011). This paper focuses on a common online advertising setting in which web publishers post display ads on their websites for an agreed-upon number of clicks (guaranteed delivery), and charge based on the cost-per-click (CPC) pricing scheme (i.e., an advertiser pays a certain price for each click made to his ad). The publishers are often faced with uncertain demand from advertisers requesting advertising space to post their ads and uncertain supply of visits from viewers whose "click behavior" is also uncertain. That is, even though the publisher guarantees to serve an ad with a certain number of clicks, the completion of the service is highly uncertain because it depends on how many viewers would visit the website and upon a visit, how much chance the ad has to be recognized and clicked on. In such an inherently uncertain environment, pricing is one of the most challenging operational decisions that web publishers face, and mostly ad-hoc approaches are currently used. It is now generally believed that the ability to determine the CPC price of display ads optimally in this highly uncertain environment is key to the web publishers' revenue increase. However, optimal pricing of display ads has not received much attention in the literature, in contrast with pricing of sponsored search ads, which is quite well researched (see, e.g., Edelman et al. 2007 and references therein).

In view of this gap, this paper has three main objectives. First, we develop a modeling framework that captures the fundamental operational challenges faced by web publishers posting ads on their websites and charging based on the CPC pricing scheme while promising to deliver a certain number of clicks on the ads posted. The publishers are faced with uncertain demand for advertising space through an advertising network (an online intermediary matching and sending advertisers to related websites), and uncertain traffic to their websites as well as uncertain "click behavior" of the visitors. Second, we use this model to determine the publishers' optimal price to charge per click and investigate the impact of various factors, such as the number of advertising slots and promised clicks on its behavior. Our third objective relates to a simple approach commonly used in practice: web

publishers convert CPM prices to CPC prices using the click-through-rate (CTR). If web publishers promise impressions (and are risk neutral), this simple conversion approach can be appropriate. However, publishers are increasingly promising certain numbers of clicks. We investigate the limitations of this simple conversion.

Advertising Networks. Advertising networks are online companies that connect web publishers who want to sell their clicks or impressions (i.e., online inventory) with advertisers who want to run their ads on the relevant websites. Large publishers often sell up to around 60% of their inventory through advertising networks, and smaller ones often sell their entire inventory. We focus on publishers that receive their demand through ad networks. However, our model also applies to the setting where direct sales channels are used with advertisers not willing to wait for advertising space to become available. This scenario is very common when there is intense competition of web publishers for attracting advertisers. The setting where a web publisher posts ads sent through an advertising network and charges based on the CPC pricing scheme captures roughly 25% of the display advertising market (Interactive Advertising Bureau 2011).

We consider a common type of advertising networks, known as blind networks. A blind network is such that advertisers clearly define their desired slot categories for their ads in advance when registering with the ad network (for instance, they may request a right-hand-side slot on a sport page). However, they do not know the exact website that their ads would be posted on. Contextweb, Valueclick, and Clicksor are examples of blind ad networks.

Advertising networks usually work with *immediate* inventories. That is, an advertiser's demand is sent to a web publisher only if it has space available to post the ad in the advertiser's requested category. Otherwise, the ad network does not offer slots with this publisher and automatically directs the demand to other available publishers.

Because advertising networks often contain thousands of websites, it is rare that an advertiser's desired slot is unavailable. However, even in such unusual cases, advertising networks do not keep advertisers waiting. Rather, they direct advertisers to available publishers that participate in one of their partner networks.

Transaction steps. The general steps for the transactions made between advertisers and publishers through an advertising network are as follows:

(1) A web publisher has slots available and approaches the advertising network. The publisher registers each group of equivalent slots (in terms of size, format, and page) as a separate "subsystem" with a different tracking code and a chosen price (per click). The price that the publisher chooses for each subsystem is often called the subsystem's (or the slots') ask-price. Publishers are mostly free to determine their ask-prices. Nevertheless, some networks, such as Clicksor, have a more selective process. In these networks, publishers are often segmented into two main groups of *premium* and *nonpremium* publishers. Premium publishers can freely choose their ask-prices, but the prices for nonpremium publishers are automatically set by advertising networks. Advertising networks often do not reveal the price information to nonpremium publishers, but they guarantee to pay no less than a preagreed minimum payment. In this paper, we restrict our attention to the common case of publishers freely setting their own ask-prices.

(2) An advertiser requests his ad to be posted on a related participating website in the network. When registering his request, the advertiser clearly defines his target slot category as well as the maximum price he is willing to pay (bid-price). In addition, most advertisers choose mostly either of two contracts: The first contract is known as guaranteed delivery (GD), in which the advertiser requests a certain number of impressions or clicks to be made to his ad. The second common contract is known as fixed-time-campaign-length (FTCL), where the advertiser specifies a start and end date and requests his ad to be posted for that fixed time frame (see, e.g., Akella et al. 2009). In this paper, we focus on the GD contract where advertisers are charged based on the CPC pricing scheme.

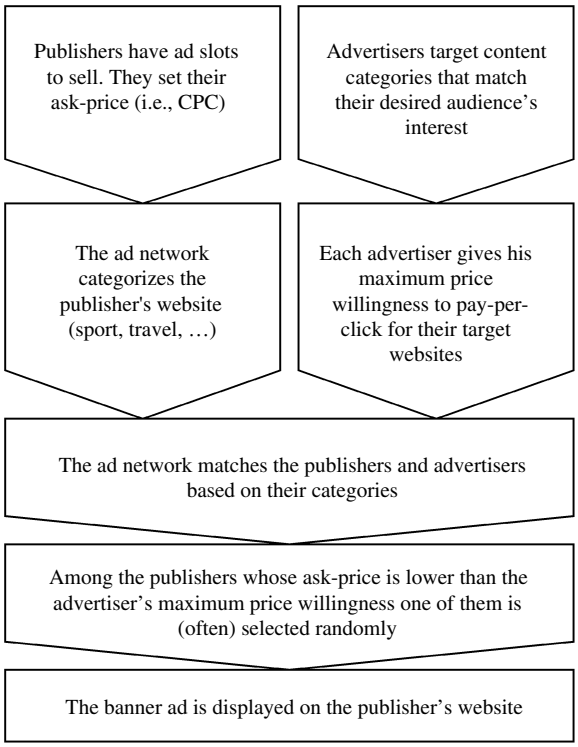
We note that in practice, there are some slot categories that no advertiser may be interested in. These categories are often referred to as *orphan categories* (Ghosh et al. 2009). Because there is little demand for orphan categories, some networks try to sell them using auctions at lower prices through advertising exchanges. Ad exchanges are platforms that facilitate buying and selling of the orphan slot categories from multiple ad networks. In this paper, we restrict our focus to GD contracts. Balseiro et al. (2013), Muthukrishnan (2009), and Celis et al. (2012) are examples of recent research for auction pricing in advertising exchanges.

(3) The ad network sends the ad to an available subsystem that is matched to the advertiser's target category and price. The relevant publisher posts the ad by using a delivery engine.

(4) The advertiser pays the ask-price to the ad network. The advertising network takes about 25%–50% of payment as its own commission and transfers the rest to the publisher's account. Figure 1 summarizes the steps in an online transaction between publishers and advertisers based on a typical ad network.

Web publishers seek to choose the CPC prices that maximize their expected revenues given the uncertain arrivals of advertisers sent from the ad network and the uncertain supply of clicks from viewers. To capture

Figure 1 The General Steps for Transactions Between Advertisers and Web Publishers Through an Advertising Network



the dynamics of the display advertising setting with advertisers approaching the publisher at any time and viewers uploading the website at any time, we model the publisher's system as a queueing system in which advertisers act as customers who arrive at the system requesting to be served with certain numbers of clicks, viewers act as servers, and the slots act as serving channels. The resulting queueing system is new and, despite having complicated dynamics, a closed-form solution of the steady-state probability can be determined.

The primary contributions of this paper are as follows:

1. We construct a modeling framework capturing the main trade-offs in the operation of a web publisher that comes from matching supply with demand. We derive a closed-form solution of the steady-state probability distribution of the number of advertisers in the web publisher's system. This enables us to determine the optimal price for the web publisher to charge advertisers and analyze the publisher's system in detail. (See §§3, 4, and 6.)

2. We demonstrate that the simple conversion rule employed in practice, in which a publisher uses the CTR to convert between the CPM and CPC prices may be misleading, resulting in a substantial revenue loss compared to the optimal policy. (See §5.)

3. We provide further insights by showing that the optimal CPC price can increase in the number of advertising slots and the number of clicks. These results may go against our common intuition from the supply-demand relationship: an increase in advertising slots in the system can be interpreted as an increase in the service capacity. However, the fact that the service rate depends on the number of active servers has the opposite impact. In addition, as the number of clicks increases, the publisher typically is expected to give a discount. However, the publisher can serve fewer advertisers because of the increased service time, which has an opposite impact on price. (See §4.)

This paper is organized as follows: Section 2 provides the relevant literature. Section 3 describes the model formulation. Section 4 discusses the web publisher's revenue maximization problem, and §5 describes how the heuristic used in practice to convert between CPM and CPC prices can be misleading. Section 6 presents some extensions to the publisher's problem, and §7 concludes and presents directions for future research.

2. Literature Review

There are two streams of literature related to our research. The first is online advertising within the *marketing* area, which is quite extensive. Hoffman and Novak (2000) provide an overview of advertising pricing schemes for the Internet. However, there is limited literature on analytical models for optimal pricing and other decision making for a web publisher with an advertising operation. (For issues faced by advertisers such as predicting audience for advertising campaigns, see, e.g., Danaher (2007) and papers referenced therein.)

The second stream of literature is in *management science*. The online advertising research within this area is limited, and there are few works directly related to online advertising pricing.

In some of the earlier work, Mangani (2004) compares the expected revenues from the CPC and the CPM schemes using a simple deterministic model. Unlike our paper, he does not consider the uncertainties involved with the advertisers' demands and viewers' supplies. Chickering and Heckerman (2003) develop a delivery system that maximizes the CTR given inventory-management constraints in the form of advertisement quotas. Both of these papers assume the prices are fixed. Fjell (2009) uses a deterministic economic model to analyze the choice between CPM and CPC when a web publisher is both a price taker in the market for display ads and faces a decreasing number of viewers visiting its website. McAfee et al. (2013) consider a deterministic model for a web publisher selling maximally representative allocations to advertisers based on the GD contract. Lewis and Reiley (2011) measure the impact of advertising on sales through an experiment performed between Yahoo! and a major

retailer. They find that online display advertising can have a significant impact on a retailer's sales.

Some authors have considered the problem of a web publisher who generates revenues not only from advertising but also from subscriptions. Baye and Morgan (2000) develop a simple economic model of online advertising and subscription fees. Prasad et al. (2003) model two offerings to viewers of a website: a lower fee with more ads and a higher fee with fewer ads. Kumar and Sethi (2009) study the problem of dynamically determining the subscription fee and the size of advertising space on a website. Unlike our paper, all these papers are focused on capacity management problems instead of pricing decisions, and the price is assumed to be fixed.

Scheduling of ads on a website has also recently become a popular topic. Kumar et al. (2006) develop a model that determines how ads on a website should be scheduled in a planning horizon to maximize revenue. Their problem belongs to the class of NP-hard problems, and they develop a heuristic to solve it. They also provide a good overview of other related papers on scheduling. In a related work, Turner et al. (2011) develop a model for the dynamic in-game ad scheduling problem faced by a leading network provider of in-game ad space.

The model studied in this paper is a queueing model with a state-dependent service rate. This class of queueing models has been extensively explored in the literature. George and Harrison (2001) study a Markovian single-server queue in which the system manager dynamically controls the service rate to minimize the long-run average cost but has no control over consumers' arrival rate. Ata and Shneorson (2006) consider a service facility in which a system manager dynamically controls the arrival and service rates to maximize the long-run average value generated. The service facility is modeled as an $M/M/1$ queue with adjustable arrival and service rates, and they find an explicit solution to the problem. The CPC system differs from this stream of literature by considering the uncertainty in when the service is delivered or who is served at each point in time. When a viewer arrives at a CPC system, it is uncertain which advertiser receives the next click. Therefore, when requesting the same number of clicks, some advertisers may by chance, stay long while others are quickly served and replaced with new advertisers.

In the CPC system, the service rate is divided among the jobs in the system. This feature is similar to the processor sharing (PS) discipline, which has been well researched in the literature. In a typical PS queue, the server divides the service rate proportionally among the jobs in the system. The jobs (are guaranteed to) receive their promised services either at the same time (with a reduced service rate), or each job receives some part of the service (depending on the service weight allocated to each job as a fraction of the bandwidth) and then waits until all other jobs receive their shares of the services in a round-robin way before the next service round is started. A common example of a PS system is an Internet provider sharing its resources (e.g., Internet bandwidth) among the existing users. Examples of these models include Jagerman and Bhaskar (1991), who consider the $G/M/1$ PS queue using a heavy traffic analysis; Zwart and Boxma (2000), who consider sojourn time asymptotics in the $M/G/1$ PS queue; and Puha et al. (2006), who consider the fluid limit of an overloaded PS queue. The CPC system is different. Unlike a typical PS queue, in the CPC system there is uncertainty in who receives the service once a viewer arrives and therefore when a job will complete its service. Hence, some jobs may by chance stay long while others are served quickly and replaced with new jobs for receiving identical services.

There is growing literature on online display advertising from a revenue management perspective, which focuses mainly on the optimal display ads allocation problems. Examples of recent works include Balseiro et al. (2011), Yang et al. (2010), and Alaei et al. (2009) (for a reference of traditional revenue management models, see, e.g., Talluri and van Ryzin 2004). Ciocan and Farias (2012) develop an algorithm for a large class of dynamic allocation problems with unknown demand, with applications in display ad slot allocation and network revenue management. Chen (2011) considers a mechanism design approach for a monopolistic web publisher that wishes to decide whether to allocate its display ad slots to guaranteed contracts or the spot market. Balseiro et al. (2011) consider a similar problem for a web publisher with a single slot, and they use a stochastic control approach to characterize an asymptotically optimal efficient allocation policy. Balseiro et al. (2013) study auctions for online display advertising exchanges and show that ignoring advertisers' budgets in these markets can result in substantial revenue losses for publishers. Araman and Popescu (2010) study the ad allocation problem for more traditional media, specifically broadcasting. Their model is concerned with how to allocate limited advertising space between up-front contracts and the so-called scatter market (i.e., a spot market). Araman and Fridgeirsdottir (2011) consider a similar web publisher setting to our paper and study pricing and capacity management for a CPM system where advertisers are willing to wait. Their setting does not allow for closed-form solutions and they derive asymptotically optimal solutions.

3. The Model

In this section, we formulate the problem of a web publisher facing uncertain demand from advertisers

requesting space to display their ads. The publisher's website consists of a single webpage with n similar slots for ads. In §6.2, we will generalize this setting and consider a website with multiple pages. The web publisher uses an ad network that supplies it with the demand. Advertisers request a certain number of viewers to click on their ads. However, the supply of viewers (in a way specified below) is uncertain as well as their clicking behavior.

Advertisers' Arrivals and Click Request. We assume that advertisers arrive through the ad network at the publisher's website according to a Poisson process with rate λ. Considering Poisson arrivals is common in service settings (see, e.g., Van Mieghem 2000, Cao et al. 2003, Savin et al. 2005). While this assumption captures the stochastic nature of advertisers' demand and may be appropriate for some ad networks, it is unlikely to be a good universal estimator of all advertisers' arrival distributions. We retain the Poisson assumption to maintain the analytical tractability.

Each advertiser arriving at the publisher's page requests his ad to be posted on one of the slots on the page until clicked by x unique viewers. In reality, the number of clicks x can be random across advertisers, i.e., x is a random variable. We will look at this as an extension in §6.3.

Moreover, advertisers arrive at the web publisher's system as long as the publisher has a slot available on the page. If no empty slot is available, the network does not send any advertiser to the publisher. This implies that the publisher's website is a *loss* system. Note that if advertisers approach a web publisher directly (not using an ad network), they may be willing to wait for an available slot. In that case, if the waiting time is short (e.g., advertisers are served shortly after their arrivals), then, since no (or little) queue is formed, the arrivals and the service mechanisms of the system would still be close to those of the loss system. Thus, the main results and managerial insights would be similar as well.

In addition, when the web publisher has a slot available, it usually does not leave it empty; rather, it places a default (filler) ad (remnant advertising). A default ad is often the publisher's own ad (house ad) or a run-of-network ad that the ad network sends to fill the place (e.g., a public service announcement). In both cases, a default ad generates minimal revenue. Hence, when a revenue-generating ad is sent to the publisher the filler ad would be replaced by a proper revenue-generating ad. Nevertheless, we note that the emergence of recent real-time ad exchanges is changing this reality. The reason is that ad exchanges enable advertisers to bid for more valuable slots that reach their ads to right customers. Hence, the quality (i.e., click chance) and the volume of slots sold through ad exchanges are both increasing (Balseiro et al. 2013, Vranica 2013).

Viewers' Arrivals and Click Behavior. We assume that viewers arrive at the publisher's website according to a Poisson process with rate μ. This assumption is consistent with empirical studies (e.g., Cao et al. 2003) that show that the viewers' traffic tends locally to Poisson distribution. An arriving viewer clicks on an ad with probability β or leaves the publisher's system without clicking on any ad with probability $1-\beta$. We denote by $\hat{\mu} := \beta\mu$ the effective rate with which viewers click on one of the ads in the publisher's system.

As for viewers' click behavior, we first assume that a viewer leaves the system after clicking on one ad. While this behavior is frequently observed in reality, it would be more realistic to consider that a viewer can click on multiple ads before leaving the publisher's system. However, it is easy to see that considering this behavior leads to intractable settings. In addition, this assumption is consistent with recent empirical studies, e.g., Jeziorski and Segal (2009), that investigate the click behavior of viewers and suggest that viewers tend to become *satiated* after clicking on good ads. These studies suggest that, in some cases of satiation, a viewer does not derive any benefit from clicking on a second ad.

Next, we assume that each similar ad has an equal chance to be clicked on. This implies that the clicking probability on a given ad decreases in the number of similar ads on the page. This assumption is consistent with recent empirical studies, e.g., Jeziorski and Segal (2009) and Gomes et al. (2009), which show that the clicking probability on a given ad tends to reduce with the existence of competitor ads.

Finally, we assume that viewers always prefer to consider and click on real ads compared to default ads when both are posted on the page. For example, CNN.com and FT.com frequently display their own default ads. These ads are often not designed to be clicked on because the publisher's major aim from displaying default ads is to strengthen its own brand recognition. In addition, this assumption is consistent with recent empirical studies (e.g., Jeziorski and Segal 2009), which suggest that viewers usually examine the *competitor* ads (i.e., ads that are similar in terms of content, format, and size) before they decide which one to click on. Thus, viewers tend not to click on filler ads as they are not competing with real ads.

The Optimization Problem. The publisher's goal is to maximize its total revenue rate by determining the right prices to charge. The revenue rate consists of the payments made by advertisers multiplied by the "actual" demand rate (defined later). Each payment consists of the price per click, denoted by p (specified below), multiplied by the number of clicks requested x.

We capture the price sensitivity of the advertisers with the price-demand function $p(\lambda, x, n)$, which is assumed to be continuous and (weakly) decreasing in the advertisers' arrival rate, the number of clicks, and the number of slots. The decreasing relation between the price and the number of clicks captures the fact that, *given that all other parameters are fixed*, advertisers receive lower prices if they purchase larger numbers of clicks. In addition, advertisers often do not want to see their ads posted on pages on which the ads are hardly considered and clicked due to a high *slot congestion* (many ads posted on the page). For this reason, advertisers request pages with a lower slot congestion because it increases the chance that an ad is considered and clicked. Advertisers often perceive such a higher chance as a higher quality of service. Publishers are aware that to deliver a desired quality of service, there is a trade-off between the website's profitability and the slot congestion. To capture this trade-off, we assume the publisher incurs a congestion penalty (i.e., in the form of a price discount) for adding an extra slot. This ensures that the publisher does not post too many ads on the page and that the click chance does not reduce significantly.

Even though it might not be trivial for the publisher to determine the price function, we assume it can do so with trial and error. For instance, ad networks often encourage publishers to start by offering low prices and then gradually increase them to the appropriate values. Furthermore, publishers such as Yahoo! have started looking into estimating the price-demand relationship. The process of advertisers being matched to web publishers based on type preference and willingness-to-pay can be modeled specifically. However, ultimately it will lead to a price-demand relationship. We will not model the process in detail here but in §1, we have provided a description of the matching process common in ad networks.

For popular websites, often only a part of the advertisers' demand can be met by the publisher. This means that the actual demand rate for each subsystem is scaled down by the probability that there are advertising slots available at the arrival time of an advertiser. However, as arrivals are Poisson, we can invoke the *PASTA* property that Poisson arrivals see time averages. Hence, the arrival-time probability of having i advertisers served by the publisher is identical to its steady-state probability (Gross and Harris 1998), which we denote by $\mathbb{P}_i$, $i \in \{0, \ldots, n\}$.

Because we have a one-to-one relationship between the prices and arrival rates of the advertisers, we optimize the revenue rate with respect to the arrival rates and then determine the prices from the price-demand function $p(\lambda, x, n)$. The optimization problem of the publisher that is maximizing its expected revenue rate can be formulated as

$$\max_{\lambda \in [0, +\infty)} R(\lambda) = \lambda(1 - \mathbb{P}_n(\lambda; x, n, \beta, \hat{\mu}))p(\lambda, x, n)x, \quad (1)$$

where $\hat{\mu} = \beta\mu$ is the viewers' effective arrival rate and $\mathbb{P}_n(\lambda; x, n, \beta, \hat{\mu})$ is the steady-state probability that all the slots on the website are occupied. Hence, $\lambda(1 - \mathbb{P}_n(\lambda; x, n, \beta, \hat{\mu}))$ is the advertisers' actual arrival rate into the system. To obtain the optimal CPC price, we first need to characterize the steady-state probability $\mathbb{P}_n$.

The Probability Distribution. Having Markovian arrival and service processes, we can now model the system using continuous-time Markov chains. Note that even though we are ultimately interested in keeping track of the number of advertisers in the system, to model the system's dynamics we need to keep track of the system at a more detailed level: the number of clicks left to be delivered for each slot.

When an advertiser arrives, he is randomly assigned to one of the available slots with an equal probability because the slots are equivalent. This random ad-to-slot allocation means that we can keep track of the dynamics of the system without distinguishing between the slots. Let us now define the state of the system and its transitions. We formulate the system as a queueing model with the state vector

$$\mathbf{k} = (k_1, k_2, \ldots, k_n), \quad 0 \le k_h \le x, \quad h = 1, 2, \ldots, n, \quad (2)$$

in which each component represents the number of clicks left to be satisfied in one of the slots without distinguishing among the slots. For instance, k_h indicates that there is an ad in the system that needs to be clicked k_h times more to leave the system. If $k_h = 0$, it indicates that the corresponding slot is empty. Alternatively, $k_h = x$ indicates that an ad of a new advertiser has just been placed in the slot. Note that because we do not distinguish among the slots (all slots in the system are equivalent) any combination of the same components does not lead to a new state. For example, $(3, 4, 2)$, $(4, 3, 2)$, and $(2, 3, 4)$ all refer to the same state. For convenience, we consider that $\mathbf{k}$'s positive components are always arranged in an increasing order followed by components whose values are zero. We illustrate how the state transitions work through the following examples.

(i) Suppose the system is in state $(k_1, k_2, \ldots, k_h, \ldots, k_i, 0, \ldots, 0)$, $k_m \neq k_l$, $m \neq l$, $m, l \in \{1, 2, \ldots, i\}$, where the first i components are positive and the rest $(n - i)$ are zero (we will consider the case of $k_m = k_l$ later). This means there are i ads in the system with the remaining clicks $k_1, \ldots, k_i$, and the rest $(n - i)$ slots are empty. The viewers consider and click on one of the ads in the system with the effective click rate $\hat{\mu} = \mu\beta$. Given

that a viewer clicks on one of the ads, each of the i equivalent ads has an equal chance $1/i$ to be clicked. For example, if the viewer chooses to click on the ad with k_h remaining clicks, then the state of the system makes a transition from $\mathbf{k}$ to the new state

$$\mathbf{k}' = (k_1, k_2, \ldots, k_h - 1, \ldots, k_i, 0, \ldots, 0) = \mathbf{k} - \mathbf{e}_h^T, \quad 1 \le h \le i$$

with rate $\hat{\mu}/i$, where $\mathbf{e}_h$ is the hth unit vector. For example, if the state of the system is $(2, 3, 4, 5, 0)$, then it can make a transition to the possible new state $(2, 3, 3, 5, 0)$ with rate $\hat{\mu}/4$.

(ii) Next, consider the state of the system to be

$$\mathbf{k} = \big(\underbrace{k_1, k_1, \ldots, k_1}_{i}, k_{i+1}, \ldots, k_h, 0, \ldots, 0\big), \quad k_m \neq k_l,\ m \neq l,\ m, l \in \{i, i+1, \ldots, h\}.$$

We observe that in state $\mathbf{k}$, i ads have the same remaining number of clicks k_1, and a total of h slots are filled. Because we do not distinguish among the ads, the viewer can click on one of the ads with k_1 remaining clicks with an i/h chance, and the other ads each have a $1/h$ chance to be clicked on. As a result, a possible transition could be to the new state

$$\mathbf{k}'' = \big(k_1 - 1, \underbrace{k_1, \ldots, k_1}_{i-1}, k_{i+1}, \ldots, k_h, 0, \ldots, 0\big) = \mathbf{k} - \mathbf{e}_1^T$$

with rate $i\hat{\mu}/h$. We note that because the ads are equivalent, the only distinguishing feature among them is their remaining clicks. Hence, the vectors $\mathbf{k} - \mathbf{e}_1^T$, $\mathbf{k} - \mathbf{e}_2^T, \ldots,$ and $\mathbf{k} - \mathbf{e}_i^T$ all refer to the same system state, where for notational convenience we represent them all with $\mathbf{k}'' = \mathbf{k} - \mathbf{e}_1^T$. For example, if $\mathbf{k} = (3, 3, 4, 0, 0)$, then it makes a transition to $\mathbf{k}'' = (2, 3, 4, 0, 0)$ with rate $2\hat{\mu}/3$.

(iii) Finally, we consider the state of the system to be

$$\mathbf{k} = \big(k_1, k_2, \ldots, k_i, \underbrace{0, \ldots, 0}_{n-i}\big).$$

Now, if an advertiser arrives at the subsystem, the publisher assigns one of the empty slots to his ad and the state would make a transition to the state

$$\mathbf{k}''' = \big(k_1, k_2, \ldots, k_i, x, \underbrace{0, \ldots, 0}_{n-i-1}\big) = \mathbf{k} + x\mathbf{e}_{i+1}^T$$

with rate λ. Once again, we note that the vectors $\mathbf{k} + x\mathbf{e}_{i+1}^T, \mathbf{k} + x\mathbf{e}_{i+2}^T, \ldots,$ and $\mathbf{k} + x\mathbf{e}_n^T$ all refer to the same system state, where for convenience we represent them all with $\mathbf{k}''' = \mathbf{k} + x\mathbf{e}_{i+1}^T$.

We note that each advertiser receives his service not with a fixed rate but with a constantly changing rate, because the probability of viewers considering an ad depends on the number of ads on display in the system at any point in time. For example, if there are three ads on display, since the ads are equivalent, each ad has a one-third chance of absorbing a viewer's attention; whereas, if there are five ads, the chance reduces to one-fifth (see, e.g., George and Harrison 2001, Ata 2005 for other queueing settings that consider state-dependent service rates).

To find $\pi_{\mathbf{k}}$, the steady-state probability that the system is in state $\mathbf{k}$, we characterize all possible states and transitions and solve the stationary flow-balance equations (see Proposition 2). Given the complex transition dynamics, one may wonder if a nondegenerate solution exists for $\pi_{\mathbf{k}}$. Proposition 1 ensures the existence of the steady state for the publisher's Markov chain.

Proposition 1. *In the publisher's Markov chain, a unique nondegenerate solution to the stationary flow-balance equations always exists.*

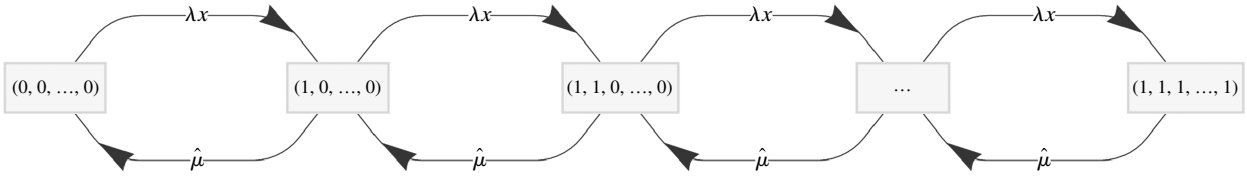
The proof of Proposition 1 relies mainly on two well-known theorems for stochastic processes. The first says that a continuous-time Markov chain (CTMC) has identical and unique limiting and stationary distributions if it is irreducible and positive recurrent, and the second says that an irreducible CTMC with finite number of states is positive recurrent. The publisher's CTMC is clearly irreducible because any state can be reached from any other. It can also be seen directly from (2) that the number of states is finite. Hence, the chain is positive recurrent. Thus, a unique steady-state distribution exists.

For the purpose of the next proposition, we define the transformation $\mathcal{G}_c(\mathbf{k})$ for the n-tuple vector $\mathbf{k} = (k_1, \ldots, k_n)$ as $\mathcal{G}_c(\mathbf{k}) = \{h \mid k_h = c\}$. That is, $\mathcal{G}_c(\mathbf{k})$ maps $\mathbf{k}$ to the set of components whose values (the number of remaining clicks) are $c \in \{0, \ldots, x\}$. For example, if $\mathbf{k} = (2, 3, 5)$, then $\mathcal{G}_2(\mathbf{k}) = \{1\}$, $\mathcal{G}_3(\mathbf{k}) = \{2\}$, and $\mathcal{G}_5(\mathbf{k}) = \{3\}$. Next, we define the real-valued function $|\mathcal{G}_c(\mathbf{k})|$ to be the size of $\mathcal{G}_c(\mathbf{k})$. In other words, $|\mathcal{G}_c(\mathbf{k})|$ is the number of slots with the remaining clicks equal to c. For instance, if $\mathbf{k} = (2, 2, 3, 5)$, then $\mathcal{G}_2(\mathbf{k}) = \{1, 2\}$ with $|\mathcal{G}_2(\mathbf{k})| = 2$, $|\mathcal{G}_3(\mathbf{k})| = 1$, and $|\mathcal{G}_5(\mathbf{k})| = 1$. The next proposition is one of our main results and gives the closed-form solution of the steady-state probability of the number of advertisers in the system.

Proposition 2. *Let k_j $(0 \le k_j \le x)$ be the number of clicks left in slot j $(0 \le j \le n)$. Define $\mathcal{G}_{c_q}(\mathbf{k}) = \{h \mid k_h = c_q\}$ (i.e., the set of components in $\mathbf{k}$ with value c_q) and $|\mathcal{G}_{c_q}(\mathbf{k})|$ as the size of $\mathcal{G}_{c_q}(\mathbf{k})$ $(0 \le q \le Q)$ (i.e., the number of components in $\mathbf{k}$ whose values are c_q), where Q is the number of the groups of slots whose remaining clicks are the same. Then the steady-state probability of the system for state $\mathbf{k}$ is*

$$\pi_{\mathbf{k}} = \frac{\left(\sum_{q=1}^{Q} |\mathcal{G}_{c_q}(\mathbf{k})|\right)!}{\left(\sum_{j=0}^{n} (rx)^j\right), \prod_{q=1}^{Q} |\mathcal{G}_{c_q}(\mathbf{k})|!} r^{\sum_{q=1}^{Q} |\mathcal{G}_{c_q}(\mathbf{k})|}, \qquad r = \frac{\lambda}{\hat{\mu}}. \tag{3}$$

Figure 2 **An Illustration of the CPC System Transition Diagram Where the Advertisers Arrival Rate and the Requested Number of Clicks Are $\lambda_1 = \lambda x$ and $x_1 = 1$, Respectively**



Furthermore, the steady-state probability of having i advertisers in the system is

$$\mathbb{P}_i = \frac{(rx)^i}{\sum_{j=0}^{n}(rx)^j}, \quad i = 0, 1, 2, \ldots, n. \tag{4}$$

It is surprising to see that Equation (4) coincides with that of an $M/M/1/n$ system with $r = \lambda/\hat{\mu}$ and $\rho = rx$. This coincidence is interesting because the two systems have considerably different characteristics. Now, let us explore the reason for the common probability distribution. We consider a similar system where the advertisers' arrival rate and the requested number of clicks take the values $\lambda_1 = \lambda x$ and $x_1 = 1$, respectively. It is easy to verify that Equation (4) remains unchanged. However, having $\lambda_1 = \lambda x$ and $x_1 = 1$ suggests that the advertisers arrive at the system with rate λ_1 and request only one click. In Figure 2 we set up a transition diagram for this system where the state vector is an n-tuple vector, with i components having the value one (indicating there are i advertisers in the system) and $n - i$ zeroes. This means that the state of the subsystem is expressed as

$$\mathbf{k} = (\overbrace{1, 1, 1 \ldots, 1}^{i}, 0, \ldots, 0)$$

and can be collapsed into one dimension. Hence, the transition diagram reduces to that of an $M/M/1/n$ system. The other interesting observation regarding Proposition 2 is that $\pi_{\mathbf{k}}$ does not depend on the actual number of clicks remaining in each slot.

In the next two propositions, we show some structural properties of the average number of advertisers in the system and the busy probability. They will be useful when considering the pricing problem of the web publisher in §4.

Proposition 3. *$\forall x, n$ the full-state probability, $\mathbb{P}_n$, defined by Equation* (4) *satisfies* (i) $\partial \mathbb{P}_n/\partial r \geq 0$, (ii) $\mathbb{P}_n(x+1) - \mathbb{P}_n(x) \geq 0$, *and* (iii) $\mathbb{P}_{n+1}(x) \leq \mathbb{P}_n(x)$.

This proposition is quite intuitive as it says that the full-state probability is increasing in the intensity rate and the number of clicks, and is decreasing in the number of slots. Nevertheless, our numerical analysis indicates that $\mathbb{P}_n$ is not necessarily concave in the number of clicks.

Proposition 4. *$\forall x, n$ the average number of advertisers, $L_n(x)$, and the increment $\Delta L_n(x) = L(x+1) - L(x)$ satisfy* (i) $\Delta L_n(x) \geq 0$; (ii) $\Delta L_n(x+1) \leq \Delta L_n(x)$, $rx > 1$; (iii) $\partial L/\partial r \geq 0$, $\partial^2 L/\partial r^2 \leq 0$, $rx > 1$; *and* (iv) $L_n(x) \leq L_{n+1}(x)$.

Parts (i) and (ii) in Proposition 4 imply that the average number of advertisers in the web publisher's system is increasing concave in the number of clicks. Furthermore, part (iii) implies that the average number of advertisers in the system is increasing concave in the intensity rate r. Part (iv) states that the average number of advertisers in the web publisher's system increases in the number of slots. Propositions 3 and 4 are crucial when solving the optimal pricing problem in §4.

4. The Optimal Price

Having fully characterized the probabilistic properties of the web publisher's operation, we now turn to the task of finding the optimal pricing policy. The web publisher's objective is to determine the optimal price to charge per click that maximizes the revenue rate defined in (11). The Proposition 5 ensures the existence of the optimal solution and gives the optimal price.

Proposition 5. *Let the price function $p(\lambda; x, n)$ be nonnegative weakly concave and decreasing with respect to λ. Then the revenue rate $R(\lambda)$ is concave with respect to λ. In addition, the optimal advertisers' arrival rate λ^* is the unique solution of*

$$\frac{\partial \Gamma(\lambda^*; x, n)}{\partial \lambda} p(\lambda^*; x, n) + \frac{\partial p(\lambda^*; x, n)}{\partial \lambda} \Gamma(\lambda^*; x, n) = 0, \tag{5}$$

where $\Gamma(\lambda; x, n) = rx(1 - \mathbb{P}_n(\lambda; x, n, \beta, \hat{\mu}))$.

To ensure concavity of the objective function, we need $p(\lambda; x, n)$ to be weakly concave. Even though this might seem a restrictive assumption, it includes a *linear price,* which is widely applied in economics and management science literature. In addition, our numerical analysis indicates that many convex price functions give a concave revenue function as well. Proposition 6 indicates that the publisher could be worse off with having more slots. In this proposition, we denote λ^* by $\lambda^*(n, x, \hat{\mu})$ to emphasize the implicit and explicit dependence of optimal arrival rate value on n, x, and $\hat{\mu}$.

Proposition 6. *Let the price function $p(\lambda; x, n)$ be nonnegative weakly concave and decreasing with respect to n and x. Then the optimal revenue rate $R(\lambda^*(n, x, \hat{\mu}), n, x)$ is concave in n.*

This proposition implies that it is possible the publisher loses revenue by serving more ads on its page. The reason for this behavior is the trade-off between the two opposing forces: (i) the number of slots and (ii) price. Adding an extra slot enables the publisher to serve more advertisers at the same time, which increases the revenue rate. However, with an additional slot, the publisher reduces the price to compensate a lower quality of service (i.e., higher slot congestion), which leads to a revenue loss. If the revenue gained from adding a slot is less than the revenue lost due to the price discount, the optimal revenue decreases.

Proposition 7. *Let the price-demand function $p(\lambda, x, n)$ be weakly concave in λ and $\partial^2 p(\lambda, x, n)/\partial\lambda\partial x \le 0$. In addition, let λ^* be the optimal advertisers' arrival rate at the system. Then,*

(i) *λ^* is decreasing in x, i.e., $\partial\lambda^*/\partial x \le 0$; and*

(ii) *$p(\lambda^*, x, n)$ is increasing in x, i.e., $dp(\lambda^*, x, n)/dx \ge 0$, if and only if $\partial\lambda^*/\partial x \le -(\partial p(\lambda^*, x, n)/\partial x)/(\partial p(\lambda^*, x, n)/\partial\lambda)$.*

This proposition is interesting because one typically expects the opposite, i.e., the optimal price to be lower when more clicks are offered. To understand what derives these results, we note that the higher the number of clicks, the longer it takes to serve each advertiser, which means that the web publisher does not need as many advertisers to keep the system busy. This results in a lower optimal arrival rate. In addition, since the price has a decreasing relationship with x and λ^*, an increase in x lowers the price while a decrease in λ^* raises it. Part (ii) suggests that if the price increase due to the lower λ^* is greater than the price decrease caused by a higher x, then the publisher finds it optimal to increase the price rather than giving a quantity discount.

Next, we consider the sensitivity of the optimal price with respect to the number of advertising slots n. When the web publisher increases n, it is typically expected to reduce the price to attract more advertisers and fill the extra space as well, because of a higher slot congestion. Nevertheless, the next proposition shows that when the publisher increases n, it may *increase* the price.

Proposition 8. *Let the price-demand function $p(\lambda, x, n)$ be decreasing and weakly concave in λ. In addition, let λ^* be the optimal arrival rate at the system. Then there exists some $w \in [\hat{\mu}/x, \exp(1)\hat{\mu}/x]$ such that if $\lambda^* \ge w$, then*

(i) *λ^* is decreasing in n, i.e., $\partial\lambda^*/\partial n \le 0$; and*

(ii) *$p(\lambda^*, x, n)$ is increasing in n, i.e., $dp(\lambda^*, x, n)/dn \ge 0$, if and only if $\partial\lambda^*/\partial n \le -(\partial p(\lambda^*, x, n)/\partial n)/(\partial p(\lambda^*, x, n)/\partial\lambda)$.*

Part (i) states that, unlike the typical expectation, when the publisher increases the number of slots it may optimally lose some advertisers rather than serving more. The main reason for this behavior is the trade-off between the service capacity and service time. Increasing the number of slots enables the publisher to serve more advertisers simultaneously. However, as more ads are being served at the same time, each ad has less chance to be recognized and clicked on by a viewer. Hence, advertisers stay longer in the system, which makes the publisher serve fewer advertisers per time unit. If the service capacity gained as a result of an extra slot is less than the capacity lost due to advertisers spending more time to complete their service, the optimal arrival rate decreases. Furthermore, we note that adding extra slots and a consequent decrease in λ^* have opposite impacts on the optimal price. Part (ii) mentions that if the price increase as a result of a lower λ^* (publisher needing fewer advertisers now) is greater than the price decrease as a result of a higher n, then the optimal price increases.

5. The Simple Conversion Rule

Many web publishers that charge per click tend to use a simple conversion rule for obtaining the CPC prices. This conversion is based on dividing the optimal CPM prices by the CTR to calculate the optimal CPC prices (for examples of publishers implementing this conversion rule, see the supplemental material available at http://dx.doi.org/10.1287/msom.2014.0491). This approach is fine (in a risk-neutral setting) if a certain number of impressions is promised. However, publishers are increasingly guaranteeing the numbers of clicks delivered. In that case, this simple rule may lead to a significant revenue loss. In this section, we study the shortcomings of this conversion rule by comparing its revenue with the optimal revenue obtained by using the "correct" CPC prices based on our model derived in §4.

As before, consider a publisher's system that has advertisers arriving with rate λ, viewers arriving with rate μ, each advertiser requesting x clicks with β as the fraction of the viewers clicking on one of the ads. If the publisher uses the simple conversion rule, it charges the scaled CPM price, $p_{\text{cpc}} = p^*_{\text{cpm}}/\text{CTR}$, as the publisher considers selling x clicks to be on average equivalent to selling $N = x/\text{CTR}$ impressions. We assume that the publisher knows how to obtain the optimal CPM price. The CTR used in practice is the observed value for the ratio of the number of people clicking on a certain ad to the total number of people visiting the publisher's system over a certain period. In other words, the CTR value that practitioners observe is the average chance that an ad would be clicked on in a steady-state condition. The following proposition

gives the CTR value that a publisher observes in the long run.

Proposition 9. *The observed CTR value in the long run converges to*

$$\mathrm{CTR}(\lambda,\hat{\mu},x,n,\beta)=\beta\sum_{i=1}^{n}\frac{1}{i}\frac{\mathbb{P}_i^{\mathrm{cpc}}(\lambda,\hat{\mu},x,n,\beta)}{1-\mathbb{P}_0^{\mathrm{cpc}}(\lambda,\hat{\mu},x,n,\beta)}, \quad (6)$$

where $r=\lambda/\hat{\mu}$, $\hat{\mu}=\mu\beta$, *and* $\mathbb{P}_i^{\mathrm{cpc}}$ *is the probability of having* i *ads in the publisher's system.*

As we can see in Proposition 9, the value of the observed CTR depends on advertisers' and viewers' arrival rates, the number of requested clicks, and the number of slots in the system. Note that in this proposition, β/i refers to the expected probability that an ad is clicked on (the state-dependent CTR) during each visit of viewers when there are i ads on display. Moreover, $\mathbb{P}_i^{\mathrm{cpc}}/(1-\mathbb{P}_0^{\mathrm{cpc}})$ refers to the proportion of the time that there are i ads in the publisher's system given that the system has at least one ad. The reason for considering the conditional probability is that the state-dependent CTR is zero when the publisher's system is empty. Proposition 9 emphasizes that the CTR's observed value changes with the price. By changing the price, the publisher is affecting λ, which affects the observed CTR.

In our comparative analysis, we set $\mu=100$, $p_{\mathrm{cpm}}(\lambda)=0.005-0.01\lambda^c$, and $\beta=0.01$, and calculate the relative revenue gap when the optimal CPC price is used compared to using the CPC price derived through the simple heuristic from the optimal CPM price. Specifically, we take the following steps: (i) We obtain the "optimal" CPM price (p^*_{cpm}) and the optimal advertisers' arrival rate of the equivalent CPM system (λ^*_{cpm}). To find these two values, we apply the steady-state probability of the number of advertisers in a CPM system (Fridgeirsdottir and Najafi-Asadolahi 2013) as follows:

$$\mathbb{P}_n^{\mathrm{cpm}}(\lambda,\mu,N,n)=\frac{\binom{N+n-1}{n}(r/(1+r))^N(1/(1+r))^{N-1}}{\sum_{j=0}^{n}\binom{x+N-1}{j}(r/(1+r))^j(1/(1+r))^{N-1+n-j}},$$

where N is the number of impressions being sold, n is the number of slots in the publisher's system, and $r=\lambda/\mu$. We then obtain λ^*_{cpm} from the following maximization problem:

$$R_{\mathrm{cpm}}(\lambda^*_{\mathrm{cpm}})=\max_{\lambda_{\mathrm{cpm}}\geq 0}\left\{\lambda_{\mathrm{cpm}}(1-\mathbb{P}_n^{\mathrm{cpm}}(\lambda_{\mathrm{cpm}};\mu,N,n))p_{\mathrm{cpm}}(\lambda_{\mathrm{cpm}})N\right\}, \quad (7)$$

$$\text{subject to}\quad N=\left[\frac{x}{\mathrm{CTR}(\lambda_{\mathrm{cpm}};\hat{\mu},x,n,\beta)}\right], \quad (8)$$

where $[\cdot]$ in (8) refers to the integer sign and $\mathrm{CTR}(\lambda_{\mathrm{cpm}};\hat{\mu},x,n,\beta)$ is obtained by (6). (ii) We obtain the resulting revenues when the simple conversion rule is used. The revenue of the CPC system when the publisher uses the corresponding scaled CPM price is

$$R_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpm}})=\lambda^*_{\mathrm{cpm}}(1-\mathbb{P}_n^{\mathrm{cpc}}(\lambda^*_{\mathrm{cpm}};\hat{\mu},x,n))\cdot\left(\frac{p_{\mathrm{cpm}}(\lambda^*_{\mathrm{cpm}})}{\mathrm{CTR}(\lambda^*_{\mathrm{cpm}};\hat{\mu},x,n,\beta)}\right)x. \quad (9)$$

Equation (9) means that the publisher sells x clicks, each with a CPC price that is obtained by scaling the optimal CPM price by the observed CTR whereas the advertisers' effective arrival rate is $\lambda^*_{\mathrm{cpm}}(1-\mathbb{P}_n^{\mathrm{cpc}}(\lambda^*_{\mathrm{cpm}};\hat{\mu},x,n))$. Note that to make a relevant comparison, we choose the structure of the CPC price-demand relationship to match the one of the CPM setting. (iii) We use the "correct" approach by determining the optimal advertisers' arrival rate of the CPC system, λ^*_{cpc}, and by obtaining the optimal revenue as follows:

$$R^*_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpc}})=\max_{\lambda_{\mathrm{cpc}}\geq 0}\left\{\lambda_{\mathrm{cpc}}(1-\mathbb{P}_n^{\mathrm{cpc}}(\lambda_{\mathrm{cpc}};\hat{\mu},x,n))\cdot\left(\frac{p_{\mathrm{cpm}}(\lambda_{\mathrm{cpc}})}{\mathrm{CTR}(\lambda_{\mathrm{cpc}};\hat{\mu},x,n,\beta)}\right)x\right\}. \quad (10)$$

(iv) We obtain the relative revenue gap as $\mathrm{Gap}=(R^*_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpc}})-R_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpm}}))/R^*_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpc}})\times 100(\%)$.

Table 1 shows this gap for different numbers of slots and different numbers of clicks. As can be seen from the table, the relative revenue gap between the optimal and the simple conversion rule ranges from 2.8% at $n=8$ and $x=50{,}000$ to 10.3% for $n=2$ and $x=3{,}000$ (with an observed CTR of about 0.5%, which is relatively common in practice).

The main reason for the revenue gap is that the publisher uses a wrong CTR value as the service processes in the CPM, and CPC systems are different. This service difference affects the publisher's revenue through the full-state probability. The observed CTR in the CPM system depends on the optimal arrival rate at the CPM system. However, this arrival rate depends on $\mathbb{P}_n^{\mathrm{cpm}}$, which is different from $\mathbb{P}_n^{\mathrm{cpc}}$. Hence, the optimal arrival rate and the CTR in the scaled CPM system are not optimal in the CPC system, which leads to

Table 1 **The Relative Performance Gap** $((R^*_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpc}})-R_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpm}}))/R^*_{\mathrm{cpc}}(\lambda^*_{\mathrm{cpc}}))\times 100$ (%)

$c=0.5$	Number of slots (n) (%)						
Number of clicks (x)	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$	$n=8$
$x=3{,}000$	10.3	10.3	9.6	7.6	8.9	7.9	7.4
$x=5{,}000$	9.0	8.9	8.0	7.3	6.6	6.2	6.0
$x=10{,}000$	7.5	7.2	6.5	6.0	5.3	5.2	4.8
$x=20{,}000$	6.2	5.8	5.1	4.7	4.2	4.1	4.1
$x=50{,}000$	4.7	4.3	3.9	3.6	3.3	3.0	2.8

a revenue loss. As will be detailed later, publishers do not typically expect the CTR to change by only switching between different pricing schemes. Hence, they may find the CTR change somewhat surprising.

Nevertheless, if in some situation $\mathbb{P}_n^{\text{cpm}}$ and $\mathbb{P}_n^{\text{cpc}}$ are sufficiently close, the use of conversion approach is fine. The reason is that when $\mathbb{P}_n^{\text{cpm}}$ and $\mathbb{P}_n^{\text{cpc}}$ are close, the publisher is not much worse off by optimizing the scaled CPM problem (7) compared to solving (10). Some instances where $\mathbb{P}_n^{\text{cpm}}$ and $\mathbb{P}_n^{\text{cpc}}$ are sufficiently close include when (i) the publisher's page has many slots (i.e., $n \to \infty$), (ii) the number of clicks is sufficiently high (i.e., $x \to \infty$), (iii) the viewers' traffic rate is either quite high (i.e., $\mu \to \infty$) or low (i.e., $\mu \to \infty$), and (iv) the publisher's page has a single slot (i.e., $n = 1$). For example, when n increases as in case (i), $\mathbb{P}_n^{\text{cpm}}$ and $\mathbb{P}_n^{\text{cpc}}$ both become close to 0, leading the relative gap to reduce (see Table 1). Similarly, when x increases, $\mathbb{P}_n^{\text{cpm}}$ and $\mathbb{P}_n^{\text{cpc}}$ both become close to 1, i.e., the publisher's system behaves like an always full system. Hence, the relative gap is reduced.

As mentioned earlier, web publishers may be puzzled as they observe the CTR value is reduced when they switch the pricing scheme from CPC to CPM. Publishers often attribute the CTR reduction to ad networks' unshared policies, which somehow lead the ads to be more visible under the CPC scheme. In Proposition 10, we provide an alternative explanation for this behavior based on the difference in the service mechanism between the CPM and CPC systems.

Proposition 10. *In switching from the CPC to an equivalent CPM scheme* (*generating the same revenue*), *if*

$$\frac{p_{\text{cpc}}(\lambda^*_{\text{cpc}})x(1-\mathbb{P}_n^{\text{cpc}}(\lambda^*_{\text{cpc}}))}{p_{\text{cpm}}(\lambda^*_{\text{cpm}})N(1-\mathbb{P}_n^{\text{cpm}}(\lambda^*_{\text{cpc}}))} \geq 1,$$

then (i) $\lambda^*_{\text{cpc}} \leq \lambda^*_{\text{cpm}}$, *and* (ii) $\text{CTR}(\lambda^*_{\text{cpm}};\ \hat{\mu},\ x, n, \beta) \leq \text{CTR}(\lambda^*_{\text{cpc}};\ \hat{\mu}, x, n, \beta)$.

Part (i) states that when the expected revenue from selling x clicks to an advertiser is greater than from selling N impressions using an equivalent CPM scheme (i.e., generating identical revenue), then by switching from CPC to CPM the publisher optimally serves more advertisers to compensate the CPM revenue shortage per advertiser compared to CPC. Technically, $\lambda^*_{\text{cpc}} \leq \lambda^*_{\text{cpm}}$ could have been expected to occur more often since the feasible solution set for λ^*_{cpc} is typically much smaller than the one for λ^*_{cpm} (i.e., $0 \leq \lambda^*_{\text{cpc}} \leq \hat{\mu} = \beta\mu$, $0 \leq \lambda^*_{\text{cpm}} \leq \mu$). Part (ii) indicates that as more advertisers are admitted into the system, the system tends to become busier. Thus, each ad has on average a lower chance to be considered and clicked on. As a result, the *CTR* is decreased.

Practical Evidence. We end this section by explaining two pieces of recent practical evidence on the revenue and CTR gaps generated by changing the pricing schemes.

(i) In a recent experiment, Quadlin (2012) (an account manager at Hanapin Marketing, a CPC ad management company) utilized Google's CPC and CPM price bidding options to assess the revenue gained by the conversion approach compared to the optimal CPC pricing. In his experiment, Quadlin (2012) first posted a particular set of ads targeted at a certain audience for a specific length of time and determined the optimal CPC price directly by utilizing Google's CPC optimization option. The CPC price he obtained was \$1.69, whereas the observed CTR and the revenue at the end of the period were 0.12% and \$268.46, respectively. He then repeated the same campaign with the only difference that this time, instead of determining the CPC price directly, he utilized Google's CPM optimization option (he turned off the CPC optimization option) and determined the CPM price recommended by the software. He then divided the CPM price by the observed CTR. The resulting CPC price was \$3.91, and the CTR value substantially decreased to 0.03%. In addition, the campaign revenue at the end of the period decreased by 5.4%, compared to the first trial, to \$253.85. As a practitioner, Quadlin (2012) suggested that there was a possibility that Google was posting the ads on better spots when he used the CPC pricing directly (i.e., Google was prioritizing CPC contracts without sharing the information). We acknowledge that we are not aware of the details of how the Google optimization toolbox obtains the CPC and CPM prices. Nevertheless, our analysis provides an alternative explanation for the CTR behavior.

(ii) In a different experiment, Jordan (2011) (an online advertising practitioner) examined the impact of the change of the pricing scheme on the CTR with identical ads posted on identical pages in Facebook. He noticed that identical ads on identical pages with the CPC pricing scheme had a CTR of 0.177%, but when a CPM scheme was used, the CTR value reduced to 0.056%. Jordan (2011) could not identify any possible explanation for the CTR differences. He mainly questioned how the difference can be possible when identical ads are used on identical pages and concluded that Facebook was probably somehow prioritizing one pricing scheme to the other. Our analysis that the two settings have different service mechanisms, i.e., advertisers are served differently, together with Proposition 10, provide an alternative explanation for this phenomenon.

Jordan's observation of the change in the CTR value was also made by other practitioners, i.e., Hobokook (2013), LaGrange (2013), Ben (2013), and RileyPool (2010). Hobokook (2013) observed that switching from CPC to CPM for his identical ads posted on Facebook reduced the CTR value from 8.1% to 0.4%. Likewise, LaGrange (2013) observed that switching from CPC to CPM for identical ads caused a CTR reduction from

1% to 0.07%. In addition, Ben (2013), who had set up a CPC campaign in Facebook to a well-targeted group, stated that the CPC campaign yielded an observed CTR of 3%. However, he realized that the average CPM price through the CPC campaign was higher than the actual CPM price in the CPM campaign. He decided to switch to CPM to lower his costs. However, he observed that after switching, the CTR was reduced to about 1%. Ben (2013) could not explain the reason and mainly questioned whether Facebook has any policy to possibly manipulate the viewers' traffic upon switching. Our result, that switching from the CPC to CPM tends to reduce the CTR (Proposition 10), provides an alternative explanation based on the CPM and CPC different service mechanisms.

6. Extensions

There are several directions the CPC model can be extended to. In this section, we discuss three extensions, leaving the rest for future research.

6.1. Rotation of Ads

In the preceding analysis we have assumed that the publisher can serve up to n ads, but in reality the publisher can often serve more advertisers than there are slots. For example, two ads could share the same slot with each ad randomly displayed to the viewers based on preassigned display weights. Random weight-based ad rotation is commonly used by ad management software such as Double-Click for Publishers by Google or AdCycle ad management software. In this section, we extend the base model to accommodate random ad rotation. Consider that the n slots on the publisher's website can be randomly shared by up to S advertisers ($S \geq n$). To accommodate this modified setting into our base model, suppose that there are m ads present in the system, which are randomly rotated across the n slots. If $0 \leq m \leq n$ and a viewer arrives at the system, then all the ads are displayed to the viewer. However, if $n < m \leq S$, then a subset of the n ads are randomly selected. We know that the number of possible subsets to select n ads out of m is $\binom{m}{n}$. To obtain the number of subsets that include a particular ad, we select that particular ad and then choose the remaining $n-1$ ads in the subset from the $m-1$ remaining total ads. Hence, the probability that a particular ad is displayed is $\mathbb{P}_{\text{Disp}} = \binom{m-1}{n-1}/\binom{m}{n} = n/m$. Moreover, each ad in the system will be clicked with the probability $\mathbb{P}_{\text{Disp}} \times (1/n) = 1/m$. Therefore, the Markovian transitions of the system become identical to those of a system with no ad rotation and S slots. That is, the probability of having m advertisers in the system with random ad rotation would become $\mathbb{P}_m = (rx)^m / \sum_{i=0}^{S} (rx)^i$, $0 \leq m \leq S$. Thus, considering a system with n slots that are rotated by up to S ads randomly is the same as considering the base system considered in §3 with S ads and no ad rotation.

6.2. Multiple Pages and Types of Ads

Until now, we have considered that the publisher's system consists of a single page with equivalent ads. In this section, we extend the base model to consider multiple pages and different types of ads. To start, we assume the publisher's website contains J pages labeled from 1 to J. For example, for a news site these pages could correspond to the business page, travel page, etc. Each page can have several groups of ads where the same price is charged within each group. For example, the top of the page may display two equally sized ads, while several small ads may be placed at the bottom (rectangles). This would lead to two ad groups. More formally, for each page j we group the ads into M^j groups of equivalent slots, which we define as *subsystems*, where each subsystem m, $1 \leq m \leq M^j$, contains $n^{j,m}$ equivalent slots. Advertisers arrive requesting a slot in subsystem (j, m) according to a Poisson process with rate $\lambda^{j,m}$. An advertiser requesting a slot in group m on page j requests his ad to be posted on the website until clicked $x^{j,m}$ times by the viewers and is charged $p^{j,m}$ dollars per click. Note that since ad networks often contain thousands of websites, it is rare that an advertiser's desired slot is unavailable. Hence, existence of substitute subsystems in the network does not affect advertisers' demand after they choose their primary subsystem to target.

We assume that viewers arrive at the publisher's website with rate μ and consider subsystem (j, m) with probability ϖ_{jm}. Having considered subsystem (j, m), a viewer clicks on an ad in the subsystem with probability $\beta^{j,m}$ or leaves the publisher's system without clicking on any ad with probability $1-\beta^{j,m}$. As mentioned before, viewers consider only real ads. If viewers consider subsystem (j, m) but it has only filler ads, we assume that they consider the ads in another subsystem (g, h), (the ads in subsystem g on page h) with probability $\alpha_{j,m}^{g,h}$, or leave the website without considering any ads with probability $1 - \sum_{\{j,m\} \neq \{g,h\}} \alpha_{j,m}^{g,h}$. In reality, a viewer of subsystem (j, m) may decide to visit (g, h) without clicking on any ad in (j, m) even if (j, m) has revenue-generating ads. Likewise, the viewer may decide to visit subsystem (g, h) after his initial click on a revenue-generating ad in subsystem (j, m). However, it is clear to see that considering these two scenarios adds several layers of complexities beyond the scope of this paper. Hence, we restrict our focus to the case where viewers of subsystem (j, m) visit some other subsystem, e.g., (g, h), if (j, m) has filler ads only. Following §6.1, we assume that up to $S^{j,m} \geq n^{j,m}$ real ads can be randomly displayed by $n^{j,m}$ slots in subsystem (j, m). We denote the price-demand function for subsystem (j, m) by $p^{j,m}(\lambda^{j,m}, x^{j,m}, S^{j,m}, n^{j,m})$, which is decreasing with respect to $\lambda^{j,m}$, $x^{j,m}$, $S^{j,m}$, and $n^{j,m}$. Let $\mathbb{P}_i^{j,m}$, $i \in \{0, \ldots, S^{j,m}\}$ be the probability of having i advertisers

in subsystem (j, m). The optimization problem for J web pages and the different types of ads on each page can be formulated as

$$\max_{\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_J} \Bigg\{ R(\mathbf{\Lambda}_1,\ldots,\mathbf{\Lambda}_J) = \sum_{j=1}^{J}\sum_{m=1}^{M^j} \lambda^{j,m}\big(1-\mathbb{P}^{j,m}_{S^{j,m}}(\lambda^{j,m}; X^{j,m}, S^{j,m}, \beta^{j,m}, \hat{\mu}^{j,m})\big) \cdot p^{j,m}(\lambda^{j,m}, x^{j,m}, S^{j,m}, n^{j,m}) x^{j,m} \Bigg\}$$

subject to

$$\mathbf{\Lambda}_j = (\lambda^{j,1},\ldots,\lambda^{j,M^j})^t \in [0,+\infty)^{M^j}, \quad j=1,\ldots,J. \quad (11)$$

In this formula $\hat{\mu}^{j,m} = \beta^{j,m}\mu^{j,m}$, and $\mathbb{P}^{j,m}_{S^{j,m}}$ is the full-state probability of the subsystem (j, m), which is given by $\mathbb{P}^{j,m}_{S^{j,m}} = (r^{j,m}x^{j,m})^{S^{j,m}} / \sum_{i=0}^{S^{j,m}} (r^{j,m}x^{j,m})^i$, $r^{j,m} = \lambda^{j,m}/\mu^{j,m}$. In addition, $\mu^{j,m}$ is the viewers' effective arrival rate at subsystem (j, m), which includes the two arrival streams: (i) from viewers who consider subsystem (j, m) as their first choice and (ii) those viewers redirected to (j, m) from other subsystems after being faced by filler ads. The following proposition gives the value of $\mu^{j,m}$ as a function of other parameters in the publisher's model.

Proposition 11. *The effective overall viewers' arrival rate at subsystem (j, m), $\mu^{j,m}$, is given by*

$$\mu^{j,m} = \mu \frac{\varpi_{j,m} + \sum_{g=1}^{J}\sum_{h=1}^{M^g} \varpi_{g,h}\, \mathbb{P}_0^{g,h}\, \alpha_{g,h}^{j,m}}{1 - \sum_{g=1}^{J}\sum_{h=1}^{M^g} \alpha_{j,m}^{g,h}\alpha_{g,h}^{j,m}\, \mathbb{P}_0^{j,m}\, \mathbb{P}_0^{g,h}}, \quad 1\le j\le J,\ 1\le m\le M^j,\ (g,h)\neq(j,m), \quad (12)$$

where μ is the rate at which viewers consider all subsystems (the traffic rate), $\mathbb{P}_0^{j,m}$ is the probability that subsystem (j, m) is empty, and $\alpha_{j,m}^{g,h}$ ($\alpha_{g,h}^{j,m}$) is the fraction of viewers that faced filler ads in subsystem (j, m) (subsystem (g, h)), who then approach subsystem (g, h) (subsystem (j, m)).

From (12), we note that the viewers' arrival rate at the subsystem (j, m) is the lowest when $\mu^{j,m} = \mu\varpi_{j,m}$. This happens when the probability of an empty system is very small and thereby, almost no interflows exist between subsystems. In such a case, each subsystem (j, m) can be considered separately, and all results of §§(3) and (4) would hold for the general system as well.

6.3. Non-Poisson Arrivals

In §3 we assumed that the advertisers' arrivals at the web publisher from the ad network follow a Poisson process, which might not be the case in reality. In addition, the viewers' arrival process might not be Poisson either while the number of requested clicks is not necessarily the same for all advertisers. In this section, we explore other distributions for both the demand and supply sides while considering each advertiser requesting a random number of clicks.

In our simulation study, we specifically examine the amount of revenue a publisher can lose by using the base model's solution obtained in §4 (based on Poisson arrivals, a single number of clicks offered, and a single price charged) to determine the CPC price, whereas the clicks requested are random across advertisers and both the advertisers' and viewers' arrivals are non-Poisson.

We consider one subsystem without rotation and let $\mu = 100$, $\beta = 0.01$ so that $\hat{\mu} = 1$. For the advertisers' interarrival time distributions, we consider the following: Normal with mean $1/\lambda$ and standard deviation $1/\lambda$, Erlang-2 with mean $1/\lambda$ and standard deviation $1/\sqrt{2}\lambda$, Erlang-4 with mean $1/\lambda$ and standard deviation $1/2\lambda$, Uniform with the two parameters 0 and $2/\lambda$, Exponential with rate λ, and finally Deterministic arrivals. For the viewers' interarrival time distributions, we consider the same distributions, with λ replaced by $\hat{\mu}$. The reason for the particular mean and variance choices is to be able to make relevant comparisons between the different non-Poisson distributions and Poisson. That is, we assume that the publisher is recording the values for the advertisers' and viewers' arrival means and variances correctly, but it wrongly considers that the arrival processes are both Poisson but they are not. Similar mean and variance choices were used by, e.g., Gross (1975) as well, who conducted an extensive numerical study to compare the average system sizes of $M/M/1$, $M/G/1$, and $G/G/1$.

The number of slots is set to be $n = 4$. We choose the pricing function to be $p(\lambda, X) = 0.5 - \lambda^{0.8} - 10^{-6}X$ where the random number of requested clicks X follows a truncated normal distribution with mean $x = \mathbb{E}(X) =$ 1,000 and standard deviation 500. Each simulation is run for 10,000 time units. The steps of each simulation are the following: (i) We obtain the advertisers' optimal arrival rate $\lambda^*_{X,\mathbf{D}_1,\mathbf{D}_2}$ when the advertisers' interarrival times follow the distribution $\mathbf{D}_1$, the viewers' interarrival times follow $\mathbf{D}_2$, and each advertiser requests a random number of clicks X. This includes simulating the publisher's system for a range of values of λ and then selecting $\lambda^*_{X,\mathbf{D}_1,\mathbf{D}_2}$, the rate that gives the maximum simulated revenue. We represent the simulated revenue corresponding to $\lambda^*_{X,\mathbf{D}_1,\mathbf{D}_2}$ with $R_{X,\mathbf{D}_1,\mathbf{D}_2}(\lambda^*_{X,\mathbf{D}_1,\mathbf{D}_2})$. (ii) We obtain the optimal value for λ using the analytical solution given by Equation (1) with Poisson demand of advertisers and viewers, where the number of clicks is the same for all advertisers, i.e., $x = 1{,}000$. We denote this optimal value with $\lambda^*_{x,\mathrm{Exp}}$. If the web publisher used our analytical solution with the deterministic demand x, for a system that does not have Poisson arrivals of advertisers and viewers and each advertiser requests X clicks, its "actual/realized" revenue would become the simulated

Table 2 The 95% Confidence Interval for the Relative Performance Gap (%)

Dist.	Viewers					
Adv	*E*-2	*E*-4	*N*	*U*	*D*	*Ex*
E-2	1.13 ± 0.45	1.07 ± 0.47	1.27 ± 0.54	2.07 ± 1.15	1.76 ± 0.77	0.88 ± 0.61
E-4	0.80 ± 0.31	1.09 ± 0.60	2.20 ± 0.94	1.35 ± 0.62	1.24 ± 0.53	1.02 ± 0.59
N	1.16 ± 0.69	1.12 ± 0.42	0.96 ± 0.47	1.47 ± 0.93	0.64 ± 0.46	0.99 ± 0.63
U	0.91 ± 0.47	0.99 ± 0.53	1.18 ± 0.72	1.00 ± 0.62	1.02 ± 0.41	1.01 ± 0.49
D	1.08 ± 0.43	0.84 ± 0.40	1.13 ± 0.52	0.92 ± 0.42	0.93 ± 0.52	0.41 ± 0.30
Ex	1.31 ± 0.71	1.88 ± 1.32	1.53 ± 0.88	1.69 ± 0.67	1.31 ± 0.79	—

revenue corresponding to $\lambda^*_{x,\text{Exp}}$, i.e., $R_{X,\mathbf{D}_1,\mathbf{D}_2}(\lambda^*_{x,\text{Exp}})$. (iii) We obtain the revenue gap as $((R_{X,\mathbf{D}_1,\mathbf{D}_2}(\lambda^*_{X,\mathbf{D}_1,\mathbf{D}_2}) - R_{X,\mathbf{D}_1,\mathbf{D}_2}(\lambda^*_{x,\text{Exp}}))/R_{X,\mathbf{D}_1,\mathbf{D}_2}(\lambda^*_{X,\mathbf{D}_1,\mathbf{D}_2}))(\%)$. (iv) We replicate steps (i)–(iii) 30 times and form a 95% confidence interval for each revenue gap.

Table 2 shows the 95% confidence intervals for the relative revenue gaps for the different interarrival time distributions considered for advertisers' and viewers' arrivals as well as the random number of requested clicks X, which results in generating instantaneously adjusted prices for each click request. We observe that the mean revenue gaps are relatively small, between 0.80%–2.2%.

We note that since the revenue depends on the full-state probability, $\mathbb{P}^{\text{cpc}}_n$, a small revenue gap means that $\mathbb{P}^{\text{cpc}}_n$ may not be sensitive to the forms of $\mathbf{D}_1$ and $\mathbf{D}_2$. One possible explanation for the $\mathbb{P}^{\text{cpc}}_n$'s robustness is that its structure is somewhat similar to Erlang loss formula, $\mathbb{P}^{\text{Er}}_n = ((rx)^n/n!)/(\sum_{j=0}^{n}(rx)^j/j!)$. In fact, when $x \to \infty$ or $n \to \infty$, both formulas converge together as both approach either 1 or 0. In addition, Erlang has this property that its structure is independent of the form of $\mathbf{D}_2$. Nevertheless, although the similarity of $\mathbb{P}^{\text{cpc}}_n$'s formula to Erlang is helpful in explaining its insensitivity to $\mathbf{D}_2$, it does not explain the reason for its insensitivity to $\mathbf{D}_1$. The second explanation is based on the possible reversibility of the CTMC defined for the publisher's system with general $\mathbf{D}_1$ and $\mathbf{D}_2$. This possible reversibility feature causes much of $\pi_{\mathbf{k}}$'s product-form structure to be insensitive to $\mathbf{D}_1$ and $\mathbf{D}_2$. To invoke the theory, we note that since the publisher's system is not generally Markovian when $\mathbf{D}_1$ and $\mathbf{D}_2$ are not Poisson, we need to define a suitable Markov process in the context of this general system. Consider the new state vector $\mathbf{k}' = (k_1, \ldots, k_n)'$, which is similar to $\mathbf{k}$ with the difference that the place of each component is important. For example, $(1, 2, 4)'$ and $(4, 1, 2)'$ refer to two different system states. Then the relationship between the steady-state probabilities $\pi_{\mathbf{k}}$ and $\pi_{\mathbf{k}'}$ is $\pi_{\mathbf{k}} = (\sum_{q=1}^{Q}|\mathcal{G}_{c_q}(\mathbf{k})|)!/\prod_{q=1}^{Q}|\mathcal{G}_{c_q}(\mathbf{k})|!\pi_{\mathbf{k}'}$ where $(\sum_{q=1}^{Q}|\mathcal{G}_{c_q}(\mathbf{k})|)!/\prod_{q=1}^{Q}|\mathcal{G}_{c_q}(\mathbf{k})|!$ accounts for all the possible rearrangements for the n occupied similar slots when they consist of Q groups with each slot in group $0 \le q \le Q$ having c_q remaining clicks. Now, we define a proper Markov process by expanding $\mathbf{k}'$ to $\mathbf{k}'_{\mathbf{t}} = (k_1^{t_1}, k_2^{t_2}, \ldots, k_n^{t_n})'$ where t_i refers to the *completed time* after the $(x - k_i)$th click was made on slot i while the remaining clicks are $k_i > 0$. It can be verified that $\mathbf{k}'_{\mathbf{t}}$ is Markovian because its future state is a function of only its current position. We conjecture that $\mathbf{k}'_{\mathbf{t}}$ process is *reversible*. (Showing the reversibility property analytically is hard, and we can only conjecture that this property holds.) With reversibility comes the critical observation that the limiting joint distribution of $(k_1^{t_1}, k_2^{t_2}, \ldots, k_n^{t_n})'$ has the proportional product form: $\pi_{\mathbf{k}'_{\mathbf{t}}} = Ca_n\bar{B}_{k_1}(t_1)\bar{B}_{k_2}(t_2)\ldots\bar{B}_{k_n}(t_n)$, where C is a constant proportional to zero-state probability, and a_n is proportional to the probability that the n slots are occupied. In addition, B_{k_i} is the service-time distribution function given that the next click is made to slot i (thus $\bar{B}_{k_i}(t_i)$ is the conditional probability that if slot i chosen next, the time that it takes for the next click to be delivered to slot i is at least t_i). Then, the marginal limiting distribution $\pi_{\mathbf{k}'} = \pi(k_1, k_2, \ldots, k_n)'$ can be obtained by taking integrals with respect to $\{t_1, \ldots, t_n\}$ from both sides. That is, $\pi_{\mathbf{k}'} = \pi(k_1, k_2, \ldots, k_n)' = Ca_n\prod_{i=1}^{n}(\int_0^\infty \bar{B}_{k_i}(t_i)\,dt_i)$. $\int_0^\infty \bar{B}_{k_i}(t_i)\,dt_i$ is the expected time for the next click to be delivered to slot i given that it is clicked next. Thus, $\int_0^\infty \bar{B}_{k_i}(t_i)\,dt_i = 1/\mu$. Hence, $\pi(k_1, k_2, \ldots, k_n)'$ with general arrivals and the service processes becomes $\pi_{\mathbf{k}'} = Ca_n(1/\mu)^n$. Therefore,

$$\pi_{\mathbf{k}} = \frac{(\sum_{q=1}^{Q}|\mathcal{G}_{c_q}(\mathbf{k})|)!}{\prod_{q=1}^{Q}|\mathcal{G}_{c_q}(\mathbf{k})|!}Ca_n\left(\frac{1}{\mu}\right)^n. \tag{13}$$

Note that the product-form structure of the formula in (13) is remarkably similar to the closed-form solution for $\pi_{\mathbf{k}}$ when the advertisers' and viewers' processes are Poisson. This similarity (given the possible reversibility) may theoretically explain why the full-state probability is robust to the forms of $\mathbf{D}_1$ and $\mathbf{D}_2$.

7. Conclusion

Optimal pricing of display ads has received minimal attention in operations and management science literature. This paper attempts to fill this gap. We present a revenue optimization model for a web publisher selling its advertising space through an ad network. The web publisher generates revenue by displaying

ads on its website and charges according to the CPC pricing scheme that it needs to optimize.

We model the web publisher's operations with a queueing system, where the arrival process corresponds to ads sent to the web publisher from the ad network, the service process corresponds to the viewers visiting the website, and the advertising slots correspond to the servers. The queueing model developed is new. Despite the complexity of the dynamics, we derive the closed-form solution for the steady-state probability distribution of the number of advertisers. This enables us to set up the revenue-maximizing problem of the web publisher and derive the optimal price to charge per click.

On the managerial side, we show that the simple conversion rule widely employed in practice, where a publisher uses the CTR to convert the CPM price to the CPC price can be misleading, resulting in a considerable revenue loss compared to obtaining and using the optimal CPC price. In addition, we provide further insights by showing that the optimal CPC price may increase as the publisher increases the advertising slots. This behavior may not seem intuitive in comparison to our common intuition from the supply-demand relationship, since an increase in the advertising slots can be interpreted as an increase in the service capacity of the system.

The results of this paper can be extended in several directions. First, given that publishers and advertisers can keep track of viewers easily, targeted advertising would be very interesting to advertisers and publishers enabling publishers to charge higher prices for a more targeted audience. Second, in this paper we have not considered the competition of web publishers. Exploring competition between two or more publishers offering CPC prices with symmetric or asymmetric information structures would be another interesting research direction. Finally, we have analyzed the publisher's cost-per-click operation from a steady-state standpoint. Dynamic pricing would also be interesting and possible to implement because advertisers often buy advertising spaces online, so it makes sense to publishers to change the prices dynamically.

Supplemental Material

Supplemental material to this paper is available at http://dx.doi.org/10.1287/msom.2014.0491.

Acknowledgments

The authors are grateful to two anonymous referees, an associate editor, and the editor for many constructive suggestions on earlier drafts of this work. They also thank Michael Harrison, Mor Harchol-Balter, and William Massey for their helpful comments and insights, as well as the participants of Stanford GSB Dynamic Pricing and Revenue Management Ph.D. seminar course.

References

Akella R, Broder A, Josifovski V (2009) Introduction to computational advertising. Course Lectures, University of California, Santa Cruz.

Alaei S, Arcaute E, Khuller S, Ma W, Malekian A, Tomlin J (2009) Online allocation of display advertisements subject to advanced sales contracts. *Proc. Third Internat. Workshop on Data Mining and Audience Intelligence for Advertising* (ACM, New York), 69–77.

Anandan R (2012) 3 predictions for display ads this year. *Bus. Standard* (March 7), http://www.business-standard.com/article/management/3-preditions-for-display-ads-this-year-112030700006_1.html.

Araman VF, Fridgeirsdottir K (2011) A uniform allocation mechanism and cost-per-impression pricing for online advertising. Working paper, American University of Beirut, Beirut, Lebanon.

Araman VF, Popescu I (2010) Media revenue management with audience uncertainty: Balancing upfront and spot market sales. *Manufacturing Service Oper. Management* 12(2):190–212.

Ata B (2005) Dynamic power control in a wireless static channel subject to a quality-of-service constraint. *Oper. Res.* 53(5):842–851.

Ata B, Shneorson S (2006) Dynamic control of an $m/m/1$ service system with adjustable arrival and service rates. *Management Sci.* 52(11):1778–1791.

Balseiro S, Feldman J, Mirrokni V, Muthukrishnan S (2011) Yield optimization of display advertising with ad exchange. *Proc. 12th ACM Conf. on Electronic Commerce* (ACM, New York), 27–28.

Balseiro S, Besbes O, Weintraub GY (2013) Auctions for online display advertising exchanges: Approximations and design. *Proc. 14th ACM Conf. on Electronic Commerce* (ACM, New York), 53–54.

Baye MR, Morgan J (2000) A simple model of advertising and subscription fees. *Econom. Lett.* 69(3):345–351.

Ben L (2013) Facebook CPM vs CPC vs CPA inventory. Accessed March 15, 2014, http://www.blackhatworld.com/.

Cao J, Cleveland WS, Lin D, Sun DX (2003) Internet traffic tends toward Poisson and independent as the load increases. *Nonlinear Estimation and Classification* (Springer, New York), 83–109.

Celis LE, Lewis G, Mobius MM, Nazerzadeh H (2012) Buy-it-now or take-a-chance: Price discrimination through randomized auctions. NBER Working Paper w18590, National Bureau of Economic Research, Cambridge, MA.

Chen YJ (2011) Optimal dynamic auctions for display advertising. Working paper, University of California, Berkeley, Berkeley.

Chickering DM, Heckerman D (2003) Targeted advertising on the web with inventory management . *Interfaces* 33(5):71–77.

Ciocan DF, Farias V (2012) Model predictive control for dynamic resource allocation. *Math. Oper. Res.* 37(3):501–525.

Danaher PJ (2007) Modeling page views across multiple websites with an application to internet reach and frequency prediction. *Marketing Sci.* 26(3):422–437.

Edelman B, Ostrovsky M, Schwarz M (2007) Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *Amer. Econom. Rev.* 97(1):242–259.

Fjell K (2009) Online advertising: Pay-per-view versus pay-per-click—A comment. *J. Revenue and Pricing Management* 8(2/3):200–206.

Fredricksen C (2011) Online advertising market poised to grow 20% in 2011. *eMarketer* (June 8), http://emarketer.com/newsroom/index.php/online-advertising-market-poised-grow-20-2011/.

Fridgeirsdottir K, Najafi-Asadolahi S (2013) Cost-per-impression pricing for display advertising. Working paper, London Business School, London.

George JM, Harrison JM (2001) Dynamic control of a queue with adjustable service rate. *Oper. Res.* 49(5):720–731.

Ghosh A, McAfee P, Papineni K, Vassilvitskii S (2009) Bidding for representative allocations for display advertising. *Internet and Network Economics* (Springer, Berlin), 208–219.

Gomes R, Immorlica N, Markakis E (2009) Externalities in keyword auctions: An empirical and theoretical assessment. *Internet and Network Economics* (Springer, Berlin), 172–183.

Gross D (1975) Sensitivity of queueing models to the assumption of exponentiality. *Naval Res. Logist. Quart.* 22(2):271–287.

Gross D, Harris CM (1998) *Fundamentals of Queueing Theory* (Wiley, New York).

Hobokook K (2013) Fb ads—CTR got killed after switch. *Warrior Forum* (blog) (May 5), http://www.warriorforum.com/.

Hoffman DL, Novak TP (2000) Advertising pricing models for the World Wide Web. *Internet Publishing and Beyond: The Economics of Digital Information and Intellectual Property* (MIT Press, Cambridge, MA).

Interactive Advertising Bureau (2011) IAB Internet advertising revenue report: 2011 first six months results. Report, IAB and Price WaterhouseCoopers. http://www.iab.net/media/file/IAB-HY-2011-Report-Final.pdf.

Jagerman DL, Bhaskar S (1991) The $g/m/1$ processor-sharing queue and its heavy traffic analysis. *Comm. Statist. Stochastic Models* 7(3):379–395.

Jeziorski P, Segal IR (2009) What makes them click: Empirical analysis of consumer demand for search advertising. Working paper, Stanford University, Stanford, CA.

Jordan B (2011) The ultimate case study: Facebook CPC vs CPM bidding—Identical Facebook ads yield 0.177% vs 0.056% CTR. How? *StackThatMoney* (blog) (April 13), http://blog.stackthatmoney.com/2011/04/13/the-ultimate-case-study-facebook-cpc-vs-cpm-bidding/.

Kumar S, Sethi SP (2009) Dynamic pricing and advertising for web content providers. *Eur. J. Oper. Res.* 197(3):924–944.

Kumar S, Jacob VS, Sriskandarajah C (2006) Scheduling advertisements on a web page to maximize revenue. *Eur. J. Oper. Res.* 173(3):1067–1089.

LaGrange A (2013) Why does the CTR drop when I change the bidding of my advert from CPC to CPM? Accessed February 20, 2014, https://www.facebook.com/help/community/question/?id=10151886726317577.

Lewis RA, Reiley DH (2011) Does retail advertising work? Measuring the effects of advertising on sales via a controlled experiment on Yahoo! CCP Working Paper 11-9, Centre for Competition Policy, Norwich, UK.

Mangani A (2004) Online advertising: Pay-per-view versus pay-per-click. *J. Revenue Pricing Management* 2(4):295–302.

McAfee P, Papineni P, Vassilvitskii S (2013) Maximally representative allocations for guaranteed delivery advertising campaigns. *Rev. Econom. Design* 17(2):83–94.

Muthukrishnan S (2009) Ad exchanges: Research issues. Leonardi S, ed. *Internet and Network Economics*, Lecture Notes in Computer Science, Vol. 5929 (Springer, Berlin), 1–12.

Peterson T (2011) Display ad spending to hit $200B, says Google VP. *Direct Marketing News* (June 9), http://www.dmnews.com/display-ad-spending-to-hit-200b-says-google-vp/article/204870/.

Prasad A, Mahajan V, Bronnenbert B (2003) Advertising versus pay-per-view in electronic media. *Internat. J. Res. Marketing* 20(1):13–30.

Puha AL, Stolyar AL, Williams RJ (2006) The fluid limit of an overloaded processor sharing queue. *Math. Oper. Res.* 31(2):316–350.

Quadlin S (2012) Is CPM bidding a waste of your money? *PPC Hero* (blog) (November 7), http://www.ppchero.com/is-cpm-bidding-a-waste-of-your-money/.

RileyPool (2010) Case study: Facebook ads—CPC versus CPM ($2163.08 spent). (blog) (November 17), http://rileypool.com/ppc/case-study-facebook-ads-cpc-versus-cpm-2163-08-spent.

Savin SV, Cohen MA, Gans N, Katalan Z (2005) Capacity management in rental businesses with two customer bases. *Oper. Res.* 53(4):617–631.

Talluri K, van Ryzin G (2004) *The Theory and Practice of Revenue Management* (Kluwer Academic Publishers, Norwell, MA).

Turner J, Scheller-Wolf A, Tayur S (2011) Scheduling of dynamic in-game advertising. *Oper. Res.* 59(1):1–16.

Van Mieghem JA (2000) Price and service discrimination in queuing systems: Incentive compatibility of $Gc\mu$ scheduling. *Management Sci.* 46(9):1249–1267.

Vranica S (2013) Automated ad buying surges online. *Wall Street Journal* (October 13), http://online.wsj.com/.

Yang J, Vee E, Vassilvitskii S, Tomlin J, Shanmugasundaraml J, Anastasakos T, Kennedy O (2010) Inventory allocation for online graphical display advertising. Technical report, Yahoo! Labs, Sunnyvale, CA.

Zwart AP, Boxma OJ (2000) Sojourn time asymptotics in the $M/G/1$ processor sharing queue. *Queueing Systems* 35(1–4):141–166.