Simulating University Application Data for Fair Matchings

Meirav Segal(✉), Anne-Marie George, and Christos Dimitrakakis

Department of Informatics, University of Oslo, Oslo, Norway
{meiravs,annemage,chridim}@ifi.uio.no

Abstract. This paper describes the design of a simulator (work in progress), that is based on Norwegian university admissions and exam data. It generates a realistic population of applicants to university programs, their preferences and study outcomes if they were admitted to the different study programs. This simulator is a versatile tool and can be used to analyse the current admission policy for Norwegian universities in terms of many fairness criteria that, e.g., take into account student preferences, gender balance, university preferences and study outcomes. More generally, it creates a benchmark for testing matching algorithms and fairness notions without revealing sensitive data.

1 Introduction

The problem of school choice, in which students are assigned to schools, is a popular research area lying in the intersection of computer science, economics and mathematics. Apart from being challenging, it has great importance due to the significant influence a school choice could have on students' future trajectories. Formally it constitutes a matching or allocation problem under preference in a bipartite graph. Algorithmic solutions are employed in many countries and similar areas, e.g., for university admissions in Hungary [5], allocation of teachers to positions in France [11] and patient-donor matches for kidneys in many countries [4]. These solutions often involve stable allocations based on students and schools preferences, capacities and other constraints enforcing formal requirements or some fairness towards subgroups. Here, stability means that no single deviation from the computed allocation is more beneficial for any party involved [6].

When designing new algorithms and methods for school choice problems, there is an obvious need to evaluate its performance in practice, preferably using real-world data. While stable allocations consider candidates' preferences, other methods might take into account future study outcomes such as dropouts or grades. Nevertheless, real-life data cannot provide outcomes for students that never participated in a study program. Thus, there is a need for a simulator that generates realistic application data and provides study outcomes for any possible allocation of students to study programs.

For example, a recent study evaluated how different policies affect dropouts in the Chilean centralized college admission system using a simulator based on

E. Zouganeli et al. (Eds.): NAIS 2022, CCIS 1650, pp. 132–138, 2022.
https://doi.org/10.1007/978-3-031-17030-0_11

real data [7]. As the simulator itself was not published, the research community cannot generate new samples to explore other questions.

We describe the planned design of a simulator based on data of the Norwegian university admission system. This data is not openly available because it contains sensitive information, but a simulator can provide reliable data for analysis while preserving privacy. This will constitute a valuable benchmark for the research community. The simulator will generate a set of applicants with demographic features (e.g. age, gender, county), educational background (e.g. high school points), their preferences over study programs and study outcomes for each of these programs. Using these attributes, decision makers can evaluate new policies. For example, the current admission system grants bonus points based on age and gender. Through simulation, we can compare students' outcomes according to assignments given by the current system, with outcomes according to assignments based on a new policy, with increased or decreased bonus points.

2 University Admissions in Norway

In Norway, the admission process for most undergraduate study programs at all public academic institutions is coordinated by the Norwegian Universities and Colleges Admission Service in a centralized manner [2]. This section describes the admission process and the available data for applications and study outcomes.

2.1 Admission Process

Candidates rank 10 study programs they wish to attend. Further, university programs specify their preferences over students by a point scheme based on grades and other factors such as age, gender or military service. In addition, candidates can apply through different quotas. For example, the first-time diplomas quota is designated for candidates who have completed and passed upper secondary school in normal time and are at most 21 years old. Other quotas are intended for underrepresented groups in specific programs. All candidates who do not fit special quotas, apply through the ordinary quota.[1] An applicant is classified as 'qualified' for a study program when they meet its minimum requirements.

In the main admission process, a specialized stable marriage algorithm is applied in order to find the candidate-optimal stable matching based on the applicants' and university programs' preferences [9]. At this point, each candidate is given at most one offer, to the highest ranked program that the candidate is qualified for (while maintaining stability).

After the candidates have accepted or declined the offers, study programs with remaining vacancies continue to make offers to available students over a period of one month in order of their preferences over the applicants.

[1] For more details of the point system and quotas see https://www.samordnaopptak.no/info/.

2.2 Data

Through the Norwegian Database for Statistics on Higher Education [1], we have been granted access to two data sets: Applications and Exams.

Applications. Application data[2] of all applications to all Norwegian university programs in the period 2017–2020. This data set includes $2,265,418$ applications of $\sim 500,000$ candidates to over $2,000$ study programs of 34 academic institutions. In each year approximately $180,000$ candidates apply, from which 50% are admitted.[3] Every application includes the following features:

- *Candidate features*: identifier, age, gender, citizenship, country of educational background, high school grades in the form of GPA and summarised language/science points, 'other points' (for other factors such as age or gender), registered municipality of residence (for applications made in 2020).
- *Program features*: identifier, department, university.
- *Application features*: year and semester of application, and quota the application is considered in.
- *Candidate preference*: Preference for the program (a number between 1–10).
- *Admission decision*: Study offer and acceptance.

Exams. Exam data[4] of all students at Norwegian universities for all their taken exams in the period 2017–2020. The exam data includes $5,321,519$ records of exams taken by students, with an average of 8 exam grades per student. For each year there are grades of approximately $30,000$ courses throughout the different study programs. More specifically, we consider the following entries:

- *Student identifier* (matched with entry in application table).
- *Program features*: identifier, department, university.
- *Exam features*: course identifier and number of credits, year and semester of exam, indication whether the student is retaking the exam.
- *Study outcome*: grade ('A'-'F' or Pass/Fail), or indication of non-attendance.

3 Simulator

In this section we describe the (planned) components of the simulator individually. Figure 1a presents the process of generating a new population given the trained components. First, we generate background attributes of candidates. In addition, we generate the candidate's underlying type. This type determines the preference profile, which together with background features sets the priorities over programs. The outcome profile and outcomes over programs are determined similarly, but also affected by the preferences. Before the release of the complete simulator, we will incorporate differential privacy throughout the pipeline.

[2] https://dbh.hkdir.no/dbh-old/dokumentasjon/tabell.action?tabellId=379.

[3] Note that about 30% of the applications are of local admission, which means that the acceptance offers are made by each institution individually and not as part of the centralised process. Local admission is performed for master's programs or for special programs in which admission is based on additional criteria such as interviews.

[4] https://dbh.hkdir.no/dbh-old/dokumentasjon/tabell.action?tabellId=472.

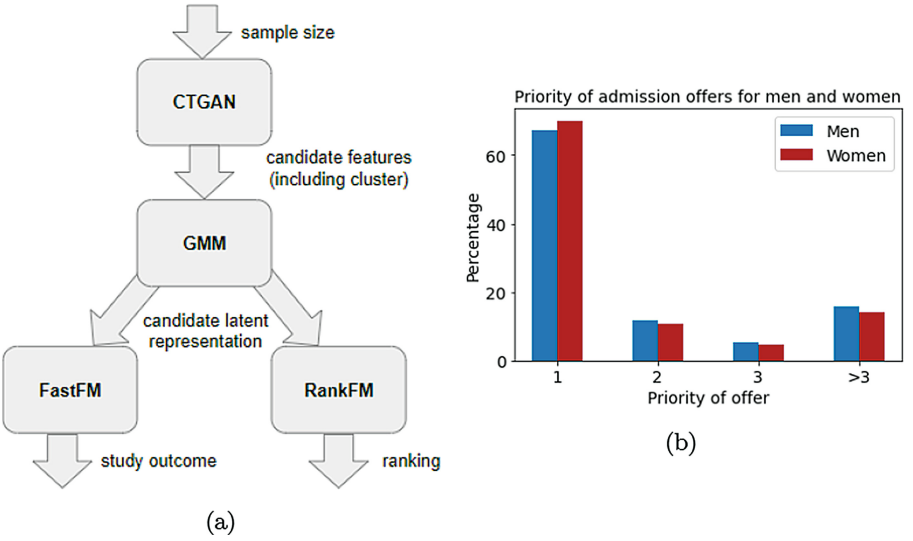

Fig. 1. (a) Simulator pipeline diagram. (b) A possible analysis to perform on the simulated data. The priority of admission offers made according to gender, using original data with 0.01-differential privacy using the Laplace mechanism.

3.1 Prepossessing and Training

We provide details of how the selected models are trained from the bottom up:

RankFM. We train rankFM[5], a factorisation machine model designed for ranked data with a loss function based on pairwise comparisons [10], to predict candidates' preferences over study programs. This model considers implicit data: a pairwise comparison is performed between programs ranked by the candidate and programs not ranked by that candidate, such that the latter are considered to have a lower priority. The comparison between ranked programs is not performed explicitly and is only addressed by giving larger confidence weights to higher ranked programs. Notably, rankFM allows us to incorporate candidates' features and study programs' features, such that their relation to candidates' preferences over programs is not lost. This model provides latent representations for candidates and study programs that, when combined, give a preference value for every student and university program pair.

FastFM. We train fastFM [3], a factorisation machine model with root-mean-square error for explicit feedback, to predict the students' study outcomes. Here, the candidates' features and the study programs' features include the latent representation obtained from rankFM. The features also include the preferences. The outcomes may be defined as average first year grade, normalised in $[0, 1]$. New latent representations of candidates and programs are provided by fastFM.

[5] https://github.com/etlundquist/rankfm.

Gaussian Mixture Model (GMM). A Gaussian Mixture Model is fitted to the concatenated latent representations of the candidates. This model allows us to sample new latent representations given a Gaussian identifier.

Conditional Tabular GAN (CTGAN). CTGAN[6] [12] is a deep learning based synthetic data generator for tabular data, that can learn from real data and generate synthetic clones with high fidelity. The CTGAN generator is trained using the candidates' feature data, including a GMM cluster identifier, which allows to generate candidate populations with similar distributions of features.

3.2 Generating Student Features, Preferences and Outcomes

To generate a new population, we can now follow Fig. 1a from top to bottom. We generate individual features for a new population of a given size using CTGAN. These features include demographic attributes such as gender and citizenship, but also the GMM cluster identifier. CTGAN is designed to generate new samples based on the train data distribution, so we expect the generated candidates to have a cluster identifier that fits their other features. Then, for each generated candidate we sample the specific pretrained Gaussian according to the their GMM cluster identifiers. As a result, we get the latent representations which holds information regarding the preferences and outcomes of candidates. Using the precalculated latent representations for study programs, we can predict the ranking and outcome of the study programs for each generated candidate.

3.3 Simulating Admission Decisions (Work in Progress)

Given the preferences of candidates and study programs, we can run the Gale-Shapley matching algorithm, a variation of Stable Marriage Matching for the hospitals-residents problem [8], to simulate the current admission system in Norway. The output will simulate the offers made in the first admission round. Given an initial offer, the candidate may decide to decline the offer. To simulate the second phase of acceptance, we simulate offers to applicants for programs in order of the programs' preferences (point scheme). We will use a classifier to predict offer acceptance by students for both first and second phase study offers.

Note that for this simulation the programs point schemes as well as their capacity has to be known. Neither are provided in the data, but can be deduced by the properties of the procedure of admissions in Norway. If a candidate has been accepted by a program (independent of whether they accept the offer and in which phase the offer was made), then

1. any other qualified applicant that was not given a study offer must have a worse point score for the program, and
2. if there exist such an applicant as in (a), then the capacity of the program is equal to the number of students that accepted the study offer.

[6] https://github.com/sdv-dev/CTGAN.

By these observations, we gain pairwise comparisons between (qualified) candidates point scores for the different programs. We can then find program point schemes that are linear functions or polynomials over the candidate features that satisfy these relations. The capacities are either determined by (b) or can simply be assumed to be the number of students that accepted the study offer.

4 Fair Matchings

The simulator, if implemented as described in Sect. 3, can be used to generate realistic instances of hospital/residents or school choice problems on which algorithmic solutions can be tested.

Fairness is particularly relevant to centralised school choice mechanisms and can be analysed for different solutions. We do not propose here a specific measure of fairness, but rather facilitate the analysis of different fairness notions. Apart from the usual notion of stability which only relies on preferences of candidates and programs, one can consider more elaborate objectives, such as equal preference satisfaction across groups based on gender or other demographic attributes. For example, Fig. 1b shows the satisfaction difference between men and women for the current admission system (real data). We can see that the percent of women who are offered admission to their first priority is higher than the equivalent percent of men. Yet, it is reversed for lower priorities. A possible explanation would be that women place 'safer' choices as their top priorities. Additional analysis could include satisfaction differences among counties or age groups, admission differences and outcome differences.

Furthermore, the possibility to predict study outcomes opens up the possibility to find allocations that offer equal predicted study success across groups. As the point scoring system of university programs is intended to rank the candidates by their capability of studying, it would be interesting to consider how much the point scheme correlates with the predicted study success of the students. One can measure how different a matching based on predicted study success instead of point schemes for university program preferences would be.

5 Conclusion

The simulator presented here is planned to use a combination of factorisation machines and Gaussian mixture models to provide a real-world-based benchmark in a countrywide scale. Using this simulated data, one could measure welfare and fairness not only with respect to students' and university's preferences, but also with respect to their outcomes. We believe this simulator has the potential to advance the research efforts in school choice and illuminate new interesting problems that exist in current school assignment systems.

Acknowledgements. This work was supported by the Research Council of Norway under project number 302203. We are thankful for the data provided by the Norwegian Directorate for Higher Education and Skills.

References

1. Homepage: Database for statistics on higher education (database for statistikk om høyere utdanning). https://dbh.hkdir.no/. Accessed 01 May 2022
2. Homepage: Norwegian universities and colleges admission service. https://www.samordnaopptak.no/info/english/. Accessed 27 Apr 2022
3. Bayer, I.: fastFM: a library for factorization machines. J. Mach. Learn. Res. **17**(1), 6393–6397 (2016)
4. Biró, P., et al.: First handbook of the cost action CA15210: European network for collaboration on kidney exchange programmes (ENCKEP). European Cooperation in Science and Technology, Brussels (2017)
5. Biró, P.: University admission practices – Hungary, MiP country profile 5 (2011). https://www.matching-in-practice.eu/higher-education-in-hungary/. Accessed 28 Apr 2022
6. Gale, D., Shapley, L.S.: College admissions and the stability of marriage. Am. Math. Monthly **69**(1), 9–15 (1962). http://www.jstor.org/stable/2312726
7. Larroucau, T., Rios, I.: Dynamic college admissions and the determinants of students' college retention. Technical report 2020 and, Do "Short-List" Students Report Truthfully (2020)
8. Manlove, D.: Algorithmics of Matching Under Preferences. Series on theoretical computer science. World Scientific (2013). https://books.google.no/books?id=7wGJMAEACAAJ
9. Samordna opptak: Wikipedia article: Norwegian universities and colleges admission service. https://en.wikipedia.org/wiki/Norwegian_Universities_and_Colleges_Admission_Service. Accessed 27 Apr 2022
10. Rendle, S., Freudenthaler, C.: Improving pairwise learning for item recommendation from implicit feedback. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 273–282 (2014)
11. Terrier, C.: Matching practices for secondary public school teachers – France, MiP country profile 20 (2014). https://www.matching-in-practice.eu/matching-practices-of-teachers-to-schools-france/. Accessed 28 Apr 2022
12. Xu, L., Skoularidou, M., Cuesta-Infante, A., Veeramachaneni, K.: Modeling tabular data using conditional GAN. In: Advances in Neural Information Processing Systems, vol. 32 (2019)