Applied Network Science

RESEARCH

Open Access

Comparative analysis of course prerequisite networks for five Midwestern public institutions

Bonan Yang[1*], Mahdi Gharebhaygloo[3*], Hannah Rachel Rondi[1], Efrosini Hortis[2], Emilia Zeledon Lostalo[4], Xiaolan Huang[4] and Gunes Ercal[1*]

*Correspondence:
byang@siue.edu;
mgharehbaygloo@mst.edu;
gercal@siue.edu

[1] Computer Science Department, Southern Illinois University Edwardsville, 1 Hairpin Dr, Edwardsville, IL 62026, USA
[2] Office of Academic Advising, Southern Illinois University Edwardsville, 1 Hairpin Dr, Edwardsville, IL 62026, USA
[3] Mathematics Department, Missouri University of Science and Technology, 400 W 12th St, Rolla, MO 65409, USA
[4] Computer Science Department, Southern Illinois University, 1230 Lincoln Dr, Carbondale, IL 62901, USA

Abstract

We present the first formal network analysis of curricular networks for public institutions, focusing around five midwestern universities. As a first such study of public institutions, our analyses are primarily macroscopic in nature, observing patterns in the overall course prerequisite networks (CPNs) and Curriculum Graphs (CGs). An overarching objective is to better understand CPN variability and patterns across different institutions and how these patterns relate to curricular outcomes. In addition to computing well known network centrality measures to capture courses of importance in the CPNs studied, we have also formulated some newer methods with specific relevance to the curricular domains and corresponding graph types at hand. We have discovered that a new graph theoretic measure of node importance which we call reach, based on the well-known concept of reachability, is needed to more accurately express the critical nature of some introductory courses in a university. Another analytical novelty that we introduce and apply to the subject of CPNs is the Longest Paths Induced sub-Graph (LPIG) of the CPN, which yields information on relatively constrained programs and pathways. Finally, we have established a new connection between clustering of the CG and meta-majors at Southern Illinois University Edwardsville (SIUE), providing clusterings of the other public institution CGs as useful heuristics of major groupings as well. This work is borne from collaboration between academic units and academic advising with hopes of practical benefits towards aiding student advising.

Keywords: Course prerequisite network, Curriculum graph, Directed acyclic graph

Introduction

Prospective students of an institution may often find the course prerequisite graphs of Science, Technology, Engineering, and Mathematics (STEM) degree programs provided alongside curricular descriptions (SIUE Civil Engineering Course 2024; SIUE Computer Engineering Course 2024; SIUE Electrical Engineering Course 2024; BYU-Idaho 2024; Macalester College 2024; Wellesley College 2024; Washington and Lee University 2024). Computer Science students may once again come across the course prerequisite graph of required courses in their program in their data structures and algorithms classes in the context of Directed Acyclic Graphs (DAGs) algorithms. Providing their relevant course prerequisite subnetwork as an example not only serves to motivate the students in the

topic of DAGs but also often serves as a great visualization and analysis tool to help them navigate their own course planning and scheduling. Despite the prevalent availability of DAGs representing the dependency structure of STEM degree programs, it is surprisingly difficult to find public datasets representing the entire course prerequisite network (CPN) of any institution, much less finding academic works providing analyses of such. To our knowledge, there are two prior works in the academic literature analyzing institutional CPNs: Stavrinides and Zuev (2023) analyzes the CPN of the California Institute of Technology (CalTech), and Aldrich (2015) analyzes the CPN of Benedictine University. Whereas the CalTech CPN is made public by the authors (Stavrinides and Zuev 2023), the Benedictine CPN was not provided publicly by Aldrich (2015).

In Aldrich (2015) the course prerequisite network at Benedictine University is encoded as a DAG visualized in Gephi (Heymann and Le Grand 2013), and some well known network science statistics are presented in relation to corresponding curricular questions. For example, node centralities express the roles of courses acting as hubs (degree centrality) or bridges (betweenness centrality) in the overall curriculum structure, while path lengths of prerequisite chains within a program yield lower bounds for completion time. The work Stavrinides and Zuev (2023) significantly extends CPN analyses for the case of the California Institute of Technology (CalTech) to additionally provide topological stratification of the CPN and interdependence analysis upon the derived curricular networks corresponding to university programs and divisions. Inter-subject relationships within the curriculum graph are implied to correspond to the fundamental relationships between the knowledge areas themselves, with high betweenness subjects appearing more interdisciplinary.

The CalTech and Benedictine CPN analyses of Stavrinides and Zuev (2023); Aldrich (2015) serve as important seminal works demonstrating the effectiveness of graph theoretic methods in understanding curricular questions. Although both CalTech and Benedictine are private institutions, the distinctions between their CPNs highlighted by Stavrinides and Zuev (2023) provide a glimpse of CPN variability. As the vast majority of undergraduate students in the United States are enrolled in public institutions (see Fig. 4 of IES NCES (2024)), a more complete picture of CPN variability and extractions of curricular patterns necessitates consideration of public institutions as well. That is the starting point for the present work.

We analyze the CPNs and derived curriculum graphs for 5 Midwestern public universities: Southern Illinois University Edwardsville (SIUE), Southern Illinois University Carbondale (SIUC), University of Illinois Urbana Champaign (UIUC), Missouri University of Science and Technology (MST), and University of Missouri Kansas City (UMKC). We include the CalTech CPN and curriculum graph in our comparative analyses as well both for context and to include additional, updated analyses of that network.

Our overarching objective is to better understand CPN variability and patterns across different institutions and how these patterns relate to curricular outcomes. As a first step towards that objective, several basic network statistical measures are compared across the different CPNs considered. Some of these measures like degrees, betweenness centralities, and diameter are immediately extracted via graph visualization tools such as Gephi (Heymann and Le Grand 2013) and relate approximately to curricular properties such as critical or important courses and critical course sequences respectively, as noted

in Aldrich (2015) and Stavrinides and Zuev (2023). Upon extracting the course nodes achieving highest degrees and highest betwenness centralities across the different CPNs, we illustrate similarities as well as dissimilarities, general patterns.

However, the importance or criticality of a course as expressed via high betweenness is of a different nature than the importance or criticality of a course as expressed via high out-degree, also noted by Stavrinides and Zuev (2023). As a more general objective, we wish to more deeply explore how different notions of node importance in a network (CPN) translate to specific notions of course criticality in the curricular landscape.

In the process of this exploration, we have discovered that a new graph theoretic measure of node importance is needed to more accurately express the critical nature of some introductory courses in a university. This notion, which we call *reach*, is simply the size of the breadth-first-search tree (reachability set) rooted at a node. In Stavrinides and Zuev (2023), PageRank centrality, which is the PageRank of the transpose network, was noted to better capture the critical nature of fundamental introductory courses compared to out-degrees and betweenness centralities. Whereas the application of PageRank centrality to the CPN has a similar motivation to reach and acts very similarly in many cases, it does not necessarily produce the same importance orderings. Reach is a meaningful notion of node importance in a DAG but less meaningful for general directed graphs and completely meaningless for undirected graphs, perhaps hinting at why reach has not been used as a measure previously. We demonstrate the importance of reach as a measure in extracting well-known critical introductory courses such as College Algebra (Goonatilake et al. 2013), and we compare the node importance rankings yielded by reach with those yielded by the PageRank centrality.

Another analytical novelty that we introduce and apply to the subject of CPNs is the *Longest Paths Induced sub-Graph* (LPIG) of the CPN. Given a length parameter d, $LPIG_d$ is the subgraph of the CPN induced by all nodes which lie on paths of length d or longer in the CPN. The LPIG is also a structure whose meaningful computation is highly dependent on the acyclic nature of the DAG: Whereas longest paths in general graphs is well-known to be NP-complete (Garey et al. 1974), the longest paths problem is linear-time computable for DAGs Sedgewick and Wayne (2011); Cormen et al. (2022). Given that each course along a prerequisite chain must be completed in a different term, the $LPIG_6$ gives information about highly constrained degree programs in a university. Comparison and contrast of LPIGs across different institutions provide further information about relative constraints of categories of degrees in addition to motivating discussion on corresponding student outcomes.

Our final novel application of graph theoretic algorithms and modeling towards understanding curricular outcomes concerns the structure and distribution of *meta-majors*. As stated in SIUE's advising website (SIUE Meta-Majors 2024), instead of declaring a major up front, first-year students are grouped into 8 meta-majors according to their stated interests, for purposes of advising and tracking. As student retention, persistence, and timely graduation are amongst the important issues that the institution continually examines, a hope concerning meta-majors is that there should be *sufficient connectedness between majors of a given meta-major* so that a student starting out with unofficial declaration in one major of a meta-major may have the opportunity to switch to another major in the same meta-major without great waste of time and credits if the

change of heart is detected soon enough. Burke (2020) We model this property in the language of complex networks as the problem of *community detection*, also called *graph partitioning* or *clustering*, in the *Curricular Graph* of majors derived from the CPN. This modeling is motivated by the fact that the intra-meta-major connectivity requirement is precisely captured by the community detection objective that the connectivity within a community be notably higher than the connectivity between communities (Girvan and Newman 2002). This brings us to our last investigation: Upon applying modularity based clustering to the Curriculum Graph, examine the relationship between the resultant clusters and the meta-major subdivisions.

While we have stated our disparate research objectives, we wish to clarify aspects of the broader motivation for this work prior to proceeding to technical aspects and results. This work represents the first step by the authors towards addressing curricular and institutional questions that have arisen in various departmental committees and university working groups over the years at the authors' respective institutions. A primary SIUE author chairs the Undergraduate Curriculum Committee in the Computer Science department and another SIUE author directs the SIUE Office of Academic Advising and architected the meta-majors at that institution: This collaboration arose during their work in a university-wide working group on Improving Persistence and Timely Graduation (IPTG). Both the institutional directives which initiated the IPTG working group and the content of the IPTG final report indicated a need to formally study prerequisite structure both within programs and across the institutional landscape with respect to properties of rigidity versus flexibility in addition to analyzing the composition of meta-majors with respect to questions of cohesiveness and minimization of excess credits upon intra-meta-major switching. Prerequisite relationships have a combination of artificially constructed and fundamental aspects, where some dependency relationships might be universally agreed upon inherent knowledge dependencies while other prerequisite dependencies may serve practical institutional advising purposes. Therefore, in the course of our investigations we discovered in the course that it is best to first analyze the *pattern and variation in prerequisite structure* across relevant institutions, which forms the major emphasis of the present study.

Towards the question of selecting relevant institutions to study: This collaboration further expanded to involve existing collaborators from SIUC Computer Science and MST Mathematics departments, hence including those neighboring institutions as well. Already with SIUE, SIUC, and MST we cover some different institutional characteristics with respect to graduation rates, STEM versus general emphases, graduate versus undergraduate emphases, rural versus suburban environment, size, and selectivity. However, as we wished to include the consistently highest ranked public university across the Illinois and Missouri regions, we include UIUC in our study. The inclusion of UMKC in our study is originally due to the implementation of meta-majors in that institution, though we were subsequently unable to obtain data on specific meta-major composition there. Nonetheless, due to its student composition and persistence problems, meta-majors have generally been used as an advising method at UMKC, yielding some similarity to SIUE despite other institutional differences between the schools with respect to selectivity, graduate research orientation, and urbanicity. Upon selecting SIUE, SIUC, MST, UIUC, and UMKC, in addition to comparing with the previous work on CalTech, our

sample incorporates sufficient variation in institutional profiles to form meaningful comparisons. With the caveat that much more work yet remains to answer many of the persistence related questions forming our original motivations, we now attempt to shed some light on broad patterns and variation within and across institutional CPN and curriculum networks for a meaningful sample of Midwestern public institutions.

Description of datasets, definitions, and methods

Datasets

The course information for the public institutions in this work are obtained from each school's online course catalog. For the CalTech data, we used the dataset provided by Stavrinides and Zuev (2023). Such data has is used to find prerequisites, co-requisites, cross-listing, and other dependency relationships. The outcome of this process is used as raw data towards generating graphs connecting the courses (CPN) and programs (CG).

Definitions and notations

CPN formation

All analyses in this work are based on the *Course Prerequisite Networks* (CPNs) extracted from the university catalogs mentioned above. As indicated in Stavrinides and Zuev (2023); Aldrich (2015), the CPN graph $G = (V, E)$ essentially captures the prerequisite relationships between courses by including, for each prerequisite $X \in V$ of course $Y \in V$, a directed edge $(X, Y) \in E$. Since prerequisites must be satisfied prior to the course itself, the CPN is a *dependency graph* and must be *acyclic*, forming a *directed acyclic graph* (DAG). Adopting the convention of Stavrinides and Zuev (2023), $X \prec Y$ denotes that course X is a prerequisite for course Y.

In this work, due to the discovery of a sizeable number of co-requisites, cross-listed courses, and other indicators of equivalent courses in the various course catalogs we have parsed, we need to modify and clarify our CPN to allow the vertex set V to be a partition of the course set. The vast majority of members of V are singleton sets whose correspondence with individual courses and the prerequisite relationship is straightforward. However, due to the existence of non-singleton sets in V we must generalize the prerequisite relationship to act between sets of courses in order to now properly define our CPN: Let S and T be disjoint sets of courses. Then

$$S \prec T \leftrightarrow \exists s \in S, t \in T \ni s \prec t \tag{1}$$

Prior to specifying the graph construction notation, we take a moment to elaborate upon a few issues surrounding the parsing of the course catalog with respect to extracting prerequisite information. First, there is the issue of different conjunctions used in expressing prerequisite information. While the vast majority of courses have standard prerequisite listings connected by the conjunction *AND*, there are also situations in which the prerequisite list is a more general logical expression involving both *AND* and *OR* connectives. We acknowledge the differing semantics induced by *OR* versus *AND* connectives acting on the prerequisite courses, as prerequisites connected via the conjunction operator are absolute requirements while the others need not be. Nonetheless, we adopt the convention in Stavrinides and Zuev (2023) in which we do not distinguish between the different types of prerequisites listed for a course in forming the CPN DAG.

As a further detail concerning CPN formation, we note the allowance of *corequisites* and *course equivalencies* in the course catalogs. Co-requisites are instances in which a course X is permitted to be taken *concurrently* with course Y. In many cases, the purpose of stating co-requisites is to allow more scheduling flexibility for students despite the existence of some degree of knowledge dependence between the respective courses. Such situations are signified in the course catalog by the listing of a course Y as "prerequisite or corequisite" for course X ***without*** the mention of X in the prerequisite listing of Y. As we are considering the underlying dependence structure without solving scheduling in this work, we treat this type of situation as signifying a directed edge from Y to X but not vice versa, hence maintaining acyclicity. Namely, $Y \prec X$ but ***not*** $X \prec Y$.

The other complicating situations involve true corequisites in addition to generalized equivalencies of course sets. The vast majority of true corequisites comprise lecture and corresponding lab pairs which must be taken in the same term such that the courses in the pair share the identical course code excepting an additional "L" following the corresponding lab. For such lecture and lab pairs of corequisites, in our CPN graph we consider the pair as a merged course node with the common course code excluding the "L" suffix of the lab code.

The last situation, which was more difficult to parse automatically from the distinct course catalogs, is the situation of courses which are treated as equivalent or cross-listed as indicated by catalog terms such as "Same as", "co-listed with", or "cross-listed with". In these cases too, consistent with the dependency characterization of the CPN structure, we have adopted the convention of merging sets of courses which are indicated to be equivalent in some catalog context. Given a set of equivalent courses $S = \{C_1, C_1, \ldots, C_3\}$, we consider the set of courses in S as a single merged course node in the CPN graph.

We note that the merging of course sets in the CPN based on lab-lecture co-requisite relations, cross-listings, co-listings, and other contexts of similarity induce an equivalence relation upon courses which are merged. Therefore, let us denote this relationship with $\equiv_C$ as follows given a pair of courses C_1 and C_2: $C_1 \equiv_C C_2 \longleftrightarrow C_1$ and C_2 are represented by the same merged vertex in the CPN. Letting $\mathfrak{C}_I$ denote all the courses in a given institution I, the equivalence relation $\equiv_C$ induces a partition on $\mathfrak{C}_I$ which we denote as V_I:

$$V_I = \{\{x \mid x \equiv_C c\} \mid c \in \mathfrak{C}_I\} \tag{2}$$

Clearly, each member of V_I is an equivalence class $[c]^{\equiv_C}$ of some course c. Courses which were not involved in any merging in the CPN have singleton equivalence classes.

Now we may denote the CPN graph for each institution I, as $CPN_I = (V_I, E_I)$ where V_I is the set of equivalence classes of courses, and the directed, unweighted edge set E_I defines the set-generalized *prerequisite* relationship $\prec$. As the CPN_I is a directed acyclic graph (DAG):

i. For any $c \in V_i, (c, c) \notin E_i$
ii. For any $x, y \in V_i$, if path $x \rightsquigarrow y$ exists in CPN_i, then there is no path from y to x.

In addition to the above notations for the CPN graph and the corresponding node and edge sets, we will refer to the adjacency matrix of a graph G as $A(G)$ or simply A when the context is clear. When discussing node centralities, especially PageRank centrality, sometimes it will be useful to refer to the *transpose CPN* which has adjacency matrix $A^T = A^T(CPN_I)$, which is the transpose of the adjacency matrix of CPN_I and defined as follows. Recall that the transpose matrix A^T is defined as follows, for any matrix A:

$$A^T_{rc} = A_{cr} \tag{3}$$

Likewise, the transpose E^T of an edge set E is simply defined as the set of edges with all arrow directions reversed, which formally may be represented as:

$$E^T = \{(x, y) \mid (y, x) \in E\} \tag{4}$$

Curriculum graph formation

As detailed in Stavrinides and Zuev (2023), in order to perform a more macro level analysis of the relationships between departments and units in the curricular landscape, we derive the Curriculum Graph (CG) from the CPN where each node represents a major code M and directed edges (M_1, M_2) between major codes M_1 and M_2 are weighted according to the number of edges in the CPN from nodes with major code M_1 into nodes with major code M_2. For example, if exactly 5 courses with major code *MATH* are immediate prerequisites of courses with major code *PHIL*, then there is an edge in the CG of weight 5 from node *MATH* to node *PHIL*. Note that the CG need not be acyclic although it is derived from an acyclic CPN, as different pairs of courses contribute to the existence and weights of edges in the CG. For example, an introductory computer science (*CS*) course may be a prerequisite to an upper level mathematics (*MATH*) course in numerical methods, while other introductory mathematics courses might be prerequisites to intermediate computer science courses, forming anti-parallel edges of different weights from *CS* to *MATH* and from *MATH* to *CS* separately, inducing a simple cycle in the CG.

As in the case of the CPNs as detailed in the prior section, we must address the treatment of courses that are in the same equivalence class *but in different majors*. Recall that two courses are only in the same equivalence class if they are either co-requisites, co-listed, crosslisted, or described to be equivalent with statements such as "same as" in the catalog. As such, the existence of courses in the same equivalence class but different majors is an indication of some level of symmetric relationship between the majors. Therefore, every instance of such an equivalent pair of courses c_1 and c_2 involving distinct majors M_1 and M_2, respectively, will contribute an additional weight of $+1$ to both the edge (M_1, M_2) and the edge (M_2, M_1).

We may refer to the curriculum graph for institution I as $CG_I = (M_I, \hat{E}_I, w_I)$ where M_I is the set of majors with distinct major codes, $\hat{E}_I$ is the set of directed, weighted edges between majors derived from CPN_I with weight function w_I. It will be useful to overload the notation to also define the major function applied to any class $c \in \mathfrak{C}$ as $M_I(c)$, meaning the major code $m \in M_I$ associated with the course. We note that the major codes also necessarily partition the course set $\mathfrak{C}$ but not the vertices of the CPN. We may therefore

also define another equivalence relation $\equiv_M$ upon courses such that $c_1 \equiv_M c_2$ if and only if $M_I(c_1) = M_I(c_2)$. Similarly, denote by equivalence class $[c]_{\overline{M}}^{\equiv}$ the set of courses in the same major as course c. And, for any $m \in M_I$, and any c such that $m = M_I(c)$, let $C(m) = [c]_{\overline{M}}^{\equiv}$.

For any major code, $m \in M_I$, we also overload our notation to extend to function $V_I(m) \subset V_I$ as the set of vertices in the CPN_I corresponding to m, namely:

$$V_I(m) = \{v \in V_I \mid \exists c \in \mathfrak{C} \ni M_I(c) = m\} \tag{5}$$

Consider again the situation of courses which are equivalent with respect to $\equiv_C$ but not equivalent with respect to $\equiv_M$: Sometimes pairs of courses in different majors that are nonetheless cross-listed with each other exist. Due to such situations, note that the set of $V_I(m)$ need not be disjoint, and in fact overlap between $V_I(m_1)$ and $V_I(m_2)$ signify a strength of connection between m_1 and m_2 in the Curriculum Graph.

Now we may exactly define $\hat{E}_I$ and w_I. For any distinct $m_1, m_2 \in M_I$ such that $m_1 \neq m_2$:

$$(m_1, m_2) \in \hat{E}_I \iff (\exists c_1 \in C(m_1), c_2 \in C(m_2), \ni ((c_1 \prec c_2) \bigvee (V_I(m_1) \bigcap V_I(m_2) \neq \emptyset))) \tag{6}$$

Regarding weight function w_I, $w_I(x, y) = 0$ if and only if $(x, y) \notin \hat{E}_I$. For any $m_1, m_2 \in M_I$ such that $(m_1, m_2) \in \hat{E}_I$:

$$w_I(m_1, m_2) = |\{(c_1, c_2) \in C(m_1) \times C(m_2) \mid c_1 \prec c_2\}| + |(V_I(m_1) \bigcap V_I(m_2))| \tag{7}$$

The adjacency matrix $A = A(CG_I)$ holds both edge and weight information as follows:

$$\forall x, y \in M_I, A_{xy} = w_I(x, y) \tag{8}$$

Node centralities

Various centrality measures are applied to rank nodes in the CPN and CG to extract information about the relative criticality of courses and majors in the curricular landscape. We follow the standard definitions of centrality measures such as betweenness centrality (BC) Brandes (2001) and out degree centrality (Freeman 1977). In applying PageRank (Page et al. 1999) to analyze the CPN, we follow the convention of Stavrinides and Zuev (2023) in denoting the *PageRank centrality* of a node (course) in the CPN to be the node's PageRank in the transpose of the CPN, which we also refer to as the *transpose PageRank* for clarity. The reason for taking the transpose of the CPN prior to application of PageRank for the purposes of extracting relative node importance is due to the meaning of edges in the CPN versus their meaning in the World Wide Web (WWW) in the original PageRank paper Page et al. (1999): A course *Y depends on* a course *X* when *X* is a prerequisite for *Y*, denoted by the edge (*X*, *Y*) in the CPN. But, a website *Y depends on* another website *X* when the direct link (*Y*, *X*) exists in the WWW. Therefore, PageRank centralities correspond to the PageRank values of the transpose CPN, namely CPN_I^T as described in Sect. 2.2.1. We elaborate on the computation of PageRank centrality in the Methods Sect. 2.3. Presently, we continue precisely defining other commonly used centrality measures.

Given a directed graph $G = (V, E)$, we use $k_{in}(i)$ and $k_{out}(i)$ to denote the in-degree and out-degree of node i, respectively. In terms of adjacency matrix A, $k_{in}(i)$ and $k_{out}(i)$ are given by (9).

$$k_{\text{in}}(i) = \sum_{j=1}^{n} A_{ji} \quad \textit{and} \quad k_{\text{out}}(i) = \sum_{j=1}^{n} A_{ij} \tag{9}$$

The betweenness centrality of node $i \in V$ can be written as (10), in which $\sigma(s, t)$ is the total number of the shortest paths from node s to node t, and $\sigma(s, t|i)$ is the number of these shortest paths which passing through i.

$$\beta(i) = \sum_{s \neq i, t \neq i} \frac{\sigma(s, t|i)}{\sigma(s, t)} \tag{10}$$

Reach

A simple measure of importance for a node x is the *number of nodes that are reachable from x*, where the reachability set is computable in linear time $\Theta(|E| + |V|)$ using breadth-first search (BFS) or depth-first search (DFS) rooted at x Cormen et al. (2022). Node y is *reachable* from node x in graph $G = (V, E)$ if either $y = x$ or there exists a path $x \rightsquigarrow y$ from x to y in G. We extend this definition naturally towards a useful graph statistic named ***reach*** as follows: Given graph $G = (V, E)$ and vertex $v \in V$

$$reach(v) = |\{u \in V \mid \exists v \rightsquigarrow u\}| \tag{11}$$

Equivalently, note that,

$$reach(v) = |BFS(v)| \tag{12}$$

where $BFS(v)$ is the BFS tree rooted at v.

In the context of a CPN, if a course d is reachable from course c, then c lies on a prerequisite chain leading to d. Therefore, the reach of a course c in the CPN is precisely the number of courses for which c is a direct or indirect prerequisite. This precise meaning yields the high relevance of reach as a measure of study in the CPN context.

While reachability is a well-known concept in classical graph theory, we are unaware of any mention of the usage of reach, or any equivalent variation under a different name, as a network statistic or centrality measure. This may be due to the relative emphasis on undirected graphs in network science due to symmetries in many complex networks. In fact, reach is not a distinguishing characteristic of a node in an undirected graph, as any two nodes in the same component will have the same reach, namely their component size. Likewise, for directed graphs that are not DAGs, the Strongly Connected Component (SCC) size of a node is a lower bound for its reach, again relating all nodes in the same SCC. *It is really in DAGs that reach is more meaningful as a distinguishing measure of node importance, hence the usage of reach in this work.*

We conclude our introduction of the measure reach by noting the uniqueness of the information conveyed by reach in a DAG compared to all other known centrality measures considered. While we shall observe some correlation between the lists of highest reach nodes and highest transpose PageRank (tPR) nodes in some results, it is not

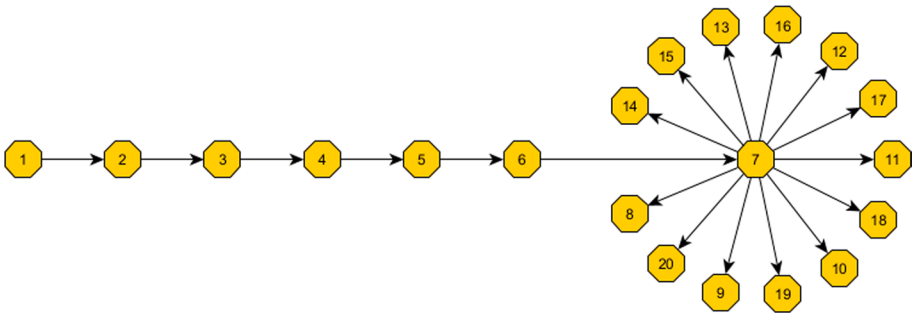


Fig. 1 Extremal example exhibiting high difference between transpose PageRank and Reach

difficult to construct an infinite class of DAGs for which the highest reach and highest tPR nodes differ for every setting of the dampening factor. A simple example of such an extremal graph is given in Fig. 1. The highest reach of that network is achieved by node 1 whereas the highest tPR is node 7, independent of dampening factor.

Longest paths induced sub-graph

As paths in the CPN represent pre-requisite chains, the length of the *longest path* leading to a course corresponds to the number of terms required to complete that course in the curriculum.[1] Courses that are sink nodes of relatively long prerequisite chains constrain the schedules of their respective degree programs. And, degree programs that have higher numbers of such constraining courses (and the chains that lead to them) are likely candidates for further analysis of how to aid student persistence throughout the completion of the curricula. Noting that 4 years is the standard time for undergraduate degree completion at a university, sample advertised curricula for undergraduate degree programs are all based on the 4 year degree goal. Moreover, with the exception of CalTech which is on the quarter system, all of the public institutions involved in this study operate on the semester system, where the standard number of terms per year (excluding the summer term) is 2, leading to 8 term graduation goals. In general, given a standard graduation goal of n terms, we wish to identify and analyze courses and degree programs which are involved in paths of length $t = n - 2$ or more. Noting that path lengths are traditionally expressed as the sum of edge weights, which for an unweighted graph is the number of edges along the path, the number of *nodes* along the path is one more than the path length. We formulate a new graph theoretic construct called ***longest paths induced sub-graph*** $LPIG_t$ that permits exactly such identification and analysis. As the $LPIG_t$ is based on the computation of *longest paths* in a graph, we first discuss the computation of longest paths.

Like reach, the longest paths problem in a graph has limited applicability for general graphs but high relevance for DAGs (like the CPNs). Unlike reach, however, the limited applicability of longest paths in general graphs arises due to computational concerns: The longest path problem in both general directed graphs and in undirected

[1] While there is some flexibility granted by edges resulting from prerequisites listed in the disjunctive "OR" form, the vast majority of links are due to the conjunctive "AND". Moreover, even "OR" based prerequisites yield that *some paths* must be chosen to complete requirements, with many of the "OR" based sub-paths being of comparable lengths.

graphs is NP-complete due to a reduction from the Hamiltonian Path problem (Cormen et al. 2022; Sedgewick and Wayne 2011). In DAGs, however, the longest path problem is linear-time solveable by updating path estimates from vertices considered in topological sort order, which results from the reverse finish time order of a depth-first search (DFS) of the graph (Cormen et al. 2022). Likewise, deciding the existence of paths of length at least k (for some given k) is in the class P when restricted to DAGs. However, the problem of computing *all* sufficiently long paths (of length at least k) is no longer in P even under the DAG restriction due to the potentially exponential number of such paths. Nonetheless, finding *all nodes involved in sufficiently long paths* (of length at least k) in a DAG is solveable in polynomial time with similar dynamic programming methods that allow the computation of a sufficiently long path in the first place as ties for predecessor nodes can also be memoized to be reconstructed during backtracking. In practice, however, when the number of sufficiently long paths in a DAG is not overwhelming, it may also be useful to simply compute all such paths, as that computation is still polynomial time in the number of such paths. Regardless of the choice of method, we emphasize that computing ***all nodes involved in sufficiently long paths*** in a DAG is a feasible problem, and this is precisely the set of nodes in the LPIG.

Notationally: Given an unweighted DAG $G = (V, E)$, with $n = |V|$, let k be given such that $1 \leq k \leq n-1$. A node $v \in V$ is said to *lie on a sufficiently long path with respect to* k iff there exists some $x, y \in V$ and a path

$p =< e_1, e_2, e_3, \ldots, e_{k+d} >=< (x, u_1), (u_1, u_2), (u_2, u_3), \ldots, (u_{k-1+d}, y) >$ such that $|p| \geq k$, meaning $d \geq 0$, and $v \in \{x, y, u_1, u_2, u_3, \ldots, u_{k-1+d}\}$

Given DAG $G = (V, E)$, let

$$V^k = \{v \in V \mid v \text{ lies on a sufficiently long path w.r.t. } k\} \tag{13}$$

Then, the induced sub-graph $LPIG_k(G) = (V^k, E^k)$ where

$$E^k = \{(x, y) \in V^k \times V^k \mid (x, y) \in E\} \tag{14}$$

Implementation of methods

The pipeline of our methods is as follows: (i) extraction of course catalog information to form the CPN graphs, (ii) construction of the Curriculum Graphs from the CPN and catalog data, (iii) computation of network centrality and importance measures on both types of networks, (iv) construction of the LPIG network from the CPN, and (v) clustering of the Curriculum Graphs. All parts of this pipeline have been implemented in Python, with **Gephi** additionally used to aid in the visualization and analyses of parts (iii) and (v).

For part (i), the Python libraries *request* and *beautifulshop4* were used to extract each school's course information from their official websites and organize it into their CPN's adjacency list. In terms of the graph construction for parts (i), (ii), and (iv) we mainly used the Python **networkX** library (Hagberg et al. 2008). For part (iii), we used Gephi to compute degree and betweenness centrality distributions and networkX to

Table 1 General data at a glance

	MST	SIUE	SIUC	UMKC	UIUC	CalTech
Size of CPN	1964	2342	3979	2304	5126	771
Number of edges in CPN	2157	2135	3044	1321	3827	772
Size of CPN_{Lcc}	1083	883	1395	624	1607	436
Diameter of CPN	8	9	7	7	10	5
Longest path of CPN	12	13	13	11	12	6
Size of CG	53	80	104	101	184	26
Size of CG_{max}	45	52	90	57	147	25

compute other measures such as reach and transpose PageRank (Page et al. 1999). For part (v), we primarily used Gephi for the clustering computation and visualization using modularity clustering based on the Louvian method (Blondel et al. 2008).

PageRank and modularity clustering are both parametrized methods. As prior work Boldi et al. (2009); Page et al. (1999) indicates $\delta = 0.85$ to be an empirically reliable parametric setting for the dampening factor of PageRank, that is the setting that we adopt. Regarding the resolution parameter for modularity clustering in Gephi, the default setting of the resolution to $r = 1.0$ is commonly used and recommended in general when the number of desired clusters is unknown. That is the setting that we also adopt for our experiments in part (v).

Results

Comparison of basic network statistics

An overview of the basic network statistics is seen in Table 1. This table also yields approximate institutional information about the number of courses and number of majors corresponding to the size of the CPN and the size of the CG, respectively. The size of the CPN is a lower bound on the actual number of distinct courses as equivalent courses are merged into one node as described in Sect. 2.2.1. On the other hand, the size of the CG is an upper bound on the number of distinct majors as it may include some codes for programs that are not currently majors as well. As the percentage of course equivalences and non-major codes are very low, the approximations provided by the CPN size and CG size are very near to the actual number of courses and majors, respectively. Therefore, this table well-encapsulates the immediate variation between the institutional sizes, with UIUC and CalTech standing out as the largest and smallest outliers respectively.

CPN centrality results

We have computed various measures of network *centrality* in the CPNs to better analyze candidates for courses important in the curricular landscape. The measures considered are betweenness centrality, out degree, reach, and transpose PageRank. The highest betweenness centrality courses of the six institutions may be found in Table 2. The highest outdegree courses are in Table 3. The highest reach courses are in Table 4. And the highest transpose PageRank courses are in Table 5. The full names of the courses in these tables are provided in the Appendix section.

Table 2 The courses with the highest BC at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
MATH1215	CHEM121A	MATH250	MATH220	MATH220	CH21ABC
MATH3304	BIOL220	CHEM200	MATH266	MATH231	PH125ABC
CIVENG2200	PHYS141(M)	MATH150	CIVENGR275	MATH241	PH2ABC
PHYSICS2135	MATH150	CHEM210(M)	PHYSICS240	MATH257	ACM95/100AB
MATH1214	NURS231	CHEM140A	CIVENGR276	CS225	CS38
MATH2222	CE240	MATH151	ECENGR276	CHEM104	ACM116
MATH1221	MS251	CHEM330	MATH210	ECE210	AE101ABC

Table 3 Courses with the highest *out degree* at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
54: ENGLISH1120	51: ENG102	37: ENGL102	98: ENGLISH225	67: MATH220	32: MA2102
50: HISTORY1310	48: ENG102N	33: (*2)	26: ENGLISH110	61: MATH221	30: ACM95/100AB
47: HISTORY1300	42: BIOL220	32: (*3)	22: MATH110(M)	54: MATH241	27: MA1ABC
45: MATH3304	30: PSYC111	28: MATH111	20: PHYSICS250	46: MATH285	22: BI8
43: HISTORY1200	29: MATH150	26: MATH305	18: (*2)	41: MATH415	19: MA3103,PH2ABC
42: MATH2222	26: CIED100	24: (*2)	17: (*2)	40: ECON302	17: PH125ABC
33: STAT3115	19: MATH125	23: (*3)	16: (*2)	37: CS225	16: CH41ABC,PH1ABC
32: STAT3117	18: (*5)	22: (*2)	15: (*2)	36: MATH257	15: (*3)
28: (*5)[1]	17: (*3)	21: (*3)	14: (*4)	35: PSYC100	14: CH21ABC
27: POLSCI1200	16: (*2)	20: (*2)	13: (*3)	31: STAT400(M)	12: (*5)

[1] If multiple courses share the highest out degree, their count is given in parentheses instead of listing course numbers

Table 4 Courses with the highest *reach* at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
MATH1140,MATH1120	**MATH120**	**MATH108**	**MATH110(M)**	**MATH112**	MA1ABC
MATH1160	MATH125	MATH111	MATH120	MATH115	MA2102
MATH1214	MATH150	MATH106	MATH125	MATH220	PH1ABC
MATH1208	MATH145	MATH109	MATH266	MATH221	CS1
MATH1210	MATH152	MATH150	MATH210	MATH231	PH2ABC
MATH1211	MATH250	MATH151	ENGLISH110	MATH241	CH1AB
MATH1215	CHEM113	MATH250	MATH220	MATH211	MA3103,PH12ABC
MATH1221	CHEM121A	MATH140	PHYSICS240	CS101	ACM95/100AB
MATH2222	PHYS140	MATH125	MATH268	MATH257	ACM11
PHYSICS1135	PHYS141(M)	MATH139	ENGLISH225	CHEM102	BI8

College Algebra or equivalent is in boldface

From Table 2, it can be seen that mathematics courses, especially those of the Calculus series, are a consistent bottleneck for curricula across the different institutions. In addition to mathematics courses, the basic sciences such as chemistry, physics, and biology are also over-represented. Several engineering courses, especially in civil engineering and electrical and computer engineering, also occupy positions of high betweenness centrality across multiple institutions. In fact, all high betweenness physics courses

Table 5 Courses with the highest *transpose PageRank* at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
MATH1215	**MATH120**	**MATH108**	ENGLISH110	MATH231	MA1ABC
MATH1140	MATH125	MATH111	**MATH110(M)**	**MATH112**	PH1ABC
MATH1120	MATH150	MATH250	ENGLISH225	MATH241	MA2102
MATH1214	MATH152	MATH150	MATH120	MATH220	CS1
MATH1160	CHEM121A	MATH151	MATH226	MATH221	CH1AB
MATH1208	MATH250	MATH106	MATH220	MATH115	BI8
MATH1221	CHEM121B	MATH109	CHEM211(M)	CHEM102	CH41ABC
CHEM1310	PHYS141(M)	ENGL101	MATH268	PHYS211	PH2ABC
MATH2222	BIOL220	CHEM200	MATH210	MATH285	CS2
ENGLISH1120	CE242	PSYC102	PHYSICS240	PHYS212	MA5105ABC

College Algebra or equivalent is in boldface

for MST, SIUE and UMKC are also *engineering physics courses*. Most notably, with the exception of one nursing course at SIUE, every course achieving high betweenness centrality is a STEM course across all institutions.

Unlike the case of betweenness centrality, the highest out degree courses are distributed across non-STEM majors in addition to STEM majors, as can be seen in Table 3. This measure also exhibits more variation across institutions. Although mathematics courses dominate the highest degree positions at UIUC and CalTech and appear as high degree courses in other institutions as well, English courses dominate the highest degree positions at MST, SIUE, SIUC, and UMKC. Closer analyses of the highest degree courses reveals that in some cases the successors of a highest degree course are in the same major, while in other cases the course is a direct prerequisite for courses across diverse majors. The next measure considered, namely *reach*, precisely captures the total immediate and downstream influence of a course in the CPN, hence well complementing the information provided by out degree.

College Algebra is the consistently highest reach class in all public institutions shown in Table 4.[2] College Algebra is also consistently amongst the top two highest transpose PageRank (tPR) courses according to Table 5. This prominent positioning of College Algebra in the CPN is supported by educational research highlighting the criticality of that course across university curricula (Goonatilake et al. 2013). Generally, introductory mathematics courses dominate both reach and tPR tables across all institutions with introductory English and basic science courses interspersed. Some introductory computer science courses also arise as high reach courses at UIUC and CalTech, where they also achieve high tPR. A notable engineering course that achieves high tPR is the civil engineering course CE242 at SIUE.

Longest paths induced graphs

In this section we present our results concerning the *longest paths induced subgraphs* (LPIGs) of each institutional CPN, specifically the $LPIG_6$ graphs which contain all

[2] While MATH110 Precalculus Algebra at UMKC is no longer titled College Algebra, some educational websites refer to that course as College Algebra, and it has a similar ALEKS placement score as College Algebra in other institutions.

Table 6 The statistics for the Longest Paths Induced sub-Graphs in each institution

	MST	SIUE	SIUC	UMKC	UIUC	CalTech
Size of $LPIG_6$	584	493	484	247	862	20
Number of paths	1151	917	863	372	1745	26
Number of source nodes	7	15	26	8	23	2
Number of sink nodes	374	276	238	121	475	6
Number of components	1	4	9	4	15	1
Top 3 highest BC	MATH1215	CHEM121A	CHEM200	MATH220	MATH220	CDS231
	MATH3304	BIOL220	MATH250	MATH266	MATH231	CMS122
	CIVENG2200	PHYS141(M)	MATH150	CIVENGR275	MATH257	CHE101

Table 7 The highest frequency majors amongst LPIG sink nodes

MST	SIUE	SIUC	UMKC	UIUC	CalTech
49: MECHENG	54: BIOL	34: ECE	27: MECENGR	56: ECE	4: CDS
37: ELECENG	23: CHEM	29: ME	24: ECENGR	44: CEE	2: CHE
32: CIVENG	22: ME	22: CE	23: CIVENGR	37: CS	
25: NUCENG	20: PHYS	16: SPAN	13: CONSVTY	24: ME	
24: ARCHENG	16: ECE,FIN	14: PLB,CHEM	8: MGT	21: MSE	

prerequisite chains of seven or more courses. As all public institutions of this study have standard eight term timelines, $LPIG_6$ gives information on highly constrained degrees and course sequences. In this section, we refer to the $LPIG_6$ graph of a given institution simply as LPIG. An overview of the network statistics for the LPIG networks are found in Table 6. This data can be taken together with the CPN longest path lengths provided by Table 1 for general comparisons. While there is some variation across the institutions with respect to LPIG sizes and the lengths of longest paths, CalTech is orders of magnitude smaller than the other institutions in both measures. This is especially striking when considering that CalTech is on a quarter system which permits a standard 12 terms instead of the 8 term standard of the other institutions. In contrast to CalTech, the public institutions involved in this study have hundreds of sink nodes in their LPIGs and longest paths comprising 12 to 14 course prerequisite chains, involving many terms more than their standard undergraduate program length. The significance of such constrained substructures in the public institutions considered warrants further analysis of which programs and categories of study are involved (Fig. 2).

Towards such investigation, Table 7 gives an overall picture of the major programs of study appearing most frequently as sink nodes of the LPIG subnetworks, while Figs. 3, 4, 5, 6, 7 and 8 show the representative $LPIG_6$ networks of each institution with color codes for categories of study given by Fig. 2. Engineering programs can be seen to dominate both the sink nodes of the institutional LPIGs of Table 7 and a significant portion of every institutional LPIG in the color coded LPIG figures as indicated by the prevalence of the sky blue subnetworks. Indeed, exactly two categories of study are common to all six LPIG networks: The mathematics & physics category in cyan and the engineering category in sky blue. However, with the exception of the CalTech and SIUE LPIG networks which include some mathematics & physics sink nodes, the cyan mathematics &

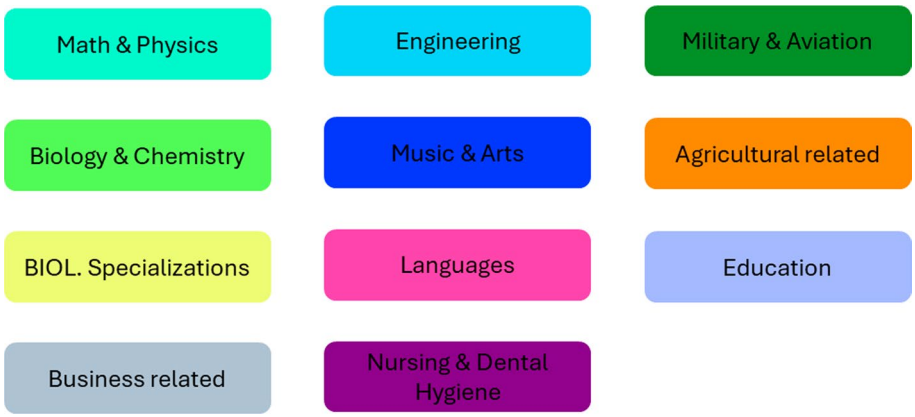


Fig. 2 LPIG color map

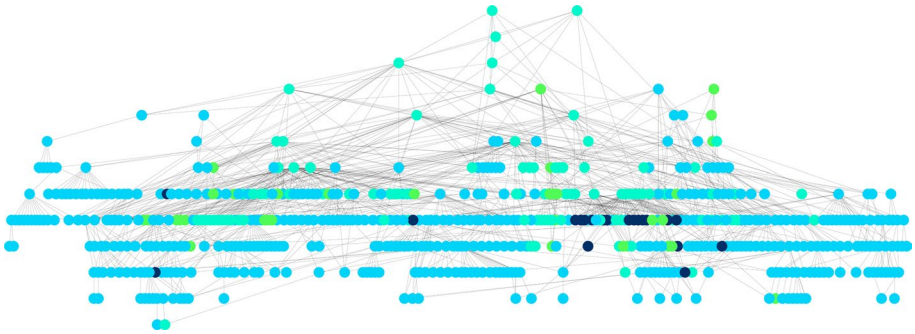


Fig. 3 LPIG at MST

Fig. 4 LPIG at SIUE

physics course category tend towards the source end of the institutional LPIG networks as prerequisites of engineering and other STEM fields. STEM fields in general comprise the vast majority of all LPIG networks with exceptions of special note in each institution. In addition to the occurrence of some non-STEM LPIG categories in some institutions, the precise composition of the STEM fields populating the LPIG networks exhibit interesting variation across institutions.

The SIUC LPIG is the most diverse as it includes all of the eleven categories of study listed. SIUE has the second most diverse LPIG in that the only excluded categories are

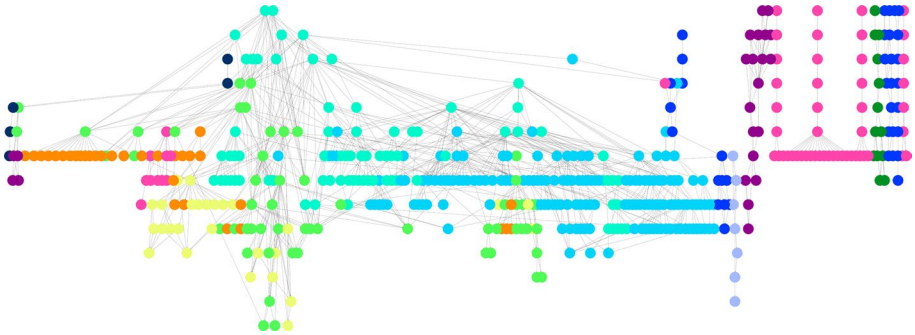


Fig. 5 LPIG at SIUC

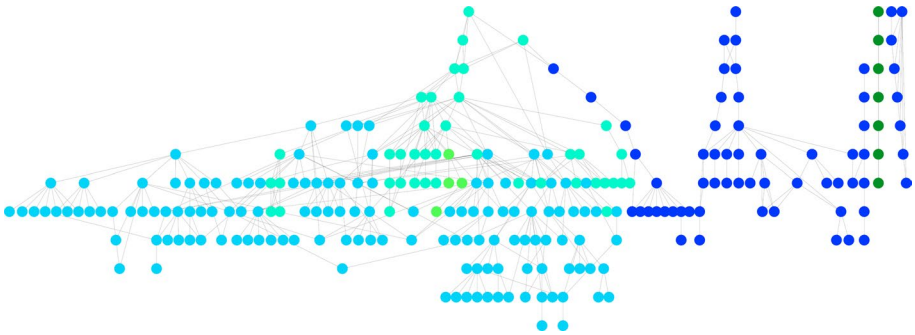


Fig. 6 LPIG at UMKC

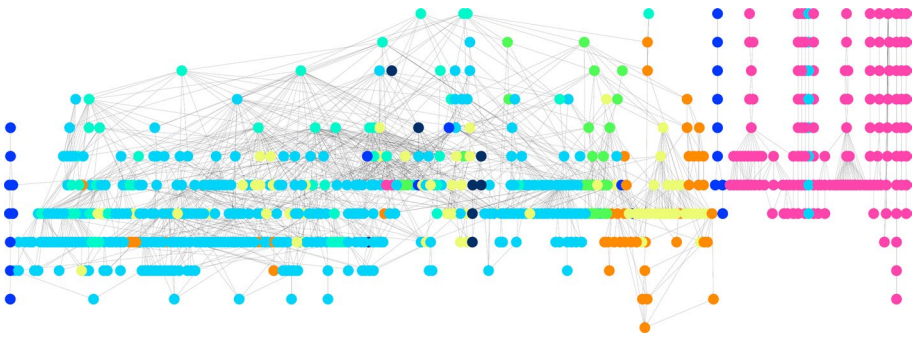


Fig. 7 LPIG at UIUC

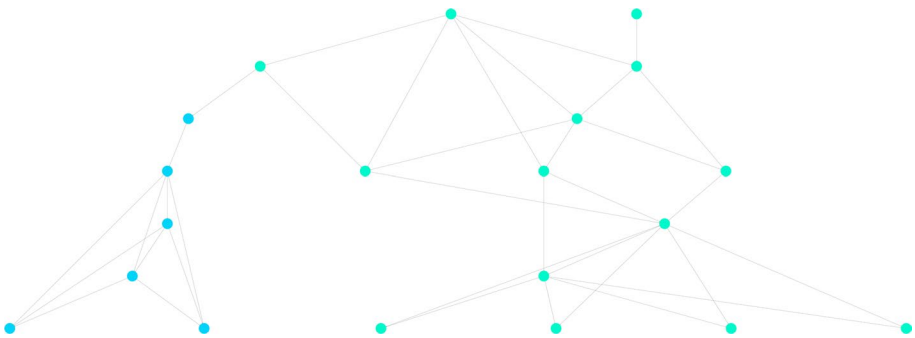


Fig. 8 LPIG at CalTech

Table 8 The highest betweenness centrality nodes in CGs at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
PHILOS	MATH	MATH	MATH	ANTH	BI
COMPENG	STAT	CI	MECENGR	PS	ACM
MATH	PHYS	PSYC	STAT	AFST	AE
HISTORY	IE	AFR	BIOLOGY	PSYC	CH
ENGMGT	CS	SCI	ACCTNG	NRES	CHE
CIVENG	ENSC	PLB	DSOM	REL	CS
COMPSCI	CHEM	ENGL	HLSC	PHIL	ME
ART	MS	BIOL	CIVENGR	CS	PH
SPMS	GEOG	HIST	CHEM	ECE	EE
ELECENG	CIED	ANS	ECON	MACS	APH

education, biology specializations, and agriculture. In fact, closer inspection of SIUE's LPIG indicates that several biology specializations and agriculture related courses are found within the green biology and chemistry category though their course codes are simply biology. UIUC also has a relatively diverse LPIG as it includes all categories except for nursing, military, business, and education. The LPIG of UMKC contains only the categories of mathematics and physics, engineering, business related, military and aviation, and music and the arts. However, UMKC's LPIG is notable in the significant size of its music and arts related subnetwork. MST's LPIG network is the least diverse amongst the public institutions considered, involving only the categories of mathematics and physics, engineering, biology and chemistry, and business related programs. Nonetheless, it is the second largest LPIG comprising 584 nodes. CalTech's LPIG of only 20 courses includes only the categories of mathematics and physics, and engineering.

Curriculum graph analysis

The Curriculum Graph (CG) is defined on the set of majors and extracted from the course dependency information encoded by the CPN as described in Sect. 2.2.2. The CG allows us to study macroscopic relationships between and among disciplines in the overall curricular landscape. Our analyses of the CG includes both the extraction of important majors with respect to network centrality metrics and the inference of groupings of majors with applications to meta-majors.

Centralities of major fields

Tables 8 and 9 list the majors obtaining the highest betweenness centralities and highest out-degrees among the institutions studied.

Fields achieving high betweenness centrality in a CG by definition have a higher tendency to *connect* other fields to each other, hence corresponding to some quantification of interdisciplinarity (Stavrinides and Zuev 2023). That the relative inter-disciplinarity of fields can vary according to the institutional context is demonstrated by the columnar variation of Table 8. While mathematics appears to connect many majors at MST, SIUE, SIUC, UMKC, and CalTech, mathematics does not appear amongst the highest betweenness major nodes of the UIUC CG, which is instead dominated by majors in the humanities and social sciences. Likewise, although mathematics and engineering majors

Table 9 The highest out-degree nodes in CGs at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
MATH	MATH	MATH	MATH	MATH	MA
CIVENG	CHEM	CHEM	ENGLISH	CWL	ACM
STAT	ENG	BIOL	STAT	GWS	PH
MECHENG	STAT	PLB	CHEM	REL	CH
COMPSCI	BIOL	AFR	PHYSICS	HIST	CMS
PHYSICS	CIED	MICR	MECENGR, BIOLOGY	CS	BI
AEROENG,ARCHENG	CE,ME	MBMB	CIVENGR, COMPSCI	ENGL	AE
ELECENG	CS	ENGR	ECENGR, HLSC	AFRO	CS
ENVENG, CHEM	MS, IE	ZOOL	DSOM, GEOG	LLS	ME
COMPENG	ENSC	CSEM	ENVSCI, LSANATO, ACCTNG	ECE	APH

are dispersed amongst MST's highest betweenness fields, humanities majors such as philosophy, history, speech and media studies, and art also occupy positions of high interdisciplinary importance. On the other hand, the highest betweenness centrality majors at SIUE and UMKC are almost entirely STEM fields. Given that MST and UIUC are well-known for their STEM programs, a possible explanation for the relatively higher betweenness of some of the non-STEM majors at those institutions may be greater curricular interaction between their STEM and non-STEM programs.

Like the case of CG betweenness distributions, math again features prominently in the out degree distributions of the institutional CGs shown in Table 9, reconfirming the critical importance of the subject in the curricular landscape overall. Out degree distributions appear more STEM oriented overall across the different institutions, including MST. However, UIUC once again includes a notable proportion of humanities majors achieving importance with respect to out degree in its CG.

Meta-majors and CG clustering

Our last network analysis concerns unsupervised inference of major groupings to better understand inter-major relationships and the macroscopic curricular landscape. Generally speaking, majors tend to be associated with specific departments or schools of an institution, and those departments are often further organized into schools or colleges of the institution. While such administrative subdivisions may reflect some of the natural groupings of the underlying fields of knowledge with which they are associated, they do not necessarily provide an accurate reflection the relationships between and amongst the curricular paths. However, relationships between curricula themselves are very important, especially as a non-negligible portion of incoming freshmen arrive with undeclared major, and another non-negligible portion of those who have declared a major switch to another major. Of course, such changes may have adverse effects on timely completion of studies when the switch occurs too late or between fields whose curricular landscapes are too dissimilar. Therefore, more accurate inference of appropriate inter-major groupings with respect to curricular relationships is useful information to present to students and advisors from the start, towards aiding in curricular planning. With precisely such concerns in mind, SIUE has implemented a system in which incoming freshmen choose

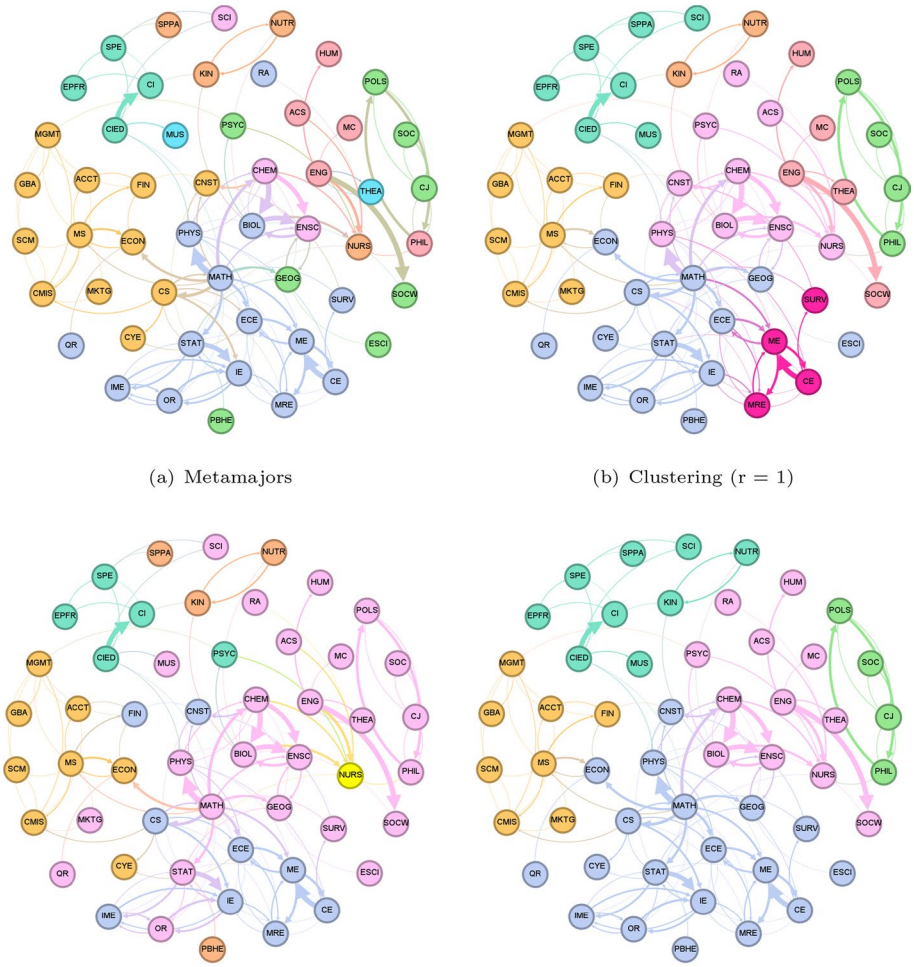
Fig. 9 SIUE metamajor color map

a general *meta-major*, which corresponds to a related group of majors, rather than specific major in their first year.

As stated in SIUE's advising website SIUE Meta-Majors (2024), first-year students are grouped into 8 meta-majors according to their stated interests, for purposes of advising and tracking: *"Meta-Majors are combinations of academic majors from different academic areas with related courses that fit within a career area. With a Meta-Major, students can explore major choices by initially following a standard first-year curriculum, and then, when they decide on their major, a four-year educational plan is followed to complete a degree without losing time and money."* It should be noted that any given meta-major is not necessarily contained in the same academic administrative unit or school within the institution but is rather constructed with commonalities in knowledge areas, skills, related careers, and student interests in mind. The names of the eight meta-majors are given in Fig. 9, which also shows meta-major color codes used in later figures.

As student retention, persistence, and timely graduation are amongst the important issues that the institution continually examines, a hope concerning meta-majors is that there should be *sufficient connectedness between majors of a given meta-major* so that a student starting out with unofficial declaration in one major of a meta-major may have the opportunity to switch to another major in the same meta-major without great waste of time and credits if the change of heart is detected soon enough. We model this property in the language of complex networks as the problem of *community detection*, also called *graph partitioning* or *clustering*, in the CG, which is the curricular network of majors derived from the CPN. This modeling is motivated by the fact that the intra-meta-major connectivity requirement for meta-majors is precisely captured by the community detection objective that the connectivity within a community be notably higher than the connectivity between communities (Girvan and Newman 2002). More simply, clustering is the problem of unsupervised inference of natural groupings of related nodes, and meta-majors are groupings of related majors. Hence, it is worthwhile to examine the relationship between the resultant clusters of the CG and the actual meta-major subdivisions.

Figure 10 provides the visualization of exactly such a comparison for the largest connected component (LCC) of the SIUE CG. The meta-majors of the LCC can be seen in part (a). The modularity based clusterings of the LCC are given in parts (b) and (d), where (b) is based on the default resolution setting. And, the administrative groupings of SIUE departments into colleges and schools can be seen in part (d). The relative matching between the meta-major groupings at SIUE and the default clustering of the SIUE CG is immediately apparent from parts (a) and (b), especially when compared to the administrative subdivisions of part (c). This is partly due to the large proportion of university departments contained in the College of Arts and Sciences (CAS) at SIUE. The meta-majors capture some of the finer-tuned groupings within CAS which are also yielded by the clustering. For example, the green political science, sociology, and criminal justice



(a) Metamajors

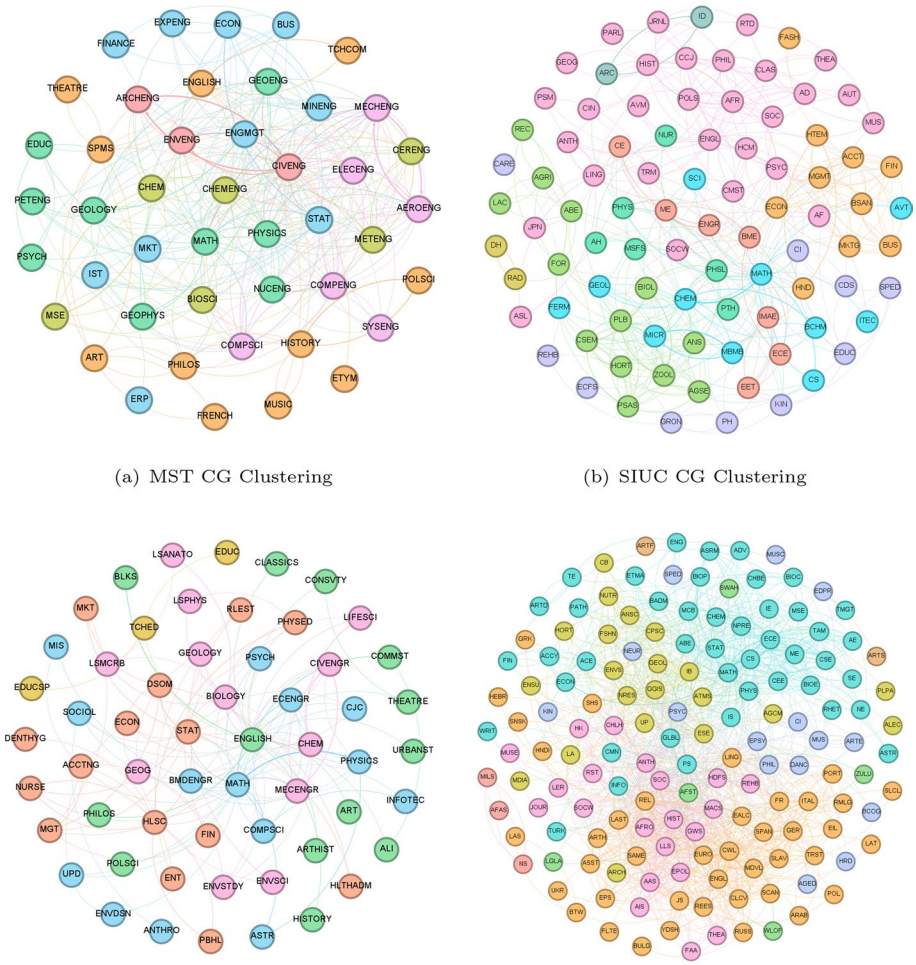
(b) Clustering (r = 1)

(c) Administrative subdivisions

(d) Clustering (r = 2.3)

Fig. 10 SIUE Curriculum Graph

majors are grouped together in both the clusterings and the meta-majors though they all reside in CAS. Ditto the pink digital humanities, mass communications, and english majors which are grouped together in both the default clustering and the meta-major groupings. Likewise, the engineering, math, and physics meta-major straddles both the School of Engineering (SoE) and CAS, but corresponds well with the clustering at resolution 2.3, which does not separate out the highly intra-connected mechanical engineering related sub-group inferred by the 1.0 resolution clustering. Some groupings of courses are stable across all of the methods of categorization, especially the majority of business related majors and the majority of engineering majors. And, as a rare exception, the teaching and learning meta-major is actually better captured by the administrative division of Education, Health, and Human Behavior than by the default clustering which includes three more majors in addition to the existing four of that meta-major. To be clear, we do not suggest that the clustering replace the meta-major groupings as the CG is not based on the finer tuned information of precise degree completion requirements of each major but only structural inter-major dependency information. However, the



(a) MST CG Clustering

(b) SIUC CG Clustering

(c) UMKC CG Clustering

(d) UIUC CG Clustering

Fig. 11 Curriculum Graph Clustering

relatively high similarity between the CG clustering and the meta-major groupings does provide macroscopic validation of the meta-major constructions.

As the institutions other than SIUE studied in this work do not appear to use advisement policies utilizing specific meta-major mappings, we are unable to analyze published meta-majors at the other institutions. Nonetheless, due to the similarity between the meta-majors and CG clustering at SIUE, we provide visualizations of the clustering of the largest connected components (LCCs) of the CGs of the other public institutions in Fig. 11.

Discussion, conclusion, and future work

We have presented the first formal network analysis of curricular networks for public institutions, focusing around 5 midwestern universities. As a first such study, our analyses are primarily macroscopic in nature, observing patterns in the overall CPN and CG networks. In addition to computing well known network centrality measures to capture courses of importance in the CPNs studied, we have also formulated some newer

methods with specific relevance to the curricular domains and corresponding graph types at hand. For example, we have proposed reach as a measure of special relevance for DAGs such as the CPNs studied herein. Another method that we propose with specific meaning and computational feasibility for the CPNs is the longest paths induced subgraph (LPIG) which yields information on relatively constrained programs and pathways. Finally, we have established a new connection between clustering of the CG and meta-majors at SIUE, providing clusterings of the other public institution CGs as useful heuristics of major groupings as well.

While certain subsets STEM courses and majors unsurprisingly feature prominently amongst the critical nodes of the CPN, CG, and LPIG networks, the consistency with which some such patterns were echoed across all studied institutions is a testament to the power of network science methods in automatically extracting such well-attested information. More interesting, perhaps, are the humanities and non-STEM courses and programs which also featured prominently in some of the curricular networks, especially for institutions well-known for their top ranking engineering programs such as UIUC. While engineering programs are known to have especially long course chains, as also confirmed by their dominant place in all the institutional LPIGs, again it was the diversity of course categories, both STEM and non-STEM, in the LPIG networks that we found to be interesting.

Looking more closely at patterns concerning non-engineering course chains which arise in the LPIG networks, the institutional variance appears to indicate some specializations offered at the respective institutions which are not widely offered in general institutions. Such examples include music education at SIUE which is well known for its Suzuki programs, the several biological specializations at SIUC which has strong introductory medical pathways leading to the SIU medical school, the highly rated music conservatory programs at UMKC, and the French, Spanish, and well-known agriculture programs at UIUC which complement that institution's enormous engineering sub-networks.

Another aspect of the LPIG networks that was surprising to us was the vast discrepancy between the length and number of very long paths in the CPNs of all of the public schools when compared to the CPN of CalTech. While we have primarily included CalTech in our analyses to compare with the prior work of Stavrinides and Zuev (2023), the vast discrepancy in this regard is more striking when considering the 12 term standard degree completion time for CalTech versus the 8 term standard in the other institutions considered. We know that a large part of this difference is due to the existence of introductory mathematics sequences from College Algebra to Pre-Calculus in the public institutions which many freshmen who seriously pursue STEM majors will test out of via the ALEKS placement exam, testing directly into some level of Calculus instead.

Such courses do not exist in the CalTech curriculum as CalTech has required its incoming freshmen to have taken a year of Calculus in highschool or to have demonstrated mastery of the same, in addition to similar requirements with physics. Of course, students at any U.S. higher education institution may skip introductory subpaths across different parts of the curricular network anyway due to AP, IB, or Honors courses in highschool, or other forms of equivalent advancement accepted from the perspective of a given institution. However, for a variety of reasons including equity

concerns, public institutions which serve their states and local regions generally cannot require their students to have already taken calculus before admission given the small minority of American highschool students who do so. It is this condition of K-12 mathematics education in the U.S. that causes a course such as College Algebra, which is of remedial content as far as university mathematics courses are concerned, to be an actual bottleneck in the advancement of a non-negligible portion of university students with negative impact on retention and graduation (Goonatilake et al. 2013). Nonetheless, even putting aside a total of two to four introductory mathematics and science courses in all of the longest paths across the public institutions, the lengths and number of the prefix-trimmed longest paths at the public institutions would still be significantly more than those at CalTech, especially when controlling for standard undergraduate degree length. Given that CalTech is a premiere STEM oriented higher education institution with many advanced offerings in engineering and the natural sciences, we believe that this discrepancy warrants further investigation as the magnitude of such curricular pathways may affect timely graduation and persistence.

Indeed, we remind that timely graduation and persistence questions in an institutional working group precisely formed the original motivation for this collaboration though the public data analyzed in this study could not directly be used towards statistically validated conclusions in that regard. Rather, we re-emphasize that our LPIG analysis concerns *lower bounds* on graduation times and *rigidity* of curricula over-represented in the LPIG. Taken together with our evidence for meta-major cohesion, this information can be used to better advise students in our respective institutions and beyond. For example, Electrical Engineering and Mechanical Engineering are consistently over-represented in the LPIGs of all institutions. They are also in the cohesive Engineering, Math, and Physics meta-major which also includes Statistics, Industrial Engineering, and other majors not represented strongly in the LPIGs. An incoming freshman interested in the Engineering, Math, and Physics meta-major without having tested into (or beyond) Calculus should be encouraged to pursue the programs such as Statistics and Industrial Engineering which yield some hope of graduation within six years. Given that the 8 year graduation rates at MST, SIUE, SIUC, UMKC, UIUC, and CalTech are 71%, 57%, 56%, 56%, 87%, and 94%, respectively (U.S. Department of Education 2024), we cannot over-emphasize the importance of leading students towards feasible graduation paths as soon as possible in public institutions. In conjunction with such advising considerations, might academic units also benefit from making graduation paths slightly more feasible given that prerequisite structure is both inherent and human-made? Without sacrificing the most universally agreed upon prerequisite relationships, some reconsideration of excessive prerequisite structures in the most rigid of curricula might broaden participation in those fields.

Future directions investigating the relationship between student outcomes and CPN pathways must involve finer tuned analysis of degree pathways in addition to longitudinal course statistic of the students in the programs. Towards this end, we wish to examine the structure of specific degree programs within and among representative institutions. As the area of formal network theoretic analyses of curricular pathways is still relatively new with a small sample of networks publicly available, it

is also worthwhile to continue providing macroscopic analyses of the CPN and CG networks of a much wider selection of institutions. Many open questions remain with respect to wider ranging institutional patterns. We hope that this work will contribute to motivating such future directions in curricular network analyses.

Appendix

See Table 10, 11, 12, 13, 14, 15, 16.

Table 10 MST course list

Course code	Course name	Course code	Course name
MATH1120	College Algebra	ENGLISH1120	Exposition & Argumentation
MATH1140	College Algebra	CIVENG2200	Statics
MATH1160	Trigonometry	CHEM1310	General Chemistry I
MATH1208	Calc. with Anal. Geometry I	HISTORY1200	Modern Western Civilization
MATH1210	Calculus I-A	HISTORY1300	American History to 1877
MATH1211	Calculus I-B	HISTORY1310	American History since 1877
MATH1214	Calculus I	POLSCI1200	American Government
MATH1215	Calculus II	PHYSICS1135	Engineering Physics I
MATH1221	Calc. with Anal. Geometry II	PHYSICS2135	Engineering Physics II
MATH2222	Calculus III	STAT3115	Engineering Statistics
MATH3304	Elementary Diff. Equ.	STAT3117	Intro. to Prob. & Stat.

Table 11 SIUE course list

Course code	Course name	Course code	Course name
MATH120	College Algebra	ENG102	English Composition II
MATH125	Pre-Calc. Math. with Trig.	ENG102N	English Composition II (Non-Native)
MATH145	Calc. for the Life Sci.	BIOL220	Genetics
MATH150	Calculus I	CE240	Statics
MATH152	Calculus II	CE242	Mechanics of Solids
MATH250	Calculus III	PHYS141(M)	Physics I for Engineers
CHEM121A	General Chemistry	PHYS140	Intro. to Phys. & Physical Reasoning
CHEM121B	General Chemistry	PSYC111	Foundations of Psychology
CHEM113	Intro. to Chemistry	CIED100	Introduction to Education

Table 12 SIUC course list

Course code	Course name	Course code	Course name
MATH106	College Algebra Enhanced	MATH250	Calculus II
MATH108	College Algebra	MATH305	Introduction to Differential Equations
MATH109	Trig. & Anal. Geometry	PSYC102	Introduction to Psychology
MATH111	Precalculus	CHEM140A	Chemistry
MATH125	Technical Math. with App.	CHEM200	Intro. to Chemical Principles
MATH139	Finite Mathematics	CHEM210	General & Inorganic Chemistry
MATH140	Short Course in Calculus	CHEM330	Quantitative Analysis
MATH150	Calculus I	ENGL101	English Composition I
MATH151	Calculus I Enhanced	ENGL102	English Composition II

Table 13 UMKC course list

Course code	Course name	Course code	Course name
MATH110	Precalculus Algebra	PHYSICS250	Physics for Sci. & Eng. II
MATH120	Precalculus	CHEM211	General Chemistry I
MATH125	Trigonometry	CIVENGR275	Engineering Statics
MATH210	Calculus I	CIVENGR276	Strength Of Materials
MATH220	Calculus II	ECENGR276	Circuit Theory I
MATH266	Accelerated Calculus I	ENGLISH110	Intro. to Academic Prose
MATH268	Accelerated Calculus II	ENGLISH225	English II: Inter. Academic Prose
PHYSICS240	Physics for Sci. & Eng. I		

Table 14 UIUC course list

Course code	Course name	Course code	Course name
MATH112	Algebra	CHEM102	General Chemistry I
MATH220	Calculus	CHEM104	General Chemistry II
MATH221	Calculus I	CHEM202	Accelerated Chemistry I
MATH231	Calculus II	ECE210	Analog Signal Processing
MATH241	Calculus III	ECON102	Microeconomic Principles
MATH257	Lin. Alg. with Computational App.	ECON302	Inter Microeconomic Theory
MATH285	Intro. Differential Equations	PHYS211	University Physics: Mechanics
MATH415	Applied Linear Algebra	PHYS212	University Physics: Elec & Mag
CS101	Intro Computing: Eng. & Sci.	PSYC100	Intro. Psych.
CS225	Data Structures	TAM251	Intro. Solid Mechanics
STAT400	Statistics and Probability I		

Table 15 CalTech course list

Course code	Course name	Course code	Course name
MA1ABC	Calc. of One & Several Var. & Lin. Alg.	BI8	Intro. to Molecular Biology
MA3103	Intro. to Probability & Statistics	CH1AB	General Chemistry
MA2102	Differential Equations	CH21ABC	Physical Chemistry
MA5105ABC	Intro. to Abstract Algebra	CH41ABC	Organic Chemistry
CS1	Intro. to Computer Programming	PH1ABC	Classical Mech. & Electromag.
CS2	Intro. to Programming Methods	PH12ABC	Waves, Quantum Phys., & Stat. Mech.
ACM11	Intro. to Computational Sci. & Eng.	PH125ABC	Quantum Mechanics
ACM95/100AB	Intro. Meth. of App. Math. for Phys. Sci.	PH2ABC	Waves, Quantum Mech., & Stat. Phys.

Table 16 The highest in-degree nodes in CGs at each institution

MST	SIUE	SIUC	UMKC	UIUC	CalTech
MECHENG	ENSC	PLB	MECENGR	CWL	GE
CIVENG	IE	CHEM	PHYSICS	GWS	AE
AEROENG	ME	MBMB	BIOLOGY, CIVENGR	CS	CS
ARCHENG	NURS	ZOOL	ASTR	HIST	EE
ELECENG	BIOL	AFR	ECENGR	REL	BE
COMPENG	PHYS	MICR	BMDENGR	ECE	CMS
COMPSCI	ECE	PSAS	HLSC	ENGL	AY
ENVENG	CI	PHSL	COMPSCI, GEOLOGY, MGT, LIFESCI	AFRO	APH, CHE
CHEMENG	MRE	CSEM	MATH, BLKS	NRES	BI
MINENG	SOCW	ME	DSOM	LLS	PH

Abbreviations

CG	Curriculum Graph
CPN	Course prerequisite network
DAG	Directed acyclic graph
LPIG	Longest paths induced sub-grap
STEM	Science, technology, engineering, mathematics
tPR	Transpose PageRank
MST	Missouri University of Science and Technology
SIUE	Southern Illinois University Edwardsville
SIUC	Southern Illinois University Carbondale
UMKC	University of Missouri Kansas City
UIUC	University of Illinois Urbana Champaign
CalTech	California Institute of Technology

Acknowledgements
GE acknowledges partial support from the Southern Illinois University Edwardsville Center for Predictive Analytics as relates to this work.

Author Contributions
BY has led the software implementation and application of all methods involved in this work, data acquisition, and reporting and visualization of all results. MG has substantially contributed to the writing and revision of this work, the data acquisition concerning MST, and the presentation, interpretation, and validation of all results. HRR contributed substantially to the data acquisition, execution of network science experiments, and reporting of the results. EH made substantial contributions to the conception of parts of the work, particularly regarding meta-majors at SIUE. EH also provided invaluable guidance on the interpretation of some results as they relate to student outcomes and observed patterns in education research. EZL and XH contributed to the data acquisition of SIUC data in addition to providing early stage guidance on administrative subdivisions and degree categories at that institution. GE is the primary architect of this work, as she formulated the original curricular questions in the language of graph theory, designed the methodology, analyzed and interpreted the results, and contributed most significantly to the writing.

Funding
Not applicable

Availability of Data and Materials
All data in this article may be found in Yang (2024).

Declarations

Conflict of interest
Not applicable

Received: 30 April 2024 Accepted: 21 June 2024
Published online: 26 June 2024

References

Aldrich PR (2015) The curriculum prerequisite network: Modeling the curriculum as a complex system. Biochem Mol Biol Educ 43(3):168–180

Blondel VD, Guillaume J-L, Lambiotte R, Lefebvre E (2008) Fast unfolding of communities in large networks. J Stat Mech Theory Exp 2008(10):10008

Boldi P, Santini M, Vigna S (2009) Pagerank: functional dependencies. ACM Trans Inform Syst (TOIS) 27(4):1–23

Brandes U (2001) A faster algorithm for betweenness centrality. J Math Sociol 25(2):163–177

Burke M (2020) Progress on guided pathways promising, but still much to do, report says, Inside Higher Ed. Available at: https://www.insidehighered.com/news/2020/09/15/progress-guided-pathways-promising-still-much-do-report-says Accessed 20-06-2024

BYU-Idaho (2024) CS Course dependency charts. https://www.byui.edu/computer-science-engineering/student-resources/course-dependency-charts

Cormen TH, Leiserson CE, Rivest RL, Stein C (2022) Introduction to algorithms. MIT press, London

U.S. Department of education (2024) College scorecard data. Available at: https://collegescorecard.ed.gov/data/ Accessed 20-06-24

Freeman LC (1977) A set of measures of centrality based on betweenness. Sociometry 40(1):35–41

Garey MR, Johnson DS, Stockmeyer L (1974) Some simplified np-complete problems. In: Proceedings of the sixth annual ACM symposium on theory of computing pp 47–63

Girvan M, Newman ME (2002) Community structure in social and biological networks. Proc Natl Acad Sci 99(12):7821–7826

Goonatilake R, Nguyen TN, Bachnak RA, San Miguel M, Garza AC (2013) All for the success of college algebra. Math Teach Res J 6:21

Hagberg A, Swart P, S Chult D (2008) Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States)

Heymann S, Le Grand B (2013) Visual analysis of complex networks for business intelligence with Gephi. In: 2013 17th international conference on information visualisation, pp. 307–312. IEEE

IES NCES (2024) Undergraduate enrollment. https://nces.ed.gov/programs/coe/indicator/cha/undergrad-enrollment

Macalester College (2024) Physics and astronomy course dependency charts. https://www.macalester.edu/physics/majorsminors/coursedependencycharts/

Page L, Brin S, Motwani R, Winograd T (1999) The pagerank citation ranking : bringing order to the web. In: The Web Conference. https://api.semanticscholar.org/CorpusID:1508503

Sedgewick R, Wayne K (2011) Algorithms, Edition 4. Addison-Wesley Professional

SIUE Civil Engineering Course (2024) Dependency chart. https://www.siue.edu/engineering/civil-engineering/current-students/undergraduate/pdf/2022-6-14-PrereqFlowchartsEnrolledAfterFall2019.pdf

SIUE Computer Engineering Course (2024) Dependency chart. https://www.siue.edu/engineering/ece/img/CompE_Dependency_Graph_Descriptions.html

SIUE Electrical Engineering Course (2024) Dependency chart. https://www.siue.edu/engineering/ece/img/EE_Dependency_Graph_Descriptions.html

SIUE meta-majors (2024) at SIUE. https://www.siue.edu/oaa/get-advised/Metamajors.shtml

Stavrinides P, Zuev KM (2023) Course-prerequisite networks for analyzing and understanding academic curricula. Appl Netw Sci 8(1):19

Washington and Lee University (2024) Math course dependency chart. https://my.wlu.edu/Documents/mathematics/math-course-dependency-chart.pdf

Wellesley College (2024) CS course prerequisite diagram. https://cs.wellesley.edu/~cs/Curriculum/dependencies.html

Yang B (2024) CPN data. https://github.com/BonanYang/Comparative-Analysis-of-Course-Prerequisite-Networks-for-Five-Midwestern-Public-Institutions

Publisher's Note