Earth Mover's Distance between Grade Distribution Data with Fixed Mean

by

Jan Kretschmann

A Thesis Submitted in
Partial Fulfillment of the
Requirements for the Degree of

Master of Science
in Mathematics

at

The University of Wisconsin-Milwaukee
August 2020

ABSTRACT

Earth Mover's Distance between Grade Distribution Data with Fixed Mean

by

Jan Kretschmann

The University of Wisconsin-Milwaukee, 2020
Under the Supervision of Professor Jeb Willenbring

The Earth Mover's Distance (EMD) is examined on all theoretically possible grade distributions with the same grade point average (GPA). The numbers of distributions with the same EMD and GPA are encoded in the coefficients of a generating function. The theoretical mean EMD for grade distributions, that are sampled uniformly and independently at random, is computed from this function, and compared to real world grade data taken from several years. The data is further examined regarding the appearance of clusters that change when varying the distance threshold.

Table of Contents

List of Figures

List of Tables

Introduction

This thesis will examine the expected value of the *Earth Mover's Distance* (EMD). To formally define the EMD, it is necessary to first define the set of joint distribution (see [BW19]) :

$$\mathcal{J}_{\mu\nu} = \left\{ J \in \mathbb{R}^{n\times n} : \begin{array}{l} J \text{ is a non-negative real number } n \text{ by } n \text{ matrix such that} \\ \sum_{i=1}^{n} J_{ij} = \mu_j \text{ for all } j \text{ and } \sum_{j=1}^{n} J_{ij} = \nu_i \text{ for all } i \end{array} \right\}.$$

where $\mathcal{P}_n$ is the set of all probability measures on a set of numbers $\{0, 1, \ldots, n\}$ and $\mu, \nu \in \mathcal{P}_n$. The EMD is defined as

$$\mathbb{EMD}(\mu, \nu) = \inf_{J\in\mathcal{J}_{\mu\nu}} \sum_{i,j=1}^{n} |i - j| J_{ij}.$$

In this thesis, the practical use of the EMD will be to measure the distance between grade distributions, specifically of classes with 30 students and the grades A, B, C, D and F. Each letter grade is assigned a number by the standard Grade Point Average (GPA): A is 4.0, B is 3.0, C is 2.0, D is 1.0 and F corresponds to 0. In order to compute the relative distance between two grades, it is only necessary to compute the absolute value of the point grade difference: for example, the distance of a B (3.0) to a D is $|3.0 - 1.0| = 2$. Some useful examples are given in [BW19]: suppose there is a class with 30 students and the five grades

A-F. Three possible grade distributions are given by

	A	B	C	D	F
X	0	19	8	2	1
Y	12	2	5	11	0
Z	2	20	2	3	3

Comparing distributions X and Y, one notices they were identical if 12 A grades in Y were changed to B, 5 C grades changed to B, 8 D grades changed to C, and one D grade changed down to F. The grade movement is encoded in the matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 12 & 2 & 5 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

where the columns and rows correspond to (A, B, C, D, F) and entry (i, j) stands for the number of grades that were moved from position i in X to position j in Y. The diagonal entries represent no grade change. The row sums return the X distribution, while the column sums return the Y distribution. The total EMD value is 26, which corresponds to the sum of the off-diagonal.

The grade movements between Y and Z are encoded in the matrix

$$\begin{bmatrix} 2 & 10 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 3 & 2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

with the EMD 23, and the movements between X and Z are encoded in

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 17 & 0 & 0 & 0 \\ 0 & 3 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

with the EMD 10. All the above distributions have the same GPA of 2.5, which shows that the EMD will distinguish between grade distributions even if the GPA is the same. In [BW19] there are three additional example distributions:

	A	B	C	D	F
U	13	13	0	0	4
V	9	1	13	2	5
W	9	7	8	6	0

this time with different GPAs, that are used to give an example for a distance matrix:

EMD	U	V	W	X	Y	Z
U	0	24	20	24	24	18
V	24	0	12	26	16	22
W	20	12	0	16	10	16
X	24	26	16	0	26	10
Y	24	16	10	26	0	26
Z	18	22	16	10	26	0

This thesis will focus on the EMD of grade distributions with a fixed GPA. Fixing the number of students to 30 and the number of grades to 5, gives a finite number of possible distributions, which will be examined theoretically. Additionally, there will be an examination of real world data from the University of Wisconsin-Milwaukee. Grade distributions from the years 2014 to 2018 will be investigated, and considering only classes with 30 students and a fixed GPA allows for a comparison to the theoretical result. Finally, the classes from one year will be examined in more detail. If some grade distributions have a particularly low EMD, they will form a connected component that is persistent through a varying number of distance thresholds. These components will be visual in EMD-based clustering of the grade data.

Background on Formal Power Series

II.1 Generating Function for the EMD

The approach in [BW19] was to encode the distribution of the discrete EMD in the coefficients of a formal power series, which is called a *generating function*. Let $a_0, a_1, a_2, ...$ be any sequence of numbers, then the *generating function* for this sequence is

$$a_0 s^0 + a_1 s^1 + a_2 s^2 + ...$$

or simply $f(s) = \sum_{n=0}^{\infty} a_n s^n$. If there is an $n* \in \mathbb{N}$ such that $\forall n > n* : a_n = 0$, the series is also called *generating polynomial* [Lan03].

The power series to encode values of the EMD is defined as

$$H_{p,q}(z,t) := \sum_{s=0}^{\infty} \left(\sum_{(\mu,\nu) \in \mathcal{C}(s,p) \times \mathcal{C}(s,q)} z^{\mathrm{EMD}_s(\mu,\nu)} \right) t^s,$$

where t, z are indeterminates and $\mathcal{C}(s,n)$ are the weak compositions of s into n parts, or

$$\mathcal{C}(s,n) = \{(a_1, a_2, \cdots, a_n) \in \mathbb{N}^n : a_1 + \cdots + a_n = s\}.$$

The coefficient of t^s is a polynomial in z, which records the distribution of the discrete EMD values.

$H_{p,q}$ is computed by:

Theorem 2.1. *For positive integers p and q,*

$$H_{p,q}(z,t) = \frac{H_{p-1,q}(z,t) + H_{p,q-1}(z,t) - H_{p-1,q-1}(z,t)}{1 - z^{|p-q|}t}$$

if $(p,q) \neq (1,1)$ *and* $H_{1,1} = \frac{1}{1-t}$.

Proof. The proof is given in [BW19]. □

For $p = q = 3$, this results in

$$H_{3,3}(t,z) = \frac{-t^3z^4 - t^2(2z+1)z^2 + t(z+2)z + 1}{(1-t)^3(1-tz)^2\left(1-tz^2\right)}.$$

Expanding until t^2 gives the polynomial:

$$H_{3,3}(t,z) = 1 + t\left(2z^2 + 4z + 3\right) + 2t^2\left(z^4 + 2z^3 + 6z^2 + 6z + 3\right) + O\left(t^3\right)$$

Now, we can see that the coefficient of, for example, t^2 is

$$C(z) = 2z^4 + 4z^3 + 12z^2 + 12z + 6z^0$$

a polynomial in z. In the context of grade distributions, we are looking at 3 possible grades ($H_{3,3}$) and classes of 2 students (t^2). Now, the monomials are structured as follows: nz^k means, that there are n possible pairs of distributions, that have an EMD of k. For example, there are 2 possible distributions with an EMD of 4.

Computing the weak compositions of 2, that consist of 3 elements gives us all the possible distributions in our scenario. Table II.1 shows a list of all the compositions.

In accordance with the polynomial C, there are only two possible pairs with an EMD of 4: $\{(2,0,0),(0,0,2)\}$ and the inverse $\{(0,0,2),(2,0,0)\}$.

$$\begin{pmatrix} 2 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 2 \end{pmatrix}$$

Table II.1: Weak compositions of 2 with 3 elements.

Main Result

To achieve the goal of recording information on the GPA in the generating function, first define $\mathcal{T}(\mu)$ as a weighted total of a distribution μ with elements $\{\mu_1, \ldots, \mu_n\}$, specifically

$$\mathcal{T}(\mu) = \sum_{i=0}^{n-1} i\mu_{i+1}$$

where n is the number of elements in μ. To include information on $\mathcal{T}$, extend the power series in [BW19] to:

$$H_{p,q}(z, t, g_1, g_2) := \sum_{s=0}^{\infty} \left(\sum_{(\mu,\nu)\in\mathcal{C}(s,p)\times\mathcal{C}(s,q)} g_1^{\mathcal{T}(\mu)} g_2^{\mathcal{T}(\nu)} z^{\mathrm{EMD}_s(\mu,\nu)} \right) t^s,$$

Now, the coefficient of t^s in $H_{p,q}$ is a polynomial in z, g_1 and g_2 whose coefficients record the distribution of the values of $\mathrm{EMD}_s(\mu, \nu)$, given the values of $\mathcal{T}(\mu)$ and $\mathcal{T}(\nu)$ saved in the exponents of g_1 and g_2.

To compute values of $H_{p,q}$, consider the following Theorem 3.2.

Theorem 3.2. *For positive integers p and q,*

$$H_{p,q}(z, t, g_1, g_2) = \frac{H_{p-1,q}(z, t, g_1, g_2) + H_{p,q-1}(z, t, g_1, g_2) - H_{p-1,q-1}(z, tg_1, g_2)}{1 - z^{|p-q|} t g_1^{p-1} g_2^{q-1}}$$

if $p > 1$ and $q > 1$, $H_{1,1} = \frac{1}{1-t}$ and $H_{p,q} = 0$ if $p < 1$ or $q < 1$.

Proof. Let

$$\mathcal{R}^s_{p,q} := \left\{ J \in \mathbb{M}_{p,q} : (\forall i,j), J_{ij} \in \mathbb{N}, \sum_{i,j} J_{ij} = s \text{ and support}(J) \text{ is a chain} \right\}.$$

be the vector space of all degree s homogeneous polynomials on p by q matrices, $\mathbb{M}_{p,q}$. By [BW19], we get a basis for $(\mathcal{R}^s_{p,q})$ by considering the monomials

$$\prod_{i=1}^{p} \prod_{j=1}^{q} x_{ij}^{J_{ij}}$$

where J is a non negative integer matrix with support on a chain.

Assigning each monomial the expression $z^{\text{EMD}_{s(\mu,\nu)}} g_1^{\mathcal{T}(\mu)} g_2^{\mathcal{T}(\nu)} t^s$ and summing them as a formal power series, the Hilbert series is obtained:

$$\sum_{s=0}^{\infty} \left(\sum_{(u,v)\in\mathcal{C}(s,p)\times\mathcal{C}(s,q)} z^{\text{EMD}_s(u,v)} g_1^{\mathcal{T}(u)} g_2^{\mathcal{T}(v)} \right) t^s$$

which coincides with the definition of $H_{p,q}(z,t,g_1,g_2)$. Like in [BW19], each monomial has non negative integer matrix J with support on a chain as exponents. For p by q matrices, this chain terminates at or before $x_{p,q}^{J_{p,q}}$. To factor in the cost of moving an element from position p to q in a distribution, the indeterminate z has to be multiplied with $z^{|p-q|}$. Additionally, to achieve the weighting of the totals in the exponents of g_i, the monomial has to be multiplied with $g_1^{p-1} g_2^{q-1}$. So in total, each monomial is multiplied with

$$z^{|p-q|} g_1^{p-1} g_2^{q-1} t$$

and contributes $\sum_{J_{p,q}=0}^{\infty} (z^{|p-q|} g_1^{p-1} g_2^{q-1} t)^{J_{p,q}}$ to all monomials.

Like in [BW19], $x_{pj} \subset H_{p,q-1}$ for some $1 \leq j \leq q$, and $x_{iq} \subset H_{p-1,q}$ for some $1 \leq i \leq p$. Since the exponent matrix has support on a chain, the monomials cannot be counted in both polynomials H.

The sum $H_{p,q-1} + H_{p-1,q}$ counts all monomials, but if there exists a $J_{ij} > 0$ with $i < p$ and $j < q$ it is counted twice, so it has to be subtracted once by subtracting $H_{p-1,q-1}$ from the total, leaving

$$H_{p,q-1} + H_{p-1,q} - H_{p-1,q-1}$$

All monomials are counted exactly once and weighted correctly in the product:

$$\frac{H_{p-1,q}(z,t,g_1,g_2) + H_{p,q-1}(z,t,g_1,g_2) - H_{p-1,q-1}(z,tg_1,g_2)}{1 - z^{|p-q|}tg_1^{p-1}g_2^{q-1}}$$

□

For $p = q = 3$, the adjusted formula for H expanded to a series in t was computed using Mathematica:

$$\begin{aligned}
H_{3,3}(z,t,g_1,g_2) = 1 + t({g_1}^2{g_2}^2 + {g_1}^2g_2z + {g_1}^2z^2 + g_1{g_2}^2z + g_1g_2 + g_1z + {g_2}^2z^2 + g_2z + 1)+ \\
t^2({g_1}^4{g_2}^4 + {g_1}^4{g_2}^3z + 2{g_1}^4{g_2}^2z^2 + {g_1}^4g_2z^3 + {g_1}^4z^4+ \\
{g_1}^3{g_2}^4z + {g_1}^3{g_2}^3 + 2{g_1}^3{g_2}^2z + {g_1}^3g_2z^2 + {g_1}^3z^3 + 2{g_1}^2{g_2}^4z^2 + 2{g_1}^2{g_2}^3z+ \\
2{g_1}^2{g_2}^2z^2 + 2{g_1}^2{g_2}^2 + 2{g_1}^2g_2z + 2{g_1}^2z^2 + g_1{g_2}^4z^3 + g_1{g_2}^3z^2 + 2g_1{g_2}^2z+ \\
g_1g_2 + g_1z + {g_2}^4z^4 + {g_2}^3z^3 + 2{g_2}^2z^2 + g_2z + 1) + O(t^3)
\end{aligned}$$

Now, the coefficient of t^2 contains the indeterminates g_1 and g_2 as well:

$$\begin{aligned}
C(z,g_1,g_2) = g_1^4g_2^4 + g_1^4g_2^3z + 2g_1^4g_2^2z^2 + g_1^4g_2z^3 + g_1^4z^4 + g_1^3g_2^4z + g_1^3g_2^3 + 2g_1^3g_2^2z + g_1^3g_2z^2+ \\
g_1^3z^3 + 2g_1^2g_2^4z^2 + 2g_1^2g_2^3z + 2g_1^2g_2^2z^2 + 2g_1^2g_2^2 + 2g_1^2g_2z + 2g_1^2z^2+ \\
g_1g_2^4z^3 + g_1g_2^3z^2 + 2g_1g_2^2z + g_1g_2 + g_1z + g_2^4z^4 + g_2^3z^3 + 2g_2^2z^2 + g_2z + 1
\end{aligned}$$

Each of the monomials in C has the structure $ng_1^ig_2^jz^k$, which encodes the number n of composition pairs with an EMD of k. However, in this case the compositions are restricted by i and j, which specify the value of $\mathcal{T}$ of the compositions counted. Specifically, g_1^i means, that $\mathcal{T}(\mu) = i$ must apply to the composition μ considered in the exponenet of g_1.

Given this information, in order to examine the EMD of compositions with a fixed value of $\mathcal{T}$, the coefficient of not only t^2, but of $g_1^i g_2^j t^2$ has to be copmuted. Let $i = j = 2$, this results again in a polynomial of z:

$$P(z) = 2z^2 + 2$$

Which means that there are 2 pairs of compositions with weighted total of 2 and a distance of 2, and there are 2 pairs of compositions with a weighted total of 0 and a distance of 0. As seen in Table II.1, the first monomial refers to the pairs $\{(1,0,1),(0,2,0)\}$ and $\{(0,2,0),(1,0,1)\}$, the second monomial to the two pairs $\{(1,0,1),(1,0,1)\}$ and $\{(0,2,0),(0,2,0)\}$.

III.1 Theoretical Mean EMD Example

To get the theoretical average EMD of classes with 30 students, where only the grades A through F (no +/-) are given out, the polynomial $H_{5,5}(t,z,g_1,g_2)$ has to be computed and expanded in t to degree 30. Because this way of computing the polynomial requires more resources than available, the polynomial will be computed in a way similar to what was briefly shown in Section II.1. Instead of finding the entire polynomial $H_{5,5}(t,z,g_1,g_2)$, the weak compositions of 30 with 5 elements were computed in Python, see Listing V.4. From all the compositions, only those with a weighted total value $\mathcal{T}$ of 90 were considered (corresponding to a B or 3.0 average grade). The required polynomial in z was then computed by counting the number n of pairs with distance i, put together to the monomials nz^i.

$$P(z) = 297z^0 + 2480z^2 + 6398z^4 + 9534z^6 + 11386z^8 + 11272z^{10} + 10412z^{12} + 8676z^{14} + 7220z^{16} + 5562z^{18} + 4372z^{20} + 3184z^{22} + 2408z^{24} + 1684z^{26} + 1218z^{28} + 820z^{30} + 552z^{32} + 348z^{34} + 206z^{36} + 108z^{38} + 50z^{40} + 18z^{42}$$

Examining the structure of $P(z)$, it can be seen that the coefficients of z sum up to the number of all pairs examined. Since the sum of these coefficients is the same as $P(1)$, the number of pairs can be acquired by computing that: $P(1) = 88205$. See Figure III.1 for a histogram showing the distributions of EMDs between all the possible compositions. Additionally, since the EMD encoded in the exponent k of nz^k is the distance between all the n distribution pairs, it follows that the coefficients n of the derivative of P with respect to z, P', sum up to a total that weighs the number of pairs by the distance between each of their elements.

This implies, that to compute the average EMD between all possible distributions consisting of 30 grades that sum up to 90, it suffices to divide the weighted sum of the numbers of pairs $P'(1)$ by the total number of pairs $P(1)$, which results in:

$$\frac{P'(1)}{P(1)} = \frac{1115148}{88205} \approx 12.64268$$

To compare this result to [BW19], it is necessary to compute the unit normalized result for the mean EMD. To achieve that, the value $\frac{P'(z)}{P(z)}$ has to be divided by the maximum possible EMD. Regarding compositions of 30 with 5 elements, the highest distance is found between the distributions $[30, 0, 0, 0, 0]$ and $[0, 0, 0, 0, 30]$, which have an EMD of 120. Therefore, the unit normalized mean EMD of classes with 30 students and 5 grades is given by:

$$\frac{P'(z)}{120P(z)} = \frac{1115148}{120 \cdot 88205} \approx 0.10536$$

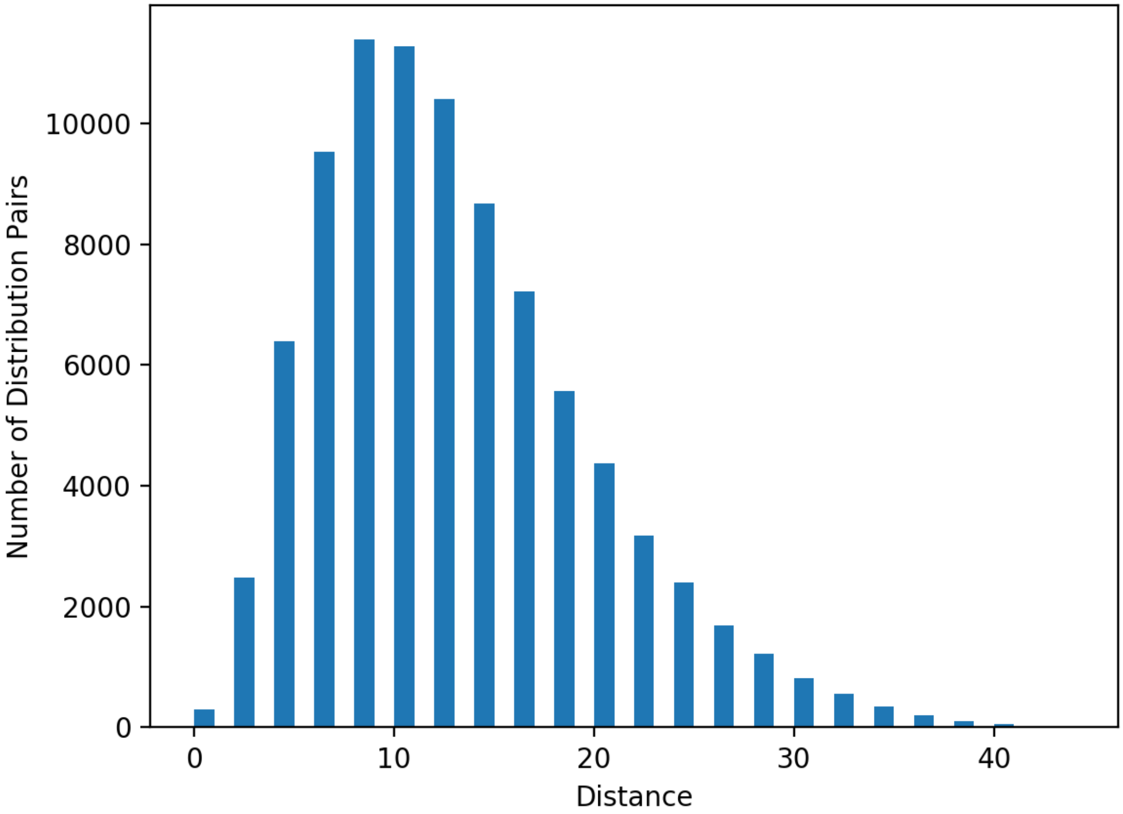

Figure III.1: Histogram of the distribution of EMDs between all possible compositions

The unit normalized EMD for classes of 30 students without grade restrictions has a value of 0.2191 and was computed in [BW19]. One can see, that the theoretical mean EMD between classes of 30 students is almost exactly twice as high when there are no grade restrictions, compared to when the GPA is restricted to be a 3.0.

Real Grade Data

In this chapter, the theoretical results of the mean EMD with a fixed GPA are compared to a real world dataset, coming from the Section Attrition and Grade Report published by the Office of Assessment and Institutional Research at the University of Wisconsin-Milwaukee [oAR20]. It contains the grade distributions of classes in the fall semesters from 2014 to 2018. In the last chapter, we computed the theoretical mean EMD for grade distributions of classes with 30 students, average GPA of 3.0, where only five grades given out, which corresponds to the letter grades A through F without plus or minus.

To compare the theoretical result to the real world data, the dataset has to be subjected to similar restrictions, without shrinking so much in size to become insubstantial.

IV.1 Examining Collected Grade Data

The dataset for the fall semester of 2018 contains data from about 3300 grade distributions. Restricting the data to only classes with exactly 30 students and an exact GPA of 3.0 leaves fewer than 10 results, so the limitations were broadened to 25 - 35 students with a GPA between 2.9 and 3.1.

Applying these restrictions leaves 71 classes to be further examined. Table IV.1 shows the first 10 entries of the data for 2018.

The data was examined using Python, and the EMD of years 2014 through 2018 can be seen in table IV.2.

In almost all the years examined, the mean EMD is always more than 20% higher than the

	Cla Subject Ldesc	Class	Enrollment	GPA	A	B	C	D	F
0	Music	127	35	2.98	11	14	7	1	1
1	Business Administration	335	30	3.047	12	9	3	3	1
2	Business Administration	404	34	3.019	11	14	8	0	1
3	Business Administration	404	25	2.987	6	13	5	0	0
4	Business Administration	409	31	2.956	5	20	4	1	1
5	Business Administration	409	34	2.961	3	27	4	0	0
6	Business Administration	451	29	2.913	3	20	2	2	0
7	Business Administration	453	25	2.988	11	9	2	2	1
8	Business Administration	454	27	3.013	7	11	7	0	0
9	Business Administration	551	34	2.921	11	12	8	3	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Table IV.1: First entries of the Fall 2018 data, restricted to classes with 25-35 students and a GPA between 2.9 and 3.1

	2014	2015	2016	2017	2018	Theory
EMD	16.2634	15.8075	13.1865	15.4164	16.3264	12.6427
Difference	28.6390%	25.0329%	4.3015%	21.9396%	29.1371%	
Classes	79	63	84	77	72	

Table IV.2: Mean EMD for the Fall Semesters 2014 through 2018, compared to the theoretical result

theoretical result, with the largest striking differences recorded in 2014 and 2018 with EMD values that are about 28.6% and 29.1% larger. At least part of the difference is accounted for by the varying class sizes that had to be considered in the real world data. For example if the two classes compared have 25 and 35 students, the difference in the number of students for a pair of classes is 10, which adds a value of 10 to the absolute EMD between this pair. The only exception is the year of 2016, where the mean EMD was only 4.3% larger than the theoretical result.

IV.2 Observations

To further examine the given grade data, it can be represented as a graph, which contains each class as a vertex. Let t be a threshold value for the EMD and define two vertices to be connected by an edge, if the EMD of two classes is less than or equal to t.

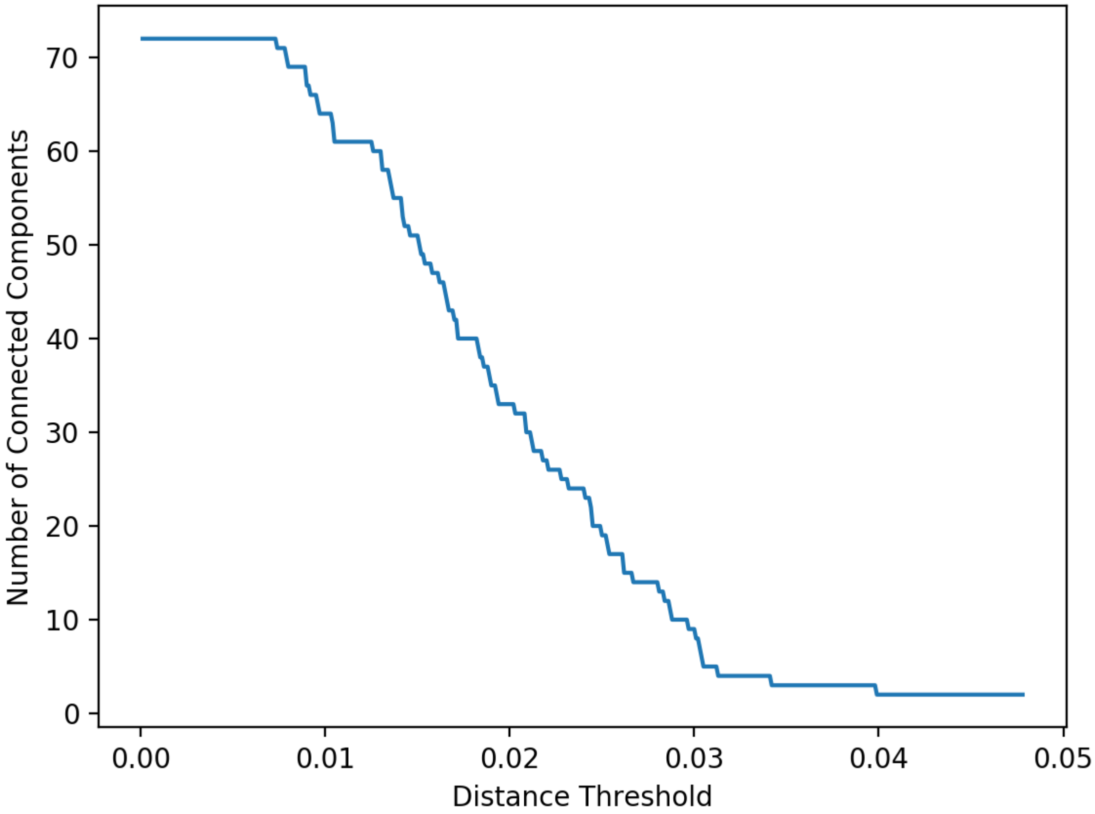

Figure IV.2: Distance threshold versus number of connected components

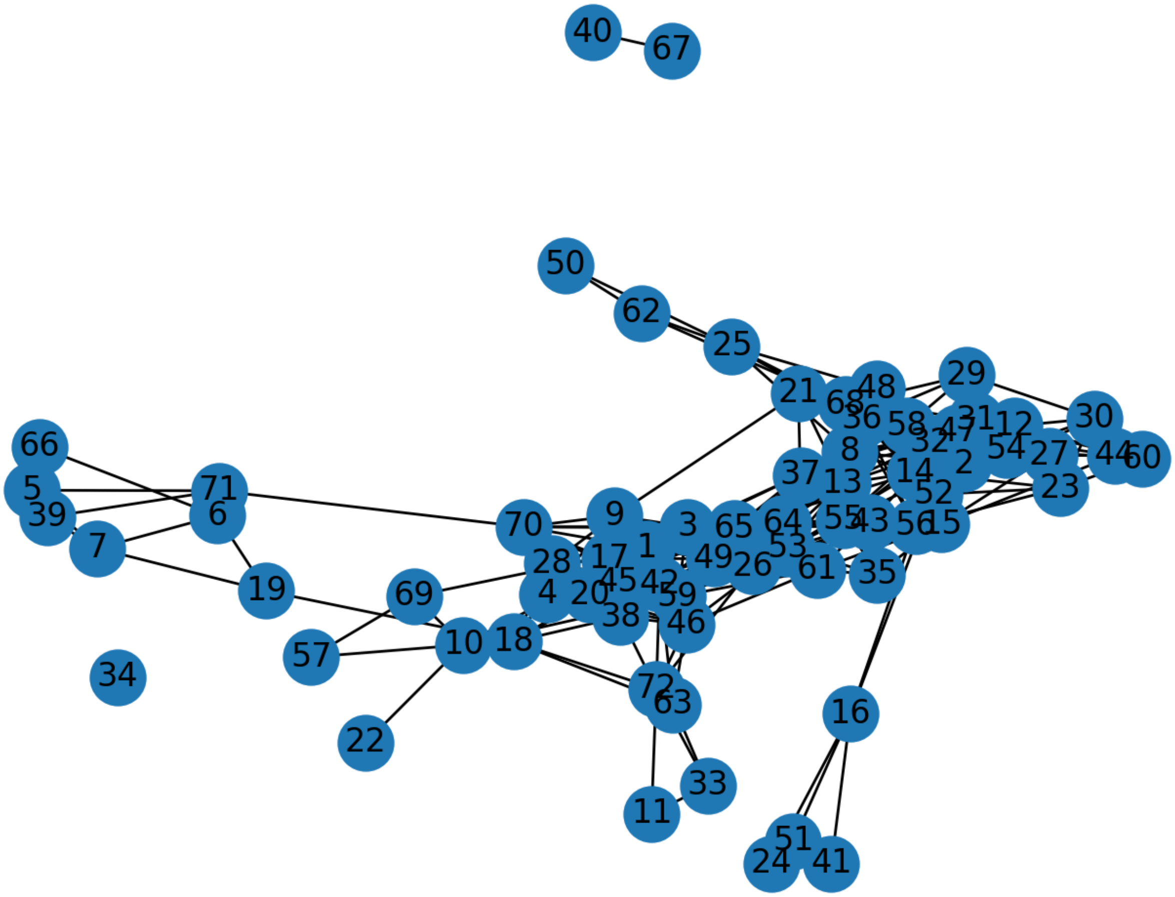

Figure IV.3: Cluster with threshold 0.035

A connected component, per definition, is a set of vertices in a graph that are connected by a walk [POM09]. Figures IV.2-IV.5 refer to the dataset of Fall 2018, and shows the number of connected components as the threshold t increases. With threshold $t = 0$, every pair of vertices is disconnected and builds an individual connected component. As the threshold increases, there is an increasing number of edges, up to a point where the entire graph builds one connected component, at around $t = 0.05$.

Figure IV.3 shows the graph when $t = 0.035$, which is about half of the unit normalized mean EMD of Fall 2018, shown in table IV.2. Still, there are only 3 connected components left, with one big component containing every vertex except for $\{40, 67\}$ and $\{34\}$. For more information on the class each vertex represents, see table F.1 in chapter IV.2.

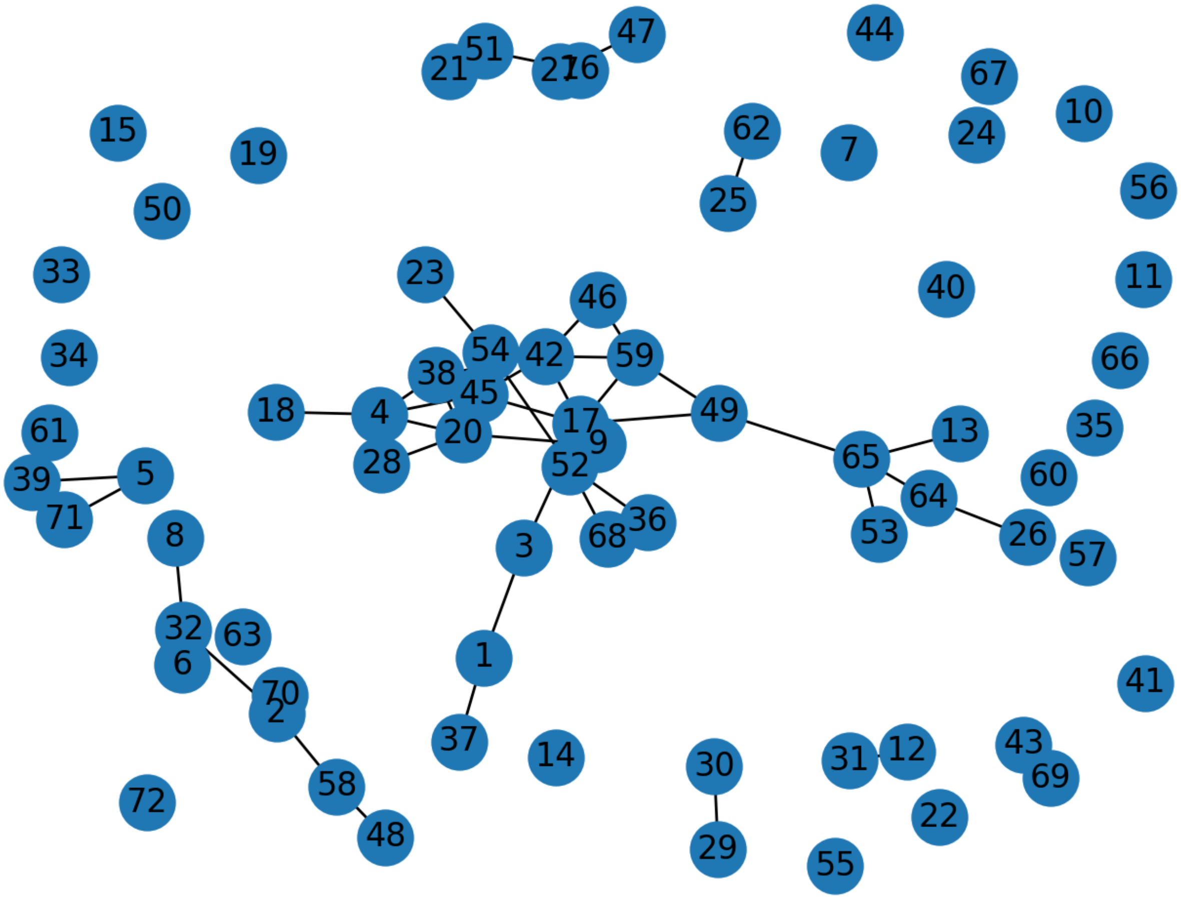

Figure IV.4: Cluster with threshold 0.0185

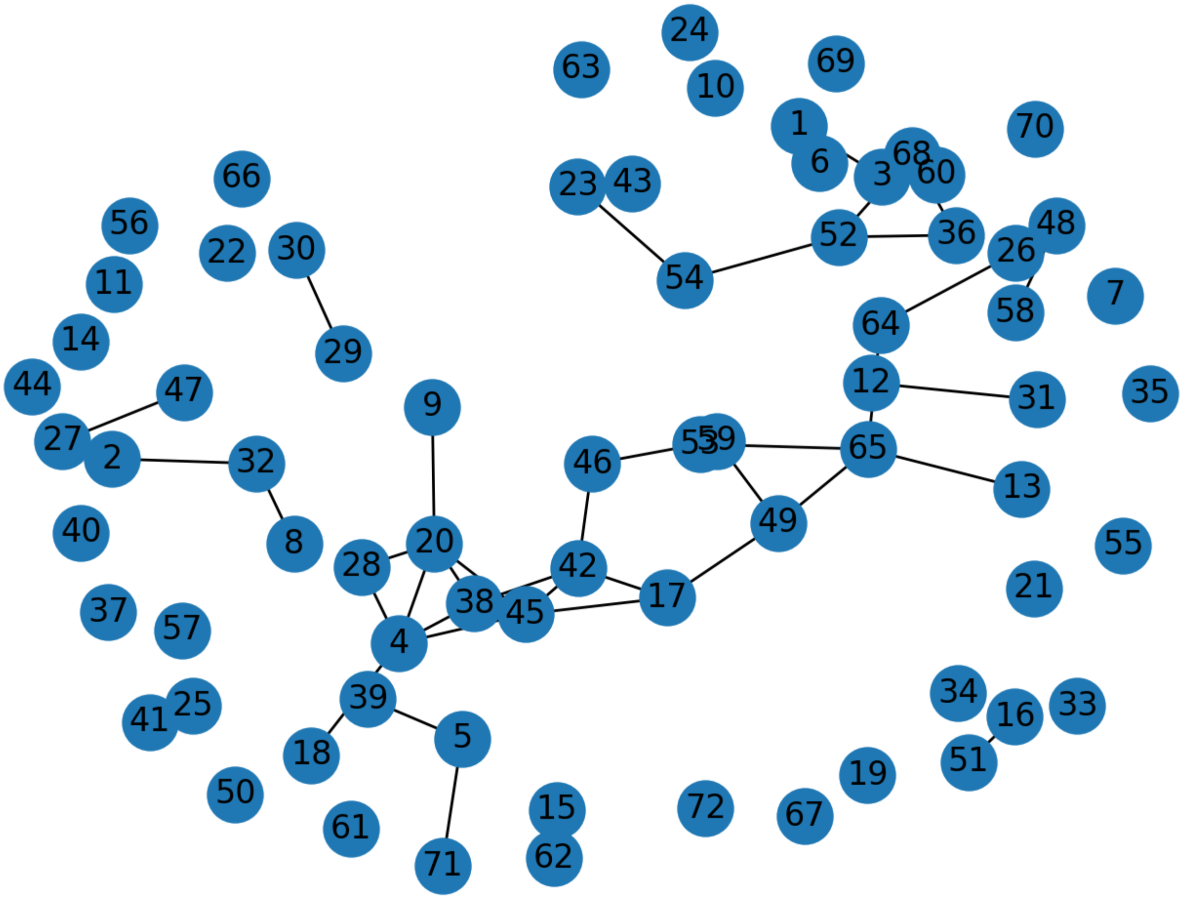

Figure IV.5: Cluster with threshold 0.017

Looking at Figure IV.2, there is a striking persistence in the number of connected components when $t \in [0.017, 0.0185]$. Figures IV.4 and IV.5 show graphs with threshold values at boths ends of the interval, and it can be seen that some of the connected components in Figure IV.4 are split apart in Figure IV.5. For instance, while $\{2, 8, 32, 48, 58\}$ form a connected component in Figure IV.4, they are broken apart into the components $\{2, 8, 32\}$ and $\{48, 58\}$ in Figure IV.5.

Ranges of persistence, for example $[0.017, 0.0185]$, can be compared to the theoretical mean EMD. The lower endpoint of the interval is 16%, the upper is approximately 18% of the theoretical mean EMD.

In general, Figures IV.3-IV.5 show the breaking apart of connected components with decreasing distance threshold. The clustering allows to examine the most persistent components among the graph with varying threshold. The two vertices with identifiers 66 and 39 built a single connected component, which turned out to be the most persistent one. For every $t \in [0, 0.0229]$, the component consisted of only those two vertices, which represent the classes *Nursing 673* and *English 205* respectively, see Table F.1 in the Appendix.

Data Science Approach

The dataset was examined in Python, using the library `pandas`. Among others, `pandas` includes functions to read the dataset from the given comma-separated-value (csv) format into a table, called a dataframe.

Dataframes consist of columns, that can be named with strings, and indexed rows, as visible in table IV.1. `pandas` includes Create-Read-Update-Delete (CRUD) operations for dataframes.

CRUD refers to the major functions of relational database management systems, and corresponding operations are also provided by the Structured Query Language (SQL, see [DD93]) or the Hypertext Transfer Protocol (HTTP, see [FGM+99]).

The accordance of `pandas` with the CRUD principles allows for efficient filtering and extracting of relevant parts of the dataset. An example for these functions can be seen in listing V.1. The data set is loaded into a dataframe in line 3.

The first filtering operation is seen in line 6, where every row in the dataframe, that does not have a value between 25 and 35 in the column `Enrollment`, is removed. That is achieved by using the function `loc[]`, which returns a set of all rows that match the given condition, which is in this case: $|e - 30| < 5$, for all e entries of the column `Enrollment`. After the data is filtered for all the restrictions, it has to be brought into the format necessary for the comparison to the theoretical result from chapter 2. Only the 5 grades `A` to `F`, without plus or minus, were considered in the theoretical approach. To get the corresponding format with the given data, all the plus and minus grades were counted as their base grade. `pandas` supports accessing entire columns by their name, as seen in lines 9 to 11, which merge the grade columns accordingly. The last line of listing V.1 extracts only the necessary grade columns from the data, that can then be used to calculate the EMD.

```
import pandas as pd

# Load data into pandas dataframe
data = pd.read_csv(’data.csv’)

# Filtering for classes with 25-35 students
data = data.loc[abs(data[’Enrollment’] - 30) <= 5]

# Merge plus/minus grades and base grade, eg. B+, B and B- all get
# counted as B
for letter in ’BCD’:
  data[letter] = data[letter+’+’] + data[letter] + data[letter+’-’]
data[’A’] = data[’A’] + data[’A-’]
data[’F’] = data[’F, F+’]

# Extract only grade information from dataset
data = data[[’A’, ’B’, ’C’, ’D’, ’F’]]

```

Listing V.1: Examples for CRUD operations in Python using `pandas`

```python
def EMD(dist1, dist2):
  dif = dist1-dist2
  result = 0

  for i in range(len(dif)):
    # Sum difference in each iteration to account for the cost
    # of moving an element further than one row/column
    result += abs(np.sum(dif[:i]))
  return result

```

Listing V.2: Python code to compute the absolute EMD of two distributions

```python
def build_distance_matrix(grades):
  # Initializing an empty matrix to save the time
  # needed to e.g. initialize everything with 0
  distance_matrix = np.empty((len(grades),
    len(grades)))
  for i in range(len(grades)):
    # Since matrix was initialized with "random" values,
    # the diagonal elements have to be set to zero here
    distance_matrix[i, i] = 0
    for j in range(i+1, len(grades)):
      # Distance Matrix is symmetric, so entry ij=ji
      distance_matrix[i, j] =
        distance_matrix[j, i] =
        EMD(grades.iloc[i].to_numpy(
        dtype=np.float64),
        grades.iloc[j].to_numpy(dtype=
        np.float64))
  return distance_matrix

```

Listing V.3: Python code to compute the distance matrix of a set of grade distributions, given as a pandas Series object

```python
import scipy.special

def rec_compositions(n, k, current, all_comps):
  # Save composition if it sums to the right number
  # and has the correct length
  if sum(current) == n and len(current) == k:
    all_comps.append(current)
  # If not, start new recursive step with every possible
  # number appended to the composition
  for i in range(n-sum(current)+1):
    if len(current) < k and (current+[i]) not in all_comps:
      rec_compositions(n, k, current + [i], all_comps)
  return all_comps


  def compositions(n, k):
    comps = rec_compositions(n, k, [], [])
    # The number of weak compositions of n with k elements
    # is known, so it can be checked here
    assert len(comps) == scipy.special.binom(n+k-1, k-1)
    return comps

```

Listing V.4: Python code to compute the distance matrix of a set of grade distributions, given as a pandas Series object.

Bibliography

[BW19] Rebecca Bourn and Jeb F Willenbring. Expected value of the one dimensional earth mover's distance. *arXiv preprint arXiv:1903.03673*, 2019.

[DD93] Christopher J Date and Hugh Darwen. Iso/iec 9075-2: 2008 (sql-part 2: Foundations), the sql standard, 1993.

[FGM+99] Roy Fielding, Jim Gettys, Jeffrey Mogul, Henrik Frystyk, Larry Masinter, Paul Leach, and Tim Berners-Lee. Hypertext transfer protocol–http/1.1. 1999.

[Lan03] Sergei K Lando. *Lectures on generating functions*, volume 23. American Mathematical Soc., 2003.

[oAR20] Office of Assessment and Institutional Research. Section attrition and grade reports. ȟttps://uwm.edu/datahub/reports/section-attrition-and-grade-reports/, 2020. Last visited on 3/17/2020.

[POM09] Mason A Porter, Jukka-Pekka Onnela, and Peter J Mucha. Communities in networks. *Notices of the AMS*, 56(9):1082–1097, 2009.

Appendix Fall 2018 Grade Dataset

Table F.1: Complete Fall 2018 grade dataset, restricted to classes with 25-35 students and 2.9-3.1 GPA

	Subject	Class	Enrollment	GPA	A	B	C	D	F
1	Music	127	35	2.98	11	14	7	1	1
2	Business Administration	335	30	347	12	9	3	3	1
3	Business Administration	404	34	319	11	14	8	0	1
4	Business Administration	404	25	2.987	6	13	5	0	0
5	Business Administration	409	31	2.956	5	20	4	1	1
6	Business Administration	409	34	2.961	3	27	4	0	0
7	Business Administration	451	29	2.913	3	20	2	2	0
8	Business Administration	453	25	2.988	11	9	2	2	1
9	Business Administration	454	27	313	7	11	7	0	0
10	Business Administration	551	34	2.921	11	12	8	3	0
11	Business Administration	600	35	391	9	22	0	1	1
12	Business Administration	703	25	2.986	10	6	6	0	1
13	Business Management	705	27	387	11	10	5	0	1
14	Curriculum and Instruction	112	30	387	12	9	4	1	1
15	Curriculum and Instruction	301	30	2.954	10	13	3	1	2
16	Curriculum and Instruction	650	30	2.953	15	5	4	1	3
17	Educational Psychology	330	35	31	10	15	7	1	0
18	Exceptional Education	303	26	354	6	15	4	0	0
19	Exceptional Education	330	31	345	4	20	5	0	0
20	Civil & Envrnmntal Engineering	456	26	338	7	13	6	0	0
21	Industrial/Manufacturing Engr	583	25	38	10	8	7	0	0
22	Mechanical Engineering	469	30	2.918	7	12	7	3	0
23	Commun Sciences & Disorders	380	29	336	15	6	6	1	1
24	Kinesiology	200	29	2.947	10	4	1	2	2
25	Information Studies	310	35	359	16	12	2	1	3
26	Information Studies	370	26	342	8	12	3	0	1
27	African & African Diaspora St	125	25	2.95	8	7	3	0	2
28	Anthropology	403	28	2.975	6	14	6	0	0
29	Art History	250	33	2.92	12	5	5	1	2
30	Art History	472	34	2.951	13	5	7	0	2
31	Biological Sciences	529	28	34	11	7	6	0	1

32	Chemistry and Biochemistry	341	33	311	14	10	4	3	1
33	Communication	101	27	3	6	13	1	1	1
34	Communication	363	25	397	10	7	4	0	0
35	Communication	410	25	2.917	7	9	1	2	1
36	Economics	210	26	2.937	13	7	3	1	2
37	Economics	325	35	2.918	12	11	7	1	1
38	English	205	26	354	7	13	5	0	0
39	English	205	26	2.903	4	16	2	1	1
40	English	205	26	2.957	14	4	0	0	5
41	English	205	25	2.914	12	6	1	1	3
42	English	215	26	3	7	12	4	1	0
43	English	215	25	344	9	10	2	1	1
44	English	233	25	2.954	8	9	2	0	3
45	English	310	25	33	7	11	5	0	0
46	English	517	25	2.931	7	12	3	2	0
47	Food & Beverage Studies	101	25	2.934	10	9	3	1	2
48	Geosciences	106	31	313	12	8	2	2	2
49	Linguistics	210	26	326	9	11	5	1	0
50	Linguistics	210	25	398	11	10	0	1	2
51	Mathematical Sciences	98	28	3	16	4	4	1	3
52	Mathematical Sciences	98	30	345	15	7	5	1	2
53	Mathematical Sciences	98	29	336	10	12	4	1	1
54	Mathematical Sciences	105	28	339	13	6	5	1	1
55	Mathematical Sciences	105	28	2.936	9	10	4	2	1
56	Mathematical Sciences	105	27	387	12	5	3	2	1
57	Mathematical Sciences	108	25	2.933	9	7	6	3	0
58	Mathematical Sciences	232	32	349	12	8	3	2	2
59	Mathematical Sciences	233	32	311	10	14	5	2	0
60	Philosophy	101	29	359	11	5	6	1	0
61	Philosophy	243	25	335	7	8	2	2	0
62	Philosophy	250	30	317	10	8	1	0	2
63	Political Science	361	33	311	8	17	2	3	0
64	Sociology	361	33	392	10	14	4	0	1
65	Women's and Gender Studies	201	35	343	11	13	6	0	1
66	Nursing	673	28	348	5	21	2	0	0
67	Criminal Justice	105	28	2.988	18	3	0	1	5
68	Criminal Justice	460	32	2.969	15	9	4	1	2
69	Criminal Justice	662	28	2.904	9	9	8	1	1
70	Social Work	753	31	2.964	7	14	6	0	1
71	Architecture	380	27	2.987	5	16	4	0	1
72	Urban Planning	316	28	343	7	14	2	1	0