Testing BayesDB on an Educational Dataset

Belhal Karimi

March 2016

1 Introduction

This document aims to show a new way of exploring the College Scorecard dataset recently published by the U.S. Department of Education: `https://collegescorecard.ed.gov/data/`.
The Economist triggered discussion about this dataset in an article describing a ranking system based only on earnings after graduation: `http://www.economist.com/blogs/graphicdetail/2015/10/value-university` To explore trends in the data, we are going to use BayesDB.

BayesDB is a Bayesian database that lets users query the probable implications of their data as easily as a SQL database lets them query the data itself. Using the built-in Bayesian Query Language (BQL), users with no statistics training can solve basic data science problems, such as detecting predictive relationships between variables, inferring missing values, simulating probable observations, and identifying statistically similar database entries.
BayesDB is suitable for analyzing complex, heterogeneous data tables with up to tens of thousands of rows and hundreds of variables. No preprocessing or parameter adjustment is required, though experts can override BayesDB's default assumptions when appropriate.
BayesDB's inferences are based in part on CrossCat, a new, nonparametric Bayesian machine learning method that automatically estimates the full joint distribution behind arbitrary data tables.
In this particular example, we analyze a dataset of 1000 thousands rows and 50 columns (reduced from the original shape of 7000 rows and 1700 columns). And along the study, we will add some more rows and variables to test our model;

2 Problem Statement

This huge dataset is going to be a good field for BayesDB to show its underlying mechanisms to the reader.

· The issue is that traditional rankings of American colleges do not focus on many variables and base its core analysis on metrics related to graduates

earnings. In the meantime, the challenge here is to be able to compute a transparent value-added for each college in order to quantify the salary boost the students are receiving from attending such schools.

· Also, current rankings prefer translating the opportunities given by a school, via its network, its partnerships... and not on the hard working and intelligence qualities of their graduates.

· In this following document, we will ask BayesDB different questions in terms of relation between variables and distribution of some metrics to compare these results to our intuition and to other iterations of the same models

As a result, we should expect from our tool to take into consideration every relation between variables and show us intuitive dependencies and simulations.

We will organize the study in different parts:

1. The first part will consist in analyzing a small subsample, given our initial selection of 50 variables, of 100 observations (colleges). We'll be plotting dependencies heatmap and be adding more and more observations with fixed and variable number of models and iterations. Both cases would translate interesting behaviors.

2. Then, using gpmcc, a new implementation of crosscat from the lens of generative population model (GPM), we'll simulate distribution of few variables and compare them with the existing distributions and our intuition. We'll try the same experiment by inferring missing values only.

3. On the same variables, we 'll ask BayesDB to find unlikely data, whether they are outliers or input errors.

4. Finally, we'll simulate colleges by size, selectivity and graduate students earnings and challenge stereotypes on colleges thanks to these results.

3 Dataset

The dataset published by the U.S. Department of Education is composed of 18 years (from 1996 to 2013) of data about more than 7,000 schools. More than 1,700 variables have been measured each year for each university. The number of universities fluctuates according to the creation of new schools and the closing of existing ones.

The overall dataset contains more than 215 million values with 43% of them missing.

Figure 1 highlights the variables measured each year. Obviously, we've been able to measure more and more variables through the years, but for some reason many data are missing in 2013 (the last year this study has been conducted). Obama's administration college database is way more transparent and better

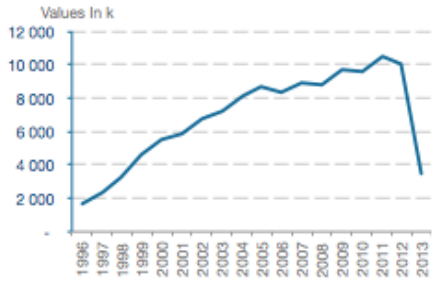

Figure 1: Data available per year

quality. It has better data on family income, family education levels of entering students and more sophisticated measures of degree completion as well as loan repayment. These numbers have been generated in particular by matching students loans to their actual tax returns. . As a result, we are able to compare professionals earnings to their students characteristics.

Note: : The data only includes students who applied for financial aid and thus is missing all the students with well-off parents. Also, as the earnings data only take into consideration 10 years after starting college, one could argue that this scope of time is too small to include future high earners, as they would still be students (e.g. Ph.D., post-doctorate).

The number of schools from a year to another is also varying Here is an overview of how many schools we have metrics of. One can tell that the overall trends is showing an increase in number of schools even though from 1996 to 1998 we can observe more schools closing than being created. One intersting result could be to caracterize the fall of trends of these schools and be able to predict future outlook of current schools based on their most recent data.

```
bdball = bayeslite.bayesdb_open("dfall.bdb", builtin_metamodels=False)
bdbcontrib.barplot(bdball, '''
SELECT year ,
COUNT(OPEID) AS "Number of schools"
FROM dfall
GROUP BY year
ORDER BY year asc
''');
```

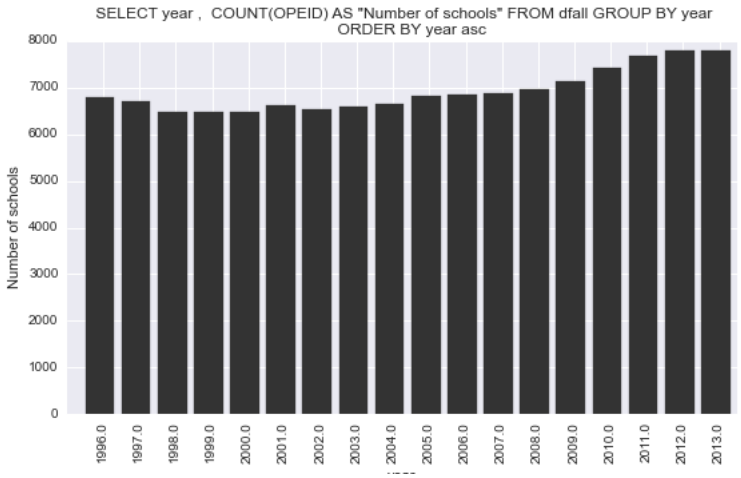

Figure 2: Number of colleges in the US per year

Our 57 variables include some financials and specs about the universities. Here is a full list of the variables selected for the study

var_name	description
Financials	
DEP_DEBT_MDN	The median debt for dependent students
LO_INC_DEATH_YR2_R	Percent of low income (between 0 and 30k in nominal family income) students who died within 2 years
MD_INC_DEATH_YR2_R	Percent of middle-income (between $30k and 75k in nominal family income) students who died within 2 years
HI_INC_DEATH_YR2_R	Percent of high income (more than 75k in nominal family income) students who died within 2 years
PELL_DEATH_YR2_RT	Percent of students who received a Pell Grant at the institution and who died within 2 years at original institution
NOPELL_DEATH_YR2_RT	Percent of students who never received a Pell Grant at the institution and who died within 2 years at original institution
LOAN_DEATH_YR2_RT	Percent of students who received a federal loan at the institution and who died within 2 years at original institution
NOLOAN_DEATH_YR2_RT	Percent of students who never received a federal loan at the institution and who died within 2 years at original institution
NOLOAN_ENRL_ORIG_YR	Percent of students who never received a federal loan at the institution and who were still enrolled at original institution within a year
DEATH_YR2_RT	Percent died within 2 years at original institution
LO_INC_COMP_ORIG_Y	Percent of female students who transferred to a 2-year institution and whose status is unknown within 8 years
MD_INC_COMP_ORIG_Y	Percent of middle-income (between $30k and 75k in nominal family income) students who died within a year
HI_INC_COMP_ORIG_Y	Percent of high-income (over in nominal family income) students who died within a year
PELL_COMP_ORIG_YR2	Percent of students who received a Pell Grant at the institution and who completed in 2 years at original
NOPELL_COMP_ORIG_YR2	Percent of students who did not receive a Pell Grant at the institution and who completed in 2 years at original
LOAN_COMP_ORIG_YR2	Percent of students who received a federal loan at the institution and who completed in 2 years
NOLOAN_COMP_ORIG_YR2	Percent of students who never received a federal loan at the institution and who were still enrolled at original institution within 2 years
MD_INC_RPY_5YR_RT	Five-year repayment rate by family income ($30k-75k)
GRAD_DEBT_MDN	The median debt for students who have completed
WDRAW_DEBT_MDN	The median debt for students who have not completed
LO_INC_DEBT_MDN	The median debt for students with family income between $0 and 30k
MD_INC_DEBT_MDN	The median debt for students with family income between $30k and 75k
HI_INC_DEBT_MDN	The median debt for students with family income between over 75k
IND_DEBT_MDN	The median debt for independent students
PCTPELL	Percentage of Pell Grant
AVGFACSAL	Average faculty salary
TUITIONFEE_PROG	TUITIONFEE_PROG
faminc	Average family income
md_faminc	Median family income
mn_earn_wne_p10	Mean earnings of students working and not enrolled 10 years after entry
md_earn_wne_p10	Median earnings of students working and not enrolled 10 years after entry
Ethnies	
PBI	Flag for predominantly black institution
AANAPII	Flag for Asian American Native American Pacific Islander-serving institution
MENONLY	Flag for men-only college
WOMENONLY	Flag for women-only college
Selectivity	
ADM_RATE_ALL	Admission rate for all campuses rolled up to the 6-digit OPE ID
SATVR25	25th percentile of SAT scores at the institution (critical reading)
SATVRMID	Midpoint of SAT scores at the institution (critical reading)
SATVR75	75th percentile of SAT scores at the institution (critical reading)
SATMT25	25th percentile of SAT scores at the institution (math)
SATMTMID	Midpoint of SAT scores at the institution (math)
SATMT75	75th percentile of SAT scores at the institution (math)
SATWR25	25th percentile of SAT scores at the institution (writing)
SATWRMID	Midpoint of SAT scores at the institution (writing)
SATWR75	75th percentile of SAT scores at the institution (writing)
SAT_AVG_ALL	Average SAT equivalent score of students admitted for all campuses rolled up to the 6-digit OPE ID
ACTCM25	25th percentile of the ACT cumulative score
ACTCMMID	Midpoint of the ACT cumulative score
ACTCM75	75th percentile of the ACT cumulative score
University specs	
st_fips	FIPS code for state
region	Region (IPEDS)
locale2	Degree of urbanization of institution
OPEID	8-digit OPE ID for institution
CCBASIC	Carnegie Classification -- basic
CCUGPROF	Carnegie Classification -- undergraduate profile
UGDS	Enrollment of undergraduate degree-seeking students
PFTFAC	Faculty Rate

Figure 3: Codebook for our variables

4 Analysis

4.1 100 observations

Let's start with a small subsample of 100 rows. We analyzed 16 models for 10 iterations:

```
#100 observations, 16 models and 10 iterations
ed100.analyze(models=16, iterations=10)
ed100.heatmap(ed100.q(
    '''ESTIMATE DEPENDENCE PROBABILITY FROM PAIRWISE COLUMNS OF %g'''))
```

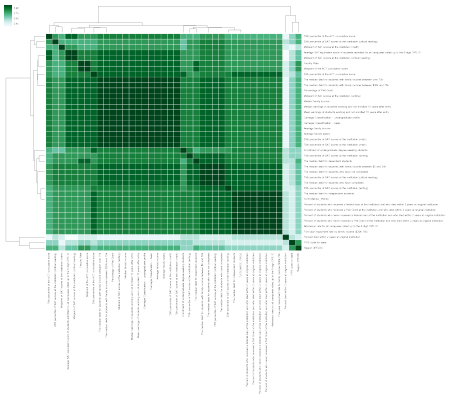

Figure 4: Dependence probability heatmap (100 obs, 16 models and 10 iterations)

Obviously, we were expecting this kind of result, meaning that no real dependant component are clear cut.
Even though variables related to SAT scores and ACT scores are expected to be strongly dependant, it is not obvious on our previous heatmap since the number of observation is too low in order for BayesDB to make such strong assumptions of dependencies Yet, some small blocks are clearly connected such as Percentage of Pell Grant, Median family income and average family income. For some reason, the "Carnegie Classification – undergraduate profile" variable is also connected to this block.

Note: The Undergraduate Profile Classification describes the undergraduate population with respect to three characteristics: the proportion of undergraduate students who attend part- or full-time; achievement characteristics of first-year, first-time students; and the proportion of entering students who transfer in from another institution. The lack of relevant observations could be

due to the small amount of observations or the lack of models and iterations Let's try to analyze more models first. We add 4 more models and do 20 more iterations and display the related heatmap.

4.2 100 observations and more models

```
#100 observations, 4 new models (20 in total) and 20 more iterations
ed100.analyze(models=20, iterations=20)
ed100.heatmap(ed100.q(
    '''ESTIMATE DEPENDENCE PROBABILITY FROM PAIRWISE COLUMNS OF %g'''))
```

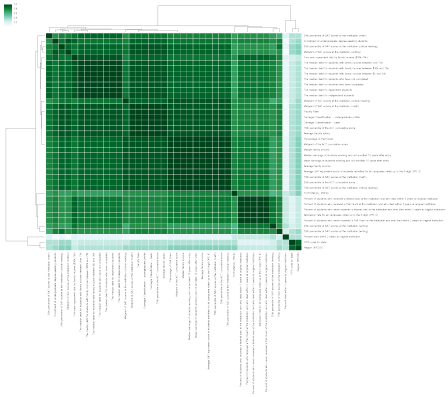

Figure 5: 100 obs, 20 models and 30 iterations

It seems to be slightly better but we still observe a raw heatmap with no real conclusions to draw... Let's add some more observations We'll check further dependencies on this latest variable (Carnegie Classification) to bring a logical explanation to this result

4.3 Increasing the number of observations to 1000

```
ed1000 = quickstart(name='df1000', bdb_path='bdb/df.bdb')
```

We can begin to explore dependence relationships between variables to test out the models.
One way to do that is, given the the models generated by CrossCat, and the iterations on these models invoked by our analysis, "counting" how many times two variables are in the same view of each model. In other words, the dependence relationship analysis will quantify the number of views generated by CrossCat including two pairwise columns
For readability purposes, the Python client for bayeslite makes it straightforward to examine the overall matrix of pairwise dependence probabilities. Cell

(i,j) in this matrix records Pr[variable i is dependent on variable j]. The matrix is reordered using a clustering algorithm to make higher-order predictive relationships — cases where some group of variables are probably all mutually independent — more visually apparent

```
#1000 observations, 16 models and 10 iterations
ed1000.analyze(models=16, iterations=10)
ed1000.heatmap(ed1000.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

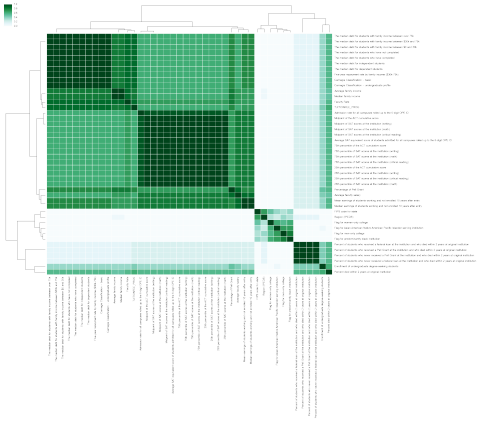

Figure 6: 1000 obs, 16 models and 10 iterations

There is a component block that can can be interesting to study in more detail. The one that shows relationship between Median family income, Mean earnings for students workings and not enrolled after 10 years of graduation and Median earnings for students workings and not enrolled after 10 years of graduation
Also, our former dependent block of four variables (reminder: Percentage of Pell Grant, Median family income, Average family income and Carnegie Classification) seems to have exploded since now the Median and Average family income are connected, together obviously, and to the Tuition fee of the school. Which makes totally sense: the richer the family the higher fee the student can afford to pay.
As far as Percentage of Pell Grant, it seems that the variable joined another block composed of Median earnings of students, Mean earnings of students and average faculty salary. The dependence between those variables and Pell Grant attribution seems strange.

Let's do 40 more iterations on the same number of models and see the evolution of the heatmap.

```
#1000 observations, 16 models and 40 iterations
ed1000.analyze(models=16, iterations=40)
ed1000.heatmap(ed1000.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

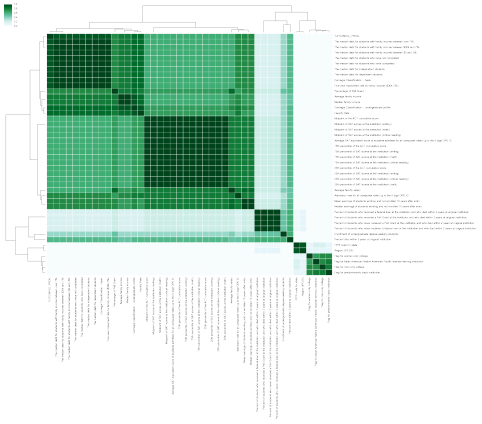

Figure 7: 1000 obs, 16 models and 50 iterations

Contrary to the evolution of the heatmap with 100 observations with respect to the increase of models and analysis, we can clearly see the improvement in terms of neat dependent blocks when we are analyzing 1000 observations.
Here the biggest dependant blocks were found after the first analysis but the fact that we added more analysis clearly refined some dependencies between few variables. Probabilities between 0.4 and 0.6, that can not lead to clear-cut interpretations, were whether strengthen by more analysis or weakened. That led to visually more dependant dark green block or white ones (meaning no dependencies).

Now that we have analyzed the behavior of doing more iterations on the same number of partitions (models), we'll iterate on more models relating to more combination of variables, in terms of mutual information. Let's create four more models and do 30 iterations on these new models and 30 more on the previous 16 models.

```
#1000 observations, 20 models and 30 iterations
ed1000.analyze(models=20, iterations=30)
ed1000.heatmap(ed1000.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

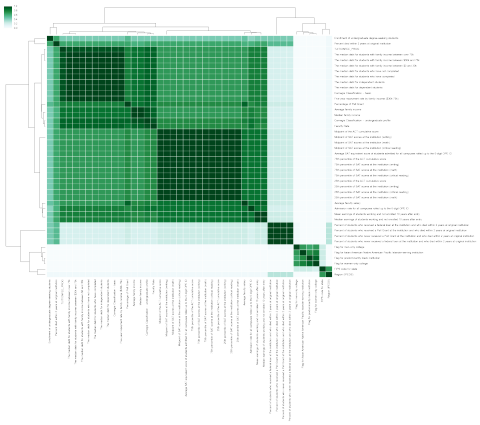

Figure 8: 1000 obs, 20 models and 80 iterations

Adding more models definitely deleted some regions where the dependency was not clear-cut not within a dependent block this time but between few variables and all of the others.
As the number of models are exactly the number of view, ie the clusters in which variables are put together, the fact that we limit the number of models freezes some depency probability bewteween variables even though there may not be a reason for that. Besides, the number of analysis won't change since the iterations is not changing the composition of the views.

4.4 Adding 2000 observations to our dataset (3000 in total)

Results are until now pretty clear since we've multiplied the number of observations by 10. Let's iterate one last time before considering the overall dataset. In this case we are taking into account 3000 observations and we'll analyze only 16 models but for different number of iterations.
In other words we'll try to highlight the evolution of BayesDB going through the data with a decent number of observations and of models.

```
#3000 observations, 16 models and 10 iterations
ed3000.analyze(models=16, iterations=10)
ed3000.heatmap(ed3000.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

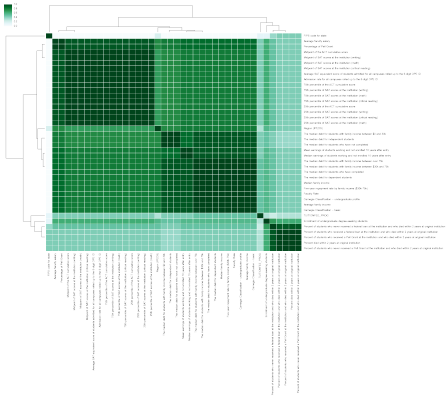

Figure 9: 3000 obs, 16 models and 10 iterations

Retrospectively, we can see how the number of observations is clearly helping the model deciding on what variable to put inside the same view. Indeed, even with very few models and analysis we have already three highly dependent connected blocks.

```
#3000 observations, 16 models and 80 iterations
ed3000.analyze(models=16, iterations=80)
ed3000.heatmap(ed3000.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

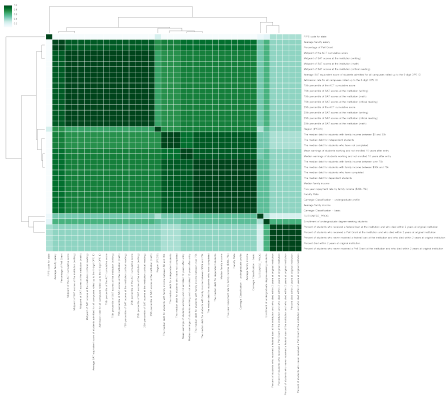

Figure 10: 3000 obs, 16 models and 90 iterations

It does not change anything. One can guess that BayesDB need more models to really think of new dependencies. We can add four more models and run some analysis.

```
#3000 observations, 20 models and 10 iterations
ed3000.analyze(models=20, iterations=10)
ed3000.heatmap(ed3000.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

Adding more models and doing a few more iterations has helped way more than analysing the same old models over and over again. As we can see, when we were dealing with 16 models, all the metrics related to ACT and SAT scores were in the same dependent block (with the same dependent probability) and it did not evolve with the time of analysis.
Whereas in the 20 models analysis, these metrics were gathered into three connected components. For instance, the SAT scores variables (25th percentile, median,...) have been gathered in one component, the ACT scores in another and a small block is putting in relation the average faculty salary and the percent of Pell Grant.

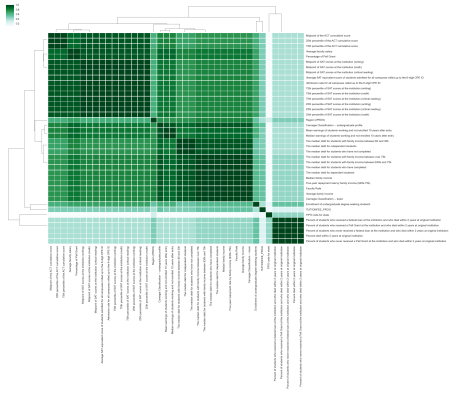

Figure 11: 3000 obs, 20 models and 100 iterations

4.5 Considering the entire dataset: 7804 colleges

The dataset we use is now including all to observations of the most recent data available about US colleges.
We begin with few iterations to see if the number of observations can help BayesDB connect some variables without going through all of it lots of time.

```
#7804 observations, 16 models and 5 iterations
edall.analyze(models=16, iterations=30)
edall.heatmap(edall.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

Figure 12: Whole dataset, 16 models and 5 iterations

This part is really interesting if our goal was to focus on what models BayesDB is built upon. Now that we are dealing with the overall dataset, 7804 observations and 57 variables, few models and very little analysis, the heatmap presents few regions of mid range probability of dependency.

We then add more models and more iterations.

```
#7804 observations, 24 models and 40 iterations
edall.analyze(models=16, iterations=50)
edall.heatmap(edall.q('''ESTIMATE DEPENDENCE PROBABILITY
  FROM PAIRWISE COLUMNS OF %g'''))
```

Figure 13: Whole dataset, 24 models and 45 iterations

4.6 Let's see the evolution of a dependent block

What think that can be interesting is to see the evolution according to the number of observations, the number of models and iterations, of a small block of variables: Admission rate, Midpoint of SAT scores at the institution (writing), The median debt for dependent students, The median debt for students who have completed.
Here is a visual evolution of this connected block and the status of the analysis at each point:

```
edx.heatmap(edx.q('''ESTIMATE DEPENDENCE PROBABILITY FROM PAIRWISE COLUMNS OF %g'''),
selectors={'MCAS': lambda name:"Admission rate" in name
    or "Midpoint of SAT scores at the institution (writing)" in name
    or "The median debt for dependent students" in name
    or "The median debt for students who have completed" in name
```

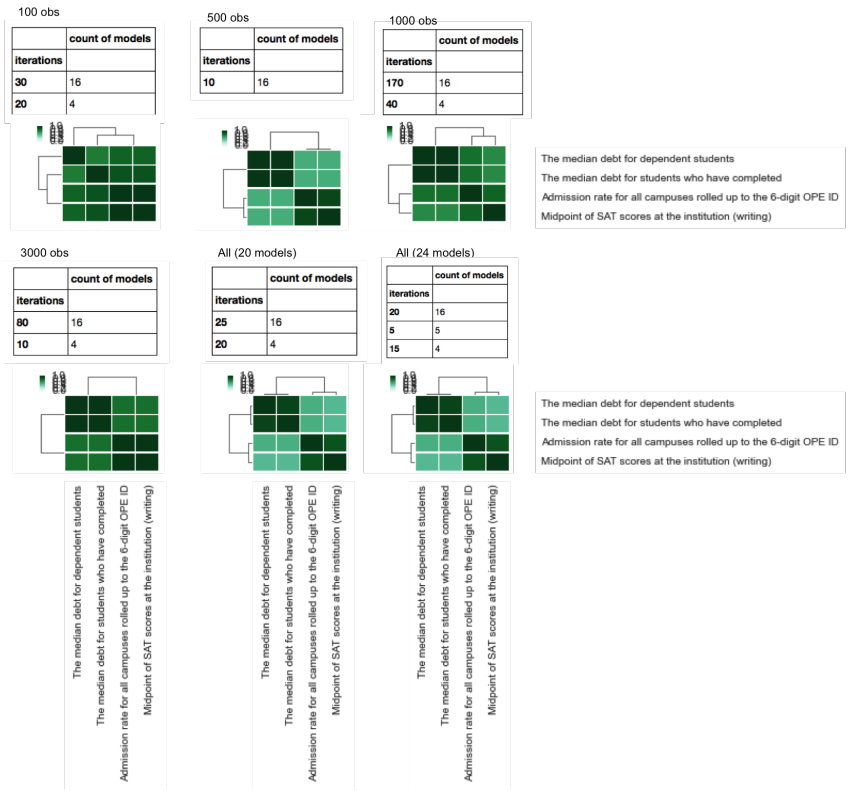

Figure 14: Evolution of a connected block

The figure above is showing us the evolution of a specific dependent block from a subsample to another and with different analysis ran on several models. Let's start with our intuition. As far as median debt variables, we can expect a strong dependency between the median debt for dependent students and for students who have completed since the dependent students can totally be included in the students who have completed and then the mutual information between those two variables is really high.
As far as Admission rate and SAT score at the institution, again we can obviously draw a relation between very selective school and a high education level (high SAT scores).

5 Gpmcc simulation versus our intuition in terms of distribution

In this part we will compare the intuitive distribution of 'the admission rates' and 'the average family income' versus the distribution that Bayes DB simulates.

5.1 Admission rate

Let's take the admission rate for all campuses.
Here is how the raw data looks like

```
bdbcontrib.histogram(bdb, '''SELECT "Admission rate for
all campuses rolled up to the 6-digit OPE ID"
                    FROM taball''', bins=35, normed=True);
```

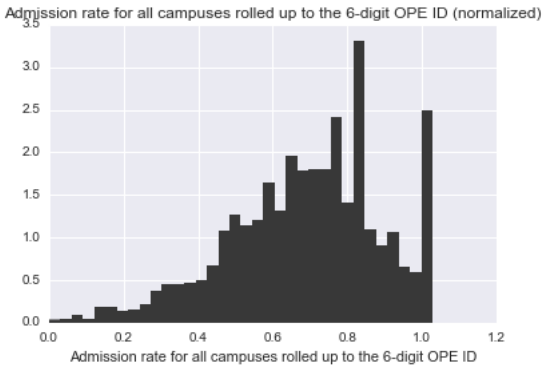

Figure 15: Distribution of the Admission rate (from the dataset)

This result is pretty far from our intuition.
Indeed, our intuition would say that the admission rate follow a normal distribution where most of the schools would have an average admission rate (whatever the average of that distribution might be, our intuition would say that most of the schools accept 50 to 60% of their applicants).
So now, let's simulate all the values of the admission rates and see what Bayes DB suggest us as a distribution.
We first create the table with all the simulated value

```
q('''
CREATE TABLE adm_rate_all AS
    SIMULATE "Admission rate for all campuses
    rolled up to the 6-digit OPE ID"
    FROM %g
    LIMIT 3000;
''');
```

And then plot it:

```
bdbcontrib.histogram(bdb, '''SELECT "Admission rate for all
campuses rolled up to the 6-digit OPE ID"
                    FROM adm_rate_all''', bins=35, normed=True);
```

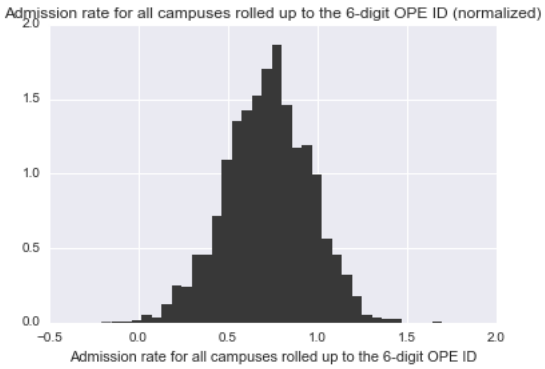

Figure 16: Distribution of the Admission rate (Simulated)

This confirms our initial intuition
Now let's see if we only infer the missing values:
Again we first create the table with the missing inferred values:

```
q('''
CREATE TABLE adm_rate_nan AS
INFER EXPLICIT "Average SAT equivalent score of
students admitted for all campuses rolled up to the 6-digit OPE ID",
```

```
    PREDICT "Admission rate for all campuses
        rolled up to the 6-digit OPE ID" AS adm_rate
            CONFIDENCE adm_rate_conf FROM %g''')
```

And plot the histograms of the values

```
bdbcontrib.histogram(bdb, '''SELECT adm_rate
                      FROM adm_rate_nan''', bins=35, normed=True);
```

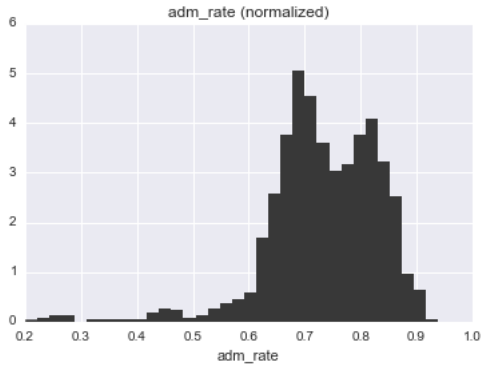

Figure 17: Distribution of the Admission rate (Missing values inferred)

If I only infer the missing values, the distribution seems to be bimodal. As if the missing values were distributed independently of the previous known data and were forming a new normal distribution. The juxtaposition of the both giving our bimodal like shape. There might be something to improve in BayesDB for that part.

5.2 Family income

The distribution of annual household income in the US is known to be this shape

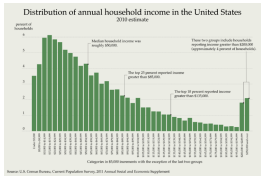

Figure 18: Distribution of the Average Family Income (Model)

Here is how our distribution of average family income looks like

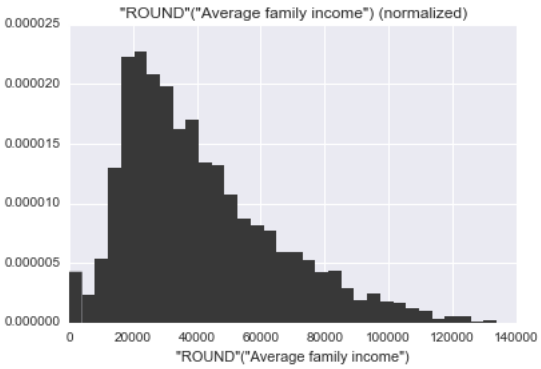

Figure 19: Distribution of the Average Family Income (from the dataset)

It seems like the tail of the shape is fitting our assumption but the beginning of the distribution is somehow confusing.
This has to be caused by the 1993 missing values (cf. 'SELECT COUNT')
Let's simulate all the missing values and plot the distribution again.

```
q('''
CREATE TABLE income_all AS
    SIMULATE "Average family income"
    FROM %g
    LIMIT 7804;
''');
bdbcontrib.histogram(bdb, '''SELECT ROUND("Average family income")
                     FROM income_all''', bins=35, normed=True);
```

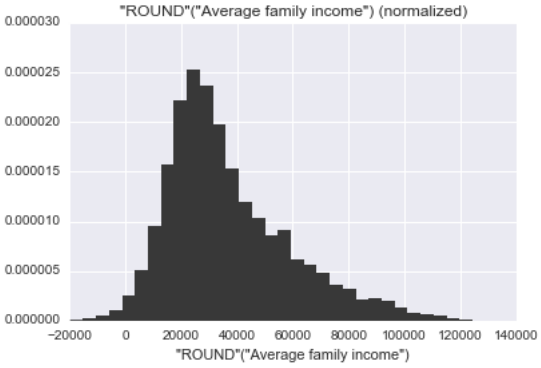

Figure 20: Distribution of the Average Family Income (Simulated)

As you can see (consider only positive values), the head of the distribution has been smoothed by the model

Though, the end of the tail, representing the little percentage of highly rich family does not seem to appear in our data. This is quite normal since there are no case of university where only highly rich family put their children in. In other words, the family income is always averaged for each school given the fact that all kind of wealth are represented accross the country

6 Finding outliars on a subset of variables

In this section, we'll be working on the overall dataset and the goal will be to find unlikely data for several variables of interest. For that, we'll be computing the 20 lowest relative probabilities of their values.
After 25 iterations on 16 models, we are able to return some relevant results.
For all the variables we'll be studying, the first step consists in creating a temporary table with the name of the colleges, the values of the current variable and its relative probability.

```
q('''
CREATE TEMP TABLE unlikely_variable AS
    ESTIMATE Name,<variable>,
    PREDICTIVE PROBABILITY OF <variable>
            AS "Relative Probability of <variable>"
    FROM \%g;
''')
```

Then we just have to select the values from that table and order by probability ascending in order to extract the lowest ones.

```
q('''
SELECT * FROM unlikely_variable
    WHERE <variable> IS NOT NULL
    ORDER BY "Relative Probability of <variable>" ASC LIMIT 20;
''')
```

6.1 Size of the colleges

Let's start with the size of the colleges ('Enrollment$_{o}f_{u}ndergraduate_{d}egree_{s}eeking_{s}tudents$')

	Name	Enrollment_of_undergraduate_degree_seeking_students	Relative Probability of Enrollment
0	Illinois State University	17667	0.000065
1	The University of Texas-Pan American	17200	0.000068
2	Chaffey College	17217	0.000068
3	Grossmont College	17042	0.000070
4	Community College of Allegheny County	17098	0.000071
5	Thomas Nelson Community College	9097	0.000073
6	Weber State University	17314	0.000073
7	Sierra College	17195	0.000074
8	University of California-Los Angeles	28667	0.000079
9	Lamar University	9175	0.000081
10	John Wood Community College	1638	0.000086
11	Los Angeles Pierce College	17528	0.000086
12	Bloomsburg University of Pennsylvania	9245	0.000086
13	Georgia Southern University	17103	0.000087
14	University of Central Missouri	9160	0.000088
15	Truckee Meadows Community College	9051	0.000090
16	University of Washington-Seattle Campus	28754	0.000091
17	Arkansas State University-Main Campus	9325	0.000094
18	Delgado Community College	17467	0.000095
19	Bucks County Community College	9347	0.000096

Figure 21: Simulation of schools by size

Examining the table, no values are strikingly unusual. All values could be plausible enrollment numbers for colleges. However, it is notable that half of these are community colleges with relatively high enrollment (Chaffey College, Grossmont College, Community College of Allegheny County, Thomas Nelson Community College, Sierra College, John Wood Community College, Los Angeles Pierce College, Truckee Meadows Community College, Delgado Community College, and Bucks County Community College). It is intuitive that community colleges wouldn't be extremely likely to have enrollments as large as 9000 to 18000 students. Community colleges tend to be smaller than, for example, public four-year colleges with many located outside of major cities.

6.2 Admission rates

As far as admission rates, we can also find some interesting results.

	Name	nb_students	adm_rate	Relative Probability of Admission Rate
0	Riverside School of Health Careers	305	0.0385	0.055331
1	College of the Ozarks	1513	0.1301	0.074803
2	Robert Morris University Illinois	2775	0.2100	0.084833
3	St Louis College of Health Careers-St Louis	295	0.3158	0.086872
4	West Virginia University Hospital Departments ...	42	0.1327	0.093155
5	Plaza College	726	0.2992	0.104145
6	St Louis College of Health Careers-Fenton	301	0.3158	0.118807
7	Corban University	902	0.3225	0.136772
8	University of Missouri-Kansas City	8309	0.3300	0.150081
9	Crossroads College	97	0.2143	0.152057
10	Corban University Puget Sound Campus	NaN	0.3225	0.153196
11	Adventist University of Health Sciences	2064	0.1336	0.162573
12	Texas Wesleyan University	1719	0.2992	0.179648
13	University of Puerto Rico-Bayamon	4847	0.2547	0.188586
14	Missouri Valley College	1411	0.2218	0.192385
15	Maria College of Albany	866	0.1603	0.193401
16	Mississippi Valley State University	1856	0.2272	0.214300
17	The Ailey School	91	0.1784	0.215720
18	Universidad del Sagrado Corazon	5097	0.3186	0.217307
19	Cass Career Center	28	0.1364	0.220802

Figure 22: Simulation of schools by selectivity

The most unlikely value here definitely stands out: 3.85%! Even Harvard's admission rate was higher than that – 5.9% – in 2014. Doing a quick Google search on Riverside School of Health Careers reveals its acceptance rate was 40.9% as of 2010. Thus, this value was almost certainly incorrectly entered. While spot checking a few other values, I also came across what seems like a possible mistake for Crossroads College's acceptance rate: here it is listed as 21.4%, but in 2014 it was reported as 90%: `https://nces.ed.gov/collegenavigator/?id=174206#admsns` Noticeably, several of these colleges are very small and have a specific focus, for example, The Ailey School is a dance school and college and a handful of these colleges are geared toward careers in health. The specificity of their programs may limit the types of students and number of students they can admit.

6.3 Tution fees

	Name	Average_faculty_salary	Median_family_income	TUITIONFEE_PROG	Relative Probability of Tuition
0	Professional Training Centers	4303	14177.5	54600	0.000010
1	Western Technical College	3482	18033.0	30746	0.000011
2	Valley Grande Institute for Academic Studies	NaN	15427.5	25138	0.000012
3	WellSpring School of Allied Health-Lawrence	2500	NaN	24000	0.000012
4	Lincoln College of Technology-Melrose Park	4741	18719.0	31502	0.000013
5	Platt College-Moore	3772	15045.0	30020	0.000013
6	Delta College Inc	NaN	14276.0	23000	0.000013
7	Medical Professional Institute	NaN	17156.0	24000	0.000013
8	Capitol City Careers	NaN	14038.0	28601	0.000014
9	Pickaway Ross Joint Vocational School District	NaN	23030.0	539	0.000015
10	Quest College	3333	13717.0	24113	0.000015
11	St Louis College of Health Careers-Fenton	4083	12191.0	28400	0.000015
12	North-West College-Pasadena	NaN	15391.5	28650	0.000015
13	Detroit Business Institute-Downriver	NaN	10906.5	26400	0.000016
14	North-West College-Pomona	NaN	16835.0	28525	0.000016
15	North-West College-Riverside	NaN	16835.0	28525	0.000016
16	Manhattan Institute	NaN	NaN	760	0.000016
17	Anamarc College-El Paso East	3750	NaN	26035	0.000016
18	PCI College	NaN	14671.0	29190	0.000016
19	Stone Academy-Waterbury	NaN	10115.0	7400	0.000016

Figure 23: Simulation of schools by tuition fees

A notable trend is that for many of these schools, the median family income is both very low and much lower than the tuition cost. This is especially true for the least likely tuition value – $54k, more than $40k higher than the median family income for the students. Assuming these families comprise two or more people, they would be below the poverty line: https://www.healthcare.gov/glossary/federal-poverty-level-FPL/ Thus, while tuition in the $20-30k range is not surprisingly high, we would expect to see a higher median family income for these schools given their tuition cost. A couple of the tuition values are surprisingly low – $760 and $539 – but they are correct based on cross-verification: `http://college-tuition.startclass.com/l/7313/Manhattan-Institute` `http://college-tuition.startclass.com/l/5161/Pickaway-Ross-Joint-Vocational-School-District` (Data on both webpages is provided by the U.S. Dept. of Education)

7 Similarities between schools as far as simple variables

In this section, we want to use crosscat generative population models to assess some of the stereotypes gravitating towards US Colleges.
To do so, we'll consider our overall dataset and build models on it.
After having ran 10 analysis on 10 models, we were able to simulate similarities with decent accuracy with respect to several variables.
The following analysis will mainly involve similarity to one school. Here we decided to start with the similarities with respect to the selectivity (Admission rate). For that we first had to select the most selective college in the US.

```
q('''
SELECT
    INSTNM,
    "key",
    "Admission rate for all campuses rolled up to the 6-digit OPE ID"
from df
WHERE
    "Admission rate for all campuses rolled up to the 6-digit OPE ID" IS NOT NULL
ORDER BY
    "Admission rate for all campuses rolled up to the 6-digit OPE ID" ASC LIMIT 15 ''')
```

	INSTNM	key	Admission rate for all campuses rolled up to the 6-digit OPE ID
0	Graham Hospital School of Nursing	893800	0.0000
1	Trinity College of Nursing & Health Sciences	622500	0.0000
2	Platt College-Aurora	3014900	0.0000
3	Riverside School of Health Careers	2140000	0.0385
4	Curtis Institute of Music	325100	0.0484
5	Stanford University	130500	0.0569
6	Harvard University	215500	0.0584
7	Yale University	142600	0.0705

Figure 24: Admission rates per college, ascending order

Obvisously the first colleges we have in our query that is ranked in ascending admission rate are colleges with 0 as admission rate (they are not NULL metrics other wise we would not have displayed it according to the query). These values are considered as NaN.
Then the first colleges with respectively 3 and 4% as admission rates are actually very specific colleges (for health and musical careers. The high selectivity

can be explained by the fact that a lot of people apply to these colleges without really knowing that they expect very specialized students in these two paths).

Then Stanford is according to that table the most selective college with a 5.7% admission rate.
Let's compute similarity to Stanford with respect to Admission rate first‘

```
q('''
ESTIMATE
"INSTNM",
SIMILARITY TO ("key" = "130500") WITH RESPECT TO
("Admission rate for all campuses rolled up to the 6-digit OPE ID") as value
FROM %g
ORDER BY value DESC ''')
```

	INSTNM	value
0	Massachusetts Institute of Technology	1.0
1	Stanford University	1.0
2	Yale University	0.9
3	University of Chicago	0.9
4	Princeton University	0.9

Figure 25: Similarity to Stanford with respect to Admission rate

With respect to Admission rate, the most selective college, that is Stanford (cf. above), is similar to MIT (1.0), Yale, University of Chicago and Princeton (0.9).
So now let's see what happened to this ranking when we add more variables that are relevant to having a selective process.
We can now add three variables and compute the similarity again with respect to:

· the mean earnings after enrollment

· the tuition fee

· Admission rate

· the median debt.

```
q('''
ESTIMATE "INSTNM", SIMILARITY TO ("key" = "130500") WITH RESPECT TO
("Admission rate for all campuses rolled up to the 6-digit OPE ID",
```

```
"TUITIONFEE_PROG",
"The median debt for dependent students",
"Mean earnings of students working and not enrolled 10 years after entry") as value
FROM %g
ORDER BY value DESC ''')
```

	INSTNM	value
0	Stanford University	1.000
1	Massachusetts Institute of Technology	0.975
2	Yale University	0.900
3	Princeton University	0.875
4	Georgetown University	0.850

Figure 26: Similarities with respect to 4 variables

Adding those variables slightly changed the similarities to Stanford. For instance, Chicago disappeared from top 5 table.
We can add one more variable : "Average family income".

```
q('''
ESTIMATE "INSTNM", SIMILARITY TO ("key" = "130500") WITH RESPECT TO
("Admission rate for all campuses rolled up to the 6-digit OPE ID",
"TUITIONFEE_PROG",
"The median debt for dependent students",
"Mean earnings of students working and not enrolled 10 years after entry",
"Average family income") as value
FROM %g
ORDER BY value DESC ''')
```

	INSTNM	value
0	Stanford University	1.00
1	Massachusetts Institute of Technology	0.98
2	Yale University	0.92
3	Princeton University	0.90
4	University of Chicago	0.88

Figure 27: Similarities with respect to 5 variables

Here, we notice that the University of Chicago became more similar to Georgetown University.
We could move on like that many times always adding new variables, and it would keep on showing that comparing a college to another has to be done with respect to several criteria, if possible as many as possible, and that those criteria have to be chosen carefully, i.e. their value has to be strongly related to the characteristics of the students post graduation (if the goal is to make a ranking of the best colleges).
Under this perspective, we can highlight the fact that whether or not you include the average family income in the similarity query, you get different result.

8 Future Work