Data Documentation for “The Large Core of College Admission Markets: Theory and Evidence”

This replication package relies on three primary data sources. Below is a summary of each source and instructions for accessing them:

Hungarian College Admissions Administrative Data

This administrative dataset contains detailed information on the bachelor’s degree admissions process in Hungary. It includes each applicant’s rank-order list (ROL), priority scores for each contract, academic performance indicators (e.g., high school grades and matriculation exam scores), and additional applicant characteristics such as gender, postal code, and high-school identifier. Detailed data on the realized priority-score cutoffs for various program–financial terms pairs are also provided.

More detailed description: https://adatbank.krtk.mta.hu/en/adatbazisok/elerheto-adatbazisok/

(> Education >> Application and admission to higher education (FELVI))

Access: https://adatbank.krtk.mta.hu/adatbazisok/adatkeresek-menete/

This dataset includes 4 files:

1. *Institutional data*: raw_data/institutional_data_2007
2. *Admission data*: raw_data/application_2007.dta
3. *Individual data*: raw_data/individual_2007.dta
4. *Capacity data*: raw_data/Capacities2007_Official.xls

Information in the Capacities2007_Official.xls has been hand-collected from the booklet titles: *Felsőoktatási felvételi tájékoztató 2007. Dr. Bakos Károly, Cseh Boglárka, Bakonyi László, Dr. Rádli Katalin, Budapest, 2006.*

If you wish to access it, please contact the corresponding author: Sandor Sovago (sovago.sandor@gmail.com)

T-STAR Dataset from the Hungarian Central Statistics Office

This dataset provides settlement-level annual information on collected income taxes and per-capita gross annual incomes. The T-STAR data were merged with the admissions dataset based on settlement identifiers to capture local economic conditions.

More detailed description: https://adatbank.krtk.mta.hu/en/adatbazisok/elerheto-adatbazisok/

(> Regional statistics and census tracts >> Regional statistics (T-STAR))

Access: https://adatbank.krtk.mta.hu/adatbazisok/adatkeresek-menete/

This dataset includes 1 file:

1. *Income data*: raw_data/tx.dta

National Assessment of Basic Competencies (NABC) Data

The NABC dataset includes survey information used to construct a standardized SES index (similar to the OECD ESCS indicator) based on parental education, home possessions, and parental labor-market status. Although the data were originally collected starting in 2008, aggregated high-school averages (calculated over 2008–2012) have been merged with the main administrative dataset.

More detailed description: https://adatbank.krtk.mta.hu/en/adatbazisok/elerheto-adatbazisok/

(> Education >> National Assessment of Basic Competencies (NABC))

Access: https://adatbank.krtk.mta.hu/adatbazisok/adatkeresek-menete/

This dataset includes 5 files:

1. *NABC data*:
    - raw_data/2008/2008_tanuloi_10.dta
    - raw_data/2009/2009_tanuloi_10.dta
    - raw_data/2010/2010_tanuloi_10.dta
    - raw_data/2011/2011_tanuloi_10.dta
    - raw_data/2012/2012_tanuloi_10.dta

Data

This document lists the datasets that are required for replicating the results (Tables and Figures) presented in the paper. These tables can be produced by executing the 0__Master_prepare_data.do file in the code\prepare_data folder.

For each dataset, we list the definition of the dataset (definition), the specific .do or .py file producing the dataset (file), the input data needed (input data), the list of tables/figures that build on this dataset (usage), and the list of variables and explanations that are included in these datasets (variables).

Stata version: 18.5
Python version: 3.12.2

StudentROLs.dta

Definition: Student-level data including information on demographics and ROL characteristics
File: 4_PrepareROLLevelData.do
Input data:

- Institutions_raw
- Application_raw

Usage:

- Table 1
- Table C4
- Table C6

Variables:

- id: applicant id
- mp: Same study program consecutively (anywhere on ROL)
- mp1: Same study program consecutively on the top of the ROL
- Pref_for_location: Preference for program characteristic: Single program location
- Pref_for_uni: Preference for program characteristic: Single university
- Pref_for_faculty: Preference for program characteristic: Single faculty
- Pref_for_field: Preference for program characteristic: Single field of study
- Pref_for_major: Preference for program characteristic: Single major

Applications.dta

Definition: Data including applicants ROLs in a long format
File: 5_PrepareApplicationsData.do

Input data:

- institutions_raw
- applicants_raw
- application_raw
- data_capacity
- ses_2008_2012
- settlement_level_income

Usage:

- Table C1
- Table C7
- Run_matching.py

Variables:

- id: applicant id
- rank: Rank on ROL
- program_id: Program id
- contract_id: Contract id
- funding: State-funded (1) or self-funded (2) contract
- priorityscore: Priority score
- admitted: Admitted to the contract (dummy)
- gender: Female
- tazon: settlement ID
- mkod2002: NUTS 3 code (2002)
- teltip8: type of settlement according to 2006 census
- omid: High-school identifier
- schooltype: type of secondary school class (3 categories)
- GPA11_3: GPA (0--5, Hungarian, math, history)
- budapest: Budapest
- county_town: County town
- town: Town
- village: Village
- settlement_missing: Settlement missing
- ses: ses
- grossincome: Gross income (1000USD, 2007 prices), settlement-level
- rollength: Length of ROL (contracts, admin)
- karkod: institution/faculty ID
- szaknev: name of program/program pair
- program_level: level of program
- munkarend: munkarend type of instruction
- rank_of_admision: ID of application where applicant was admitted
- field_study: Field of study
- szekhely: location of institution headquarter
- szakkod: ID of program/program pair

- kepztelep: location of program
- priorityscorecutoff: priority score cutoff
- applicants: number of applicants
- admitted_applicants: number of admitted students
- capacity_min: capacity (minimum)
- capacity: Capacity (tidy)
- no_capacity_in_booklet: The contract does not show up in the booklet
- no_official_capacity: No official capacity
- capacity_official: Official capacity
- capacity_class: Field-level funded capacity code
- capacity_realized: Funded students
- capacity_booklet: Field-level fnded capacity in the booklet
- capacity_total: Total capacities per program
- rejected: Rejected by the contract (dummy)
- disadv2: Disadvantaged student (severe)
- disadv1: Disadvantaged student (normal)
- rejected_stududents: Number of rejected students
- program_county: Program county
- dual: Dual program

Matching_under_Heuristics_SR.dta

Definition: Applicant-level data that includes various allocations (incl. Realized, benchmark, preference flip, greedy reject, combined algorithm, and SP-DA)
File: 7_PrepareMatchingData.do
Input data:

- stable_matchings_abc_remove_11.csv
- Student.dta
- Applications.dta

Usage:

- Table 2
- Table 3
- Table C2
- Table D2
- Table D3
- Table D4
- Table E1

Variables:

- id: Applicant id
- contract_id_real: Contract id (realized)
- contract_id_orig: Contract id (benchmark)

- contract_id_h1: Contract id (Greedy reject algo.)
- contract_id_h2: Contract id (Preference flip algo.)
- contract_id_h3: Contract id (Combined algo.)
- contract_id_h4: Contract id (SP-DA)
- program_id_orig: Program id (benchmark)
- program_id_h1: Program id (Greedy reject algo.)
- program_id_h2: Program id (Preference flip algo.)
- program_id_h3: Program id (Combined algo.)
- program_id_h4: Program id (SP-DA)
- program_county_orig: Program county (benchmark)
- program_county_h1: Program county (Greedy reject algo.)
- program_county_h2: Program county (Preference flip algo.)
- program_county_h3: Program county (Combined algo.)
- program_county_h4: Program county (SP-DA)
- schedule_orig: Schedule (benchmark)
- schedule_h1: Schedule (Greedy reject algo.)
- schedule_h2: Schedule (Preference flip algo.)
- schedule_h3: Schedule (Combined algo.)
- schedule_h4: Schedule (SP-DA)
- funding_real: State-funded (1) or self-funded (2) contract (realized, 2007)
- funding_orig: State-funded (1) or self-funded (2) contract (benchmark)
- funding_h1: State-funded (1) or self-funded (2) contract (Greedy reject algo.)
- funding_h2: State-funded (1) or self-funded (2) contract (Preference flip algo.)
- funding_h3: State-funded (1) or self-funded (2) contract (Combined algo.)
- funding_h4: State-funded (1) or self-funded (2) contract (SP-DA)
- fulltime_orig: Matched to full-time program (benchmark)
- fulltime_h1: Matched to full-time program (Greedy reject algo.)
- fulltime_h2: Matched to full-time program (Preference flip algo.)
- fulltime_h3: Matched to full-time program (Combined algo.)
- fulltime_h4: Matched to full-time program (SP-DA)
- rank_of_admission_orig: Rank of admission (benchmark)
- rank_of_admission_h1: Rank of admission (Greedy reject algo.)
- rank_of_admission_h2: Rank of admission (Preference flip algo.)
- rank_of_admission_h3: Rank of admission (Combined algo.)
- rank_of_admission_h4: Rank of admission (SP-DA)
- winner_h1: Winner (Greedy reject algo.) relative to benchmark
- winner_h2: Winner (Preference flip algo.) relative to benchmark
- winner_h3: Winner (Combined algo.) relative to benchmark
- loser_h1: Loser (Greedy reject algo.) relative to benchmark
- loser_h2: Loser (Preference flip algo.) relative to benchmark
- loser_h3: Loser (Combined algo.) relative to benchmark

Student.dta

Definition: Student-level data including information on demographics and ROL characteristics
Input data: replication_package/data/Applications.dta
File: 8_PrepareStudentLevelData.do
Usage:

- Table 1
- Table 3
- Table C3
- Table C4
- Table C5
- Table C6
- Table D3
- Table D4
- Table E1

Variables:

- budapest: Capital
- county_town: County capital
- disadv: Disadvantaged
- funded_only: State-funded contract exclusively
- gender: Female
- GPA11_3: 11th-grade GPA (standardized)
- grossincome: Per-capita annual gross income (1000 USD, 2007 prices)
- mkod2002: County code
- Program_number: Number of programs on ROL (observed)
- rollength: Number of alternatives in ROL
- ses: NABC-based SES index
- town: Town
- unfunded_only: Self-funded contract exclusively
- village: Village

Capacities.dta

Definition: Contract-level capacity in the benchmark allocation excluding students whose allocation is fixed (due to data corruption)

File: clean_analysis.py
Input data:

- Applications.csv

Usage: Table C1
Variables:

- Cid: contract id

- capacity_wo_ABC_benchmark: contract-level capacity in the benchmark allocation excluding students whose allocation is fixed (due to data corruption)

Capacities_in_benchmark.csv

Definition: Contract-level capacity in the benchmark allocation
File: clean_analysis.py
Input data:
- Applications.csv

Usage: Table C1
Variables:
- Cid: contract id
- capacity: contract-level capacity in the benchmark allocation

Benchmark.dta

Definition: Applicant-level data including the benchmark allocation
File: clean_analysis.py
Input data:
- Applications.csv

Usage: Table C1
Variables:
- Id: applicant id
- Contract_id_orig: Contract id in the benchmark allocation

Narrow_market_power.csv

Definition: Summary of narrow market power
File: clean_analysis.py
Input data:
- Applications.csv

Usage:
- Table D1

Variables:
- Fraction: Fraction subject to market power
- Assigned: Assigned under preference flip algorithm
- Change: Change relative to SR-DA

Tables and Figures

Tables and figures in the paper can be replicated by running the 0_Master_replication.do file in the code/analyze_data folder.

Stata version: 18.5

For Appendix Figure F1, run FigureF1.R.

R version: 4.1.1

Packages:

- vroom version 1.5.5
- tidyverse version 1.3.1
- glue version 1.4.2
- broom version 0.7.9
- haven version 2.4.3