Identifying Hubs in Undergraduate Course Networks Based on Scaled Co-Enrollments

Gary M. Weiss, Nam Nguyen, Karla Dominguez and Daniel D. Leeds
Department of Computer and Information Science
Fordham University, New York, NY
{gaweiss, nnguyen56, kdominguezmelo, dleeds}@fordham.edu

ABSTRACT

This study uses eight years of undergraduate course enrollment data from a major university to form networks of courses based on student co-enrollments. The networks are analyzed to identify "hub" courses often taken with many other courses. Two notions of hubs are considered: one based on raw popularity and another on proportional likelihoods of co-enrollment with other courses. Network metrics are calculated to describe the course networks. Academic departments and high-level academic categories (e.g., humanities), are studied for their influence over course groupings. The identification of hub courses has practical applications, since it can help better predict the impact of changes in course offerings and in course popularity, and in the case of interdisciplinary hub courses, can be used to increase or decrease interest and enrollments in specific academic departments and areas.

Keywords

Graph mining, network analysis, educational data mining.

1. INTRODUCTION

Universities typically offer thousands of different courses across dozens of departments. The interrelationships between courses that are taken together, especially those in different departments, is often not well understood. This paper addresses this deficiency by forming course networks, connecting courses often taken by the same students. Each course is represented as a node in the graph. Several network analyses are pursued. This work also studies “hub” courses, defined as network nodes that are connected to many other nodes, resulting in a high degree count [2]. This study utilizes three popular centrality metrics to identify course hubs and compares the results when using each metric.

Network analyses utilized in this paper have been proven useful to other domains. Analysis of social networks like Facebook identify hubs corresponding to influencers with an outsized impact on other users’ purchasing behaviors [3]. Network analysis metrics pursued in the present work have been applied to the World Wide Web, particularly for web searches [5, 8, 9].

Identifying and analyzing hub courses can provide concrete benefits. Courses heavily associated with other courses can be used for better resource planning, particularly when changes are made in the frequency or capacity of such courses. Furthermore, hub courses may be adjusted to drive (or diminish) student interest in an area or academic discipline. For example, there is a current need for more STEM (Science, Technology, Engineering, and Math) professionals. If a hub course is well connected to STEM courses, promoting this course may lead to increased STEM enrollments—even if the hub course is not a STEM course.

The course network analyzed in this study is based on eight years of undergraduate student course enrollment data from Fordham University. An edge connects two courses if the number of students taking both courses is above a threshold. Two types of thresholding mechanisms are considered: (1) a static threshold that is the same for all pairs of courses and (2) a dynamic threshold set to link together only courses taken together *relatively* frequently (i.e., relative to their popularity). We find the dynamic threshold shifts hub courses from humanities to STEM disciplines. Also, tighter course groupings are found within STEM and looser groupings within the humanities and social sciences. An extended version of this paper is available [12].

2. DATASET DESCRIPTION

Our study uses course enrollment data to generate a course-pair dataset, which is then used to form the course networks analyzed in this paper. This course enrollment data contains eight years of undergraduate data from Fordham University, where each record corresponds to one student in one course section. Student grades are also available and used in two of our other studies, one of which analyzes the impact of course sequencing on student grades [4], and the other that forms course networks based on the correlation of grades between courses, and then analyzes the networks [6]. This later study performs a somewhat similar analysis to the one provided in this paper, but with a very different notion of course similarity/linkage.

The course-pair dataset aggregates the course enrollment data to the course level and then extracts information about each course pair. Each course-pair record includes identifying information about two courses and the number of students that took each course and both courses (not necessarily at the same time). The department associated with each course is mapped to one of the six major course categories. The course-pair dataset contains 78,173 records, which are formed from 1,763 distinct courses. The dataset does not contain all possible pairings because pairs with fewer than 20 common students are excluded. The course-pair dataset, and the network metrics provided later, are generated from the course enrollment data using a publicly available Python-based software tool developed by our research group [10].

3. NETWORK ANALYSIS METRICS

The course-pair dataset is used to form course networks by viewing each course as a node and connecting nodes that have a sufficient number of common students. Table 1 provides the network analysis metrics used in this paper. The first three, density, diameter, and average clustering coefficient (ACC) [1], are computed using an entire network or subnetwork. Our data shows subnetworks of courses within single departments have a higher density, smaller diameter, and higher average clustering coefficient than the network based on all undergraduate courses, because courses within a discipline are more tightly connected (see Table 2). The last three metrics are defined for each *node* in the network and can be used to help identify hubs. These metrics consist of three centrality measures: degree centrality, eigenvector centrality [11], and betweenness centrality [7]. Each measure can be used to identify a different type of hub course.

Table 1. Summary of network analysis metrics

Metric	Summary Description	Range
Density	Fraction of possible edges present.	0 - 1
Diameter	Maximum distance between any pair of nodes in network.	$\mathbf{Z}^+$
Ave. Clustering Coefficient	Fraction of pairs of neighbor nodes that are connected to each other.	0 - 1
Degree Centrality	Number of edges to node (degree).	$\mathbf{Z}^+$
Eigenvector centrality	Based on centrality of node's neighbors.	≥ 0
Betweenness centrality	Measure all shortest paths passing through node.	≥ 0

4. EDGE INCLUSION METHODOLOGY

To form a course network, each course is represented by a node, and an edge is added between two nodes if the courses, across all sections, have enough common students. Static and dynamic thresholds specify a minimum number of common students.

The static threshold is based on the number of common students between two courses, independent of how many students take each course. The distribution of common students by course pair is provided in Figure 1 in the appendix. Most course-pairs have very few common students, since few students take upper-level courses in disparate disciplines. A threshold of 20 students maintains 11% of all course-pairs with at least one student in common, and this is the static threshold utilized in this study. The static threshold is heavily biased towards popular courses, taken very frequently, even if only a few students in the popular course take specific other courses.

We also define a dynamic threshold relying primarily on the *co-occurrence rate* of courses. The dynamic threshold is determined by multiplying the co-occurrence threshold rate *k* by the number of students in the larger course within each course-pair. To ensure a minimum number of common students, a static threshold of 20 students is used as the floor for the dynamic threshold. The dynamic threshold, *d-thresh*, associated with two courses, C_1 and C_2, is provided in Equation 1, where C_x.students represents the number of students who have taken class C_x.

$$\text{d-thresh}(C_1, C_2) = \max(20, k \times \max(C_1.\text{students}, C_2.\text{students})) \quad [1]$$

The dynamic threshold is heavily dependent on the co-occurrence rate *k*, defined as the number of common students divided by the number of students in the larger course. The co-occurrence rate distribution is displayed in Figure 2 of the appendix, which shows that a co-occurrence rate threshold $k = 0.017$ discards 39% of the edges that satisfy the static threshold. This threshold is used because it leads to the most stable centrality measures while excluding the fewest number of edges. Table 5 in the appendix shows how this dynamic threshold impacts an Art History course.

5. RESULTS

This section analyzes course networks using the metrics presented in Table 1 and through the identification of hub courses. Static and dynamic thresholds are considered. Hub results are analyzed within academic departments and broader course categories. This study utilizes six course categories: Arts, Communication and Media Studies, Humanities, Modern Languages, Social Sciences, and STEM. The mapping from academic department to course category is partially provided in Table 2.

5.1 Network Metric Results

Table 2 presents the values of the previously defined network metrics for the course network and subnetworks at the department and category levels. Course categories are denoted in bold, with a subset of two selected associated departments listed below it (see [12] for the full table). The category level value reflects the median values across the member departments. The first row of data provides the values over all courses in the course network. The color of the cells reflects the magnitude of the cell value, with red (green) used for the highest (lowest) values. The colors for the departments and categories are determined independently.

The network covering all courses has a high diameter and low density compared to the subnetworks, since it includes many diverse courses that are loosely connected. Courses associated with a specific department are typically associated with a major; students within the major will take many of these courses. The dynamic threshold decreases the density, average clustering coefficient, and number of edges, while increasing the diameter.

Study of departmental subnetworks shows dynamic thresholding most dramatically decreases edges for Philosophy (52% decrease), English (44% decrease), and Theology (35% decrease), which are fields of study that include many core curriculum courses. This drop is mirrored by ACC. Conversely, the diameter maintains similar values for most departments, regardless of threshold. Overall, dynamic thresholding has a substantial impact on density and ACC of Humanities and Social Science courses, and only minimal impact on other categories, likely reflecting the core curriculum's emphasis on humanities and social science courses.

The STEM courses have much higher density and form much more dense clusters (based on ACC) than humanities courses, for both thresholds. This indicates that humanities students are less likely to take the same group of courses in their discipline. In our university, humanities majors have fewer required courses than STEM majors. Humanities departments have the highest number of nodes (distinct courses taken), closely followed by Social Science, suggesting that those disciplines allow more flexibility in course choices. The Modern Languages category also has a relatively high density and ACC. Language courses, like science courses, typically rely on prerequisite course requirements for proper student preparation.

5.2 Hub Analysis

Hubs play a special role in network structures and play an important role in understanding and utilizing the information in course co-enrollment networks. Table 3 identifies the top-17 hubs using the median of the ranks of the three centrality metrics, "Combined Rank". The top half of the table provides the top-7 hubs when using the static threshold, while the bottom half provides the top-7 for the dynamic threshold. Note that the best combined rank when using the dynamic threshold is 3—no course consistently ranks above third on all the centrality metrics. While only the combined rank for the static (dynamic) threshold is used

Table 2. Summary course network statistics based on category and selected departments

Category/ Department	Nodes	Static Threshold				Dynamic Threshold			
		Edges	Density	Diam.	ACC	Edges	Density	Diam.	ACC
ALL	1763	39968	0.03	4	0.74	24323	0.02	6	0.40
Arts	**41.5**	**239**	**0.32**	**3**	**0.56**	**231**	**0.29**	**2.5**	**0.56**
Dance	54	1236	0.86	3	0.95	1236	0.86	3	0.95
Music	24	87	0.32	3	0.51	73	0.26	2	0.52
Comm and Media Studies	**24**	**25**	**0.20**	**2**	**0.16**	**25**	**0.19**	**2**	**0.16**
Comm and Media Studies	94	862	0.20	3	0.72	828	0.19	4	0.58
New Media & Digital Design	6	8	0.53	2	0.00	8	0.53	2	0.00
Humanities	**81**	**179**	**0.06**	**3**	**0.21**	**104**	**0.04**	**3**	**0.08**
African & African Amer Studies	28	34	0.09	2	0.11	34	0.09	2	0.11
English	167	462	0.03	3	0.59	258	0.02	3	0.12
Modern Languages	**9**	**19**	**0.53**	**2**	**0.49**	**19**	**0.53**	**2**	**0.38**
Greek	4	6	1.00	2	0.00	6	1.00	2	0.00
Spanish	40	118	0.15	3	0.49	98	0.13	2	0.33
STEM	**34**	**295**	**0.47**	**3**	**0.76**	**288**	**0.45**	**3**	**0.75**
Biological Sciences	30	274	0.63	2	0.77	274	0.63	2	0.77
Physics	38	286	0.41	4	0.73	277	0.39	3	0.74
Social Science	**74**	**329**	**0.18**	**2.5**	**0.51**	**285**	**0.16**	**2.5**	**0.44**
Economics	45	325	0.33	2	0.64	270	0.27	2	0.59
Sociology	90	236	0.06	3	0.37	206	0.05	3	0.30

to select the entries in the top (bottom) half of the table, both combined ranks are provided to help compare differences between the thresholding mechanisms. Courses exhibit very different ranks for the two thresholds.

The first few entries for the static threshold in Table 3 vary only slightly depending on which of the three centrality metrics is used. The first four entries cover core curriculum requirements that can only be satisfied by a single course. Most of the remaining top hub courses also satisfy a core requirement, but can be satisfied by several courses. The very few STEM courses listed are introductory and satisfy a core requirement (e.g., *Finite Mathematics*). Thus, we see that hubs identified using the static threshold are based on raw popularity. Most courses identified using the dynamic threshold also satisfy a core requirement, but often many courses can satisfy the requirement. There are no courses that appear in the top-7 lists for both thresholds. For static threshold hubs, most connections to other courses may be incidental, due to so many students taking the popular course.

Table 3. Top-7 static and dynamic course hubs

Courses	Combined Rank		Centrality Rank		
	Static	Dyn.	Deg.	Btw.	Eig.
Static Threshold: Top Hubs					
Philosophical Ethics	1	45	1	1	2
Faith & Critical Reason	2	76	2	2	1
Philos. of Human Nature	3	75	3	3	3
Composition II	4	78	4	5	4
Banned Books	5	49	5	4	5
Finite Mathematics	7	56	6	7	7
Spanish Lang and Lit	7	29	7	6	8
Dynamic Threshold: Top Hubs					
Biopsychology	31	3	3	20	3
Phys. Sci.: Today's World	30	4	4	2	63
Latin American History	44	5	2	6	5
Intro World Art History	22	5	5	26	2
Intro Phys. Anthropol.	41	6	1	9	6
Intro Cultural Anthropol.	18	6	6	33	1
Films of Moral Struggle	55	8	8	7	54

Table 3 allows further comparisons among the centrality metrics. When using the static threshold, course ranks are quite consistent across all the centrality metrics. This ensures that the combined rank is also highly correlated with each of the individual metrics, and that the degree centrality is usually equal to the combined rank. This correlation is weaker when examining the dynamic threshold; degree centrality sometimes differs substantially from the combined dynamic rank; *Calculus II* has degree 7 and combined rank 19. Nonetheless, degree centrality is still generally close to combined rank and is identical in 5 of the first 7 cases.

We focus on degree centrality as our metric for identifying hubs under both thresholds. This is attractive since degree centrality is the simplest and most common metric for identifying hubs. We utilize a degree count threshold of 200 to identify hub courses. This retains all entries in Table 3, which have degree count of at least 245 [12]; the underlying data ensures that a degree count of 200 will retain the top fifty courses associated with each metric).

Table 4 shows the distribution of hub edges between the six course categories using a degree centrality threshold of 200, helping to consider connections across categories. The table displays the percentage of total hub edges from one category (row) to both hub and non-hub courses in another category (column), for each threshold. The percentage of total edges, as well as the actual number of edges, associated with each category (row), are also provided. A color scale is applied to the rows to highlight where the hub connections are directed (red is high percentage and green low percentage). For example, the first row indicates that, using a static threshold, 5% of all Arts hub courses are connected to other Arts courses and 14% are connected to Communication courses. Furthermore, Arts courses have 1,520 edges, comprising 5% of all edges in the course network.

Table 4 shows that for both thresholds, Humanities, STEM, and Social Sciences have the most hub edges, while Arts, Communications, and Modern Language have many fewer. Notably, the static threshold associates more edges with humanities courses than STEM courses (35% to 27%), whereas the dynamic threshold

Table 4. Percent distribution of hub edge linkage by course category (hubs with degree ≥200) with edge info

Category	Static threshold								Dynamic threshold							
	Arts	Comm	Hum	Lang	STEM	SocSci	#Edges	%Edges	Arts	Comm	Hum	Lang	STEM	SocSci	#Edges	%Edges
Arts	5	14	27	7	26	22	1520	5	5	9	21	10	36	20	650	5
Comm	11	24	22	7	20	15	1426	5	8	27	19	9	21	15	954	7
Hum	10	12	31	6	21	20	10892	35	6	11	20	10	33	21	2758	19
Lang	10	16	28	5	18	23	3219	10	9	17	23	5	22	24	1773	12
STEM	5	8	27	7	32	20	8479	27	5	6	24	8	36	21	5543	39
SocSci.	7	10	25	8	25	25	5543	18	3	6	23	10	29	29	2717	19

reverses this trend (19% humanities to 39% STEM). Most core curriculum requirements are associated with humanities and the dynamic threshold has an outsized impact removing courses that are hubs simply due to their popularity.

It is especially notable that more edges link humanities to STEM courses than to other humanities courses. Examining the underlying data, we find that the humanities courses *Introduction to Cultural Anthropology*, *Introduction to Physical Anthropology*, and *Introduction to Art History* all connect to STEM hub courses. Similarly, most connections for courses in the Anthropology and Art History departments go towards the Biological Sciences and Natural Science departments. While *Introduction to Physical Anthropology* is part of the Natural Science major requirement, it also satisfies a science core curriculum requirement for non-Science majors. It is interesting to observe this course's popularity with Science students. The course is a general survey of the biological focus of Anthropology.

Also notable is that Communications and Social Sciences have more links to themselves than to any other category, for both static and dynamic thresholds, even though these categories do not have as many total links as other categories. The Languages category have mostly internal links, and an intermediate number of edges overall for both thresholds.

Social Science hubs in Table 4 have a significant number of connections to STEM courses, commensurate with connections back towards Social Science. Most of the connections to STEM refer to courses in Biological Sciences, particularly from the Psychology course *Foundations of Psychology*. This course is a requirement for the Psychology major but is not part of the core curriculum. This course also has a significant number of connections with the Natural Science department. Overall, the number of connections from non-STEM to STEM courses when using the dynamic threshold is a bit of a surprise. Conversely, STEM hubs made many connections to the Social Science category in Table 4; these connections are largely directed towards the Economics department, which requires a strong mathematical base.

6. CONCLUSIONS

This study analyzed course network graphs using eight years of undergraduate course-grade data from Fordham University. General network statistics and course hub statistics were generated using a publicly available Python-based tool created by our research group [10]. Network structure and hub identity are strongly influenced by the definition of edges between courses, and whether static or dynamic threshold were applied to course co-enrollments. We gain important insights on relations among courses, departments, and categories, and on metrics naturally applied to characterize these relations.

All three common network centrality metrics (degree centrality, betweenness centrality, and eigenvector centrality) identify a similar set of hub courses using static thresholding to define edges. However, the metrics behave much less similarly when dynamic thresholding is used, requiring careful consideration in future analyses. Nonetheless, degree centrality yields a reasonable approximation of the other two metrics for both thresholds, favoring its future use to study course co-enrollment networks.

The static and dynamic thresholds yield very different course networks and hubs. Static thresholds place more emphasis on course popularity, highlighting courses that uniquely satisfy a core requirement. The dynamic threshold reduces, but does not eliminate, popularity bias. Due to the many mandatory humanities core courses, and the variety of core options in STEM, the dynamic threshold substantially shifted apparent hub focus from Humanities to STEM. Future analyses of course relations and discipline relations must continue to carefully weigh the influence of popularity or the mandatory nature of courses. For both thresholds, STEM courses have the highest density and form tightly connected clusters, while humanities courses have the opposite behavior; this is likely due to the more extensive use of prerequisites in STEM disciplines in our university.

Our analysis also identified large numbers of edges between the different course categories. Edge distributions shifted between thresholds, favoring humanities for the static threshold and STEM for the dynamic threshold. Study of courses forming individual edges provided additional insights. The strong connection between humanities and STEM courses was driven by humanities courses like *Introduction to Physical Anthropology*, which has a strong STEM component; the connection between social sciences and STEM was driven by courses like *Foundations of Psychology* which is linked to STEM courses in Biology (Psychology students must take several biology courses).

This study provides a better understanding of course co-enrollment patterns, suggesting directions for valuable practical applications. Strong models of co-enrollment patterns can help with course planning and ensuring enough of course sections are offered. Our course networks reveal valuable details and quantitative relationships among courses. This work is a foundational step in better understanding course co-enrollments.

There are many ways in which this work can be extended and improved. The dynamic threshold could incorporate underlying probabilities of each course being taken, so courses are linked only where their co-occurrence is much more likely than chance. We also can consider additional methods for clustering courses. Future analyses may extend to course ordering information. It may be useful to reduce the influence of popular departments in repeated analysis of category-level network patterns. More fundamentally, our present results may be validated by partitioning the underlying student enrollment records into distinct subsets, to create training and testing data for our network models.

7. REFERENCES

[1] Arif, T. 2015. Mining and Analyzing Academic Social Networks, *International Journal of Computer Applications Technology and Research*, 4, 878-883. 10.7753/IJCATR0412.1001.

[2] Barabási, A. 2016. *Network Science*, Cambridge University Press.

[3] Catanese, S. A., De Meo, P., Ferrara, E., Fiumara, G., and Provetti, A. 2011. Crawling Facebook for social network analysis purposes, *Proceedings of the International Conference on Web Intelligence, Mining and Semantics*, 1-8.

[4] Gutenbrunner, T., Leeds, D.D., Ross, S., Riad-Zaky, M., and Weiss, G.M. 2021. Measuring the academic impact of course sequencing using student grade data. In *Proc. of the 14th International Conference on Educational Data Mining*.

[5] Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46.5 (1999): 604-632.

[6] Leeds, D. D., Zhang, T., and Weiss, G. M. 2021. Mining course groupings using academic performance. In *Proceedings of the 14th International Conference on Educational Data Mining.*

[7] Marsden, P. V. 2005. *Encyclopedia of Social Measurement*.

[8] Page, L., Brin, S., Motwani, R., and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Stanford InfoLab.

[9] Park, H. W., and Thelwall, M. 2003. Hyperlink analyses of the World Wide Web: A review. *Journal of Computer-Mediated Communication*, 8(4).

[10] Riad-Zaky, M., Weiss, G.M., and Leeds, D.D. Course Grade Analytics with Networks (CGAN) [computer software], April 2021.

[11] Ruhnau, B. 2000. Eigenvector-centrality a node-centrality? *Social networks*, 22, 357–365.

[12] G. M. Weiss, N. Nguyen, K. Dominguez, and D. D. Leeds. Identifying hubs in undergraduate course networks based on scaled co-enrollments, arXiv:2104.14500 [cs.SI]

APPENDIX

Figure 1 shows the distribution of common students by course pair (each bin covers a range of common students). The orange curve is a cumulative curve that corresponds to the y-axis values listed to the right (varying between 0% and 80%) and represents the percentage of course-pairs that are *maintained* for each common student threshold value (e.g., a threshold of 20 maintains 11% of all course-pairs with at least one student).

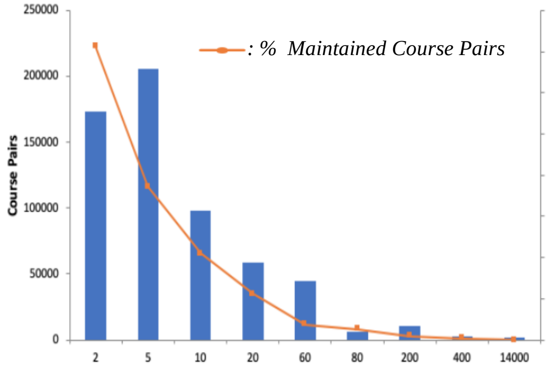

Figure 1. Distribution of common students by course-pair

The dynamic threshold is heavily dependent on the co-occurrence rate *k*. To help set this value appropriately, Figure 2 shows the distribution of course-pairs for each co-occurrence rate, for the course pairs that satisfy the static threshold of 20. The co-occurrence rate is the number of common students divided by the number of students in the course with more students. The co-occurrence rate distribution is heavily skewed to the smaller values, just as the number of common students was skewed to the smaller values in Figure 1. The bar at the far right at x=1.0 is associated with course pairs with the same course in both positions and should be ignored. After some experimentation we decided on a co-occurrence rate threshold *k* = 0.017, which is the value that leads to the most stable centrality measures while excluding the fewest number of edges. The orange curve, which shows the fraction of edges discarded, indicates that this value of *k* discards 39% of the edges that satisfy the static threshold.

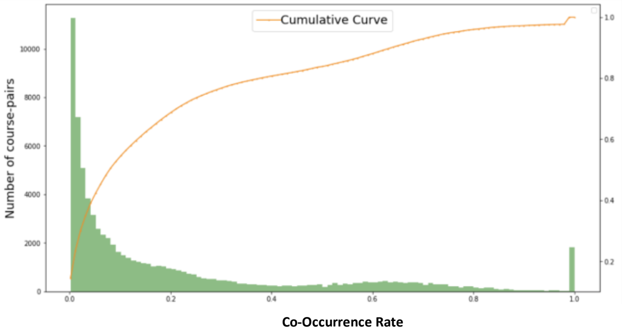

Figure 2. Co-Occurrence Rate Distribution

To illustrate the dynamic threshold, we apply it to the course *Art History Seminar*, which has 123 students. There are 22 courses that share at least 20 students in common with this course, satisfying the static threshold. However, 9 courses have fewer common students than the computed dynamic threshold, and hence are pruned. Half of these 22 courses are displayed in Table 5, and five of these, denoted in **bold**, are pruned since the number of common students is less than the dynamic threshold. As anticipated, the courses affected by the dynamic threshold have a large number of students (third column). In this example, every course that satisfies the static threshold, but is pruned by the dynamic threshold, fulfills a core curriculum requirement.

Table 5. Dynamic threshold for Art History seminar course

Course2	Common Students	Students Course2	Dynamic Threshold
Intro Cultural Anthro.	**23**	**2514**	**43**
Ancient American Art	21	34	20
17th Century Art	22	47	20
20th Century Art	43	130	20
Age of Cathedrals	20	39	20
Aztec Art	22	61	20
Composition II	**58**	**12446**	**211**
Intermediate French II	20	1329	23
Finite Math	**42**	**4976**	**85**
Philosophical Ethics	**58**	**11218**	**191**
Faith & Critical Reason	**56**	**13317**	**226**