# Research Statement

Peng Shi

My research vision is to develop mathematical models and techniques that can significantly benefit society. My current focus is prediction and optimization in matching markets, which include systems that match students to schools, applicants to public housing, workers to jobs, and patients who need organ transplants to donors.

Much of my PhD work is motivated by designing a better school choice system in Boston, so that students have equitable chances to go to the schools they want, the city's busing cost is controlled, and local communities remain cohesive. One of the plans I proposed was implemented across Boston in 2014. I also developed theories and methodologies to better assist such reforms in the future. In the future, I plan to study prediction and optimization in other matching markets such as public housing and two-sided matching websites. In the following, I will elaborate on the school choice problem, my research contributions, and my future plans.

## The School Choice Problem

Many cities in the US allow students to choose schools within the public system. Since 1988, Boston Public Schools (BPS) has assigned students to elementary and middle schools by dividing the city into three geographic zones. About six months before school starts, new applicants submit a ranked-order list of any number of schools in their zone, plus a one-mile radius "walk-zone." Given the submitted preferences, BPS uses a system of priorities and lotteries to determine the assignment. Similar systems are in place in Cambridge, Charlotte-Mecklenburg, Chicago, Denver, Minneapolis, Miami-Dade, New York City, New Orleans, Newark, and San Francisco.

One problem is how much choice to allow and how to prioritize between students. The trade-off is between efficiency, equity, and system costs. Choice can help students to find a better match, but is costly to the city since the city needs to provide school busing. In 2012, Boston spent about 90 million dollars a year on busing, which represented almost 10% of the entire school board budget. When limiting choice, the city has to decide how to equitably allocate among neighborhoods, which is difficult as schools differ in perceived qualities. Other possible problems include lack of predictability, because lotteries are used, and community dispersion, as students from the same neighborhood go to different schools.

In Boston, the choice system since 2006 used the Gale-Shapley Deferred Acceptance (DA) algorithm, which guarantees that students have no incentives to misreport their preferences. The algorithm allows BPS two policy levers: the subset of schools each student can rank and his/her priority at each school. To limit busing and encourage closer-to-home assignments, the city attempted to introduce more zones in 2004 and 2009, but both attempts failed due to public concern over equity of the proposed zoning.

### My Involvement

When the city appointed a committee to try again to reform the 3-zone plan in 2012, I began attending the committee meetings, which were open to the public. Through these, I dialogued

with city committee members, school board staff, concerned parents, and activists. I also began formally collaborating with BPS, which gave me access to student-level data. This led to a series of research projects, which I describe below.

## A Predictive Model for School Choice

A pressing need of the city committee was a way to predict how any given assignment plan would perform. Working with BPS and other academic researchers, I developed a multinomial logit (MNL) discrete-choice model of how students would rank schools given new sets of choice options. This allowed us to predict the outcome of any proposed plan by simulation. The metrics we examined included equity of access to quality, proximity to home, variety of choice, predictability, bus coverage area, socio-economic diversity, and community cohesion.

### Proposing a Simple Alternative Plan

Besides building the simulation engine, I also proposed my own plan. The city committee decided to measure quality of schools using a test-score-based metric. I proposed that every student can rank the union of the following sets: a) every school within a one-mile radius; b) the closest 2 top 25% schools; c) the closest 4 top 50% schools; d) the closest 6 top 75% schools; e) the closest 3 capacity schools, which were to be chosen by BPS to accommodate excess demand.

The logic was that if a student already lives near the highest quality schools, then the sets would intersect, since a top 25% school is by definition also top 50% and 75%. However, if the closest schools do not rank well according to the metric, then the set of options would expand to include the closest alternatives. A side benefit of this plan is that it can smoothly adapt to changes in schools, whereas zone boundaries are hard to change when drawn, since those on the boundary would be greatly affected. This was later called the Home-Based Plan.

### Results and Impact

We used the simulation engine to analyze the Home-Based Plan and various zone-based plans, and found that while no plan dominated another, the Home-Based Plan performed well throughout the portfolio of metrics. After seeing the result of our analysis as well as independent analyses by others, the city implemented the Home Based Plan across Boston in 2014. A detailed description of this project can be found in my paper, "Guiding School Choice Reform through Novel Application of Operations Research" (*Interfaces*, 2015), which won the 2013 INFORMS Doing Good with Good OR Student Paper Competition.

### Further Validation by Field Experiment

The implementation of the Home-Based Plan also provided a natural experiment to validate the predictive model. A concern was that if students' choices were affected more by behavioral issues such as framing than by underlying preferences, then a predictive model based on past data might no longer be valid after the reform, since the reform might have altered the framing. We found that after controlling for changes in the demographics of applicants, the predictive model based on pre-reform data performed almost as well as the predictive model based on post-reform data. We also compared the MNL model to a mixture-of-logit model, which theoretically can capture more complicated subsitution patterns, but we found little difference, possibly because of the lack of data correlating with actual substitution patterns. This validated MNL as a reasonable way

to model students' preferences. The corresponding paper, "Demand Modeling, Forecasting and Counterfactuals," is under preparation.

## Optimizing the Assignment Plan

Having arrived at a reasonable predictive model, we turn to optimization: suppose that students chose according to a given utility model, what incentive compatible assignment system would maximize a weighted sum of utilitarian welfare and max-min welfare, while staying within a given busing constraint?

### Structure of Valid Mechanisms in the Fluid Model

In "Optimal Allocation without Money: an Engineering Approach" (*Management Science*, 2015), we show that in a fluid approximation with infitessimal students, any assignment system that satisfies certain regularity conditions either looks like competitive equilibrium with virtual money, or Deferred Acceptance (DA) with a single random number for each student at all schools to break ties. The latter matches what is used in Boston since 2006.

In a follow-up paper, "Socially-Optimal Assortment Planning" (in preparation), we relax one of the regularity conditions, and so allow a richer class of mechanisms, which we show is equivalent to the class of stable-matching based mechanisms: prioritize the students in a neighborhood dependent way, and finding a stable match. The difference from above is that this allows different random numbers for each student at different schools.

### Socially-Optimal Assortment Planning

Using these structures, both papers reduce the school choice optimization to a Linear Program (LP) that can be efficiently solved by repeatedly solving the "socially-optimal assortment planning" problem: for each neighborhood, given students' utility distributions and a cost of assigning to each school, find the optimal assortment of schools to offer to maximize value to students minus costs. These costs represent opportunity costs for other students, and arise from the shadow prices of the capacity and busing constraints.

While the first paper also solves this subproblem under the MNL model, the follow-up paper solves it under many other utility models. The observation is that many of the results on revenue-maximizing assortment planning in the operations management literature can be extended to this setting, although the problems are not equivalent because we have here also consumer welfare in the objective. We give efficient algorithms for socially-optimal assortment planning under MNL utilities and matroid constraints, under d-level Nested Logit utilities and a cardinality constraint at each nest, and under a Markov-chain-based model.

### Evaluation by Simulation

In both papers, we evaluate the optimized plan by simulation in the discrete model, and find significant improvements over Home-Based in average expected utilities, minimum expected utility of a neighborhood, and probability of getting first choice, while staying within the same busing budget. The second paper also evaluates the robustness of the results to possible errors in parameters and the convergence rate to the fluid approximation.

Earlier versions of the first paper won the 2013 INFORMS Public Sector OR Best Paper Competition, and the 2014 MIT ORC Best Student Paper Competition.

## Correlated-Lottery Implementation

An orthogonal direction to optimizing assignment probabilities is to optimize how these probabilities map to actual assignments. In "Improving Community Cohesion in School Choice via Correlated-Lottery Implementation" (*Operations Research*, 2014), we measure community cohesion as proportional to the number of pairs of students from the same community going to the same school, and we study how much we can increase cohesion without changing anyone's assignment probabilities. We show that while maximizing community cohesion is NP-hard even with 2 schools, we have a heuristic that performs well in practice.

### Empirical Findings in Boston

We test our heuristic on real data from Boston and show significant improvements. In fact, the improvement under the 3-zone system from correlated-lottery alone is greater than the improvement from changing to any of the assignment plans considered by the city committee. However, the improvement is not equitably distributed across the city, with communities near the center of the city seeing little improvement. When we applied both the Home Based Plan and correlated lottery, all communities improved significantly in cohesion. The total gain is also greater than the sum of applying the two separately, being a factor of 3 improvement for grade K1 and a factor of 2 for K2.

### Theoretical Analysis in Fluid Model

To understand how the benefit of lottery-correlation depends on the primitives of the school district, we analyze a fluid approximation with infitessimal students, and show that under certain regularity conditions, the benefit from lottery-correlation can be written as a constant minus the Herfindal index of school sizes, minus the between-community and the within-community variations in assignment probabilities. This suggests that lottery-correlation matters more when school sizes are similar, and when preference heterogeneity is low. The latter explains our empirical findings because the communities near the center of Boston have greater preference heterogeneity, so lottery-correlation initially does not help much; under the Home-Based Plan, preference heterogeneity in those communities is reduced, so the benefit from lottery-correlation increases.

## Future Research

I plan to study prediction and optimization in other matching markets. One idea is to apply socially-optimal assortment planning to systems that allocate subsidized housing. Similar to school choice, such systems can decide what options to offer to each subpopulation. The difference is that the allocation is dynamic and may involve differential waiting times. I have obtained partial data on the Build-to-Order public housing system in Singapore and have ideas on how to optimize the assortment of options offered at each round to improve social welfare.

Another idea is to optimize search results in online matching platforms such as Airbnb, Upwork, and eHarmony, in order to increase platform revenue and value to customers. The difference between this and classical assortment optimization is that customers are on both sides of the market. A related application is suggesting interviews in centralized labor markets such as university recruiting. As a start, I have collected internship matching data from a Master's program at MIT, estimated demand models, and have ideas on how to optimize.

My style of research is application-driven, and I am most interested in applications that have positive societal impact.