Designing Marketplaces and Civic Engagement Platforms

Nikhil Garg

Many crucial societal interactions are now mediated by algorithmic systems. We buy goods, find work and hire each other, discuss current events, and make public decisions through online platforms. Non-profit and government actors further use such systems to assign kids to schools [1], organs to patients [11], and food to food banks [10]. The promise of these socio-technical systems is that they enable coordination at scale. Each participant can act locally according to their own incentives, information, and constraints – and make global connections and impact. When designed correctly, the system helps people to together achieve some shared goal, and ensures that the benefits are divided fairly; meanwhile, bad designs waste resources and privilege some participants to the detriment of others.

Building these systems is challenging. The designer must decide who can participate, what participants can do, and how they communicate to each other and the platform – while respecting business, legal, and human constraints. Each decision affects participant incentives and information, and hence behavior and system outcomes. The specific context determines good design. For example, commodity markets such as ride-hailing centrally set prices, while ones with more heterogeneous products like lodging allow hosts to set their own, reflecting the relative amount of information available to participants and the platform.

Many disciplines now consider the design of such systems. Practitioners at large companies run thousands of experiments (A/B tests) a day, to evaluate everything from user interfaces and site design to algorithms for matching, recommendation, and pricing [8, 12]. Whether a particular change is deployed depends on how it performs on a suite of metrics. Such methods are best when theory and intuition is not precise enough to choose among similar designs. For example, Google tested 40 shades of blue for web links [2].

However, experimental and data-driven methods are insufficient to design modern socio-technical systems. It is often too expensive or simply infeasible to experimentally consider too many options. Some changes may be too sensitive or risky to deploy, even in a limited environment. Such difficulties are compounded for changes that are public-facing or have effects that may not be apparent on the time scale of an experiment. For example, in ride-hailing markets, substantial changes to how drivers are paid should be paired with public communication, naturally limiting how often such changes can be made.

Theoretical mechanism and market design models shed insight in such cases; under assumptions on participant utility functions, they seek to predict behavior under large design changes. Such approaches are especially successful at constructing system designs that provably are optimal under some objective (such as welfare), with attractive properties (such as strategy-proofness). Different policies are compared by connecting individual actions under the model to a global objective. Papers shed high-level insight on practical problems, and researchers apply their training to design mechanisms outside academia, across domains [3].

However, as the aphorism, "all models are wrong but some are useful" [6] suggests, design optimality guarantees are useful to the extent that the underlying model approximates reality. Real-world behavior often differs substantially from that assumed by theoretical models, and mechanisms in practice often face subtle business, legal, and human constraints that are difficult to model. Such factors make it difficult to predict a mechanism's practical impact *a priori*, before it is deployed or evaluated in a particular context.

These competing aspects of data-driven and theoretical approaches lead to a division of labor to design and build socio-technical systems across domains; the latter provides high-level guidance and ideas, and the former evaluates specific proposals and optimizes them in context. However, there are gaps at the interface of these approaches – where platforms must make fine-grained design choices that are infeasible to wholly evaluate experimentally, but for which coarse theoretical insights are insufficient:

- Optimal mechanisms may not be implementable in practice, for context-specific reasons. How do we test assumptions and analyze good-enough approximations for a given application?
- Some experimentally-made design decisions affect downstream decisions and platform objectives. How do we connect short-lived experimental measures to these long-run effects?

Motivated by such gaps, I demonstrate several approaches to connect the practically feasible to the theoretically optimal, in service of building useful systems and solving central socio-technical design challenges. One recurring lesson is that this approach requires focusing on a particular application and developing requisite domain knowledge; through collaborations with practitioners, my work has informed systems at Uber, a large online labor platform, and in participatory budgeting elections across the United States. The dissertation is organized as follows.

Part I, "Pricing in Online Marketplaces" This part is composed of a single chapter (Chapter 2),[1] "Driver Surge Pricing," which presents a surge pricing scheme for drivers in ride-hailing platforms. The work grew out of my summer at Uber; I was a data science intern on the team building a new driver surge mechanism (now deployed across the US), in which the surged component of a trip payment is *additive* (independent of trip length) as opposed to *multiplicative* (proportional to trip length), the historical standard.

The chapter presents the theoretical foundation that informed this change. Due to the temporal dynamics of surge – in which certain time periods are more valuable than other periods, to balance the supply to available drivers with the demand for rides – trips of different lengths have different driver opportunity costs. We model surge evolution as a continuous-time Markov chain; in our model, we show that, with traditional, multiplicative pricing schemes, strategically rejecting certain trip requests may maximize an individual driver's earnings, to the detriment of riders and other drivers. For example, it may be advantageous to reject short trips during surge in the hopes of getting a longer surge trip. We then develop an incentive compatible pricing scheme with an approximately affine, closed-form expression. Such simplicity is important in practice to enable transparency and

[1] With H. Nazerzadeh. Major revision at *Management Science*. In the 2020 *Conference on Economics and Computation (EC)*.

communication of surge prices to drivers through a heat-map, and stands in contrast to previous works which resolve such strategic concerns through prices that emerge from a global optimization framework [4, 5, 9]. Finally, through both calibrated simulations and by analyzing counter-factual earnings from more than 500,000 ride-hailing trips, we validate that our proposal would increase incentive compatibility and driver earnings stability in practice.

Part II, "Designing Rating Systems in Online Marketplaces" This part considers rating inflation: sellers overwhelmingly receive positive ratings (on AirBnB, for example, almost 95% of hosts have an average rating of at least 4.5/5 stars [13]). Such inflation leads to uninformative rating systems in which noise dominates.

In **Chapter 3**,[2] we study how the platform can choose the multiple choice question that it asks raters. Each potential question induces a joint distribution between the seller's true underlying quality and the ratings they receive. For example, asking "How did this freelancer compare to others you've hired" versus, "Please rate the seller from 1 to 5 stars," yields different responses, for the same quality seller. We develop a large deviations based framework to quantify how quickly the platform recovers the true ranking of sellers, given this joint distribution. This framework provides the ideal metric for an A/B test, connecting a design to a platform's long-term goals, without needing to run a long experiment to measure the outcome directly.

We then run an experiment on a large online labor market and show that platforms can get informative ratings, by leveraging *positive-skewed, verbal* label scales (e.g., one scale ranges from *Below Average* to *Best Freelancer I've Hired*). In the experiment, clients rate freelancers through various verbal and numeric scales. These scales induce substantially different rating behaviors: while 80.6% of freelancers receive the best numeric rating, less than 35.8% receive top verbal ratings; furthermore, clients are up to 31.8% more likely to rehire the freelancer after giving them a top rating on a verbal scale than after giving them the top numeric score. Our theoretical framework quantifies the resulting information gain and provides a principled way to choose among the scales given the behavioral data. These results serve as a positive contrast to a long line of work proposing various changes that do not prevent inflation.

In **Chapter 4**,[3] we show how a platform's informational priorities should affect the rating system design. In commodity markets such as ride-hailing, it is essential to separate unacceptable from acceptable participants as quickly as possible. In superstar markets, on the other hand, fine differentiation among the best sellers is most important. We formalize such goals as weighted versions of Kendall τ distance between the estimated and true participant rankings. Then, in a setting in which rater responses are binary, we develop an efficient non-convex optimization algorithm to find the optimal joint distribution, i.e., the relationship between the participant's quality and the probability at which they should receive a positive rating. This joint distribution maximizes the asymptotic weighted accuracy and the large deviations rate at which it is reached. Our algorithm exploits a dimensionality reduction in which only ranking mistakes between similar participants can dominate the large deviations rate at

[2] With R. Johari. In *M&SOM* (Accepted) and the 2020 *Conference on Economics and Computation (EC)*.
[3] With R. Johari. In the 2019 *International Conference on Artificial Intelligence and Statistics* (AISTATS).

which the error decays. We find, for example, that it is optimal for most participants to receive primarily positive ratings when the goal is to identify unacceptable ones.

Part III, "Designing Voting Mechanisms on Civic Engagement Platforms" Civic engagement platforms enable people to collectively make complex, public decisions in an axiomatically fair way, applying the social choice tradition of Kenneth Arrow in a computational age. In participatory budgeting, people vote on how to allocate millions of dollars across many candidate projects. A key challenge in such systems is to design the *elicitation mechanism*: participants must be able to share their opinions in a manner that is simple, expressive enough for decisions that lie in high-dimensional spaces, and yet enables provably efficient aggregation. This part develops such mechanisms in a theory-driven way and then tests them through synthetic experiments and deployments on real municipal PB elections.[4]

In **Chapter 5**,[5] we develop a new method for people to collectively make a decision on a societal budget. Voters are sequentially asked for their ideal budget within a constraint set determined by the previous voter's answer. This process simulates stochastic gradient descent, and the asymptotic output provably maximizes societal welfare in certain settings. In particular, if each voter's dis-utility for a budget is its ℓ_p distance from their ideal budget, then asking each voter for their favorite budget in a local ℓ_q dual ball provides a stochastic gradient for the societal welfare function.[6] Then, sequentially querying voters in this manner simulates stochastic gradient descent. We tested our method by running elections on Amazon Mechanical Turk,[7] demonstrating that (a) one can build an intuitive user interface, and (b) the procedure converges to a consistent point across several runs, with a small number of voters.

Finally, in **Chapter 6**,[8] we show how to optimize an existing elicitation mechanism – K Approval, in which each voter identifies their favorite K candidates – in a principled manner. We extend the approach of the work in Part II, showing how the value of K (e.g., eliciting 3 candidates versus 4) determines the large deviations rate at which the asymptotic outcome is learned, even when it does not change the outcome. Then, with real voter data from over thirty elections (including from our PB platform), we demonstrate that many multi-candidate elections that select W winners are run sub-optimally; whereas voters are typically asked to identify their $K=W$ favorite candidates (e.g., $K=1$ in a winner-takes-all election), it is learning rate optimal to ask voters to identify their favorite $M>K$ candidates. This small change matters: in one election, asking for each voter's two favorite candidates versus single favorite would have been the difference between identifying the ultimate winner with a 99.9% vs 80% probability after 400 voters. This rule-of-thumb has influenced our recommendations for the elections run by our partner cities, demonstrating the impact of combining theory with data analysis.

[4] Run on our group's platform, `pbstanford.org`. The platform has been used in over fifty elections, allocating tens of millions of dollars with tens of thousands of voters. I was not involved in initial platform development, but contribute to its continued use.

[5] With V. Kamble, A. Goel, D. Marn, and K. Munagala. In *Journal of Artificial Intelligence Research* (JAIR) in 2019, and the 2017 *Conference on World Wide Web* (WWW).

[6] For $p\in\{1,2,\infty\}$. All $p,q\in[1,\infty)\cup\infty$ s.t. $1/p+1/q=1$ work for a more restricted voter behavior model.

[7] A demo of the system I built is available at `http://54.183.140.235/radius/50/mechanism/l2/`.

[8] With L. Gelauff, S. Sakshuwong, and A. Goel. In 2019 *AAAI Conference on Human Computation and Crowdsourcing* (HCOMP).

Together, this thesis demonstrates several approaches to overcome gaps at the interface between theoretical and data-driven approaches to design socio-technical systems.

(Approach 1) Empirically designing implementable mechanisms that approximate ideal ones.

Mechanism design solutions are often theoretically elegant but may violate real-world constraints. For example, they may be too complicated to be understood by regular people, and so participants may not trust the system or act strategically (sub-optimally) even when the mechanism is provably "strategy-proof" [7]. A platform designer must then implement a mechanism that obeys its constraints and best approximates the first-best solution, even when analyzing such mechanisms is theoretically intractable. Chapter 2 proceeds by first constructing an optimal solution, and then empirically and numerically comparing different approximations of it to make a practical recommendation.

(Approach 2) Deriving principled outcome measurements for experiments.

Some design challenges, especially concerning platform user interfaces, both (a) are so contingent on idiosyncratic human behavior that an experiment is necessary and (b) affect long-run platform objectives in ways that may not be transparent in a short experiment. To address this challenge in various contexts, Chapters 3, 4 and 6 each start with models of downstream platform decisions and how they're differentially affected by the distribution of responses from the user interface, yielding a statistic mapping such distributions to a long-run platform objective. Then, we use both experiments and historical data to compare specific user interface designs with respect to this statistic.

(Approach 3) Building systems to evaluate mechanism usability and to test assumptions.

Some proposed mechanisms simply need to be tested, either in the lab or in a real-world setting. While a mechanism may have desirable theoretical properties, the best proof of its practical usability is a working demonstration. In Chapter 5, we derive a new voting mechanism with theory, build a usable system, and test it through controlled experiments. Such end-to-end design and analysis is the promise of a truly interdisciplinary approach.

References

[1] Atila Abdulkadiroğlu, Parag A Pathak, and Alvin E Roth. The New York City High School Match. *American Economic Review*, 95 (2):364–367, 2005.

[2] Charles Arthur. Marissa Mayer's Appointment: What Does It Mean for Yahoo? *The Guardian*, July 2012. ISSN 0261-3077. URL https://www.theguardian.com/technology/2012/jul/16/marissa-mayer-appointment-mean-yahoo.

[3] Susan Athey and Michael Luca. Economists (and Economics) in Tech Companies. *Journal of Economic Perspectives*, 33(1):209–30, 2019.

[4] Omar Besbes, Francisco Castro, and Ilan Lobel. Spatial Capacity Planning. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3292651. URL https://www.ssrn.com/abstract=3292651.

[5] Kostas Bimpikis, Ozan Candogan, and Daniela Saban. Spatial Pricing in Ride-Sharing Networks. (ID 2868080), November 2016. URL https://papers.ssrn.com/abstract=2868080.

[6] George EP Box. Robustness in the Strategy of Scientific Model Building. In *Robustness in Statistics*, pages 201–236. Elsevier, 1979.

[7] Avinatan Hassidim, Assaf Romm, and Ran I Shorrer. "Strategic" Behavior in a Strategy-Proof Environment. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, pages 763–764, 2016.

[8] Michael Luca and Max Bazerman. *The Power of Experiments: Decision Making in a Data-Driven World* . MIT Press, 2020.

[9] Hongyao Ma, Fei Fang, and David C. Parkes. Spatio-Temporal Pricing for Ridesharing Platforms. January 2018. URL http://arxiv.org/abs/1801.04015.

[10] Canice Prendergast. The Allocation of Food to Food Banks. *EAI Endorsed Trans. Serious Games*, 3(10):e4, 2016.

[11] Alvin E. Roth, Tayfun Sönmez, and M. Utku Ünver. Kidney Exchange. *The Quarterly Journal of Economics*, 119(2):457–488, 05 2004. ISSN 0033-5533. doi: 10.1162/0033553041382157. URL https://doi.org/10.1162/0033553041382157.

[12] Matthew Salganik. *Bit by Bit: Social Research in the Digital Age*. Princeton University Press, 2019.

[13] Georgios Zervas, Davide Proserpio, and John Byers. A First Look at Online Reputation on Airbnb, Where Every Stay Is Above Average. Technical Report ID 2554500, Social Science Research Network, January 2015.