How Should Marketplaces Decide Who to Show?

PENG SHI
USC Marshall School of Business

Across three papers, I study how digital marketplaces should decide which providers to recommend when providers set their own prices. The central lesson is that prices let providers manage demand and capacity, so the platform often need not ration attention or track capacities directly. This yields three simple prescriptions. When expected customer surplus is measured well, ranking by that score maximizes social welfare, defined as the sum of customer surplus, provider surplus, and platform revenue. When quality information is unreliable, sell customer attention at a market-clearing fee, or equivalently rank ads by Bid $\cdot$ CTR. When both signals are available, add them: Bid $\cdot$ CTR $+ \mu \cdot \widehat{\text{Quality}}$, interpretable as a quality discount on paid attention. The weight $\mu \in [0, 1]$ reflects how much the platform values customer surplus relative to provider and platform surplus. This letter explains the economic logic and connects it to practice on platforms such as Amazon, Alibaba, Google, Yelp, and HomeAdvisor.

Categories and Subject Descriptors: K.4.4 [**Computers and Society**]: Electronic Commerce

General Terms: Design, Economics, Reliability, Theory

Additional Key Words and Phrases: Market design, matching, ranking, sponsored advertising, two-sided marketplace

1. WHY THE QUESTION MATTERS

Digital platforms now mediate a large and growing share of how customers find goods, services, stays, and freelance professionals: products on Amazon and Alibaba; local services such as plumbers, auto shops, and restaurants on Yelp, Google Maps, and HomeAdvisor; stays on Airbnb; and freelance work on Upwork. The economic footprint is large: Digital Commerce 360 estimates that the top 100 global online retail marketplaces reached roughly $3.8 trillion in gross merchandise value in 2024, having roughly doubled in six years [Digital Commerce 360 2024]. The true scope is even larger, since this aggregate omits the many service, labor, and local-search platforms consumers turn to throughout daily life. Yelp alone reports tens of millions of monthly app users and hundreds of thousands of paying advertisers [Yelp 2026].

I focus on platforms that *recommend* options for the customer to compare and choose from, rather than platforms that assign a single provider (e.g., ride-hail dispatch). Across these platforms, a key value proposition is to elevate providers likely to suit the customer. The interface design varies widely across platforms: Yelp and Google search return ranked lists; Amazon arranges products in a grid of varied prominence, with sponsored cards, brand banners, and video creatives interleaved among conventional listings; Google Maps shows sponsored pins more prominently than organic ones; and lead-generation platforms often forward each customer request to a small assortment of providers. On all of these interfaces, a shared operational decision is: among the providers who could serve a customer, which ones should the platform put in front of that customer?

Author's address: Peng Shi, USC Marshall School of Business, pengshi@usc.edu.

The natural impulse is to show the "best" providers first. But what does "best" mean? Suppose a homeowner searches for a kitchen remodel contractor. Should the platform always promote the contractor with the strongest reviews? Reviews may not capture whether that contractor is a good fit for the homeowner's specific needs. But even if they did, a contractor shown first to everyone may raise prices, schedule far into the future, cherry-pick jobs, or take on more jobs than they can personally supervise. These responses may not be what homeowners want when searching for a hands-on contractor. The platform therefore needs some way to distribute recommendations across providers based on capacity. But providers may have incentives to overstate capacity, since visibility can create pricing power. At the same time, providers also want to avoid poorly matched leads because answering calls and giving quotes are time-consuming.

A natural academic framing is two-sided assortment optimization [Ashlagi et al. 2022; Aouad and Saban 2023; Housni et al. 2026; Rios and Torrico 2026]. This literature was motivated in part by dating markets, where each participant has limited attention and the platform rations how often each profile is shown. The optimization problem is computationally hard, so the literature develops sophisticated submodular optimization algorithms with constant-factor approximation guarantees. These algorithms look quite different from the familiar ranking and ad-allocation systems used in commercial marketplaces. A more fundamental gap is that this literature takes prices, and hence provider attractiveness, as exogenous. The papers discussed here instead study commercial marketplaces where providers set their own prices in response to platform policy, and may also choose how much attention to seek or accept.

The three papers discussed here address this gap. They build a tractable theory of match recommendations in marketplaces where providers choose prices in response to platform policy. The papers are [Shi 2024] (EC'22), [Shi 2026a] (EC'24), and [Shi 2026b] (EC'25, Exemplary Applied Modeling Track Paper Award). The models differ, but share a common idea: when providers set their own prices, they use price to balance demand against their own capacity, so the platform need not artificially ration attention to scarce providers, and simple recommendation and ad-pricing rules suffice. The remainder of the letter develops three prescriptions: rank by expected customer surplus when quality is measured well; use willingness to pay when it is not; and combine the two through quality-adjusted ad pricing. I then explain how the same logic applies across different platform interfaces and close with limitations and open directions.

2. RANKING BY EXPECTED CUSTOMER SURPLUS

What is the ideal metric by which a platform should rank providers? For customer type t and provider j, define

$$\begin{aligned}\text{Quality}_{tj} &:= \mathbb{E}[\text{customer surplus} \mid t, \text{showing } j] \\ &= \text{CTR}_{tj} \cdot \text{CVR}_{tj} \cdot \mathbb{E}[\text{customer surplus} \mid \text{transaction with } j].\end{aligned}$$

In words, Quality_{tj} is the expected customer surplus from directing one unit of attention from a type-t customer to provider j. (Section 5 makes "unit of attention" precise; for now, think of it as one listing impression.) The customer type may

represent a search query, geographic segment, or richer user context. Customer surplus is measured in dollars: it is the value the customer keeps after paying the provider, relative to an outside option such as not purchasing, choosing another provider, or leaving for another platform. On a platform where the customer clicks through to inspect a listing, the score decomposes into click-through rate (CTR), conversion rate conditional on click (CVR), and expected surplus from the resulting transaction. On a platform without a click step, the CTR term can be set to one.

This quality score is broader than ratings or service quality. It depends on price, as surplus is evaluated at the prices customers actually face. A higher provider price may reduce clicks or conversions, and it lowers the surplus captured by customers who still transact. A provider can therefore improve Quality_{tj} by providing a better product or service, or by setting a price that leaves customers more surplus.

When Quality_{tj} is measured exactly, ranking by this score achieves first-best social welfare in all three papers. Here, "social welfare" means the total value generated by the platform, including customer surplus, provider surplus, and platform revenue. "First-best" means the best outcome a perfectly informed central planner could achieve when given the power to choose both recommendations and provider prices. The simple policy of ranking by Quality_{tj} attains the first-best benchmark because the quality-score is evaluated at equilibrium prices and availability. Providers' own decisions internalize their capacity, even when the platform cannot observe it. As providers maximize profit, those with scarce capacity may raise prices, decline jobs, or cap incoming attention — as a busy contractor does by pausing new leads. Raising prices or declining jobs lowers the conversion rate and hence the quality score, while pausing leads reduces attention directly; either way, the provider receives less attention. Therefore, ranking by quality automatically distributes attention based on provider capacity, mimicking the first-best.

In practice, platforms rarely observe Quality_{tj} exactly. CTR and CVR can often be predicted from logs, but dollar-equivalent surplus conditional on transaction is harder to measure. Ratings, reviews, returns, complaints, and post-transaction surveys are noisy, sparse, delayed, and sometimes manipulated. Offline service platforms may not even observe whether a transaction occurred. Shi [2026b] addresses this by using a conservative lower estimate, denoted $\widehat{\text{Quality}}_{tj}$. The true Quality_{tj} is assumed to lie in the range

$$[\widehat{\text{Quality}}_{tj},\ \widehat{\text{Quality}}_{tj}/\rho],$$

where $\rho \in (0, 1]$ summarizes the reliability of the quality estimate. The paper shows that ranking by $\widehat{\text{Quality}}_{tj}$ guarantees at least a ρ fraction of first-best social welfare, and the bound is tight in the worst case. The platform need not know ρ to implement the ranking; ρ governs the welfare guarantee, not the algorithm.

The conservatism matters. If a platform ranks by optimistic estimates, it may reward providers whose reviews are sparse, manipulated, or merely lucky. A conservative score gives credit only to what the platform can support with evidence. This resembles a familiar way to interpret ratings: a provider with 4.1 stars and 1000 reviews is typically ranked above one with 4.8 stars and 5 reviews, because the former signal is more reliable. Better measurement and validation correspond to a higher reliability parameter ρ and a stronger welfare guarantee.

3. RANKING BY WILLINGNESS TO PAY

Some platforms may not have reliable quality estimates. In home services, lead-generation platforms such as Angi, HomeAdvisor, Bark, Google Local Services, Modernize, Porch, and Thumbtack often do not directly observe whether transactions occur, since contractors visit customers' homes, give quotes, and close offline. Reviews may also be compressed near the top of the scale and sparse; Raval [2024] documents that 96% of HomeAdvisor businesses have average ratings above 4 stars. Related measurement problems can arise in other markets with infrequent or hard-to-observe outcomes, or where customers expect to have future face-to-face interactions with the provider and are therefore reluctant to leave bad reviews.

A common response in lead and advertising systems is to charge providers for the opportunity to be shown, using willingness to pay as a signal of who can profitably serve the customer. HomeAdvisor connects each customer with up to a small number of providers and charges category-specific lead fees: a small repair job may cost a few dollars per lead, whereas a large remodel may cost hundreds. In the model, the fee is a market-clearing price f_t for one unit of attention from customer type t. The platform can implement this without detailed information on market parameters, simply by adjusting fees over time — raising the lead fee where leads are oversubscribed, lowering it where they go unsold.

Shi [2026a] analyzes this policy and shows that it achieves a constant fraction of first-best social welfare in the worst case. The policy is quality-agnostic: providers self-select by buying leads only when they expect to convert the customer and earn sufficient margin. Thus, when the reliability ρ of quality estimates is low, willingness to pay may signal provider relevance better than a noisy organic ranker.

The specific constant in the guarantee depends on demand assumptions: under standard regularity conditions on demand, Shi [2026a] obtains $1/(e-1) \approx 58\%$ assuming additively separable customer utilities with a vertical component and a horizontal component whose distribution is symmetric across providers; Shi [2026b] obtains $1/e \approx 37\%$ allowing arbitrarily different demand distributions across providers. Both are tight worst-case guarantees: they hold across arbitrarily many providers, heterogeneous capacities, and multiple customer types.

These partial guarantees reflect a real limitation: willingness to pay is an imperfect welfare signal. A high bid may reflect high margins, strong sales ability, or the ability to extract surplus from customers, not just customer value. Bid ranking is therefore a robust fallback when quality measurement is weak, not a substitute for quality information when it is available.

The same market-clearing logic appears in sponsored advertising. In a reduced-form view of an ad system, ads are ranked by an Ad-Rank score and admitted when the score clears a threshold θ_t, determined by the supply and demand for attention of type t. The threshold θ_t represents the marginal price of customer attention. The cost per click is the minimum bid needed for the Ad-Rank score to clear θ_t. A common baseline formula is

$$\text{AdRank} = \text{Bid} \cdot \text{CTR},$$

which was behind Google's early sponsored-search auctions [Edelman et al. 2007; Varian 2007], and remains a recognizable building block of many advertising sys-

tems. Multiplying the per-click bid by the click-through rate converts it into per-impression willingness to pay. Under this policy, the effective per-impression fee is $f_t = \theta_t$. When the platform sets the Ad-Rank threshold θ_t to be market-clearing, which approximates an auction system with no reserve price, the above Ad-Rank formula achieves the same constant-fraction welfare guarantees as described above.

This view of ads as welfare-enhancing connects to, but is distinct from, the literature on advertising as a signal [Kihlstrom and Riordan 1984; Milgrom and Roberts 1986]. In classic signaling stories, advertising may change what customers infer about product quality. Here, the ad is not assumed to persuade customers; rather, the platform uses providers' willingness to pay to select the most motivated and available providers to show.

The mechanism also depends on customers making informed conversion decisions. In practice, that requires accountability systems such as fraud detection, complaint handling, and removal of bad actors. With those safeguards, having providers pay for customer attention can be a welfare-enhancing discovery mechanism rather than merely a distortion.

4. COMBINING QUALITY AND WILLINGNESS TO PAY

When the platform has both conservative quality estimates and providers' willingness to pay, neither pure organic ranking nor pure bid ranking is best. Shi [2026b] studies a robust-optimization problem: the platform knows only that true quality lies in $[\widehat{\text{Quality}}_{tj},\ \widehat{\text{Quality}}_{tj}/\rho]$, and policies are evaluated by their worst-case fraction of first-best weighted welfare. Here weighted welfare is (supply-side surplus) + $\mu \cdot$ (customer surplus). Supply-side surplus is provider profit before platform fees, equivalently provider surplus plus platform revenue. The parameter $\mu \in [0, 1]$ represents how much the platform values the customer-side relative to the supply-side.[1] Strikingly, across all policies monotone in bid and estimated quality — including nonlinear ones — the best worst-case guarantee is achieved by a simple linear sum whose quality weight is exactly the μ from the weighted objective:

$$\text{AdRank}_{tj} = \text{Bid}_{tj} \cdot \text{CTR}_{tj} + \mu \cdot \widehat{\text{Quality}}_{tj}. \tag{1}$$

Both terms are in dollars per unit of customer attention: the first is per-impression willingness to pay, and the second is a conservative estimate of customer surplus per impression, weighted by μ. Noisier quality evidence lowers the conservative estimate, so bids matter more; more reliable evidence raises the quality term.

The parameter μ is a policy choice. Raising μ moves the platform toward customer surplus; lowering it moves toward supply-side surplus. Setting $\mu = 0$ recovers the policy in the previous section, which in the model attains the first-best supply-side surplus exactly. Extensive simulations further show that the platform can optimize the breakdown of supply-side surplus into provider surplus and platform revenue by utilizing additional transaction fees or commissions. Under optimal choice of such fees, $\mu = 0$ generally achieves the highest platform revenue among the policies considered. Larger μ trades some platform revenue for customer and

[1] For $\mu > 1$, the weighted objective overcounts customer surplus, so a central planner could increase it without bound by lowering prices to transfer surplus from providers to customers, implying that the first-best benchmark is infinite and therefore uninteresting. [Shi 2026b] hence restricts $\mu \leq 1$.

provider surplus, making it more attractive when the platform is willing to forgo short-term revenue for long-term customer value.

The formula also has a price interpretation. A provider's AdRank_{tj} clears the threshold θ_t when their per-impression willingness to pay $\text{Bid}_{tj} \cdot \text{CTR}_{tj}$ exceeds

$$f_{tj} = \theta_t - \mu \cdot \widehat{\text{Quality}}_{tj},$$

the equilibrium per-impression fee. Higher estimated quality therefore acts as a discount on customer attention, which in practice may be delivered through lower advertising costs, reduced commissions, rebates, or credits. Such discounts also give providers a reason to improve the quality estimate in welfare-relevant ways, including by setting prices that leave customers more surplus.

This prescription is concrete: it tells platforms what to estimate (a conservative dollar-valued customer surplus per impression) and how to incorporate it into ad prices (as a linear discount, weighted by μ). Quality-based discounts on ad fees already appear in industry practice: Alibaba's published cost-per-click formula (as of January 2025) discounted advertising costs by a promotion-quality score [Alibaba 2025], and on Amazon, marketing-agency analysis reported that products with stronger organic rank paid much less per click for comparable sponsored positions [Signalytics 2023]. What the theory adds is a welfare rationale for using such discounts and a concrete formula for platforms to calibrate and test.

5. THE SAME LOGIC ACROSS INTERFACES

The framework operates at the level of allocating customer attention. In the formulas above, an "impression" is a unit of customer attention rather than a raw page event, and different placements may correspond to different amounts: a large slot at the top might count as 3.2 impressions, a small one near the bottom as 0.5. This view aligns with a familiar industry decomposition: Varian [2006] writes the raw click-through rate as a multiplicatively separable model with a position-specific factor (how much attention the slot delivers) and an ad-specific factor (how appealing the listing is, given that attention). The CTR_{tj} in the Ad-Rank formula corresponds to the ad-specific factor alone; the position factor is absorbed into the impression count and can be estimated from data.

This abstraction makes the framework implementation-agnostic across both format and billing convention. The Ad-Rank formula in (1) requires only per-impression willingness to pay as its first term. Different formats — ranked lists, map pins, product grids, banners, and video ads — enter through different impression counts, leaving the per-impression economics unchanged. Different billing conventions differ only in the conversion factor that translates a per-unit bid into per-impression willingness to pay: CTR_{tj} does this for per-click billing; under pay-per-transaction, replace it with the rate at which a unit of attention converts to a transaction.

A more speculative extension is LLM-driven search. As AI assistants mediate how people find products and services, "attention" may include the prominence of a mention, the wording used, and whether the assistant offers a direct "book now" action. The same prescription suggests charging per unit of attention, with a discount for conservatively estimated customer surplus. When responses include clickable links, the cost-per-click form in (1) applies directly, with CTR_{tj} capturing

appeal conditional on attention. The existing results do not model conversational persuasion or trust formation; those are open directions.

6. LIMITATIONS AND NEXT STEPS

To derive clean results, the theory fixes the total amount of customer attention and studies how that attention should be allocated. An assortment extension in Shi [2026b] allows providers shown together to affect one another's demand and equilibrium prices, and the above results extend to that setting. This still leaves room for richer models of customer search in which platform policies affect whether customers use the platform in the first place and how much attention they are willing to devote. It would be interesting to analyze the proposed policies in more complex models of endogenous customer search and platform competition.

The theory is also stationary and abstracts from fluctuations in market conditions. Moreover, it characterizes long-run equilibrium outcomes, but not the transient path by which providers learn to bid and platforms adjust estimates. Nevertheless, the underlying economic logic can be incorporated into dynamic implementations, by adding a conservative quality term in the Ad-Rank formula and letting existing auction systems determine allocations. When estimating quality, more weight should be put on recent data so estimates do not become stale. Formally studying learning, cold start, and rate of convergence would further bridge the gap between theory and practice, and is a direction for future work.

The framework also uses a stylized treatment of capacity. Shi [2026b] extends the results to richer settings, including allowing providers to choose their own capacity and accommodating platform-imposed allocation constraints such as diversity requirements. Even so, capacity binds only in expectation: providers are assumed to have enough scheduling flexibility to serve demand as long as the overall rate fits within their capacity. This approximation is less applicable when matches are time-sensitive and commit a large share of a provider's capacity for an extended period — for example, finding a long-term nanny.

The most important implementation challenge is measuring Quality_{tj}. The appendix of Shi [2026b] gives two starting points. A survey-based approach asks customers whether they transacted, with whom, and how much more they would have paid; the responses train a predictive model of expected surplus per impression. A revealed-preference approach uses logged impressions, prices, and transactions to estimate demand and dollar-valued customer surplus, drawing on randomized price variation or valid instruments such as cost shifters. In both cases, the appendix applies conservative shrinkage before the estimate enters the Ad-Rank formula.

Production systems add another layer. Real ranking stacks combine candidate generation, relevance filters, safety rules, pacing, reserves, and reranking. None of this complexity is fatal to the theory's applicability: the formula serves as a benchmark for allocating and pricing attention among eligible providers, after the platform has applied its other operational constraints. Translating that benchmark into a tested production rule remains substantial work. Because the theory describes long-run equilibrium, short A/B tests are unlikely to be informative; field tests should run long enough, and announce policy changes clearly enough, for providers to respond by adjusting prices, ad spend, and capacity. Autobidders and

pricing agents ease this by pushing provider behavior toward the profit optimization the theory assumes; platforms could go further by offering decision-support tools that jointly optimize pricing and ad spend.

These limitations point to a rich agenda spanning empirical estimation, theoretical extensions, and production design. The SIGecom community — with its blend of economic modeling, algorithm design, market design, and empirical platform research — is well placed to advance it, and I welcome collaboration in closing the gap between theory and practice.

REFERENCES

Alibaba. 2025. How will keyword advertising charge me? Available at https://so.alibaba.com/s/ggs/category?categoryId=1000045920&questionId=1000100938. Accessed January 2025.

Aouad, A. and Saban, D. 2023. Online assortment optimization for two-sided matching platforms. *Management Science 69,* 4, 2069–2087.

Ashlagi, I., Krishnaswamy, A. K., Makhijani, R., Saban, D., and Shiragur, K. 2022. Assortment planning for two-sided sequential matching markets. *Operations Research 70,* 5, 2784–2803.

Digital Commerce 360. 2024. Top global online marketplaces: Key data and statistics. Available at https://www.digitalcommerce360.com/top-online-marketplaces-data-stats/. Accessed May 2026.

Edelman, B., Ostrovsky, M., and Schwarz, M. 2007. Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords. *American Economic Review 97,* 1, 242–259.

Housni, O. E., Hennebelle, U., and Torrico, A. 2026. Two-sided assortment optimization: Adaptivity gaps and approximation algorithms.

Kihlstrom, R. E. and Riordan, M. H. 1984. Advertising as a signal. *Journal of Political Economy 92,* 3, 427–450.

Milgrom, P. and Roberts, J. 1986. Price and advertising signals of product quality. *Journal of Political Economy 94,* 4, 796–821.

Raval, D. 2024. Do bad businesses get good reviews? Evidence across several online review platforms. *Working paper*. Available at https://deveshraval.github.io/reviews.pdf.

Rios, I. and Torrico, A. 2026. The dating heuristic: A provably strong matching algorithm for dating platforms. *Manufacturing & Service Operations Management*.

Shi, P. 2024. Optimal match recommendations in two-sided marketplaces with endogenous prices. *Management Science 71,* 9, 7431–7448.

Shi, P. 2026a. The welfare effects of selling leads in a two-sided marketplace. *Management Science, Forthcoming.* Preprint available at https://ssrn.com/abstract=4727198.

Shi, P. 2026b. Welfare-optimal policies for sponsored advertising in a two-sided marketplace. Working paper, updated May 2026. Available at https://ssrn.com/abstract=5132218.

Signalytics. 2023. What you didn't know about Amazon's "second-price" auction. Available at https://www.linkedin.com/pulse/what-you-didnt-know-amazons-second-price-auction-signalytics/. Accessed January 2025.

Varian, H. R. 2006. The economics of internet search. *Rivista di politica economica 96,* 11/12, 177–191.

Varian, H. R. 2007. Position auctions. *International Journal of Industrial Organization 25,* 6, 1163–1178.

Yelp. 2026. Investor relations overview. Reports 28 million monthly average app unique devices and 496,000 paying advertising locations for 2025. Available at https://www.yelp-ir.com/overview/default.aspx.