UC Irvine
UC Irvine Previously Published Works

Title
Planning Online Advertising Using Gini Indices

Permalink
https://escholarship.org/uc/item/6hx3s47w

Authors
Lejeune, Miguel
Turner, John

Publication Date
2015

DOI
10.2139/ssrn.2702590

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

Peer reviewed

eScholarship.org Powered by the California Digital Library
University of California

Planning Online Advertising Using Gini Indices

December 14, 2018

Miguel A. Lejeune
Department of Decision Sciences, George Washington University, Washington, DC 20052, mlejeune@gwu.edu

John Turner
The Paul Merage School of Business, University of California at Irvine, Irvine, CA 92697, john.turner@uci.edu

We study an online display advertising planning problem in which advertisers' demands for ad exposures (impressions) of various types compete for slices of shared resources, and advertisers prefer to receive impressions that are evenly-spread across the audience segments they target. We use the Gini coefficient measure and formulate an optimization problem that maximizes spreading of impressions across targeted audience segments while limiting demand shortfalls. First, we show how Gini-based metrics can be used to measure spreading that publishers of online advertising care about, and how Lorenz curves can be used to visualize Gini-based spread so that managers can effectively monitor the performance of a publisher's ad delivery system. Second, we adapt an existing ad planning model to measure Gini-based spread across audience segments, and compare and contrast our model to this baseline with respect to key properties and the structure of the solutions they produce. Third, we introduce a novel optimization-based decomposition scheme which efficiently solves our instances of the Gini-based problem up to 60 times faster than the commercial solver CPLEX solves a basic formulation directly. Finally, we present a number of model and algorithmic extensions, including (1) an online algorithm which mirrors the structure of our decomposition method to serve well-spread ads in real-time, (2) a model extension which allows an aggregator buying impressions in an external market to allocate them to advertisers in a well-spread manner, and (3) a multi-period model and decomposition method which spreads impressions across both audience segments and time.

Key words: Online Advertising, Gini Index, Lorenz Curve, Decomposition Method, Spreading Impressions

1. Introduction

Online advertising continues to be a fast-growing market in the United States, having grown 21.64% in 2016 to reach $72.5 billion annually (c.f., Internet Advertising Bureau 2017). Forty-four percent ($31.7B) of this market is classified as *display advertising*, a category that includes banner ads, digital video ads, rich media, and sponsorships; moreover, one-third of all online advertising ($25B) is impression-based, the majority of which is display advertising. Publishers (e.g., Google, Yahoo, Facebook) must choose, at each point in time, which ad to show to which arriving viewer – a challenging problem in which advertisers' demands for *impressions* (ad exposures) of specific types compete for slices of shared resources (the supply of impression opportunities from viewers of specific types). In many cases, advertisers prefer to receive impressions that are evenly-spread across the audience segments they target, as well as evenly-spread across time. Gini coefficients and their distributional

analogs, Lorenz curves, are well-studied in the field of economics for the purpose of measuring how well-spread resources, such as wealth and income, are across individuals in a population; however, to the best of our knowledge, there is no study that uses Gini-based metrics to measure the performance of online advertising campaigns. Our goal is to (1) propose that Gini-based metrics can be used to measure spreading that publishers of online advertising care about; (2) show how a popular model for spreading impressions across audience segments from the ad planning literature can be adapted to measure Gini-based performance; (3) compare and contrast our model to this baseline with respect to key properties and the structure of the solutions they produce, including how well revenue can be traded off to receive better spreading; and (4) introduce a novel optimization-based decomposition scheme based on Dantzig-Wolfe decomposition and subgradient optimization to efficiently exploit the nonlinear structure of our Gini objective to solve our proposed formulation on average 60 times faster than solving a simpler linear programming reformulation directly. By using Gini-based performance metrics and Lorenz curves, we argue that managers can better visualize and understand the performance of a publisher's ad delivery system.

Our paper is organized as follows. We begin by providing the requisite background in online advertising and economics, and review the relevant literature. Then, we discuss how Gini coefficients and Lorenz curves can be used by both publishers and advertisers. Next, we develop Gini-based metrics for measuring impression spread across audience segments, and formulate a single-period ad planning problem which maximally spreads impressions across targeted audience segments while minimizing demand shortfalls. We then define five key properties and show that our Gini-based model satisfies each. Using numerical experiments, we compare and contrast the solutions from our Gini-based model with a baseline model developed at Yahoo, and show that our Gini-based model is generally better at trading off revenue (i.e., incurring some demand shortfalls) to achieve better spread. Then, we introduce our novel decomposition method, evaluate its computational performance, and describe the structure of the solutions that it produces. Finally, we extend our model and solution method so it can be used (1) to solve the multi-period case which spreads impressions across both audience segments and time, (2) to help an aggregator decide which impressions to buy from the market to allocate to advertisers, and (3) to serve well-spread ads in real-time, using an online algorithm that works in conjunction with our decomposition method.

2. Background

2.1. Advertising Background

There are many forms of targeted display advertising, from banner ads on websites, to digital video ads, dynamic in-game ads placed in video games, ads in social networks, and promotional ads in mobile apps. Despite this, it is usually possible to classify an ad campaign as either impression-based, click-based, or conversion-based; and, orthogonally, guaranteed or non-guaranteed. Our focus in this paper is on guaranteed impression-based ads.

Impression-based ads, such as banner, dynamic in-game, and digital video ads, incur a cost to the advertiser whenever the publisher places the ad on some viewer's screen. In contrast, click-based and conversion-based ads only charge the advertiser for clicks and conversions (installing an app, or buying a product), respectively. Although our focus is impression-based ads, in many cases the models we describe can be adapted for click and conversion-based advertising (e.g., by introducing scaling factors to convert from impressions to clicks).

Guaranteed ads are purchased by advertisers from publishers in advance, and involve a promise to deliver a given number of impressions over a particular time window; see Bharadwaj et al. (2010, 2012), Chen et al. (2012), Turner et al. (2011), Turner (2012), Hojjat et al. (2017). In contrast, non-guaranteed ads are more opportunistic, and can use complex strategies to decide if, when, where, and to whom they are shown; see Muthukrishnan (2009), Chakraborty (2010), Dütting et al. (2011), Yuan et al. (2013), Balseiro et al. (2015a), Golrezaei et al. (2017). Many publishers manage both guaranteed and non-guaranteed ads, and researchers have studied how to optimally apportion exposures across these two main channels; see Araman and Popescu (2010), Chen (2013), Balseiro et al. (2014). We focus on guaranteed ads because they are commonly used for branding initiatives where spreading impressions across audience segments and time are important, and because in this context it is natural for publishers to use math programs to explicitly allocate a number of impressions of each audience segment to each advertiser. Nevertheless, it is worth pointing out that although non-guaranteed ads lack explicit impression goals and are typically allocated via real-time auctions, advertisers increasingly enlist the help of intermediaries to manage their bidding strategies (Balseiro et al. 2015b, Allouah et al. 2017), and some of these intermediaries use mathematical programming techniques to manage their clients' (the advertisers') budgets over time. Thus, Gini-based metrics could be used to serve non-guaranteed ads as well.

Multiple papers have studied the operational question of how to generate well-spread ad allocation plans for multiple advertisers. In one research stream, the focus is on producing so-called *representative* allocations, i.e., the goal is to spread each advertiser's impressions across targeted audience segments such that all opportunities for the advertiser's ad to be displayed have the same probability of showing the ad. This is the strategy used by major websites historically (c.f., McAfee et al. 2013), and provides quality impressions to advertisers who

implicitly want a broad swath of their targeted audience, otherwise they would have chosen narrower targeting criteria. That is, if an advertiser asks for his ad to be shown nationwide across the U.S.A., he would not be happy to get all his impressions from women in Florida. Receiving a narrower level of targeting than requested could be undesirable by the advertiser, for at least three reasons. First, the advertiser may have stocked product in stores across the country; in this case, limiting ads to one state has high mismatch costs and is operationally ineffective. Second, the advertiser may worry that the publisher could choose to satisfy its campaign in the cheapest way possible, i.e., by showing its ad to the least-sought-after (thus cheapest) subset of American women; in this case, the advertiser's request to evenly-spread impressions over its full target audience keeps any potentially misaligned publisher incentives in check. Finally, when traffic is stochastic and the allocation plan is used to serve impressions over a single period without re-solving the plan, Turner (2012) showed that representative solutions can minimize the variance of the number of impressions delivered while maximizing the number of unique individuals that see the advertiser's ad.[1] Note that while representative allocations spread impressions *within* a target market chosen by the advertiser, this is very different from mass marketing, which does not restrict ads to targeted audience segments. As well, representative allocations do not give each audience segment the same number of impressions; rather, larger segments are awarded proportionally more impressions than smaller ones.

In general, the conflicting impression demands of advertisers make achieving perfectly-representative solutions nearly impossible. In practice, maximally representative solutions are typically produced by minimizing a quadratic function that penalizes the L2-distance from the perfectly-representative solution, while satisfying supply constraints. In this line of research, Yang et al. (2010) propose a multi-objective goal programming model to incorporate guaranteed and non-guaranteed campaign contracts into the same planning problem. Bharadwaj et al. (2012) consider only guaranteed campaigns, but develop a fast scalable algorithm called SHALE which produces well-spread impression allocations. Ghosh et al. (2009) show how representative impression allocations can be purchased for guaranteed campaigns by randomly bidding for supply in the spot market of non-guaranteed impressions. Finally, McAfee et al. (2013) provide results which cross-cut these papers, and additionally consider minimizing an objective based on Kullback-Leibler divergence rather than L2-distance. Like these papers, we focus on spreading impressions across audience segments. However, our Gini-based approach is fundamentally different and provides an intuitive and appealing graphical representation of spread (i.e., Lorenz curves), as well as a model which satisfies important properties, and a tailored efficient algorithm that exploits Gini structure.

2.2. Economics Background

The Gini coefficient (Gini 1912) and Lorenz curve (Lorenz 1905) are commonly-used in economics to measure the dispersion of an endowment of resources; for example, to measure income inequality (see, e.g., Atkinson 1975, Campano and Salvatore 2006). While the Lorenz curve provides in some sense a distribution of the endowment of resources, the Gini coefficient is a summary statistic that summarizes the shape of the Lorenz curve and ranges from 0 (complete equality) to 1 (complete inequality).

Given an n-dimensional endowment vector x that describes an assignment of resources to individuals $i = 1, \ldots, n$, the Gini coefficient is computed by taking the average absolute difference of the endowments across all pairs of individuals, and normalizing by twice the mean endowment. That is, the Gini coefficient is defined as:

$$G = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j|}{2n^2 \mu}, \tag{1}$$

where $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the mean endowment.

The Gini Mean Difference (GMD) is a closely-related measure of dispersion defined as the average absolute difference in endowments across all pairs of individuals, and is written (Yitzhaki 1982) as:

$$GMD = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} |x_i - x_j| \; = \; \frac{2}{n^2} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} |x_i - x_j| \; = \; 2\mu G. \tag{2}$$

The above definition (2) highlights that Gini's Mean Difference is dependent on the spread of the values of the variable of interest, but not on deviations from some central value (e.g., the mean). Note that there are other variants of GMD; for these, we refer the reader to an extended review (Yitzhaki and Schechtman 2013) that discusses how to compute GMD under different assumptions (e.g., non-negativity of the variables of interest).

The Gini coefficient has been used in various fields and for multiple purposes, including to measure income distribution (Campano and Salvatore 2006), asset allocation and investment (Ogryczak and Ruszczyński 2002a, Shalit and Yitzhaki 1984, 2005), service equity (Drezner et al. 2009), and health inequality measurement (Lai et al. 2008), to name a few. The Gini measure and its derivatives, such as the Gini Mean Difference (GMD) risk measure (Shalit and Yitzhaki 2005, Yitzhaki 1982), have appealing features. The GMD metric is a coherent risk measure and is consistent with utility maximization theory and provides necessary conditions for second degree stochastic dominance (Ogryczak and Ruszczyński 2002a). Gini measures do not require the utility function to be quadratic or the distribution of the uncertain variables to be normally distributed. The GMD measure has also been extended to include decision-makers' preferences toward risk (Ogryczak and Ruszczyński 1999, 2002a)– see, for example, the use of the extended Gini (Yitzhaki 1983) as a measure of risk in the Mean-Extended Gini model (Shalit and Yitzhaki 1984, Yitzhaki and Schechtman 2013). As indicated in the above literature

review, Gini-based metrics are usually used in finance to model risk and analyzed with respect to risk axioms and properties. The motivation for using Gini metrics in this paper is quite different, as the primary objective is to provide spreading and diversification in the display of online advertising and the properties that it should exhibit and that are used to analyze its suitability are thus different (see Table 1 in Section 4 for a summary).

We now illustrate the connection between Lorenz curve and Gini coefficient, and their practical interpretations. A Lorenz curve of a country's income distribution can be produced by sorting the citizens from poorest to richest, computing the cumulative income of the n-poorest citizens, and finally normalizing both the population on the x-axis and the total accumulated income from all citizens on the y-axis to 1. Such a Lorenz curve plots the cumulative percentage share of income recipients, ranked from poorest to richest, against the cumulative percentage share of total income. In an online advertising context, we are interested in the distribution of the number of impressions that an ad campaign gets from each audience segment that it targets (i.e., desires). Consequently, a Lorenz curve can be used to plot the cumulative percentage share of the desired impressions, ranked from least-allocated to most-allocated, against the cumulative percentage share of allocated impressions. More specifically, this Lorenz curve is produced by sorting the targeted audience segments from least-allocated to most-allocated, computing the cumulative allocation of the n-least-allocated targeted impressions, and finally normalizing both the targeted impression supply on the x-axis and total impression allocation on the y-axis to 1.

Figure 1 depicts two Lorenz curves for two distinct impression allocations. The blue curve represents a well-spread (i.e., low inequality) allocation where 36% of the impressions allocated to this ad campaign are assigned to the least-allocated 50% of the targeted impression supply. Meanwhile, the red curve represents a poorly-spread (i.e., high inequality) allocation where only 17% of the impressions allocated to this ad campaign are assigned to the least-allocated 50% of the targeted impression supply. The straight dotted diagonal line represents perfect equality (i.e., a perfectly-spread allocation), and the farther a Lorenz curve is below the line of perfect equality, the more unequal is the corresponding impression distribution. Formally, the Gini coefficient of a Lorenz curve measures the extent of this inequality, and is computed by taking the area between the line of equal distribution and the Lorenz curve, and dividing it by the entire area below the line of equal distribution. In our example, the blue curve has a Gini coefficient of 0.20=A/(A+B+C), and the red curve has a Gini coefficient of 0.50=(A+B)/(A+B+C).

For more context, assume a campaign targets only 3 audience segments, e.g., Washington, Nevada, and Oregon, and receives 19%, 30%, and 40% of the corresponding impression traffic of 5M, 3M, and 2M, yielding 0.95M, 0.9M, and 0.8M allocated impressions, respectively. Although normalized, the x-axis represents the 10M impressions this campaign targets (i.e., desires). With audience segments sorted from least-to-most allocated,

the cumulative share of desired impressions on the x-axis corresponds to Washington between 0% and 50%, to Nevada between 50% and 80%, and to Oregon between 80% and 100%. Moreover, of the 2.65M impressions allocated, 36% come from Washington, and 70% come from Washington and Nevada. Hence, the resulting Lorenz curve is similar to the blue one in Figure 1 in that it also passes through $(0.5, 0.36)$ and $(0.8, 0.7)$. However, since in this case there are only three audience segments, it would be piecewise-linear with three segments.

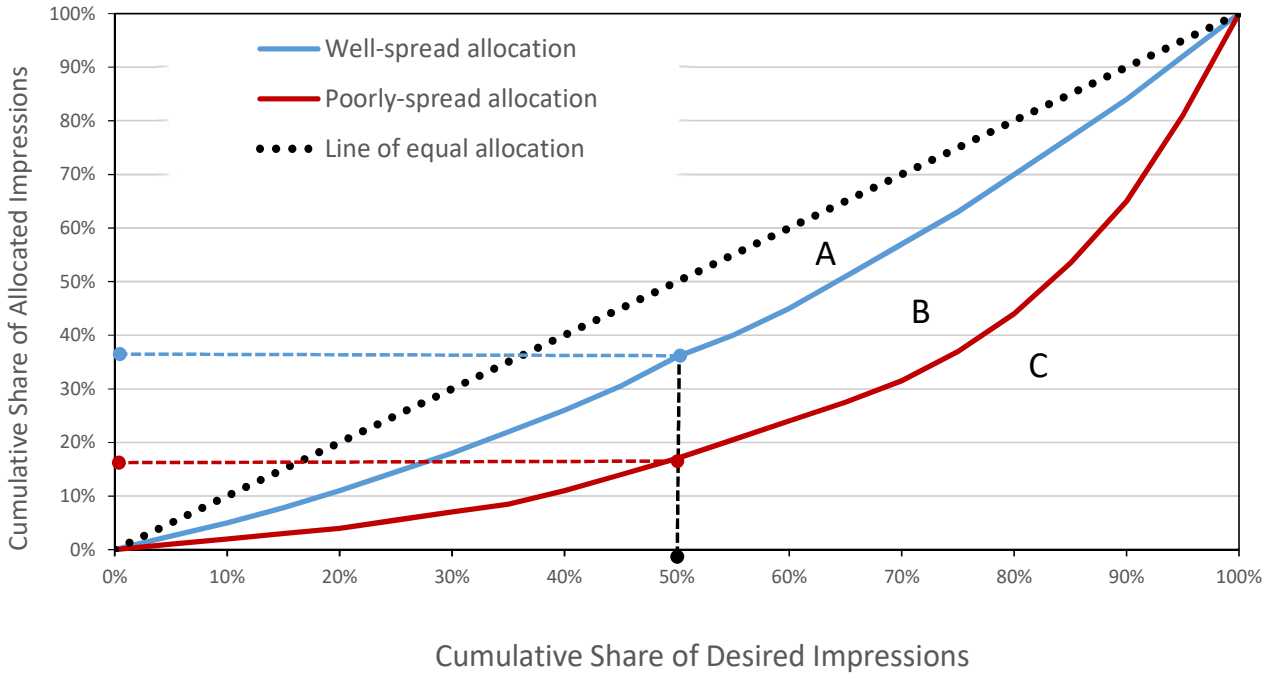

Figure 1 Lorenz curve.

A Lorenz curve provides both an order-independent and a scaling-independent graphical representation of an advertising plan's spread, and is superior to a bar graph for these two reasons. Indeed, bar graphs can look different depending on how audience segments, which typically have no natural ordering (e.g., demographic groups or geographic regions), are arbitrarily ordered. Moreover, audience segments can represent different-sized groups in a population, and a bar graph which plots total impressions allocated does not properly normalize for heterogeneity in audience segment sizes. For details, see Figure EC.1 and associated description in Appendix B.1.

A Lorenz curve also provides a systematic way to graphically compare impression spread across different dimensions and campaigns. We envision that a publisher managing multiple advertisers' campaigns would produce a report like Figure 2, where each row depicts how well impressions have been spread for each advertiser, and each column depicts how well impressions have been spread across various orthogonal dimensions such as demographic, geography, and time. Each advertiser would be able to view only the rows corresponding to

their own campaigns, while the publisher would be able to view all rows to troubleshoot and view allocations across campaigns. Since Lorenz curves are by definition suitably normalized with both axes ranging from 0 to 1, an advertiser can quickly and easily see which dimension (e.g., demographic, geographic, time) is the best or worst-spread, and compare the relative amount of spreading across these dimensions. As well, the publisher can also compare the quality of the spread across advertisers for the same or across different spread dimensions.

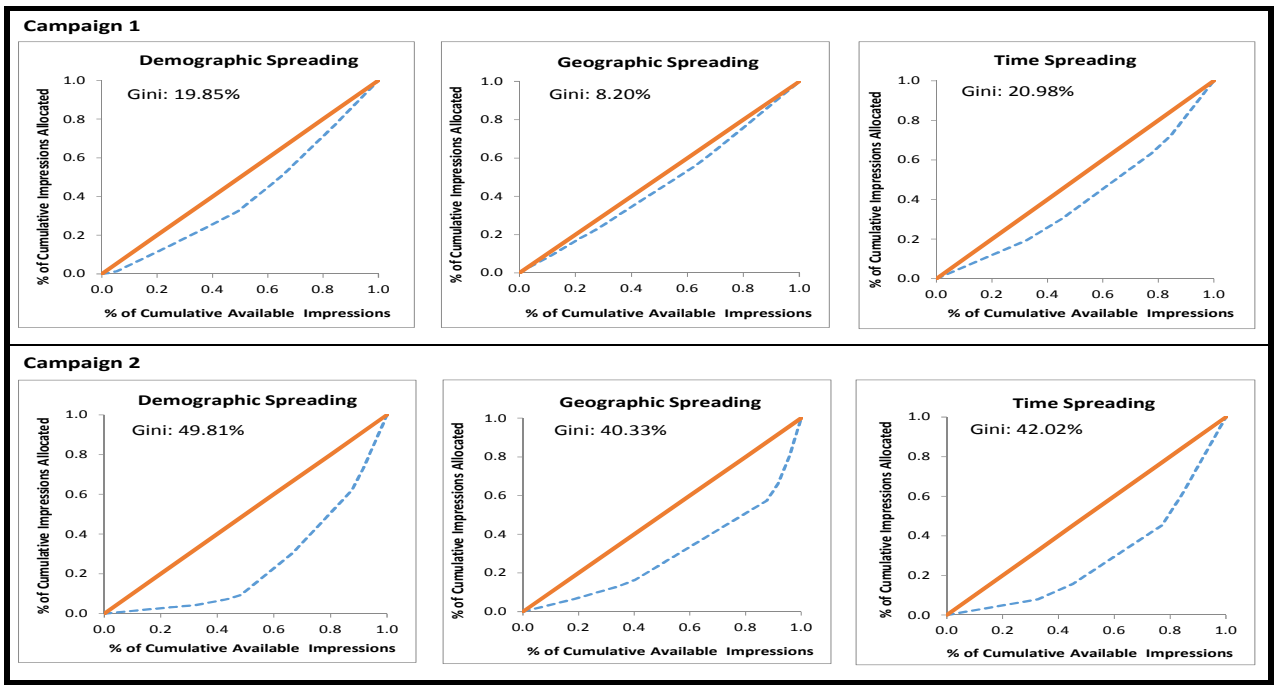

Figure 2 Impression spread by demographic, time, and geography, visually represented using Lorenz curves.

Our Gini-based model provides solutions with optimized Lorenz curves, in the sense that the Gini coefficients, i.e., the areas above the Lorenz curves, are minimized.

3. Model

We begin by stating the most popular model in the literature for spreading impressions across targeted audience segments in a single-period context. This model, which we use as our baseline for comparison, was developed at Yahoo and is described in the stream of literature by Ghosh et al. (2009), Yang et al. (2010), Bharadwaj et al. (2012), and McAfee et al. (2013). The primary objective in this model is to spread impressions as evenly as possible across targeted audience segments, to the extent that supply allows. After introducing this baseline model, we derive an appropriate Gini metric and a related Gini-based model.

3.1. Baseline Model

The baseline model (c.f., Bharadwaj et al. 2012) can be represented by a bipartite graph with supply nodes on one side representing audience segments, and demand nodes on the other representing ad campaigns (see Figure 3). Note that an audience segment can be defined in a way that it subsumes the demographic and geography dimensions (e.g., 35-45 years old in Florida) mentioned earlier and depicted in Figure 2. For clarity of presentation, we do not explicitly model the time dimension in the body of this paper; however, we provide an extensive formal treatment of the multi-period extension of our Gini model in Appendix E.

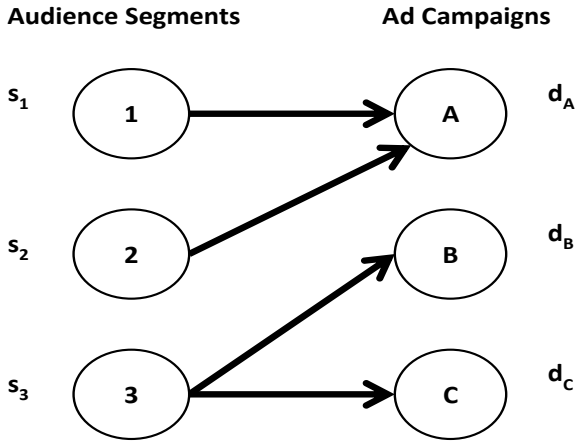

Figure 3 Bipartite graph.

Supply node i is connected to demand node j with an edge (i, j) whenever audience segment i is *targeted* (desired, or eligible for display) by campaign j. We denote the full set of edges as Γ, the set of audience segments targeted by campaign j as $\Gamma(j)$, and the set of campaigns that target audience segment i as $\Gamma(i)$. The impression goal (demand) of campaign j is denoted d_j, and the number of impressions generated (supply) from audience segment i is denoted s_i. Both demand and supply are assumed to be deterministic. The principal decision variables are denoted x_{ij}, and measure the proportion of audience segment i that is assigned to campaign j. Demand shortfalls are possible if there is insufficient supply to meet all campaigns' impression demands. We use y_j to measure the planned demand shortfall for campaign j, and penalize shortfalls in the objective function. The constant p_j is the penalty incurred per unit of shortfall, and in practice is either equal to the price per impression paid by the advertiser (i.e., a full refund is given for any impressions not delivered as planned), or slightly larger (i.e., to additionally compensate the advertiser for the loss of goodwill). This single-period baseline model, which we

will refer to as (SB), is formulated[2] as follows (Table EC.1 in the Appendix provides a reference to the notation):

$$\text{(SB)} \quad \min \sum_{(i,j)\in\Gamma} \frac{V_j}{2\theta_j} s_i (x_{ij} - \theta_j)^2 + \sum_{j\in J} p_j y_j \tag{3a}$$

$$\text{s.t.} \sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j \qquad \forall j \in J \qquad \text{(demand)} \tag{3b}$$

$$\sum_{j\in\Gamma(i)} x_{ij} \leq 1 \qquad \forall i \in I \qquad \text{(supply)} \tag{3c}$$

$$x_{ij}, y_j \geq 0 \qquad \forall (i,j) \in \Gamma \qquad \text{(non-negativity)} \tag{3d}$$

The objective function has two terms. The first term is the so-called *non-representativeness* term, which minimizes the L2 distance from the perfectly representative solution in which each campaign receives an equal proportion of each targeted audience segment. The second term minimizes the cost of demand shortfalls. The parameter $V_j > 0$ allows the modeler to apply more or less weight to the quadratic terms, to control how much effort is placed on spreading relative to attaining low demand shortfalls.

In the special case where supply exceeds demand causing the supply constraints to be nonbinding, the perfectly representative solution can be obtained by solving (SB). Given $\hat{s}_j = \sum_{i\in\Gamma(j)} s_i$ as the number of impressions eligible for campaign j, and the demand intensity $\theta_j = d_j/\hat{s}_j$, it is easy to show that the perfectly representative solution assigns campaign j a θ_j-proportion of the impressions from each audience segment $i \in \Gamma(j)$ that campaign j targets. In general, of course, supply constraints may bind, and the objective of (SB) finds a *maximally representative* allocation which is as close as possible to the perfectly representative allocation (for further details, see either Ghosh et al. (2009) or McAfee et al. (2013)).

Note that the decision variables x_{ij} are expressed as proportions, rather than absolute impressions. This is convenient, as it provides a natural way to operationalize the solution to this allocation plan within a real-time ad-serving system. When a user from audience segment i *arrives* (i.e., views a webpage that requires an ad to be chosen), we can assign this user ad j with probability x_{ij}. This equivalence between a proportional allocation and a probabilistic assignment is common in the advertising allocation literature[3].

3.2. Gini-Based Model

3.2.1. Gini-Based Metric. We now derive a Gini-based metric to measure, for each campaign j, how well-spread impressions are allocated across the audience segments $i \in \Gamma(j)$ that campaign j targets. For notational convenience, we order the audience segments and define $\Gamma_0(j) = \{(h,i) \in \Gamma(j)^2 : h < i\}$, which indexes all distinct audience segment pairs targeted by campaign j. Recall from (2) that to compute the GMD of endowments across individuals, we take the average absolute difference in endowments across all pairs of individuals. In

our context, audience segments take the place of individuals, and for each campaign j we have an endowment vector made up of the proportions x_{ij} of each audience segment $i \in \Gamma(j)$ targeted by campaign j. Consequently, we define the GMD for campaign j as the average absolute difference in proportional impression endowments across all pairs of audience segments, with the added twist that since audience segments differ in their size s_i, some additional audience-size factors enter into the expression. Formally, the GMD of a campaign is generalized from (2) as:

$$GMD_j = \frac{2}{\hat{s}_j^2} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}|. \tag{4}$$

The $s_h s_i$ factor enters into this expression since the number of ways we can pick a pair of impressions such that one is from audience segment h and the other is from audience segment i is precisely s_h times s_i. Note that one can also derive the above expression by directly computing the average absolute difference in endowments across the individual impressions which comprise the mutually exclusive audience segments, and then aggregating the impression-level quantities into audience segment sizes s_i and $\hat{s}_j$ (see Appendix A.3 for details).

Next, we define $w_j = \sum_{i \in \Gamma(j)} s_i x_{ij}$ as the total number of impressions planned for campaign j, and $\mu_j = w_j / \hat{s}_j$ as the ratio of the number of impressions allocated for campaign j to the number of impressions targeted by campaign j. Note that μ_j corresponds to the mean endowment for campaign j, and generalizes the unweighted average of (1) to a weighted average of endowments x_{ij}, where again the weights are the sizes s_i of the targeted audience segments.

Finally, following (2) the Gini coefficient for campaign j, which measures how well-spread impressions of campaign j are across the audience segments campaign j targets, is:

$$G_j = \frac{GMD_j}{2\mu_j} = \frac{\hat{s}_j GMD_j}{2w_j} = \left(\frac{1}{\hat{s}_j}\right) \frac{\sum_{(h,i) \in \Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}|}{\sum_{i \in \Gamma(j)} s_i x_{ij}}. \tag{5}$$

3.2.2. Gini-Based Model. Given the scalar $\alpha > 0$, we propose the objective function

$$\min \alpha \sum_{j \in J} w_j G_j + \sum_{j \in J} p_j y_j, \tag{6}$$

which has the nice property to be linear in GMD, since $w_j G_j = \frac{\hat{s}_j GMD_j}{2}$. Moreover, GMD_j as defined by (4) can be linearized in terms of the x_{ij} variables by introducing auxiliary variables x^+_{hij} for all $(h,i) \in \Gamma_0(j)$, and simplifying GMD_j to:

$$GMD_j = \frac{2}{\hat{s}_j^2} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i x^+_{hij}, \tag{7}$$

where we introduce the constraints:

$$x^+_{hij} \geq x_{hj} - x_{ij} \qquad \forall j \in J,\ \forall (h,i) \in \Gamma_0(j) \tag{8a}$$

$$x^+_{hij} \geq x_{ij} - x_{hj} \qquad \forall j \in J,\, \forall (h,i) \in \Gamma_0(j) \tag{8b}$$

to model the absolute value terms that were present in the original definition of GMD.

Making the substitution $w_j G_j = \frac{\hat{s}_j GMD_j}{2}$ in (6) yields the following Gini-based formulation:

$$\text{(SG)} \qquad \min \alpha \sum_{j \in J} \frac{1}{\hat{s}_j} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i x^+_{hij} + \sum_{j \in J} p_j y_j \tag{9}$$
$$\text{s.t. } (3b), (3c), (3d), (8a), (8b).$$

This is a linear program, which can be solved directly by a linear programming solver. However, the linearization step introduces many auxiliary variables and constraints. Later, in §6, we develop a decomposition method that sidesteps this large set of auxiliary variables and constraints while simultaneously illuminating problem structure.

There are two typical use cases for such a deterministic optimization model, which assumes point forecasts of impression supply are readily available. The first helps managers decide which contracts to accept and which to reject, and to perform what-if analysis while measuring the displacement effects of accepting particular contracts. The second deploys the model within a rolling-horizon decision-support system which re-optimizes periodically and provides guidance to a real-time ad-serving system that selects which ad to show to each viewer. A number of papers have discussed how deterministic planning models such as ours can be operationalized when ad requests arrive stochastically over a fixed time horizon (Chen et al. 2012, Yang et al. 2010, Turner et al. 2011, Hojjat et al. 2017). Our primary focus in this paper is to define the key properties that we believe a Gini-based model for online advertising should satisfy, understand the structure of the solutions that the Gini model produces, and introduce a novel decomposition method for solving our Gini-based optimization problem.

It is worth pointing out that our Gini model is quite different than Gini portfolio optimization models used in finance, despite the fact that both encourage allocations to be spread across numerous buckets. First, because Gini portfolio optimization models are typically used by one investor at a time, the corresponding bipartite graph (c.f., Figure 3) has only one demand node, in contrast to the multiple demand nodes (one per advertiser) that we have. Second, all Gini-based financial models we are aware of implicitly assume assets are liquid and consequently supply constraints may be ignored; however, in our context audience segments are commonly modeled as scarce resources, and indeed supplies of popular audience segments can be binding in practice, which presents additional challenges. Third, the form of the Gini metrics differ. Whereas Gini-based financial models promote stability and smoothness of portfolio returns under any foreseeable scenario, our advertising-related Gini metrics are not scenario-based and instead focus on diversification of allocation across audience segments. Finally, our Gini metrics incorporate particular weights which explicitly account for heterogeneity in the sizes of

audience segments and the volume of ad campaigns, both of which are specific to online advertising. For these reasons, ours is also a substantially more challenging problem to solve computationally.

3.3. Model Comparison

Both the baseline and Gini models have objectives that induce good spreading, yet are different. To understand the structural similarities between these objectives, we can focus on a single campaign's spread term and compare the interpretations of their functional forms. To derive our first comparison, we begin by defining a random variable to capture the variation in campaign j's allocation; specifically, let $X_j = x_{ij}$ with probability $s_i/\hat{s}_j$ for all $i \in \Gamma(j)$. Here, X_j selects a targeted audience segment at random, proportional to the size of each audience segment, and returns the allocation x_{ij} of the chosen audience segment. We can compute the variance of X_j as $\mathrm{Var}(X_j) = (1/\hat{s}_j)\sum_{i\in\Gamma(j)} s_i(x_{ij}-\mu_j)^2$, which is very similar to the spread terms in the baseline objective. Indeed, ignoring campaign-specific weighting factors, the baseline objective's spread terms can be written as $(1/\hat{s}_j)\sum_{i\in\Gamma(j)} s_i(x_{ij}-\theta_j)^2$. The interpretation is when there is no shortfall ($w_j = d_j$), the baseline objective's spread terms are essentially variance terms; this follows since no shortfall implies $\theta_j = d_j/\hat{s}_j = w_j/\hat{s}_j = \mu_j$.

To compare this to the spread term in our Gini objective, we need to also consider a second random variable $\tilde{X}_j$, defined exactly the same way as X_j, so that X_j and $\tilde{X}_j$ represent two independent draws from the same distribution. Yitzhaki (2003) defines the GMD of a random variable as the expectation of the absolute difference of two i.i.d. such random variables, i.e., $\mathrm{GMD}(X_j) = \mathbb{E}[|X_j - \tilde{X}_j|]$. It is straightforward to show that $\mathrm{GMD}(X_j) = GMD_j$, where GMD_j is the GMD measure we defined in (4). Ignoring campaign-specific weighting factors, this is exactly the spread term in our Gini objective. Finally, it is interesting to note (c.f., Yitzhaki 2003) that the variance of a random variable can also be written as the expectation of the *squared* difference of two i.i.d. random draws from that random variable's distribution; therefore, the baseline objective's spread terms (when there is no shortfall) can be expressed as $\mathrm{Var}(X_j) = (1/2)\,\mathbb{E}[(X_j - \tilde{X}_j)^2]$, which is similar in spirit to our Gini objective's spread terms, $GMD_j = \mathbb{E}[|X_j - \tilde{X}_j|]$.

In some contexts, it is helpful to express the deviations from the mean allocation as residuals, defined for each (i,j) pair as $r_{ij} = x_{ij} - \mu_j$. Moreover, if we define a randomly-chosen residual for campaign j as $R_j = X_j - \mu_j$, then it is straightforward to show that $\mathrm{Var}(R_j) = \mathrm{Var}(X_j)$ and $\mathrm{GMD}(R_j) = \mathrm{GMD}(X_j)$. Consequently, we may also think of the baseline objective (in the case of no shortfall) as a *sum of squared residuals*, $(1/\hat{s}_j)\sum_{i\in\Gamma(j)} s_i r_{ij}^2$, and the Gini objective as a *GMD of residuals*, $GMD_j = \frac{2}{\hat{s}_j^2}\sum_{(h,i)\in\Gamma_0(j)} s_h s_i |r_{hj} - r_{ij}|$. Stated another way, the baseline objective's spread terms (when there is no shortfall) have the form $\mathrm{Var}(R_j) = (1/2)\,\mathbb{E}[(R_j - \tilde{R}_j)^2]$, which is similar in spirit to our Gini objective's spread terms, $GMD_j = E[|R_j - \tilde{R}_j|]$. This connection to

residuals is important for two reasons. First, in Section 5 we will empirically compare solutions from the baseline and Gini models by plotting their corresponding residual distributions. Second, this observation connects our Gini model to the existing literature on R-regression. Developed by Hettmansperger (1984) and mentioned in Yitzhaki and Schechtman (2013), R-regression is used to solve robust parameter estimation problems by minimizing the GMD of residuals in a similar spirit to how an ordinary linear regression model is estimated by minimizing squared residuals. Although there are some structural differences between an R-regression model and our Gini model, one could view our model as a special case of an R-regression model which additionally has supply and demand constraints, and treats x_{ij}'s as variables rather than input data. Finally, it may also be worth pointing out that there is a link between the Gini measure and L-moments (see, e.g., Hosking 1990), which are specific linear combinations of order statistics. In particular, the second-order L-moment, also called the sample L-scale, is defined as half of Gini's mean difference. Further examination of these connections is beyond the scope of this paper, and are left to future research.

4. Model Properties

In this section, we introduce five key properties which we believe an online advertising optimization model that balances shortfalls with spreading should possess, and show when and how the baseline (SB) and Gini (SG) models satisfy these properties. Table 1 summarizes our results, which we demonstrate in the remainder of this section. Three of these properties are shared among both the baseline and Gini model; this is to be expected, since the baseline model was developed with some of these properties in mind (c.f., Bharadwaj et al. 2012, McAfee et al. 2013 in particular). This section verifies that our Gini model also satisfies these properties, and shows that our model additionally satisfies two more.

Property	Baseline Model	Gini Model
1 - Efficient Solvability	Yes	Yes (via algorithm of §6)
2 - Ideal Allocation when Possible	No	Yes, when $p_j = p \, \forall \, j \in J$
3 - Sufficient Orthogonality	No	Yes
4 - Split-and-Merge Invariance	Yes, as long as V_j's are chosen independently of d_j's	Yes, always
5 - Weight by Campaign Size	Yes, size is impressions demanded (d_j)	Yes, size is impressions allocated (w_j)

Table 1 Summary of Key Properties

Property 1 *The optimization model should be efficiently solvable.*

The baseline model (SB) is a (convex) quadratic program, and our Gini model (SG) can be formulated as a linear program, albeit with a large number of variables and constraints. Consequently, both the baseline and Gini models are efficiently solvable, and in particular for the Gini model we develop a specialized solution method in §6 to intelligently sidestep the large number of variables and constraints.

Property 2 *When possible, the optimal solution should be an **ideal allocation**, which we define as an allocation that (i) provides perfect spreading, and (ii) all impression shortfall is unavoidable.*

We formally define *perfect spreading* and *unavoidable impression shortfall* below. Intuitively, if an ideal allocation is feasible, it is clear that we cannot improve either spread or impression shortfall, which are the two objectives of (SB) and (SG).

DEFINITION 1. **Perfect Spreading**. A solution $\mathbf{x} = \{x_{ij}\ \forall (i,j) \in \Gamma\}$ is *perfectly spread* if for each campaign $j \in J$, the proportional allocations of each targeted audience segment are equal, i.e., $x_{ij} = x_j$ for all $i \in \Gamma(j)$.

DEFINITION 2. **Unavoidable Impression Shortfall**. Without loss of generality, let $I = \bigcup_{j \in J} \Gamma(j)$ be the set of audience segments targeted by at least one campaign (audience segments not targeted by any campaign can be safely deleted). The *unavoidable impression shortfall* is the constant

$$Y = \left[\sum_{j \in J} d_j - \sum_{i \in I} s_i \right]^+, \tag{10}$$

which is the positive part of the difference between the total demand and total supply of impressions.

The following proposition provides us with a sufficient condition for the existence of an ideal allocation.

PROPOSITION 1. *If the system of linear inequalities*

$$\sum_{i \in \Gamma(j)} s_i x_j + y_j = d_j \qquad \forall j \in J \qquad \textit{(link shortfall to allocation)} \tag{11a}$$

$$\sum_{j \in J} y_j = Y \qquad \textit{(all shortfalls are unavoidable)} \tag{11b}$$

$$\sum_{j \in \Gamma(j)} x_j \leq 1 \qquad \forall j \in J \qquad \textit{(supply constraints)} \tag{11c}$$

$$y_j, x_j \geq 0 \qquad \forall j \in J \qquad \textit{(non-negativity)} \tag{11d}$$

admits a feasible solution, then $\mathbf{x}$ *is an ideal allocation.*

Proposition 1 ensures perfect spreading by using only one allocation variable x_j per campaign j, which we can expand to a full solution to (SB) or (SG) by taking $x_{ij} = x_j$ for all $(i,j) \in \Gamma$. Constraint (11a) defines the impression shortfalls y_j, $j \in J$, while Constraint (11b), along with the non-negativity of y_j in Constraint (11d), ensures shortfall is at its minimum, i.e., equal to the level of unavoidable shortfall Y.

PROPOSITION 2. *Let $p_j = p, j \in J$. If the system* (11a)-(11d) *is feasible,*

1. *The optimal solution of the Gini model (SG) is always an ideal allocation. This statement is valid regardless of the emphasis placed on spreading determined by the parameter $\alpha > 0$.*
2. *There is no guarantee that the optimal solution of the baseline model (SB) is an ideal allocation.*

See Appendix A.4 for the proof. We illustrate the above property using the following toy problem with three audience segments $I = \{1, 2, 3\}$ and four ad campaigns $J = \{A, B, C, D\}$, in which $d_A = 20, d_B = 400, d_C = 200, d_D = 350$, $s_1 = s_2 = 200, s_3 = 300$, $p_A = p_B = p_C = p_D = 0.01$, $V_A = V_B = V_C = V_D = 1$, $\Gamma(A) = \Gamma(B) = \{1, 2, 3\}$, $\Gamma(C) = \{1, 2\}$, and $\Gamma(D) = \{3\}$. Unavoidable shortfall Y is equal to 270. The optimal solution of the Gini model provides perfect spreading ($x_{1A} = x_{2A} = x_{3A} = 0.0079$, $x_{1B} = x_{2B} = x_{3B} = 0.5554$, $x_{1C} = x_{2C} = 0.4367$, and $x_{3D} = 0.4367$) with minimal shortfall of 270 units ($y_A = 14.45, y_B = 11.25, y_C = 25.31$, and $y_D = 218.98$). While the baseline model's optimal solution also provides minimal shortfall ($y_A = 4.76, y_B = 95.17, y_C = 18.18$, and $y_D = 151.89$), it does not provide perfect spreading ($x_{1A} = x_{2A} = 0.0260, x_{3A} = 0.0162, x_{1B} = x_{2B} = 0.5195, x_{3B} = 0.3235, x_{1C} = x_{2C} = 0.4545$, and $x_{3D} = 0.6604$).

Property 3 *The measures of spread and shortfall should be **sufficiently orthogonal** (defined formally below).*

Intuitively, this property says that the two objectives that we minimize, the cost of (poor) spreading and the cost of impression shortfall, should be independent; or, if a dependency exists, this dependency should not run in the "wrong" direction. More specifically, if we start from a feasible solution and seek one with better spread, it is natural to expect that we may need to give up some revenue (i.e., incur additional shortfall). Conversely, if we start from a feasible solution and move to a lower-revenue (i.e., higher shortfall) solution, we would not expect spread to worsen. Although the interplay between spread and shortfall can be quite complex in a constrained optimization problem such as this, we can nevertheless prefer what we call *sufficiently orthogonal* measures, i.e., for each campaign, the spread cost should not increase when we increase shortfall cost.

Formally, let $\mathbf{x_j} = \{x_{ij}\}_{i \in \Gamma(j)}$ denote campaign j's allocation vector. Then the baseline (3a) and Gini (6) objectives, which minimize the cost of (poor) spreading plus the cost of shortfalls can be written in the general form $\sum_{j \in J} \left(f_j^{SPREAD}(\mathbf{x_j}) + f_j^{SHORTFALL}(\mathbf{x_j}) \right)$. In particular, $f_j^{SPREAD}(\mathbf{x_j}) = \sum_{i \in \Gamma(j)} \frac{V_j}{2\theta_j} s_i (x_{ij} - \theta_j)^2$ for the baseline objective, $f_j^{SPREAD}(\mathbf{x_j}) = \alpha w_j G_j$ for the Gini objective, and $f_j^{SHORTFALL}(\mathbf{x_j}) = p_j y_j$ for both. Recall that w_j, G_j, and y_j are all functions of $\mathbf{x_j}$, since $w_j = \sum_{i \in \Gamma(j)} s_i x_{ij}$; $y_j = d_j - w_j$; and G_j is defined by (5). Finally, define $f_j^{SPREAD*}(c) = \left\{ \min_{\mathbf{x_j}} f_j^{SPREAD}(\mathbf{x_j}) : f_j^{SHORTFALL}(\mathbf{x_j}) = c \right\}$ as the minimum spread cost achievable when the shortfall cost is exactly c.

DEFINITION 3. **Sufficiently Orthogonal**. We say that spread measure $f_j^{SPREAD}(\mathbf{x_j})$ is *sufficiently orthogonal* to shortfall measure $f_j^{SHORTFALL}(\mathbf{x_j})$ if $f_j^{SPREAD*}(c)$ is nonincreasing in c for all $c > 0$.

PROPOSITION 3. *The Gini objective (6) has sufficiently orthogonal spread and shortfall measures, while the baseline objective (3a) does not.*

See Appendix A.5 for the proof, which shows that the Gini objective has $f_j^{SPREAD*}(c) = 0$ for all $c \geq 0$, while the baseline objective has $f_j^{SPREAD*}(c)$ strictly increasing in c for all $c > 0$.

We illustrate the importance of the sufficient orthogonality property with the following example, which has three audience segments $I = \{1, 2, 3\}$ and two ad campaigns $J = \{A, B\}$, such that $d_A = 13000$, $d_B = 6000$, $s_1 = 1000$, $s_2 = 60000$, $s_3 = 15000$, $p_A = p_B = 0.1$, $V_A = V_B = 1$, $\Gamma(A) = \{1, 2, 3\}$, and $\Gamma(B) = \{1, 2\}$. There is no unavoidable shortfall ($Y = 0$). We observe the following with the:

- **Gini model**. If we prioritize minimizing shortfall by setting the parameter α to a low value (i.e., < 0.1), the optimal solution has no shortfall, provides perfect spreading for campaign 2 ($x_{B1} = x_{B2} = 0.857$), and imperfect spreading for campaign 3 ($x_{A1} = x_{A2} = 0.143, x_{A3} = 0.8$). On the other hand, if we give higher priority to spreading by setting $\alpha \geq 0.1$, the optimal solution has 3136 units of shortfall exclusively supported by campaign 2 and provides perfect spreading for both campaigns ($x_{B1} = x_{B2} = 0.409$ and $x_{A1} = x_{A2} = x_{A3} = 0.591$).

- **Baseline model.** If we prioritize minimizing shortfall by setting all V_j parameters to a low value (in this case, any value between 0 and 0.0895) the optimal solution of the baseline model has no shortfall, provides perfect spreading for campaign 2 ($x_{B1} = x_{B2} = 0.857$), and imperfect spreading for campaign 3 ($x_{A1} = x_{A2} = 0.143, x_{A3} = 0.8$). However, if we increase the importance of spreading by setting $V_j \geq 0.09, j \in J$, the optimal solution of the baseline model involves some shortfall whose magnitude increases as the value of V_j increases. Increasing the V_j parameters beyond 50 has minimal effect; i.e., spreading no longer improves and shortfall does not increase much either (shortfall stays between 2131 and 2132 for values of V_j between 50 and 100000). The baseline model does not produce a perfectly-spread solution.

The inability of the baseline model to produce a perfectly-spread solution despite a high importance placed on spreading is due to the fact that the baseline objective is not sufficiently orthogonal. Because the baseline objective's spread term minimizes a distance from the nominal solution θ_j, $j \in J$, it not only strives to achieve good spread but also serves to minimize shortfall. Consequently, the baseline objective is limited in its ability to trade revenue (i.e., incur shortfall) to get better spread. On the other hand, our Gini objective is sufficiently orthogonal. Each demanded impression is either allocated, or not. If allocated, an impression incurs the servicing cost αG_j, and if not it incurs the shortfall cost p_j. Such a cost breakdown is not possible for the baseline objective.

Property 4 *The objective function value should be invariant to campaigns being arbitrarily split or merged.*

This property says that if we take two campaigns with the same targeting and merge them together (adding together their demands), that the objective value does not change. Similarly, if we take a campaign and split it into two campaigns so that each have the same targeting and the total demand remains the same as before the split, then the objective value does not change. This property is important for several reasons. First, it ensures that campaigns do not get better or worse service (i.e., quality of spread) simply because they have large or small demands. Second, it eliminates the perverse incentive that advertisers may have to book many small campaigns instead of one large one. Indeed, if advertisers could increase their contribution to the objective function and therefore get better spread by splitting a large campaign into many small ones, they would be tempted to do that. Clearly, this behavior is not only unfair to advertisers who do not game the system, it also places a burden on the publisher who would end up managing many small campaigns instead of a few large ones. Third, for computational tractability, a publisher may wish to merge all campaigns that share the same targeting together as a pre-processing step before optimizing its ad allocation (undoing this aggregation after solving the optimization problem). But, if the objective is not invariant to arbitrary merges, some campaigns will inadvertently contribute more or less to the objective as a result. For these reasons, it is important for the objective function to be invariant to arbitrary splits and merges. Formally, we define an arbitrary split as follows (an arbitrary merge of two campaigns sharing the same targeting is the reverse operation):

DEFINITION 4. **Arbitrary Split**. Given a campaign C, we say that C is *arbitrarily split* into the two smaller campaigns A and B if (i) campaign C is replaced by campaigns A and B; (ii) campaigns A and B inherit campaign C's targeting, i.e., $\Gamma(A) = \Gamma(B) = \Gamma(C)$; and (iii) the sum of the demands from campaigns A and B equal the demand of campaign C, i.e., $d_C = d_A + d_B$.

PROPOSITION 4. *(1) If the parameters V_j in the baseline objective are chosen independently of campaign demands d_j, then the baseline model's optimal solution is not affected by arbitrary campaign splits or merges. (2) Our Gini model's optimal solution is not affected by arbitrary campaign splits or merges.*

The proof of Proposition 4, provided in Appendix A.6, is constructive. For the Gini model, it shows that when campaign C is arbitrarily split into campaigns A and B, we always have $GMD_C = GMD_A + GMD_B$, and consequently $w_C G_C = w_A G_A + w_B G_B$. The situation is not as straightforward for the baseline model. Indeed, McAfee et al. (2013), who identified the importance of making the objective invariant to arbitrary splits and merges, showed that the coefficients in the baseline objective must be carefully chosen for this invariance to occur. Indeed, the campaign-specific scaling factor of $1/\theta_j$ in the baseline objective (3a) is motivated by the

desire to satisfy this invariance property. More specifically, in the spread term each campaign is scaled by the constant $W_j = V_j/(2\theta_j)$, which can be written as $W_j = Q_j/d_j$ with $Q_j = V_j\hat{s}_j/2$. Our proof of Proposition 4 shows that so long as the Q_j values are chosen independently of the demands d_j, split-and-merge invariance holds. The insight here is that, for the baseline model, the scaling factors W_j should be linear in $1/d_j$.

Property 5 *Larger campaigns should be given proportionally larger weight than smaller campaigns.*

Because our models minimize the sum total of spread and shortfall costs for multiple ad campaigns, and because different campaigns have different sizes, it is desirable to weight each campaign by its size in the objective, so that the needs of small campaigns do not disproportionately affect the quality of the solution for large campaigns. Both the Gini and baseline objectives weight campaigns by their size, but use different metrics for size. The Gini model considers impressions planned as the campaign size, whereas the baseline model counts both planned and unplanned impressions (i.e., the total demand) in its measure of campaign size.

PROPOSITION 5. *(1) Assuming $V_j = 1$ for all campaigns $j \in J$, the baseline objective weights campaigns by their size (i.e., impressions demanded d_j), and penalizes the average squared percentage deviation from the target allocation ($x_{ij} = \theta_j \; \forall i \in \Gamma(j)$). (2) Our Gini objective weights campaigns by their size (i.e., impressions allocated w_j), and minimizes $w_j G_j$, i.e., a weight times the Gini coefficient for campaign j.*

The proof of part 1 of the above proposition is in Appendix A.7 (this interpretation is not at all obvious from looking at the baseline objective). Part 2 follows by definition of our Gini objective (6).

Alternative Models. When constructing a model, there are often many decisions to be made, and having such a list of important properties directs this process and makes it less arbitrary. Indeed, ours is not the only possible Gini model that one could propose; but, it is one that satisfies our five properties. While (SB) is the most popular single-period spreading model in the literature, alternative baseline models could also be considered. In Appendix B.2, we discuss several alternative spreading objectives, and show how seemingly small structural modifications affect the ability of the corresponding model to satisfy the five key properties listed in Table 1.

5. Numerical Comparison

We now run some numerical experiments to compare the solutions of the baseline (SB) and Gini (SG) models. We focus on illustrating the impacts of Properties 2 and 3 (Ideal Allocation when Possible and Sufficient Orthogonality), which are satisfied by the Gini model but not the baseline. Our principal goal is to understand how the baseline and Gini model differ in the way that they trade off revenue (i.e., incur shortfall) to achieve good spreading. We begin by illustrating the impact of Property 2, and show that instances which admit ideal

solutions can produce poor solutions when using the baseline model. Next, we illustrate the impact of Property 3, which is applicable to all instances although its effect may not always be obvious due to the fact that we solve constrained optimization problems which involve tradeoffs in multidimensional space (i.e., tradeoffs that involve multiple campaigns, and for each a tradeoff between spread and shortfall). We have defined Property 3 in such a way that the effect of a single campaign's tradeoff between spread and shortfall can be isolated within this multidimensional space. Finally, we plot residual distributions as a function of revenue to understand the full impact of how each model trades off revenue (i.e., incurs shortfall) to achieve good spreading.

For all experiments in this section, we vary the magnitude of effort applied to spreading (the α parameter in the Gini model, and ν in the expression $V_j = \nu$ for the baseline model), and compute optimal solutions for each model at different levels of effort applied to spreading. When α and ν are both set to zero, the objective functions of both models coincide, no effort is applied to spreading, and both models simply minimize impression shortfall. At the other extreme, when α and ν are large, spread is maximized while little effort is applied to minimizing shortfall. Since the two models define spread differently, their solutions differ most when α and ν are large. However, even when α and ν are very small (e.g., set to $\epsilon = 10^{-6}$) the solutions to the two models, which are both shortfall-minimizing, can be quite different. This is because there are typically many solutions with the same shortfall, and each model optimizes its own spread function within the set of shortfall-minimizing solutions.

5.1. Data Instances

We generate three families of instances which correspond to the presence of different structures in the underlying bipartite graph (c.f., Figure 3):

1. **Loose instances** have low *sell-through* (the ratio of aggregate demand to aggregate supply) and high *targeting percentage* (the average proportion of viewer types targeted by a campaign, which corresponds to the link density of the bipartite graph). Because loose instances have both a high volume of aggregate slack as well as high link density, they commonly admit solutions that are both well-spread and low-shortfall.

2. **Globally tight instances** have high sell-through and high targeting percentage. Despite having little aggregate slack, the high link density in this family of instances provides some flexibility with regard to how impressions are distributed across audience segments.

3. **Locally tight instances** have high sell-through and low targeting percentage. The combination of a low volume of aggregate slack with low link density makes this family of instances the most constrained.

We used the following scheme to generate 10 instances from each of the 3 families above, each having 20 campaigns and 100 viewer types. First, we randomly assign each campaign and each viewer type a link intensity,

and with probability (campaign link intensity) $\times$ (viewer type link intensity), create a link between campaign and viewer type in the underlying bipartite graph. Next, we randomly generate the supply for each viewer type by multiplying 4 impressions per arrival by a random Normal$(1000, 1000^2)$ number of arrivals[4], and merge viewer types that target the same set of campaigns (summing together supplies). Finally, we generate demands by constructing an allocation that is only used for this purpose. We begin with no impressions allocated, and produce an allocation by sequentially allocating campaigns to the available supply. When the process is complete, we consider the total number of impressions allocated to a campaign as its demand. More specifically, to construct this allocation, we iterate through the list of campaigns, and for each campaign we attempt to book the same proportion r of supply s_i from each viewer type i that it targets (we call r its "reservation proportion"). For some viewer types, remaining supply may be less than rs_i, and in those cases the campaign settles for less by booking the remaining supply. Figure 4 lists the three families of test cases along with the uniform distributions used to generate their link intensities and the reservation proportions used to generate their demands. These datasets may be downloaded from the author's website[5].

Test Case Family	Sell-through	Targeting %	Campaign Link Intensity	Viewer Type Link Intensity	Reservation Proportion
Loose	71.4%	19.3%	U[0.1, 0.4]	U[0.5, 1]	20%
Globally Tight	94.6%	19.6%	U[0.1, 0.4]	U[0.5, 1]	48%
Locally Tight	90.7%	7.8%	U[0.05, 0.2]	U[0.25, 0.5]	$*$

Figure 4 The three families of test cases, with average sell-through and targeting percentages. In cases marked $*$, reservation proportions for each campaign were drawn from {80%, 90%, and 100%} with probabilities 0.6, 0.3, and 0.1, respectively.

In what follows, we illustrate concepts which broadly apply to all three families of test cases using instances from a single representative family (usually, the family in which the concept is most clearly demonstrated). When there are differences in the structure of the solutions across families, we also comment on these differences.

5.2. Perfect Spreading at the Cost of Avoidable Shortfall in the Baseline Model

Recall from Property 2 that while the Gini model produces an ideal allocation when one is feasible, the baseline model may not. We now experimentally measure the extent to which the baseline model's solution deviates from the ideal allocation when one exists. For this purpose, we minimally modify the 10 globally tight instances so that an ideal allocation is always feasible. We then solve each instance with the Gini model and record the amount of unavoidable shortfall Y (recall from Proposition 2 that for any value of the spread importance parameter α, the Gini model provides perfect spreading with minimal, i.e., unavoidable, shortfall). Next, we use the baseline

model to solve each instance ten times with spread importance parameters $V_j = \nu$ ranging from 0.2 to 2 in increments of 0.2 (larger ν means a higher importance is placed on spreading relative to shortfall minimization). Over the 100 solutions we obtain from the baseline model, on average 21.2% of campaigns do not receive perfect spreading across their targeted segments. Moreover, the baseline model provides perfect spreading with all shortfall unavoidable in only 8 cases (4%) with low V_j values. For 96% of the cases, we may modify the solution to obtain perfect spreading at the cost of incurring higher shortfall than necessary (i.e., higher than Y, which the Gini model provides). Specifically, we additionally constrain the baseline model to produce only perfectly-spread solutions and re-optimize each of the 100 cases, recording for each case the shortfall amount s_B and the relative magnitude of avoidable shortfall $\%S = (s_B - Y)/Y$ (if all shortfall is unavoidable, then $s_B = Y$ implies $\%S = 0$).

Figure 5 displays the average, minimum, and maximum value of $\%S$ over the 10 instances for each of the 10 values taken by the parameters V_j. The average magnitude of avoidable shortfall $\%S$ ranges from 10.01% ($V_j = 0.2$) to 48.07% ($V_j = 2$), and is equal to 35.59% over all 100 cases. The maximum magnitude of avoidable shortfall over all 100 cases is 172.01%. Figure 5 shows that requiring perfect spreading with the baseline model produces solutions with significantly larger shortfall than those obtained with the Gini model.

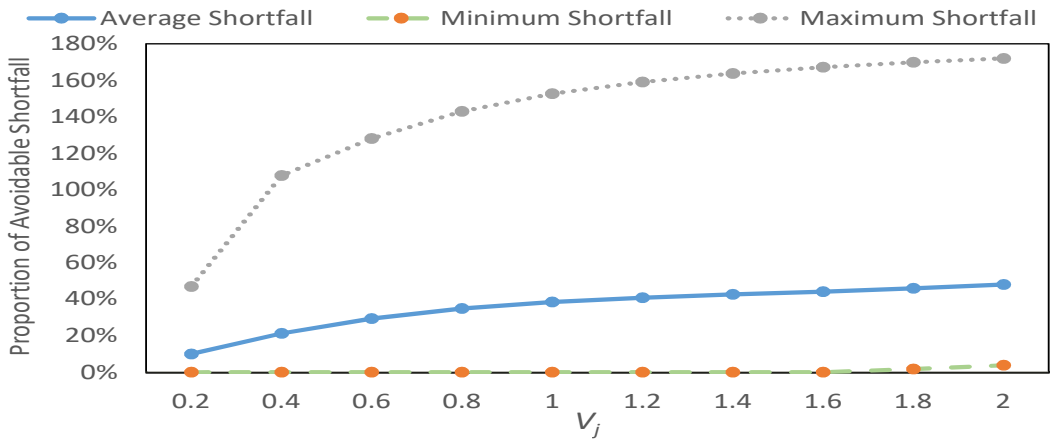

Figure 5 Avoidable Shortfall with Perfectly-Spread Solution using Baseline Model. Note: Gini Model has $\%S = 0$.

5.3. Characterizing the Tradeoff Between Revenue and Spread

The tradeoff between shortfall and spread can also be measured in a way that does not require an ideal allocation to be feasible. The question we seek to answer is, 'Using each model, how much revenue would the publisher need to sacrifice (by incurring shortfall) to produce solutions with a given level of spread?' To answer this

question, we define the revenue of a solution as $R = \sum_j p_j(d_j - y_j)$; i.e., we treat p_j as both the price per impression and the shortfall penalty, so that given demand d_j and shortfall y_j, the publisher collects $\sum_j p_j d_j$ from the advertiser, minus $\sum_j p_j y_j$ refunded for shortfalls.[6] We can illustrate this tradeoff for a single instance by computing the optimal solutions for each model at different levels of effort applied to spreading (the α parameter in the Gini model, and ν in the expression $V_j = \nu p_j$ for the baseline model). For example, Figure 6 plots the scaled spread cost as a function of revenue for one loose instance (for reporting, spread costs are normalized by dividing the spread term from the baseline objective by ν, and the spread term from the Gini objective by α).

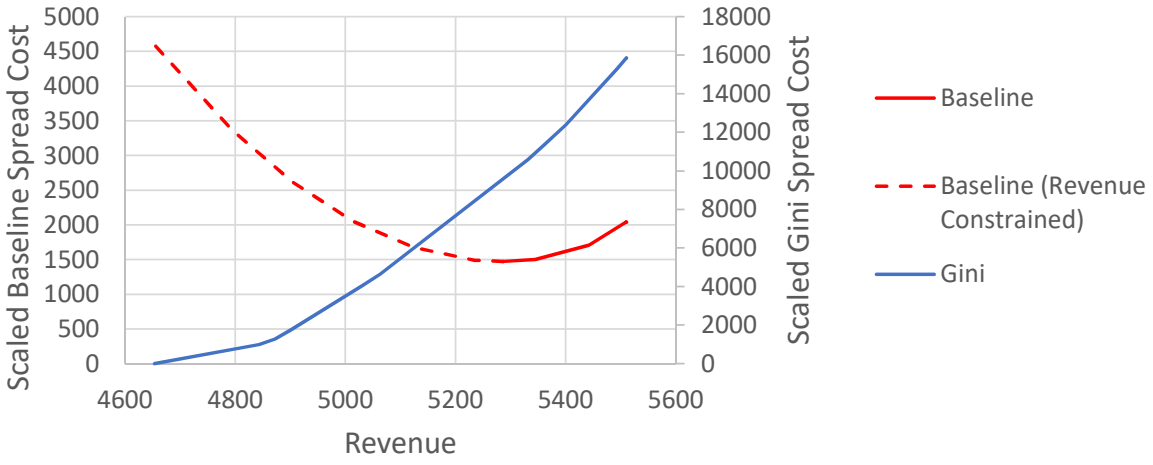

Figure 6 The tradeoff between revenue and spread cost for a single instance.

The impact of Property 3 (Sufficient Orthogonality) is evident in Figure 6. For the Gini model (blue line), as we increase the effort applied to spreading α, revenue decreases (moving from right to left), and spread cost monotonically decreases until we get to a point (at revenue = 4650) where the spread cost is zero. Initially, the baseline model (solid red line) behaves similarly, and as we increase the effort applied to spreading ν, the spread cost monotonically decreases as revenue decreases (again, moving from right to left). However, once we get to a revenue of 5275, further increases in ν will not budge the solution. Constraining the revenue with an equality constraint to values below 5275 produces lower-revenue solutions (dashed red line), but these solutions are not helpful since they do not lower the spread cost below 1500. Because the baseline objective is not sufficiently orthogonal, there is a limit to the amount that revenue can be sacrificed to produce well-spread solutions. This observation is robust, as we observe qualitatively similar behavior in globally tight and locally tight instances.

Although Figure 6 is helpful to visualize the comparative statics when we change the effort applied to spreading, one must be careful not to compare the absolute magnitudes between the baseline and Gini spread costs. These are on different scales, and are represented on different vertical axes. Indeed, there is no meaning to the

crossing point between the baseline and Gini curves. To compare the structure of the Gini and baseline solutions, we instead plot how the residual distributions change as we vary the effort applied to spreading.

5.4. Comparing Residual Distributions

Recall from §3.3 that we may define residuals $r_{ij} = x_{ij} - \mu_j$ for each audience segment and campaign pair (i, j), where $\mu_j = \frac{1}{\hat{s}_j} \sum_{i \in \Gamma(j)} s_i x_{ij}$ is the mean allocation across audience segments campaign j targets. Notice that if all residuals are zero, i.e., $r_{ij} = 0 \ \forall i \in \Gamma(j)$, we have a perfectly-spread solution, i.e., $x_{ij} = \mu_j \ \forall i \in \Gamma(j)$. Moreover, since $x_{ij} \in [0, 1]$, it follows that $\mu_j \in [0, 1]$, and therefore all residuals r_{ij} are in the range $[-1, 1]$.

For each of the three instance families (loose, globally tight, and locally tight), we proceed as follows. We repeatedly solve all ten instances to produce optimal solutions for the baseline and Gini models at different levels of effort applied to spreading α and ν. Defining the maximum revenue attainable R_{max} as the total revenue attained over all 10 instances when $\alpha = \nu = 0$, we normalize the total revenue R achieved at a particular value of α or ν to $\%R = R/R_{max}$, which we call the proportion of maximum revenue attained. We compute residuals for each instance from their respective optimal solutions[7], and combine residuals from all ten instances to produce a residual distribution for each model (baseline and Gini) at each level of revenue ($\%R$).

Figure 7 provides a residual distribution comparison for the family of locally tight instances. The main plot, top left, shows how the residual distributions from the baseline (red) and Gini (blue) models change as a function of revenue ($\%R$). Each line (different dashed pattern) corresponds to a different percentile of the distribution, as defined by the legend at the bottom right. The legend also applies to the plot at the bottom left, which provides a zoomed-in section of the main plot for $\%R$ between 0.95 and 1 and residual values between -0.1 and 0.1. Finally, at the top right we use three histograms to plot the residual distributions at $\%R = 1$, 0.95, and 0.9; each histogram corresponds to one of the cross-sections in the main plot highlighted by grey vertical bars.

From Figure 7 we observe that for any given revenue level $\%R$, the residual distribution under the Gini model is narrower in the middle and slightly wider in the tails, relative to the baseline. Indeed, Gini solutions have significantly more near-zero residuals, especially when revenue is low (i.e., when the emphasis on spreading is high). Although the residual distributions differ the most for the locally tight instances, structurally this observation is robust, with the globally tight and loose instances also producing Gini solutions with more near-zero residuals than the baseline; this result is summarized by Figure 8. Note that property 3 (Sufficient Orthogonality) is also evident in Figure 7, as we observe the baseline model cannot produce better-spread solutions by lowering

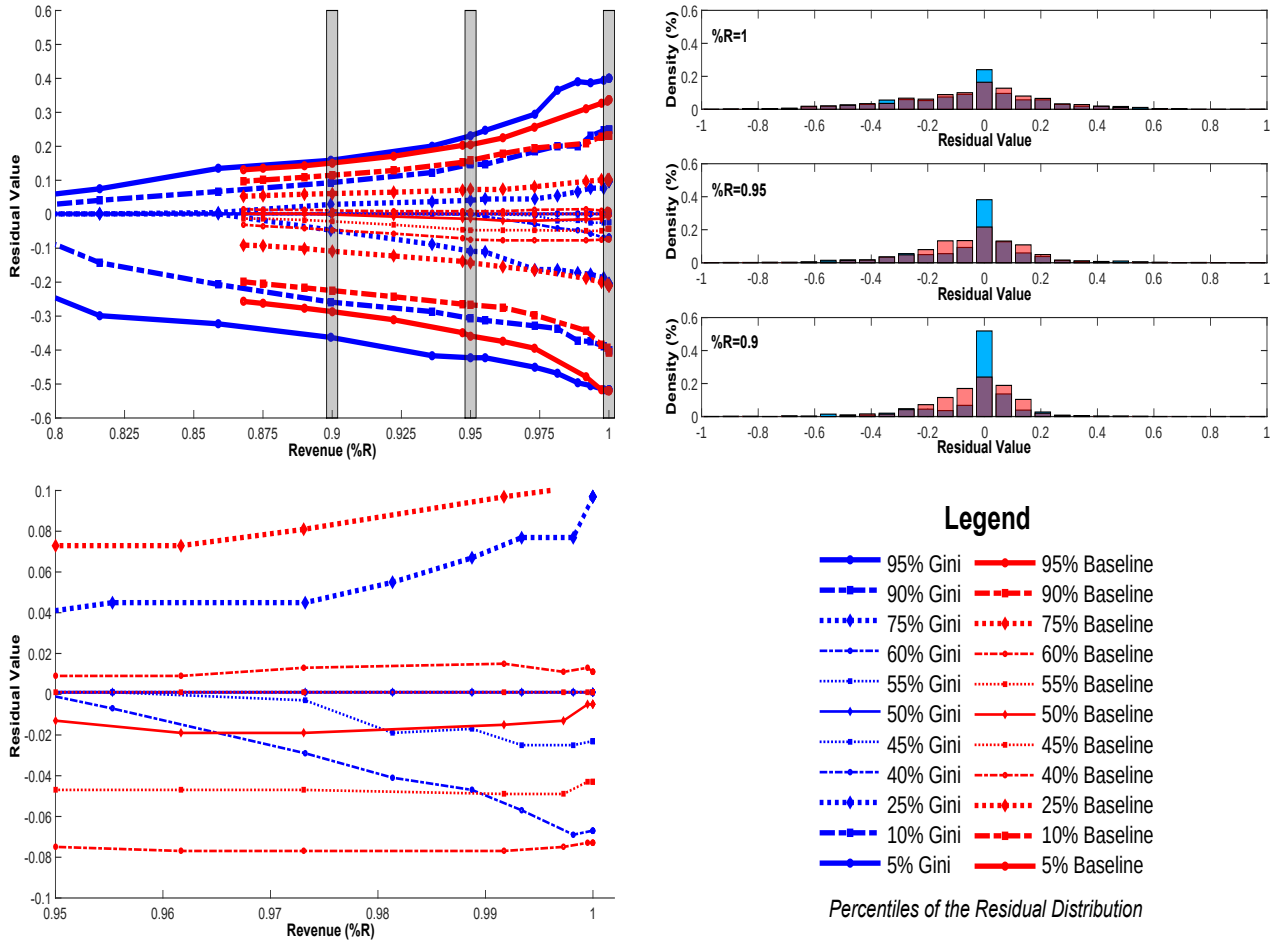

Figure 7 Residual Distribution Comparison (Locally Tight Instances). **Top left**: Percentiles of the residual distributions from the baseline (red) and Gini (blue) models, as a function of revenue ($\%R$). **Bottom right**: The legend represents each percentile by a different dashed pattern, and applies to both plots at left. **Bottom Left**: Zoomed-in section of the main (top left) plot for $\%R$ between 0.95 and 1, and residual values between -0.1 and 0.1. **Top right**: Histograms depict residual distributions at $\%R = 1$, 0.95, and 0.9; each histogram corresponds to one of the cross-sections in the main (top left) plot highlighted by grey vertical bars. Histograms for the Gini (blue) and baseline (red) distributions are overlayed, with purple bars indicating where the blue and red bars overlap.

revenue beyond $\%R = 0.87$, while the Gini model can. For a more detailed description of Figure 7 along with plots for the globally tight and loose instances, see Appendix B.4.

5.5. Practical Implications

There are several practical reasons why a publisher may prefer solutions from our Gini model to the baseline. First, a publisher may win contracts with advertisers or other ad partners based on its ability to deliver impressions in a precise manner. In a related example, the supply-side network Chitika was asked by a large ad aggregator to show that it could deliver ads with a click-through rate (CTR) of exactly 1.5%; it achieved 1.51% and won a large contract as a result (c.f., Mookerjee et al. 2016). Relative to the baseline, our Gini model pro-

Instance Family	Revenue (%R)	Baseline	Gini	Improvement (Gini - Baseline)
Locally Tight	1	16.4%	24.1%	7.7%
	0.95	21.6%	38.2%	16.6%
	0.90	23.9%	51.9%	28.0%
Globally Tight	1	30.5%	35.6%	5.1%
	0.95	37.9%	42.7%	4.8%
	0.90	45.3%	54.7%	9.4%
Loose	1	29.8%	53.8%	24%
	0.95	52.3%	71.9%	19.6%
	0.90	N/A	87.5%	N/A

Figure 8 Proportion of Near-Zero Residuals. For the locally tight and globally tight cases, near-zero residuals are those in the range $[-0.035, 0.035]$ (i.e., corresponding to the middle bar in Figure 7's histograms). For the loose case, nearly all residuals are in the range $[-0.035, 0.035]$, so we report the number of near-zero residuals in the range $[-0.010, 0.010]$ instead. Moreover, because the baseline model violates Property 3 (Sufficient Orthogonality), no solutions at $\%R = 0.9$ exist in the loose case.

duces solutions with more audience segments that get exactly the mean allocation (i.e., residuals are zero), which some advertisers and ad partners may view as an important precision metric. Second, the Gini model is more efficient in the way it allocates impressions from audience segments to campaigns, in the sense that it is much more likely than the baseline model to cede priority to a high-revenue campaign (i.e., one with a high shortfall penalty p_j) when an audience segment is in high demand. Indeed, the baseline model's quadratic objective all but assures most campaigns get some amount of each audience segment, even when the opportunity costs of allocating impressions from some audience segments are very high.

To compare the efficiency of the Gini and baseline models, we modify our test cases (c.f., Figure 4) by introducing an outside option (i.e., secondary channel) where the publisher may sell impressions of each audience segment i at the price $\tilde{\beta}_i$, which we generate uniformly at random from the interval $[0, 0.03]$. By construction, because the prices p_j for satisfying existing contracts are uniformly distributed in $[0.01, 0.02]$, in expectation half the impressions are more profitable to assign to the outside option than to a randomly-chosen campaign, at least one-quarter of the impressions are more profitable to assign to the outside option than to any campaign, and at least one-quarter of the impressions are more profitable to assign to any campaign than to the outside option.

We repeat our residual analysis, but with revenue now coming from both guaranteed campaigns as well as from slack impressions sold to the outside option[8], i.e., $R = \sum_j p_j(d_j - y_j) + \sum_i \tilde{\beta}_i s_i \left(1 - \sum_{j \in \Gamma(i)} x_{ij}\right)$. For the locally tight family of instances, whose solutions differ the most across the two models, Figure 9 shows that the Gini distribution is significantly narrower and more concentrated than the baseline at nearly all revenue levels. Indeed, the mass of residuals in the near-zero $[-0.035, 0.035]$-range is much higher for the Gini model than for the baseline (72.0% versus 16.4% at $\%R = 0.95$, and 94.7% versus 21.6% at $\%R = 0.9$). This demonstrates

that when slack impressions have value, the Gini model is more efficient at producing well-spread solutions. For completeness, we provide plots for the globally tight and loose cases in Appendix B.4.

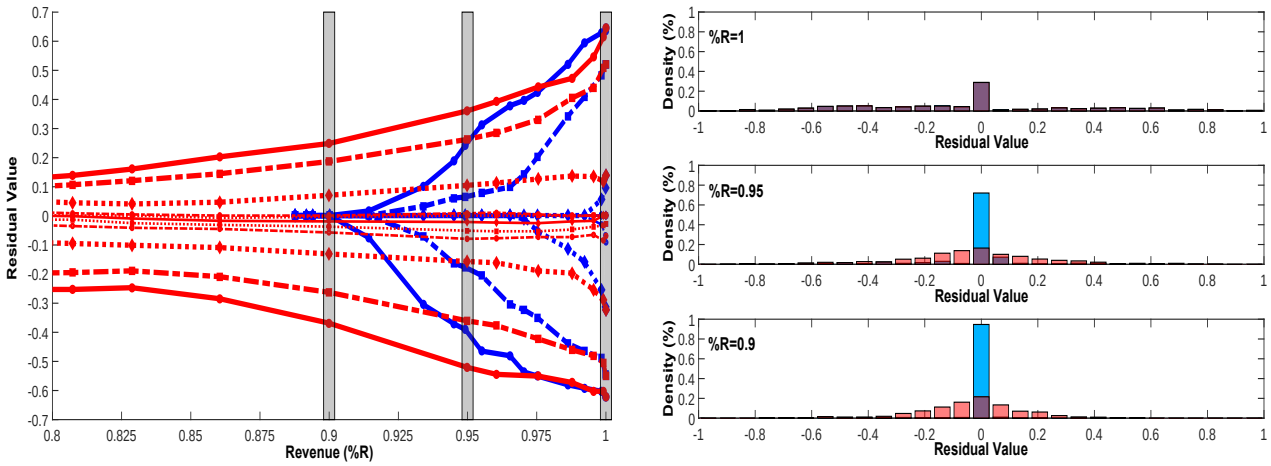

Figure 9 Residual Distribution Comparison with Outside Option (Locally Tight Instances). **Left:** Values of residuals at set percentiles as a function of revenue $\%R$ (see legend in Figure 7). **Right:** Histograms of the residual distributions at $\%R = 1$, 0.95, and 0.9; each histogram corresponds to a cross-section in the main (left) plot highlighted by vertical grey bars. Values from the Gini-based model are blue, values from the baseline model red, and the purple bars in the histogram plots indicate areas where the blue (Gini) and red (baseline) histograms overlap.

6. Decomposition Method

In this section, we introduce a novel decomposition method that leverages the structure of our Gini model to solve large instances quickly and efficiently. This is particularly important for practice, where publishers and ad aggregators manage thousands of ad campaigns simultaneously. In our numerical experiments, our decomposition method is able to solve instances on average 60 times faster than CPLEX. Moreover, it uses significantly less memory and subproblems are paralellizable; these are highly-desirable properties for handling large instances.

We begin this section by describing the key algorithmic ideas behind our decomposition method. We then sketch our algorithm, and describe the structure of the solutions that it produces.

6.1. Key Algorithmic Ideas

A key Gini structure our algorithm exploits is the fact that we may compute a Gini coefficient in two separate ways. Recall from §2.2 that we may either (1) suitably normalize the average absolute difference of endowments across all pairs of "individuals" (in our case, audience segments), or (2) calculate the area under the Lorenz curve. Only the first approach yields a formula for the Gini coefficient that can be directly embedded into a math program, and thus far, we followed the first approach by defining the Gini coefficient G_j as in Equation 5.

Producing a campaign's Lorenz curve to employ the second approach would require us to first sort the audience segments from least allocated to most allocated. But herein lies the difficulty, since prior to solving the allocation problem we do not know the allocations, and so cannot sort audience segments by their allocations! Although this argument would seem to rule out using the second approach, it turns out that if we relax the supply constraints so each campaign can buy any amount of impressions from each audience segment i at the fixed price $\hat{\beta}_i$, then for this relaxed problem we can infer how the audience segments should be ordered. With this ordering in hand, we may express each Gini coefficient as the area under a Lorenz curve, or equivalently we may dramatically simplify our original definition of G_j from Equation 5 (the ordering of the allocation vector's components tells us the sign of each term inside the absolute value, which allows us to drop absolute values and collect like terms). Moreover, this alternative view of how to mathematically express the Gini coefficient G_j also permits a useful change of variables into what we call the "increment space", allowing us to define our supply-relaxed problem in terms of the *rate* that the Lorenz curve increases from one ordered audience segment to the next. This is particularly useful, since in the increment space the supply-relaxed problem may be reformulated into a set of continuous knapsack problems (one per campaign). We then leverage the fact that the continuous knapsack problem has a closed-form solution, derive this solution in the increment space, and transform it back into the original space. This process is described formally immediately following Theorem 1, which characterizes the optimal solution to the supply-relaxed problem.

Except for the special case where supply constraints truly are non-binding and impressions of all types can be acquired on the open market at the fixed prices $\hat{\beta}_i$, $i \in I$, solving the original supply-constrained Gini ad allocation problem (SG) requires a more sophisticated approach. However, we can embed the supply-relaxed problem as a subproblem within a decomposition scheme that proposes at each iteration a new vector of prices $\hat{\beta} \equiv \{\hat{\beta}_i\}_{i \in I}$, and combines the supply-relaxed solutions $\{x^n_{ij}\}_{(i,j) \in \Gamma}$ from past iterations $n = 1..N$ to produce a near-optimal (and feasible) solution for the full Gini ad allocation problem (SG). We do this by treating the supply constraints as so-called coupling constraints, and employ Dantzig-Wolfe decomposition (Dantzig and Wolfe 1960). Moreover, since our subproblem is extremely fast to solve (computationally, its bottleneck is sorting the price vector $\hat{\beta}$), we also use subgradient optimization (Fisher 2004, Held et al. 1974, Anstreicher and Wolsey 2009, Gustavsson et al. 2015) to produce several cheaply-computed proposed solutions each time we solve the Dantzig-Wolfe master problem. Due to space considerations, we present only our main results without proof in the body of this manuscript; see Appendix C for full derivations and proofs.

6.2. The Supply-Relaxed Subproblem

We now formally derive our supply-relaxed subproblem, and introduce Theorem 1 which characterizes its solution. Defining $\psi_j = \alpha/\hat{s}_j$, our single-period Gini ad allocation problem (SG) is equivalently formulated as:

$$\begin{aligned}
\text{(P-ORIG)} \qquad z^* = \min & \sum_{j\in J} \psi_j \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \sum_{j\in J} p_j y_j \\
\text{s.t.} & \sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j \qquad \forall j \in J \qquad \text{(demand)} \\
& \sum_{j\in\Gamma(i)} s_i x_{ij} \le s_i \qquad \forall i \in I \qquad \text{(supply)} \\
& x_{ij} \ge 0 \,\forall (i,j) \in \Gamma; \qquad y_j \ge 0 \,\forall j \in J
\end{aligned}$$

Notice that in (P-ORIG) we have scaled both sides of the supply constraint by s_i so that the marginal value of increasing the right-hand side of the i^{th} supply constraint is naturally expressed as a per-impression price.

For every given vector of per-impression prices $\hat{\beta} \equiv \{\hat{\beta}_i\}_{i\in I}$, we can solve the following supply-relaxed subproblem to produce a lower bound for (P-ORIG):

$$\begin{aligned}
\text{(PS)} \qquad z^{LB}(\hat{\beta}) = \min & \sum_{j\in J} \psi_j \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \sum_{j\in J} p_j y_j + \sum_{i\in I} \hat{\beta}_i \left(\sum_{j\in\Gamma(i)} s_i x_{ij} - s_i \right) \\
\text{s.t.} & \sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j \qquad \forall j \in J \qquad \text{(demand)} \\
& x_{ij} \ge 0 \,\forall (i,j) \in \Gamma; \qquad y_j \ge 0 \,\forall j \in J
\end{aligned}$$

The supply-relaxed subproblem (PS) decomposes by campaign, allowing us to solve a sequence of smaller subproblems, one for each campaign $j \in J$, rather than solving (PS) directly. Using $z_j^{LB}(\hat{\beta})$ to represent the optimal value of the j^{th} supply-relaxed subproblem, we have the following relationship: $z^{LB}(\hat{\beta}) = -\sum_{i\in I} s_i \hat{\beta}_i + \sum_{j\in J} z_j^{LB}(\hat{\beta})$, where each subproblem looks like:

$$\begin{aligned}
\text{(PS-j)} \qquad z_j^{LB}(\hat{\beta}) = \min \ & \psi_j \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + p_j y_j + \sum_{i\in\Gamma(j)} \hat{\beta}_i s_i x_{ij} \\
\text{s.t.} & \sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j \qquad \text{(demand)} \\
& x_{ij} \ge 0 \,\forall i \in \Gamma(j); \qquad y_j \ge 0
\end{aligned}$$

Each campaign-j supply-relaxed subproblem can be solved analytically after sorting the viewer types in increasing order of $\hat{\beta}_i$ value, computing some important quantities, and then reading off the optimal solution. The following theorem characterizes the structure of the campaign-j subproblem's optimal solution.

THEOREM 1. *Assume that viewer types $i \in \Gamma(j)$ are ordered according to the prices $\hat{\beta}_1 \le \hat{\beta}_2 \le \cdots \le \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$. Define the following important quantities:*

- *Let* $s_i^{bj} = \sum_{i'=1..i-1} s_{i'}$ *and* $s_i^{aj} = \sum_{i'=i+1..m_j} s_{i'}$ *be the number of impressions that are rank ordered before and after viewer type* i*'s impressions, respectively.*
- *Let* $c_{ij} = \psi_j s_i (s_i^{aj} - s_i^{bj}) + s_i \hat{\beta}_i$ *represent the "cost" of assigning viewer type* i *to campaign* j *in its entirety.*
- *Let* $\tilde{c}_{ij} = \sum_{i'=1..i} c_{i'j}$ *and* $\tilde{s}_{ij} = \sum_{i'=1..i} s_{i'}$ *represent cost and supply usage "increments" that apply to the* i^{th} *increment of campaign* j *in a transformed space where the subproblem is easier to solve.*
- *Let* $\pi_{ij} = \tilde{c}_{ij}/\tilde{s}_{ij}$ *represent the cost of assigning one unit of the* i^{th} *increment to campaign* j*, in the transformed space.*
- *Let* $i^* = \arg\min_{i \in \{1..m_j\}} \pi_{ij}$ *be a lowest-cost increment in the transformed space.*

Then, the optimal solution to (PS-j), represented as (x_j^*, y_j^*) *with* $x_j^* \equiv \{x_{ij}^*\}_{i=1..m_j}$*, takes the following form. If* $\pi_{i^*j} > p_j$*, then* $y_j^* = d_j$*,* $x_{ij}^* = 0 \; \forall i = 1..m_j$*, and the corresponding optimal value is* $p_j d_j$*. Otherwise, the optimal value is* $\pi_{i^*j} d_j$ *with corresponding optimal solution* $y_j^* = 0$*, and*

$$x_{ij}^* = \begin{cases} d_j/\tilde{s}_{i^*j} & \text{for } i \leq i^* \\ 0 & \text{for } i > i^* \end{cases}.$$

The proof of Theorem 1 proceeds in several steps, which are described in full in Appendix C. At a high level, we (1) define an impression-based version of (P-ORIG) which allows us to omit the sizes of the viewer types (s_i) from several stages of our proof; (2) prove that, in the impression space, if the targeted impressions $r \in \Lambda(j)$ are ordered without loss of generality according to the prices β_r for each impression (i.e., $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{\hat{s}_j}$), then optimal solutions of the campaign-j subproblem in the impression space, represented as (x_j^*, y_j^*) where $x_j^* \equiv \{x_{rj}^*\}_{r=1..\hat{s}_j}$, must satisfy the constraints $x_{1j}^* \geq x_{2j}^* \geq \cdots \geq x_{\hat{s}_j j}^*$; (3) prove that, in the impression space, if $\beta_q = \beta_{q'}$ for two impressions (q, q'), then the optimal solutions for those impressions must also coincide ($x_{qj}^* = x_{q'j}^*$); (4) impose $x_{1j} \geq x_{2j} \geq \cdots \geq x_{\hat{s}_j j}$ as an optimality cut, allowing us to remove the absolute values in the Gini-based objective, which then allows us to linearize and simplify our subproblem's objective so that it contains $O(\hat{s}_j)$ terms rather than $O(\hat{s}_j^2)$ terms; (5) use the fact that $\beta_q = \beta_{q'} \implies x_{qj}^* = x_{q'j}^*$ to aggregate this simplified impression-based formulation with $O(\hat{s}_j)$ objective terms into a simplified viewer-type-based formulation with $O(|\Gamma(j)|)$ objective terms; (6) transform the resulting viewer-type-based formulation into one that has variables in so-called "increment space" by instead considering decision variables that model the increments $\delta_{ij} = x_{ij} - x_{i+1,j}$ for $i = 1..|\Gamma(j)| - 1$ rather than the x_{ij} variables directly; (7) read off the solution in the increment space, and translate it back into the viewer-type space.

Theorem 1 characterizes the solution to the supply-relaxed campaign-j subproblem (PS-j). If shortfall costs are relatively low ($p_j < \pi_{i^*j}$), then it is optimal to allocate no impressions to campaign j, and incur full shortfall ($y_j^* = d_j$). On the other hand, if shortfall costs are relatively high ($p_j \geq \pi_{i^*j}$), then it is optimal to have no

shortfall ($y_j^* = 0$) and to spread impressions proportionally across the viewer types $\{1, \ldots, i^*\}$. Since viewer types $i \in \Gamma(j)$ are ordered according to the per-impression prices $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$, we can interpret the solution as one that proportionally spreads impressions across the cheapest set of viewer types $\{1, \ldots, i^*\}$, where the costs of the viewer types are nontrivial and computed using Theorem 1. The viewer type i^* is at the threshold of affordability, given the values for $\{\hat{\beta}_i\}_{i=1..m_j}$.

6.3. Algorithm

Our decomposition method leverages the analytical result of Theorem 1 to solve our original Gini ad allocation problem (P-ORIG). It additionally provides (1) a method to choose "good" price vectors $\hat{\beta}$; and (2) a method to convert one or more solutions from supply-relaxed campaign-j subproblems into a solution $\{x_{ij}\}_{(i,j)\in\Gamma}$ that is near-optimal for (P-ORIG) and satisfies all supply constraints $i \in I$ of the form $\sum_{j\in\Gamma(i)} x_{ij} \leq 1$.

To get an initial feasible solution, we first run a two-stage heuristic. In stage one, our heuristic iterates through the list of campaigns $j = 1..|J|$, and attempts to assign $s_i\theta_j$ impressions of each viewer type $i \in \Gamma(j)$ to campaign j. Such allocations are processed in sequence, and at some point the remaining supply of viewer type i may be less than the $s_i\theta_j$ impressions campaign j is requesting. At such a point, campaign j gets the remaining supply of viewer type i, and subsequent campaigns do not get any impressions from this viewer type. When stage one is completed, some viewer types may be completely allocated, and yet some campaigns may still have demand that has not been allocated. This can happen whenever a campaign cannot grab all $s_i\theta_j$ impressions from each viewer type $i \in \Gamma(j)$ that it wants. Stage two then iterates through the list of campaigns once again, but this time each campaign j checks to see how many impressions in each matching viewer type $i \in \Gamma(j)$ are still available, and grabs the same proportion of slack from each matching viewer type. Ideally, after booking this proportion of slack, the demand for campaign j is fully allocated. However, it could be that the proportion of slack needed to satisfy demand would exceed one; in that case, the campaign grabs all remaining slack from each matching viewer type, and remains partially unallocated. We denote the solution computed by this heuristic as $\left\{\{x_{ij}^0\}_{(i,j)\in\Gamma}, \{y_j^0\}_{j\in J}\right\}$, and compute the values corresponding to each j-component $\{v_j^0\}_{j\in J}$ using the terms corresponding to campaign j in the Gini objective (6); i.e., for $n = 0$ we compute:

$$v_j^n := \psi_j \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj}^n - x_{ij}^n| + p_j y_j^n. \tag{12}$$

We then use this heuristic solution to initialize the first set of columns, corresponding to $n = 0$, in the following Dantzig-Wolfe master problem, which has parameters x_{ij}^n, v_j^n, $n = 0..N$, $j \in J$, $i \in \Gamma(j)$, and decision variables

$\lambda_{jn}, n = 0..N, j \in J$:

$$
\begin{aligned}
\text{(PM)} \qquad z^{UB} = \min & \sum_{j \in J, n=0..N} v_j^n \lambda_{jn} & & \\
\text{s.t.} & \sum_{n=0..N, j \in \Gamma(i)} x_{ij}^n \lambda_{jn} \leq 1 & \forall i \in I & \quad \text{(supply)} \\
& \sum_{n=0..N} \lambda_{jn} = 1 & \forall j \in J & \quad \text{(convexity)} \\
& \lambda_{jn} \geq 0 \, \forall j \in J, \, \forall n = 0..N & &
\end{aligned}
$$

A naïve version of our decomposition method can be described as follows. Starting with $N = 0$, we solve the master problem (PM). Then, we set the price vector $\hat{\beta} \equiv \{\hat{\beta}_i\}_{i \in I}$ to be equal to the dual values of the supply constraints for each viewer type $i \in I$, and solve the supply-relaxed subproblems (PS-j) for all campaigns $j \in J$ analytically, using $\hat{\beta}$ and Theorem 1. We increment the iteration counter $N := N + 1$, and record the solutions to the supply-relaxed subproblems as $\left\{\{x_{ij}^N\}_{(i,j) \in \Gamma}, \{y_j^N\}_{j \in J}\right\}$. We then compute the values $\{v_j^N\}_{j \in J}$ of these solutions using Equation (12), and add the columns (variables and associated terms) corresponding to iteration N for all campaigns $j \in J$ to the master problem, and iterate. At each iteration, we solve the master problem (PM) and subproblems (PS-j) for all campaigns $j = 1..|J|$ in this fashion until we satisfy a termination criterion.

The master problem (PM) encodes a solution to our original problem (P-ORIG) that is a convex combination of all subproblem-j solutions produced in past iterations $n = 0..N$, i.e., $x_{ij} = \sum_{n=0..N} \lambda_{jn} x_{ij}^n$, for all $(i, j) \in \Gamma$ (the λ_{jn} variables determine the weight to apply to the solution produced by subproblem j in iteration n). This solution is feasible in (P-ORIG), since the demand and non-negativity constraints are satisfied by all subproblem solutions and by construction the master problem ensures that the convex combination additionally satisfies the supply constraints that were relaxed in the subproblems. Moreover, the initial solution $\left\{\{x_{ij}^0\}_{(i,j) \in \Gamma}, \{y_j^0\}_{j \in J}\right\}$ is feasible in (P-ORIG), which guarantees (PM) always has a feasible solution.

The master problem (PM) yields an upper bound and the subproblems collectively yield the lower bound $z^{LB}(\hat{\beta})$. Therefore, at each iteration, we compute the optimality gap $(z^{UB} - z^{LB})/z^{LB}$, and terminate when it is below a desired threshold. In general, Dantzig-Wolfe decomposition is known to converge to optimality; however, its rate of convergence depends on the problem being solved. In our case, because the subproblems are analytically solvable, the master problem is the computational bottleneck. Therefore, it makes sense to either avoid solving the master problem at each iteration, or to reduce the size of the master problem by limiting the number of solutions N that we take a convex combination of. It turns out that a good way to speed up our Dantzig-Wolfe-based decomposition method is to solve the master problem (PM) only every fifth iteration, and to employ subgradient optimization to update the prices $\hat{\beta}_i$, $i \in I$, for four out of every five iterations.

To use subgradient optimization, we note that the i^{th} component of the subgradient is simply the negative slack of the i^{th} supply constraint; i.e., it is $g_i := \sum_{j \in \Gamma(i)} x_{ij} - 1$. Using the step size σ_n at iteration n, we update $\hat{\beta}$ as follows: $\hat{\beta}_i := \hat{\beta}_i + \sigma_n g_i$. We use step sizes $\sigma_n = 2(z^{UB} - z^{LB}(\hat{\beta}))/(\sum_{i \in I} s_i g_i^2)$, which have been shown (c.f., Fisher 2004) to make steady progress toward the optimal $\hat{\beta}$. After updating $\hat{\beta}$, we solve subproblems (PS-j) for all campaigns $j \in J$ just as before, and add columns (solutions proposed by the subproblems) to the master problem. Note that, for an iteration that we use subgradient optimization, even though we do not solve the master problem it still grows in size. We now have all the ingredients to state our algorithm for solving (P-ORIG).

Gini-Based Decomposition Method. Starting at iteration $N = 0$, we (1) initialize the master problem (PM) with a feasible heuristic solution, and then (2) increment the iteration counter ($N := N + 1$). Every fifth iteration, we (3a) solve the master problem (PM) to get a new price vector $\hat{\beta}$, a near-optimal feasible solution, and the value of this solution which serves as an upper bound z^{UB} for (P-ORIG). For four out of every five iterations, we instead (3b) employ subgradient optimization to update the values of $\hat{\beta}_i$, $i \in I$, computed in the past iteration. We then (4) use Theorem 1, which leverages the Gini structure of our problem, to solve all campaign-j subproblems analytically. We record the proposed solutions (x_j^n, y_j^n) for each campaign j as well as the lower bound z^{LB}, and then (5) update either the lower bound or the upper bound, or both, if they are tighter than in the previous iteration. We (6) compute the optimality gap $(z^{UB} - z^{LB})/z^{LB}$, and stop if we have attained the termination criterion (e.g., optimality gap $< 1\%$), or continue to iterate (go to step 2) otherwise. Our decomposition method terminates with the feasible solution $x_{ij} = \sum_{n=0..N} \lambda_{jn} x_{ij}^n$, $(i,j) \in \Gamma$, which is near-optimal.

6.4. Solution Structure

Not only is our decomposition method an efficient algorithm for finding near-optimal solutions to the Gini ad allocation problem (SG), but it also characterizes the structure of the optimal solution, as we will now show. Let (x^*, y^*) with $x^* \equiv \{x_{ij}^*\}_{(i,j) \in \Gamma}$ and $y^* \equiv \{y_j^*\}_{j \in J}$ be an optimal solution of (SG), and denote the campaign-j component of the optimal solution as (x_j^*, y_j^*), where $x_j^* \equiv \{x_{ij}^*\}_{i \in \Gamma(j)}$. We begin with the following definition.

DEFINITION 5. **Basic Solution**. An allocation (x_j, y_j) to campaign j is considered *basic* if either

(i) No demand is met, i.e., y_j = d_j and $x_{ij} = 0 \, \forall i \in \Gamma(j)$, or

(ii) All demand is met, and the demand is spread evenly over a subset of the audience segments that campaign j targets. Formally, this means $y_j = 0$ and there exists a subset $B_j \subseteq \Gamma(j)$ of the audience segments such that $x_{ij} = d_j / \tilde{s}_j$ for all $i \in B_j$ and $x_{ij} = 0$ for all $i \notin B_j$. Moreover, we have used $\tilde{s}_j = \sum_{i \in B_j} s_i$ to denote the total supply coming from all audience segments in the set B_j.

Our decomposition method may be used to express the optimal allocation for each campaign j, i.e., (x_j^*, y_j^*), as a convex combination of one or more basic solutions. This follows directly from the structure of our decomposition method because (1) Theorem 1 tells us that each supply-relaxed campaign-j subproblem produces a basic solution, (2) at each iteration, the Dantzig-Wolfe master problem (PM) produces a feasible solution which is a convex combination of (possibly) a heuristic solution (which may be chosen basic) and some subproblem solutions (which are themselves basic), and (3) given sufficiently many iterations, the Dantzig-Wolfe method converges to optimality. Note that using the heuristic solution described in §6.3 yields good convergence but is not basic. Although the heuristic solution is often no longer part of the convex combination when our algorithm terminates, strictly speaking if one wants to ensure convergence to a solution that is a convex combination of basic solutions, one can instead initialize the Dantzig-Wolfe master problem with the trivially basic solution $x_{ij} = 0$ for all $(i, j) \in \Gamma$, $y_j = d_j$ for all $j \in J$.

Assuming we initialize our algorithm with a basic solution and converge to optimality after N iterations, our algorithm produces (i) prices $\hat{\beta}_i^n$ for all viewer types i and all iterations n, (ii) weights λ_{jn} for all campaigns j and iterations n, and (iii) basic solutions (x_j^n, y_j^n) for all campaigns j and iterations n. Denoting the set N_j as the subset of iterations $n = 0..N$ for which the weight λ_{jn} is strictly positive, we can express the optimal allocation for campaign j as:

$$x_{ij}^* = \sum_{n \in N_j} \lambda_{jn} x_{ij}^n \text{ for all } i \in I, \text{ and} \tag{13}$$

$$y_j^* = \sum_{n \in N_j} \lambda_{jn} y_j^n. \tag{14}$$

The above characterization allows us to view the optimal solution as being generated by randomizing over a set of prices. Specifically, with probability λ_{jn}, campaign j is subject to internal prices $\hat{\beta}_i^n$ for audience segments $i \in I$, and must "pay" these prices for the impressions it gets from each audience segment. These internal prices correspond to the dual variables for the supply constraints, and one vector of internal prices is computed in each iteration of the Dantzig-Wolfe procedure. Given a vector of internal prices $\hat{\beta} \equiv \{\hat{\beta}_i\}_{i \in I}$, we can identify, using Theorem 1, the audience segments that are relatively cheap for campaign j (this becomes the set B_j), and the audience segments that are relatively expensive for campaign j (these remain in the set $\Gamma(j) \setminus B_j$). In Appendix D we provide an online algorithm that exploits this interpretation to serve well-spread ads in real-time.

The following simple example provides an illustration. We are given 2 campaigns (A, B) and 2 audience segments (1, 2), demands are $d_A = 60$ and $d_B = 800$, supplies are $s_1 = 80$ and $s_2 = 800$, campaign A targets segment 1 only, and campaign B targets segments 1 and 2. Per-impression shortfall penalties p_A and p_B are both 0.01, and we fix $\alpha = 0.009$. For campaign A, our algorithm produces one basic solution, $\{x_A^1 = (3/4, 0), y_A^1 = 0\}$, with

corresponding weight $\lambda_{A1} = 1$. The optimal solution (x_A^*, y_A^*) is precisely this basic solution (i.e., it is a degenerate convex combination of one basic solution). For campaign B, our algorithm produces two basic solutions, $x_B^1 = (10/11, 10/11)$, $x_B^2 = (0, 1)$, $y_B^1 = y_B^2 = 0$, with weights $\lambda_{B1} = 11/40$ and $\lambda_{B2} = 29/40$. The optimal solution is a convex combination of these basic solutions, i.e., $x_B^* = (11/40)(10/11, 10/11) + (29/40)(0, 1) = (1/4, 39/40)$, and $y_B^* = (11/40)(0) + (29/40)(0) = 0$. That is, with probability $\lambda_{B1} = 11/40$, campaign B is subject to low internal prices in both audience segments ($\hat{\beta}_1^1 = \hat{\beta}_2^1 = 0$), making both audience segments affordable. And with probability $\lambda_{B2} = 29/40$, campaign B is subject to a high internal price in segment 1 ($\hat{\beta}_1^2 = 0.1$) as well as a low internal price in segment 2 ($\hat{\beta}_2^2 = 0$), making only segment 2 affordable.

6.5. Computational Results

We now evaluate the computational performance of our decomposition method. For this, we generate a number of test problems using the method described in detail in §5.1. Our test cases here have three sizes: small (100 campaigns $\times$ 100 viewer types), medium (100 campaigns $\times$ 200 viewer types), and large (100 campaigns $\times$ 500 viewer types). As before, we have three categories of test case: loose, globally tight, and locally tight. We represent these three categories notationally as Lx, GTx, and LTx, where the x represents the extent to which demand is inflated, and is either $x = 0$ or $x = 20$ percent. When $x = 20$, we inflate demands so that they are 20% higher than what they would ordinarily be for such an instance, leading to demand shortfalls of up to 20%. Figure 10 lists the properties of this set of instances.

Test Case	Sell-through	Targeting %	Campaign Link Intensity	Viewer Type Link Intensity	Reservation Proportion
Loose (L0)	67.1%	19.6%	U[0.1, 0.4]	U[0.5, 1]	3.75%
Globally Tight (GT0)	95.4%	19.0%	U[0.1, 0.4]	U[0.5, 1]	6.6%
Locally Tight (LT0)	92.4%	3.4%	U[0.05, 0.2]	U[0.1, 0.4]	$*$
Loose (L20)	73.2%	19.5%	U[0.1, 0.4]	U[0.5, 1]	3.75%
Globally Tight (GT20)	106.7%	18.4%	U[0.1, 0.4]	U[0.5, 1]	6.6%
Locally Tight (LT20)	108.0%	3.5%	U[0.05, 0.2]	U[0.1, 0.4]	$*$

Figure 10 Six batches of test cases used to evaluate our decomposition method. In cases marked $*$, reservation proportions for each campaign were drawn from {40%, 70%, and 100%} with probabilities 0.6, 0.3, and 0.1, respectively. Globally tight cases have high sell-through, while locally tight cases have both high sell-through and low targeting percentage.

As seen in Table 2, our decomposition method is very effective. Solution times are on average 60 times faster than when CPLEX is used to solve (SG) directly. Furthermore, we note that our decomposition algorithm can solve all instances in at most 11 minutes, while 22% of the instances cannot be solved by CPLEX in under 2 hours. All tests were run on a 64-bit desktop with Intel(R) Core(TM) i7-2600 processor, running at 3.4GHz CPU

and with 16GB RAM, using CPLEX 12.6 and AMPL 2016.03.10. For all tests, we used a single processor core and terminated the decomposition method when the optimality gap was less than 1%.

Instance	Small (100x100)		Medium (100x200)		Large (100x500)	
Instance	Regular	Decomposition	Regular	Decomposition	Regular	Decomposition
L0	8	1	78	3	7200+	20
GT0	26	1	922	10	7200+	152
LT0	1	1	1	1	28	4
L20	4	1	82	3	7200+	36
GT20	73	1	2132	11	7200+	396
LT20	1	1	1	1	52	5

Table 2 Comparison of solution times (in CPU seconds) with and without the decomposition method. Solution times marked 7200+ indicate cases where CPLEX did not find a solution within 2 hours and was aborted.

7. Extensions

We have also derived several model and algorithmic extensions. In Appendix §B.5, we show how the supply-relaxed subproblem defined in §6.2 can be used directly to solve a market-based variant of our problem, where an ad aggregator does not own any impression traffic itself and must purchase the entire supply of each audience segment from the market, at fixed (estimated) prices. In Appendix §D, we introduce an online algorithm that can be used in conjunction with our decomposition method to serve well-spread ads in real-time. It operates by reconstructing, as-needed on-the-fly, all iterations' solutions x_{ij}^n, $n = 0..N$, according to the probabilities λ_{jn}, and randomizing over these solutions as the Dantzig-Wolfe master problem dictates. Finally, in Appendix §E, we develop a multi-dimensional (i.e., multi-period) Gini model as well as a nested decomposition method that repeatedly invokes Theorem 1 to spread impressions across both audience segments and time.

8. Conclusions

Online advertising is a $72.5 billion market, forty-four percent of which is display advertising, which can be further subdivided into impression-based, click-based, or conversion-based advertising, and orthogonally into guaranteed or non-guaranteed advertising. This paper's focus is on planning guaranteed impression-based ads, in which advertisers' demands for impressions of various types compete for shared resources and advertisers prefer to receive impressions that are evenly-spread across targeted audience segments and across time. We employ the Gini coefficient measure to quantify what is meant by "good spreading", and derive specific Gini-based metrics to measure several dimensions of spreading that advertisers and publishers care about.

We propose a new single-period optimization planning problem that maximizes the spreading of impressions across targeted audience segments while limiting demand shortfalls. We identify five key properties which are

essential in the online display advertising context and have shown that our Gini-based model satisfies all of them. We design a decomposition method that allows for the exact and efficient solution of the Gini formulations that are, to the best of our knowledge, of a size not precedently handled in the literature. Our decomposition scheme incorporates ingredients from the Dantzig-Wolfe decomposition and subgradient optimization methods, and exploits the special structure of the Gini-based ad planning problems. Crucially, we are able to show that after a suitable problem transformation, the subproblems can be solved analytically, which leads to a substantial speedup. Numerical tests on our single-period model show that our decomposition method is on average 60 times faster than solving the Gini-based formulation directly with CPLEX. Finally, we extend our main result, which allows us to solve single-period subproblems analytically, to a multi-period model using a Lagrangian Decomposition scheme. We also discuss how our decomposition framework can be used by an aggregator that buys impressions in the market and allocates them to advertisers, sketch an online algorithm which can be used with our decomposition method to serve well-spread ads in real-time, and suggest that our decomposition method, which is versatile enough to be used for the solution of other Gini-related optimization problems, may be employed as the foundation for large-scale parallelized implementations.

The analysis of the results provided by our model and their comparison with results derived from a popular ad planning model developed at Yahoo allow us to draw insights about the usefulness of the Gini coefficient and to quantify the potential benefits of using Gini-based metrics for planning online advertising. Using numerical experiments, we compare and contrast the solutions from our Gini-based model with this baseline model, and show that our Gini model is generally better at trading off revenue to achieve better spreading. As we have shown, the extent to which a campaign's impressions are well-spread can be easily visualized using a Lorenz curve. By minimizing Gini coefficients, we effectively minimize the area above the Lorenz curves. We interpret these Lorenz curves, and discuss how they can be used in practice, both by the publisher, to monitor ad delivery performance, and by the advertiser, to review past ad delivery performance.

Acknowledgments

The authors thank both area editors who handled this manuscript, Harikesh Nair and Gustavo Vulcano, as well as the anonymous associate editor and three referees, whose comments decidedly improved the manuscript. Many thanks also to those who provided useful feedback at POMS 2016, RM&P 2016, CORS 2016, INFORMS 2016, INFORMS 2018, and the Southern California OR/OM Day 2016, as well as during seminars at Stanford GSB, University of Chicago (Booth), University of Washington (Foster), University of Miami Business School, Uni-

versity of Illinois at Chicago CBA, Santa Clara University (Leavey), University of California at Irvine (Merage), and University of California at San Diego (Rady).

Endnotes

1. Advertisers often prefer if 2 individuals each receive 1 impression rather than for 1 individual to receive 2 impressions. Although display advertising is commonly planned using impressions, the number of unique individuals that see an ad, called *reach*, is another metric that is also important to advertisers.

2. Although Bharadwaj et al. (2012) define the demand constraint as $\sum_{i\in\Gamma(j)} s_i x_{ij} + y_j \geq d_j$ for convenience when deriving the dual, it is easily proven that it is never optimal to over-deliver; hence, we can equivalently represent the demand constraint with an equality, which is a clearer representation for our purposes.

3. Strictly speaking, this assumes each arriving user sees exactly one impression. Generalizing the model to multiple impressions per arrival is possible (e.g., see Turner 2012), but we ignore that possibility to keep the model accessible.

4. If we draw a negative realization we discard it, and continue drawing until we get a positive one.

5. For datasets, please visit `https://faculty.sites.uci.edu/turnerjg/datasets/gini/`.

6. To model loss of goodwill inherent in issuing partial refunds for less-than-contracted service, we may generalize the revenue metric to $R = \sum_j p_j d_j - \eta \sum_j p_j y_j$ by using $\eta > 1$.

7. Since both audience segments and campaigns have different sizes, we count each residual r_{ij} exactly $s_i \times w_j$ times. This normalization is equivalent to splitting each supply node of size s_i into s_i supply nodes of size 1, and each demand node with size w_j into w_j demand nodes of size 1.

8. Note that the minimization objectives of both models (Equations 3a, 6) now include an additional term that subtracts off the revenue from the outside option, $\sum_i \tilde{\beta}_i s_i \left(1 - \sum_{j\in\Gamma(i)} x_{ij}\right)$.

References

Agrawal S., Z. Wang, Y. Ye. 2014. A dynamic near-optimal algorithm for online linear programming. *Operations Research* 62(4): 876-890.

Allouah A., Besbes O. 2017. Auctions in the online display advertising chain: A case for independent campaign management. *Working Paper*: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2919665`.

Anstreicher K.M., L.A. Wolsey. 2009. Two well-known properties of subgradient optimization. *Mathematical Programming* 120, 213-220.

Araman V. F., K. Fridgeirsdottir. 2015. Cost-per-Impression Pricing and Campaign Delivery for Online Display Advertising. Working Paper. Available at `http://pages.stern.nyu.edu/~varaman/Research/OnlineAd.pdf`

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

Atkinson A.B. *The economics of inequality*. Oxford: Clarendon Press, 1975.

Balseiro S., J. Feldman, V. Mirrokni, S. Muthukrishnan. 2014. Yield optimization of display advertising with ad exchange. *Management Science* 60(12), 2886-2907.

Balseiro S., O. Besbes, G. Weintraub. 2015. Repeated auctions with budgets in ad exchanges: Approximations and design. *Management Science* 61(4), 864-884.

Balseiro S., O. Candogan, H. Gurkan. 2015. Multi-stage intermediation in online internet advertising (September 15, 2015). Available at SSRN: `http://ssrn.com/abstract=2661459` or `http://dx.doi.org/10.2139/ssrn.2661459`

Bhalgat V, J. Feldman, V. Mirrokni. 2012. Online allocation of display ads with smooth delivery. *Proceedings of the 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*: 1213-1221.

Bharadwaj V., W. Ma, M. Schwarz, J. Shanmugasundaram, E. Vee, J. Xie, J. Yang. 2010. Pricing guaranteed contracts in online display advertising. *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*.

Bharadwaj V., P. Chen, W. Ma, C. Nagarajan, J. Tomlin, S. Vassilvitskii, E. Vee, J. Yang. 2012. SHALE: An efficient algorithm for allocation of guaranteed display advertising. In: *Proceedings of the* 18^{th} *ACM International Conference on Knowledge Discovery and Data Mining*, 1195-1203.

Biglione S. 2016. “To target or not to target, that is not the question." *LinkedIn Pulse* April 15, 2016. `https://www.linkedin.com/pulse/target-question-shann-biglione`

Bollapragada S., H. Cheng, M. Phillips, M. Garbiras, M. Scholes, T. Gibbs, M. Humphreville. 2002. NBC’s optimization systems increase its revenues and productivity. *Interfaces* 32(1) 47-60.

Campano F., D. Salvatore. 2006. *Income distribution*. Oxford: Oxford University Press.

Chakraborty T., E. Even-Dar, S. Guha, Y. Mansour, S. Muthukrishnan. 2010. Selective call out and real time bidding. In *Internet and Network Economics* pp. 145-157. Springer Berlin Heidelberg.

Chen P., W. Ma, S. Mandalapu, C. Nagarjan, J. Shanmugasundaram, S. Vassilvitskii, E. Vee, M. Yu, J. Zien. 2012. Ad serving using a compact allocation plan. *Proceedings of the 13th ACM Conference on Electronic Commerce*. pp. 319-336.

Chen Y. 2013. Optimal dynamic auctions for display advertising (February 12, 2013). Available at SSRN: http://ssrn.com/abstract=2216361 or http://dx.doi.org/10.2139/ssrn.2216361

Dantzig G.B., P. Wolfe. 1960. Decomposition principle for linear programs. *Operations Research* 8 (1): 101-111.

Devanur N.R., T.P. Hayes. 2009. The adwords problem: Online keyword matching with budgeted bidders under random permutations. In *EC’09 Proceedings of the 10th ACM Conference on Electronic Commerce*. Stanford, CA, 71-78.

Drezner T., Z. Drezner, J. Guyse. 2009. Equitable service by a facility: Minimizing the Gini coefficient. *Computers & Operations Research* 36 (12), 3240-3246.

Dütting P., M. Henzinger, I. Weber. 2011. An expressive mechanism for auctions on the web. In *Proceedings of the 20th International Conference on World Wide Web*, 127-136. .

Feldman J., M. Henzinger, N. Korula, V.S. Mirrokni, C. Stein. 2010. Online stochastic packing applied to display ad allocation. In *Proceedings of European Symposium on Algorithms*, 182-194.

Fisher M. 2004. The Lagrangian relaxation method for solving integer programming problems. *Management Science* 50 (12): 1861-1871.

Ghosh A., P. McAfee, K. Papineni, and S. Vassilvitskii. 2009. Bidding for representative allocations for display advertising. In: *Workshop on Internet and Network Economics (WINE)*, LNCS 5929, 208-219. Berlin: Springer.

Gini C. 1912. *Variabilitá e mutabilita*. Reprinted in memorie di metodologica statistica. Eds: Pizetti, E., T. Salvemini. Rome: Libreria Eredi Virgilio Veschi 1955.

Golrezaei N., Lin M., Mirrokni V., Nazerzadeh H. 2017. Boosted second price auctions: Revenue optimization for heterogeneous bidders. *Working Paper*: `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3016465`.

Gustavsson E., M. Patriksson, A.-B. Stromberg. 2015. Primal convergence from dual subgradient methods for convex optimization. *Mathematical Programming* 150 (2): 365-390.

Hettmansperger, T.P. 1984. Statistical Inference based on Ranks. *Series in Probability and Mathematical Statistics*, John Wiley & Sons Inc., New York.

Held M., P. Wolfe, H.D. Crowder. 1974. Validation of subgradient optimization. *Math. Programming* 6: 62-88.

Hojjat A., J. Turner, S. Cetintas, J. Yang. 2017. A Unified Framework for the Scheduling of Guaranteed Targeted Display Advertising under Reach and Frequency Requirements. *Operations Research*. Published Online in Advance of Print at `http://dx.doi.org/10.1287/opre.2016.1567`.

Hosking J.R.M. 1990. L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society, Series B* 52: 105-124.

Internet Advertising Bureau. 2017. IAB Internet Advertising Revenue Report, April 2017. (Available online at `http://www.iab.net/research`).

Lai D., Huang J., Risser J., Kapadia A.S. 2008. Statistical properties of generalized Gini coefficient with application to health inequality measurement. *Social Indicators Research* 87(2): 249-258.

Lee K.-C., A. Jalali, A. Dasdan. 2013. Real time bid optimization with smooth budget delivery in online advertising. *Proceedings of the The Seventh International Workshop on Data Mining for Online Advertising*.

Lorenz M.O. 1905. Methods for measuring the concentration of wealth. *Journal of the American Statistical Association* 9: 209-219.

Mansini R., W. Ogryczak, M.G. Speranza. Tail Gini's risk measures and related linear programming models for portfolio optimization. *Research Paper*. `http://www.aueb.gr/pympe/hercma/proceedings2007/H07-ABSTRACTS-1/MANSINI-OGRYCZAK-SPERANZA-1.pdf`

McAfee R. Preston, K. Papineni, and S. Vassilvitskii. 2013. Maximally representative allocations for guaranteed delivery advertising campaigns. *Review of Economic Design* 17.2: 83-94.

Mehta A. 2012. Online matching and ad allocation. *Theoretical Comput. Sci.* 8(4): 265-368.

Mookerjee R., S. Kumar, V. Mookerjee. 2016. Optimizing Performance-Based Internet Advertisement Campaigns. *Operations Research* Published online in advance of print on December 5, 2016. DOI:10.1287/opre.2016.1553.

Muthukrishnan S. 2009. Ad exchanges: Research issues. In *Internet and network economics*. Springer Berlin Heidelberg, 1-12.

Ogryczak W., A. Ruszczyński. 1999. From stochastic dominance to mean-risk models: Semi-deviations as risk measures. *European Journal of Operational Research* 116: 33-50.

Ogryczak W., A. Ruszczyński. 2002. Dual stochastic dominance and related mean-risk models. *SIAM Journal on Optimization* 13: 60-78.

Ogryczak W., A. Ruszczyński. 2002. Dual stochastic dominance and quantile risk measures. *International Transactions in Operational Research* 9: 661-680.

Shalit H., S. Yitzhaki. 1984. Mean-Gini, portfolio theory, and the pricing of risky assets. *Journal of Finance* 39: 1449-1468.

Shalit H., S. Yitzhaki. 2005. The Mean-Gini efficient portfolio frontier. *The Journal of Financial Research* 18: 59-95.

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

Turner J. 2012. The planning of guaranteed targeted display advertising. *Operations Research* 60(1): 18-33.

Vee E., S. Vassilvitskii, J. Shanmugasundaram. 2010. Optimal online assignment with forecasts. Parkes D.C., C. Dellarocas, M. Tennenholtz, eds. *Proc. 11th ACM Conf. Electronic Commerce, EC '10* (ACM, New York), 109-118.

Yang J., E. Vee, S. Vassilvitskii, J. Tomlin, J. Shanmugasundaram, T. Anastasakos. 2010. Inventory allocation for online graphical display advertising. *Yahoo! Labs Technical Report* YL-2010-004. Available at `https://arxiv.org/abs/1008.3551`.

Yitzhaki S. 1982. Stochastic dominance, mean variance and Gini's mean difference. *American Economic Review* 72: 178-185.

Yitzhaki S. 1983. On an extension of the Gini inequality index. *International Economic Review* 24: 617-628.

Yitzhaki, S. 2003. Gini's mean difference: A superior measure of variability for non-normal distributions. *Metron* 61(2): 285–316.

Yitzhaki S., E. Schechtman. 2013. *The Gini methodology: A primer on a statistical methodology*. New York: Springer.

Yuan S., J. Wang, X. Zhao. 2013. Real-time bidding for online advertising: measurement and analysis. In *Proceedings of the Seventh International Workshop on Data Mining for Online Advertising* p. 3. ACM.

Xu J., K.-C. Lee, W. Li, H. Qi, Q. Lu. 2015. Smart pacing for effective online ad campaign optimization. *Proceedings of the 21st ACM SIGKDD Conference on Knowledge Discovery and Data Mining*.

Miguel A. Lejeune is a Professor of Decision Sciences at the George Washington University. He is a recipient of the CAREER Grant from the Army Research Office and of the IBM Smarter Planet Faculty Innovation Award. He is a member of the Board of the Stochastic Programming Society. His research interests include stochastic programming, distributionally robust optimization, applied and data-driven optimization, and financial risk.

John Turner is an Associate Professor of Operations and Decision Technologies at the Paul Merage School of Business, University of California at Irvine. He is a recipient of the George B. Dantzig Dissertation Award, the

William Pierskalla Award for Health Care Management Science, and was named a Yahoo Faculty Research & Engagement Program Scholar. Turner's research interests include applied optimization, decomposition methods for solving large-scale math programs, advertising planning, revenue management, and healthcare optimization.

Appendices

Appendix A: Notation, Derivations, and Proofs

A.1. Notation for Audience-Level Models & Expressions

Table EC.1 is a reference for the main notation used in the paper, which concerns models that plan blocks of impressions known as 'audience segments' or 'viewer types.'

Index	**Description**
h, i	Audience segment
j	Ad campaign
t, τ	Time period
Parameter (Single-Period Models)	**Description**
I	Set of audience segments
J	Set of ad campaigns
$\Gamma(i)$	Set of campaigns that target audience segment i
$\Gamma(j)$	Set of audience segments targeted by campaign j
Γ	Set of (audience segment, campaign) (i,j) pairs that define all campaigns' targeting requirements
d_j	Demand (in impressions) of ad campaign j
p_j	Penalty assessed for each impression of demand shortfall for campaign j
s_i	Supply (in impressions) of audience segment i
$\hat{s}_j$	Supply (in impressions) eligible for campaign j
θ_j	Demand intensity of campaign j
Decision Variable (Single-Period Models)	**Description**
x_{ij}	Proportion of the impressions of audience segment i to assign to campaign j
y_j	Demand shortfall (in impressions) for campaign j

Table EC.1 Notation

A.2. Notation for Impression-Level Models & Expressions

Table EC.2 is a reference to additional notation used in parts of this Appendix, where audience segments are treated as collections of individual impressions to express impression-level quantities for the proofs.

A.3. Full Derivation of Gini-Based Metric

We now derive our Gini-based metric (5), which measures the degree to which impressions are spread across audience segments, from first principles. We begin in the disaggregated impression-level space, where x_{rj} is the proportion of impression (arrival) r to assign to campaign j, and can be alternatively interpreted as the probability of assigning impression r to campaign j. With $\Lambda(j)$ defined as the set of impressions that match the

Index	**Description**
q, r	Impression (arrival)
Parameter (Single-Period Models)	**Description**
R	Set of impressions (arrivals)
R_i	Set of impressions that comprise audience segment i
$\Lambda(r)$	Set of campaigns that target impression r
$\Lambda(j)$	Set of impressions targeted by campaign j
Λ	Set of (impression, campaign) (r, j) pairs that define all campaigns' targeting requirements
Decision Variable (Single-Period Models)	**Description**
x_{rj}	Proportion of impression r to assign to campaign j (can be interpreted as the probability of assigning impression r to campaign j)
y_j	Demand shortfall (in impressions) for campaign j

Table EC.2 Notation for Impression-Level Expressions

targeting of campaign j, the following GMD metric, which measures how evenly-spread a campaign's ads are across impressions in a campaign's target audience, follows directly from the definition of GMD (2):

$$GMD_j = \frac{1}{|\Lambda(j)|^2} \sum_{q \in \Lambda(j)} \sum_{r \in \Lambda(j)} |x_{qj} - x_{rj}|. \tag{EC.1}$$

By aggregating impressions into mutually exclusive audience segments, and using x_{ij} as the proportion of audience segment i to assign to campaign j, we simplify equation (EC.1) as follows. For notational convenience, we order the audience segments and define $\Gamma_0(j) = \{(h,i) \in \Gamma(j)^2 : h < i\}$, which indexes all distinct audience segment pairs.

PROPOSITION EC.1. *In the aggregated audience-level space, the GMD metric corresponding to* (EC.1) *is:*

$$GMD_j = \frac{2}{\hat{s}_j^2} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}|. \tag{EC.2}$$

Proof. Let R_i be the set of impressions represented by the audience segment i. Then we have (i) $s_i = |R_i|$, (ii) $\Lambda(j) = \cup_{i \in \Gamma(j)} R_i$, (iii) $\hat{s}_j = |\Lambda(j)| = \sum_{i \in \Gamma(j)} |R_i| = \sum_{i \in \Gamma(j)} s_i$, and (iv) $x_{rj} = x_{ij}\ \forall r \in R_i$, allowing us to simplify equation (EC.1) as follows:

$$\begin{aligned}
GMD_j &= \frac{1}{|\Lambda(j)|^2} \sum_{q \in \Lambda(j)} \sum_{r \in \Lambda(j)} |x_{qj} - x_{rj}| = \frac{1}{|\Lambda(j)|^2} \sum_{q \in \{\cup_{h \in \Gamma(j)} R_h\}} \sum_{r \in \{\cup_{i \in \Gamma(j)} R_i\}} |x_{qj} - x_{rj}| \\
&= \frac{1}{|\Lambda(j)|^2} \sum_{h \in \Gamma(j)} \sum_{i \in \Gamma(j)} \left(\sum_{q \in R_h} \sum_{r \in R_i} |x_{qj} - x_{rj}| \right) = \frac{1}{|\Lambda(j)|^2} \sum_{h \in \Gamma(j)} \sum_{i \in \Gamma(j)} \left(\sum_{q \in R_h} \sum_{r \in R_i} |x_{hj} - x_{ij}| \right) \\
&= \frac{1}{\hat{s}_j^2} \sum_{h \in \Gamma(j)} \sum_{i \in \Gamma(j)} s_h s_i |x_{hj} - x_{ij}| = \frac{2}{\hat{s}_j^2} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}|.
\end{aligned}$$

□

The $s_h s_i$ factor enters into this expression since the number of ways we can pick a pair of impressions such that one is from audience segment h and the other is from audience segment i is precisely s_h times s_i.

Moreover, in a similar manner we can also derive the average proportion μ_j of an impression assigned to campaign j from impression-level quantities as follows:

$$\mu_j = \frac{1}{|\Lambda(j)|}\sum_{r\in\Lambda(j)} x_{rj} = \frac{1}{|\Lambda(j)|}\sum_{i\in\Gamma(j)}\sum_{r\in R_i} x_{rj} = \frac{1}{|\Lambda(j)|}\sum_{i\in\Gamma(j)}\sum_{r\in R_i} x_{ij} = \frac{1}{\hat{s}_j}\sum_{i\in\Gamma(j)} s_i x_{ij}. \tag{EC.3}$$

Finally, from (2), our Gini-based metric is $G_j = GMD_j/(2\mu_j)$.

A.4. Proof of Proposition 2: Existence of Ideal Impression Allocations

Proposition 2: Let $p_j = p, j \in J$. If the system (11a)-(11d) is feasible,

1. The optimal solution of the Gini model (SG) is always an ideal allocation. This statement is valid regardless of the emphasis placed on spreading determined by the parameter $\alpha > 0$.

2. There is no guarantee that the optimal solution of the baseline model (SB) is an ideal allocation.

Proof. Part 1) Regardless of whether perfect spreading is possible or not, the quantity pY is a valid lower bound on the optimal value of the objective function (9) of the Gini model. Any solution $\mathbf{x}$ feasible for (11a)-(11d) gives a value of (9) equal to pY and is therefore optimal. Additionally, any solution $\mathbf{x}'$ not feasible for (11a)-(11d) will either: 1) violate the minimum shortfall constraint (11b), thereby increasing the shortfall cost component in (9) without reducing the spreading cost component (i.e., since any feasible $\mathbf{x}$ has perfect spreading and minimal spreading cost equal to 0); or 2) violate the perfect spreading condition, thereby increasing the spreading cost component in (9) without reducing the shortfall cost component (i.e., since any feasible $\mathbf{x}$ has minimal spreading cost).

Part 2) Any solution $\mathbf{x}$ feasible for (11a)-(11d) minimizes the shortfall cost component in the objective function (3) of the baseline model. However, there is no guarantee that the spreading term in (3) is minimized by $\mathbf{x}$. A numerical counterexample is provided after Proposition 2. □

A.5. Proof of Proposition 3: Existence of Sufficient Orthogonality

Proposition 3:

1. The Gini objective (6) has sufficiently orthogonal spread and shortfall measures.

2. The baseline objective (3a) doesn't have sufficiently orthogonal spread and shortfall measures.

Proof. Part 1) To solve for a minimizer $\mathbf{x_j^*}$ of $f_j^{SPREAD*}(c)$, we begin by noting that the feasibility condition $f_j^{SHORTFALL}(\mathbf{x_j}) = c$ can be phrased as a restriction on the mean allocation $\mu_j = \sum_{i\in\Gamma(j)} s_i x_{ij}/\hat{s}_j$, since:

$$\begin{aligned} &f_j^{SHORTFALL}(\mathbf{x_j}) = c \implies p_j y_j = c \implies p_j(d_j - w_j) = c \\ &\implies (d_j - w_j)/\hat{s}_j = c/(p_j\hat{s}_j) \implies (d_j/\hat{s}_j) - (w_j/\hat{s}_j) = c/(p_j\hat{s}_j) \\ &\implies \theta_j - \mu_j = c/(p_j\hat{s}_j) \implies \mu_j = \theta_j - c/(p_j\hat{s}_j). \end{aligned}$$

Next, we claim that the solution $x_{ij}^* = \theta_j - c/(p_j\hat{s}_j)$ for all $i \in \Gamma(j)$ is a minimizer of $f_j^{SPREAD*}(c)$. By definition, the mean allocation μ_j^* under the solution x_{ij}^* is:

$$\mu_j^* = \sum_{i\in\Gamma(j)} s_i x_{ij}^*/\hat{s}_j = \sum_{i\in\Gamma(j)} s_i\left(\theta_j - c/(p_j\hat{s}_j)\right)/\hat{s}_j = \theta_j - c/(p_j\hat{s}_j),$$

which clearly satisfies $f_j^{SHORTFALL}(\mathbf{x_j}) = c$. As well, $f_j^{SPREAD*}(c) = \alpha w_j^* G_j^* = \alpha \hat{s}_j GMD_j^*/2 = 0$, where the last equality follows because $GMD_j^* = 0$ whenever x_{ij}^* takes the same value for all $i \in \Gamma(j)$ (see Eq. 4). Because $GMD_j \geq 0$, we must always have $f_j^{SPREAD*}(c) \geq 0$. Consequently, the solution $\{x_{ij}^* = \theta_j - c/(p_j \hat{s}_j)$ for all $i \in \Gamma(j)\}$ minimizes $f_j^{SPREAD*}(c)$ with the value of $f_j^{SPREAD*}(c) = 0$.

Part 2) To solve for the minimizer $\mathbf{x_j}$ in the definition of $f_j^{SPREAD*}(c)$, we use the fact that $y_j = d_j - \sum_{i\in\Gamma(j)} s_i x_{ij}$ to define the Lagrangian $L(\mathbf{x_j}, \zeta_j) = \sum_{i\in\Gamma(j)} \frac{V_j}{2\theta_j} s_i (x_{ij} - \theta_j)^2 + \zeta_j (p_j (d_j - \sum_{i\in\Gamma(j)} s_i x_{ij}) - c)$. From the stationarity condition $\partial L(\mathbf{x_j^*}, \zeta_j^*)/\partial x_{ij} = 0$, we have $\frac{V_j}{\theta_j} s_i (x_{ij}^* - \theta_j) - \zeta_j^* p_j s_i = 0$; thus, $x_{ij}^* = \theta_j + \zeta_j^* \frac{\theta_j p_j}{V_j}$ for all $i \in \Gamma(j)$. Substituting this x_{ij}^* into the feasibility condition $f_j^{SHORTFALL}(\mathbf{x_j}) = c$, which in this case is $p_j(d_j - \sum_{i\in\Gamma(j)} s_i x_{ij}^*) = c$, yields $\zeta_j^* = -cV_j/(d_j p_j^2)$. Since $x_{ij}^* = \theta_j + \zeta_j^* \frac{\theta_j p_j}{V_j}$ for all $i \in \Gamma(j)$, we have $x_{ij}^* - \theta_j = \zeta_j^* \frac{\theta_j p_j}{V_j} = -c/(p_j \hat{s}_j)$, and therefore $f_j^{SPREAD*}(c) = \sum_{i\in\Gamma(j)} \frac{V_j}{2\theta_j} s_i (x_{ij}^* - \theta_j)^2 = \frac{V_j}{2\theta_j} \hat{s}_j (-c/(p_j \hat{s}_j))^2 = c^2 V_j/(2 d_j p_j^2)$. Finally, $\partial f_j^{SPREAD*}(c)/\partial c = cV_j/(d_j p_j^2) > 0$ for all $c > 0$. Since $f_j^{SPREAD*}(c)$ is an increasing function of c for all $c > 0$, we have a contradiction which proves the proposition. □

A.6. Proof of Proposition 4: Existence of Split-and-Merge Invariance

Proposition 4:

1. If the parameters V_j in the baseline objective function (3a) are chosen independently of campaign demands d_j, then the optimal solution to the baseline model (SB) is not affected by arbitrary campaign splits or merges.
2. The optimal solution to the Gini model (SG) is not affected by arbitrarily splitting or merging campaigns.

Preliminaries: The following proofs use the following setup. Consider a publisher that minimizes the cost of spreading impressions over audience segments, defined as the sum of campaign-specific terms $F_j := W_j f_j$, where W_j is a campaign-specific scaling factor and f_j is either $\sum_{i\in\Gamma(j)} s_i (x_{ij} - \theta_j)^2$ for the baseline model or GMD_j for the Gini model. We have a large campaign C which we are considering splitting into two smaller campaigns A and B so that (i) campaigns A and B inherit the targeting of campaign C, i.e., $\Gamma(A) = \Gamma(B) = \Gamma(C)$, implying $\hat{s}_A = \hat{s}_B = \hat{s}_C$; and (ii) the demand of campaign C is equal to the total demand of campaigns A and B, i.e., $d_C = d_A + d_B$. Therefore, by definition $\theta_A = (d_A/d_C)\theta_C$ and $\theta_B = (d_B/d_C)\theta_C$. Furthermore, assume the allocation x_{iC} given to campaign C is proportionally split across campaigns A and B; i.e., $x_{iA} = (d_A/d_C) x_{iC}$ and $x_{iB} = (d_B/d_C) x_{iC}$, for all audience segments i. Since $x_{iA} + x_{iB} = x_{iC}$, an advertiser should be indifferent between purchasing both campaigns A and B, or just campaign C. We will now show when the objective function $\sum_j F_j$ exhibits this indifference.

Proof of Part 1 (Baseline Model): We begin with a technical lemma, then present the main proposition, and finally communicate our result using a corollary.

LEMMA EC.1. $\sum_{i\in\Gamma(C)} s_i (x_{iC} - \theta_C)^2 = \left(\frac{d_C}{d_A}\right) \sum_{i\in\Gamma(A)} s_i (x_{iA} - \theta_A)^2 + \left(\frac{d_C}{d_B}\right) \sum_{i\in\Gamma(B)} s_i (x_{iB} - \theta_B)^2$.

Proof. We simplify the left-hand-side expression as follows:

$$\sum_{i\in\Gamma(C)} s_i (x_{iC} - \theta_C)^2$$

$$
\begin{aligned}
&= \left(\frac{d_A}{d_C}\right) \sum_{i\in\Gamma(C)} s_i(x_{iC}-\theta_C)^2 + \left(\frac{d_B}{d_C}\right) \sum_{i\in\Gamma(C)} s_i(x_{iC}-\theta_C)^2, \text{ since } d_A+d_B=d_C \\
&= \left(\frac{d_A}{d_C}\right) \sum_{i\in\Gamma(A)} s_i \left(\left(\frac{d_C}{d_A}\right) x_{iA} - \left(\frac{d_C}{d_A}\right)\theta_A\right)^2 + \left(\frac{d_B}{d_C}\right) \sum_{i\in\Gamma(B)} s_i \left(\left(\frac{d_C}{d_B}\right) x_{iB} - \left(\frac{d_C}{d_B}\right)\theta_B\right)^2 \\
&= \left(\frac{d_C}{d_A}\right) \sum_{i\in\Gamma(A)} s_i \left(x_{iA}-\theta_A\right)^2 + \left(\frac{d_C}{d_B}\right) \sum_{i\in\Gamma(B)} s_i \left(x_{iB}-\theta_B\right)^2,
\end{aligned}
$$

which concludes the proof of this lemma. □

PROPOSITION EC.2. *If $W_A > W_C\left(\frac{d_C}{d_A}\right)$ and $W_B > W_C\left(\frac{d_C}{d_B}\right)$, then $F_C < F_A + F_B$; thus, an advertiser prefers to have both campaigns A and B over having only campaign C. Similarly, if $W_A < W_C\left(\frac{d_C}{d_A}\right)$ and $W_B < W_C\left(\frac{d_C}{d_B}\right)$, then $F_C > F_A + F_B$; thus, an advertiser prefers to have only campaign C over having both A and B. Moreover, if $W_A = W_C\left(\frac{d_C}{d_A}\right)$ and $W_B = W_C\left(\frac{d_C}{d_B}\right)$, then $F_C = F_A + F_B$; in this case, an advertiser is indifferent to having only C versus having both A and B.*

Proof. We prove only the first case, as all three cases are similar. Assume $W_A > W_C\left(\frac{d_C}{d_A}\right)$ and $W_B > W_C\left(\frac{d_C}{d_B}\right)$. Then:

$$
\begin{aligned}
F_C &= W_C \sum_{i\in\Gamma(C)} s_i(x_{iC}-\theta_C)^2 \\
&= W_C\left(\frac{d_C}{d_A}\right) \sum_{i\in\Gamma(A)} s_i \left(x_{iA}-\theta_A\right)^2 + W_C\left(\frac{d_C}{d_B}\right) \sum_{i\in\Gamma(B)} s_i \left(x_{iB}-\theta_B\right)^2, \text{ by Lemma EC.1} \\
&< W_A \sum_{i\in\Gamma(A)} s_i \left(x_{iA}-\theta_A\right)^2 + W_B \sum_{i\in\Gamma(B)} s_i \left(x_{iB}-\theta_B\right)^2 = F_A + F_B,
\end{aligned}
$$

which concludes the proof of this proposition. □

COROLLARY EC.1. *If $W_j = Q_j/d_j$ where the given Q_j values do not change when campaigns are split (i.e., $Q_C = Q_A = Q_B$), advertisers are indifferent toward buying small or large campaigns.*

Proof. We have $W_A d_A = Q_A = Q_C = W_C d_C$, and similarly $W_B d_B = W_C d_C$. Therefore, $W_A = W_C\left(\frac{d_C}{d_A}\right)$ and $W_B = W_C\left(\frac{d_C}{d_B}\right)$. Invoking Proposition EC.2 concludes the proof of this corollary. □

Finally, we note that as long as the weights V_j in the baseline objective (3a) are chosen independently of campaign demands d_j, then $W_j = V_j/2\theta_j$ satisfies the assumption of Corollary EC.1 with $Q_j = V_j\hat{s}_j/2$. Therefore, so long as V_j's are chosen independently of d_j's, the baseline objective is invariant to arbitrary splits and merges.

Proof of Part 2 (Gini Model): Recall from Section 3.2.2 that the spread term in our Gini objective is $\alpha\sum_j w_j G_j$, which is equivalent to $\sum_j \frac{\alpha\hat{s}_j}{2} GMD_j$. With $W_j = \frac{\alpha\hat{s}_j}{2}$, $f_j = GMD_j$, and $F_j = W_j f_j$, we aim to show that $F_C = F_A + F_B$.

Proof. Since $\hat{s}_A = \hat{s}_B = \hat{s}_C$, it is clear that $W_A = W_B = W_C$. Therefore, it remains to show that $f_C = f_A + f_B$.

$$
\begin{aligned}
f_C = GMD_C &= \frac{2}{\hat{s}_C^2} \sum_{(h,i)\in\Gamma_0(C)} s_h s_i \left| x_{hC} - x_{iC} \right| \\
&= \left(\frac{d_A}{d_C}\right) \frac{2}{\hat{s}_C^2} \sum_{(h,i)\in\Gamma_0(C)} s_h s_i \left| x_{hC} - x_{iC} \right| + \left(\frac{d_B}{d_C}\right) \frac{2}{\hat{s}_C^2} \sum_{(h,i)\in\Gamma_0(C)} s_h s_i \left| x_{hC} - x_{iC} \right| \\
&= \left(\frac{d_A}{d_C}\right) \frac{2}{\hat{s}_A^2} \sum_{(h,i)\in\Gamma_0(A)} s_h s_i \left| \left(\frac{d_C}{d_A}\right)(x_{hA} - x_{iA}) \right| + \left(\frac{d_B}{d_C}\right) \frac{2}{\hat{s}_B^2} \sum_{(h,i)\in\Gamma_0(B)} s_h s_i \left| \left(\frac{d_C}{d_B}\right)(x_{hB} - x_{iB}) \right| \\
&= \frac{2}{\hat{s}_A^2} \sum_{(h,i)\in\Gamma_0(A)} s_h s_i \left| x_{hA} - x_{iA} \right| + \frac{2}{\hat{s}_B^2} \sum_{(h,i)\in\Gamma_0(B)} s_h s_i \left| x_{hB} - x_{iB} \right| \\
&= GMD_A + GMD_B = f_A + f_B,
\end{aligned}
$$

which concludes the proof. □

Moreover, if the spread term $\alpha \sum_{j\in J} w_j G_j$ in our Gini objective is generalized to $\sum_{j\in J} \alpha_j w_j G_j$, then the above proposition holds so long as $\alpha_A = \alpha_B = \alpha_C$, i.e., the campaign-specific priority weighting factors α_j do not change when campaigns are arbitrarily split or merged.

A.7. Proof of Proposition 5, Part 1: Interpretation of Weights in Baseline Objective

Proposition 5, Part 1: Assuming $V_j = 1$ for all campaigns $j \in J$, the baseline objective (3a) weights campaigns by their size (i.e., impressions demanded d_j), and penalizes the average squared percentage deviation from the target allocation ($x_{ij} = \theta_j \; \forall i \in \Gamma(j)$).

Proof. Since $\frac{1}{\theta_j} = \left(\frac{1}{\theta_j}\right)^2 \cdot \theta_j = \left(\frac{1}{\theta_j}\right)^2 \cdot \frac{d_j}{\hat{s}_j}$, we have:

$$
\frac{1}{\theta_j} \sum_{i\in\Gamma(j)} s_i (x_{ij} - \theta_j)^2 = \left(\frac{1}{\theta_j}\right)^2 \cdot \frac{d_j}{\hat{s}_j} \sum_{i\in\Gamma(j)} s_i (x_{ij} - \theta_j)^2 = d_j \cdot \left(\sum_{i\in\Gamma(j)} \frac{s_i}{\hat{s}_j} \left(\frac{x_{ij}}{\theta_j} - 1\right)^2 \right).
$$

Notice that, for campaign j, the expression $\left(\frac{x_{ij}}{\theta_j} - 1\right)^2$ is the squared percentage deviation from the target allocation of audience segment i. We average this over all audience segments $i \in \Gamma(j)$ by appropriately weighting larger segments more than smaller segments (note that $\sum_{i\in\Gamma(j)} \frac{s_i}{\hat{s}_j} = 1$). □

Appendix B: Additional Material: Background, Plots, Descriptions, and Extended Analysis

B.1. Bar Graph Visualization of Impression Spread

Consider an advertiser who wants to visualize how well-spread the impressions she got are over some dimension of interest, for example, geography. Her ad campaign targeted only the three western states and received a total of 10 million impressions, with the breakdown being 7, 1, and 2 million from California, Oregon, and Washington, respectively. Although we can visualize this data using a bar graph, as it could be done by using the results of the baseline model, this is problematic for at least two reasons. First, the geographic dimension does not have a natural order, and thus, there are six different permutations of the bars (see Figure EC.1), all of which are equally valid. Second, a typical bar graph is not normalized for differing population sizes across audience segments. Indeed, in our example California, Oregon, and Washington have populations of 39, 4, and 7 million

people, respectively; this corresponds to 78%, 8% and 14% of the western region's population of 50 million. An evenly-spread ad campaign of 10 million impressions would assign 7.8, 0.8, and 1.4 million impressions, respectively, to California, Oregon, and Washington. Meanwhile, our original allocation of 7, 1, and 2 million impressions is 10.3% below, 25% above, and 42.9% above the equal-proportion solution. In essence, the bar graph is misleading since it incorrectly compares the 7 million impressions in California with the 2 million in Washington, without accounting for the fact that California and Washington's audience sizes are on different scales (indeed – California's allocation is in fact below average, not above average!). A Lorenz curve, on the other hand, visually represents the spread of a categorical distribution in such a way that, unlike the more widely used (and abused) bar graph, does not suffer from (a) multiple possible visual representations due to arbitrary ordering of categories, and (b) poor scaling.

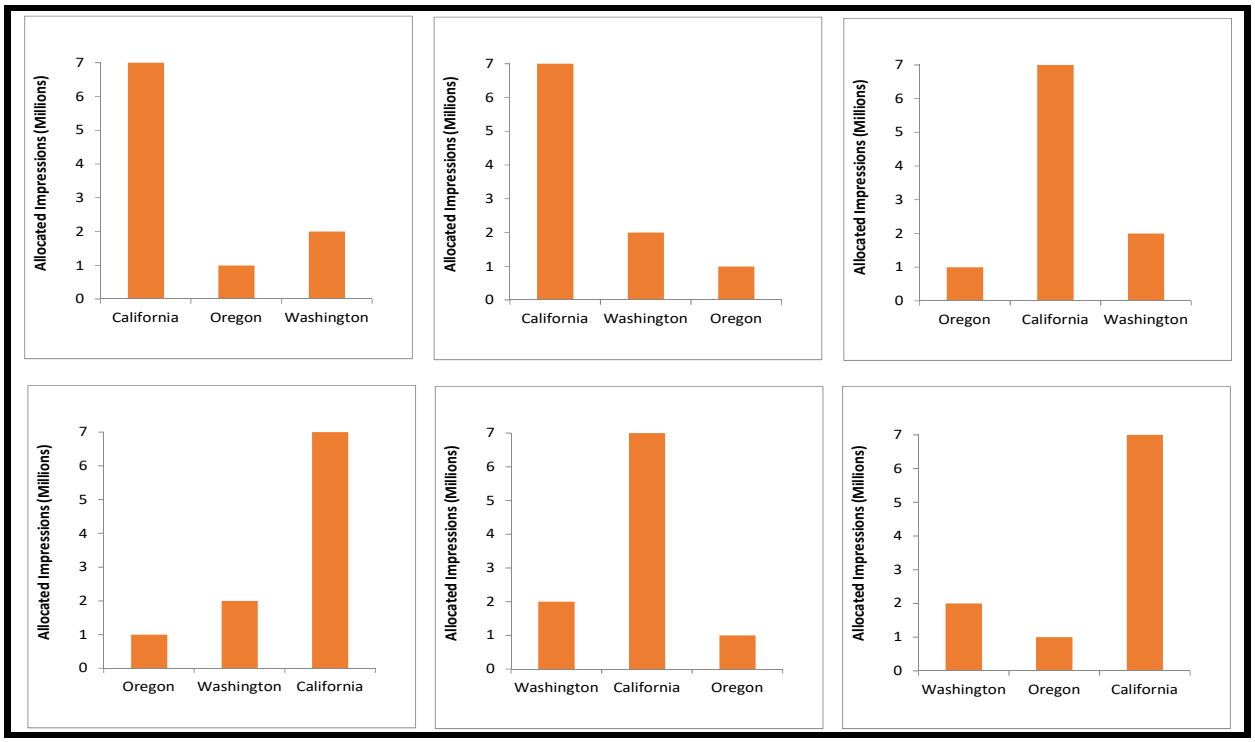

Figure EC.1 Impression spread by geographic region, visually represented using bar graphs.

Note that the practical importance of visualizing spread using a Lorenz curve instead of a bar graph is compounded when comparing spread across multiple ad campaigns. This is because while it may be possible for users to agree on a particular ordering of categories (e.g., listing all geographic regions alphabetically by name from left-to-right), different campaigns target different subsets of categories along a dimension. For example, in a bar graph, the third bar from the left could represent California for one campaign, and Kansas for another. Indeed, a bar graph would only properly represent spread of impressions when (a) all campaigns target all audience segments, (b) users can agree on a single arbitrary ordering of the segments within a bar graph, and (c)

all audience segments represent audiences of the same size. Clearly, this special case is quite rare in practice, which motivates the Lorenz curve as a means of visually representing the spread of impressions across audience segments.

B.2. Alternative Models

In this section, we discuss several alternative spreading objectives, and show how seemingly small structural modifications affect the ability of the corresponding model to satisfy the five key properties listed in Table 1.

Campaign-specific scaling factors, i.e., the w_j's in the Gini objective $\alpha \sum_{j\in J} w_j G_j$ and W_j's in the baseline objective $\sum_{(i,j)\in\Gamma} W_j s_i (x_{ij} - \theta_j)^2$, must be chosen appropriately for several key properties to hold. Consider the following model variants with modified scaling factors, and the resulting effects:

- **Alternative Gini Objective,** $\alpha \sum_{j\in J} d_j G_j$: If we weight the Gini coefficients by the demands d_j instead of by impressions allocated $w_j = d_j - y_j$, the Gini term in the objective would be equivalent to $\alpha \sum_{j\in J}(y_j + w_j) G_j$. Consequently, demand shortfalls y_j would be penalized by the spread metric αG_j, violating Property 3 (Sufficient Orthogonality). Moreover, whereas $\sum_{j\in J} w_j G_j$ is linearizeable via the formulation (7)-(9), the expression $\sum_{j\in J} d_j G_j$ is non-convex in the primary decision variables x_{ij}, causing this model variant to additionally violate Property 1 (Efficient Solvability).

- **Alternative Baseline Objective,** $\sum_{(i,j)\in\Gamma} V_j \theta_j s_i (x_{ij} - \theta_j)^2$: In this model variant, the scaling factor is $W_j = V_j \theta_j$ rather than $W_j = V_j/(2\theta_j)$, and as before we assume V_j is independent of demand d_j. This objective has a nice interpretation, since with $V_j = 1$ it simplifies to $\sum_{(i,j)\in\Gamma} d_j (s_i/\hat{s}_j)(x_{ij} - \theta_j)^2$, which is a weighted average of the squared deviations $(x_{ij} - \theta_j)$, computed using the audience segment sizes s_i as weights, and then scaling by the size of the campaign d_j (recall that by definition, $\sum_{i\in\Gamma(j)} (s_i/\hat{s}_j) = 1$). Unfortunately, this objective violates Property 4 (Split-and-Merge Invariance), which can be verified by tracing the proof of Proposition 4. In this case, merged campaigns receive improved performance, and to avoid this distortion the publisher would need to refrain from merging campaigns together to make the planning problem more efficient to solve.

- **Alternative Baseline Objective,** $\sum_{(i,j)\in\Gamma} V_j/(d_j\theta_j) s_i (x_{ij} - \theta_j)^2$: In this model variant, the scaling factor is $W_j = V_j/(d_j\theta_j)$ rather than $W_j = V_j/(2\theta_j)$, and as before we assume V_j is independent of demand d_j. Again, Property 4 (Split-and-Merge Invariance) is violated. But this time, split campaigns receive improved performance. In Appendix B.3 we show that the incentive that advertisers have to split a large campaign into many smaller ones can be significant. Indeed, splitting one campaign into 10 smaller copies can yield 12% lower spread cost and 69% lower shortfall, while splitting it into 100 smaller copies yields 87% lower spread cost and 96% lower shortfall, all at the expense of other campaigns.

Other spreading objectives could also be considered, but many have similar issues. For example:

- **Alternative Baseline Objective,** $\sum_{(i,j)\in\Gamma} V_j/(2\mu_j) s_i (x_{ij} - \mu_j)^2$: This is the same as our original baseline objective, except we have replaced the nominal target θ_j (a constant) with the mean allocation $\mu_j = (1/\hat{s}_j) \sum_{i\in\Gamma(j)} x_{ij}$ (a dependent variable). Unlike the original, because this objective measures spread as a distance from the mean allocation rather than an a priori-defined target, it satisfies both Properties 2 and 3 (Ideal Allocation when Possible and Sufficient Orthogonality). However, because the associated model is no longer a

quadratic program, Property 1 (Efficient Solvability) is violated. Although the model is technically still convex, the $(1/\mu_j)$ factors lead to some numerical difficulties, especially when one or more μ_j's are small. On the other hand, the $(1/\mu_j)$ factors are required for Property 4 (Split-and-Merge Invariance). Consequently, there does not exist a quadratic spreading objective that simultaneously (i) penalizes deviations from μ_j, (ii) is efficiently solvable, and (iii) is split-and-merge invariant.

Finally, we mention two spread objectives that involve a sum of absolute values like our Gini objective but are structurally more similar to our baseline, in that they minimize absolute rather than squared deviations:

- **Alternative Absolute Deviation Objective,** $\sum_{(i,j)\in\Gamma} V_j s_i |x_{ij} - \theta_j|$: Because this objective, like our baseline, measures spread as a distance from an a priori-defined target, it violates Properties 2 and 3 (Ideal Allocation when Possible and Sufficient Orthogonality). However, the other three properties are satisfied. The corresponding model is representable as a linear program, and therefore efficiently solvable. The scaling factor V_j satisfies split-and-merge invariance, and although not immediately obvious, a rearrangement like that in the proof of Proposition 5 part 1 shows that this objective can be interpreted as weighting each campaign by its size, d_j.
- **Alternative Absolute Deviation Objective,** $\sum_{(i,j)\in\Gamma} V_j s_i |x_{ij} - \mu_j|$: Because this objective measures spread as a distance from the mean allocation, it satisfies Properties 2 and 3 (Ideal Allocation when Possible and Sufficient Orthogonality). Moreover, the other three properties are satisfied as well.

Note that, in the existing literature, there are no single-period ad allocation models that we are aware of which spread impressions using either of the two objectives above. Some multi-period models minimize expressions similar to $|x_{ij} - \theta_j|$, and for this reason we use such a baseline in our multi-period extension in Appendix E. We are not aware of any ad planning models that have previously explored the use of the second objective, which minimizes absolute distances to mean allocations, μ_j. We leave a full analysis of this new model to future work, and focus here on comparing our Gini model to the most commonly-used single-period baseline, (SB), as well as deriving computational results which are specific to deploying the Gini model (i.e., our decomposition method and related structural results). One particular point of departure between the absolute deviation models above and our Gini model that could be important in practice is that the absolute deviation models do not produce unique solutions (it is easy to construct examples with multiple optimal solutions). On the other hand, we conjecture that our Gini model produces unique solutions (as it has in all examples that we tested).

B.3. Quantifying the Importance of Split-and-Merge Invariance

We now introduce an alternative baseline model to estimate the inefficiency associated with a model that is not merge-and-split invariant (recall that both the baseline and Gini models are merge-and-split invariant). The objective function, which induces an incentive for advertisers to split their campaigns to get better performance, is:

$$\min \sum_{(i,j)\in\Gamma} V_j/(d_j\theta_j) s_i (x_{ij} - \theta_j)^2 + \sum_{j\in J} p_j y_j \tag{EC.4}$$

Note that in this model variant, the scaling factor W_j in the spread term $\sum_{(i,j)\in\Gamma} W_j s_i (x_{ij} - \theta_j)^2$ of the baseline objective has been set to $W_j = V_j/(d_j\theta_j)$ rather than $W_j = V_j/(2\theta_j)$. Furthermore, we fix V_j to be equal

to 16887 (the average campaign demand $(1/|J|)\sum_{j' \in J} d_{j'}$ for all campaigns, $j \in J$, of the original non-split instance), which produces solutions that balance both shortfall and spread.

Using this alternative baseline model, we first solve a simulated instance with 20 campaigns and 100 viewer types (this is one of the locally tight instances described in Section 5.1). Then, we split campaign #1 into 10 campaigns that each have 1/10th the demand of the original, and re-solve. Finally, we split campaign #1 into 100 campaigns that each have 1/100th the demand of the original, and re-solve. We then look at the combined contribution from all copies of campaign #1 to the alternative baseline's optimal value, and compare this to the contribution from all other campaigns. The results are in Table EC.3 below.

Optimal Value of the Alternative Baseline Objective			
	Original (No Splitting)	After 10 Splits	After 100 Splits
Contribution of All Copies of Campaign #1	102.3 (21.5%)	44.7 (8.3%)	6.0 (1.1%)
Contribution of All Other Campaigns	372.9 (78.5%)	493.2 (91.7%)	554.4 (98.9%)
Total	475.3 (100%)	537.9 (100%)	560.4 (100%)

Table EC.3 Contribution of campaign #1 to the alternative baseline's optimal value. Numerical values are optimal values of (SB) using the alternative objective function (EC.4).

Recall that the baseline objective minimizes spread cost and shortfall cost. Therefore, when campaign #1 contributes less to the optimal value, it means that it receives better spread and/or lower shortfall. We can also repeat the above analysis, isolating the effect on the spread cost and shortfall cost components of the objective. Doing so, we find that both spreading and shortfall are improved for campaign #1, at the expense of the other campaigns, and the magnitude of this improvement increases as a function of the number of times that campaign #1 is split. Details are in Tables EC.4 and EC.5 below.

Spread Cost Component of the Alternative Baseline Objective			
	Original (No Splitting)	After 10 Splits	After 100 Splits
Contribution of All Copies of Campaign #1	23.2 (8.6%)	20.5 (5.8%)	3.0 (0.8%)
Contribution of All Other Campaigns	247.3 (91.4%)	333.8 (94.2%)	371.0 (99.2%)
Total	270.5 (100%)	354.3 (100%)	374.0 (100%)

Table EC.4 Contribution of campaign #1 to the spread cost component of the alternative baseline's optimal value. Numerical values are spread costs as defined in the first term of the alternative objective function (EC.4).

If advertisers are fully rational, one can imagine that they could split each campaign into tens or hundreds of thousands of tiny campaigns to get even better service from a publisher that uses an objective function that is not invariant to arbitrary splits (i.e., if they wrote code on their end to perform this splitting, and then recombine the

Shortfall Cost Component of the Alternative Baseline Objective			
	Original (No Splitting)	After 10 Splits	After 100 Splits
Contribution of All Copies of Campaign #1	79.1 (38.6%)	24.2 (13.2%)	2.9 (1.6%)
Contribution of All Other Campaigns	125.6 (61.4%)	159.4 (86.8%)	183.4 (98.4%)
Total	204.8 (100%)	183.6 (100%)	186.4 (100%)

Table EC.5 Contribution of campaign #1 to the shortfall cost component of the alternative baseline's optimal value. Numerical values are shortfall costs as defined in the second term of the alternative objective function (EC.4).

results for reporting). This is why we suggest that publishers should use objectives that are invariant to arbitrary splits and merges.

B.4. Detailed Description of Residual Distribution Comparison

In this section, we describe the residuals of the baseline and Gini models in detail, and additionally include plots for the Globally Tight and Loose instances, which were omitted from the main body of the paper to save space.

We begin with an in-depth description of Figure 7, which illustrates, for the family of Locally Tight instances, how the residual distributions of the Gini and baseline models change as a function of revenue ($\%R$). From the top-left of Figure 7, we see that although the Gini model (blue) produces solutions that correspond to all revenue levels $\%R \in [0.8, 1]$, the baseline model (red) only produces solutions in the range $\%R \in [0.868, 1]$. The baseline model's revenue range is more limited than the Gini's because the baseline objective is not sufficiently orthogonal (c.f., Property 3). Indeed, starting from $\nu = \epsilon$, where ϵ is a very small positive quantity, yields minimal shortfall and consequently maximal revenue ($\%R = 1$) for the baseline model. Increasing ν exerts more effort on spreading and so, up to a point, shortfall increases and consequently revenue decreases. However, after some point it is not possible to get better spreading by reducing revenue further, and as ν grows large we converge to a solution with $\%R = 0.868$. On the other hand, the Gini objective is separable in shortfall and spread; thus, it is always possible to achieve better spreading by increasing shortfall. For the Gini model, as we increase α above ϵ we can always reach a point where all residuals shrink down to zero, yielding perfect spreading. This point may be very low-revenue, but it is always attainable, and often such a point exists well before $\%R$ drops to zero (for the loose family of instances, the Gini model attains perfect spreading at approximately $\%R = 0.85$).

Continuing to describe the solution structure in the locally tight case, we observe that in the top-left of Figure 7 the top-most blue line (95th-percentile of Gini) is above the top-most red line (95th percentile of baseline). As well, the bottom-most blue line (5th percentile of Gini) is below the bottom-most red line (5th percentile of baseline). This relationship holds for all revenue levels over which both models provide solutions ($\%R \in [0.868, 1]$). Thus, for all revenue levels, the residual distribution is wider under the Gini model than under the baseline model. That said, although the Gini's distribution has wider tails, it also has more mass in its midsection. Compare the third pair of lines from the bottom (25th percentiles, dotted) and notice that the blue Gini line is above the red baseline, with the Gini line approaching the zero-level faster than the baseline as revenue decreases. The second pair of lines from the top (90th percentiles, dashed) exhibit the same pattern, with the blue Gini line below (i.e., closer to the zero-level) than the baseline. The midsection of the distribution is quite concentrated, as

can be seen more clearly by zooming in on the portion of the plot where $\%R \in [0.95, 1]$ and residuals are in the range $[-0.1, 0.1]$, presented in the bottom-left panel of Figure 7. Notice that the 40th percentile of Gini (blue, dashed) starts (at $\%R = 1$) just above the 40th percentile of baseline (red, dashed), at a value of -0.0669. As revenue decreases, we see that the 40th percentile of Gini converges to the zero-level at $\%R = 0.95$, whereas the 40th percentile of baseline actually drops a little from -0.0729 at $\%R = 1$ to -0.0749 at $\%R = 0.95$. Notably, the median (50th percentile) of Gini (thin solid blue line) is always at the zero-level for all revenue levels, whereas the median of baseline (thin solid red line) is always slightly below the zero-level. The 60th percentile of Gini (dashed blue line) is also horizontal, at the zero-level for all revenue levels, compared to the 60th percentile of baseline (dashed red line) which runs from 0.0110 at $\%R = 1$ to 0.0090 at $\%R = 0.95$.

Another way to compare the residual distributions is by overlaying histograms of their density functions. Continuing to describe the locally tight instance, we observe that the top-right of Figure 7 has three such overlayed histograms, with blue (Gini) histograms overlayed onto red (baseline) histograms, and areas of agreement showing in purple. The first overlayed histogram compares the residual distributions at $\%R = 1$, and corresponds to the cross-section in the top-left plot indicated by the right-most vertical grey bar. From the tallest bar at the center of the histogram, we can see that 24.1% of Gini's residuals are in the range $[-0.035, 0.035]$, compared to 16.4% from the baseline model. The second overlayed histogram compares the residual distributions at $\%R = 0.95$, and corresponds to the cross-section in the top-left plot indicated by the middle vertical grey bar. Here, we can see that 38.2% of Gini's residuals are in the range $[-0.035, 0.035]$, compared to 21.6% from the baseline model. Finally, the third overlayed histogram compares the residual distributions at $\%R = 0.9$, and corresponds to the cross-section in the top-left plot indicated by the left-most grey bar. Here, we can see that 51.9% of Gini's residuals are in the range $[-0.035, 0.035]$, compared to 23.9% from the baseline model.

In summary, the Gini's residual distribution is significantly more concentrated in the middle near zero, and slightly fatter in the tails than the baseline. This difference is the most striking for the locally tight family of instances; however, this structural result is robust and holds for the globally tight and loose families of instances as well (see Figure 8). For completeness, Figure EC.2 plots the residual distributions for the globally tight and loose families of instances.

Finally, in Section 5.5 we investigated the residual distribution under the case of having an outside option that we can also sell impressions to. Figure 9 plotted the residual distribution for the locally tight family of instances. To save space in the main body of the manuscript, we omitted the globally tight and loose cases; these are plotted in Figure EC.3.

B.5. The Supply-Relaxed Subproblem as a Market-Based Model

Small to medium ad aggregators, who do not operate their own websites but instead purchase all of their impression traffic from other publishers and sell it to advertisers, may also use our Gini-based model (SG) to plan and allocate impressions to advertisers who have purchased guaranteed contracts and expect to receive well-spread impressions. In this context, we assume the aggregator has estimates for the market prices $\hat{\beta}_i$ and market

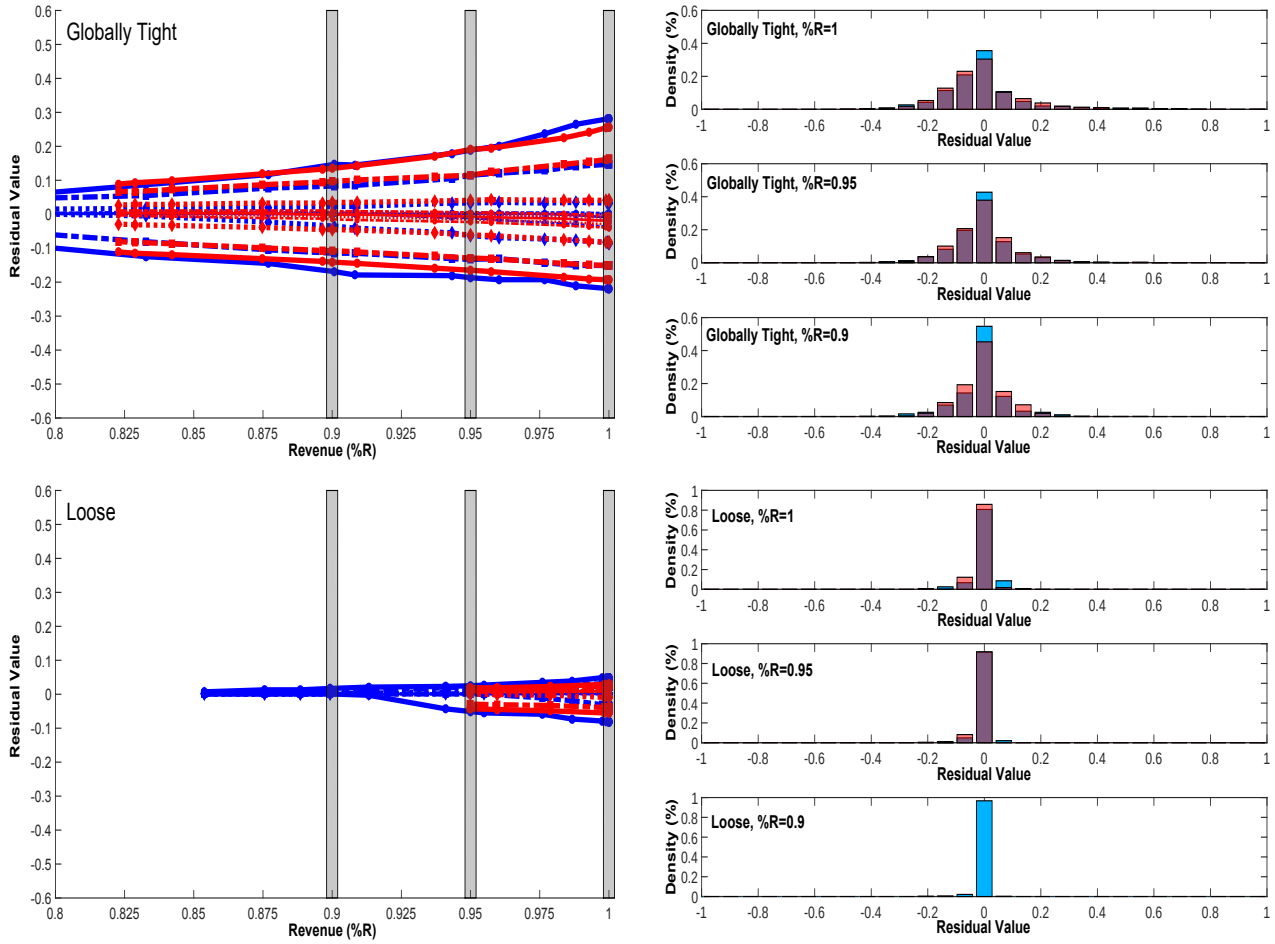

Figure EC.2 Residual Distribution Comparison (Globally Tight and Loose Instances). **Top Row**: Results for the globally tight instances. **Bottom Row**: Results for the loose instances. **Left Column**: Percentiles of the residual distributions from the baseline (red) and Gini (blue) models, as a function of revenue ($\%R$); refer to the legend in Figure 7 of §5.4, which represents each percentile by a different dashed pattern. **Right Column**: Histograms depict residual distributions at $\%R = 1$, 0.95, and 0.9; each histogram corresponds to a cross-section in the left column highlighted by a vertical grey bar. Histograms for the Gini (blue) and baseline (red) distributions are overlayed, with purple bars indicating where the blue and red bars overlap.

supplies s_i of each audience segment. In addition, we assume the volume of impression traffic the aggregator handles is small relative to the size of the market, so that all supply constraints (3c) are nonbinding. The resulting Gini-based model is precisely (SG) with the objective $\min \alpha \sum_{j\in J} \frac{1}{\hat{s}_j} \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \sum_{j\in J} p_j y_j + \sum_{(i,j)\in\Gamma} \hat{\beta}_i s_i x_{ij}$ and the supply constraints dropped.

It is important to note that this market-based model's optimal solution can be obtained by solving (PS) with $\psi_j = \alpha/\hat{s}_j$. Indeed, (PS) is exactly this market-based model with additional constant terms $-\sum_{i\in I} \hat{\beta}_i s_i$ in the objective, and the presence of these constant terms does not affect the optimal solution. Moreover, since (PS) decomposes by campaign, Theorem 1 tells us the optimal allocation for each campaign j in this market-based model. Indeed, since in this case prices $\hat{\beta}_i$ are exogenously given and supply constraints are nonbinding, there is no need to iterate through Dantzig-Wolfe decomposition, and the solution is immediate. Using the optimal

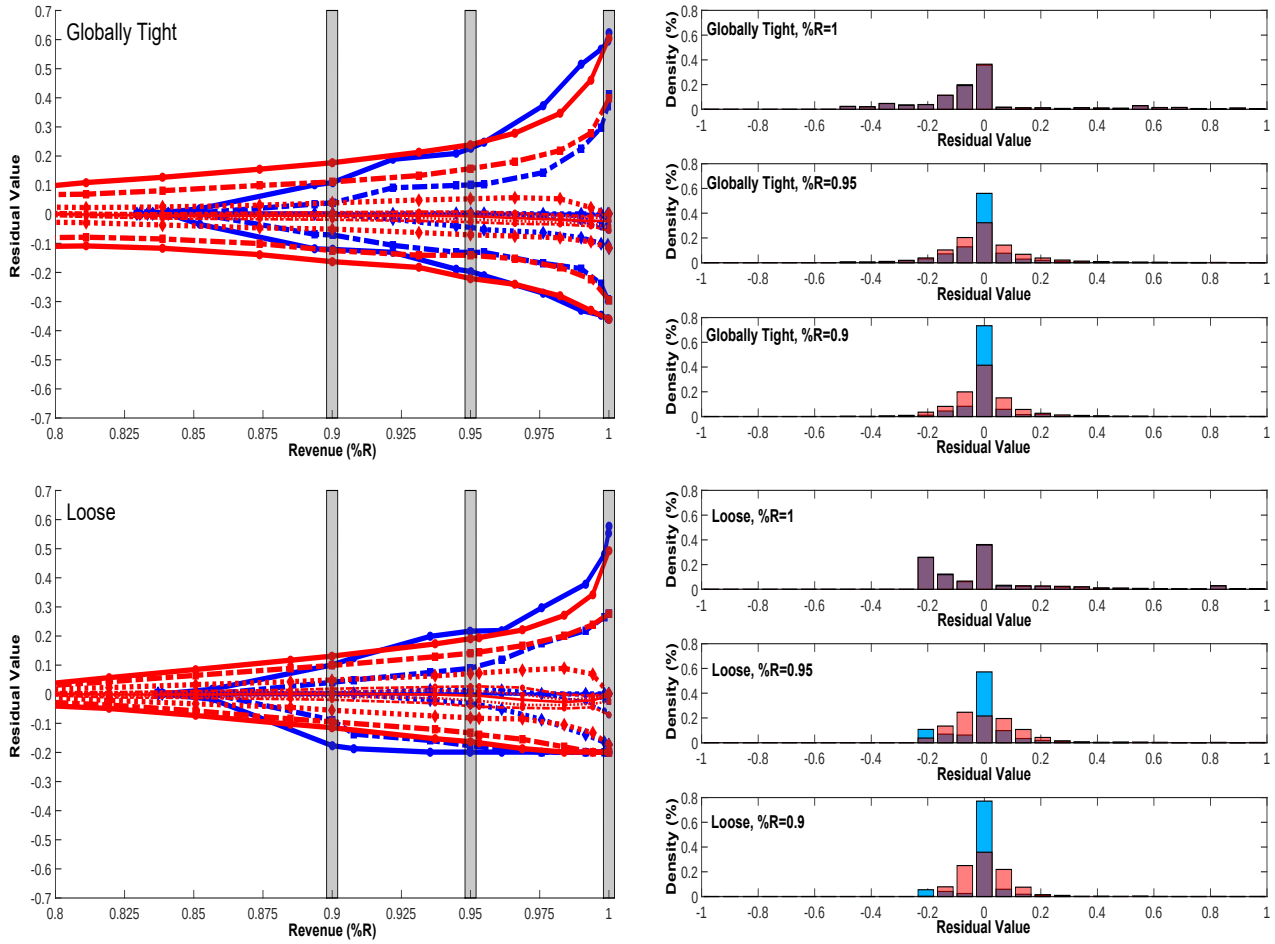

Figure EC.3 Residual Distribution Comparison with Outside Option (Globally Tight and Loose Instances). **Top Row**: Results for the globally tight instances. **Bottom Row**: Results for the loose instances. **Left Column**: Percentiles of the residual distributions from the baseline (red) and Gini (blue) models, as a function of revenue ($\%R$); refer to the legend in Figure 7 of §5.4, which represents each percentile by a different dashed pattern. **Right Column**: Histograms depict residual distributions at $\%R = 1$, 0.95, and 0.9; each histogram corresponds to a cross-section in the left column highlighted by a vertical grey bar. Histograms for the Gini (blue) and baseline (red) distributions are overlayed, with purple bars indicating where the blue and red bars overlap.

solution x^*_{ij}, the aggregator can then establish impression targets $w^*_{ij} = s_i x^*_{ij}$ for each audience segment and campaign pair, which it would attempt to achieve by acquiring impressions at or near predicted market prices.

We believe this model provides a salient first-order characterization of an aggregator's optimal policy, and that suitable data exists to estimate the model's parameters. Although market prices may fluctuate over time, we expect that historical data could be used to estimate the prices that other market participants will charge in expectation for impressions from popular audience segments. As well, estimates of audience sizes can be obtained from third-party audience measurement firms such as ComScore or Nielsen. Note that since there are no supply constraints in this variant of our model, audience size estimates are used exclusively in the objective function to measure the quality of spread.

There are of course many nuanced complications concerning intermediation in online advertising. For more details, see Balseiro et al. (2015b) and Allouah et al. (2017).

Appendix C: Decomposition Method

We now derive in detail our decomposition scheme for our Gini ad allocation problem (SG). We begin by re-stating the general formulation of this single-period problem, which we presented in §6:

$$\begin{aligned}
\text{(P-ORIG)} \qquad z^* = \min \quad & \sum_{j\in J,(h,i)\in\Gamma_0(j)} \psi_j s_h s_i |x_{hj} - x_{ij}| + \sum_j p_j y_j \\
\text{s.t.} \quad & \sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j && \forall j \in J && \text{(demand)} \\
& \sum_{j\in\Gamma(i)} s_i x_{ij} \le s_i && \forall i \in I && \text{(supply)} \\
& x_{ij} \ge 0 \,\forall (i,j) \in \Gamma; \qquad y_j \ge 0 \,\forall j \in J
\end{aligned}$$

First, in Section C.1, we describe how to disaggregate the audience segments of (P-ORIG) into agglomerations of individual impressions. Then, in the subsequent subsection, we use this impression-based formulation to formally derive our decomposition method.

C.1. Translation to Disaggregate Impression-Level Space

Because some of our proofs are clearer when the quantities of interest are impressions rather than buckets of impressions aggregated into audience segments, we first translate (P-ORIG) into an equivalent impression-level optimization problem. Let R be the set of all impressions, R_i be the set of impressions that comprise audience segment i, Λ be the set of (impression, campaign) pairs that define the targeted instance, $\Lambda(r)$ be the set of campaigns that are targeted by impression r, $\Lambda(j)$ be the set of impressions targeted by campaign j, and $\Lambda_0(j) = \{(q,r) \in \Lambda(j)^2 : q < r\}$ index all pairs of impressions targeted by campaign j. Note that by definition, (i) $|R_i| = s_i$, (ii) $|\Lambda(j)| = \hat{s}_j$, and (iii) $\Lambda(r) = \Gamma(i)$ when impression r is in audience segment i. Our impression-level formulation is as follows (Table EC.2 provides a quick reference to all impression-level notation):

$$\begin{aligned}
\text{(P2)} \qquad \min \quad & \sum_{j\in J,(q,r)\in\Lambda_0(j)} \psi_j |x_{qj} - x_{rj}| + \sum_j p_j y_j \\
\text{s.t.} \quad & \sum_{r\in\Lambda(j)} x_{rj} + y_j = d_j && \forall j \in J && \text{(demand)} \\
& \sum_{j\in\Lambda(r)} x_{rj} \le 1 && \forall r \in R && \text{(supply)} \\
& x_{rj} \ge 0 \,\forall (r,j) \in \Lambda; \qquad y_j \ge 0 \,\forall j \in J
\end{aligned}$$

The following proposition formalizes the connection between the audience-level formulation (P-ORIG) and the impression-level formulation (P2).

PROPOSITION EC.3. *Problem (P2) with (i) additional variables $x_{ij}, (i,j) \in \Gamma$, and (ii) additional constraints $x_{rj} = x_{ij}\ \forall r \in R_i, \forall j \in J$, which force all impression allocations within the same audience segment to have the same value, is equivalent to (P-ORIG).*

Proof. We have previously shown (c.f., Appendix A.3) that the objective functions of (P2) and (P-ORIG) are equivalent when $x_{rj} = x_{ij}\ \forall r \in R_i, j \in J$. In a similar manner, the left-hand-side of (P2)'s demand constraint aggregates into an audience-level expression as follows:

$$\sum_{r\in\Lambda(j)} x_{rj} = \sum_{i\in\Gamma(j)} \sum_{r\in R_i} x_{rj} = \sum_{i\in\Gamma(j)} \sum_{r\in R_i} x_{ij} = \sum_{i\in\Gamma(j)} |R_i| x_{ij} = \sum_{i\in\Gamma(j)} s_i x_{ij},$$

and the left-hand-side of (P2)'s supply constraint aggregates into an audience-level expression as follows:

$$\sum_{j\in\Lambda(r)} x_{rj} = \sum_{j\in\Lambda(r)} x_{ij} = \sum_{j\in\Gamma(i)} x_{ij}.$$

Since the objective function and all constraints are logically equivalent, the result follows. □

Technically, (P2) is a relaxation of (P-ORIG), and is equivalent to (P-ORIG) only when we additionally impose the constraints $x_{rj} = x_{ij}\ \forall r \in R_i, j \in J$. However, it turns out that we can impose these constraints in (P2) without loss of optimality. That is, there always exists an optimal solution to (P2) where, for each campaign $j \in J$, the allocation x_{rj} is equal across all impressions $r \in R_i$ within each audience segment $i \in I$. Consequently, (P2) and (P-ORIG) can be treated as essentially equivalent. To prove this result, we first prove the following technical lemma, which shows that if an unequal solution is made more equal, then the value of the Gini term in the objective function of (P2) improves (reduces). The proof requires several steps to be rigorously shown, and formalizes how we may incrementally improve a solution by making it more equal.

LEMMA EC.2. *Assume we are given an impression-level allocation vector $\mathbf{x}_j \equiv \{x_{rj}\}_{r\in\Lambda(j)}$ for some campaign $j \in J$ whose components are not all equal, i.e., there exists a pair of impressions (q, q') such that $x_{qj} < x_{q'j}$. Let $\Delta \in (0, \frac{1}{2}(x_{q'j} - x_{qj})]$ and construct an alternative impression-level allocation vector $\overline{\mathbf{x}}_j \equiv \{\overline{x}_{rj}\}_{r\in\Lambda(j)}$ such that $\overline{x}_{q'j} = x_{q'j} - \Delta$, $\overline{x}_{qj} = x_{qj} + \Delta$, and $\overline{x}_{rj} = x_{rj}\ \forall r \notin \{q, q'\}$. Then $\sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}| < \sum_{(r,r')\in\Lambda_0(j)} |x_{rj} - x_{r'j}|$, i.e., the GMD value of $\overline{\mathbf{x}}_j$ is strictly lower (better) than the GMD value of $\mathbf{x}_j$.*

Proof. Without loss of generality, assume the impressions are ordered by increasing value of x_{rj}; that is, $x_{1j} \le x_{2j} \le \cdots \le x_{\hat{s}_j j}$. Partition the impressions $r \in \Lambda(j) \setminus \{q, q'\}$ into five sets:

$$\begin{aligned}
A &= \{r \in \Lambda(j) : x_{rj} \le x_{qj}, r \ne q\}, \\
B &= \{r \in \Lambda(j) : x_{qj} < x_{rj} \le \overline{x}_{qj}\}, \\
C &= \{r \in \Lambda(j) : \overline{x}_{qj} < x_{rj} \le x_{q'j}\}, \\
D &= \{r \in \Lambda(j) : \overline{x}_{q'j} < x_{rj} \le x_{q'j}, r \ne q'\}, \text{ and} \\
E &= \{r \in \Lambda(j) : x_{q'j} < x_{rj}\}.
\end{aligned}$$

By construction, these sets are mutually exclusive, none of them contain the impressions q or q', and $A \cup B \cup C \cup D \cup E \cup \{q, q'\} = \{1, \ldots, \hat{s}_j\} = \Lambda(j)$. Moreover, impressions are ordered according to $A \prec q \prec B \prec C \prec$

$D \prec q' \prec E$, where $A \prec B$ means that all impressions in A precede all impressions in B with respect to the established ordering. Using these definitions, we can simplify the GMD value of $\overline{\mathbf{x}}_j$ as follows:

$$
\begin{aligned}
\sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}| = & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_{r\in A, r'=q} |\overline{x}_{rj} - \overline{x}_{r'j}| \\
& + \sum_{r=q, r'\in B} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_{r=q, r'\in C\cup D\cup E} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_{r\in A\cup B\cup C, r'=q'} |\overline{x}_{rj} - \overline{x}_{r'j}| \\
& + \sum_{r\in D, r'=q'} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_{r=q', r'\in E} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_{r=q, r'=q'} |\overline{x}_{rj} - \overline{x}_{r'j}| \\
= & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |x_{rj} - x_{r'j}| + \sum_{r\in A} |x_{rj} - \overline{x}_{qj}| \\
& + \sum_{r'\in B} |\overline{x}_{qj} - x_{r'j}| + \sum_{r'\in C\cup D\cup E} |\overline{x}_{qj} - x_{r'j}| + \sum_{r\in A\cup B\cup C} |x_{rj} - \overline{x}_{q'j}| \\
& + \sum_{r\in D} |x_{rj} - \overline{x}_{q'j}| + \sum_{r'\in E} |\overline{x}_{q'j} - x_{r'j}| + |\overline{x}_{qj} - \overline{x}_{q'j}| \\
= & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |x_{rj} - x_{r'j}| + \sum_{r\in A} (\overline{x}_{qj} - x_{rj}) \\
& + \sum_{r'\in B} (\overline{x}_{qj} - x_{r'j}) + \sum_{r'\in C\cup D\cup E} (x_{r'j} - \overline{x}_{qj}) + \sum_{r\in A\cup B\cup C} (\overline{x}_{q'j} - x_{rj}) \\
& + \sum_{r\in D} (x_{rj} - \overline{x}_{q'j}) + \sum_{r'\in E} (x_{r'j} - \overline{x}_{q'j}) + (\overline{x}_{q'j} - \overline{x}_{qj}) \\
= & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |x_{rj} - x_{r'j}| + \sum_{r\in A} (x_{qj} + \Delta - x_{rj}) \\
& + \sum_{r'\in B} (x_{qj} + \Delta - x_{r'j}) + \sum_{r'\in C\cup D\cup E} (x_{r'j} - x_{qj} - \Delta) + \sum_{r\in A\cup B\cup C} (x_{q'j} - \Delta - x_{rj}) \\
& + \sum_{r\in D} (x_{rj} - x_{q'j} + \Delta) + \sum_{r'\in E} (x_{r'j} - x_{q'j} + \Delta) + ((x_{q'j} - \Delta) - (x_{qj} + \Delta)) \\
= & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |x_{rj} - x_{r'j}| + \sum_{r\in A} (x_{qj} - x_{rj}) \\
& + \sum_{r'\in B} (x_{qj} - x_{r'j}) + \sum_{r'\in C\cup D\cup E} (x_{r'j} - x_{qj}) + \sum_{r\in A\cup B\cup C} (x_{q'j} - x_{rj}) \\
& + \sum_{r\in D} (x_{rj} - x_{q'j}) + \sum_{r'\in E} (x_{r'j} - x_{q'j}) + (x_{q'j} - x_{qj}) - 2\Delta(|C| + 1) \\
= & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |x_{rj} - x_{r'j}| + \sum_{r\in A} |x_{qj} - x_{rj}| \\
& - \sum_{r'\in B} |x_{qj} - x_{r'j}| + \sum_{r'\in C\cup D\cup E} |x_{r'j} - x_{qj}| + \sum_{r\in A\cup B\cup C} |x_{q'j} - x_{rj}| \\
& - \sum_{r\in D} |x_{rj} - x_{q'j}| + \sum_{r'\in E} |x_{r'j} - x_{q'j}| + |x_{q'j} - x_{qj}| - 2\Delta(|C| + 1) \\
< & \sum_{(r,r')\in\Lambda_0(j): r\neq q, r'\neq q'} |x_{rj} - x_{r'j}| + \sum_{r\in A} |x_{qj} - x_{rj}| \\
& + \sum_{r'\in B} |x_{qj} - x_{r'j}| + \sum_{r'\in C\cup D\cup E} |x_{r'j} - x_{qj}| + \sum_{r\in A\cup B\cup C} |x_{q'j} - x_{rj}| \\
& + \sum_{r\in D} |x_{rj} - x_{q'j}| + \sum_{r'\in E} |x_{r'j} - x_{q'j}| + |x_{q'j} - x_{qj}|
\end{aligned}
$$

$$= \sum_{(r,r')\in\Lambda_0(j)} |x_{rj} - x_{r'j}|,$$

which is the GMD value of $\mathbf{x}_j$. The first equality expands $\sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}|$ into eight terms, allowing us to treat pairs of impressions (r, r') differently based on their membership in the sets $\{A, B, C, D, E\}$. The second equality follows since $\overline{x}_{rj} = x_{rj}$ for all $r \in A \cup B \cup C \cup D \cup E$. The third equality follows from the definitions of the sets $\{A, B, C, D, E\}$, and the fourth is by substitution. The fifth equality follows by collecting all Δ terms and simplifying. The sixth equality re-establishes absolute values, and follows by the defined ordering $A \prec q \prec B \prec C \prec D \prec q' \prec E$. The inequality in the second-last line must be strict since by definition $\Delta > 0$. Finally, the last equality undoes the earlier expansion and collects all eight terms which are, by definition, equal to $\sum_{(r,r')\in\Lambda_0(j)} |x_{rj} - x_{r'j}|$. □

With this technical lemma in hand, we are now ready to show that the restriction that all impression allocations within the same audience segment are the same, i.e., $x_{rj} = x_{ij}\ \forall r \in R_i, j \in J$, can be applied to (P2) without loss of optimality.

PROPOSITION EC.4. *All optimal solutions* $(\mathbf{x}^*, \mathbf{y}^*)$ *to (P2) satisfy the condition* $x^*_{rj} = x^*_{r'j}\ \forall (r, r') \in R_i^2, \forall (i, j) \in \Gamma$*, i.e., all impressions within the same audience segment have the same allocation.*

Proof. Assume there exists an optimal solution $(\mathbf{x}, \mathbf{y})$ to (P2) that does not satisfy the stated condition. We will construct an alternative solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ which (i) satisfies the stated condition, (ii) is feasible in (P2), and (iii) has a strictly lower (better) objective value, thereby contradicting our original assumption that $(\mathbf{x}, \mathbf{y})$ was optimal. We construct the alternative solution $(\hat{\mathbf{x}}, \hat{\mathbf{y}})$ as follows, which involves one or more invocations of the following improvement step.

(Improvement Step) Select an audience segment i and a campaign j for which there exists a pair of impressions (r^1, r^2) within audience segment i that have unequal allocations, i.e., $x_{r^1 j} < x_{r^2 j}$. Let $\mu_{ij} = (1/|R_i|) \sum_{r\in R_i} x_{rj}$ be the average allocation to campaign j from impressions of audience segment i. Since not all impressions $r \in R_i$ have the same campaign-j allocation x_{rj}, there must exist at least one impression q with a below-average allocation, and another impression q' with an above-average allocation; i.e., $x_{qj} < \mu_{ij} < x_{q'j}$. Let $\Delta = \min\{x_{q'j} - \mu_{ij},\ \mu_{ij} - x_{qj}\}$, and modify the values of the allocation $\mathbf{x}$ such that $x_{q'j} \leftarrow x_{q'j} - \Delta$, and $x_{qj} \leftarrow x_{qj} + \Delta$.

(Termination Criterion) Repeatedly invoke the Improvement Step until $x_{rj} = x_{r'j}\ \forall (r, r') \in R_i^2, \forall (i, j) \in \Gamma$, i.e., all impressions within the same audience segment have the same allocation. Upon termination, denote the resulting allocation as $\hat{\mathbf{x}}$. By construction, this allocation satisfies $\hat{x}_{rj} = \mu_{ij}\ \forall r \in R_i,\ \forall j \in J$.

(Finite Convergence) Termination in a finite number of steps is assured, since (i) at each Improvement Step, at least one impression has its allocation x_{rj} updated to a corresponding mean μ_{ij}, (ii) there are a finite number of impressions and campaigns, and (iii) the means μ_{ij} do not change throughout this procedure.

(Feasibility) We now show that the solution $(\hat{\mathbf{x}}, \mathbf{y})$ is feasible in (P2). First, $\hat{x}_{rj} \geq 0\ \forall (r, j) \in \Lambda$ is assured, since the original solution $(\mathbf{x}, \mathbf{y})$ satisfied $x_{rj} \geq 0\ \forall (r, j) \in \Lambda$, and whenever x_{rj} was decreased, its new level was

never set to below μ_{ij}, a non-negative quantity. Next, the left-hand side of the demand constraint has the same value under both solutions $(\mathbf{x}, \mathbf{y})$ and $(\hat{\mathbf{x}}, \mathbf{y})$, since, for each campaign $j \in J$,

$$\sum_{r \in \Lambda(j)} \hat{x}_{rj} = \sum_{i \in \Gamma(j)} \left(\sum_{r \in R_i} \hat{x}_{rj} \right) = \sum_{i \in \Gamma(j)} \left(\sum_{r \in R_i} x_{rj} \right) = \sum_{r \in \Lambda(j)} x_{rj}.$$

The middle equality holds because $\sum_{r \in R_i} x_{rj}$ remains unchanged after each Improvement Step (although two x_{rj} values are updated, the net change to the sum is $\Delta - \Delta = 0$). Since the demand constraint was satisfied by $(\mathbf{x}, \mathbf{y})$, it is also satisfied by $(\hat{\mathbf{x}}, \mathbf{y})$. Finally, since the supply constraint is satisfied by the solution $(\mathbf{x}, \mathbf{y})$, the following simplifications show it must also be satisfied by $(\hat{\mathbf{x}}, \mathbf{y})$:

$$\sum_{j \in \Lambda(r)} x_{rj} \leq 1 \, \forall r \in R \implies \sum_{j \in \Lambda(r)} x_{rj} \leq 1 \, \forall r \in R_i, \forall i \in I \implies \sum_{r \in R_i} \sum_{j \in \Lambda(r)} x_{rj} \leq |R_i| \, \forall i \in I$$
$$\implies \sum_{j \in \Gamma(i)} (1/|R_i|) \sum_{r \in R_i} x_{rj} \leq 1 \, \forall i \in I \implies \sum_{j \in \Gamma(i)} \mu_{ij} \leq 1 \, \forall i \in I \implies \sum_{j \in \Gamma(i)} \hat{x}_{ij} \leq 1 \, \forall i \in I.$$

(Improved Solution) Lemma EC.2 shows that each Improvement Step strictly lowers (improves) the value of the Gini term in the objective of (P2). Since we perform at least one Improvement Step, the solution $(\hat{\mathbf{x}}, \mathbf{y})$ has a strictly lower (better) value for its Gini term than $(\mathbf{x}, \mathbf{y})$. Moreover, both these solutions share the same $\mathbf{y}$, and thus have the same shortfall cost. Therefore, $(\mathbf{x}, \mathbf{y})$ is clearly suboptimal, which contradicts our initial assumption. This completes our proof. □

C.2. Derivation of Decomposition Method

Having just established that (P2) is the impression-level analog to the audience-segment-based (P-ORIG), we now derive our decomposition method by relaxing (P2) and making a number of observations about the resulting relaxation.

Using β_r as the dual value for impression r's supply constraint, after dualizing the supply constraints, (P2) becomes:

$$\begin{aligned} \text{(P3)} \qquad z^{LB} = \min & \sum_{j \in J, (q,r) \in \Lambda_0(j)} \psi_j |x_{qj} - x_{rj}| + \sum_{j \in J} p_j y_j + \sum_{r \in R} \beta_r \Big(\sum_{j \in \Lambda(r)} x_{rj} - 1 \Big) \\ \text{s.t.} & \sum_{r \in \Lambda(j)} x_{rj} + y_j = d_j \qquad \forall j \in J \qquad \text{(demand)} \\ & x_{rj} \geq 0 \, \forall (r,j) \in \Lambda; \qquad y_j \geq 0 \, \forall j \in J \end{aligned}$$

Since (P3) is a relaxation of (P2), its optimal value is a lower bound for the optimal value of (P-ORIG), i.e., $z^{LB} \leq z^*$. Moreover, (P3) decomposes by campaign and its optimal value can be written as $z^{LB} = -\sum_{r \in R} \beta_r + z_j^{LB}$, where each campaign-$j$ subproblem has optimal value z_j^{LB} and takes the following form:

$$\begin{aligned} \text{(P3-j)} \qquad z_j^{LB} = \min \; & \psi_j \sum_{(q,r) \in \Lambda_0(j)} |x_{qj} - x_{rj}| + p_j y_j + \sum_{r \in \Lambda(j)} \beta_r x_{rj} \\ \text{s.t.} & \sum_{r \in \Lambda(j)} x_{rj} + y_j = d_j \qquad \text{(demand)} \\ & x_{rj} \geq 0 \, \forall r \in \Lambda(j); \qquad y_j \geq 0 \end{aligned}$$

Although it is somewhat difficult to characterize the optimal solution to (P3-j) in general, the following proposition characterizes the optimal solution to (P3-j) for the special case where all dual values β_r are zero.

PROPOSITION EC.5. *If $\beta_r = 0$ for all $r \in \Lambda(j)$, then the optimal solution to (P3-j) is $x_{rj} = \theta_j$ for all $r \in \Lambda(j)$, and $y_j = 0$. That is, when all supply constraints of (P2) are non-binding (i.e., $\beta_r = 0$), the optimal solution to (P3-j) is to spread the campaign's impression allocation proportionally across its targeted supply of impressions, and to have no shortfall.*

Proof. The proposed solution is feasible, since the demand constraint holds, i.e., $\sum_{r\in\Lambda(j)} \theta_j + 0 = |\Lambda(j)|\theta_j = \hat{s}_j(d_j/\hat{s}_j) = d_j$. Moreover, this solution has objective value 0, which must be optimal since solutions of negative value are not possible. □

The remainder of this section derives a number of transformations of (P3-j), the last of which allows us to fully characterize the optimal solution of (P3-j). The following theorem and corollary establish a partial characterization for the optimal solution of (P3-j), and are an important step towards fully characterizing the optimal solution of (P3-j).

THEOREM EC.1. *Without loss of generality, order the impressions of $\Lambda(j)$ by β_r such that $\beta_1 \le \beta_2 \le \cdots \le \beta_{\hat{s}_j}$, where $\hat{s}_j = |\Lambda(j)|$. If the solution (x_j^*, y_j^*) where $x_j^* \equiv \{x_{rj}^*\}_{r=1..\hat{s}_j}$ is optimal for (P3-j), then we must have $x_{1j}^* \ge x_{2j}^* \ge \cdots \ge x_{\hat{s}_j j}^*$.*

Proof. Assume for a contradiction that (x_j^*, y_j^*) with $x_j^* \equiv \{x_{rj}^*\}_{r=1..\hat{s}_j}$ is optimal for (P3-j) yet for some pair of impressions (q, q') we have $\beta_q \le \beta_{q'}$ and $x_{qj}^* < x_{q'j}^*$. Let $\Delta = \frac{1}{2}(x_{q'j}^* - x_{qj}^*)$ and construct the solution $(\overline{x}_j, y_j^*)$, where $\overline{x}_j \equiv \{\overline{x}_{rj}\}_{r=1..\hat{s}_j}$, by taking $\overline{x}_{q'j} = x_{q'j}^* - \Delta$, $\overline{x}_{qj} = x_{qj}^* + \Delta$, and $\overline{x}_{rj} = x_{rj}^*\ \forall r \notin \{q, q'\}$. Note that by construction, $\Delta > 0$ and $\overline{x}_{q'j} = \overline{x}_{qj}$. We will now show that $(\overline{x}_j, y_j^*)$ has a lower value than (x_j^*, y_j^*) when evaluated in the objective of (P3-j), contradicting our assumption that (x_j^*, y_j^*) was optimal. First, it is clear that $\sum_{r\in\Lambda(j)} \beta_r \overline{x}_{rj} \le \sum_{r\in\Lambda(j)} \beta_r x_{rj}^*$, since:

$$\begin{aligned}
\sum_{r\in\Lambda(j)} \beta_r \overline{x}_{rj} &= \sum_{r\in\Lambda(j):r\notin\{q,q'\}} \beta_r \overline{x}_{rj} + \beta_q \overline{x}_{qj} + \beta_{q'} \overline{x}_{q'j} \\
&= \sum_{r\in\Lambda(j):r\notin\{q,q'\}} \beta_r x_{rj}^* + \beta_q (x_{qj}^* + \Delta) + \beta_{q'} (x_{q'j}^* - \Delta) \\
&= \sum_{r\in\Lambda(j)} \beta_r x_{rj}^* + \Delta(\beta_q - \beta_{q'}) \\
&\le \sum_{r\in\Lambda(j)} \beta_r x_{rj}^*.
\end{aligned}$$

Moreover, $\sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}| < \sum_{(r,r')\in\Lambda_0(j)} |x_{rj}^* - x_{r'j}^*|$, by Lemma EC.2. Thus, $\psi_j \sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_j p_j y_j^* + \sum_{r\in\Lambda(j)} \beta_r \overline{x}_{rj} < \psi_j \sum_{(r,r')\in\Lambda_0(j)} |x_{rj}^* - x_{r'j}^*| + \sum_j p_j y_j^* + \sum_{r\in\Lambda(j)} \beta_r x_{rj}^*$, contradicting our assumption that (x_j^*, y_j^*) was optimal. It follows that if the impressions are ordered as $\beta_1 \le \beta_2 \le \cdots \le \beta_{\hat{s}_j}$, then any optimal solution (x_j^*, y_j^*) to (P3-j) must satisfy $x_{1j}^* \ge x_{2j}^* \ge \cdots \ge x_{\hat{s}_j j}^*$. □

COROLLARY EC.2. *Without loss of generality, order the impressions of $\Lambda(j)$ by β_r such that $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{\hat{s}_j}$, where $\hat{s}_j = |\Lambda(j)|$. Given any two impressions (q, q') with equal dual values $\beta_q = \beta_{q'}$, the optimal allocations for these impressions in (P3-j) are also equal, i.e., $x^*_{qj} = x^*_{q'j}$.*

Proof. Without loss of generality, assume $q' < q$ in the impression ordering. By Theorem EC.1, $x^*_{q'j} \geq x^*_{qj}$. Assume for a contradiction that $x^*_{q'j} > x^*_{qj}$, let $\Delta = \frac{1}{2}(x^*_{q'j} - x^*_{qj})$, and construct the solution $(\overline{x}_j, y^*_j)$, where $\overline{x}_j \equiv \{\overline{x}_{rj}\}_{r=1..m}$, by taking $\overline{x}_{q'j} = x^*_{q'j} - \Delta$, $\overline{x}_{qj} = x^*_{qj} + \Delta$, and $\overline{x}_{rj} = x^*_{rj}\ \forall r \notin \{q, q'\}$. First, it is clear that $\sum_{r\in\Lambda(j)} \beta_r \overline{x}_{rj} = \sum_{r\in\Lambda(j)} \beta_r x^*_{rj}$, since:

$$\begin{aligned}
\sum_{r\in\Lambda(j)} \beta_r \overline{x}_{rj} &= \sum_{r\in\Lambda(j):r\notin\{q,q'\}} \beta_r \overline{x}_{rj} + \beta_q \overline{x}_{qj} + \beta_{q'} \overline{x}_{q'j} \\
&= \sum_{r\in\Lambda(j):r\notin\{q,q'\}} \beta_r x^*_{rj} + \beta_q (x^*_{qj} + \Delta) + \beta_{q'} (x^*_{q'j} - \Delta) \\
&= \sum_{r\in\Lambda(j)} \beta_r x^*_{rj} + \Delta(\beta_q - \beta_{q'}) \\
&= \sum_{r\in\Lambda(j)} \beta_r x^*_{rj},
\end{aligned}$$

where the last line follows since $\beta_q = \beta_{q'}$. Moreover, Lemma EC.2 provides $\sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}| < \sum_{(r,r')\in\Lambda_0(j)} |x^*_{rj} - x^*_{r'j}|$. Thus, $\psi_j \sum_{(r,r')\in\Lambda_0(j)} |\overline{x}_{rj} - \overline{x}_{r'j}| + \sum_j p_j y^*_j + \sum_{r\in\Lambda(j)} \beta_r \overline{x}_{rj} < \psi_j \sum_{(r,r')\in\Lambda_0(j)} |x^*_{rj} - x^*_{r'j}| + \sum_j p_j y^*_j + \sum_{r\in\Lambda(j)} \beta_r x^*_{rj}$, implying (x^*_j, y^*_j) is not optimal. This is a contradiction; hence, we conclude that $x^*_{q'j} = x^*_{qj}$. □

Corollary EC.2 is important for two reasons. First, if we restrict ourselves to considering β_r values that are homogenous across all impressions in each viewer type, i.e., $\beta_r = \hat{\beta}_i\ \forall r \in R_i$, then any optimal solution $\{x^*_{rj}\}_{r\in\Lambda(j)}$ to (P3-j) satisfies $x^*_{rj} = x^*_{r'j}\ \forall (r, r') \in R_i^2$ and can be represented in the viewer type space using variables $\{x^*_{ij}\}_{i\in\Gamma(j)}$ such that $x^*_{rj} = x^*_{ij}\ \forall r \in R_i$. Second, we can further exploit this property by aggregating viewer types by dual value $\hat{\beta}_i$, since any viewer types that have the same dual values ($\hat{\beta}_i = \hat{\beta}_{i'}$) must also have equal solutions $x^*_{ij} = x^*_{i'j}$. Finally, we note that we can assume without loss of optimality that β_r values are homogenous across all impressions in each viewer type, since (i) Proposition EC.4 shows that without loss of optimality we can restrict consideration to solutions that satisfy $x_{rj} = x_{r'j}\ \forall (r, r') \in R_i^2$, (ii) under this restriction the objective term $\sum_{r\in R} \beta_r \left(\sum_{j\in\Lambda(r)} x_{rj} - 1\right)$ of (P3-j) simplifies to $\sum_{i\in I} \left(\sum_{r\in R_i} \beta_r\right) \left(\sum_{j\in\Gamma(i)} x_{ij} - 1\right)$, and (iii) in this form, it is clear that the individual values of β_r do not matter, just the sum $\sum_{r\in R_i} \beta_r$, and thus without loss of generality we can assume $\beta_r = \hat{\beta}_i\ \forall r \in R_i$.

Theorem EC.1 is important, as it allows us to simplify (P3-j) in the manner defined by the following proposition.

PROPOSITION EC.6. *Defining $c_{rj} = \psi_j(\hat{s}_j - 2r + 1) + \beta_r$ for all impressions $r \in \Lambda(j)$ ordered according to $\beta_1 \leq \beta_2 \leq \cdots \leq \beta_{\hat{s}_j}$, where $\hat{s}_j = |\Lambda(j)|$, we can write subproblem (P3-j) as follows:*

$$\textit{(P4-j)} \qquad z^{LB}_j = \min \sum_{r=1..\hat{s}_j} c_{rj} x_{rj} + p_j y_j$$

$$\begin{aligned} &\textit{s.t.} \sum_{r=1..\hat{s}_j} x_{rj} + y_j = d_j && \textit{(demand)} \\ &x_{1j} \geq x_{2j} \geq \cdots \geq x_{\hat{s}_j j} \geq 0; \qquad y_j \geq 0 \end{aligned}$$

Proof. Theorem EC.1 allows us to impose the optimality cut $x_{1j} \geq x_{2j} \geq \cdots \geq x_{\hat{s}_j j} \geq 0$. These constraints on the variables then allow us to simplify the objective of (P3-j) by removing the absolute values as follows:

$$\begin{aligned} &\psi_j \sum_{(q,r)\in\Lambda_0(j)} |x_{qj} - x_{rj}| + p_j y_j + \sum_{r\in\Lambda(j)} \beta_r x_{rj} = \psi_j \sum_{\substack{q=1..\hat{s}_j-1 \\ r=q+1..\hat{s}_j}} (x_{qj} - x_{rj}) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left(\sum_{\substack{q=1..\hat{s}_j-1 \\ r=q+1..\hat{s}_j}} x_{qj} - \sum_{\substack{r=2..\hat{s}_j \\ q=1..r-1}} x_{rj} \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left(\sum_{q=1..\hat{s}_j-1} (\hat{s}_j - q) x_{qj} - \sum_{r=2..\hat{s}_j} (r-1) x_{rj} \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left(\hat{s}_j \sum_{q=1..\hat{s}_j-1} x_{qj} - \sum_{q=1..\hat{s}_j-1} q x_{qj} - \sum_{r=2..\hat{s}_j} r x_{rj} + \sum_{r=2..\hat{s}_j} x_{rj} \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left(\hat{s}_j \sum_{r=1..\hat{s}_j-1} x_{rj} - \sum_{r=1..\hat{s}_j-1} r x_{rj} - \sum_{r=2..\hat{s}_j} r x_{rj} + \sum_{r=2..\hat{s}_j} x_{rj} \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left(\hat{s}_j \left(x_{1j} + \sum_{r=2..\hat{s}_j-1} x_{rj} \right) - \left(x_{1j} + \sum_{r=2..\hat{s}_j-1} r x_{rj} \right) \right. \\ &\qquad \left. - \left(\sum_{r=2..\hat{s}_j-1} r x_{rj} + \hat{s}_j x_{\hat{s}_j j} \right) + \left(\sum_{r=2..\hat{s}_j-1} x_{rj} + x_{\hat{s}_j j} \right) \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left((\hat{s}_j - 1) x_{1j} + \sum_{r=2..\hat{s}_j-1} (\hat{s}_j - 2r + 1) x_{rj} + (-\hat{s}_j + 1) x_{\hat{s}_j j} \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} \\ &= \psi_j \left(\sum_{r=1..\hat{s}_j} (\hat{s}_j - 2r + 1) x_{rj} \right) + p_j y_j + \sum_{r=1..\hat{s}_j} \beta_r x_{rj} = \sum_{r=1..\hat{s}_j} c_{rj} x_{rj} + p_j y_j, \end{aligned}$$

as required. □

Next, the following proposition describes how to aggregate (P4-j) to the viewer type space. Further simplifications of (P4-j) will be presented in the viewer type space, allowing us to read off our final result in the viewer type space.

PROPOSITION EC.7. *Without loss of generality, order all viewer types $i \in \Gamma(j)$ according to $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$. Define $c_{ij} = \psi_j s_i(s_i^{aj} - s_i^{bj}) + s_i\hat{\beta}_i$, where $s_i^{bj} = \sum_{i'=1..i-1} s_{i'}$ and $s_i^{aj} = \sum_{i'=i+1..m_j} s_{i'}$ are the number of impressions that are rank ordered before and after viewer type i's impressions, respectively (thus, $\hat{s}_j = s_i^{bj} + s_i + s_i^{aj}$). Taking $\beta_r = \hat{\beta}_i \, \forall r \in R_i$, we represent subproblem (P4-j) in the viewer type space by aggregating all impressions of each viewer type together. The resulting formulation is:*

$$\begin{aligned}
\textit{(P5-j)} \qquad z_j^{LB} = \min & \sum_{i=1..m_j} c_{ij}x_{ij} + p_j y_j \\
\textit{s.t.} & \sum_{i=1..m_j} s_i x_{ij} + y_j = d_j && \textit{(demand)} \\
& x_{1j} \geq x_{2j} \geq \cdots \geq x_{m_j j} \geq 0; \qquad y_j \geq 0
\end{aligned}$$

Proof. By Corollary EC.2, since $\beta_r = \hat{\beta}_i \, \forall r \in R_i$, we must have $x_{rj} = x_{ij} \, \forall r \in R_i$. The left-hand-side of the demand constraint aggregates as follows:

$$\sum_{r \in \Lambda(j)} x_{rj} = \sum_{i \in \Gamma(j)} \sum_{r \in R_i} x_{rj} = \sum_{i \in \Gamma(j)} \sum_{r \in R_i} x_{ij} = \sum_{i \in \Gamma(j)} |R_i| x_{ij} = \sum_{i=1..m_j} s_i x_{ij}.$$

Moreover, for each viewer type i we have one term in the objective which was aggregated from the impression space as follows. Let us assume that viewer type i corresponds to the impressions $\{q, \ldots, q'\}$. Therefore, by construction, $s_i = q' - q + 1$ is the size of viewer type i created by aggregating impressions $\{q, \ldots, q'\}$ together. By definition, we have:

$$\begin{aligned}
\sum_{r=q..q'} c_{rj}x_{rj} &= \sum_{r=q..q'} \left(\psi_j(\hat{s}_j - 2r + 1) + \beta_r\right) x_{rj} \\
&= \left(\psi_j \sum_{r=q..q'} (\hat{s}_j - 2r + 1) + \sum_{r=q..q'} \beta_r\right) x_{ij} \\
&= \left(\psi_j \left(\sum_{r=q..q'} (\hat{s}_j + 1) - 2\sum_{r=q..q'} r\right) + \sum_{r=q..q'} \hat{\beta}_i\right) x_{ij} \\
&= \left(\psi_j \left(s_i(\hat{s}_j + 1) - 2\left(s_i(q+q')/2\right)\right) + s_i\hat{\beta}_i\right) x_{ij} \\
&= \left(\psi_j s_i \left(\hat{s}_j + 1 - q - q'\right) + s_i\hat{\beta}_i\right) x_{ij} \\
&= \left(\psi_j s_i \left((\hat{s}_j - q') - (q-1)\right) + s_i\hat{\beta}_i\right) x_{ij} \\
&= \left(\psi_j s_i \left(s_i^{aj} - s_i^{bj}\right) + s_i\hat{\beta}_i\right) x_{ij} = c_{ij}x_{ij},
\end{aligned}$$

as required. □

Next, we reformulate (P5-j) using the following variable transformation, which will make it easier to read off the optimal solution.

PROPOSITION EC.8. *Given viewer types $i \in \Gamma(j)$ ordered according to $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$, define $\tilde{c}_{ij} = \sum_{i'=1..i} c_{i'j}$ and $\tilde{s}_{ij} = \sum_{i'=1..i} s_{i'}$. Let $\delta_{m_j j} = x_{m_j j}$, and $\delta_{ij} = x_{ij} - x_{i+1,j}$ for all $i = 1..m_j - 1$. We can reformulate (P5-j) using the decision variables $\{\delta_{ij}\}_{i=1..m_j}$ as follows:*

$$\begin{aligned} \textit{(P6-j)} \qquad z_j^{LB} = \min & \sum_{i=1..m_j} \tilde{c}_{ij}\delta_{ij} + p_j y_j \\ \textit{s.t.} & \sum_{i=1..m_j} \tilde{s}_{ij}\delta_{ij} + y_j = d_j \qquad \textit{(demand)} \\ & \delta_{ij} \geq 0 \; \forall i = 1..m_j;\; y_j \geq 0 \end{aligned}$$

Proof. By definition, $x_{ij} = \sum_{i'=i..m_j} \delta_{i'j}$. The left-hand side of the demand constraint from (P5-j) transforms via substitution as follows:

$$\sum_{i=1..m_j} s_i x_{ij} = \sum_{i=1..m_j} s_i \left(\sum_{i'=i..m_j} \delta_{i'j} \right) = \sum_{i'=1..m_j} \delta_{i'j} \left(\sum_{i=1..i'} s_i \right) = \sum_{i'=1..m_j} \tilde{s}_{i'j}\delta_{i'j}.$$

Moreover, the constraints $x_{ij} \geq x_{i+1,j}$ for $i = 1..m_j - 1$ in (P5-j) imply $x_{ij} - x_{i+1,j} \geq 0$, i.e., $\delta_{ij} \geq 0$ for $i = 1..m_j - 1$. And the constraint $x_{m_j j} \geq 0$ implies $\delta_{m_j j} \geq 0$. Finally, the objective of (P5-j) transforms via substitution as follows:

$$\sum_{i=1..m_j} c_{ij} x_{ij} = \sum_{i=1..m_j} c_{ij} \left(\sum_{i'=i..m_j} \delta_{i'j} \right) = \sum_{i'=1..m_j} \delta_{i'j} \left(\sum_{i=1..i'} c_{ij} \right) = \sum_{i'=1..m_j} \tilde{c}_{i'j}\delta_{i'j}. \qquad \square$$

We can now read off the optimal solution of (P6-j) in the "increment space" defined by the δ_{ij} variables.

PROPOSITION EC.9. *Assume viewer types $i \in \Gamma(j)$ are ordered according to the supply duals $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$. Let $\pi_{ij} = \tilde{c}_{ij}/\tilde{s}_{ij}$ $\forall i = 1..m_j$ and define $i^* = \arg\min_{i \in \{1..m_j\}} \pi_{ij}$. Then, the optimal solution and value to (P6-j) take the following form. If $\pi_{i^*j} > p_j$, then $y_j^* = d_j$, $\delta_{ij}^* = 0 \; \forall i = 1..m_j$, with corresponding optimal value $p_j d_j$. Otherwise, the optimal value is $\pi_{i^*j} d_j$ with corresponding optimal solution $y_j^* = 0$, and*

$$\delta_{ij}^* = \begin{cases} d_j/\tilde{s}_{ij} & \textit{for } i = i^* \\ 0 & \textit{for } i \neq i^* \end{cases}$$

Proof. The δ_{ij} variable with the lowest cost per unit of demand satisfied is indexed by i^*. If this cheapest variable is too expensive ($\pi_{i^*j} > p_j$), we prefer not to satisfy this campaign at all, and set shortfall equal to demand, i.e., $y_j^* = d_j$. On the other hand, if using this cheapest variable is no more expensive than incurring shortfall ($\pi_{i^*j} \leq p_j$), we satisfy demand fully with this variable by using the solution $\{\delta_{ij}^*\}_{i=1..m_j}$ as defined. In the former case, the optimal value is $p_j y_j^* = p_j d_j$. In the latter case, the optimal value is $\tilde{c}_{i^*j}\delta_{i^*j}^* = \tilde{c}_{i^*j}(d_j/\tilde{s}_{i^*j}) = (\tilde{c}_{i^*j}/\tilde{s}_{i^*j})d_j = \pi_{i^*j} d_j$. $\square$

Finally, we can translate the solution represented by the δ-variables in the so-called "increment space" back to the original space represented by x-variables. The following theorem summarizes our main structural result.

THEOREM EC.2. *Assume viewer types $i \in \Gamma(j)$ are ordered according to the supply duals $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$. Let $\pi_{ij} = \tilde{c}_{ij}/\tilde{s}_{ij}\ \forall i = 1..m_j$ and define $i^* = \arg\min_{i \in \{1..m_j\}} \pi_{ij}$. Then, the optimal solution to (P5-j) takes the following form. If $\pi_{i^*j} > p_j$, then $y_j^* = d_j$ and $x_{ij}^* = 0\ \forall i = 1..m_j$. Otherwise, $y_j^* = 0$ and*

$$x_{ij}^* = \begin{cases} d_j/\tilde{s}_{i^*j} & \text{for } i \leq i^* \\ 0 & \text{for } i > i^* \end{cases}$$

Proof. From Proposition EC.8, we know that $x_{ij}^* = \sum_{i'=i..m_j} \delta_{i'j}^*$. Combining this fact with the optimal solution $\{\delta_{ij}^*\}_{i=1..m_j}$ from Proposition EC.9 yields the desired result. □

Theorem EC.2 fully characterizes the solution to (P5-j), and equivalently to (P3-j). If shortfall costs are relatively low ($p_j < \pi_{i^*j}$), then it is optimal to allocate no impressions to campaign j, and incur full shortfall ($y_j^* = d_j$). If shortfall costs are relatively high ($p_j \geq \pi_{i^*j}$), then it is optimal to have no shortfall ($y_j^* = 0$) and to spread impressions proportionally across the viewer types $\{1, \ldots, i^*\}$. Recall that since viewer types $i \in \Gamma(j)$ are ordered according to the supply duals $\hat{\beta}_1 \leq \hat{\beta}_2 \leq \cdots \leq \hat{\beta}_{m_j}$, where $m_j = |\Gamma(j)|$, and so it is possible to interpret the solution as one that proportionally spreads impressions across the cheapest set of viewer types $\{1, \ldots, i^*\}$. The viewer type i^* is at the threshold of affordability, given the values for $\{\hat{\beta}_i\}_{i=1..m_j}$.

We can now solve (P3-j) efficiently by sorting the viewer types in order of $\hat{\beta}_i$, and then invoking Theorem EC.2 to compute the optimal solution to (P3-j) analytically. Notice that Theorem 1 of §6 is essentially Theorem EC.2 with a number of parameters which we formally defined in previous propositions now written out in the statement of the theorem. The full decomposition method, which involves a way to choose $\hat{\beta}_i$'s as well as a way to produce a near-feasible solution to (P-ORIG) from one or more optimal solutions of (P3-j) computed from different $\{\hat{\beta}_i\}_{i \in I}$ values, is described in §6.

Appendix D: Online Algorithm

While the solution to our Gini-based model (SG) can be used directly to serve ads in real-time by periodically re-solving (SG) over a rolling horizon as supply forecasts are updated, and for each arriving user of type i serving ad $j \in \Gamma(i)$ with probability x_{ij}, we can also use information generated by our decomposition method to construct an online algorithm that adapts the ad to show based on the types of impressions which actually materialize. This approach is in line with that of Devanur and Hayes (2009), Feldman et al. (2010), Vee et al. (2010), Mehta (2012), and Agrawal et al. (2014), who have developed online algorithms for a variety of different (non-Gini-based) ad allocation problems. In what follows, we provide a sketch of an online algorithm whose structure is in line with our Gini-based model and associated decomposition method. Since a full formal analysis of the behavior of an online algorithm can be quite lengthy, we leave this to future work. Nevertheless, the following description is useful to draw structural parallels between our decomposition method and the online algorithm it suggests.

Recall that our decomposition method (Section 6) produces the following outputs:

- A vector $\lambda_j = \{\lambda_{j0}, \lambda_{j1}, \ldots, \lambda_{jN}\}$ for each campaign j whose elements λ_{jn}, $n = 0..N$, are all non-negative and satisfy $\sum_{n=0..N} \lambda_{jn} = 1$. That is, λ_j is a probability vector.

- A vector $\hat{\beta}^n = \{\hat{\beta}_1^n, \hat{\beta}_2^n, \ldots, \hat{\beta}_m^n\}$ for each iteration $n = 0..N$ whose elements $\hat{\beta}_i^n$, $i \in I \equiv \{1..m\}$, are the dual prices computed in iteration n for each audience segment i. That is, $\hat{\beta}^n$ is a price vector.

We propose the following online algorithm that decides, on-the-fly, what ad to show to each arriving user.

(Initialization) Set $x_{ij}^S := 0$ for all $(i,j) \in \Gamma$. We will use x_{ij}^S to keep track of any surplus allocation from one arrival to the next (the 'S' superscript denotes "surplus").

(Repeat) For each arriving user,

1. Identify the user as a member of a particular audience segment i, and loop through all campaigns $j \in \Gamma(i)$ which target this audience segment, performing the following:

   (a) Randomly draw $n' := n$ with probability λ_{jn}, and let $\hat{\beta} := \hat{\beta}^{n'}$ be the price vector that campaign j faces. That is, campaign j faces the price vector $\hat{\beta}^n$ with probability λ_{jn}.

   (b) Invoke Theorem 1 using price vector $\hat{\beta}$ to compute x_{ij}.

2. At this stage, we have a proposed allocation x_{ij}, $j \in \Gamma(i)$. Augment this allocation with the surplus allocation from previous arrivals of the same type, i.e., compute $x_{ij}^A := x_{ij} + x_{ij}^S$ for all $j \in \Gamma(i)$.

3. Compute $\xi := 1/\max(1, \sum_{j\in\Gamma(i)} x_{ij}^A)$. Then for all $j \in \Gamma(i)$, let $x_{ij}^P := \xi x_{ij}^A$ and $x_{ij}^S := (1-\xi)x_{ij}^A$. Notice that, by construction, $\xi \in [0,1]$ and $\sum_{j\in\Gamma(i)} x_{ij}^P \leq 1$.

4. Show ad $j \in \Gamma(i)$ with probability x_{ij}^P, and show no ad with probability $1 - \sum_{j\in\Gamma(i)} x_{ij}^P$. Any surplus allocation x_{ij}^S is carried forward to the next user arrival of type i.

By design, this online algorithm over-serves excess impressions to a campaign if the supply matching a campaign exceeds the forecasted traffic volume. On the other hand, if we wish to avoid over-serving, then we can keep track of the total number of impressions delivered to each campaign and modify step 1b so that Theorem 1 is invoked to compute x_{ij} when fewer than the demanded d_j impressions have been served to campaign j, and $x_{ij} := 0$ once the campaign's demand d_j is satisfied.

The structure of our online algorithm follows directly from the Dantzig-Wolfe master problem (PM). Indeed, we are reconstructing, on-the-fly, all solutions x_{ij}^n, $n = 0..N$, according to the probabilities λ_{jn}, and randomizing over these solutions as the Dantzig-Wolfe master problem dictates. By construction, if our online algorithm is given a sufficiently large number of user arrivals of type i, then by the law of large numbers the actual proportion of arrivals of type i which see the solution x_{ij}^n converges to λ_{jn}, and thus with probability 1 the resulting allocation is feasible (i.e., satisfies the supply constraint in (PM)). Moreover, if we get exactly s_i arrivals of each type i, then by a similar asymptotic argument the value achieved by our online algorithm as measured by our Gini objective (6) converges to the master problem's optimal value. It follows that if our decomposition method is run to optimality, and its inputs are used by our online algorithm, then our online algorithm is asymptotically optimal when the number of arrivals of each type are equal to their forecasted quantities. More generally, if our decomposition method is terminated early with a measurable optimality gap, i.e., with N smaller than it would be if the decomposition was run to completion, and/or our forecasts for the supplies s_i have some error, then our online algorithm produces a robust yet near-optimal solution.

Appendix E: Multi-Period Model

In this section, we extend our single-period Gini-based ad planning model (SG) to a multi-period (and thus, multi-dimensional) model. We begin in §E.1 with a short literature review of models that spread impressions across time, as well as those that spread across both audience segments and time. We then introduce a multi-period baseline model (§E.2), define relevant Gini-based metrics (§E.3), use these metrics to formulate a multi-period Gini model (§E.4), and illustrate the Lorenz curves produced by our multi-period model (§E.5). We then introduce a novel decomposition method for our multi-period Gini problem that nests the single-period decomposition method of §6 into a Lagrangian Decomposition scheme (§E.6), and conclude by demonstrating tail Gini metrics that may be used to soften the penalties for deviating from impression targets (§E.7).

E.1. Multi-Period Spreading Literature

As we have described in §2.1, advertisers prefer for the impressions that they get to be well-spread across audience segments. But, this is not the only dimension on which advertisers care to have their impressions spread. In practice, advertisers also often want their campaigns to be well-paced, i.e., with impressions delivered evenly across time. Pacing ensures ads reach a wide audience, have a sustained impact, and support complementary ads that are being run in other media (e.g., television). Bhalgat et al. (2012) use several nested packing constraints to develop a tight $(1-1/e)$-competitive online algorithm for pacing impressions over time. Meanwhile, Lee et al. (2013) and Xu et al. (2015) employ control-based heuristics to throttle the rate impressions are purchased from a spot market to satisfy guaranteed campaigns, with Lee et al. (2013) adjusting the value of bids over time and Xu et al. (2015) keeping bids constant but instead throttling the auction participation rate (both have been tested with real data, and the latter describes a real implementation). Araman and Fridgeirsdottir (2015) use a queuing model and fluid analysis to determine asymptotically optimal prices and display frequencies for well-paced ads.

Finally, Bollapragada et al. (2002) and Turner et al. (2011) have studied the problem of spreading impressions across *both* audience segments and time, for television and dynamic in-game advertising, respectively. Methodologically, all of these papers are different than ours, since they do not consider Gini metrics. To the best of our knowledge, ours is the first paper to propose using Gini metrics to spread online ad impressions.

E.2. Multi-Period Baseline Model

We use the model of Turner et al. (2011), developed for Microsoft (then Massive, Inc.) to plan and schedule dynamic in-game advertising, as our baseline multi-period online ad planning model. In this model, each advertiser defines, in its contract with the publisher, a set of goals that should be met, along with weights that capture the relative importance of each goal. The primary goal is always the end-of-campaign impression goal, which, as in (SB) and (SG), states that each campaign j should get a total of d_j impressions by its end date, otherwise the shortfall y_j is penalized linearly in the objective. In addition, there are a number of secondary goals that ensure impressions are also well-spread (i) over targeted audience segments, (ii) over time, and (iii) over targeted audience segments at each point in time. Note that because (iii) is the most difficult to achieve, criteria (i) and (ii) are given more weight, and thus prioritized ahead of (iii). Consequently, (i) and (ii) are not subsumed by (iii).

Formally, spreading goals (i), (ii), and (iii) are modeled in the following general way. First, each goal is expressed as an impression target d. Then, lower and upper bounds (ℓ, u) are defined which bracket the impression target d; i.e., $\ell \leq d \leq u$. Oftentimes, the bounds are defined in a symmetric manner, e.g., $\ell = \alpha d$ and $u = (1/\alpha)d$, for $\alpha \in [0, 1]$; however, that need not be the case. As long as the number of allocated impressions w falls within the bounds, this is perceived as "good enough spreading", and the publisher is not penalized. However, if w is above or below the bounds, then the extent to which the allocation falls outside of the bounds is penalized linearly. Formally, the objective function minimizes a sum of terms of the form py, where y is the extent to which the bounds are violated and is a decision variable. The impression allocation $w = sx$ is equal to the supply of impressions s multiplied by the proportion of impressions allocated x. Constraints of the form $\ell - y \leq w \leq u + y$; $w \geq 0$; $y \geq 0$ link w with y. We use various forms of x and y as decision variables, and only implicitly model $w \equiv sx$. Indexing time periods with $t \in T$, we get a model with the essential structure of Turner et al. (2011), that captures the primary impression goals (modeled by the demand constraints) as well as the secondary spreading goals of types (i), (ii), and (iii). The full model, which uses variables x and y, as well as parameters p, d, s, ℓ, and u with different subscripts as appropriate from the context, is as follows:

$$\text{(MB)} \quad \min \sum_{j \in J} p_j y_j + \sum_{(i,j) \in \Gamma} p_{ij} y_{ij} + \sum_{j \in J, t \in T_j} p_{jt} y_{jt} + \sum_{(i,j) \in \Gamma, t \in T_j} p_{ijt} y_{ijt} \tag{EC.5a}$$

$$\text{s.t.} \sum_{i \in \Gamma(j), t \in T_j} s_{it} x_{ijt} + y_j = d_j \qquad \forall j \in J \qquad \text{(demand)} \tag{EC.5b}$$

$$\sum_{j \in \Gamma(i) \cap J_t} x_{ijt} \leq 1 \qquad \forall i \in I, t \in T \qquad \text{(supply)} \tag{EC.5c}$$

$$\ell_{ij} - y_{ij} \leq \sum_{t \in T_j} s_{it} x_{ijt} \leq u_{ij} + y_{ij} \qquad \forall (i,j) \in \Gamma \qquad \text{(spreading type (i))} \tag{EC.5d}$$

$$\ell_{jt} - y_{jt} \leq \sum_{i \in \Gamma(j)} s_{it} x_{ijt} \leq u_{jt} + y_{jt} \qquad \forall j \in J, t \in T_j \qquad \text{(spreading type (ii))} \tag{EC.5e}$$

$$\ell_{ijt} - y_{ijt} \leq s_{it} x_{ijt} \leq u_{ijt} + y_{ijt} \qquad \forall (i,j) \in \Gamma, t \in T_j \qquad \text{(spreading type (iii))} \tag{EC.5f}$$

$$x_{ijt}, y_j, y_{ij}, y_{jt}, y_{ijt} \geq 0 \qquad \forall (i,j) \in \Gamma, t \in T_j \qquad \text{(non-negativity)} \tag{EC.5g}$$

The impression targets, i.e., sub-demands, for the three spread constraints are defined as (i) $d_{ij} = s_i \theta_j = \frac{s_i}{\hat{s}_j} d_j$, (ii) $d_{jt} = (1/|T_j|) d_j$, and (iii) $d_{ijt} = \frac{s_{it}}{\hat{s}_{jt}} d_j$, respectively, where T_j represents the set of time periods which campaign j spans, J_t represents the campaigns active in period t, and $\hat{s}_{jt} = \sum_{i \in \Gamma(j)} s_{it}$ is the number of impressions eligible for campaign j at time t. Thus, reasonable choices for bounds are $\ell_{ij} = 0.9 d_{ij}$, $u_{ij} = 1.1 d_{ij}$, $\ell_{jt} = 0.9 d_{jt}$, $u_{jt} = 1.1 d_{jt}$, $\ell_{ijt} = 0.9 d_{ijt}$, and $u_{ijt} = 1.1 d_{ijt}$, assuming we allow a deviation from all impression targets of ten percent.

E.3. Multi-Period Gini-Based Metrics

We now define Gini-based metrics for the three spread dimensions. In practice, spreading over audience segments is defined proportionally, whereas spreading over time is defined using absolute impressions; therefore, the Gini metric for spread type (ii) takes a different form than (i) and (iii). In what follows, T_j is the set of time periods

campaign j spans; J_t is the set of campaigns active in period t; s_{it} is the impression supply of audience segment i in time period t; and $\hat{s}_{jt} = \sum_{i \in \Gamma(j)} s_{it}$ is the number of impressions eligible for campaign j at time t. Our primary decision variables are x_{ijt}, the proportion of audience segment i to assign campaign j in time period t.

The Gini-based metric for dimension (i), which spreads impressions proportionally across audience segments, is the same as in the single-period model (c.f., Eq. (5)). Analogously, the Gini-based metric for dimension (iii), which spreads impressions proportionally across audience segments within a single time period, is:

$$G_{jt} = \frac{GMD_{jt}}{2\mu_{jt}} = \frac{\frac{2}{\hat{s}_{jt}^2} \sum\limits_{(h,i) \in \Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}|}{2\left(\frac{1}{\hat{s}_{jt}} \sum_{i \in \Gamma(j)} s_{it} x_{ijt}\right)} = \left(\frac{1}{\hat{s}_{jt}}\right) \frac{\sum\limits_{(h,i) \in \Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}|}{\sum_{i \in \Gamma(j)} s_{it} x_{ijt}}. \tag{EC.6}$$

In contrast, the Gini metric for constraint (ii) that spreads impressions across time follows from (1) and, using $T_0(j) = \{(t,\tau) \in T_j^2 : t < \tau\}$, is defined as:

$$G_j^T = \frac{GMD_j^T}{2\mu_j^T} = \frac{\frac{2}{|T_j|^2} \sum\limits_{(t,\tau) \in T_0(j)} |w_{jt} - w_{j\tau}|}{2\left(\frac{1}{|T_j|} w_j\right)} = \left(\frac{1}{|T_j| w_j}\right) \sum_{(t,\tau) \in T_0(j)} |w_{jt} - w_{j\tau}|, \tag{EC.7}$$

where $w_{jt} = \sum_{i \in \Gamma(j)} s_{it} x_{ijt}$ and $w_j = \sum_{t \in T_j} w_{jt}$ are the number of impressions allocated to campaign j in period t and over the planning horizon, respectively. This Gini metric strives for solutions that have the same absolute impression allocation each time period, e.g., 100,000 impressions each week, regardless of whether impression supply $\{\hat{s}_{jt}, t \in T_j\}$ is uniform across time. In contrast, we could define the alternate Gini metric:

$$G_j^{ALT} = \frac{GMD_j^{ALT}}{2\mu_j^{ALT}} = \frac{\frac{2}{\hat{s}_j^2} \sum\limits_{(t,\tau) \in T_0(j)} \hat{s}_{jt} \hat{s}_{j\tau} |x_{jt} - x_{j\tau}|}{2\left(\frac{1}{\hat{s}_j} w_j\right)} = \frac{1}{\hat{s}_j w_j} \sum_{(t,\tau) \in T_0(j)} \hat{s}_{jt} \hat{s}_{j\tau} |x_{jt} - x_{j\tau}|, \tag{EC.8}$$

where $x_{jt} = w_{jt} / \hat{s}_{jt}$ is the proportion of eligible impressions in time period t allocated to campaign j. If $\hat{s}_{j2}$ is twice that of $\hat{s}_{j1}$, this alternate Gini metric will strive to allocate twice as many impressions to the second period as the first. This alternate metric is akin to the proportional spreading that we have defined for constraints (i) and (iii); however, in practice advertisers care more about the absolute spread over time and so we use (EC.7).

E.4. Multi-Period Gini-Based Model

We now introduce our multi-period Gini-based model. Following the spirit of our single-period Gini model (SG), we propose the following objective function:

$$\min \ \alpha_1 \sum_{j \in J} w_j G_j + \alpha_2 \sum_{j \in J} w_j G_j^T + \alpha_3 \sum_{j \in J, t \in T_j} w_{jt} G_{jt} + \sum_{j \in J} p_j y_j, \tag{EC.9}$$

which, defining the total Gini penalty $G_{jt}^{TOTAL} = \alpha_1 G_j + \alpha_2 G_j^T + \alpha_3 G_{jt}$, may be expressed as:

$$\min \sum_{j \in J, t \in T_j} G_{jt}^{TOTAL} w_{jt} + \sum_{j \in J} p_j y_j.$$

This objective again has the useful interpretation (c.f., Property 3) of assigning a cost to each impression demanded by each campaign j, since $\sum_{t \in T_j} w_{jt} + y_j = d_j$. Unallocated impressions (corresponding to a demand

shortfall) incur the cost p_j, while impressions allocated to time period t incur the cost G_{jt}^{TOTAL}, which can be viewed as a charge for less-than-perfect service (spreading). Putting all the pieces together yields the following Gini-based linear program[9] over the variables x_{ij}, x_{ijt}, x^+_{hij}, x^+_{hijt}, w_{jt}, $w^+_{jt\tau}$, and y_j:

$$\text{(MG)} \quad \min \ \alpha_1 \sum_{j\in J} \frac{1}{\hat{s}_j} \sum_{(h,i)\in\Gamma_0(j)} s_h s_i x^+_{hij} + \alpha_2 \sum_{j\in J} \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} w^+_{jt\tau}$$
$$+ \alpha_3 \sum_{j\in J, t\in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} x^+_{hijt} + \sum_{j\in J} p_j y_j \qquad \text{(EC.10a)}$$

s.t. (8a), (8b) (linking)

$$\sum_{i\in\Gamma(j), t\in T_j} s_{it} x_{ijt} + y_j = d_j \quad \forall j \in J \qquad \text{(demand)} \qquad \text{(EC.10b)}$$

$$\sum_{j\in\Gamma(i)\cap J_t} x_{ijt} \le 1 \quad \forall i \in I, t \in T \qquad \text{(supply)} \qquad \text{(EC.10c)}$$

$$s_i x_{ij} = \sum_{t\in T_j} s_{it} x_{ijt} \quad \forall (i,j) \in \Gamma \qquad \text{(linking)} \qquad \text{(EC.10d)}$$

$$w_{jt} = \sum_{i\in\Gamma(j)} s_{it} x_{ijt} \quad \forall j \in J, \forall t \in T_j \qquad \text{(linking)} \qquad \text{(EC.10e)}$$

$$x^+_{hijt} \ge x_{hjt} - x_{ijt} \quad \forall j \in J, \forall t \in T_j, \forall (h,i) \in \Gamma_0(j) \qquad \text{(linking)} \qquad \text{(EC.10f)}$$

$$x^+_{hijt} \ge x_{ijt} - x_{hjt} \quad \forall j \in J, \forall t \in T_j, \forall (h,i) \in \Gamma_0(j) \qquad \text{(linking)} \qquad \text{(EC.10g)}$$

$$w^+_{jt\tau} \ge w_{jt} - w_{j\tau} \quad \forall j \in J, \forall (t,\tau) \in T_0(j) \qquad \text{(linking)} \qquad \text{(EC.10h)}$$

$$w^+_{jt\tau} \ge w_{j\tau} - w_{jt} \quad \forall j \in J, \forall (t,\tau) \in T_0(j) \qquad \text{(linking)} \qquad \text{(EC.10i)}$$

$$x_{ijt}, y_j \ge 0 \quad \forall j \in J, \forall t \in T_j, \forall (h,i) \in \Gamma_0(j) \qquad \text{(non-negativity)} \qquad \text{(EC.10j)}$$

Constraint (EC.10d) links the proportions of supply allocated in each period with the proportions of supply used across all periods, and makes use of the fact that $s_i = \sum_{t\in T_j} s_{it}$. Constraint (EC.10e) links the auxiliary w_{jt} variables to x_{ijt} variables; we note that it is possible to eliminate this constraint as well as the w_{jt} variables by substituting it into (EC.10h) and (EC.10i).

We also point out that (MB) is somewhat less restrictive than (MG), since for each impression target in (MB) there was always a range around the target over which no penalty was incurred. If one would prefer to mimic a similar logic in a Gini-based model, one can generalize the Gini-based metrics defined here to their tail-Gini equivalents, as we illustrate in Appendix E.7.

E.5. Lorenz Curves of Multi-Period Model

Figure EC.4 plots three campaigns (one per row) as well as six Lorenz curves per campaign (one per column), to illustrate the output from our multi-period model (MG). Each Lorenz curve corresponds to a single Gini metric. The first column corresponds to G_j, the Gini metric which measures spreading of impressions across audience segments for the entire time horizon; columns 2-5 correspond to G_{jt} for the time periods $t = 1..4$, which measure the spreading of impressions across audience segments for each time period $t = 1..4$; and the last

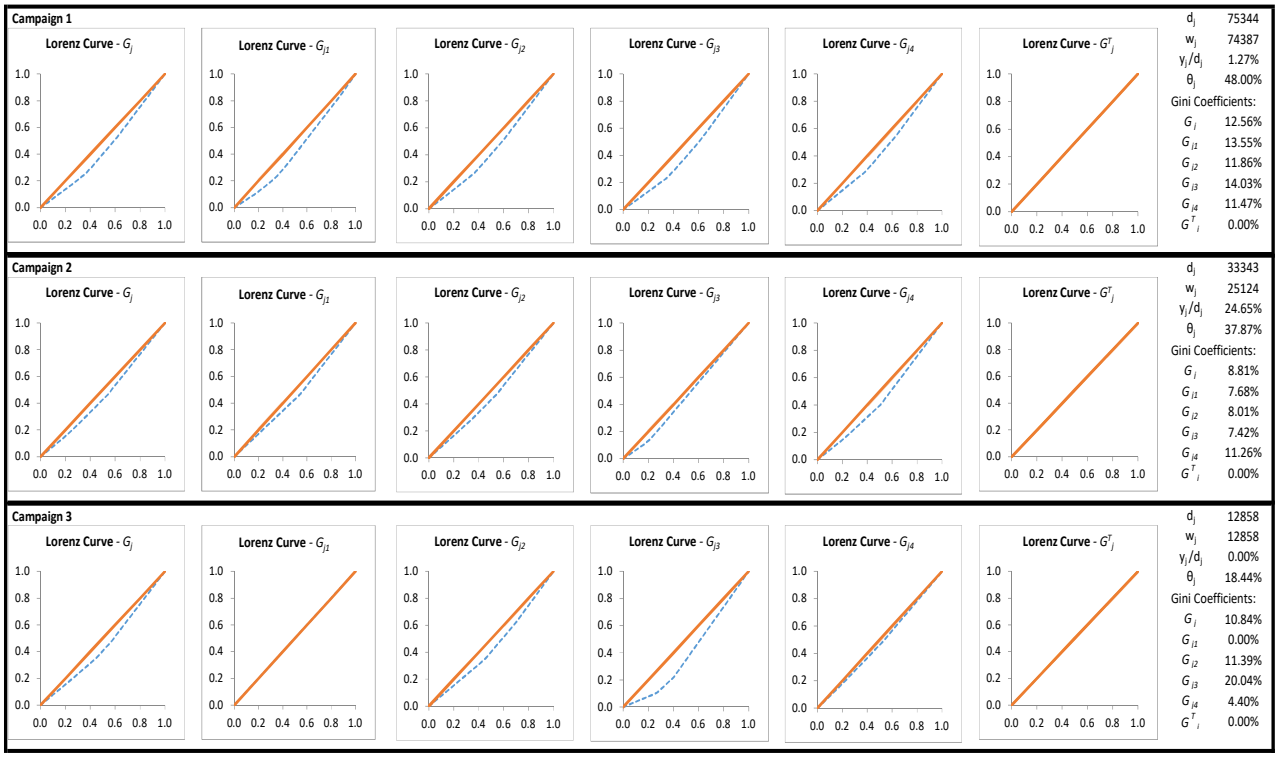

Figure EC.4 Lorenz curves for three campaigns and six Gini metrics.

column corresponds to G_j^T, which measures the spreading of impressions across time. Each advertiser would be able to see only its own Lorenz curves, i.e., a single row of Figure EC.4.

Each panel of our figure illustrates a single Lorenz curve, corresponding to a single Gini metric for a single campaign. The Lorenz curve itself is the dotted blue line, while the orange $45°$ line represents what the Lorenz curve would look like if impressions were perfectly spread. As we can see, in this instance, we are able to achieve perfect spreading across time (see column 6), as well as perfect spreading across audience segments for campaign 3 in week 1. The worst spreading occurs for campaign 3 in week 3, which is reflected by the highest G_{jt} value. The area between the Lorenz curve and its nominal $45°$ line is, by definition, the Gini metric corresponding with that Lorenz curve. Therefore, by minimizing our Gini metrics, we are equivalently minimizing the areas between each Lorenz curve and its nominal $45°$ line.

E.6. Multi-Period Decomposition Method

Our multi-period ad allocation problem (MG) can be decomposed and solved using repeated invocation of Theorem 1. Here we provide one particular way to achieve a decomposition in this general vein. Computational performance of this particular multi-period decomposition scheme compares favorably with using CPLEX directly. For example, an instance with 4 time periods, 100 campaigns, and 100 viewer types is solved to within 1% of optimality in 12093 seconds using our decomposition method, versus 22483 seconds using CPLEX directly.

Structurally, our proposed multi-period decomposition scheme uses Lagrangian Decomposition to split the problem into a number of single-period-type problems with different definitions for $\hat{\beta}$ and ψ_j, for which we

use Theorem 1 to analytically solve all resulting subproblems. To summarize, we break each campaign-j subproblem into two dimensions, a viewer-type dimension and a time dimension. Then, we further decompose the time-dimension subproblem into a series of "viewer-type and time" subproblems. The viewer-type subproblem handles spreading of type (i), the time subproblem handles spreading of type (ii), and the "viewer-type and time" subproblem handles spreading of type (iii). To coordinate the solutions across these subproblems, we use a Dantzig-Wolfe master problem, which now takes convex combinations over both the time-dimension solutions produced at each iteration, and the viewer-type-dimension solutions produced at each iteration. There are additional equality constraints in the Dantzig-Wolfe master problem that ensure that the solutions produced in the time dimension are consistent with those produced in the viewer type dimension; these require additional dual variables and form the basis of the so-called Lagrangian Decomposition. A key structural result is that the "viewer-type and time" subproblem is nested within the time-dimension subproblem; consequently, the solution method uses back-substitution, which turns out to be quite elegant and allows the optimal value of the lower-level "viewer-type and time" subproblem to be analytically embedded into the objective function of the higher-level time-dimension subproblem.

We now describe the details of our decomposition method for our multi-period ad allocation problem (MG). We begin by casting our multi-period ad planning problem (MG) as the following equivalent formulation:

$$
\begin{aligned}
\text{(P-MULT-ORIG)} \qquad \min\ & \alpha_1 \sum_{j\in J} \frac{1}{\hat{s}_j} \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \alpha_2 \sum_{j\in J} \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} |w_{jt} - w_{j\tau}| \\
& + \alpha_3 \sum_{j\in J, t\in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}| + \sum_{j\in J} p_j y_j \\
\text{s.t.}\ & \sum_{j\in\Gamma(i)\cap J_t} s_{it} x_{ijt} \le s_{it} && \forall i\in I,\ \forall t\in T && \text{(supply)} \\
& \sum_{i\in\Gamma(j), t\in T_j} s_{it} x_{ijt} + y_j = d_j && \forall j\in J && \text{(demand)} \\
& w_{jt} = \sum_{i\in\Gamma(j)} s_{it} x_{ijt} && \forall j\in J,\ \forall t\in T_j && \text{(linking (i))} \\
& s_i x_{ij} = \sum_{t\in T_j} s_{it} x_{ijt} && \forall (i,j)\in\Gamma && \text{(linking (ii))} \\
& x_{ij} \ge 0\ \forall (i,j)\in\Gamma; \qquad x_{ijt} \ge 0\ \forall j\in J, \forall i\in\Gamma(j), \forall t\in T_j \\
& w_{jt} \ge 0\ \forall j\in J,\ \forall t\in T_j; \qquad y_j \ge 0\ \forall j\in J
\end{aligned}
$$

Noticing that there are two dimensions to the planning problem, the viewer-type dimension and the time dimension, we exploit this structure by not only designing a decomposition scheme that decomposes the problem by campaign, but also producing a scheme that further decomposes each campaign's subproblem into a viewer-type subproblem and a time subproblem. Additionally, since we want to leverage the results provided by Theorem 1, we need both the viewer-type subproblem for campaign j and the time subproblem for campaign j to each have its own set of decision variables and its own demand constraint. Therefore, we introduce the decision variables y_j^T, $j\in J$, which we initially fix to be equal to their respective y_j, $j\in J$, counterparts. The y_j variables

will measure demand shortfalls in the viewer-type subproblems, while y_j^T will measure demand shortfalls in the time subproblems (the T superscript denotes the time dimension). While it is the case that $y_j = y_j^T$ for all campaigns $j \in J$ must hold for any feasible plan, it could be the case that at any iteration, the subproblems produce solutions that have $y_j \neq y_j^T$. This can and often does happen because the viewer-type subproblems are solved independently from the time subproblems. We will describe later how we use Lagrangian Decomposition to produce solutions that get close to satisfying $y_j = y_j^T \; \forall j \in J$; for now, it is sufficient to understand that while we would like y_j and y_j^T to take on the same values, we model these variables separately so that we can decouple the campaign-j subproblems into viewer-type and time dimensions. Thus, we produce two logically equivalent sets of demand constraints, one using the y_j variables which will end up in the viewer-type subproblems, and the other using the y_j^T variables which will end up in the time subproblems. With these additions to our formulation, (P-MULT-ORIG) can be equivalently represented as:

$$
\begin{aligned}
\text{(P-MULT-ORIG2)} \qquad \min \; & \alpha_1 \sum_{j \in J} \frac{1}{\hat{s}_j} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \alpha_2 \sum_{j \in J} \frac{1}{|T_j|} \sum_{(t,\tau) \in T_0(j)} |w_{jt} - w_{j\tau}| \\
& + \alpha_3 \sum_{j \in J, t \in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i) \in \Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}| + \frac{1}{2} \sum_{j \in J} p_j (y_j + y_j^T) \\
\text{s.t.} \; & \sum_{j \in \Gamma(i) \cap J_t} s_{it} x_{ijt} \leq s_{it} \qquad \forall i \in I, \; \forall t \in T \qquad \text{(supply)} \\
& \sum_{t \in T_j} w_{jt} + y_j^T = d_j \qquad \forall j \in J \qquad \text{(demand (i))} \\
& \sum_{i \in \Gamma(j)} s_i x_{ij} + y_j = d_j \qquad \forall j \in J \qquad \text{(demand (ii))} \\
& w_{jt} = \sum_{i \in \Gamma(j)} s_{it} x_{ijt} \qquad \forall j, \; \forall t \in T_j \qquad \text{(linking (i))} \\
& s_i x_{ij} = \sum_{t \in T_j} s_{it} x_{ijt} \qquad \forall (i,j) \in \Gamma \qquad \text{(linking (ii))} \\
& y_j = y_j^T \qquad \forall j \in J \qquad \text{(linking (iii))} \\
& x_{ij} \geq 0 \; \forall (i,j) \in \Gamma; \qquad x_{ijt} \geq 0 \; \forall j \in J, \forall i \in \Gamma(j), \forall t \in T_j \\
& w_{jt} \geq 0 \; \forall j \in J, \; \forall t \in T_j; \qquad y_j \geq 0 \; \forall j \in J; \qquad y_j^T \geq 0 \; \forall j \in J
\end{aligned}
$$

We now produce a relaxation of (P-MULT-ORIG2) which dualizes the supply constraint, as well as linking constraints (ii) and (iii). We use dual variables $\hat{\beta}_{it}$ for the supply constraints, as well as dual variables η_{ij} and γ_j for linking constraints (ii) and (iii), respectively. Note that γ_j and η_{ij} can take negative or positive values, whereas $\hat{\beta}_{it} \geq 0$. The relaxed problem is:

$$
\begin{aligned}
\text{(PS-MULT)} \quad z^{LB} = \min \; & \alpha_1 \sum_{j \in J} \frac{1}{\hat{s}_j} \sum_{(h,i) \in \Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \alpha_2 \sum_{j \in J} \frac{1}{|T_j|} \sum_{(t,\tau) \in T_0(j)} |w_{jt} - w_{j\tau}| \\
& + \alpha_3 \sum_{j \in J, t \in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i) \in \Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}| + \frac{1}{2} \sum_{j \in J} p_j (y_j + y_j^T) + \sum_{j \in J} \gamma_j (y_j - y_j^T)
\end{aligned}
$$

$$
\begin{aligned}
&+ \sum_{(i,j)\in\Gamma} \eta_{ij} \left(s_i x_{ij} - \sum_{t\in T_j} s_{it} x_{ijt} \right) + \sum_{i\in I, t\in T} \hat{\beta}_{it} \left(\sum_{j\in\Gamma(i)\cap J_t} s_{it} x_{ijt} - s_{it} \right) \\
\text{s.t.} &\sum_{t\in T_j} w_{jt} + y_j^T = d_j \qquad \forall j \in J \qquad \text{(demand (i))} \\
&\sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j \quad \forall j \in J \qquad \text{(demand (ii))} \\
&w_{jt} = \sum_{i\in\Gamma(j)} s_{it} x_{ijt} \qquad \forall j,\, \forall t \in T_j \quad \text{(linking (i))} \\
&x_{ij} \geq 0 \,\forall (i,j) \in \Gamma; \qquad x_{ijt} \geq 0 \,\forall j \in J, \forall i \in \Gamma(j), \forall t \in T_j \\
&w_{jt} \geq 0 \,\forall j \in J,\, \forall t \in T_j; \qquad y_j \geq 0 \,\forall j \in J; \qquad y_j^T \geq 0 \,\forall j \in J
\end{aligned}
$$

We can also simplify the objective function further, to get:

$$
\begin{aligned}
\min\ & \alpha_1 \sum_{j\in J} \frac{1}{\hat{s}_j} \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \alpha_2 \sum_{j\in J} \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} |w_{jt} - w_{j\tau}| \\
&+ \alpha_3 \sum_{j\in J, t\in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}| + \sum_{j\in J} \left(\frac{1}{2} p_j + \gamma_j \right) y_j \\
&+ \sum_{j\in J} (\frac{1}{2} p_j - \gamma_j) y_j^T + \sum_{j\in J, i\in\Gamma(j)} \eta_{ij} s_i x_{ij} + \sum_{j\in J, i\in\Gamma(j), t\in T_j} (\hat{\beta}_{it} - \eta_{ij}) s_{it} x_{ijt} - \sum_{i\in I, t\in T} \hat{\beta}_{it} s_{it}
\end{aligned}
$$

With supply, linking (ii), and linking (iii) constraints dualized, we can see that the relaxed problem (PS-MULT) decomposes into one viewer-type subproblem and one time subproblem for each campaign. The viewer-type subproblem for campaign j has decision variables $\{x_{ij}\}_{(i,j)\in\Gamma}$ and y_j, has optimal value $z_j^{VLB}(\hat{\beta}, \eta, \gamma)$, and is defined as:

$$
\begin{aligned}
\text{(PS-V-j)} \qquad z_j^{VLB}(\hat{\beta}, \eta, \gamma) = \min\ & \alpha_1 \frac{1}{\hat{s}_j} \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hj} - x_{ij}| + \left(\frac{1}{2} p_j + \gamma_j \right) y_j + \sum_{i\in\Gamma(j)} \eta_{ij} s_i x_{ij} \\
\text{s.t.} &\sum_{i\in\Gamma(j)} s_i x_{ij} + y_j = d_j \qquad \text{(demand (ii))} \\
&x_{ij} \geq 0 \,\forall i \in \Gamma(j); \qquad y_j \geq 0
\end{aligned}
$$

Similarly, the time subproblem for campaign j has decision variables $\{x_{ijt}\}_{i\in\Gamma(j),\forall t\in T_j}$, $\{w_{jt}\}_{t\in T_j}$, and y_j^T, has optimal value $z_j^{TLB}(\hat{\beta}, \eta, \gamma)$, and is defined as:

$$
\begin{aligned}
\text{(PS-T-j)} \qquad z_j^{TLB}(\hat{\beta}, \eta, \gamma) = \min\ & \alpha_2 \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} |w_{jt} - w_{j\tau}| + \alpha_3 \sum_{t\in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}| \\
&+ \left(\frac{1}{2} p_j - \gamma_j \right) y_j^T + \sum_{i\in\Gamma(j), t\in T_j} (\hat{\beta}_{it} - \eta_{ij}) s_{it} x_{ijt} \\
\text{s.t.} &\sum_{t\in T_j} w_{jt} + y_j^T = d_j \qquad \text{(demand (i))} \\
&w_{jt} = \sum_{i\in\Gamma(j)} s_{it} x_{ijt} \quad \forall t \in T_j \quad \text{(linking (i))} \\
&x_{ijt} \geq 0 \,\forall i \in \Gamma(j),\, \forall t \in T_j; \qquad w_{jt} \geq 0 \,\forall t \in T_j; \qquad y_j^T \geq 0
\end{aligned}
$$

The relaxed problem (PS-MULT) is related to the set of $|J|$ subproblems (PS-V-j) and (PS-T-j) via the following formula, which expresses the optimal value of (PS-MULT) in terms of the optimal values of (PS-V-j) and (PS-T-j):

$$z^{LB} = -\sum_{i\in I, t\in T} \hat{\beta}_{it} s_{it} + \sum_{j\in J} z_j^{VLB} + \sum_{j\in J} z_j^{TLB}$$

Theorem 1 tells us how to solve (PS-V-j), the viewer-type subproblem for campaign j. We take $\psi_j^V = \alpha_1/\hat{s}_j$, treat $p_j^V = (\frac{1}{2}p_j + \gamma_j)$ as the p_j defined in the theorem, and treat $\hat{\beta}_i^{Vj} = \eta_{ij}$ as the $\hat{\beta}_i$ defined in the theorem. Invoking the theorem yields the following. Assuming without loss of generality that all $i \in \Gamma(j)$ viewer types are sorted by $\hat{\beta}_i^{Vj}$, i.e., $\hat{\beta}_1^{Vj} \leq \hat{\beta}_2^{Vj} \leq \cdots \leq \hat{\beta}_{m_j}^{Vj}$, where $m_j = |\Gamma(j)|$, we define $s_i^{Vbj} = \sum_{i'=1..i-1} s_{i'}$; $s_i^{Vaj} = \sum_{i'=i+1..m_j} s_{i'}$; $c_{ij}^V = \psi_j^V s_i (s_i^{Vaj} - s_i^{Vbj}) + s_i \hat{\beta}_i^{Vj}$; $\tilde{c}_{ij}^V = \sum_{i'=1..i} c_{i'j}^V$; $\tilde{s}_{ij}^V = \sum_{i'=1..i} s_{i'}$; $\pi_{ij}^V = \tilde{c}_{ij}^V / \tilde{s}_{ij}^V$; and $i^* = \arg\min_{i\in\{1..m_j\}} \pi_{ij}^V$. If $\pi_{i^*j}^V > p_j^V$, then $y_j^* = d_j$, $x_{ij}^* = 0\,\forall i = 1..m_j$, and the corresponding optimal value is $p_j^V d_j$. Otherwise, the optimal value is $\pi_{i^*j}^V d_j$ with corresponding optimal solution $y_j^* = 0$, and

$$x_{ij}^* = \begin{cases} d_j / \tilde{s}_{i^*j}^V & \text{for } i \leq i^* \\ 0 & \text{for } i > i^* \end{cases}$$

Solving (PS-T-j), the time subproblem for campaign j, is more involved. For this, we make use of a very nice further decomposition. Indeed, we break (PS-T-j) into a number of "viewer-type and time" subproblems, each of which is responsible for spreading impressions across viewer types for a single time period. Each such "viewer-type and time" subproblem has the required structure to invoke Theorem 1, and thus can be solved analytically. Moreover, via backwards substitution, we can encode the optimal values from these "viewer-type and time" subproblems into the objective function of the time subproblem, finally solve the time subproblem via another invocation of Theorem 1, and then analytically compute the optimal solution to each "viewer-type and time" subproblem given the optimal solution to the time subproblem. To derive this construction, we first define (PS-VT-jt), the campaign-j subproblem which spreads impressions across viewer types within a single time period t. The subproblem (PS-VT-jt) has decision variables $\{x_{ijt}\}_{i\in\Gamma(j)\cap T_j}$, has optimal value $z_{jt}^{VTLB}(\hat{\beta}, \eta, w_{jt})$, and is defined as:

$$\text{(PS-VT-jt)} \qquad z_{jt}^{VTLB}(\hat{\beta}, \eta, w_{jt}) = \min\ \alpha_3 \sum_{t\in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} |x_{hjt} - x_{ijt}| + \sum_{i\in\Gamma(j), t\in T_j} (\hat{\beta}_{it} - \eta_{ij}) s_{it} x_{ijt}$$
$$\text{s.t.} \sum_{i\in\Gamma(j)} s_{it} x_{ijt} = w_{jt} \quad \text{(linking (i))}$$
$$x_{ijt} \geq 0\ \forall i \in \Gamma(j) \cap T_j$$

Using the optimal value from (PS-VT-jt), we can write the campaign-j time subproblem (PS-T-j) equivalently as:

$$\text{(PS-T2-j)} \qquad z_j^{TLB}(\hat{\beta}, \eta, \gamma) = \min\ \alpha_2 \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} |w_{jt} - w_{j\tau}| + (\frac{1}{2}p_j - \gamma_j) y_j^T + \sum_{t\in T_j} z_{jt}^{VTLB}(\hat{\beta}, \eta, w_{jt})$$
$$\text{s.t.} \sum_{t\in T_j} w_{jt} + y_j^T = d_j \quad \text{(demand (i))}$$
$$w_{jt} \geq 0\ \forall t \in T_j; \qquad y_j^T \geq 0$$

With this representation, (PS-T2-j) encodes $|T_j|$ inner optimization problems of the form (PS-VT-jt) into its objective function.

Theorem 1 tells us how to solve (PS-VT-jt), the campaign-j subproblem for spreading impressions across viewer types within time period t. We take $\psi_j^t = \alpha_3/\hat{s}_{jt}$, treat $\hat{\beta}_i^{tj} = (\hat{\beta}_{it} - \eta_{ij})$ as the $\hat{\beta}_i$ defined by the theorem, and treat w_{jt} as d_j. Moreover, since no shortfall is allowed for the "demand" constraint $\sum_{i\in\Gamma(j)} s_{it}x_{ijt} = w_{jt}$, we treat p_j as defined by the theorem as if it is $+\infty$ so that no shortfall occurs in the solution. Invoking the theorem yields the following. Assuming without loss of generality that all $i \in \Gamma(j)\cap T_j$ viewer types are sorted by $\hat{\beta}_i^{tj}$, i.e., $\hat{\beta}_1^{tj} \le \hat{\beta}_2^{tj} \le \cdots \le \hat{\beta}_{m_j}^{tj}$, where $m_j = |\Gamma(j)\cap T_j|$, we define $s_i^{tbj} = \sum_{i'=1..i-1} s_{i't}$; $s_i^{taj} = \sum_{i'=i+1..m_j} s_{i't}$; $c_{ij}^t = \psi_j^t s_{it}(s_i^{taj} - s_i^{tbj}) + s_{it}\hat{\beta}_i^{tj}$; $\tilde{c}_{ij}^t = \sum_{i'=1..i} c_{i'j}^t$; $\tilde{s}_{ij}^t = \sum_{i'=1..i} s_{i't}$; $\pi_{ij}^t = \tilde{c}_{ij}^t/\tilde{s}_{ij}^t$; and $i^* = \arg\min_{i\in\{1..m_j\}} \pi_{ij}^t$. Then, the optimal value of (PS-VT-jt) corresponding to campaign-j and time period t is $\pi_{i^*j}^t w_{jt}$, with corresponding optimal solution:

$$x_{ijt}^* = \begin{cases} w_{jt}/\tilde{s}_{i^*j}^t & \text{for } i \le i^* \\ 0 & \text{for } i > i^* \end{cases} \tag{EC.11}$$

Moreover, denoting $\pi_{jt}^* \equiv \pi_{i^*j}^t$, we have $z_{jt}^{VTLB}(\hat{\beta}, \eta, w_{jt}) = \pi_{jt}^* w_{jt}$, where Theorem 1 gives us π_{jt}^* as a function of $\hat{\beta}$ and η. Thus, the campaign-j time subproblem (PS-T2-j) can be simplified to:

$$\text{(PS-T3-j)} \qquad z_j^{TLB}(\hat{\beta}, \eta, \gamma) = \min\ \alpha_2 \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} |w_{jt} - w_{j\tau}| + \left(\frac{1}{2}p_j - \gamma_j\right) y_j^T + \sum_{t\in T_j} \pi_{jt}^* w_{jt}$$

$$\text{s.t.} \sum_{t\in T_j} w_{jt} + y_j^T = d_j \qquad \text{(demand (i))}$$

$$w_{jt} \ge 0\ \forall t \in T_j; \qquad y_j^T \ge 0$$

Now we notice that the time subproblem (PS-T3-j) is in the form required to also solve it using Theorem 1. We take $\psi_j^T = \alpha_2/|T_j|$, treat $p_j^T = (\frac{1}{2}p_j - \gamma_j)$ as the p_j defined in the theorem, and treat $\hat{\beta}_t^{Tj} = \pi_{jt}^*$ as the $\hat{\beta}_i$ defined in the theorem. Moreover, w_{jt} takes the place of x_{ij} and correspondingly we assume $s_i = 1$ for all viewer types $i \in I$. Invoking the theorem yields the following. Assuming without loss of generality that all $t \in T_j$ time periods are sorted by $\hat{\beta}_t^{Tj}$, i.e., $\hat{\beta}_1^{Tj} \le \hat{\beta}_2^{Tj} \le \cdots \le \hat{\beta}_{m_j}^{Tj}$, where $m_j = |T_j|$, we define $s_t^{Tbj} = \sum_{t'=1..t-1} 1 = t-1$; $s_t^{Taj} = \sum_{t'=t+1..m_j} 1 = m_j - t$; $c_{tj}^T = \psi_j^T(s_t^{Taj} - s_t^{Tbj}) + \hat{\beta}_t^{Tj}$; $\tilde{c}_{tj}^T = \sum_{t'=1..t} c_{t'j}^T$; $\tilde{s}_{tj}^T = \sum_{t'=1..t} 1 = t$; $\pi_{tj}^T = \tilde{c}_{tj}^T/\tilde{s}_{tj}^T$; and $t^* = \arg\min_{t\in\{1..m_j\}} \pi_{tj}^T$. If $\pi_{t^*j}^T > p_j^T$, then $y_j^{T*} = d_j$, $w_{jt}^* = 0\ \forall t = 1..m_j$, and the corresponding optimal value is $p_j^T d_j$. Otherwise, the optimal value is $\pi_{t^*j}^T d_j$ with corresponding optimal solution $y_j^{T*} = 0$, and

$$w_{jt}^* = \begin{cases} d_j/t^* & \text{for } t \le t^* \\ 0 & \text{for } t > t^* \end{cases} \tag{EC.12}$$

Finally, once we have solution (EC.12) for (PS-T-j), we can substitute it into (EC.11) to yield the solutions for (PS-VT-jt) corresponding to all $t \in T_j$.

We have now shown how to completely solve all subproblems analytically, which can be summed up as, for each campaign j, (1) solving (PS-V-j) analytically using Theorem 1; (2) solving (PS-VT-jt) analytically using Theorem 1 for all time periods $t \in T_j$; (3) using the optimal values π_{jt}^* computed for (PS-VT-jt), $t \in T_j$, to solve

(PS-T-j) analytically using Theorem 1; and then finally (4) numerically computing the solutions to (PS-VT-jt), $t \in T_j$, using the optimal solution from (PS-T-j) by substituting (EC.12) into (EC.11). Equivalently, we are able to solve (PS-MULT) analytically.

There are two major elements of our decomposition method that still need to be developed before we can solve the original problem, (P-MULT-ORIG). As in the single-period decomposition scheme discussed in Section 6 and Appendix C, we additionally need (1) a method for choosing a "good" price vector $(\hat{\beta}, \eta, \gamma)$; and (2) a method to convert one or more solutions to (PS-MULT) into a near-optimal solution for (P-MULT-ORIG) that satisfies all of the constraints that were dualized when relaxing (P-MULT-ORIG) to produce (PS-MULT) (i.e., the supply constraints, as well as linking constraints (ii) and (iii)). We proceed as in the single-period case, and develop a Dantzig-Wolfe decomposition scheme that we modify for performance reasons to solve the Dantzig-Wolfe master problem only once every five iterations, instead updating the price vector $(\hat{\beta}, \eta, \gamma)$ using subgradient optimization in the intervening four of five iterations.

Our solution method begins by first solving a heuristic to get an initial feasible solution. Our heuristic has two stages. In stage one, the heuristic iterates through the list of campaigns $j = 1..|J|$, and attempts to assign $s_{it}\theta_j$ impressions of each viewer type $i \in \Gamma(j)$ in each time period $t \in T_j$ to campaign j. Such allocations are processed in sequence, and at some point the remaining supply of viewer type i for time period t may be less than the $s_{it}\theta_j$ impressions campaign j is requesting. At such point, the campaign j gets the remaining supply of s_{it}, and subsequent campaigns do not get any impressions from viewer type i at time t. When stage one is completed, some viewer types in some time periods may be completely allocated, and yet some campaigns may still have demand that has not been allocated. This can happen whenever a campaign cannot grab all $s_{it}\theta_j$ impressions from each viewer type $i \in \Gamma(j)$ and time period $t \in T_j$ that it wants. Stage two then iterates through the list of campaigns once again, but this time each campaign j checks to see how many impressions in each matching viewer type $i \in \Gamma(j)$ and time period $t \in T_j$ are still available, and grabs the same proportion of slack from each matching viewer type in each time period. Ideally, after booking this proportion of slack, the demand for campaign j is fully allocated. However, it could be that the proportion of slack needed to satisfy demand would exceed one; in that case, the campaign grabs all remaining slack from each matching viewer type in each matching time period, and remains partially unallocated. We denote the solution computed by this heuristic as $\left\{\{x_{ijt0}\}_{(i,j)\in\Gamma}, \{y_{j0}\}_{j\in J}\right\}$, and compute the values corresponding to each j-component $\{v_{j0}\}_{j\in J}$ using the terms corresponding to campaign j in the Gini-based objective (EC.9). In other words, for $n = 0$, we use linking constraints (i), (ii), and (iii) to first compute $x_{ijn} = (1/s_i)\sum_{t\in T_j} s_{it}x_{ijtn}$; $w_{jtn} = \sum_{i\in\Gamma(j)} s_{it}x_{ijtn}$; and $y^T_{jn} = y_{jn}$; and then we substitute those values into the following formulas:

$$v^V_{jn} := \alpha_1 \frac{1}{\hat{s}_j} \sum_{(h,i)\in\Gamma_0(j)} s_h s_i |x_{hjn} - x_{ijn}| + \frac{1}{2} p_j y_{jn}, \text{ and} \tag{EC.13}$$

$$v^T_{jn} := \alpha_2 \frac{1}{|T_j|} \sum_{(t,\tau)\in T_0(j)} |w_{jtn} - w_{j\tau n}| + \alpha_3 \sum_{t\in T_j} \frac{1}{\hat{s}_{jt}} \sum_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} |x_{hjtn} - x_{ijtn}| + \frac{1}{2} p_j y^T_{jn}. \tag{EC.14}$$

We then use this heuristic solution to initialize the first set of columns, corresponding to $n = 0$, in the Dantzig-Wolfe master problem. The Dantzig-Wolfe master problem, which has parameters x_{ijtn}, x_{ijn}, y_{jn}, y^T_{jn}, v^T_{jn}, and v^V_{jn}, $n = 0..N$, $j \in J$, $i \in \Gamma(j)$, $t \in T_j$, and decision variables λ^V_{jn}, λ^T_{jn}, $n = 0..N$, $j \in J$, is formulated as follows:

$$
\begin{aligned}
\text{(PM-MULT)} \qquad z^{UB} = \min \; & \sum_{j \in J, n=0..N} (v^V_{jn}\lambda^V_{jn} + v^T_{jn}\lambda^T_{jn}) & & \\
\text{s.t.} \; & \sum_{n=0..N, j \in \Gamma(it)} x_{ijtn}\lambda^T_{jn} \leq 1 & \forall i \in I, \forall t \in T & \quad \text{(supply)} \\
& \sum_{n=0..N} s_i x_{ijn}\lambda^V_{jn} = \sum_{n=0..N} \sum_{t \in T_j} s_{it} x_{ijtn}\lambda^T_{jn} & \forall (i,j) \in \Gamma & \quad \text{(linking (ii))} \\
& \sum_{n=0..N} y_{jn}\lambda^V_{jn} = \sum_{n=0..N} y^T_{jn}\lambda^T_{jn} & \forall j \in J & \quad \text{(linking (iii))} \\
& \sum_{n=0..N} \lambda^V_{jn} = 1 & \forall j \in J & \quad \text{(convexity)} \\
& \sum_{n=0..N} \lambda^T_{jn} = 1 & \forall j \in J & \quad \text{(convexity)} \\
& \lambda^V_{jn} \geq 0, \; \lambda^T_{jn} \geq 0 & \forall j \in J, \forall n = 0..N &
\end{aligned}
$$

In a classical Dantzig-Wolfe decomposition, we would begin with $N = 0$; solve the master problem (PM-MULT); take $\hat{\beta}_{it}$ to be the dual values of the supply constraints for each viewer type $i \in I$ at each time period $t \in T$; take η_{ij} to be the dual values of the "linking (ii)" constraints for each $(i,j) \in \Gamma$; take γ_j to be the dual values of the "linking (iii)" constraints for each $j \in J$; re-solve the subproblems (PS-V-j), (PS-VT-jt), and (PS-T-j) for all campaigns $j \in J$ and time periods $t \in T_j$ (analytically, via repeated use of Theorem 1, as we have described); increment the iteration counter $N := N + 1$ and record the subproblem solutions as $\{x_{ijtN}\}_{j \in J, i \in \Gamma(j), t \in T_j}$, $\{x_{ijN}\}_{j \in J, i \in \Gamma(j)}$, $\{y_{jN}\}_{j \in J}$, $\{y^T_{jN}\}_{j \in J}$, with values $\{v^V_{jN}\}_{j \in J}$ and $\{v^T_{jN}\}_{j \in J}$, computed again using (EC.13) and (EC.14); add the subproblem solutions as a new set of columns to the master problem; and then iterate, repeatedly solving the master problem (PM-MULT) and subproblems (PS-V-j), (PS-VT-jt), and (PS-T-j) for all $j = 1..|J|$, $t \in T_j$, in this fashion until a termination criterion is attained.

We use the master problem (PM-MULT) to produce a near-optimal (and feasible) solution to (P-MULT-ORIG), where the variables λ^V_{jn} and λ^T_{jn} determine the weights to apply to the solutions produced by the viewer-type and time subproblems for campaign j in iteration n, respectively. The master problem (PM-MULT) yields an upper bound and the subproblems collectively yield the lower bound $z^{LB}(\hat{\beta}, \eta, \gamma)$. Therefore, at each iteration, we can compute the optimality gap $(z^{UB} - z^{LB})/z^{LB}$, and terminate when it is below a desired threshold.

As in the single-period case, we solve the master problem (PM) only every fifth iteration, and employ subgradient optimization to update the prices $\hat{\beta}_{it}$, $i \in I$, $t \in T$; η_{ij}, $(i,j) \in \Gamma$; γ_j, $j \in J$ for four out of every five iterations. After we update the price vector using subgradient optimization, we solve the subproblems (PS-V-j), (PS-VT-jt), and (PS-T-j) for all campaigns $j \in J$ and time periods $t \in T_j$, just as before, and update the master problem's columns using the subproblem solutions.

To use subgradient optimization, we note that the $(i,t)^{th}$ component of the subgradient corresponding to the $(i,t)^{th}$ relaxed supply constraint is $g_{it} := \sum_{j \in \Gamma(i) \cap J_t} s_{it} x_{ijt} - s_{it}$; the $(i,j)^{th}$ component of the subgradient corresponding to the $(i,j)^{th}$ relaxed "linking (ii)" constraint is $g_{ij} := s_i x_{ij} - \sum_{t \in T_j} s_{it} x_{ijt}$; and the j^{th} component of the subgradient corresponding to the j^{th} relaxed "linking (iii)" constraint is $g_j := y_j - y_j^T$. Using the step size σ_n at iteration n, we update the price vector $(\hat{\beta}, \eta, \gamma)$ using subgradient optimization as follows:

$$\hat{\beta}_{it} := \max(0, \hat{\beta}_{it} + \sigma_n g_{it});$$
$$\eta_{ij} := \eta_{ij} + \sigma_n g_{ij};$$
$$\gamma_j := \gamma_j + \sigma_n g_j.$$

We use step sizes

$$\sigma_n = \frac{2(z^{UB} - z^{LB}(\hat{\beta}, \eta, \gamma))}{\sum_{i \in I, t \in T} g_i^2 + \sum_{(i,j) \in \Gamma} g_{ij}^2 + \sum_{j \in J} g_j^2},$$

which have been shown in the literature (e.g., Fisher 2004) to make steady progress toward the optimal dual prices.

To summarize, we solve (P-MULT-ORIG) to near-optimality by (1) initializing the master problem (PM-MULT) at iteration $n = 0$ using a heuristic solution; (2) incrementing the iteration counter $(N := N + 1)$; (3a) solving the master problem (PM-MULT) every fifth iteration to get a new price vector $(\hat{\beta}, \eta, \gamma)$, a near-optimal feasible solution, and the value of this solution which serves as an upper bound z^{UB} for (P-MULT-ORIG); (3b) employing subgradient optimization for four out of every five iterations to update the values of $(\hat{\beta}, \eta, \gamma)$, from the past iteration; (4) solving all campaign-j subproblems for the viewer-type, time, and "viewer-type and time" dimensions analytically using Theorem 1 to get new columns to be placed into the master problem (PM-MULT) as well as a new lower bound z^{LB}; (5) updating either the lower bound or the upper bound, or both, if they are tighter than they were last iteration; (6) computing the optimality gap $(z^{UB} - z^{LB})/z^{LB}$ and terminating if the termination criterion has been attained (e.g., optimality gap $< 1\%$); or continuing to iterate by going to step 2, otherwise.

E.7. Tail Gini

In the baseline multi-period Gini model (MB) from §E.4, penalties are incurred if the impression allocation does not fall within the impression target intervals $[l_{ij}, u_{ij}]$, $[l_{jt}, u_{jt}]$, and $[l_{ijt}, u_{ijt}]$ defined by the constraints (EC.5d), (EC.5e), and (EC.5f). To formulate Gini-based models that incorporate the idea of impression target ranges, where no penalty is incurred when the impression allocation is within the target range, we use the tail-Gini concept (see the related studies of Mansini et al. 2007; Ogryczak and Ruszczyński 2002a,b). Consider an n-dimensional vector x. The tail-Gini coefficient of x is the conditional average of the absolute differences between the components x_i and x_j of each pair $(x_i, x_j), i, j = 1, \ldots, n$ divided by twice the mean value $\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$:

$$TG = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} (|x_i - x_j| - r)^+}{2n^2 \mu},$$

where r is a user-defined threshold value. The above formula shows that the tail-Gini measure is based on the idea of conditional deviation and is equal to the standard Gini (1) if r is equal to 0. A pair (x_i, x_j) has an impact on the value of TG if and only if the absolute value of the difference between x_i and x_j is strictly larger than r.

Using the above definition, we may formulate tail-Gini metrics corresponding to (5), (EC.6), and (EC.7) as:

$$TG_j = \left(\frac{1}{\hat{s}_j}\right) \frac{\sum\limits_{(h,i)\in\Gamma_0(j)} s_h s_i \left(|x_{hj} - x_{ij}| - r_{hi}\right)^+}{\sum\limits_{i\in\Gamma(j)} s_i x_{ij}} ,$$

$$TG_{jt} = \left(\frac{1}{\hat{s}_{jt}}\right) \frac{\sum\limits_{(h,i)\in\Gamma_0(j)} s_{ht} s_{it} \left(|x_{hjt} - x_{ijt}| - r_{hit}\right)^+}{\sum_{i\in\Gamma(j)} s_{it} x_{ijt}} ,$$

and

$$TG_j^T = \left(\frac{1}{|T_j| w_j}\right) \sum_{(t,\tau)\in T_0(j)} \left(|w_{jt} - w_{j\tau}| - r_t\right)^+ ,$$

where r_{hi}, r_{hit}, and r_t are fixed parameters.

Using the tail-Gini metric, we obtain the following objective function:

$$\min\ \alpha_1 \sum_{j\in J} w_j TG_j + \alpha_2 \sum_{j\in J} w_j TG_j^T + \alpha_3 \sum_{j\in J, t\in T_j} w_{jt} TG_{jt} + \sum_{j\in J} p_j y_j .$$