Crowdsourcing City Government: Using Tournaments to Improve Inspection Accuracy

By Edward L. Glaeser, Andrew Hillis, Scott Duke Kominers, and Michael Luca*

Big data and predictive analytics have the potential to make cities more efficient. Yet cities often lack the resources to use new data and methods effectively, and private researchers often lack the incentives to help solve the problems that cities face.

New platforms—such as Kaggle, TopCoder, and DrivenData—now enable governments and other organizations to outsource large-scale prediction problems via open tournaments. *But can open tournaments really help solve public problems?*

In this paper, we theoretically and empirically explore the potential of prediction tournaments to improve city operations and translate data science insights into practice.

In Section I, we present a formal model that compares tournaments to outsourcing via consultancy. We show that tournaments are preferable when (1) the city is comfortable with project risk and (2) there is an abundant supply of labor willing to work in exchange for a chance of public recognition and a moderate prize.

In Section II, we describe the design and initial results of an open tournament that we ran in collaboration with the City of Boston, Yelp, and DrivenData. Our contest awarded financial prizes for the algorithms that most effectively used Yelp review text to predict Boston restaurant health and sanitation violations.[1] Over seven hundred people signed up for the tournament, and fifty-five ultimately contributed at least one set of predictions, submitting a total of 449 sets of predictions. We then tested twenty-three "final" algorithm submissions out-of-sample, comparing their predictions to the true results of the 364 restaurant inspections conducted over a six-week period after the close of submissions. The evidence suggests that using the winning algorithms to identify restaurants to inspect could increase inspection efficacy significantly: We estimate that the City of Boston could be 30%-50% more productive if it allocated inspections as suggested by a top-performing algorithm from the tournament.

In Section III, we discuss general lessons for cities seeking to implement prediction tournaments. We conclude in Section IV.

I. A Brief Model of Tournaments

This section develops a framework for cities (and other organizations) that are de-

* Glaeser: Department of Economics and John F. Kennedy School of Government, Harvard University, and National Bureau of Economic Research, Littauer Center, 1805 Cambridge Street, Cambridge, MA 02138, eglaeser@harvard.edu. Hillis: Department of Economics, Harvard University, and Harvard Business School, Baker Library, Soldiers Field, Boston, MA 02163, ahillis@hbs.edu. Kominers: Society of Fellows, Department of Economics, Center of Mathematical Sciences and Applications, Center for Research on Computation and Society, and Program for Evolutionary Dynamics, Harvard University, Harvard Business School, and National Bureau of Economic Research, Room 219, Arthur Rock Center for Entrepreneurship, Soldiers Field, Boston, MA 02163, kominers@fas.harvard.edu. Luca: Harvard Business School, Baker Library, Soldiers Field, Boston, MA 02163, mluca@hbs.edu. The authors are deeply grateful for the collaboration of the City of Boston (especially Ben Batorsky, Matthew Mayrl, and Commissioner William Christopher), Yelp (Artem Avdacev, Luther Lowe, and Aaron Schur), and DrivenData (Peter Bull and Greg Lipstein), without whom the project described herein would not have been possible. Additionally, the authors gratefully acknowledge the helpful comments of Benjamin Edelman, Anthony Goldbloom, Mitchell Weiss, and especially Susan Athey, as well as the support of Yelp, the National Science Foundation (grants CCF-1216095, DGE-1144152, and SES-1459912), the Harvard Milton Fund, the Taubman Center for State and Local Government, the Rappaport Institute for Greater Boston, and the Wu Fund for Big Data Analysis.

[1] Earlier research has provided evidence suggesting that Yelp text could be used to make inspections more efficient (Kang et al., 2013); however, prior to our work, (to our knowledge) this insight had not been incorporated into city inspection processes.

ciding whether to use a tournament to develop a product—in our application context, a predictive algorithm. The tournaments we study have two essential features: (1) they are open to all, and (2) they task all participants with the same goal.[2]

As an alternative to running a tournament, the government can choose to contract with a consultancy that will receive fixed compensation as long it produces a product above some specified quality level. We constrain both the consulting and the tournament contract to have simple contracting structures, in line with historical norms and the non-verifiable nature of innovation quality.

The city chooses an option that maximizes the expected value of $V(q) - \text{Cost}$, where q is the produced quality level. All actors are assumed to be risk neutral. Consulting companies compete for government contracts, and earn no expected rents in equilibrium; hence, they deliver the lowest-cost means of achieving any fixed level of quality. In a tournament, workers will enter to the point at which their expected returns equal their opportunity costs of time.

Each worker is of either *high* or *low* skill. High-skilled workers and low-skilled workers respectively have opportunity costs of time equal to $\overline{w}$ and $\underline{w}$. When performing the task, high-skilled workers offer a minimum quality level of $\overline{q}$ and low-skilled workers offer a minimum quality level of $\underline{q}$. With probability φ either type of worker can achieve a "breakthrough" that increases output quality to $q_{\max}$, which is greater than $\overline{q}$.[3]

In a consulting contract, there are three plausible values for the minimum quality level: $\underline{q}$, $\overline{q}$, and $q_{\max}$. As consulting companies offer to fulfill the contract at its expected cost, the contract that specifies $\underline{q}$ will cost $\underline{w}$; the contract that specifies $\overline{q}$ will cost $\overline{q}$; and the contract that specifies $q_{\max}$ will cost $\frac{\underline{w}}{\varphi}$. We assume that $\frac{\underline{w}}{\varphi} > \overline{w}$. If $V(\overline{q}) - V(\underline{q}) > \frac{\overline{w}-\underline{w}}{1-\varphi}$ and $\frac{\frac{\underline{w}}{\varphi}-\overline{w}}{1-\varphi} > V(q_{\max}) - V(\overline{q})$, then the city will prefer the high-quality consulting contract to either the low-quality contract or the maximum-quality contract.

If the city runs a tournament, then it posts reward value R that is granted to the participant who delivers the highest-quality project.[4] A tournament with reward R attracts $N = N(R)$ participants. In theory, the participants may be all high-skilled workers (in which case $N = \frac{R}{\overline{w}}$ and $\underline{w} > \overline{w} - \overline{w}(1-\varphi)^{\frac{R}{\overline{w}}}$) or all low-skilled workers (in which case $N = \frac{R}{\underline{w}}$ and $\overline{w} - R(1-\varphi)^{\frac{R}{\underline{w}}} > \underline{w}$) or a mixture of both types (in which case $\frac{R(1-(1-\varphi)^N)}{N} + \frac{R(1-\varphi)^N}{N_H} = \overline{w}$ and $\frac{R(1-(1-\varphi)^N)}{N} = \underline{w}$).

We focus on tournaments that only attract low-skilled workers; such tournaments arise when wage inequality is large. In this case, the tournament sponsor chooses R to maximize

$$(1-\varphi)^{\frac{R}{\underline{w}}} V(\underline{q}) + \left(1 - (1-\varphi)^{\frac{R}{\underline{w}}}\right) V(q_{\max}) - R.$$

We consider the tradeoff between a consulting contract and a tournament that both cost the same amount.

If the consultancy and the tournament both pay $\overline{w}$, and wage inequality is high, the tournament attracts $\frac{\overline{w}}{\underline{w}}$ unskilled workers. In this case, the tournament dominates the consultancy if and only if $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}} - 1 > \frac{V(\overline{q})-V(\underline{q})}{V(q_{\max})-V(\overline{q})}$.[5] We then find:

[2] Thus, our tournaments roughly correspond the settings of most online tournament platforms, as well as tournaments used by companies like Netflix and researchers in fields like computational biology. Such tournaments are sometimes modeled as all-pay auctions, but the prior work has focused on questions of optimal mechanism/prize design, rather than on the types of problems best solved via contests (see, e.g., Che and Gale (2003); Siegel (2009)). By contrast, our tournaments are quite different from those historically modeled by economists (Lazear and Rosen, 1981), in which workers compete to contribute to a firm's productivity by performing non-identical tasks and being promoted within the firm based on their relative achievement.

[3] Our risk assumption serves to make tournament more attractive—the case for tournaments relies on the existence of workers with some upside potential who have low opportunity costs of time.

[4] If multiple workers "win" the tournament by delivering the same highest level of quality, then they split the reward.

[5] If the consulting contract pays $\underline{w}$, then a tournament that pays the same amount yields the same result as the consultancy, as both draw only one low-skilled

PROPOSITION 1: *There exists a value of φ, denoted φ^*, at which the returns to the tournament are the same as the returns to the consulting contract. For values of $\varphi > \varphi^*$, the tournament dominates the consultancy and for values of $\varphi < \varphi^*$, the consultancy dominates the tournament. The value of φ^* increases in $V(\overline{q})$ and decreases with $V(\underline{q})$, $V(q_{\max})$, and $\frac{\overline{w}}{\underline{w}}$.*

Proposition 1 tells us that tournaments make sense when the probability of a breakthrough, φ, is relatively high. The range of values which make tournaments attractive increases with $V(\underline{q})$, $V(q_{\max})$, and $\frac{\overline{w}}{\underline{w}}$. That is, tournaments are more appealing when the baseline low-skilled outcome is not that bad, and when the best-possible outcome is particularly good. Wage inequality also makes tournaments more appealing, as tournaments attract workers with a particularly low opportunity cost of time. When cities want to ensure that they achieve at least the middle outcome, $\overline{q}$, tournaments are less attractive.

Our model suggests that the appeal of tournaments depends on wage inequality and the public sector tolerance for risk. Tournaments thus may be particularly attractive in the 21st century, because there are now many information technology workers, particularly in developing countries, with relatively low opportunity costs of time. The second factor that drives the appeal of tournaments is the public tolerance for risk. When running a tournament, the city must be willing to trade a reduced chance of getting a middling outcome for an increased probability of getting an outcome in the upper and lower tails.

II. A Restaurant Hygiene Prediction Tournament

Building on preliminary evidence of Kang et al. (2013), we partnered with the City of Boston, Yelp, and DrivenData to run an open tournament to source algorithms for predicting restaurant hygiene and sanitation violations from Yelp reviews. Participants in the tournament had twelve weeks to develop algorithms for predicting hygiene violations from Yelp data. While developing their algorithms, participants had access to a dataset recording 34,879 City of Boston hygiene inspections, dating back to April 2006, and a linked set of Yelp.com reviews, ratings, and business attributes for Boston restaurants recorded over the same time period. In Phase I ("Development Phase"), participants developed predictive algorithms based on historical data. During this phase, participants could share their predictive performance publicly on the DrivenData website, which ranked the highest performers to date. Over seven-hundred people registered for the tournament. Fifty-five competitors completed the Development phase, submitting a total of 449 sets of predictions. At the end of the Development Phase, participants submitted "final algorithms" for evaluation. Twenty-three competitors, submitted a total of 36 separate final algorithms. In Phase II ("Evaluation Phase"), the final algorithms were evaluated according to their effectiveness in predicting the outcomes of inspections conducted in a six-week test period that started after final algorithm submission. Algorithm performance was measured by root mean squared logarithmic error (RMSLE).[6] The winning algorithm's designer received $3,000; the second- and third-place algorithms' designers each received $1,000; prize money was provided by Yelp.

The winner of the tournament was a data scientist based in the United Kingdom; her submission used the statistical program R to implement an average of pre-

worker. If the consulting contract pays $\frac{w}{\varphi}$, then the tournament is clearly dominated as it does not guarantee maximal quality, while the consulting contract does.

[6] To compute this scoring function, we first collapsed the prediction for each restaurant i into a unidimensional prediction $\hat{Y}_i$, by weighting the number of minor (1x), major (2x), and severe (5x) violations. We also computed the actual (weighted) numbers of violations Y_i that were found during the test period. The performance metric was

$$\text{RMSLE} = \sqrt{\frac{1}{\#_{\text{Restaurants}}} \sum_i \Big(\log(\hat{Y}_i + 1) - \log(Y_i + 1)\Big)^2}.$$

dictions from a random forest model and gradient-boosted model with decision trees. The runner-up was a Ph.D. student in Marketing Analytics from the Netherlands; his submission used Python to implement a random forest model. Just as our theory model predicts, the tournament was effective in large part because it sourced contributions from a broad array of participants.

While the tournament itself scored submissions according to a standard prediction performance metric—RMSLE—the outcome for the City of Boston is best measured in terms of productivity. That is: *How much can using predictive algorithms improve inspector allocation?*

During the testing period, the City inspected 364 restaurants, uncovering 1,593 minor violations, 153 major violations, and 341 severe violations, for a total of 3,604 total weighted violations. We predict that if the City had used the algorithms to prioritize 364 restaurants to inspect from the universe of restaurants available in the data, it would have found 5,406 weighted violations (4,756 using the runner-up)—50% more violations than were found using the baseline inspector allocation system (32% more using the runner-up).[7] Thus, we estimate that the City of Boston would be 30%-50% more productive using a top-performing algorithm from the tournament. We are currently testing the winning algorithms' efficacy in practice, using a field experiment that integrates the winning algorithms into Boston's process for allocating inspectors.

III. Designing Prediction Problems and Tournaments

In this section, we draw upon our theory work and experience in Boston to offer general advice cities should consider when designing prediction tournaments.

[7]Alternatively, had the city used the winning algorithm to prioritize restaurants for inspection, it could have inspected only 219 restaurants (249 using the runner-up), reducing the number of inspections by 40% (32%) while identifying the same number of weighted violations and risks.

A. Problem Selection and Setup

Prediction tournaments are most effective for solving well-defined prediction problems for which large data sets are available (either to the tournament organizer, or through external sources). For the case of hygiene prediction, for example, deciding which restaurants to inspect directly involves an element of prediction, yet inspections had not incorporated systematic predictive efforts in the past. Moreover, hygiene prediction provided us with an opportunity to incorporate new digital data sources (Yelp reviews), offering at least the possibility of significant improvements in predictive accuracy.

Cities can incorporate algorithms into many of their operational processes that involve prediction. Because algorithms are extremely literal and do not make implicit tradeoffs the way that a policymaker would, cities should be explicit about all of their tournament objectives, keeping in mind both intended and unintended consequences. A simple example from the hygiene prediction context is the need for specifying tradeoffs between minor, major, and severe violations. Through conversations with stakeholders, a complete set of design objectives can be identified and then formally integrated into the tournament scoring function (for further discussion, see Luca, Kleinberg and Mullainathan (2016)).

B. Choosing Data

Finding the relevant data for a prediction tournament requires a systematic approach to determining the value of different data sets, as well as collaboration with partner organizations. Beginning with internal data is sensible, but that data may lack critical information or be low-frequency. In the hygiene prediction competition, prior violations provided a good signal about future violations, but the frequency and scope of the Yelp data allowed for finer predictions.

C. Incentives and Information

Competitors are driven by a variety of motives, from prize money to job market signaling to just the opportunity to work on interesting and important problems; cities have considerable flexibility in leveraging all of these incentives. For example, city governments have the ability to generate publicity, partially because they have members (like the Mayor) who are covered regularly by the media. Tournament designers must decide how and at which stages to use publicity—upfront media attention can serve to inform prospective participants and generate competition, while ex post publicity serves to increase the signaling value of winning.

D. Choosing a Platform

As our theory indicates, the talent pool largely determines the success of the tournament. At this point, the main prediction tournament platforms have developed their own distinct user followings. On larger platforms such as Kaggle, there are more than 450,000 registered data scientists who all receive an email when each competition begins. A tournament's designer should look for a platform that has run similar tournaments in the past, and should look at prior tournaments' outcome statistics to get estimates of participation and expected performance.

E. Measuring Success

It is essential to choose metrics for evaluating both (1) tournament participants and (2) the overall value of the tournament itself. Whenever possible, it is optimal to score entries according to information collected after the close of submissions—as we did in the Evaluation Phase—so as to have a true out-of-sample test.[8] And of course, tournaments are not free, and thus should only be used if they deliver value that exceeds costs. As we mentioned in Section II, we have evaluated (and are continuing to evaluate) the value of the prediction tournament itself by asking how much the winning algorithms can improve the allocation of inspector time in Boston in practice.

[8] Additionally, it is important to figure out a clear way of communicating the scoring system to tournament participants.

IV. Conclusion

Open tournaments are a new and exciting tool for leveraging latent, low-cost talent to solve cities' problems. However, tournaments are not a panacea, as they have downside risk and involve duplication of effort. Tournaments are thus most effective when (1) the organizers are comfortable with project risk and (2) an abundant supply of low-cost labor is available. In the case of Boston, the needed conditions were met, and our hygiene prediction tournament successfully sourced algorithms that can improve inspector allocation.

REFERENCES

Che, Yeon-Koo, and Ian Gale. 2003. "Optimal design of research contests." *American Economic Review*, 93(3): 646–671.

Kang, Jun Seok, Polina Kuznetsova, Michael Luca, and Yejin Choi. 2013. "Where Not to Eat? Improving Public Policy by Predicting Hygiene Inspections Using Online Reviews." In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. 1443–1448.

Lazear, Edward P., and Sherwin Rosen. 1981. "Rank-Order Tournaments as Optimum Labor Contracts." *The Journal of Political Economy*, 89(5): 841–864.

Luca, Michael, Jon Kleinberg, and Sendhil Mullainathan. 2016. "Algorithms Need Managers, Too." *Harvard Business Review*, 94: 96–101.

Siegel, Ron. 2009. "All-Pay Contests." *Econometrica*, 77(1): 71–92.

Appendix – Proof of Proposition 1

The value of $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}}-1$ is monotonically increasing in φ and goes from 0 to ∞ as φ goes from 0 to 1. Hence, there must exist a value of φ at which $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}}-1$ equals $\frac{V(\overline{q})-V(\underline{q})}{V(q_{\max})-V(\overline{q})}$, a constant. The value of $\frac{V(\overline{q})-V(\underline{q})}{V(q_{\max})-V(\overline{q})}$ is rising with $V(\overline{q})$ and falling with $V(\underline{q})$ and $V(q_{\max})$; hence, φ^* is rising with $V(\overline{q})$ and falling with $V(\underline{q})$ and $V(q_{\max})$. For a given φ, the value of $(1-\varphi)^{1-\frac{\overline{w}}{\underline{w}}}-1$ is rising with $\frac{\overline{w}}{\underline{w}}$; hence, φ^* must be falling with $\frac{\overline{w}}{\underline{w}}$.