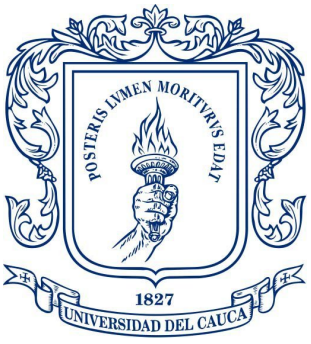
Contribuciones al descubrimiento de patrones de comportamiento de estudiantes en plataformas MOOC



Propuesta de Grado de Maestría

Ing. Luis Alejandro Cruz Ordóñez

Director: Ph.D. Mario Fernando Solarte

Universidad del Cauca

Facultad de Ingeniería Electrónica y Telecomunicaciones

Maestría en Ingeniería Telemática

Línea de Investigación en Tele-Educación

Popayán, Octubre de 2022

Página de Aceptación

Firma del jurado 1 - Evaluador

Firma del jurado 2 - Evaluador

Firma del Director

Resumen Estructurado

Este trabajo de investigación presenta una aproximación de la representación de las rutas de aprendizaje de estudiantes en plataforma en línea masivas. El desarrollo del mismo aportó un modelo de representación para aplicarlo en plataformas MOOC, una herramienta de visualización para docentes y estudiantes que quieran identificar las rutas en cada módulo y sesión de clase. La adaptación de algoritmos para identificar patrones de comportamiento y la aplicación de un caso de estudio en la universidad del Cauca.

Los eventos de los estudiantes en la plataforma, específicamente los relacionados con actividades en vídeos, foros y evaluaciones; se pudieron representar secuencialmente en grafos dirigidos cuyos vértices fueron el tipo de actividad realizada y las aristas la transición o saltos entre los contenidos. Para la construcción de los grafos se tuvo como base un modelo de representación propuesto en este trabajo y adaptado a los datos recolectados de la actividad de clickstream de los estudiantes en las plataformas. El proceso de recolección de la actividad de los estudiantes se basó en un trabajo previo de Maestría de la Universidad del Cauca del estudiante Daniel Jaramillo quień logró implementar n recolector de eventos extraídos de archivos log de las plataformas.

El modelo propuesto permite identificar 4 tipos de grafos diferentes: IW: individual por periodo, IWS: individual por periodo y por sesión, IS: Individual por sesión e GW: grupal por período. La investigación logró la transformación de las interacciones de los estudiantes con MOOCs en grafos que pueden representar patrones de comportamiento con una base de datos centradas en grafos.

Fue posible también desarrollar una herramienta de visualización, orientada para que los docentes puedan revisar la interacción de sus estudiantes con los contenidos del curso. Se definió una arquitectura software, un modelo de procesado enfocado en grafos con el fin de que los docentes puedan interpretar fácilmente las rutas de aprendizajes por la que los estudiantes optaron y se desplegó en un servidor Web para que el docente pueda acceder a la herramienta desde cualquier lugar con conexión a internet.

Pensando en una interpretación de los grafos obtenidos, se propuso un análisis estadístico proponiendo algoritmos para la comparación de las rutas de los estudiantes frente a la ruta definida por el docente. También se presentó una forma de identificar los saltos más comunes entre los contenidos en cada período junto con el cálculo de las medidas de centralidad que son posibles obtener con la teoría de grafos.

Palabras claves: MOOC, Rutas de navegación, Análisis de aprendizaje, Análisis de datos, Grafos.

Structured Abstract

This research work presents an approximation of the representation of student learning paths in a massive online platform. Its development provided a representation model to apply it on MOOC platforms, a visualization tool for teachers and students who want to identify the routes in each module and class session. The adaptation of algorithms to identify behavior patterns and the application of a case study at the University of Cauca.

Student events on the platform, specifically those related to video activities, forums, and assessments; they could be represented sequentially in directed graphs whose vertices were the type of activity carried out and the edges the transition or jumps between the contents. The construction of the graphs was based on a representation model proposed in this work and adapted to the data collected from the clickstream activity of the students on the platforms. The process of collecting the activity of the students was based on a previous work of the Master's Degree of the University of Cauca of the student Daniel Jaramillo who managed to implement an event collector extracted from the log files of the platforms.

The proposed model allows identifying 4 different types of graphs: IW: individual by period, IWS: individual by period and by session, IS: Individual by session and GW: group by period. The research achieved the transformation of student interactions with MOOCs into graphs that can represent behavioral patterns with a graph-centric database.

It was also possible to develop a visualization tool, oriented so that teachers can review the interaction of their students with the course contents. A software architecture was defined, a processing model focused on graphs so that teachers can easily interpret the learning paths that students chose and it was deployed on a Web server so that the teacher can access the tool from anywhere with an internet connection.

Thinking about an interpretation of the obtained graphs, a statistical analysis was proposed proposing algorithms for the comparison of the students' routes against the route defined by the teacher. A way to identify the most common jumps between the contents in each period was also presented, together with the calculation of the centrality measures that are possible to obtain with graph theory.

Keywords: MOOC, Learning Paths, Data Analytics, Learning Analytics, Graphs.

Índice general

1. Introducción 2
   1.1. Planteamiento del problema 3
2. Marco Conceptual 6
   2.1. Massive Online Open Courses-MOOCs 6
   2.2. Arquitectura MOOC 7
      2.2.1. Heterogeneidad Estudiantil 8
      2.2.2. Learning Analytics 9
   2.3. Grafos 10
      2.3.1. Conceptos preliminares 10
      2.3.2. Bases de Datos de Grafos 11
      2.3.3. Patrones en grafos a partir de Clickstream 13
3. Estado del arte 14
   3.1. Selección de trabajos 14
   3.2. Análisis de comportamiento en MOOC 16
   3.3. Descubrimiento de patrones de secuencia 18
   3.4. Herramientas de Visualización 20
   3.5. Aporte Investigativo 23
4. Objetivos 24
   4.1. Objetivo General 24
   4.2. Objetivos Específicos 24
   4.3. Metodología 25
5. Representación basado en grafos de las actividades de aprendizaje de los estudiantes en Open edX 26
   5.1. Fundamentos de modelado de interacción 26
      5.1.1. Estructura y actividades de un curso 26
      5.1.2. Modelado de grafos 27
   5.2. Diseño e implementación del modelo de representación 28
      5.2.1. Open EdX 30
   5.3. Modelado de Procesado 31
   5.4. Moockly 32
6. Similitud de grafos según el comportamiento de los estudiantes 35
   6.1. Centralidad 36
      6.1.1. Matriz de adyacencia y probabilidad 37
      6.1.2. Cálculo de centralidad 39
   6.2. Similitud de rutas 41
      6.2.1. Distancia de Levenshtein 41
   6.3. Patrón de comportamiento 43

6.4. Patrones de estados 49
7. Validación y evaluación del modelo y similitud de grafos en un curso en línea con reconocimiento académico 51
7.1. Contexto Educativo y Recopilación de Datos 51
7.2. Grafos generados 51
7.3. Análisis de resultados 57
7.4. Exploración de patrones 61
7.4.1. Patrones de Navegación Lineal 62
7.4.2. Patrones de Navegación No lineal 63
8. Conclusiones y Trabajos futuros 74

Índice de figuras

1. Tendencia en plataformas MOOC 7
2. Recursos MOOC 8
3. Representación de un grafo 11
4. Metodología 25
5. Diagrama de clases 29
6. Arquitectura Software 30
7. Grafo de período 38
8. Camino ideal 43
9. Múltiples estudiantes 45
10. Patrón de saltos 48
11. Módulos del curso 52
12. Filtro para grafo IWS 53
13. Filtro para grafo IW 53
14. Filtro para grafo IS 53
15. Grafo IWS estudiante activo, sesión 1 54
16. Grafo IWS estudiante activo, sesión 2 54
17. Grafo IWS estudiante inactivo 55
18. Grafo IW 56
19. Grafo IS estudiante activo 57
20. Módulos del curso 58
21. Contenidos visitados por módulo 59
22. Contenidos visitados totales 59
23. Centralidad y Visitas módulo 3 61
24. Navegación Lineal 1 62
25. Navegación Lineal 2 63
26. Navegación Lineal 3 64
27. Navegación Lineal 4 64
28. Navegación Lineal 5 65
29. Navegación No Lineal Vídeos Continuos 1 65
30. Navegación No Lineal Vídeos Continuos 2 66
31. Navegación No Lineal Vídeos Continuos 3 67
32. Navegación No Lineal Vídeos Continuos 4 67
33. Navegación No Lineal Nodos Discontinuos 1 68
34. Navegación No Lineal Nodos Discontinuos 2 68
35. Navegación No Lineal Nodos Discontinuos 3 69
36. Navegación No Lineal Nodos Discontinuos 4 70
37. Navegación No Lineal múltiples lazos 1 70

38. Navegación No Lineal múltiples lazos 2 71
39. Navegación No Lineal múltiples lazos 3 71
40. Navegación No Lineal múltiples lazos 4 72
41. Navegación No Lineal múltiples lazos 5 72
42. Navegación No Lineal múltiples lazos 6 73
43. Navegación No Lineal Sin Quiz 1 73
44. Navegación No Lineal Sin Quiz 2 74

Índice de tablas

1. GDB vs RDBMS 12
2. Documentos encontrados 15
3. Documentos finales. 15
4. Artículos relevantes 17
5. Artículos relevantes 21
6. Columnas Dataset 32
7. Medidas de centralidad 37
8. Centralidad de G 41
9. Distancia de Levenshtein 43
10. Muestra estudiantes 46
11. Resumen estadísticas de patrones 74

1. Introducción

Desde el 2012, la oferta de MOOCs (Massive Open Online Courses) ha venido creciendo a través de diferentes plataformas para soportar educación en línea en entornos de masividad. Los MOOCs se caracterizan por ser oportunidades académicas que superan las limitaciones de disponibilidad y distancia entre los estudiantes y los profesores. La principal razón de este incremento de uso es su bajo costo y la calidad de los cursos, poniéndose a disposición a personas de cualquier lugar del mundo suscribiéndose cada vez más todos los años. En este contexto, una de las problemáticas para los docentes, es no poder identificar el comportamiento, las características y las acciones de los estudiantes en su interacción con los diferentes contenidos. Por ejemplo, entre las muchas posibilidades, está el análisis de cómo se mueven los estudiantes por el curso, que puede permitir a los docentes la navegación que han seguido por los diferentes recursos en una sesión de clase. De los retos identificados en los MOOC, este es uno de los más significativos porque al conocer detalladamente la interacción de los alumnos con los cursos, se pueden proponer mejoras a la forma cómo los docentes diseñan los mismos.

Para afrontar este problema de la identificación de características propias de alumnos, se han explorado soluciones desde diferentes enfoques. Un enfoque es desde el análisis de sentimientos haciendo uso de la información que los estudiantes proveen en comentarios de foros, vídeos, chats, etc. [1, 2, 3]. Otro enfoque es desde el tiempo que los estudiantes le dedican a los cursos, el tiempo activo en las plataformas y la frecuencia con la que visitan el curso [4, 5]. Otro posibilidad es identificar perfiles desde sus resultados de calificaciones en pruebas y evaluaciones, relacionando estos resultados con la temática evaluada, proponiendo así un perfil característico del estudiante [6, 7]. Sin embargo, un enfoque muy poco explorado y que es precisamente el aquí propuesto es identificar los perfiles desde las rutas de navegación por las que los estudiantes optan [8, 9]; por ejemplo es posible identificar y recopilar los clickstream de cada sesión en las plataformas, cuando ingresa a una sección, cuando reproduce un vídeo o cuando lo para, cuando envía la respuesta de un examen, etc.

Con estos datos se puede generar un modelo con base a las secuencias de rutas de navegación del estudiante en el curso.

Este trabajo propone la aproximación de un modelo de representación del comportamiento estudiantes en una plataforma. y a partir del modelo permitir visualizar las rutas de navegación y descubrir patrones de comportamientos entre los estudiantes. El presente documento esta conformado por los siguientes capítulos: 1) Introducción, 2) Contexto, 3) Estado del arte, 4) Modelo de representación de rutas de aprendizaje 3) Similitud de grafos según el comportamiento de estudiantes. 4) Evaluación de resultados en un curso ofrecido en Open edX de la Universidad del Cauca.

1.1. Planteamiento del problema

En el 2008 se acuñó por primera vez el termino MOOC (Massive Online Open Courses) [10], como una vertiente significativa de E-learning [11] para referirse a los cursos ofrecidos en línea, abiertos y masivos. Desde el 2012 hasta la actualidad ha mostrado un crecimiento exponencial de su acogida a través de diferentes plataformas que soportan educación en línea en ambientes de masividad[12]. Los MOOCs se caracterizan por ser oportunidades académicas que superan limitaciones de disponibilidad y distancia entre los estudiantes y los profesores. Según el reporte de Class Central, durante el 2020 más de 19400 MOOCs fueron desarrollados en más de 950 universidades con un registro cerca de 220 millones de estudiantes [13]. La principal razón de este incremento de uso es su bajo costo y la calidad de los cursos, poniéndose a disposición a personas de cualquier lugar del mundo suscribiéndose cada vez más todos los años.

A pesar de los beneficios que han traído los MOOC a la sociedad, aún se presentan problemas de investigación abiertos, estos problemas se relacionan con la deserción académica, la deshonestidad, la adaptación tecnológica para usuarios acostumbrados a métodos convencionales, la heterogeneidad estudiantil, entre otras. La heterogeneidad estudiantil es uno de los problemas más grandes y poco estudiados que se centra específicamente en los perfiles de los estudiantes. Es claro que todos los estudiantes de un curso tienen un perfil diferente que se conforma de características propias según la región en la que se encuentra, edad, sexo, estilo de aprendizaje, cultura,

nivel académico y cualquier otra variable que haga parte de su perfil. Por eso, es difícil para los profesores tener un acompañamiento a los estudiantes y aún más, un diseño ideal de curso según los perfiles de los mismos. Este problema es posible afrontarlo desde diferentes disciplinas pero la más significativa han sido las ciencias de la computación[14]. De esta manera se han podido mejorar las plataformas con interfaces de usuario compuestas de recursos y componentes cada vez más completos para que los profesores los puedan utilizar con sus estudiantes.

Una de las soluciones aportadas para la heterogeneidad estudiantil desde las ciencias de la computación son los Sistemas de Aprendizaje Adaptativos (ALS), que es una contribución substancial de la inteligencia artificial (AI)[15] [16], generalmente los ALS organizan el contenido según las preferencias de aprendizaje de los alumnos individuales con el objetivo de maximizar el rendimiento del aprendizaje a través de una continua realimentación inteligente. Esto es posible implementarlo siempre y cuando se tenga un conocimiento de los perfiles de los alumnos y su comportamiento. Una actividad abarcada en el "E-Learning" para el acceso y gestión de la información es "Learning Analytics", en donde con técnicas y herramientas se logra recopilar las acciones desarrolladas por los estudiantes en el entorno web y procesando dicha información, poder inferir correlaciones y comportamientos que permiten entender mejor los procesos educativos para posteriormente mejorarlos. Gracias a esto, es posible la medición, recopilación y análisis de datos de interacción entre alumno y plataforma comprendiendo y optimizando el aprendizaje y sus entornos de producción.

La identificación de perfiles estudiantiles es un reto que los investigadores han aceptado afrontarlos desde diferentes enfoques. Un enfoque es desde el análisis de sentimientos haciendo uso de la información que los estudiantes proveen en comentarios de foros [17], vídeos, chats, etc. [18, 1, 2, 3]. Otro enfoque es desde el tiempo que los estudiantes le dedican a los cursos, el tiempo activo en las plataformas y la concurrencia con la que visitan el curso [4, 5]. Otro enfoque posible, es identificar perfiles desde sus resultados de calificaciones en pruebas y evaluaciones, relacionando estos resultados con la temática evaluada, proponiendo así un perfil característico del estudiante [6, 7]. Sin embargo, un enfoque muy poco explorado es identificar los perfiles desde las rutas de navegación por las que los estudiantes optan[8, 9].

En el contexto del aprendizaje en línea, existen estudiantes que se sienten seguros en entornos de aprendizaje no lineales, es decir en la definición de sus propias rutas de aprendizaje, lo que indica que navegan libremente sin seguir necesariamente lo sugerido por los creadores de contenido. Hay otros estudiantes, que prefieren seguir un camino de aprendizaje definido externamente, como lo impone un profesor o el entorno de aprendizaje en línea [19]. En el mejor de los conocimientos no se ha encontrado un análisis de comportamientos desde este último enfoque haciendo uso de bases de grafos, los cuales son una muy buena alternativa para representar la información, almacenarla y utilizarla para identificar patrones. Los grafos han demostrado ser un motor potencial para optimizar la alta conexión de datos y basados en una base de datos de grafos es posible acceder a la información de la transición entre los nodos sin considerar el tamaño del grafo o lo escalable que puede convertirse el mismo. Estos pueden ser una solución proponiendo un modelo de representación útil tanto para los docentes como para los investigadores logrando acercarse al comportamiento de los estudiantes.

Nos hemos planteado que no se ha encontrado un consenso claro acerca del comportamiento de los estudiantes desde un enfoque en donde se analicen las secuencias y rutas de navegación representadas en bases de datos de grafos y lo relacionen con diversas variables mediante las cuales se puede categorizar a un estudiante según su perfil: región, edad, sexo, estilos de aprendizaje, cultura, datos demográficos, rendimiento académico, etc. Estas variables demográficas han demostrado tener una importante influencia en el comportamiento y desempeño de un estudiante en el curso [20], de este modo podrían identificarse patrones de grupos de estudiantes clasificándolos por alguna de ellas.

Teniendo en cuenta lo mencionado anteriormente, se plantea la siguiente pregunta de investigación:

¿Cómo identificar patrones de comportamiento de estudiantes en un ambiente en línea masivo desde sus rutas de navegación usando grafos?

2. Marco Conceptual

En esta sección se presenta el marco conceptual de esta propuesta de maestría. En primera medida se aborda el concepto de dominio más significativo (MOOCs), en él, las problemáticas identificadas y el enfoque específico al que se quiere investigar. Como segunda medida se aborda el concepto de grafos, identificado como relevante en el punto de partida a la solución del problema.

2.1. Massive Online Open Courses-MOOCs

El término MOOC es acuñado por primera vez por Dave Cormier en 2008 [10] para referirse a los cursos ofrecidos en línea, abiertos y masivos. Los MOOC fueron utilizados por primera vez por George Siemens y Stephen Downes en el 2008 y desde ese entonces se han dado a conocer como una oportunidad de aprendizaje extra para los alumnos. ¿Quién puede acceder a un MOOC? Básicamente, un estudiante que tenga conexión a internet y se una a uno de los cursos ofrecidos en las plataformas MOOC. Es por eso que su nombre define sus características. Masivo (Massive), referido a la cantidad de estudiantes registrados por MOOC, que puede oscilar desde los cientos de estudiantes hasta cerca de los 150.000 estudiantes [21] Abierto (Open), definido así por la libertad con la que los estudiantes pueden registrarse a ellos independientemente de su localización, edad, nivel de aprendizaje, cultura o cualquier otro factor que los distinga entre ellos. Él termino de abierto también puede hacer referencia los recursos abiertos utilizados en cada MOOC por los profesores [22], como diapositivas, videos, apuntes, imágenes, etc.

El otro término es Online (En línea), relacionado directamente con la accesibilidad de los cursos, que se resume al hecho de tener como mínimo una conexión a internet para poder acceder al mismo y de este modo establecer una relación síncrona de interacción y asíncrona entre los participantes del curso [23]. Por último, el término Curso (Course) permite organizar los anteriores términos en un plan de estudios estructurado, que no sólo consiste en escuchar una clase pregrabada sino en el seguimiento detallado de la brecha Estudiante-Profesor para alcanzar el objetivo principal que es el aprendizaje. La plataformas MOOCs están compuestas con interfaces de

usuarios que brindan recursos y componentes organizados a los estudiantes para cada sección de clase. Estos abarcan las diferentes actividades de formación como lo son la difusión de información, discusión y evaluación. La figura 1 muestra en términos de nuevos estudiantes por plataforma, un gran número de estudiantes inscritos en cursos por año, mostrando también el fenómeno de la pandemia de 2020 donde cerca de 60 millones de personas se inscribieron al menos a un curso MOOC. Con estos datos es clara la importancia que se debe darle al diseño de las plataformas, conocer su estructura para posteriormente proponer mejoras. La siguiente sección amplía precisamente esta estructuración.

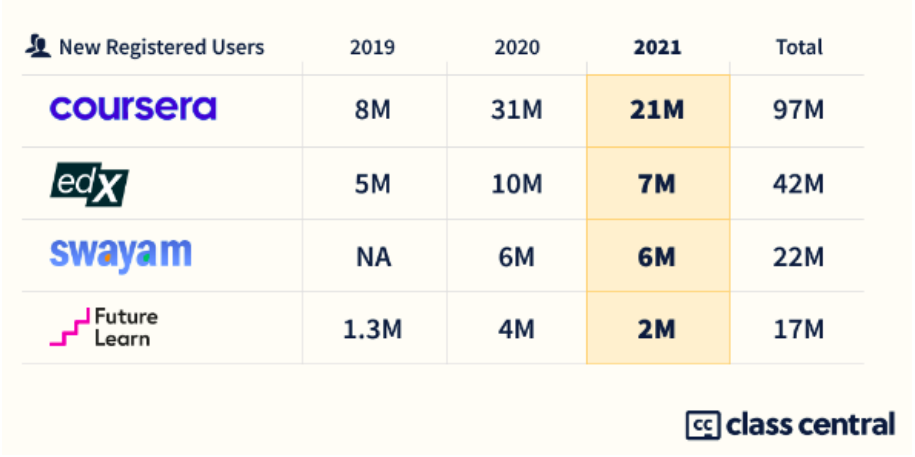


Figura 1: Tendencia en plataformas MOOC

2.2. Arquitectura MOOC

Al incrementar el uso de los MOOCs en los diferentes países, también fue incrementando el desarrollo de las plataformas por parte de universidades o compañías. Por eso fue importante estandarizar un modelo que le permita a profesores diseñar sus cursos y poderles brindar además todas las herramientas necesarias que ellos utilizarían con sus estudiantes. Los MOOCs están organizados en secciones y estas a su vez en sub-secciones. Cada sub-sección puede disponer de diferentes bloques llamados unidades para brindarle a los estudiantes recursos para el curso. Los recursos pueden estar caracterizados de tres tipos: Enseñanza, retroalimentación y evaluación. Los profesores pueden tener la libertad de combinar dichos recursos según su metodología y el objetivo del curso. La figura 2 muestra los recursos más comunes que se

pueden encontrar. En la enseñanza, el uso de de vídeos, presentaciones y lecturas. En la retroalimentación, herramientas como chats, foros de discusión, gamificación y laboratorios virtuales. y Finalmente en la evaluación recursos de exámenes y quiz.

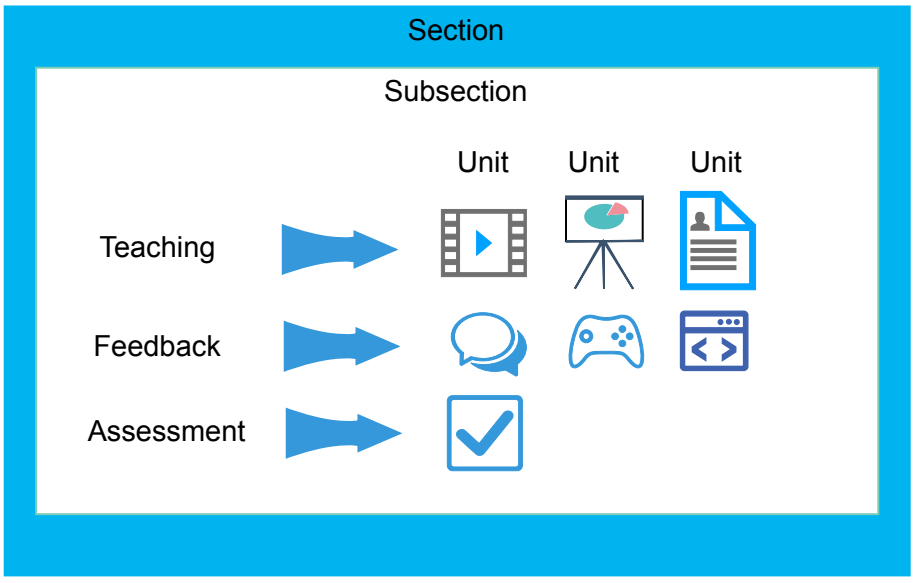


Figura 2: Recursos MOOC

El patrón de diseño en todas las plataformas es prácticamente el mismo, dividir el contenido del curso en módulos y proponer una secuencia al estudiante para revisar los contenidos e ir evaluándolos; paralelamente hay otros recursos como chats o foros de discusión que pueden estar ubicados en cada subsección o en una sección especial para ellos.

La razón por la que le es difícil a los docentes tener un acercamiento a los estudiantes a través de las plataformas es la heterogeneidad estudiantil, a continuación presentaremos este concepto en la siguiente sección.

2.2.1. Heterogeneidad Estudiantil

Un factor relevante en el auge de los MOOCs fueron los cientos de miles de estudiantes inscritos en los primeros cursos, esto contribuyó al aumento de la popularidad y la influencia de los MOOC, lo que iba a motivar cada vez más muchos estudiantes a inscribirse [24]. Cada uno de los estudiantes inscritos tiene un perfil específico caracterizado por su lugar de procedencia, cultura, condición social, nivel académico y cualquier otra variable que influya en el comportamiento y personalidad del mismo.

La heterogeneidad estudiantil precisamente es una problemática que reconoce la variedad de perfiles existentes en los cursos. Esto incentiva a un diseño de MOOCs centrados en adaptabilidad, pensando en soluciones para las plataformas a partir del análisis de personalidades [25], sentimientos [26] y en general en el perfil de los usuarios.

Daniel et al. [27] expresa que "Una solución posible, pero aún no desarrollada, que probablemente, estará disponible en un futuro próximo es la de implementar técnicas de aprendizaje adaptativo para que los cursos MOOC sean más personalizados". También, con respecto a la adaptabilidad pronuncia que "Los agentes que analizan el perfil del alumno pueden personalizar un curso de la siguiente manera: ajustar el contenido del curso de acuerdo con los requisitos previos de los participantes o la formación académica; cambiar el contenido del curso según la ubicación del participante o el país de origen, por ejemplo, el idioma, las unidades de medida, el símbolo de moneda, las estaciones, etc. y así mostrar estudios de caso relevantes o lecturas adicionales según el país o región de origen o interés".

Por los motivos mencionados anteriormente, ha sido propuesto el termino aMOOC a aquellos cursos pensados en ofrecer recursos y actividades dependiendo del perfil y las preferencias de cada participante. Sin embargo, es viable pensar en plataformas con diseños simples que no requieran de conocimientos especializados por parte de los diseñadores del curso (profesores); Esto contrasta con otras propuestas las cuales son generalmente más elaboradas pero en las cuales sólo pueden acceder a ellas expertos tecnológicos [8].

2.2.2. Learning Analytics

El análisis de aprendizaje mejor conocido en la literatura como Learning Analytics (LA), es una disciplina que busca poder inferir correlaciones y comportamientos a partir de datos recolectados durante una actividad de aprendizaje específica. Según Dietz, es una disciplina emergente relacionada con el desarrollo de métodos para explorar series de datos procedentes de ecosistemas educativos. Y con el uso posterior de los resultados del análisis para entender mejor al alumnado, sus comportamientos y así mejorar el diseño de los entornos en los que aprenden [28]. En un proceso en

le cual se aplique LA, es importante tener en cuenta que dentro de los procesos de aprendizaje se deben identificar primero, cuáles son las variables que se van a medir. Segundo, con qué mecanismo serán medidas y recolectadas identificando criterios e indicadores mensurables como la selección de herramientas y procedimientos de análisis y finalmente, es importante tener una buena interpretación de los datos con el fin de que pueda contribuir a la solución del problema planteado.

En el contexto de este trabajo LA ha sido aplicado en varios frameworks que permiten la inclusión de técnicas adaptativas en MOOCs. Sonwalkar [8] propone, utilizando servicios web y una arquitectura de computadora, un sistema de aprendizaje adaptativo (ALS) que se adapta a los cinco estilos de aprendizaje a través de las evaluaciones de diagnóstico sobre las preferencias y objetivos de los participantes. Onah y Sinclair [29] usan un sistema de recomendación, el framework ayuda a los usuarios a crear sus propios rutas, permitiéndoles tomar decisiones informadas sobre los recursos apropiados basados en sus objetivos y preferencias actuales. Este framework se enfoca en capturar el conocimiento de los usuarios utilizando pruebas basadas en conceptos. Teixeira [30] agrega a su modelo pedagógico MOOC, denominado iMOOC, la adaptación del contenido teniendo en cuenta el conocimiento previo de los participantes y el dispositivo que utilizan para acceder al curso.

2.3. Grafos

2.3.1. Conceptos preliminares

Un grafo G(V,E), según [31], esta compuesto de dos conjuntos: Un numero finito V de elementos llamados vértices o nodos y un número finito E de elementos llamados enlaces o aristas (Edges). Tanto los nodos como las relaciones tienen la capacidad de almacenar cualquier tipo de propiedad. Un grafo es de orden n, si su conjunto de nodos tiene n elementos. Un grafo sin enlaces es llamado un grafo vacío y un grafo sin nodos (y por consiguiente sin enlaces) es llamado un grafo nulo. Existen diferentes formas de representar un grafo, además de la geométrica y muchos métodos para almacenarlos en una computadora. La estructura de datos usada depende de las características del grafo y el algoritmo usado para manipularlo.

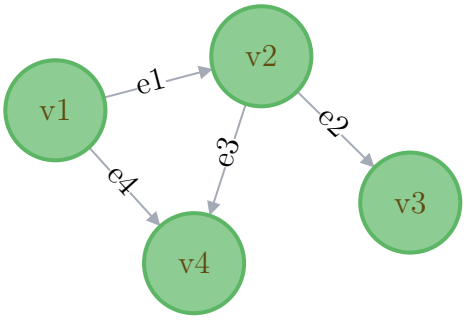


Figura 3: Representación de un grafo

La representación de un grafo geométricamente consiste en caracterizar nodos con círculos y enlaces con lineas que unen dichos círculos. Por ejemplo, si:

$$V=(v1,v2,v3,v4) \text{ y } E=(e1,e2,e3,e4,)$$

tal que

$$e1=(v1,v2),\ e2=(v2,v3),\ e3=(v2,v4),\ e4=(v1,v4)$$

Luego, la representación gráfica del grafo sería la presentada en la Figura 3. Un grafo puede también formar caminos, un camino de un grafo es un conjunto de vértices interconectados por aristas. Dos vértices están conectados si hay un camino entre ellos. Estos caminos pueden o no tener una dirección específica indicando algún tipo de característica en el sistema.

2.3.2. Bases de Datos de Grafos

Las bases de datos de grafos (GDB) hoy día se muestra como una posible alternativa para los Sistemas de Gestión de Bases de Datos Relacionales (RDBMS), generando grafos con propiedades en sus nodos y enlaces que representan toda la información. Las bases de datos relacionales son representadas por tablas con filas y columnas; una fila se puede percibir como un objeto mientras que las columnas como atributos o propiedades de los objetos [32]. Sin embargo una de las debilidades de este modelo es la capacidad de capturar explícitamente requerimientos semánticos [33]. De esta forma, reconociendo que los grafos se caracterizan por su flexibilidad y esca-

labilidad,los desarrolladores reconocen las bases de datos de grafos como un motor potencial para optimizar la alta conexión de datos.

La tecnología GDB es una herramienta efectiva para el modelo de datos y hace parte de las bases de datos NoSQL creadas para direccionar la limitaciones encontradas en las bases de datos relacionales. A continuación la principal diferencia entre el modelo GDB y los sistemas RDBMS, lo que puede justificar lo apropiado que es escoger la tecnología GDB como parte de la solución de esta propuesta maestral.

Por parte de las GDB, cada nodo contiene de manera directa y física, una lista de registros de relaciones que representan las relaciones con otros nodos. Estos registros de relaciones están organizados por tipo y dirección y pueden contener atributos adicionales. Siempre al ejecutar una operación de "MATCH", la base de datos de grafos usa esta lista, accediendo directamente a los nodos conectados y eliminando la necesidad de costosos cálculos de búsqueda y coincidencia. En cambio por parte de los sistemas de bases de datos relacionales, cuando se producen relaciones de muchos a muchos en el modelo, debe introducir una tabla ÚNICA (o tabla de entidades asociativas) que contenga las llaves externas de ambas tablas participantes, lo que aumentará aún más los costos de operación de la combinación.

La tabla 1 presenta una síntesis de comparación entre los modelos anteriormente mencionados.

Tabla 1: GDB vs RDBMS

Base de datos Relacionales	Base de datos de Grafos
1. Complejidad para modelar y almacenar relaciones.	1. Fácil para modelar y almacenar relaciones.
2. El rendimiento se degrada con el aumento de los datos.	2. El rendimiento de la relación se mantiene constante con el crecimiento del tamaño de los datos.
3. Las consultas se hacen largas y complejas.	3. Las consultas son más cortas y legibles.
4. Difícil mantenimiento.	4. Agregar propiedades y relaciones adicionales es un proceso rápido y sin migraciones.

Este diseño permite la construcción de modelos predictivos, y detección de correlaciones y patrones [34]. Esto dinamiza altamente el modelo de datos, en el cual todos

los nodos están conectados por relaciones permitiendo un recorrido rápido a lo largo de la relaciones entre nodos. Un particular beneficio es el hecho de recorrer, localizar y no tomar en cuenta conjunto de datos no relacionados. Problema que es inherente en SQL.

2.3.3. Patrones en grafos a partir de Clickstream

La recolección de la información en datasets de la interacción de los usuarios con una aplicación o sitio web, es una práctica valiosa que las compañías o individuos han comenzado a realizar. Esto con el fin de poder comprender comportamientos o intensiones de las personas y a partir de esa información generar ideas que ayuden a mejorar su experiencia de usuario. Precisamente la información recolectada es conocida como clickstream. Es la información grabada de la secuencia de un usuario durante la interacción con una página web o aplicación [35]. Para analizar el clickstream, los investigadores han propuesto diferentes técnicas de visualización[36][37][38], entre ellas se encuentran técnicas de clustering, análisis vectorial, análisis a partir de matrices, entre otras. Sin embargo, una manera potencial de visualizar la información es a través de la generación de grafos.

Ahora, muchos autores han propuesto la consolidación de eventos usando grafos. Ham [39] propuso que un diagrama de transisción de estados de un usuario en una página web puede ser representado como un grafo general de nodos estáticos que muestran directamente todas las rutas entre páginas. Pero este enfoque no muestra información de la transición del paso entre páginas y no escala a grafos de complejidad mayor. Para algunos casos la escalabilidad es un motivo importante de descubrimiento de patrones o repeticiones en los grafos (Network motifs), de esta manera comenzaron a proponerse trabajos [40][41] que identifican patrones a partir de una red de grafos escalables, en donde es posible acceder a la información de la transición entre los nodos sin considerar el tamaño del grafo o nivel de escalabilidad a la que ha llegado. Por ejemplo, el diagrama de Sankey, entre otras propuestas, es una tipo de grafo ordenado para mostrar las transiciones paso a paso. Es un tipo específico de diagrama de flujo, en el que la anchura de las flechas se muestra proporcional a la cantidad de flujo[19]. A partir de esto, una serie de variantes de Sankey han sido desarrolladas [42] con el fin de lograr algoritmos de reconocimiento

de patrones.

3. Estado del arte

Esta sección presenta una revisión de trabajos relacionados al de esta propuesta, para ellos se ha dividido esta sección en 4 partes. Una primera parte del proceso de búsqueda de documentos utilizado y las métricas seleccionadas. Segundo, se presentan los trabajos propuestos para el análisis de comportamientos de estudiantes MOOC. Tercero, trabajos enfocados en el descubrimiento de patrones de secuencia y finalmente una recopilación de los trabajos mas significativos con las respectivas brechas existentes.

3.1. Selección de trabajos

Es importante una buena definición de palabras claves para la búsqueda de documentos. Para ello se ha decidido realizar la búsqueda de publicaciones científicas tanto en la base de datos de Scopus [43] como de la de WoS (Web of Science)[44].A continuación se muestra el criterio de búsqueda utilizado para las bases de datos:

TITLE-ABS-KEY ("MOOC" OR "Massive Open Online Course*" OR "Massive Online Open Course*").*

Debido a que el término MOOC es relativamente nuevo, es muy probable que no se encuentre un gran numero de publicaciones como en otras temáticas más conocidas, sin embargo, se quiere agotar cualquier término con el que se puede también encontrar en las publicaciones. El asterisco mostrado en la búsqueda hace referencia a buscar no solamente el término MOOC sino también aquellos que van seguidos de otros caracteres, por ejemplo: MOOCS, MOOC-Education, MOOC-Learning, etc. Se ha utilizado el conector "OR" para también buscar aquellas publicaciones en donde sus autores no han utilizado la sigla MOOC sino las palabras que conforman el término y además se ha considerado algún cambio en el orden de las palabras que puede variar según algunos países. Estas consideraciones ayudarán a tener un campo más amplio de datos. Estos datos fueron extraídos el el 15 de Julio de 2018

y la Tabla 2 muestra el número total de documentos encontrados en ambas bases de datos según el tipo de documento.

Tabla 2: Documentos encontrados

Source	Conference Paper	Article	Review	Article in Press	Total
WoS	0	1029	29	0	1058
Scopus	1977	1370	83	54	3484

Cabe la posibilidad que en el total de documentos mostrado en la subsección anterior, hayan duplicados en las bases de datos, o que se encuentre el mismo documento en ambas bases de datos. Para esto, Se ha utilizado la herramienta Scientopy para aplicar una técnica de Pre-procesamiento de datos que elimina todos los documentos duplicados. La tabla 3 muestra que en total, fueron encontradas 704 muestras duplicadas. Otra de las funciones de esta técnica es la simplificación de nombres de autores. Un problema común que se ha percibido en las publicaciones es la inconsistencia de los nombres y apellidos de los autores. A través del preProces, ScientoPy es capaz de simplificar caracteres especiales en los nombres, acentos y abreviaturas utilizadas.

Las abreviaturas o simplificación de nombres también se percibe en los demás tópicos como países, keywords, instituciones, por lo que se recomienda definir a través de un solo término cada país para hacer un buen procesamiento posteriormente. Una vez obtenidos los documentos de las respectivas bases de datos, se proceden a estudiar, cuáles de ellos enfrentan la problemática expuesta en esta propuesta y se seleccionan aquellos trabajos relacionados. La siguiente sección muestra dichos documentos.

Tabla 3: Documentos finales.

Source	Conference Paper	Article	Review	Article in Press	Total
WoS	0	1027	29	0	1056
Scopus	1947	729	53	53	2782

3.2. Análisis de comportamiento en MOOC

El estudio del comportamiento de los estudiantes frente a las plataformas, es uno de los tópicos más explorados en los últimos años entre las otras problemáticas. Suhang Jiang[45] propone analizar los comportamiento de los estudiantes a partir de la combinación entre la primera semana de clase en el MOOC y la interacción social de sus recursos, de esta manera propone predecir el rendimiento del alumno al finalizar el curso a partir de sus conocimientos, aplicándolo posteriormente a enfrentar el problema de la deserción en la primera semana. En la actualidad, hay trabajos significativos que buscan predecir el desempeño académico de un estudiante a través de modelos análisis factorial exploratorio, regresiones lineales, análisis de conglomerados y correlación, como el trabajo de Bravo, [46] que mediante variables de acceso, cuestionarios, tareas y datos de los estudiantes como edad, entran a analizar si contribuyen positiva o negativamente en la predicción. La correlación todas las variables que se pueden extraer de la interacción con un curso han aportado al mejoramiento del diseño de los cursos y las plataformas mismas.

Por ejemplo, Balakrishnan realizó una investigación implicada directamente con el comportamiento de un alumno para poder medir la retención del mismo[47]. En dicho trabajo, el investigador se valió de variables como el tiempo que gastaba viendo vídeos, el número de publicaciones que veía en los foros y el tiempo en la interacción con el progreso de las unidades. De este modo, él con los resultados, a partir de ese comportamiento plantea identificar un valor que indique la retención de dicho estudiante. Ahora, un trabajo relacionado en cuanto a la metodología con la cual se quiere obtener los objetivos es el de Kizilcec [4], en donde propone disminuir el "dis-engagement" a través del análisis de los alumnos. Es decir a partir de variables de interacción como los vídeos revisados, la participación en los foros y en la evaluaciones el investigador relaciona los datos con el compromiso de dicho estudiante. La aplicación de este tipo de análisis pueden llevar en un futuro a poder construir herramientas como la de Andreas [48], que propuso un framework que permite a través de un análisis de datos predecir los estudiantes con una alta probabilidad de abandonar el curso, lo que puede ser un alerta para los creadores MOOC de buscar la forma de atraer más a dichos perfiles. Así como a la deserción, la identificación de los perfiles de los estudiantes pueden aportar a la solución de otras problemáti-

cas identificadas en la educación en línea como los comportamientos deshonestos y detección de fraude.

La tabla 4, muestra además de los ya mencionados, los artículos más importantes que buscan el comportamiento de estudiantes desde diferentes métodos. En ella se presentan en orden de columnas: la referencia del trabajo, el título, el enfoque seleccionado para el análisis de los estudiante, las plataformas de donde han obtenido los datos y el principal objetivo que quieren cumplir.

Tabla 4: Artículos relevantes

Ref.	Título	**Enfoque**	**Plataforma**	**Objetivo**
[5]	Engaging with Massive Online Courses	Tareas, lecturas y foros	Coursera	Desarrollar una taxonomía del comportamiento individual de un estudiante.
[49]	Studying learning in the worldwide classroom research into edX's first MOOC	Evaluaciones, foros, encuestas y navegación	EDX	Encontrar modelos predictivos para el éxito del estudiante
[4]	Deconstructing Disengagement: Analyzing Learner Subpopulations in Massive Open Online Courses	Tiempo de actividad y participación	Coursera	Proponer un método de clasificación usando clustering para encontrar subpoblaciones de estudiantes según su compromiso
[18]	Performance in e-learning: online participation and student grades	Foros	E-learning (Virtual Classroom)	Demostrar una relación directamente proporcional entre participación y rendimiento académico
[1] [2] [50]	Improved Learning in a Large-Enrollment Physics Class. Does the discussion help? The impact of a formally assessed online discussion on final student results. Factors that influence participation in online learning	Foros	No indica	Descubrir patrones de comportamiento a partir de las discusiones y aportes en los foros de discusión.
[51]	Relationship Between Learning Time in an Online Course and Learning Behavior and Outcomes	Tiempo de actividad vs. Secuencias	No indica	Descubrir patrones de comportamiento según el tiempo que dedicaron en las plataformas frente a los contenidos a través de la clasificación en grupos

Como brechas principales de esta propuesta maestral con los trabajos en esta sección exploradas se encuentras las siguientes:

- El objetivo general de estos trabajos es relacionar los resultados con una variable específica, ya sea el rendimiento académico, la deserción o el acople del estudiante, no precisamente como se busca en esta propuesta abarcar variables de interés no solo académicas sino además demográfica, culturales, psicológicas, y cualquier otra que involucre el perfil del estudiante como tal.
- Se considera que los trabajos no utilizan un conjunto de datos que determinen secuencias de navegación de los estudiantes, por el contrario, utilizan o imple-

mentan sus algoritmos con base a los recursos utilizados en los MOOC como los foros, tareas, lecturas o tiempo de permanencia activa en un vídeo.

Como se puede evidenciar, los trabajos que han optado por un descubrimiento de patrones de comportamiento difieren al presentado en esta propuesta por el método utilizado.

Por otro lado una alternativa para representar la información de interacción de los estudiantes MOOC almacenada en dataSets es la representación mediante grafos, poco explorada en la literatura. La generación de grafos como autómatas finitos de la secuencia de los estudiantes no ha sido punto de partida para investigaciones que se enfoquen en el descubrimiento de patrones de comportamiento para estudiantes MOOC.

Es decir no se encuentran la implementación de algoritmos basados en grafos para obtener dichos patrones y además se basen en la secuencia de navegación del estudiante. Teniendo en cuenta esto, se comenzaron a explorar aquellos trabajos que sí involucren un conjunto de datos que representen la navegación del estudiante o secuencia utilizada durante sus secciones con el fin de analizar sentimientos, comportamientos y acople con el curso. Es decir, trabajo más similares al que estamos proponiendo. La siguiente sección se dedica precisamente estudiar dichos trabajos.

3.3. Descubrimiento de patrones de secuencia

Analizar a un usuario desde las rutas optadas en la ejecución de una serie de tareas, es uno de los mayores aportes al descubrimiento de patrones de comportamiento en la minería de datos [52], sin embargo es considerado todo un reto debido al gran volumen de datos a procesar. Este análisis de secuencias ha contribuido también en entornos de educación [53] [54] analizando el comportamientos de alumnos durante las sesiones de clase. Específicamente en MOOCs, desde sus inicios, analizar las “Learning Paths"(Rutas o secuencias por la que los estudiantes se mueven), se destacó como un importante foco para los investigadores en minería de datos [55][56] para proponer mejoras a los diseñadores de cursos. La solución que se quiere proponer en esta propuesta es descubrir patrones de comportamiento de los estudiantes a partir

de las secuencias generadas con la interacción de las plataformas MOOC. Para ello se han encontrado los trabajos dispuestos en la tabla 5, la cual muestran propuestas a partir de la navegación o secuencia de un estudiante para la obtención de patrones de comportamiento. Dicha tabla está organizada de la siguiente manera: Referencia, año de la publicación, Título del trabajo, y método implementado. En este contexto se ha partido de la afirmación que además de los estudiantes que siguen la trayectoria propuesta por el docente, también existen estudiantes que se sienten seguros en entornos de aprendizaje no lineales, lo que indica que navegan libremente sin seguir necesariamente lo sugerido por los creadores de contenido, definiendo externamente un camino de aprendizaje alternativo. Esto puede comprobarse en algunos trabajos [57] [14], en el cual se realizan seguimientos a las rutas a partir de los datos almacenados en las bases de datos de una plataforma.

En contraste a los trabajos anteriormente mencionados, este artículo abarca varios tipos de actividades y no se enfoca en un solo tipo. Por ejemplo en actividades relacionadas con evaluación, Köck y Paramythis [54] presentan un agrupamiento de secuencias para modelar el comportamiento de los estudiantes mientras resuelven un problema. Algo similar plantea Shanabrook [56] analizando mientras los estudiantes resuelven un problema, secuencias de acciones (por ejemplo, jugar con el sistema, adivinar por frustración, abusar de pistas) en un sistema de tutoría inteligente que emplea el descubrimiento de motivos basado en secuencias. Entre los pocos trabajos encontrados que abarcan varios tipos de actividad, hay algunos que se centran en identificar los recursos más visitados por los estudiantes y el impacto de variar el orden de los contenidos o alterar las rutas propuestas por los docentes [58]. Wen y Rosé presentan un estudio [59] donde pueden identificar el salto más común que hace un estudiante entre dos contenidos. Mercado [60], presenta un análisis de visualización de secuencias de la plataforma Coursera, que permite establecer la importancia de la presencia de expertos en los vídeos como el recurso más concurrido y la participación de los foros como un recurso frecuentemente usado independiente de la ruta que tome el alumno. Otro grupo de trabajos aprovechan las rutas y secuencias de los estudiantes para proponer sistemas de recomendaciones para la definición de su propia ruta antes de comenzar un curso según el perfil del alumno y además evitar la futura deserción [61]. Ardchir [62], ha propuesto un marco de trabajo que brinda una ayuda al estudiante a definir la ruta de aprendizaje, para fortalecer el

compromiso con el curso y disminuir la tasa de deserción académica muy frecuente en los MOOC.

El trabajo de Davis [63] es de gran referencia para el enfoque seleccionado en este artículo, porque se ha considerado la representación de las secuencia en forma de grafos. Como conclusión de su trabajo muestra hasta qué punto los alumnos (como un grupo completo y divididos en alumnos que pasan y no pasan) siguen el camino prescrito por el docente. El presente trabajo quiere aprovechar este mismo enfoque pero pensado en la visualización e interacción de los datos a través de una novedosa interfaz de visualizaciones donde los usuarios sean los docentes. En la siguiente sección se presentan trabajos relacionados con respecto a las herramientas de visualización.

Este enfoque de análisis de datos a través del clickstream puede haber sido utilizado en otros trabajos de las ciencias de la computación que no pertenecen a la temática MOOC. Estos trabajos fueron mencionados alguno en el marco conceptual en donde para diferentes aplicaciones y entornos, se busca descubrir patrones de grafos a partir del clickstream de usuario almacenado. Los algoritmos que en estos trabajos han sido implementados, también se consideran guía para seleccionar el de nuestro trabajo, es por eso que vale la pena no pasar por alto la revisión de dichos trabajos.

3.4. Herramientas de Visualización

Para poder validar el modelo de representación que propone este trabajo, se necesita una interfaz gráfica de interacción para el usuario, una herramienta de visualización de datos para cursos de la plataforma, la cual pueda ofrecer información relevante del comportamiento de un estudiante. En la literatura han sido encontradas herramientas que presentan el análisis con diferentes tipos de indicadores en entornos de aprendizaje. Lo interesante es poder proponer visualización de datos que las plataformas no ofrecen aún. Por ejemplo ALAS-KA [70] es creada como un apoyo a la plataforma Khan Academy, ampliando su funcionalidad a más de 20 nuevas visualizaciones de datos relevantes sobre el rendimiento y la interacción del estudiante con la plataforma. Cada herramienta de visualización para entornos virtuales de aprendizaje en línea trata de identificar el perfil tanto del estudiante como del maestro

Tabla 5: Artículos relevantes

Ref.	Año	Título	Enfoque
[64]	2014	Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions	Ayuda psicológica cognitiva
[9]	2016	Exploring Differences in How Learners Navigate in MOOCs Based on Self-Regulated Learning and Learning Styles	Análisis de perfiles SRL y estilos de aprendizaje
[65]	2017	Discovery of Navigation Patterns in Open Edx - An Architectural Approach	Matrices en espacio vectorial
[66]	2018	Predicting Learners' Success in a Self-paced MOOC Through Sequence Patterns of Self-regulated Learning	Estudio de variables de ritmo propio.
[67]	2018	Mining Theory-Based Patterns from Big Data: Identifying Self-Regulated Learning Strategies in Massive Open Online Courses	Teorías basada en aprendizaje autoregulado y estrategias de clustering.
[68]	2022	Interrelated analysis of interaction, sequential patterns and academic achievement in online learning	Análisis por agrupación de estudiantes: orientados a evaluaciones, certificación y orientados a premios,frente a la interacción de fragmentos de semanas de contenido.
[69]	2022	Do individual characteristics affect online learning behaviors? An analysis of learners sequential patterns	Influencia del comportamiento según cracterísticas cognitivas de los estudiantes

a través de diferentes tópicos comparándolo con los datos de la muestra estudiada. Es por eso que estas herramientas tienen un enfoque estadístico, por ejemplo el identificar los eventos más comunes de los estudiantes en un curso, como lo analiza GLASS [71], y visualizar estos resultados para el interés de los maestros o los mismos estudiantes.

Hay una herramienta interesante [72] que se centra en analizar a través de redes de alto nivel las visitas en los vídeos de los cursos, y a través de análisis de redes (cuyos nodos son cada vídeo) comenzar a identificar cuáles fueron los vídeos más favorables, o más tediosos, de esta forma la herramienta le permite al usuario predecir razones por lo cual los estudiantes abandonan el curso, la herramienta permite relacionar el comportamiento y el rendimiento del estudiante todo a partir de los flujos de clics realizados en los vídeos. Este es un importante aporte debido a que el recurso de vídeo es el contenido más utilizado en las plataformas pero es necesario un buen flujo de datos que puedan revelar eventos realizados en los vídeos.

Así como hay herramientas centradas en vídeos también hay centradas en foros de discusión, por ejemplo se han desarrollado herramientas [73] [74] pensando en la dificultad que tiene la participación de los foros en ser evaluados al momento de darle un valor en su calificación. El primer trabajo es una herramienta pensada para los docentes y realiza un cálculo automático de la calificación de los foros considerando la dimensión cuantitativa y la relevancia de sus contribuciones en el mismo. En la herramienta el docente puede visualizar la participación de los estudiantes en los foros juntos con una calificación calculada por el algoritmo. La segunda herramienta detecta automáticamente el compromiso emocional y cognitivo a partir de la clasificación de textos que los estudiantes escribieron en los foros.

Entre las herramientas de visualización desarrollada para estudiantes, está la propuesta por Tobías [75]. Debido a que el aprendizaje auto regulado(SRL) es un reto para los estudiantes se han desarrollado herramientas como esta que buscan incluir un dashboard de aprendizaje en las plataformas para que los estudiantes visualicen el progreso, su comportamiento, crear conciencia y fomentar la autoreflexión. De esta manera los estudiantes se sienten motivados para lograr y alcanzar sus objetivos de aprendizaje.

Según lo expuesto, es importante que la herramienta desarrollada para la visualización del comportamiento de los estudiantes pueda ser utilizada tanto como por docentes como por estudiantes, poruqe por un lado los docentes pueden analizar y evaluar el rendimiento de los estudiantes y el diseño del curso y por otro lado los estudiantes pueden revisar su progreso, su compromiso con el curso y auto evaluarse.

Para Open edX existe ANALYSE [76], una herramienta basada en las características propias de un MOOC, la retroalimentación de los maestros y fundamentos pedagógicos, pero que no considera como característica relevante, rutas de aprendizaje a través del curso. El enfoque de la herramienta que aquí se propone está relacionado directamente con la organización y orden de los recursos educativos, el plan de estudio o la agenda del curso. Según Dipace [77], este tipo de herramientas deben comparar el tiempo gastado con el rendimiento académico logrado, para motivar a los estudiantes a invertir mejor sus tiempos activos de sesiones y seleccionar mejor las rutas de aprendizaje. Por lo tanto, a pesar de que hay diversas interfaces de visualización en la literatura, la mayoría de ellas omiten las secuencias de acciones de su análisis.

Según la información brindada se considera que lo más importante es identificar las brechas existentes entre estos diferentes trabajos y el aquí propuesto, para ello, se identificaron las siguientes brechas:

- Ninguno de los trabajos propone una base de datos de grafos para la representación de la información de navegación de los estudiantes frente a las plataformas MOOC.
- En los trabajos No se opta por la utilización de algoritmos basados en teoría de grafos que ayuden a descubrir patrones de comportamiento.
- Los patrones encontrados por los trabajos no se relacionan a mas de 2 variables de interés con respecto a los estudiantes.
- De los trabajos encontrados en ninguno es prioridad proponer una herramienta de visualización de las secuencias de los estudiantes partiendo desde una base de datos de grafos.

3.5. Aporte Investigativo

El aporte de este trabajo de maestría puede contribuir a facilitar el descubrimiento de patrones de comportamiento permitiendo entender mejor los procesos de aprendizaje en estos ambientes y a partir de allí entrar a realizar mejoras en el mismo. A partir

de esto, y teniendo en cuenta las brechas encontradas en relación a otros trabajos, esta propuesta plantea los siguientes aportes:

- La representación como grafos del comportamiento de un conjunto de estudiantes en un curso en línea masivo, cuyos nodos, enlaces y propiedades comprenden la información contenida en un dataset de navegación.

- Un algoritmo de reconocimiento de patrones para estudiar el comportamiento de los estudiantes en un curso en plataforma MOOC.

- Verificación del algoritmo propuesto en un conjunto de datos ya recolectados en un curso en línea en la Universidad del Cauca.

Para dirigir la pregunta de investigación presentada en en el planteamiento del problema de la sección 1.1, esta propuesta de maestría plantea la siguiente hipótesis: es posible el descubrimiento de patrones de comportamiento de estudiantes a partir de la representación como grafos de la navegación en plataforma MOOC.

4. Objetivos

4.1. Objetivo General

Construir una aproximación de un mecanismo para el descubrimiento de patrones de comportamiento de estudiantes en plataforma MOOC.

4.2. Objetivos Específicos

1. Proponer un esquema de representación basado en grafos de las actividades de aprendizaje de los estudiantes en Open edX.

2. Evaluar, seleccionar y adaptar un algoritmo que permita el descubrimiento de patrones de similitud de grafos según el comportamiento de los estudiantes.

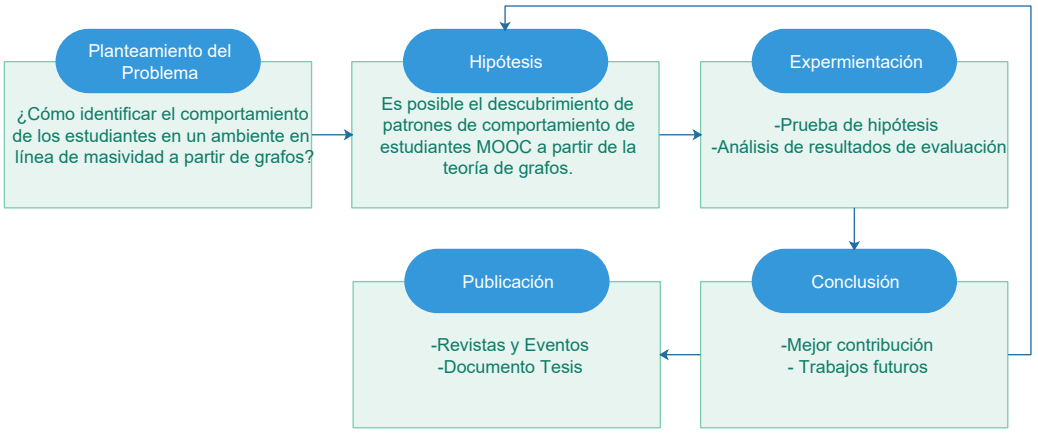


Figura 4: Metodología

3. Verificar el algoritmo propuesto aplicado al descubrimiento de patrones en un conjunto de datos en un curso en línea con reconocimiento académico en la Universidad del Cauca.

4. Evaluar el resultado de los patrones de comportamiento encontrados respecto a la experiencia de profesores y tutores de cursos en plataforma MOOC.

4.3. Metodología

Para la elaboración del cronograma de actividades y con el fin de dar cumplimiento a los objetivos planteados, se usó como referencia la descomposición jerárquica WBS (Work Breakdown Structure) sugerida en la metodología PMBOK (Project Management Base of Knowledge) por el PMI (Project Managament Institute), específicamente en el área de gestión del alcance (Scope Management) [78].

La figura 4 muestra el proceso científico propuesto para alcanzar el objetivo final de esta propuesta. Dicho proceso está compuesto de las siguientes fases: Planteamiento del Problema, Construcción de Hipótesis, Experimentación, Conclusión y Publicación.

5. Representación basado en grafos de las actividades de aprendizaje de los estudiantes en Open edX

Para alcanzar el primer objetivo del trabajo se propone un modelo de representación del comportamiento el cual sea válido para una cualquier plataforma con su arquitectura convencional de MOOC. Las características principales del modelo son dos: El modelo debe permitirle a los docentes la facilidad de visualizar las secuencias de los estudiantes durante sus secciones activas de clase y dos, el modelo debe estar soportado en grafos lo cual permitirá identificar a futuro patrones de comportamiento entre los estudiantes. En la siguiente sección se comienza por los fundamentos iniciales para la construcción del modelo.

5.1. Fundamentos de modelado de interacción

Inicialmente se presenta la estructura y componentes involucrados en un curso siguiendo la estructura del Sistema de Gestión de Aprendizaje Open edX, una de las plataformas con características convencionales en cuanto a la estructura revisada en el marco teórico. En el contexto de este trabajo será tomada esta plataforma como referencia debido a que cumple con las características mas comunes entre las demás plataformas, es una de las plataformas más usadas en el mundo y además tenemos acceso a la extracción de datos por parte del servicio ofrecido en algunos cursos de la Universidad del Cauca, lo que va a permitir verificar y evaluar lo implementado. Seguidamente vamos a caracterizar la teoría de grafos para la creación del modelo de representación del comportamiento del estudiante hasta proponer unos grafos a partir de los cuales se pueda evidenciar las secuencias de los estudiantes.

5.1.1. Estructura y actividades de un curso

La interfaz de usuario de un MOOC en una plataforma está usualmente definida por una estructura organizada en módulos y tipos contenidos y servicios. El docente del MOOC en Open edX, al realizar su diseño, define la distribución de contenido a partir

de unos módulos denominados secciones, subsecciones y unidades. La sección es el módulo de mayor nivel, usualmente los docentes definen una temática diferente por cada sección. En cada sección se pueden establecer cuantas subsecciones se considere necesario (es decir desglosar las temáticas en sub-temáticas) y finalmente, en cada subsección el docente tiene la posibilidad de distribuir sus contenidos y servicios por unidades por ejemplo vídeos, documentos, foros, evaluaciones, etc. Los contenidos se pueden dividir en diferentes tipos, algunos son de tipo enseñanza que hace referencia a la tarea de divulgar el conocimiento desde el docente al estudiante. En este se encuentran los vídeos, los documentos, presentaciones, infografías o algún recurso exterior, entre otros, que desee usar el docente. Otros son de tipo retroalimentación, ahí están incluidos los foros, chats, simulaciones o juegos y por último están los de tipo evaluación, estos son componentes que se usan para evaluar a los estudiantes. Diseñar un curso en MOOC implica establecer una trayectoria definida por el docente a través de todo el material de aprendizaje. Cada cierto periodo determinado por el docente es liberada una parte del contenido la cual los estudiantes deben consumir, junto al contenido se disponen los demás recursos de retroalimentación y finaliza evaluando el contenido liberado. Este patrón de diseño es el mismo en la mayoría de las otras plataformas. El siguiente paso después de considerar estos conceptos, es diseñar un grafo dirigido que represente claramente el movimiento de un estudiante por los diferentes contenidos, para esto se ha contextualizado la teoría de grafos en este modelo.

5.1.2. Modelado de grafos

Al estudiar tanto la estructura y organización de los componentes de un MOOC como los posibles metadatos a extraer de la actividad de secuencia de clics, se comienza a construir un modelo de visualización enfocado en grafos, el cual pueda abarcar los eventos más representativos y presentar al usuario como rutas de aprendizaje combinando parámetros de módulos (secciones), sesiones y estudiantes. El modelo propuesto consiste en la transformación del conjunto de acciones realizadas en las actividades, en un grafo dirigido caracterizando la ubicación en la estructura jerárquica del curso y el tipo de actividad realizada.

Según la teoría de grafos, un grafo dirigido o dígrafo está definido por un par de

conjuntos $G = (V, E)$, en donde:

- $V \neq \emptyset$: Es un conjunto no vacío denominado vértices
- $E \subseteq \{(a, b) \in V \times V : a \neq b\}$: Es el conjunto de pares ordenados de elementos de V, denominados aristas, donde por definición una arista va del primer nodo (a) al segundo nodo (b) dentro del par.

En nuestro contexto, el conjunto de nodos está definido por los diferentes contenidos del curso, mientras que las aristas establecen trayectorias entre los contenidos. Ahora, considerando que se trata de trayectorias secuenciales, el grafo es identificado como conexo, en donde el extremo de una arista es el origen de la otra. Teniendo en cuenta las situaciones que se pueden presentar, se establece que los dígrafos que generaremos no serán completos, es decir, es posible que se presenten bucles o lazos cual es característica de un grafo simple, y además no necesariamente habrá por lo menos una arista por cada par de nodos.

5.2. Diseño e implementación del modelo de representación

De acuerdo con los fundamentos del modelado de interacción y los objetivos planteados, la figura 5 presenta el diseño del modelo de representación de rutas de estudiantes en un grafo dirigido a través de un diagrama de clases. Este modelo define la transformación de un conjunto de datos de eventos registrados en una plataforma MOOC a por lo menos 3 tipos de grafos dirigidos de rutas de navegación involucrando uno o más estudiantes. El modelo es validado desarrollando de una herramienta de visualización denominada Moockly desplegada en un servidor web desarrollada en la presente investigación.

Con el modelo y la forma de cómo validar lo definido, se presenta la implementación de Moockly, una herramienta que ha sido construida reuniendo los conceptos anteriormente presentados y que su desarrollo parte desde la extracción, organización y procesamiento hasta la visualización de los datos. La figura 6 muestra la arquitectura software para el proceso de transformación de los datos y cómo lograr visualizarlos en a interfaz de usuario, además de la interacción de los actores con los elementos.

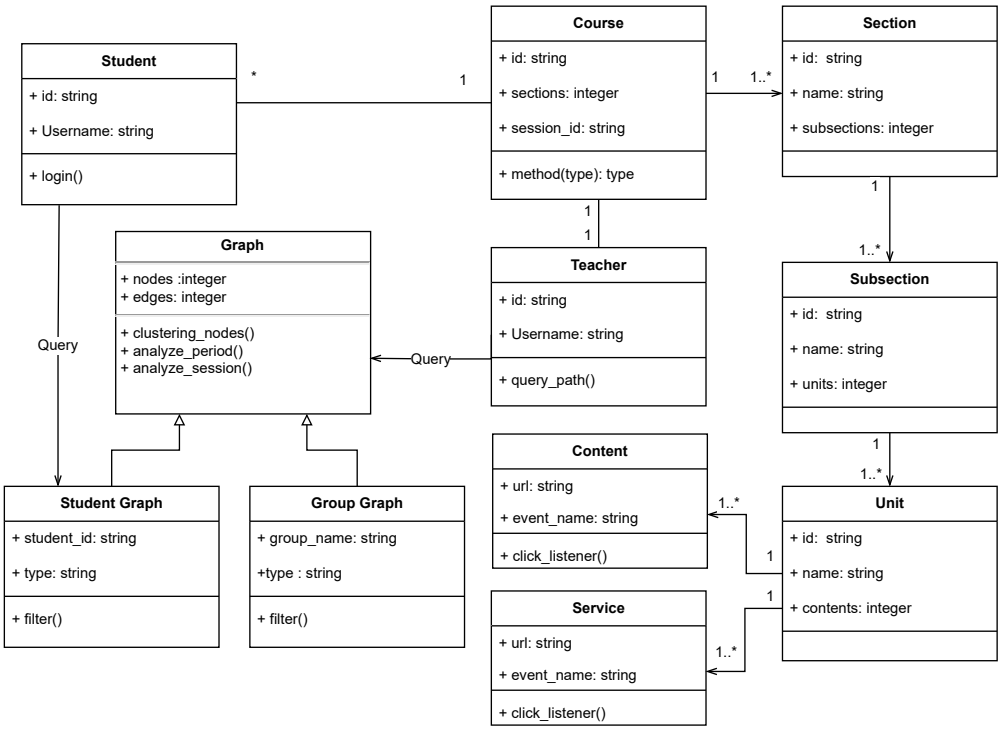


Figura 5: Diagrama de clases

La arquitectura se compone de 3 elementos, el primero, denominado Open EdX, fuente de los datos y donde se encuentra alojado el MOOC, el segundo, Modelado de procesado, elemento que acondiciona los datos y los proporciona a la herramienta y finalmente Moockly como la herramienta de visualización que se le despliega al usuario final. Aunque la herramienta de Moockly está pensada principalmente para los docentes, es posible que, como un ejercicio de autoconocimiento, los estudiantes puedan ver sus rutas de navegación por el curso por los que se han considerado a ambos como Usuarios. A continuación, revisemos más detalladamente cada elemento de esta arquitectura:

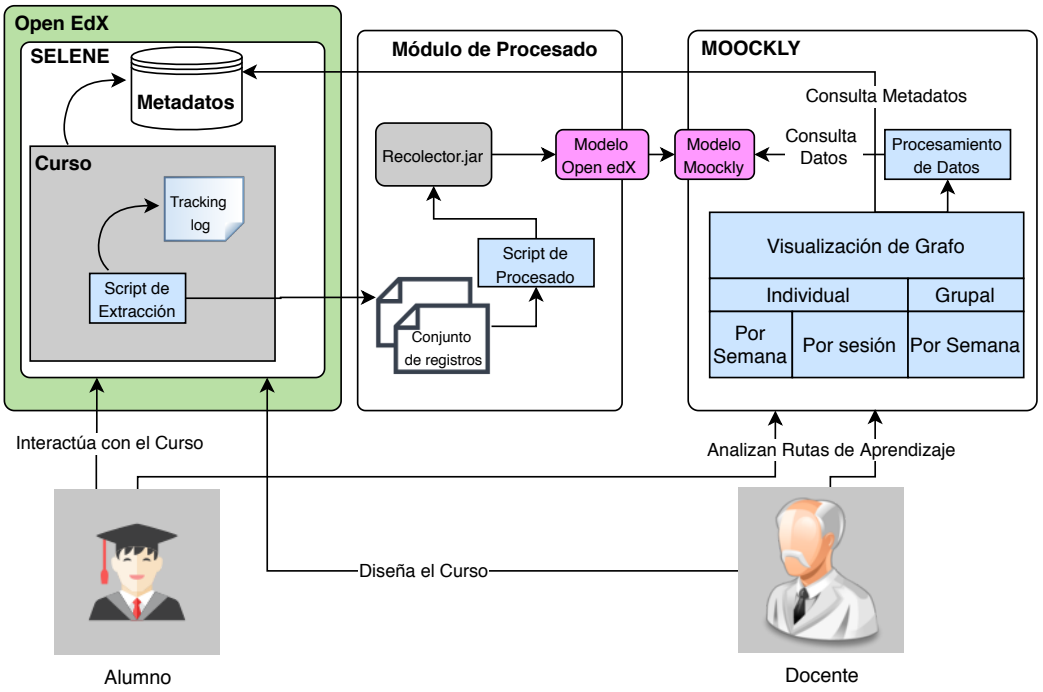


Figura 6: Arquitectura Software

5.2.1. Open EdX

Este elemento es el encargado de la recolección y almacenamiento de los datos. Este a su vez contiene un elemento llamado Selene que es una instancia de Open edX a través de la cual la Universidad del Cauca ofrece los MOOCs. Selene está instalada sobre un servidor físico propio de la universidad. Es así como se tiene acceso a todos los datos de los estudiantes en la plataforma. Por un lado se encuentran los metadatos que se quiere utilizar para la identificación de secciones y nombres de contenidos y por otro lado, está el curso donde se recolectan los datos.

El diseño del mecanismo para la recolección de datos se realizó teniendo en cuenta un trabajo previo [14] que utiliza los siguientes elementos:

- **Tracking.log:** La plataforma Selene se despliega sobre un equipo servidor con una versión de Ubuntu server 14.04 y es una instancia de Open edX. En la plataforma se genera un archivo de registro llamado tracking.log. En este archivo se registran todas las interacciones de los estudiantes con el entorno de aprendizaje Selene, tales como: inicio de sesión, ingreso en los cursos, navegación en contenidos, presentación de foros e información de la interacción con

las actividades evaluativas.

- **Script de Extracción:** Los scripts son un conjunto de instrucciones generalmente almacenadas en un archivo de texto que deben ser interpretados línea a línea en tiempo real para su ejecución. El script de extracción lo que contiene son los comandos utilizados para realizar una copia de toda la carpeta que contiene los registros tracking.log de la plataforma y una línea de sincronización de dicha carpeta en un equipo externo. De esta manera todos los registros pueden ser procesados en tiempo real. El script es ejecutado cada 5 minutos.

5.3. Modelado de Procesado

Este segundo elemento procesa los datos recolectados, para entregárselos posteriormente a Moockly, este elemento se compone del conjunto de registros obtenidos en el elemento anterior y los siguientes componentes:

- **Script de Procesamiento:** Una vez con todos los registros disponibles en el equipo externo, el script de procesamiento captura el archivo tracking.log que es el archivo que tiene la información de las interacciones en tiempo real, realiza una copia del mismo, lo ubica en una dirección específica en donde el Recolector.jar puede procesarlo y repite esta operación periódicamente cada 5 minutos a través del programa Cron presente en Ubuntu. De esta manera los datos procesados de la interacción pueden ser utilizados para realizar seguimiento a los estudiantes en todo momento. Para obtener los datos de eventos pasados, el script fue modificado a fin de que descomprimiera uno por uno los archivos de registro y ordenara la ejecución del Recolector.jar para cada registro presente en la carpeta sincronizada, así se logró crear un juego de datos con la interacción de los estudiantes desde el momento en que se creó la plataforma.

- **Recolector.jar:** Es un ejecutable creado en código Python, que al ejecutarse busca el archivo tracking.log en una ubicación específica, lo lee tomando evento por evento presente en el archivo de registro, guarda en un buffer los eventos relacionados a las actividades que se desean capturar para luego procesarlos,

obtiene la información de manera ordenada y entendible para un humano, y los guarda en una base de datos (MySQL). El archivo tracking está escrito en JSON, por lo que el recolector debe interpretar dicha información de manera adecuada. En la tabla 6 se muestran las actividades capturadas.

Tabla 6: Columnas Dataset

Tipo	**Evento**
Video	Play
	Pause
	Stop
Lectura	Abrir documento
Foro	Abrir hilo en foro
	Comentario en foro
	Respuesta en comentario de foro
Evaluación	selección de una respuesta
	Envío de respuesta

5.4. Moockly

Este otro elemento corresponde a la interfaz de visualización con el que los usuarios finales van a interactuar. La herramienta desarrollada es una aplicación web que recibe datos procedentes de Open edX generados de la forma especificada en el elemento de la arquitectura anterior.

Las diferentes plataformas MOOC ofrecen la posibilidad de acceder a metadatos recopilados de la actividad de los estudiantes, en el caso de Open edX, se obtienen los archivos de tracking log en formato JSON, se filtra y se organizan los datos de tal manera que cada registro brinde información precisa para nuestro interés. El dataset contiene la siguiente información: Nombre de usuario, código del curso, código de sesión, fecha y hora del evento, sección, subsección y unidad del contenido; por último el nombre del evento. Los datos de este modelo son procesados y visualizados según lo muestran los dos componentes a continuación:

- Procesamiento de Datos: El procesamiento consiste en la transformación del conjunto de datos de Open edX en una estructura de grafos. Las actividades

existentes son caracterizadas como nodos y los eventos registrados como enlaces. Cada evento tiene dos propiedades temporales (fecha y hora), espaciales (Sección, subsección y unidad) y el usuario para identificar al estudiante.

- Visualización de Grafos: El motor de la aplicación recibe los nodos y los enlaces correspondientes y construye los grafos que serán visualizados por el usuario. Además se han construido los respectivos filtros para obtener grafos por semana y por sesión.

Como parte del proceso de implementación de la herramienta se partió desde los dos elementos anteriormente mencionados. Para el procesamiento de datos era necesario realizar una transformación del dataset, de tal manera que no se perciban como registros unitarios recolectados sino como secuencias de eventos realizados en cada sesión por un estudiante. De esta forma se puede realizar un acoplamiento entre los datos y la estructura de un estudiante. Para esta transformación se utilizaron librerías de procesamiento de datos en Python como Pandas en donde se pueden manipular, filtrar y reorganizar los mismos según las propiedades o columnas con las que se componen. Pensando en la aplicación de la herramienta, a los docentes les podría interesar visualizar las rutas de navegación desde dos enfoques, uno en el que se analice al estudiante dentro de un módulo y se ignoren otros módulos, otro enfoque en donde se analice una sesión completa considerando otros módulos en los que se haya navegado. Como resultado del modelo propuesto y el planteamiento de la herramienta de visualización, se presenta específicamente qué tipo de grafos es posible generar. Igualmente, la distribución cuantitativa de los patrones de comportamiento y la comparación con el patrón diseñado por el docente. Los roles del usuario del modelo son 2: estudiante y docente. Los grafos generados pueden ser grupales o individuales, ambos tipos para descubrir patrones de comportamiento. Los grafos grupales son útiles para que el docente pueda visualizar y analizar una población especifica filtrando por alguna variable y los grafos individuales pueden ser aprovechados por ambos roles. Un estudiante puede analizar su secuencia individual, así como también lo puede hacer el docente. Para el grafo individual es válido visualizar las rutas de navegación desde tres enfoques, uno en el que se analice al estudiante dentro de un único período o módulo, todas las sesiones que realizó en dicho período y se ignoren otros módulos; otro donde sea el mismo enfoque mencio-

nado pero analizando sólo una sesión activa en la plataforma, y un tercer enfoque en donde se analice una sesión activa completa considerando no únicamente un período sino otros períodos externos en los que se haya navegado. En muchas ocasiones un módulo equivale a una semana de clase sin embargo en otras no, por lo que se ha determinado utilizar mejor el término período.

Gracias al modelo definido fue posible generar cuatro tipos de grafos. El primer tipo consiste en analizar al estudiante en cada período liberado definido de la siguiente forma:

Considere $W = \{w_1, w_2, ..., w_n\}$ al conjunto de n períodos de un curso, donde $w_i = \{G_1(V, E), G_2(V, E), ..., G_s(V, E)\}$, corresponde al conjunto de dígrafos para s sesiones dedicadas a cada periodo. Cada sesión del periodo correspondiente, está representada por un dígrafo compuesto por $V = \{a_1, a_2, ..., a_m\} \cup \{other\}$, el conjunto de m actividades de la semana unido con un elemento "other"que representa cualquier otra actividad que no haga parte de la semana analizada. Estos vértices entre otras propiedades, posee una propiedad llamada tipo de actividad dado por $T = \{Video, Lectura, Foro, Evaluacion\}$. Ahora bien, E es el conjunto de aristas que representa las trayectorias adyacentes entre las actividades tales que $E \subseteq \{(a_i, a_j) \in V \times V : a_i \neq a_j\}$, la condición aclara que las transiciones del grafo solo representan eventos para cambios de actividad y no un evento dentro de una misma actividad (por ejemplo pausar y reproducir el mismo vídeo) este tipo de grafo es denominado individual por periodo (IW).

El segundo tipo de grafo es similar al primero con la condición de que $w_i = \{G_u(V, E)\}$ es decir, se analiza una única sesión del período, este tipo de grafo es denominado Individual por Período por Sesión (IWS).

El tercer tipo de grafo también analiza eventos realizados en un lapso de tiempo de una sesión activa, pero con la diferencia de considerar las actividades de otros períodos. es decir, el conjunto O = {other} es transformado en $O = \{a_{1,1}, a_{1,2}, ..., a_{2,1}, a_{2,2}, ...a_{m,n}\}$ donde m es el número de actividades presentes en la semana n. De igual forma se considera el conjunto T definido en la sección anterior al igual que los las trayectorias E. Este tipo de grafo es denominado Individual por Sesión (IS).

Finalmente, un cuarto tipo de grafo a obtener con este modelo sirve para analizar un

grupo de estudiantes el cual es identificado por algún parámetro, por lo que se define como un conjunto de grafos de tipo IW donde se reutilicen los nodos y se distingan los diferentes enlaces para cada estudiante. Los estudiantes pertenecen a un mismo grupo si y sólo si tienen una propiedad en común entre ellos ya sea demográfica o académica. Este tipo de grafo implementado es denominado Grupal por Período (GW).

Para poder clasificar a los estudiantes por algún grupo se han identificados en trabajos como en el de Maldonado [79] que agrupan a los estudiantes como comprensivos, los que siguen la ruta diseñada por el docente y focalizados, los que solo buscan información necesaria para aprobar. Además dentro de cada grupo se pueden aplicar filtros por variables proporcionadas por los mismos estudiantes en las encuestas antes de iniciar el curso.

Visualizar estos tipos de grafos obtenidos como resultado del modelo es efectivamente la forma cómo se valida la herramienta Moockly. Presentaremos en el caso de uso del capítulo 6 algunas de las secuencias de estudiantes de un curso real en un caso de estudio representadas en grafos aplicando el modelo definido y generando los cuatro tipos de grafos diferentes.

A partir del modelo implementado podemos ahora si comenzar a proponer análisis estadísticos que me permitan relacionar los grafos de cada tipo y poder encontrar correlaciones entre ellos. En el siguiente capítulo se expone la selección de un algoritmo con el que se logre identificar dicha correlación.

6. Similitud de grafos según el comportamiento de los estudiantes

Teniendo un modelo de representación en grafos de los caminos realizados por los estudiantes en la interacción con sus cursos en las plataformas MOOC, puede aprovecharse la estructura los datos para proponer análisis estadísticos que permitan identificar correlaciones entre los datos y la posible identificación de patrones de comportamientos. Los análisis estadísticos se proponen a partir de grafos dirigidos,

más específicamente paseos, en donde los caminos secuenciales pueden visitar un nodo una o varias veces.

Los análisis que en este capítulo se proponen pueden aplicarse a cualquiera de los tipos de grafos generados y los resultados serán interpretados de manera independiente

6.1. Centralidad

La centralidad en un grafo se refiere a un valor medido que tiene un nodo en un grafo. La posibilidad de comparar este valor a una escala acorde con el tamaño del grafo puede determinarse la importancia que tiene dicho nodo dentro del grafo. Conocer la centralidad permite determinar el impacto que este causa dentro del conjunto del que pertenece. Este concepto en el contexto educativo de las rutas de navegación de los estudiantes, puede ser utilizado para entender diversas relaciones que hay entre las decisiones que los estudiantes toman. Según el tipo de grafo del modelo de nuestro de trabajo la centralidad puede dar a conocer la relevancia de un tipo de contenido dentro de un período de estudio.

La centralidad no es un atributo intrínseco de los nodos en el grafo, sino un atributo estructural, es decir, un valor asignado que depende estrictamente de su localización en la red. La centralidad mide según un cierto criterio la contribución de un nodo según su ubicación en la red, independientemente de si se esté evaluando su importancia, influencia o relevancia.

En un grafo como los generados en el modelo, el grado se subdivide en grados de entrada y salida. El grado de entrada hace referencia a los enlaces incidentes en el nodo, es decir a la actividad que vaya a realizar el estudiante, mientras que el grado de salida es el número de enlaces dirigidos a otros nodos desde un nodo en particular. Cuanto mayor sea el grado de un nodo, más crucial se vuelve en el grafo.

Las medidas de centralidad pueden separarse en dos grupos: medidas radiales y mediales [80] Las radiales toman como punto de referencia un nodo, que inicia o termina un recorrido por el grafo, mientras que las medidas mediales toman como referencia los recorridos que pasan a través de un nodo dado. Las medidas radiales

a su vez se pueden clasificar en medidas de volumen y de longitud, según el tipo de recorridos que consideran. Las primeras miden el volumen (o el número) de recorridos limitados a dicha longitud prefijada, en tanto que las segundas miden la longitud de los recorridos necesarios para alcanzar un volumen prefijado. en este contexto asumiremos en el análisis estadístico 4 medidas de centralidad: Centralidad de grado, cercanía, intermediación y vector propio. La tabla 7 muestra la categpria de cada medida de centralidad y la siguiente sección presenta la obtención de estas medidas en los grafos que representan la navegación de los estudiantes.

Tabla 7: Medidas de centralidad

Nombre de medida	**Nombre en inglés**	**Categoría de la medida**		
		Radial		**Medi**
		De volumen	**De longitud**	
Centralidad de grado	Degree centrality	Si	No	No
Centralidad de cercanía	Clossness centrality	No	Si	No
Centralidad de intermediación	Betweenness centrality	No	No	Sí
Centralidad de vector propio	Eigenvector centrality	Si	No	No

Sea $G(V, E)$ un grafo representado en la figura 7 donde V es su conjunto de nodos y E es su conjunto de enlaces. Este grafo es un grafo IW generado con el modelo de esta propuesta del movimiento de un estudiante en un período. Tomemos este grafo de referencia para usar la teoría de grafos y poder identificar los valores estadísticos propuestos en esta sección.

6.1.1. Matriz de adyacencia y probabilidad

Cada grafo dirigido se puede representar con su respectiva matriz de adyacencia dada de tamaño N*N siendo N el número de vértices del grafo. Los elementos de dicha matriz son 0 ó 1 de manera que el elemento de la matriz de adyacencia será 1 si existe un enlace del nodo j a la página i. De esta forma la matriz de adyacencia A de $G(V, E)$ será:

$$V_1 \quad V_2 \quad V_3 \quad F \quad Q$$

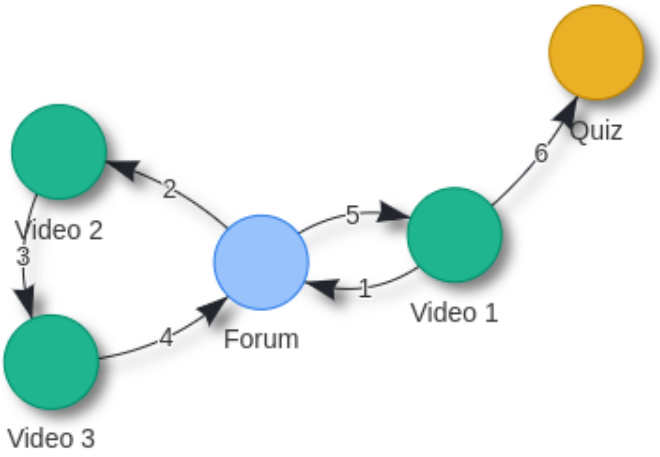


Figura 7: Grafo de período

$$\begin{array}{c} V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Una vez identificada la matriz de adyacencia se puede hallar la matriz de probabilidad de transición que establece la distribución de probabilidades de un salto entre un nodo y otro. Esta matriz es obtenida dividiendo cada elemento de A por la suma de los elementos de la columna correspondiente.

$$\begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{array}{c} \begin{array}{ccccc} V_1 & V_2 & V_3 & F & Q \end{array} \\ \begin{bmatrix} 0 & 0 & 0 & 0{,}5 & 0 \\ 0 & 0 & 0 & 0{,}5 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0{,}5 & 0 & 1 & 0 & 0 \\ 0{,}5 & 0 & 0 & 0 & 0 \end{bmatrix} \end{array}$$

Si navegamos a través de los contenidos educativos de cierto período, el vector de estado $x^{(k)}$ es un vector columna cuyo elemento i-ésimo es la probabilidad de que

estemos en el contenido i después de k "clicks". Supongamos, en el grafo G que salimos desde el nodo corresponsiet al Foro, en ese caso el vector de estado inicial será:

$$x_0 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

Ahora, si queremos conocer cuáles son las probabilidades de estar en otro nodo del grafo partiendo desde el nodo de Foro, basta con multiplicar la matriz de probabilidad de transición por el vector de estado x_0:

$$\begin{bmatrix} 0 & 0 & 0 & 0{,}5 & 0 \\ 0 & 0 & 0 & 0{,}5 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0{,}5 & 0 & 1 & 0 & 0 \\ 0{,}5 & 0 & 0 & 0 & 0 \end{bmatrix} * \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0{,}5 \\ 0{,}5 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

Esto indica que hay un 1/2 de probabilidad de salir al video 1 o al video 2 y 0 de probabilidad de salir al video 3, al Quiz o permaneces en el foro, ya que no existen enlaces en el grafo dirigidos hacia ellos. De estas probabilidades parten muchos de los algoritmos que buscan encontrar los contenidos más relevantes según sus aristas de entradas y salidas. En la siguiente sección haremos el cálculos de las principales medidas de centralidad para el grafo G.

6.1.2. Cálculo de centralidad

La primera medida de centralidad y la más básica es la centralidad de grado corresponde sencillamente al grado de un nodo o actor, esto es, al número de aristas o lazos que posee un nodo con los demás. A partir de la matriz de adyacencia del grafo, la centralidad de grado de un nodo i se pueden medir dos medidas de centralidad de grado diferentes, grado de entrada y grado de salida, es decir, respectivamente:

$$C_D^-(v) = \delta^-(v) \; y \; C_D^+(v) = \delta^+(v)$$

y como matrices de adyacencia:

$$C_D^-(i) = \sum_j a_{ji} \; C_D^+(i) = \sum_j a_{ij} \; C_D^+(i) = \sum_j a_{ij}$$

Este es un primer indicio de saber cuáles son los principales contenidos en el periodo a analizar o en una sesión específica de clase. Los nodos con una centralidad de grado mayor sobresalen en importancia en el grafo y depende del contexto interpretar esta importancia. por ejemplo el grado de entrada podría interpretarse como una medida de popularidad del grafo, mientras que el grado de salida como una de actividad o sociabilidad de ese contenido con otros.

La segunda medida de centralidad es la centralidad de cercanía que se encarga de determinar las rutas más eficientes que se deben recorrer para llegar desde un nodo a otro dentro de un grafo. En el contexto de las rutas de navegación de los estudiantes puede ser útil al calcular las rutas con la condición de que cada nodo del grafo sea visitado por lo menos una vez. Debido a que esta medida es de longitud, no tendrá un valor relevante en el análisis del grafo analizado.

La tercera medida es la intermediación o betweeness que cuantifica la frecuencia o el número de veces que un nodo se encuentra entre las geodésicas o caminos más cortos de otros actores. Formalmente, la intermediación $C_B(i)$ de un nodo i en una red o grafo se define como:

$$C_B(i) = \sum_{j \neq k \in V} \frac{b_{jik}}{b_{jk}}$$

Para los docentes esta medida puede servir para analizar que contenidos han servido más veces como puentes entre otros contenidos.

Por último la centralidad de vector propio mide la influencia de un nodo en una red. Los nodos de alta puntuación representan a aquellos vértices que tienen mayores conexiones y por lo tanto poseen un nivel de relevancia superior. Mientras que en

el caso de la centralidad de grado, cada nodo pesa lo mismo dentro de la red, en este caso la conexión de los nodos pesa de forma diferente. Esta medida permite comprender que que no solo es importante la cantidad de conexiones salientes o entrantes, sino la relevancia del nodo que da origen a los enlaces. La siguiente tabla muestra los valores obtenidos de las respectivas centralidades del grafo G dado como ejemplo en esta sección:

Tabla 8: Centralidad de G

nodo	**Grado**	**Cercanía**	**Intermediación**	**Vector propio**
Video 1	3	0.5714286	3	0.8269434
Video 2	2	0.4000000	2	0.5622851
Video 3	2	0.5000000	3	0.5622851
Foro	4	0.6666667	7	1.0000000
Quiz	1	0.2000000	0	0.2976268

6.2. Similitud de rutas

En este punto se comienza a estudiar una forma efectiva para comparar los diferentes grafos obtenidos en el modelo a través de un algoritmo implementado o propuesto por nosotros mismos. En la siguiente sección presentamos el algoritmo de la distancia de Levenshtein

6.2.1. Distancia de Levenshtein

La distancia de Levenshtein es un valor obtenido por un algoritmo que determina el número mínimo de operaciones requeridas para transformar una cadena de caracteres en otra, se usa ampliamente en teoría de la información y ciencias de la computación [81]. Se entiende por operación, bien una inserción, eliminación o la sustitución de un carácter. Esta distancia recibe ese nombre en honor al científico ruso Vladimir Levenshtein, quien se ocupó de esta distancia en 1965. Según la definición planteada y teniendo en cuenta que los grafos obtenidos están representando una secuancia de eventos ordenadas tal cual como una cadena de string, podemos proponer este algoritmo para comparar una ruta de aprendizaje de un estudiante frente a a la ruta de aprendizaje planteada por el docente. De este modo sabríamos con un indicio que

tan similares son los dos caminos.

Lo que indica la distancia de Levenshtein es la edición mínima que necesita una cadena para convertirse en otra a través de una inserción, sustitución o eliminación de un caracter. Así mismo, en el contexto educativo, lo ideal es saber con cuántos cambios mínimos una ruta de navegación de convierte en la planteada por el docente. Entre menos sea esa distancia mas parecida será.

Matemáticamente, la distancia de Levenshtein entre dos cadenas de caracteres a y b, cuya longitud son respectivamente $|a|$ y $|b|$, se puede expresar como $lev_{a,b}(|a|, |b|)$

$$lev_{a,b}(i,j) = \begin{cases} max(i,j) & si\ min(i,j) = 0, \\ min \begin{cases} lev_{a,b}(i-1,j) + 1 \\ lev_{a,b}(i,j-1) + 1 \\ lev_{a,b}(i-1,j-1) + 1_{a_i \neq b_j} \end{cases} & si\ min(i,j) = 0, \end{cases}$$

Veamos, la Figura 8 representa el camino ”ideal"planteado por el docente para recorrer los contenidos al que pertenece el grafo G. La idea es poder identificar que tan semejantes es el camino optado por el estudiante en ese período comparado con el establecido por el docente. Para eso comparamos las dos secuencias que se pueden identificar en el grafo G y en el grafo ideal y se plasman en un diagrama de Leveinshtein para determinar la distancia. La tabla 9 muestra el algoritmo aplicado y da como resultado una distancia de 2. Lo que equivale a que se deben hacer mínimo dos ediciones para que la ruta del estudiante del grafo G se convierta en la ruta establecida por el docente. <Cabe resaltar que las ediciones pueden ser sustitución, inserción o eliminación por lo que contemplan una relación de variaciones en un camino entre los diferentes saltos entre contenidos.

Este algoritmo se puede aplicar en las bases de datos de los grafos generados y sacar un promedio de las distancias de Levenshtein por periodo. De esta forma el docente podrá hacerse una idea, qué tan parecidas están siendo las rutas de los estudiantes con la suya.

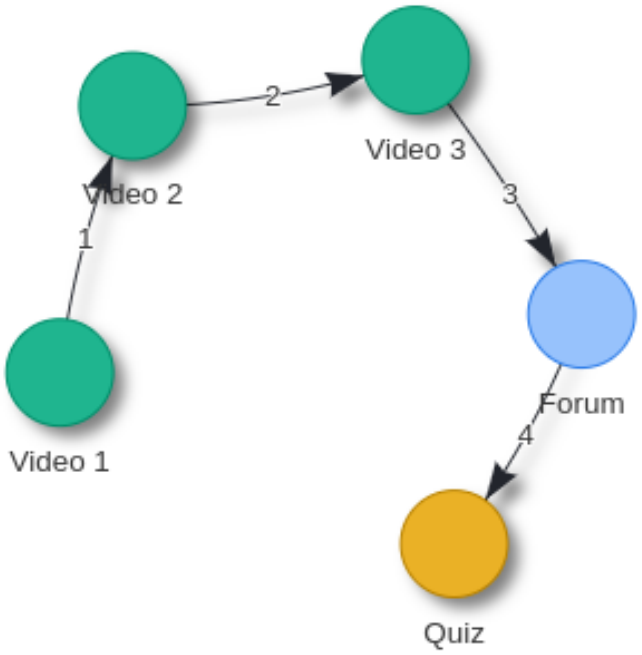


Figura 8: Camino ideal

6.3. Patrón de comportamiento

En la sección anterior, hemos logrado comparar una ruta de aprendizaje de un estudiante frente la establecida por el docente. En esta sección se quieren hacer análisis comparando las rutas de los estudiantes entre si mismas. Al tener grafos dirigidos que pertenezcan a un período definido, podrían compararse las transiciones más frecuentes de los estudiantes y definir un patrón de comportamiento por módulo

Tabla 9: Distancia de Levenshtein

		Y	V1	V2	V3	F	Q
		0	1	2	3	4	5
X	0	0	1	2	3	4	5
V1	1	1	0	1	2	3	4
F	2	2	1	1	2	2	3
V2	3	3	2	1	2	3	3
V3	4	4	3	2	1	2	3
F	5	5	4	3	2	1	2
V1	6	6	5	4	3	2	2
Q	7	7	6	5	4	3	2

del curso.

Dado que cada grafo dirigido se puede representar mediante su matriz de adyacencia A, es posible superponer las matrices de adyacencia que pertenezcan al periodo que se está analizando y generar un grafo H(V,E) que representa el patrón de comportamiento de dicho período. Matemáticamente podemos encontrar el grafo de patrón de comportamiento con el siguiente proceso.

Sea P el conjunto de grafos de todos los estudiantes del período:

$$P = \{G_1(V,E), G_2(V,E), G_2(V,E), ..., G_n(V,E\}$$

Cada elemento de P tiene asociada una matriz de adyacencia correspondiente A_{mxn} definida así:

$$A_{mxn} = \sum(a_{ij})$$

Estas matrices pueden sumarse o superponerse obteniendo una nueva matriz de orden mxn denotada con B:

$$B_{mxn} = \sum_{n}(A_{mxn})$$

Entonces, el valor más alto de cada fila de la matriz representará la transición más frecuente entre los nodos del grafo. Por lo que se comienza a contruir una matríz patrón donde el valor del elemento será 1 si es el mayor entre los demás de la fila y 0 en un caso distinto esta matríz estará definida por C_{mxn}:

$$C_{mxn} = \begin{cases} (a_{i,j}) = 1 & si \quad a_{i,j} > a_{nxm} \\ (a_{i,j}) = 0 & si \quad a_{i,j} < a_{nxm} \end{cases}$$

La nueva matriz C_{mxn} generada es la matriz de adyacencia del grafo patrón de

comportamiento del período analizado, Grafo representado como:

$$H_{i,j}(V, E)$$

Podemos ejemplarizar el algoritmo anterior planteado tomando una muestra de 6 estudiantes analizándolos en la un período específico, aprovechando el grafo G ya definido tomaremos 5 estudiantes más de ese período. La muestra ha sido pequeña para poder visualizar el grafo. Usando la reutilización de los nodos y generando el grafo C para los 6 estudiantes se obtiene el grafo presentado en la figura 9 y cuyas rutas se pueden identificar más claramente en la tabla 10. La ruta de aprendizaje del estudiante E1 es la del grafo G, la del estudiante 2 es la ruta establecida por el docente y los otros 4 estudiantes tomados de muestra del período analizado.

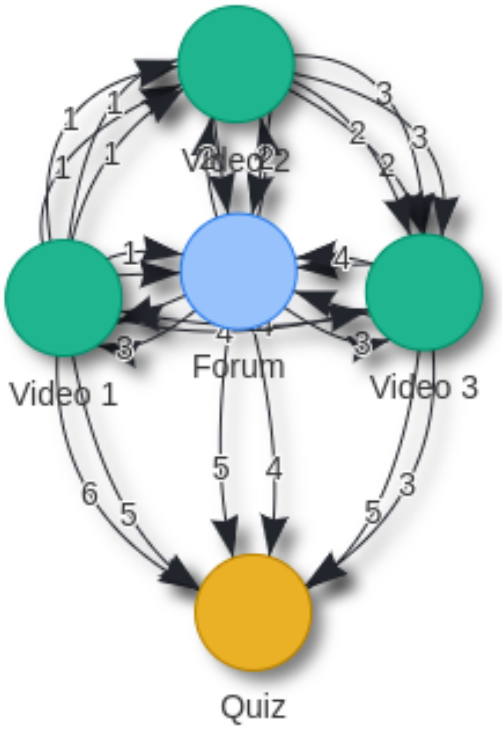


Figura 9: Múltiples estudiantes

Una vez obtenidos los grafos de los estudiantes, podemos extraer su matrices de adyacencia correspondientes.

Tabla 10: Muestra estudiantes

Identificación	Ruta de aprendizaje	Saltos
E1	V1-V2-F-V3-F-V1-Q	6
E2	V1-V2-V3-F-Q	4
E3	V1-V2-F-V3-V1-Q	5
E4	V1-F-V2-V3-F-Q	5
E5	V1-V2-F-V1-V3-Q	5
E6	V1-V2-V3-Q	3

$$E_1 = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} V_1 & V_2 & V_3 & F & Q \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E_2 = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} V_1 & V_2 & V_3 & F & Q \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E_3 = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} V_1 & V_2 & V_3 & F & Q \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E_4 = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} V_1 & V_2 & V_3 & F & Q \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E_5 = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} V_1 & V_2 & V_3 & F & Q \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$E_6 = \begin{array}{c} \\ V_1 \\ V_2 \\ V_3 \\ F \\ Q \end{array} \begin{bmatrix} V_1 & V_2 & V_3 & F & Q \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Al sumar las 6 matrices de adyacencia cuyas dimensiones son iguales obtenemos la matriz B_{5x5} como resultado:

$$B_{5x5} = \begin{bmatrix} 0 & 4 & 1 & 3 & 2 \\ 0 & 0 & 4 & 2 & 0 \\ 1 & 0 & 0 & 3 & 2 \\ 2 & 1 & 1 & 0 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Aunque la muestra es tan pequeña y es posible mayores repetidos como se puede observar en la matriz B obtenida en la fila 4, con una gran cantidad de datos sería muy poco probable que pasara. sin embargo, para el ejemplo que se tiene a conti-

nuación basta con tomar uno de los valores mayores. Escogiendo el salto F-Q en la fila 4, la transformación de la matriz B a una matriz de adyacencia C será:

$$C_{5x5} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Esta matriz C es la asociada al grafo de la figura 10, uno de los patrones de comportamiento encontrados para el periodo analizado. Algo importante de los patrones que se descubren con esta mecanismo es que representan los saltos mas comunes y no están estableciendo ningún orden especial como si lo establece los grafos de muestra. Sin embargo, con un patrón referencia ya es posible comparar que tan parecidos son los saltos y en general el camino optado por por el estudiante frente a los caminos establecido por sus demás compañeros.

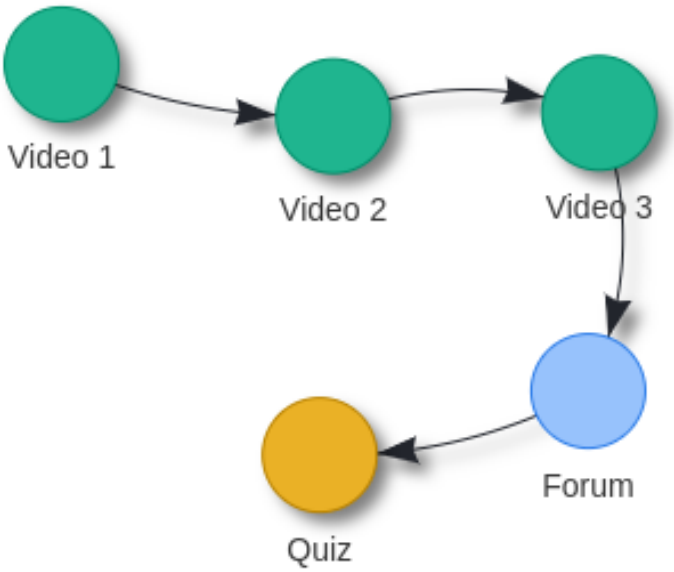


Figura 10: Patrón de saltos

6.4. Patrones de estados

En la sección anterior se pudo observar que es difícil comparar grafos secuenciales, cuyos caminos no sean del mismo tamaño, es decir que varían el número de saltos. Para poder abordar algún tipo de patrón de comportamiento que si tenga en cuenta el orden y la secuencia que ha usado el estudiante se propone en esta sección descubrir patrones según el número de estados que posee. Si un grupo de estudiantes dieron el mismo número de saltos, sus respectivos grafos pueden compararse hasta encontrar una similitud entre cada salto que dio. Para esto se deben comparar el par de nodos cuya arista del salto pertenecen.

Como ejemplo, y reutilizando los datos de muestra de la sección anterior vamos a descubrir los patrones de comportamiento de los 5 estados de los estudiantes E3, E4 y E5 quienes tiene el mismo número de estados igual a 5.

Para el estado inicial se tiene el siguiente conjunto de pares de nodos:

$$D_0 = \{V1 - V2, V1 - F, V1 - V2\}$$

Si sacamos la moda de este conjunto, podemos observar que la pareja de nodos que más se repite es $V1 - V2$, por tanto la arista $E_0 = V1 - V2$, Así mismo hallaremos la dupla para los estados $E_1E_2E_2, E_4E_5$

$$D_1 = \{V2 - F, F - V2, V2 - F\}$$

$$E_1 = V2 - F$$

$$D_2 = \{F - V3, V2 - V3, F - V1\}$$

En esta ocasión no hay una moda dentro del conjunto de duplas, para poder tomar la decisión de la dupla a escoger revisamos el nodo emisor que mas se repite y el nodo receptor que más se repite, o sea $E_2 = F - V2$

$$D_3 = \{V3 - V1, V3 - F, V1 - V3\}$$

Para este caso el nodo receptor no tiene un conjunto, por lo pequeño que es la muestra tomaremos uno aleatorio así $E_3 = V3 - V1$

$$D_4 = \{V1 - Q, F - Q, V3 - Q\}$$

$$E_4 = V1 - Q$$

De esta forma podemos identificar en orden de secuencia los estados más comunes entre los estudiantes que hicieron el mismo número de saltos, otro análisis importante que podría dar ideas de qué orden están tomando lo estudiantes antes o después de un nodo, interesante por ejemplo revisar los nodos que se visitan antes de realizar una evaluación o después de de ella. O si mientras realizan una evaluación revisan otros contenidos se podrían identificar alguna sospecha de comportamientos deshonestos.

Los análisis estadísticos presentados en este capítulo fueron ejemplificados analizando los períodos o módulos de los cursos por razón de aplicación, pues pueden interpretarse más fácil, sin embargo pueden aplicarse a los diferentes tipos de grafos que la herramienta de visualización permite obtener. El siguiente capitulo presenta un caso de estudio con una muestra significativa real de la Universidad del Cauca, en el cual podremos validas y evaluar el modelo de representación diseñado, podemos visualizar los diferentes tipos de grafos generados y analizar los grafos obtenidos a través de los algoritmos y operaciones estadísticas propuestas.

7. Validación y evaluación del modelo y similitud de grafos en un curso en línea con reconocimiento académico

7.1. Contexto Educativo y Recopilación de Datos

Los datos usados para este caso de prueba fueron recopilados del curso electivo en modalidad virtual “Introducción al emprendimiento con Lean Startup"soportado en la plataforma Selene de la Universidad del Cauca correspondiente a una instancia del sistema de gestión de aprendizaje de software libre Open edX. Se consideró una muestra de 98 estudiantes (matriculados) de la cohorte 2018-2, conjunto de datos disponible para la investigación que no incluye información personal que permita identificar a los individuos. Se recopilaron sólo los eventos relevantes establecidos en el cuadro 1 obteniendo una muestra de N=34859 eventos. Este curso se compone de los módulos presentados en la figura 11, sin embargo como períodos del curso están establecidos seis módulos y aparte los módulos descripción, calificaciones, actividades externas. Cada módulo contiene subsecciones dedicadas a enseñanza a través de vídeos y lecturas; discusión a través de foros, en donde también se presentaban actividades de aprendizaje y una evaluación al finalizar cada módulo. El concepto de período en este trabajo hace referencia a los contenidos liberados cada cierto tiempo por el docente, que para nuestro caso de estudio corresponde con los módulos. Habiendo verificado a través de un filtro que cada evento de la muestra son exclusivamente los de nuestro interés, se procedió a organizarlos cronológicamente y generar los archivos necesarios para que la herramienta pueda construir los diferentes grafos a través del modelo definido en esta propuesta.

7.2. Grafos generados

Inicialmente son analizados los grafos que muestren las trayectorias en un período específico de la clase o en este contexto, en un módulo del curso. Moockly dispone para este enfoque dos filtros diferentes, uno como el presentado en la figura 12 para seleccionar el módulo y una sesión específica y otro como el presentado en la figura

Figura 11: Módulos del curso

13 para seleccionar solo el módulo, es decir la unión de todas las sesiones de ese módulo en un grafo. Con el primer filtro es posible generar grafos IWS en donde el proceso del algoritmo de la herramienta es seleccionar el período, actualizar el dataframe, seleccionar el estudiante, actualizar el dataframe y por último seleccionar la sesión actualizando por última vez el dataframe. Para el caso de querer visualizar un grafo IW, se utiliza el segundo filtro, se opta por no seleccionar ninguna sesión y la herramienta representa el tiempo desde la primera hasta la última sesión de ese período, es decir, sin distinguir sesiones. El total de grafos IWS generados serán todas las posibles combinaciones de este filtro, que teóricamente se resume al número total de sesiones registradas equivalentes a 324. En otra pestaña de la herramienta se brinda la posibilidad de analizar sesiones de un estudiante, sin distinguir entre períodos. Un ejemplo de un filtro para generar un grafo IS que es de este tipo, es el mostrado en la figura 14. Únicamente basta con seleccionar el estudiante y una sesión. Ahí el docente puede analizar si en algunas sesiones de un módulo correspondiente, revisó otros módulos anteriores.

Para ejemplificar un análisis con ayuda de la herramienta, vamos a visualizar los grafos considerando el comportamiento de un estudiante con registro de actividad alto frente a un estudiante con registro de actividad bajo durante el Módulo 3 del curso denominado “Unidad temática 2: Validación de problemas”. Las figuras 15 y 16 muestran los grafos generados para el estudiante activo de dos sesiones consecutivas respectivamente. Es posible identificar los diferentes tipos de contenido por colores y la etiqueta de cada nodo y las trayectorias con la etiqueta “Step” para identificar el orden de la secuencia. Todo grafo generado que represente una

Figura 12: Filtro para grafo IWS

Figura 13: Filtro para grafo IW

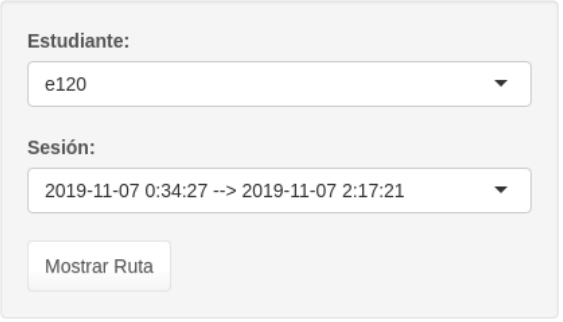


Figura 14: Filtro para grafo IS

sesión, necesariamente comienza y termina con un nodo “Signin” y un “Signout” respectivamente. De esta manera el grafo puede ser entendido fácilmente por los docentes sin tener un conocimiento técnico en teoría de grafos, puede visualizarse un nodo con el nombre other, que indica que el estudiante ha salido del módulo que se está analizando y puede haber ido a cualquier otro recurso pero de un módulo

diferente.

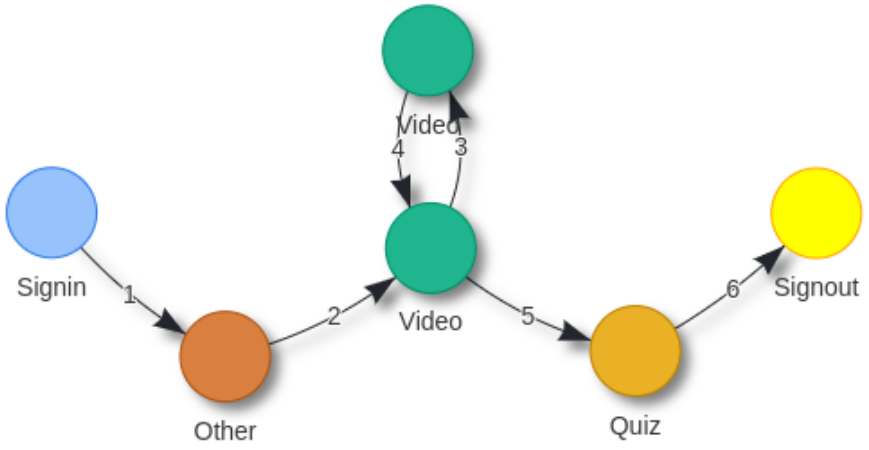


Figura 15: Grafo IWS estudiante activo, sesión 1

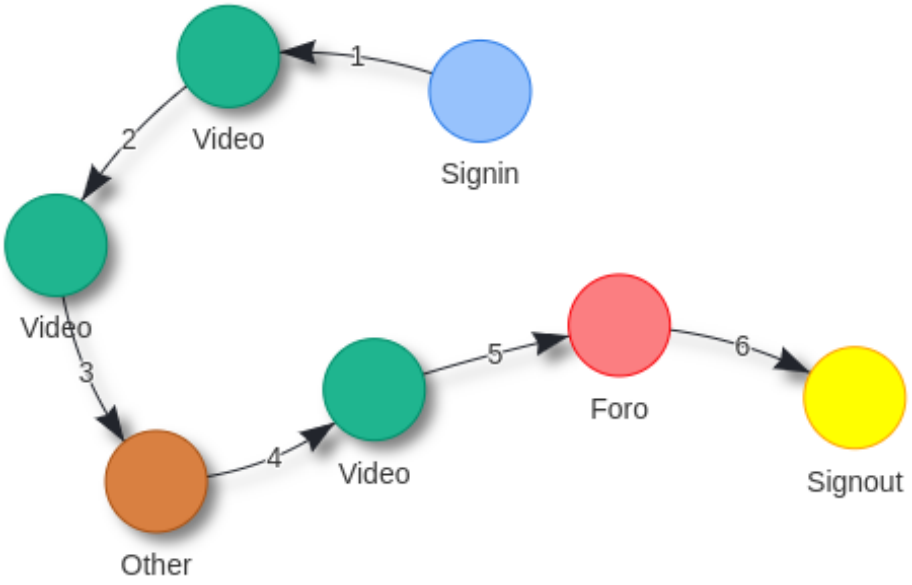


Figura 16: Grafo IWS estudiante activo, sesión 2

Es interesante analizar e interpretar estos grafos y compararlos con comportamientos pedagógicos comunes. Por ejemplo, el nodo "Other" justo al iniciar sesión es muy común en muchos de los grafos y se puede explicar que los estudiantes antes de entrar a algún módulo específico del curso revisan secciones como guías introductorias, descripción del curso o la sección de calificaciones. En la figura 6 se puede evidenciar también que, aunque el estudiante había realizado el Quiz en la sesión anterior aún sigue revisando vídeos y consultado o participando en el Foro. Según la actividad realizada en el foro al final del grafo, estos vídeos se pueden interpretar como temas

de discusión complementario que, aunque no son evaluados son necesarios para discusiones en foros o también por ejemplo que el estudiante quiera verificar en el vídeo el contenido preguntado en el Examen. Por otra parte, la figura 17 presenta el grafo del estudiante con baja actividad en el curso. Este grafo describe la actividad en el mismo módulo en que se analizó el estudiante comprometido. Se puede percibir que este tipo de grafos es un patrón muy probable en los MOOC de aquellos estudiantes que se limitan únicamente a cumplir el compromiso que en la mayoría de casos es realizar la evaluación. En los enlaces de los grafos también se puede visualizar una etiqueta con la fecha y hora en la que se realizó la transición, dato útil que los docentes pueden aprovechar para identificar el tiempo que están gastando los estudiantes en las actividades.

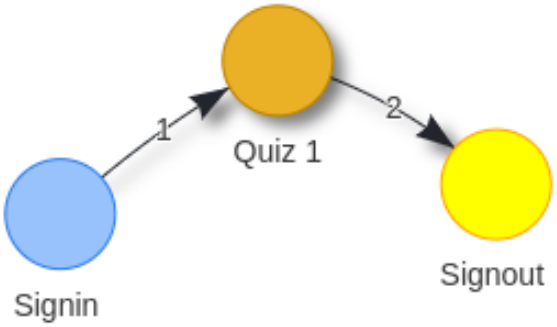


Figura 17: Grafo IWS estudiante inactivo

La figura 18 presenta un grafo IW, que representan la unión de la actividad de las 2 sesiones que tuvo en dicho período, nótese que como es un grafo IW, no tiene los nodos de inicio y fin de sesión y a partir del nodo Quiz que termina la primera sesión seguidamente este se ve enlazado con el video con el que inicia la segunda sesión. Si quisiéramos darle una interpretación a este comportamiento podríamos plantear que en la primera sesión, el estudiante ingresó a cumplir a cumplir el requisito de la evaluación y en la segunda sesión el estudiante ingresó a cumplir el requisito de por lo menos tener una participación en el foro.

La idea de seleccionar alguna característica en común como se hizo para analizar los grafos (estudiante activo e inactivo), puede llevar a agrupar cierto número de grafos según esa característica. Esto conlleva a pensar en multitudes de grafos generados denominados grupales que compartan algo en común. Algunas de las características más comunes pueden ser: Rangos de edades, ubicación geográfica, género, poseer co-

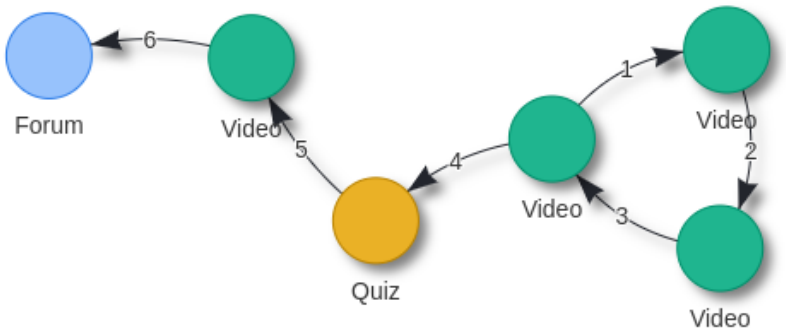


Figura 18: Grafo IW

nocimientos previos de la temática, rendimiento académico, estilos de aprendizaje, entre otros. Estas características deben ser suministradas por el modelo a la herramienta para así mismo tener un criterio de clasificación y visualizar algún tipo de grafo grupal. Por ejemplo, para este caso de estudio se realizó un criterio de clasificación según el número de ingresos a la plataforma identificando al estudiante como activo e inactivo. En este enfoque el modelo permite al usuario seleccionar el grupo de estudiantes y el período a analizar, el grafo generado incluye las actividades de la semana y las rutas de los estudiantes de este grupo. Los grafos generados hasta aquí, permiten descubrir patrones de comportamiento en cuánto a un módulo específico y permiten al docente conocer los hábitos preferidos de los estudiantes para consultar los contenidos para, si es del caso, rediseñar la estructura del curso. Sin embargo, la herramienta también puede generar grafos que no están centrados en las semanas o módulos de clase sino en las sesiones completas que pueden incluir uno o más módulos del curso como se analiza a continuación. Analicemos una sesión específica de un estudiante sin distinguir módulos o períodos, el docente solo debe seleccionar el estudiante y una sesión completa que tuvo dicho estudiante. Para este caso de prueba se seleccionó un estudiante comprometido y una sesión registrada en las últimas semanas, esto con el fin de abarcar contenidos de módulos anteriores. El resultado del grafo obtenido se visualiza en la figura 19. Cada color en estos grafos representa una semana o módulo diferente del curso y los números que acompañan la etiqueta del nodo corresponden a la unidad de la subsección. Es decir que ahora

se puede identificar en qué modulo, en qué unidad y qué tipo de actividad realizó el estudiante de una forma muy sencilla para el docente.

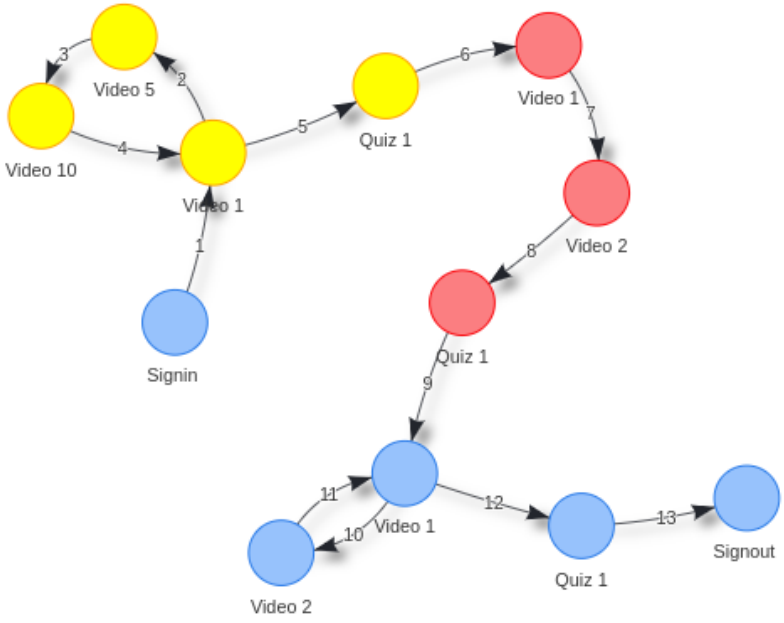


Figura 19: Grafo IS estudiante activo

7.3. Análisis de resultados

El modelo propuesto efectivamente ha generado grafos de diferentes tipos para analizar. Con este caso de estudio fueron generados en total 2527 grafos, sin tener en cuenta grafos grupales distribuidos según el tipo, de la siguiente forma: IWS= 1386, IS= 651, IW= 490. En función del tiempo disminuye en cada periodo, inicia con 786 grafos, en el segundo periodo ya se visualizan 622, en el tercero 485, en el cuarto 338 y para el examen final se registraron 296 grafos. Además se visualiza un ejemplo de agrupación en el que se seleccionaron sólo estudiantes que enviaron respuestas a evaluación que suman GW= 362. Estas cantidades indican que la amplitud de la muestra para un análisis estadístico depende directamente del número de estudiantes y sesiones existentes. Sin embargo, también tiene una gran dependencia de lo mucho que se mueva el estudiante en la plataforma. La interpretación de estas cantidades de grafos por periodo es que disminuye la audiencia y/o la actividad de interacción con la plataforma.

Al comparar la cantidad de contenidos visitados visualizados a través de los grafos,

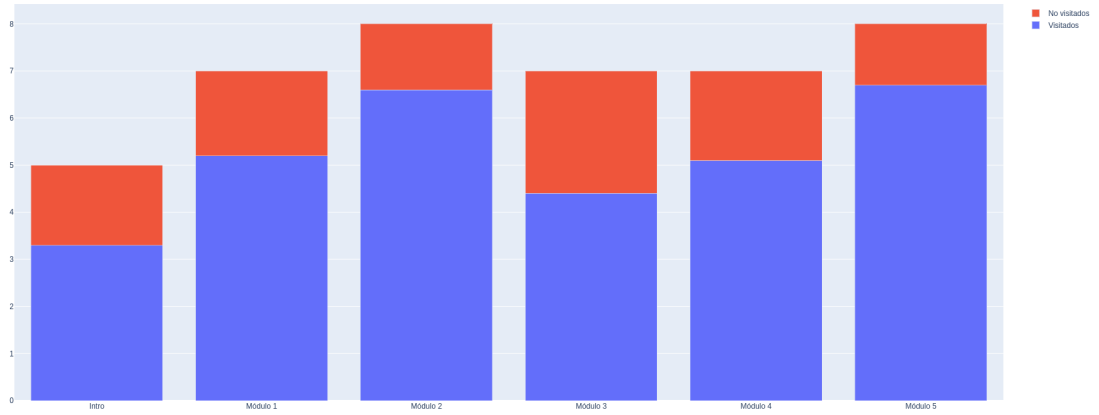


Figura 20: Módulos del curso

con el total de contenidos por módulo. Es posible obtener un promedio de la cantidad de contenidos visitados frente el total. La figura 20 muestra dicha información. La gráfica aporta que la participación de los estudiantes ha sido sobresaliente por lo menos en cada módulo del curso.

Se han podido clasificar a los estudiantes con dos categorías: activos e inactivos, gracias a los grafos obtenidos podemos visualizar cuántos contenidos visitó cada estudiante en cada módulo, este análisis se hace con el fin de comparar cual es la correlación entre el tiempo activo y la participación de cada estudiante frente a la cantidad de contenidos que visitó registrados en los nodos de los grafos generados. la figura 21 presenta la cantidad de nodos visitados de todos los estudiantes de la muestra activos e inactivos en cada módulo inactivo. Se puede analizar que los estudiantes inactivos tienen muy poca participación en los últimos contenidos del curso, lo que sucede normalmente cuando hay decersión por un estudiante. También se ha querido analizar las visitas en general, unificando los módulos. La figura 22 muestra esta actividad, en la gráfica los puntos rojos representan los estudiantes inactivos y los azules los activos.

Algunos datos obtenidos gracias al análisis de los grafos de la actividad de los estudiantes en el Módulo 2 fueron: El promedio de sesiones por estudiante en este módulo es de 2.72 sesiones, el número de grafos en este módulo difiere en un 38.29 % menos que el Módulo inicial lo cual indica una posible deserción en etapas tempranas

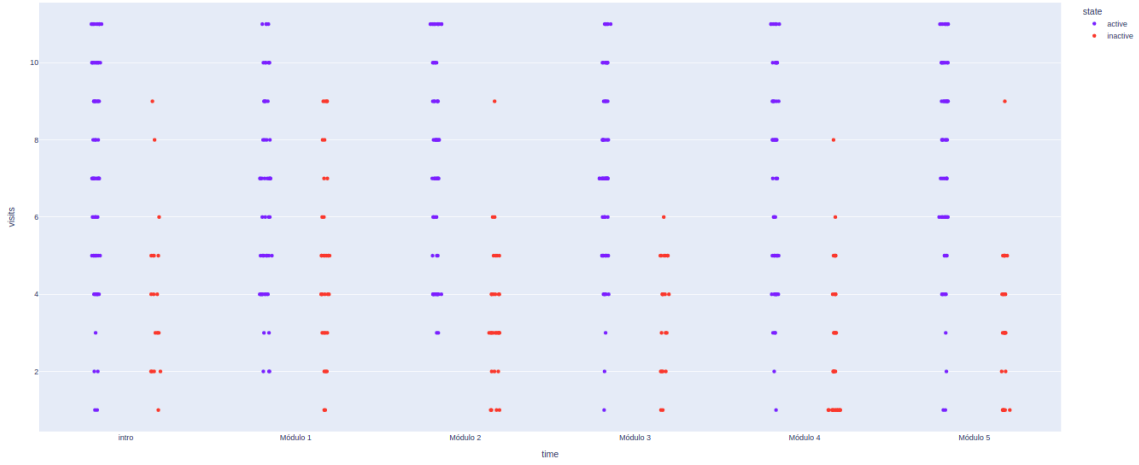


Figura 21: Contenidos visitados por módulo

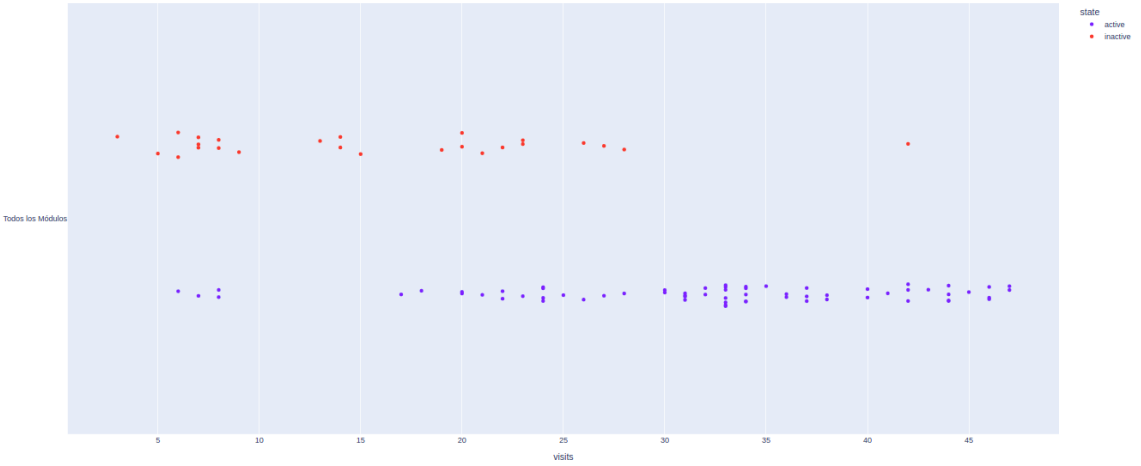


Figura 22: Contenidos visitados totales

del curso. La herramienta también sirve para detectar el grado de centralidad de un grafo; por ejemplo, de todos los grafos generados, se observó en el nodo correspondiente al Video número 2 de la Unidad Temática 1, con una centralidad promedio de 3.85 entre todos los estudiantes, es el más importante pues fue el que registró más entradas y salidas. Dicho nodo también es el más concurrido como puente para saltar a otro nodo (el de evaluación en muchos casos) lo que en teoría de grafos es definido como el Betweeness o intermediación y que para este caso su valor es de 1.82, el número de veces en promedio que sirvió como puente.

Para poder entender mejor el concepto de las centralidades de los grafos, se ha tomado como ejemplo la Unidad temática 3. La figura 23 es la gráfica que representa en el eje x los contenidos de dicho período o módulo y el eje y el valor de la centralidad de intermediación. Además el área o tamaño de la burbuja es proporcional a las veces que fue visitado el nodo del total de los estudiantes sin distinguir si era activo o inactivo.

El nodo de Quiz es un claro ejemplo de que aunque sea visitado muchas veces un nodo no necesariamente sea el más influyente en el grafo, pues casi todos los estudiantes pretenden cumplir con el requisito de la evaluación, sin embargo si se compara con el nodo de foro este además de ser altamente visitado es un puente importante de saltos entre contenidos.

En general, entre otros cálculo posibles, con la generación de los grafos fue posible obtener información valiosa para analizar, por ejemplo, por mencionar:

- Para IWS puede conocerse el porcentaje de estudiantes que recorrieron todos los contenidos en una única sesión y el tipo de actividades más comunes en la primera sesión del período.

- Para IW el porcentaje de estudiantes que siguieron el patrón diseñado por el docente durante el período analizado, el número de nodos posibles a recorrer contra el número de nodos abarcado por el estudiante en ese período y el tipo de actividad más y menos frecuentada del período. También el nodo destino más común en saltos entre contenidos del período, es decir el contenido que fue visitado más veces por el mismo estudiante.

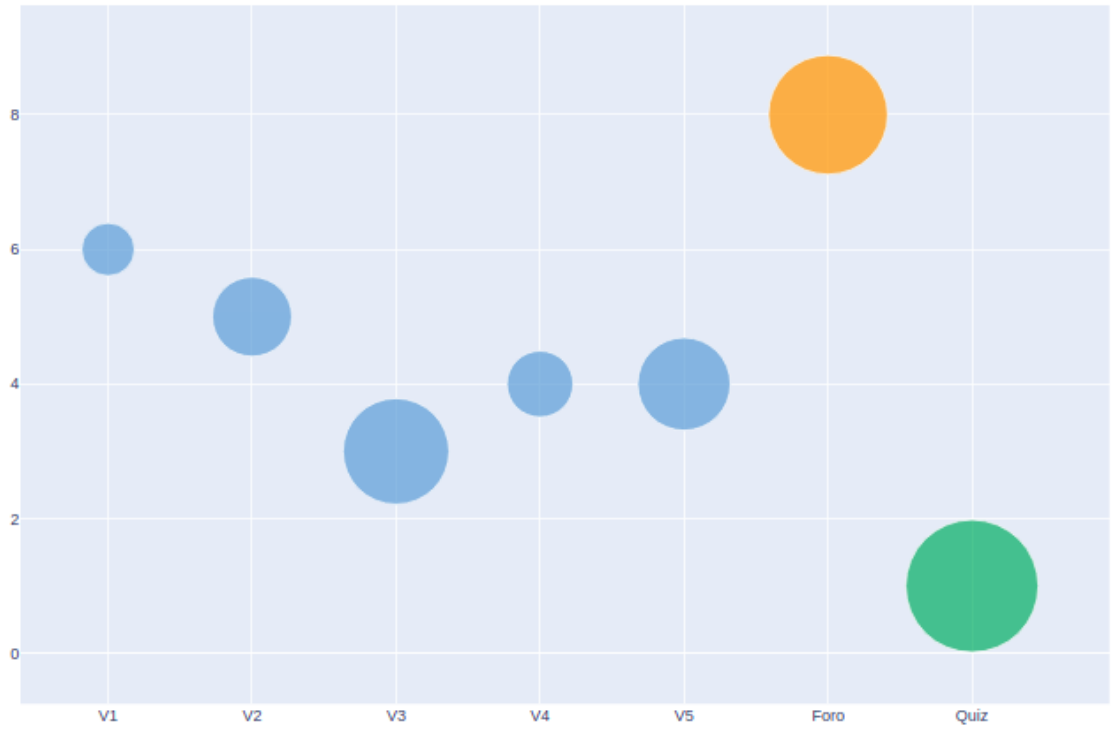


Figura 23: Centralidad y Visitas módulo 3

- Para IS, los módulos externos que más se volvieron a visitar después de haberlos cursado y el contenido más visitados previo y posterior a una evaluación.

- Para IG el patrón de comportamiento generado al acotar la muestra de los estudiantes según una característica específica.

7.4. Exploración de patrones

Ahora bien se han realizado los análisis estadísticos propuestos en el capítulo 5, sin embargo en esta sección se quiere realizar una exploración de grafos con el fin de descubrir patrones de comportamiento a través de la aplicación Moockly cuyos datos del caso de estudio pueden ser analizados por los docentes que impartieron el curso. Para focalizar un poco la exploración se ha seleccionado únicamente el período denominado "Módulo 4, unidad temática 2ƨ se le ha pedido al docente de dicho período realizar un análisis, clasificando a los estudiantes según los contenidos revisados y la ruta de aprendizaje que decidieron seguir durante el período.

Los resultados más significativos se muestran en la siguiente clasificación de patrones descubiertos.

7.4.1. Patrones de Navegación Lineal

En el mejor de los casos, lo ideal en el aprendizaje de un estudiante durante un curso sería que siguiendo la metodología del docente y la ruta que él ha propuesto, comprendiera cada tema y no tuviera la necesidad de devolverse o en los contenidos o repetir vídeos por ejemplo. A los estudiantes que se observó con este comportamiento se les ha clasificado en este grupo de navegación lineal, la figura 24 muestra el patrón de comportamiento de uno de los estudiantes que siguió la ruta definida por el docente y revisó todos los contenidos en orden. Aquí, los grafos son secuencias cuyos nodos no son visitados más de una vez, o no se presentan ciclos en las rutas. Los estudiantes revisan en secuencia los contenidos y finalizan realizando la respectiva evaluación del período.

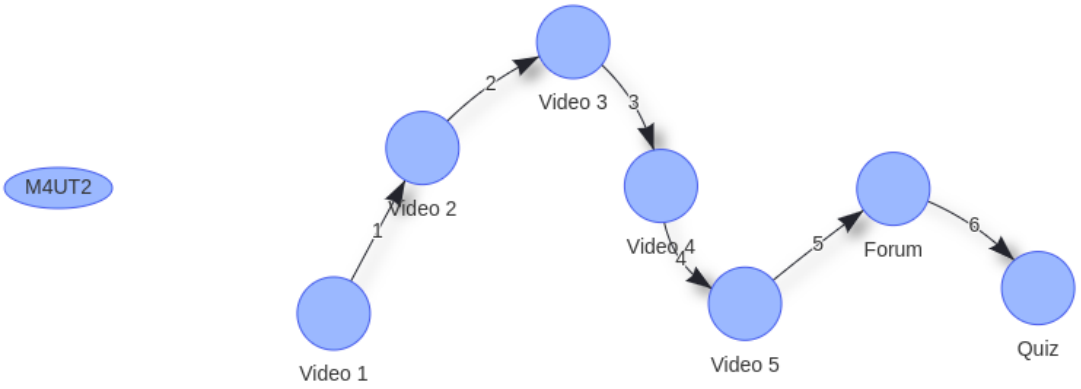


Figura 24: Navegación Lineal 1

Otro patrón similar importante para destacar es de los estudiantes que tienen una perfil focalizado en aprobar el Quiz. Visitaron los contenidos necesarios linealmente para terminar en el Quiz, ignorando otro tipo de contenido opcional o tal vez no relevante en la evaluación. este es el caso de la figura 25 donde el grafo es similar al patrón anterior con la diferencia de no visitar el foro del periodo.

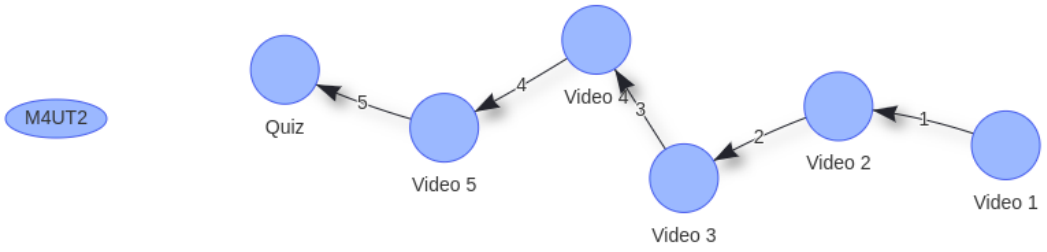


Figura 25: Navegación Lineal 2

Un patrón válido correspondiente a estudiantes con una buena navegación lineal, es el mostrado en la figura 26. Que además de revisar todos los contenidos en secuencia deciden ir a participar en el foro de discusión antes de entrar a la evaluación. A nivel cognitivo esta es una buena práctica en los cursos en línea, reforzar o ampliar los conocimientos aprendidos antes de ser evaluados.

Los grafos de navegación lineal también pueden mostrar patrones que involucren contenidos de otros períodos, como lo muestra la figura 28. Muy seguramente es un comportamiento de estudiantes que han dejado acumular los contenidos y en el momento de cumplir el compromiso de un periodo en específico, ingresan a la plataforma y se desatrasan. Sin embargo parece ser que el patrón de la figura mencionada, es un ejemplo de un estudiante que antes de revisar los contenidos del periodo a estudiar a decidido ver los vídeos del periodo anterior a manera de repaso o contextualización, algo válido y común de estudiantes con índices bajos de retención o atención.

7.4.2. Patrones de Navegación No lineal

En esta sección se exploran patrones de grafos de comportamientos no lineales en la forma como navegan los estudiantes. Con no lineales se refieren a grafos donde al menos uno de los nodos tienen 2 o más aristas que entrando en ellos, es decir, cuando un estudiante revisa un contenido más de una vez. Al ocurrir estos comportamientos, se forman ciclos, o bien llamados lazos cerrados. La figura 29 es un ejemplo sencillo de un ciclo formado al revisar el video anterior nuevamente y después ya realizar

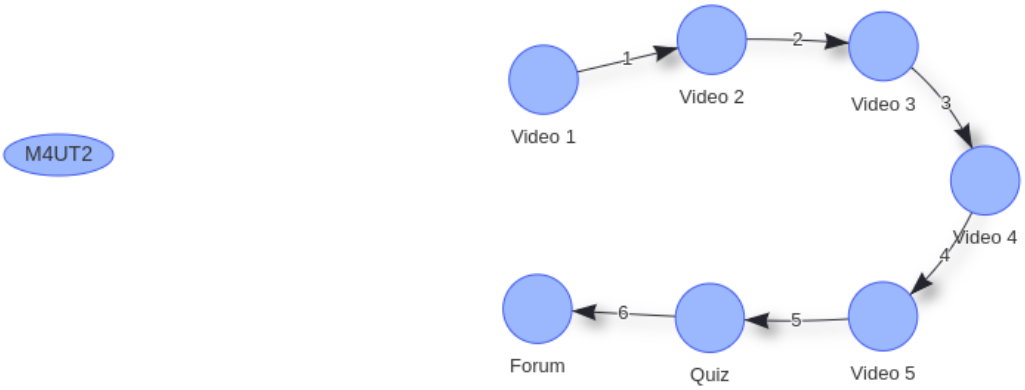


Figura 26: Navegación Lineal 3

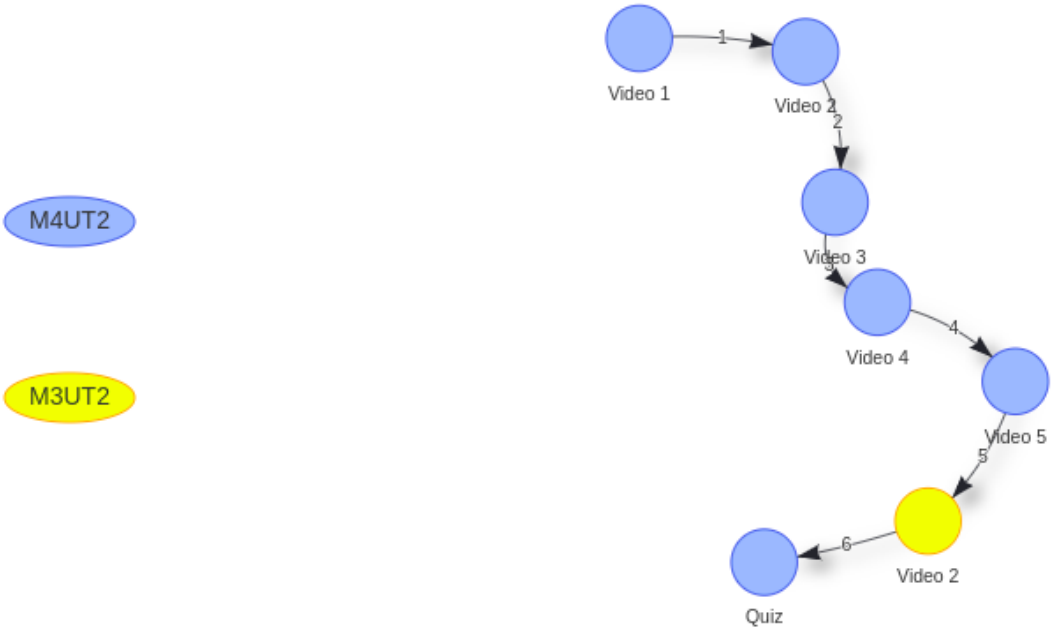


Figura 27: Navegación Lineal 4

la evaluación. Así como este grafo se encontraron muchos con el mismo patrón. Por ejemplo la figura 30 muestra un grafo similar y lo curioso es el mismo nodo que representa al video 4 que ha sido visitado nuevamente. Puede ser una tarea importante identificar las razones por lo cual se crean estos ciclos con el video 4. Una de las razones puede ser que el video no haya sido lo suficiente claro, también que

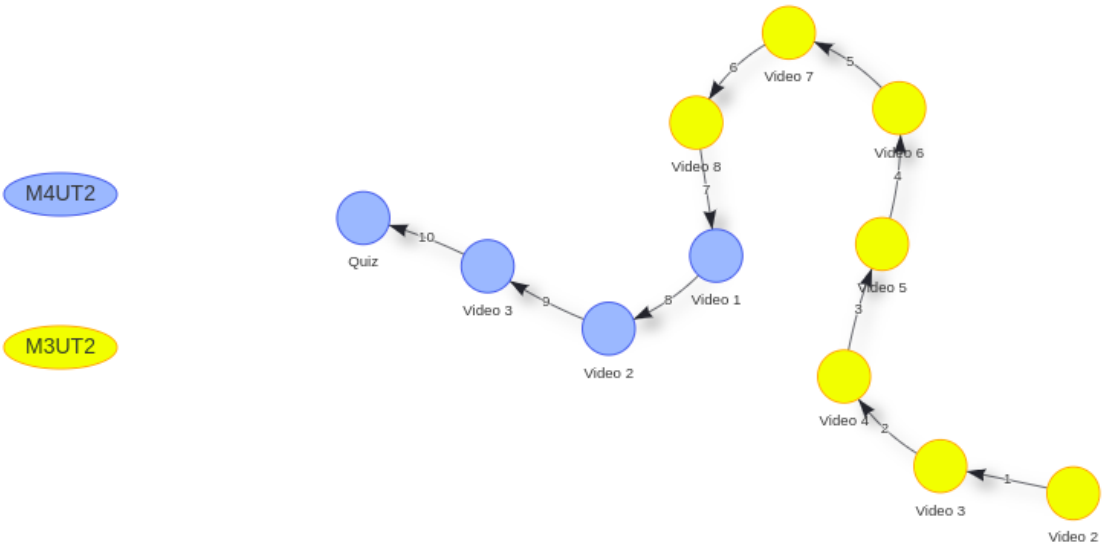


Figura 28: Navegación Lineal 5

el video siguiente esté relacionado con dicho video y haya sido necesaria una revisión al contenido anterior, en fin, esto podría hacer reflexionar al docente y revisar dicho video. Estas prácticas de exploración de patrones tienen como aplicación aportar a la mejora del diseño tanto del curso como de sus contenidos.

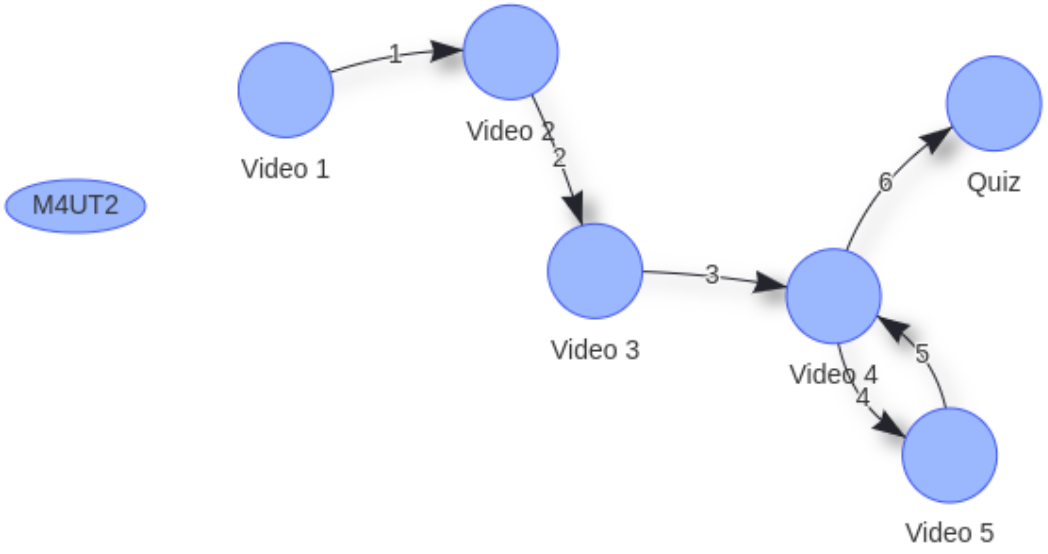


Figura 29: Navegación No Lineal Vídeos Continuos 1

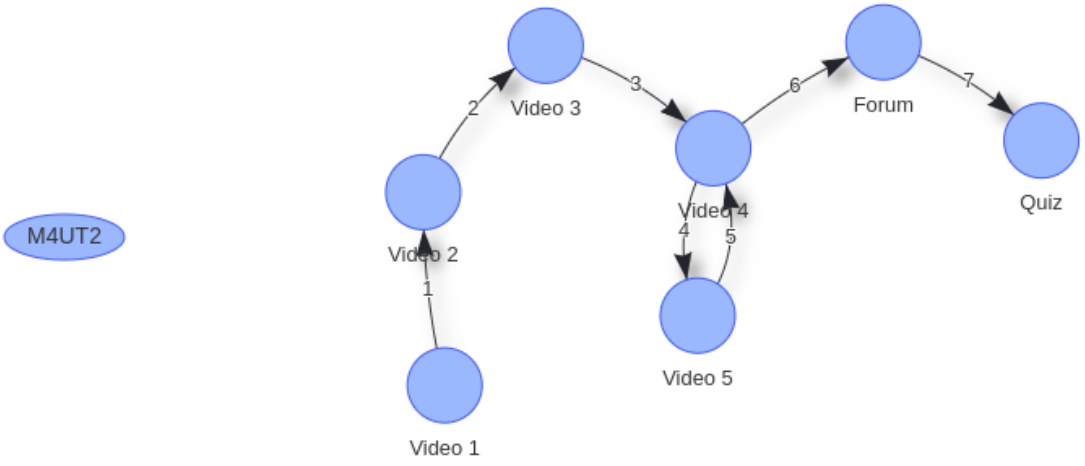


Figura 30: Navegación No Lineal Vídeos Continuos 2

El video 4 puede ser un protagonista importante para revisión, sin embargo otro de los patrones encontrados es el mostrado en la figura 31, donde el ciclo se presenta con el vídeo 1 también presente en la figura 32, donde curiosamente este estudiante ha ingresado a inicialmente a realizar la evaluación sin revisar ningún contenido anteriormente. Aunque posteriormente haya revisado los vídeos no quedó registrado (por lo menos durante las fechas de este período) el ingreso nuevamente a la evaluación. Los ciclos presentados en estos ejemplo son pequeños, esto es debido a que los nodos visitados nuevamente están continuos a la secuencia, estos saltos son los más comunes, es decir entre videos consecutivos, es por eso que estos patrones han sido clasificados además como No lineales, con ciclos de nodos continuos, cuya continuidad la define la secuencia establecida por el docente.

Por otra parte, también se han descubierto patrones donde los ciclos formados no tienen la continuidad presentada en el grupo anterior. La figura 33 que el ciclo se forma al momento de devolverse del video 3 al video 1, en este caso como no se devolvió al video anterior hay una discontinuidad en los saltos. Se puede notar que entre más distancia haya entre los contenidos en que se presenta la No linealidad, más grande será el lazo que se forma en el grafo. Analizar el patrón de esta figura puede llamar la atención que el video 1 suele ser visitado más de una vez desde cualquier otro contenido por lo que conlleva a pensar que tiene información relevante para todo el período.

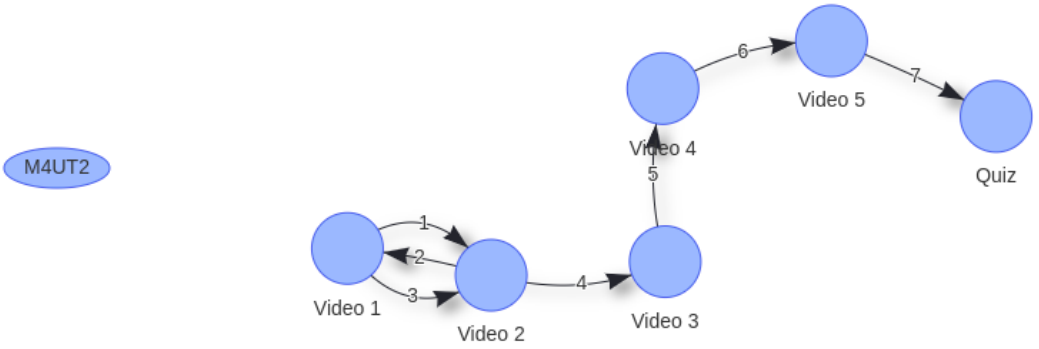


Figura 31: Navegación No Lineal Vídeos Continuos 3

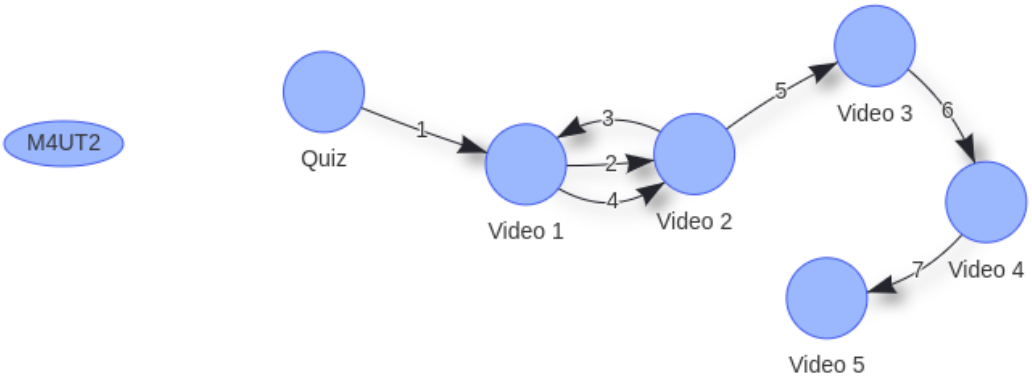


Figura 32: Navegación No Lineal Vídeos Continuos 4

Para escalar otro grado en la magnitud de los ciclos se ha identificado el patrón de la figura 34 donde el salto ya es desde el video 5 al video 2, se puede observar por las secuencias que no solo se devolvió a ese video en especial y siguió su plan, sino que de nuevo volvió a revisar en orden los contenidos desde el video 2 hasta la evaluación. El video que aparece al finalizar el grafo pertenece al capítulo introductorio, estas visitas a contenidos pueden obviarse siempre y cuando no tengan relación alguna entre los temas que se traten en el período analizado, es decir si se quiere en un futuro hacer una comparación de patrones similares, podrían eliminarse para reducir el ruido en el análisis de grafos.

En la figura 35 se puede observar que los foros también hacen parte importante de

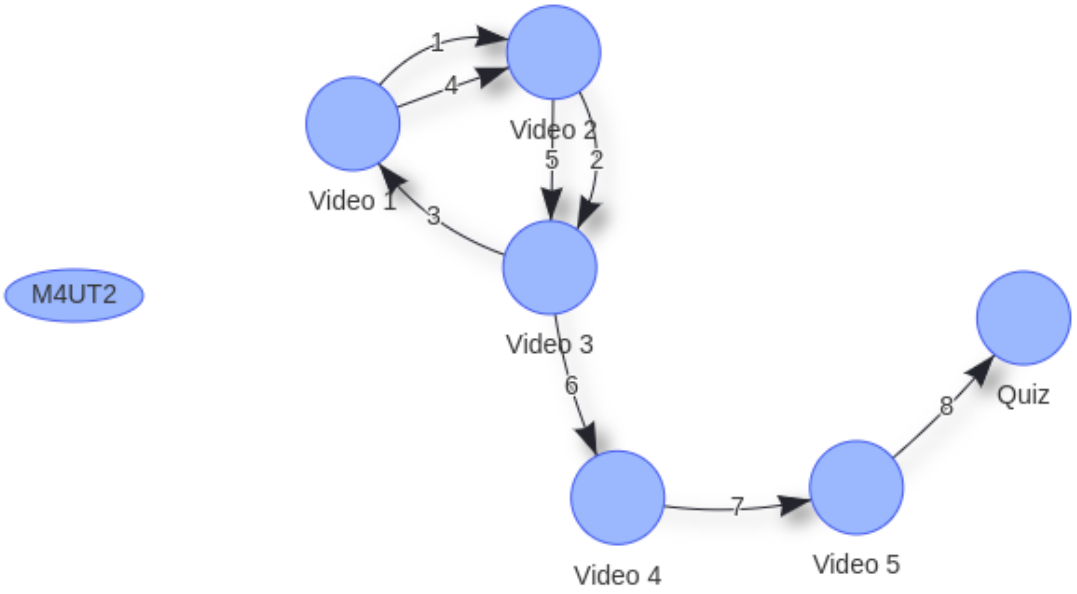


Figura 33: Navegación No Lineal Nodos Discontinuos 1

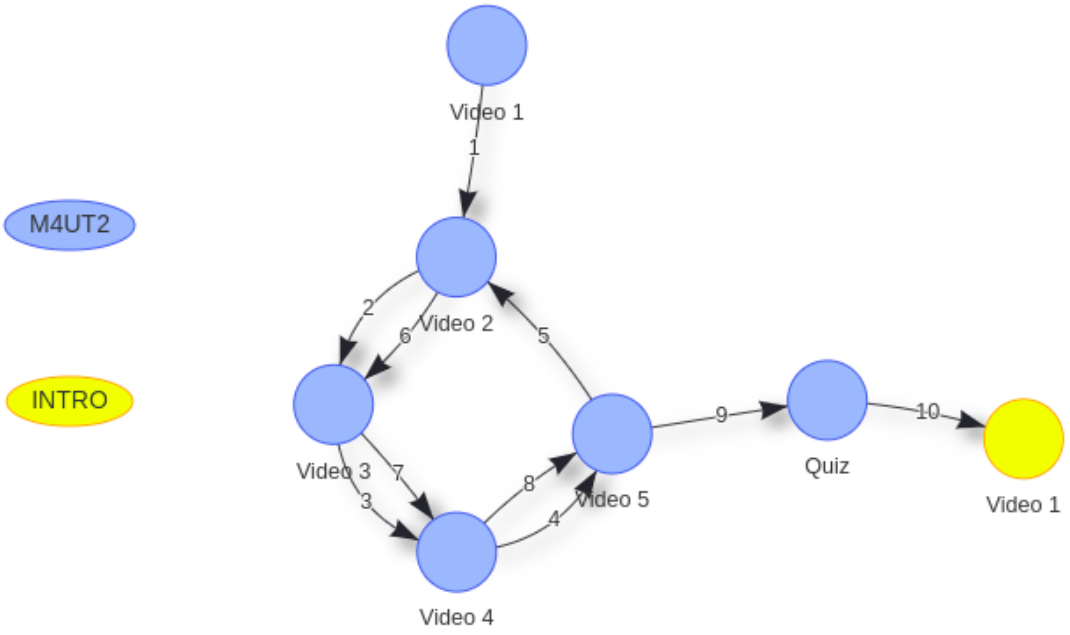


Figura 34: Navegación No Lineal Nodos Discontinuos 2

los ciclos, ya que estando en al alguno de los videos en cualquier momento puede presentarse un tema interesante para compartirlo en los foros, esta acción provocaría la creación de un lazo cerrado en el grafo y puede contribuir a a descubrir las temáticas que mas se discuten en los foros siguiendo el nodo anterior visitado.

La figura 36 muestra patrón con el salto mas grande entre los videos que se puede

presentar, ahí el estudiante salta del video 5 al video 1 y recorre nuevamente cada contenido de siguiendo la secuencia que ha planteado el docente. Este patrón es poco común entre los encontrados, sin embargo se puede reflexionar que el estudiante o le ha dedicado tiempo a los contenidos del curso para no fallar en la evaluación, o simplemente no ha comprendido bien en resumen todos los videos que visitó y decide volverlos a ver en el orden establecido. Este patrón también finaliza con la evaluación lo que significa que cualquiera que sea su caso busca cumplir con el requisito del período que es aprobar.

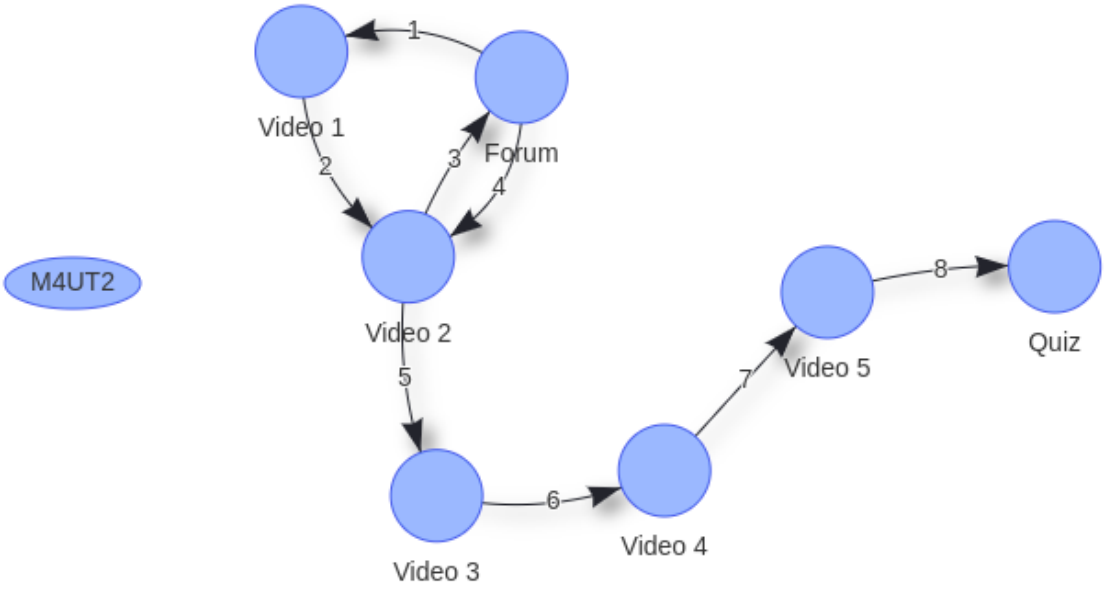


Figura 35: Navegación No Lineal Nodos Discontinuos 3

Hasta ahora, han sido descubiero patrones de comportamientos que tienen una estructura de grafo simple, porque a lo sumo poseen un ciclo durante la secuancia de navegación. A continuación se presentarábn patrones descubierto con una estructura más compleja, donde los ciclos son 2 o más. La figura 37 muestra el ejemplo para dos lazos cerrados, patrón encontrado dándole un valor especial al nodo del video 2, recordemos que como lo expusimos anteriormente, entre más aristas estén conectadas a un nodo, una mayor medida de centralidad tendrá lo que implica un valor de importancia en el grafo.

La figura 38 ya muestra un patrón con un nivel de complejidad más alto, en este tipo de patrones se puede evidenciar como el estudiante salta entre videos de for-

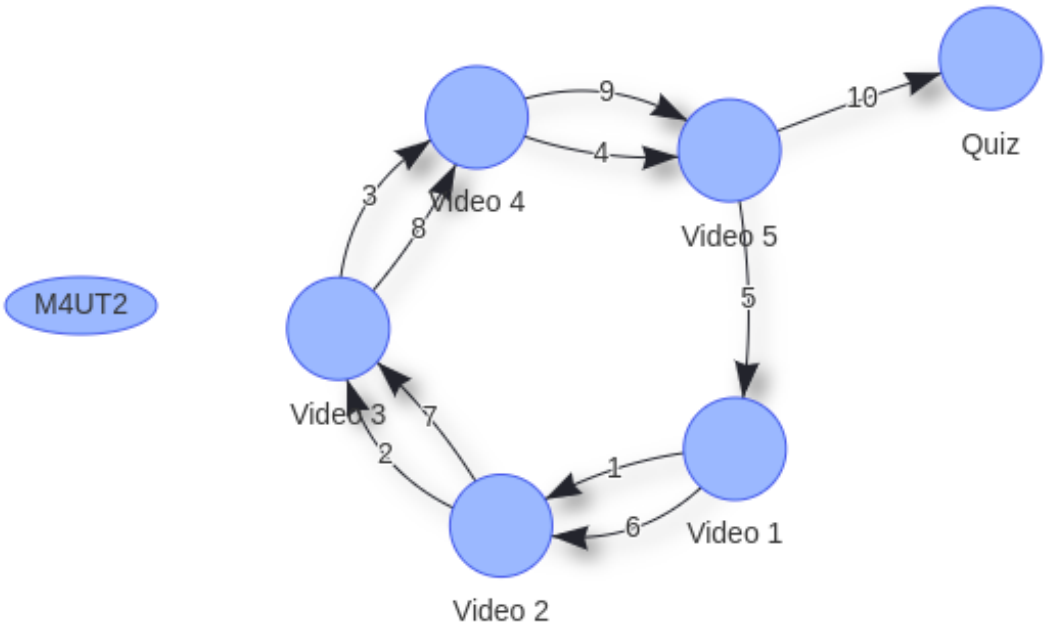


Figura 36: Navegación No Lineal Nodos Discontinuos 4

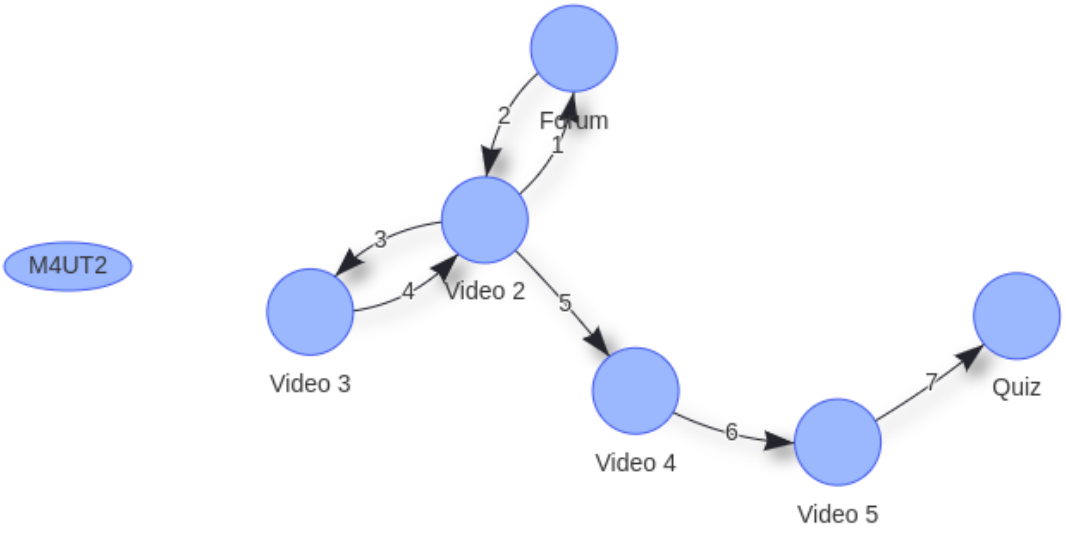


Figura 37: Navegación No Lineal múltiples lazos 1

ma discontinua múltiples veces. Entonces podríamos clasificar a estos patrones en el grupo de navegación lineal con nodos discontinuos con mas de 2 de lazos cerrados, y ahí estarán todos los estudiantes con un perfil de actividad alto y valdría la pena a futuro observar cuál es la relación entre este tipo de patrones con el desempeño aca-

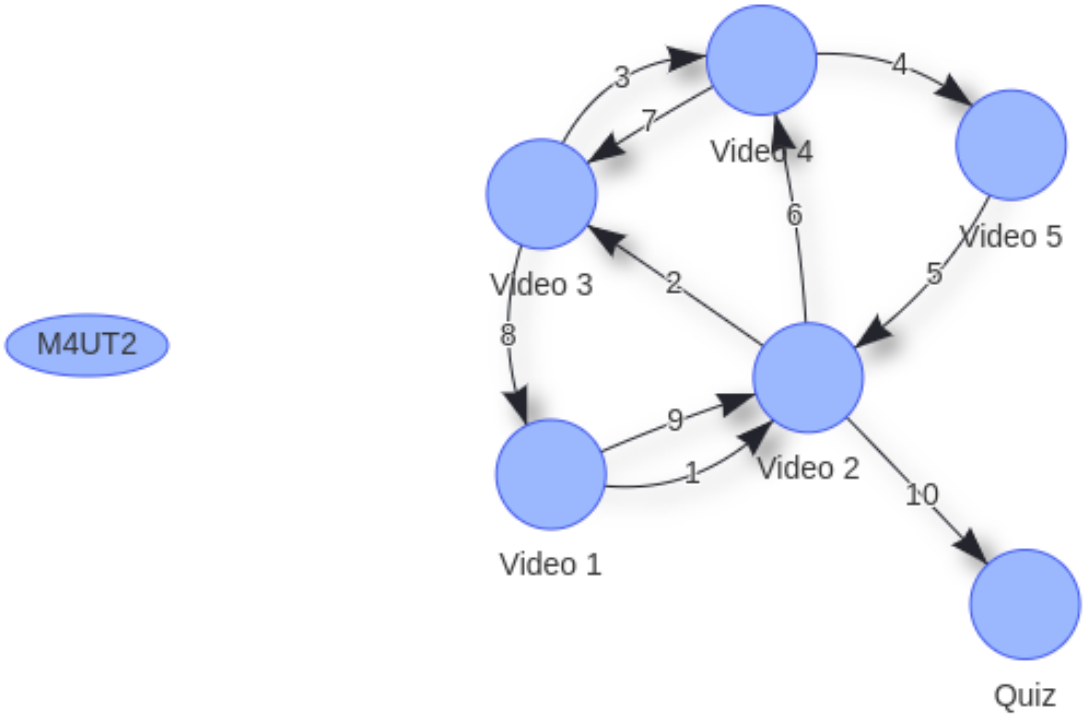


Figura 38: Navegación No Lineal múltiples lazos 2

démico obtenido al finalizar el período. Otros ejemplos de estos patrones se pueden visualizar en la figura 39, 40, 41 y 42

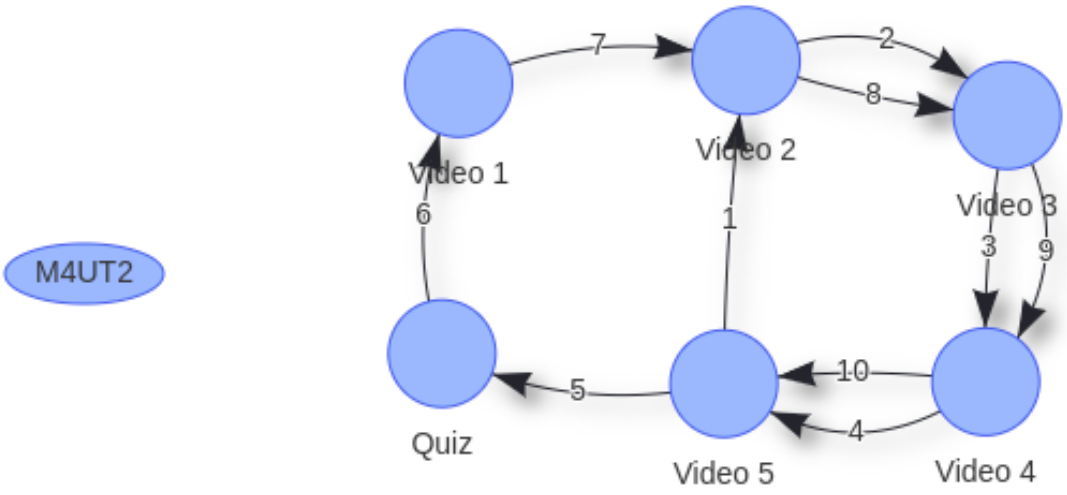


Figura 39: Navegación No Lineal múltiples lazos 3

Finalmente una última clasificación de los patrones encontrados son los grafos que no

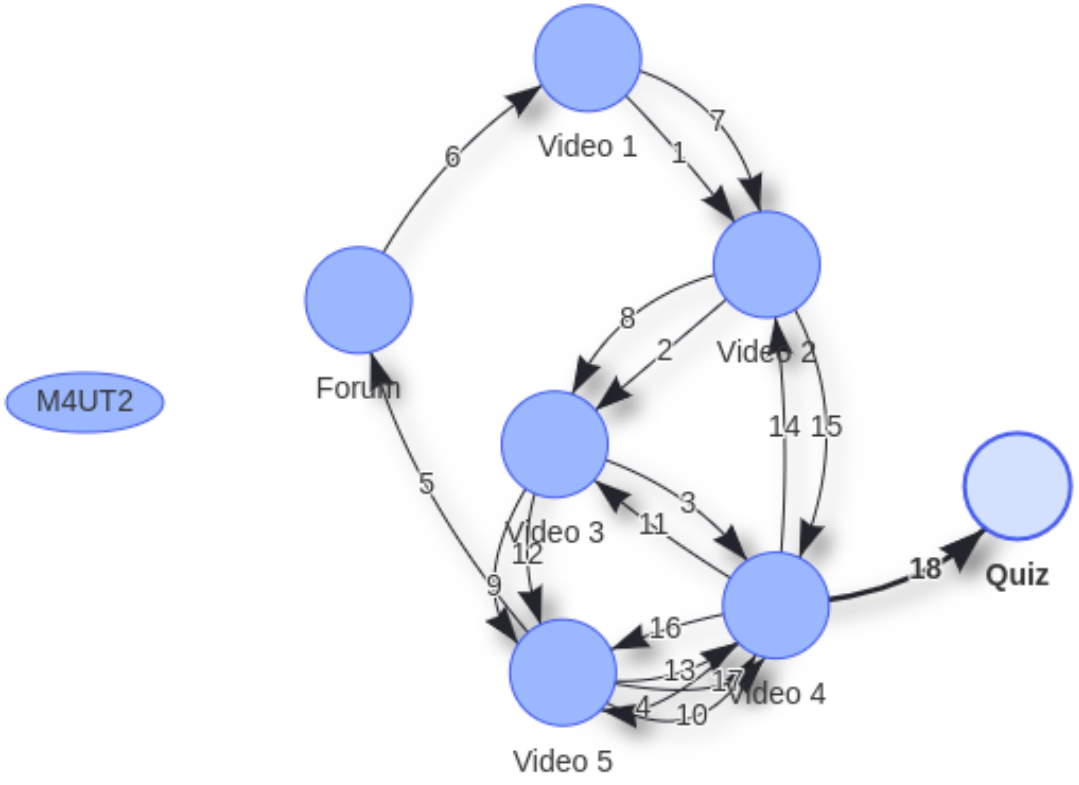


Figura 40: Navegación No Lineal múltiples lazos 4

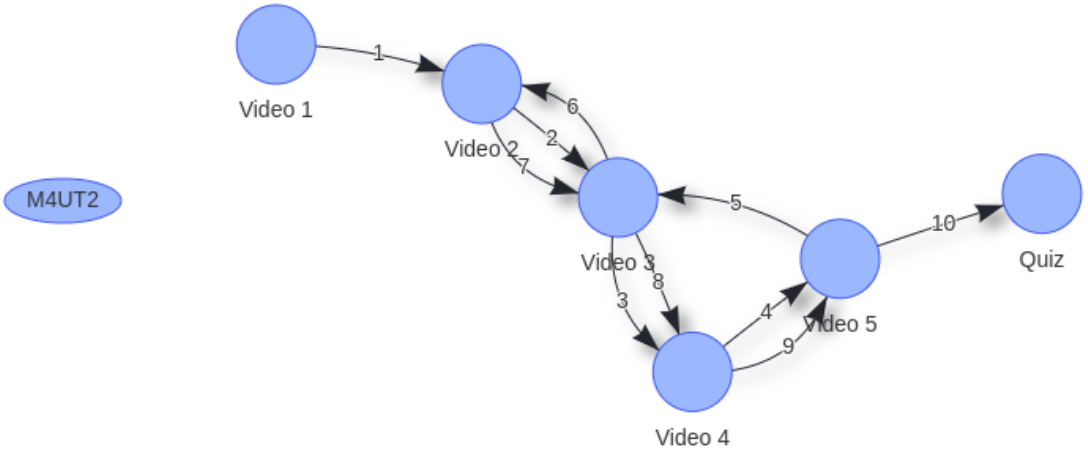


Figura 41: Navegación No Lineal múltiples lazos 5

poseen nodos equivalentes a evaluación, aquellos estudiantes que no presentaron el Quiz, por lo menos en el intervalo de tiempo especificado por el docente. La figura 43 y 44 muestran ejemplos de este grupo de grafos, ahí los estudiantes revisan vídeos y

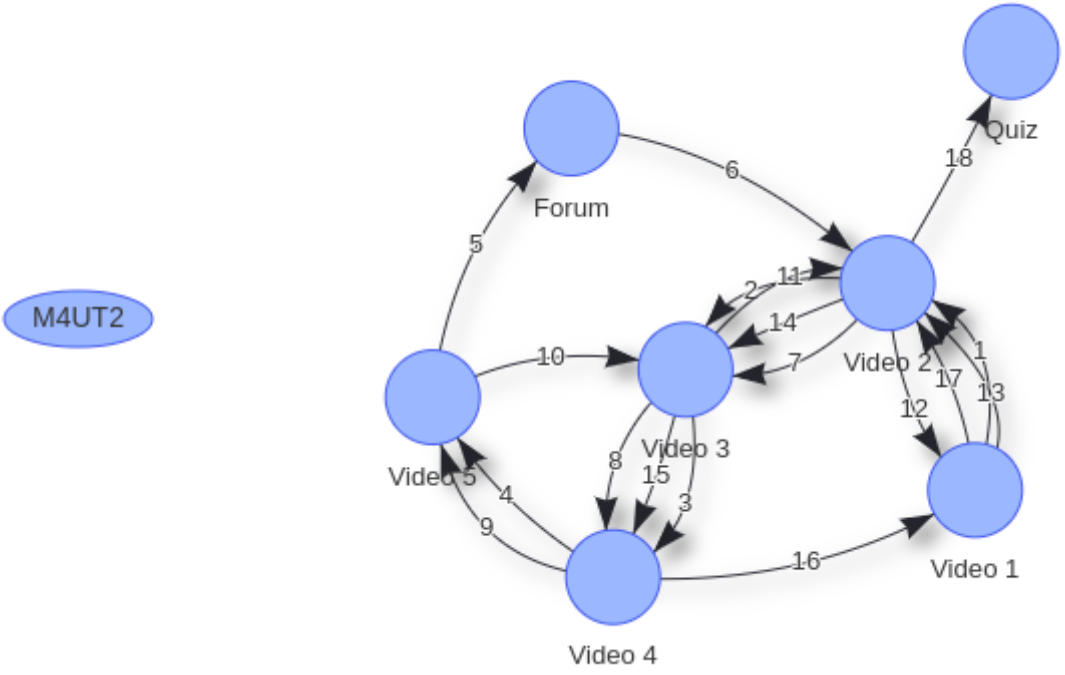


Figura 42: Navegación No Lineal múltiples lazos 6

hasta forman lazos cerrados pero no se ve evidenciada la realización de la evaluación.

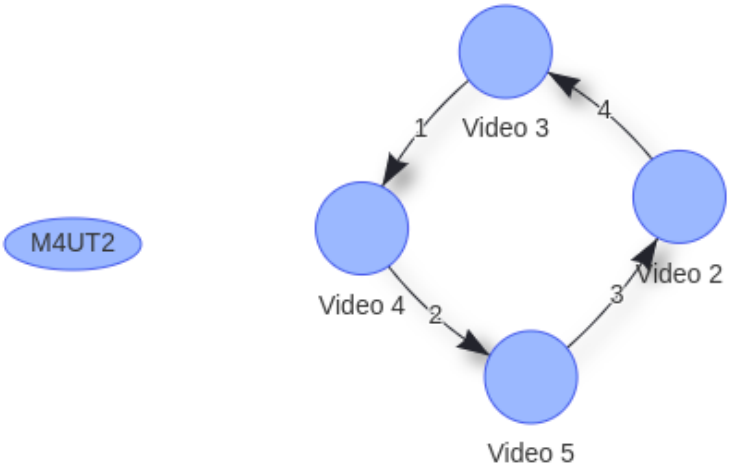


Figura 43: Navegación No Lineal Sin Quiz 1

La tabla 11 presenta un resumen estadístico de los patrones encontrados del período analizado, estos patrones están clasificados según la forma de navegación, lineal o no lineal, y la presencia o no de vídeos y nodos correspondientes a evaluaciones.En la

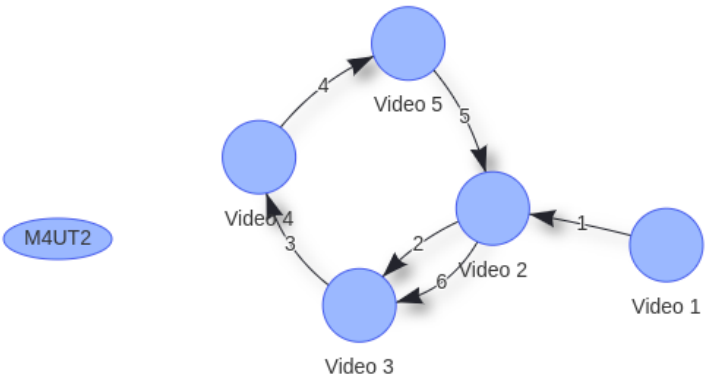


Figura 44: Navegación No Lineal Sin Quiz 2

tabla se puede evidencia que el 59,67 % de los grafos corresponden a navegación lineal y el 40,32 % a navegación no lineal. Además en el grupo de grafos de navegación no lineal, casi la mitad de los estudiantes tienen a lo sumo un ciclo cerrado mientras que la otra mitad tienen navegaciones con la presencia de 2 o más ciclos.

Tabla 11: Resumen estadísticas de patrones

Patrón	Vídeos	Sin vídeos	Sin Quiz	Sin Foro	Subtotal	Total
Navegación lineal	23			3	26	
Navegación lineal		5		1	6	37
Navegación lineal			2	3	5	
Navegación no lineal – 1 bucle (videos continuos)	4				4	
Navegación no lineal – 1 bucle (videos discontinuos)	5		2	1	8	25
Navegación no lineal – 2 o más bucles	8			5	13	
Totales	40	5	4	13		

8. Conclusiones y Trabajos futuros

Este trabajo tuvo como objetivo tener una aproximación del comportamiento de estudiantes de cursos en línea masivos. Esta tesis ha demostrado que los grafos y en general los cálculos que se pueden hacer gracias a su teoría matemática es una buena elección de representación de la información de las rutas de los estudiantes. Los grafos visualmente ayudan a un docente a interpretar mejor los movimientos de un estudiante por los contenidos del curso de este modo, pueden construirse grafos de diferentes tipo ya sea centrado en sesiones o módulos del curso, incluso caracterizando un grupo de estudiantes por una variable que tengan en común.

A través de un modelo de representación fue posible transformar eventos de click registradas en archivos de Tracking Log de las plataformas en una base de datos de grafos de estudiantes. Con los grafos generados se encontró similitud de la ruta de un estudiante con el patrón de diseño propuesto por el docente. La similitud de rutas entre los estudiante. Los contenidos más visitados y finalmente los saltos más comunes de cada módulo del curso. De igual forma, se propuso una forma de analizar a un grupo de estudiantes que tuvieran el mismo número de saltos en un período para compararlos y obtener un patrón característico de su comportamiento.

Con la implementación de la representación del comportamiento de los estudiantes, se llevó a nivel de aplicación la investigación. Realizamos un caso de estudio en la Universidad del Cauca con un curso en linea ofrecido como electiva de la cohorte 2019 con una muestra de 98 estudiantes. La plataforma utiliza es una instancia de Open edX. Fue posible representar el comportamiento de los estudiantes, aplicar los análisis estadísticos propuestos. Seguidamente, el docente del curso pudo interactuar con la herramienta de visualización y pudo explorar e identificar patrones de comportamiento. Con base a los resultados del caso de estudio se clasificaron los patrones encontrados y presentamos visualmente los ejemplos más significativos.

A futuro se propone mejorar el modelo de representación teniendo en cuanta que se pueden aprovechar muchas de las características de los perfiles para clasificar a los estudiantes y definir grafos adaptados a un grupo en especial. De igual forma, a nivel estadístico a futuro se propone ampliar la exploración de algoritmos que permitan proponer sistemas de recomendación para los docentes en el diseño del curso con base a los grafos encontrados de los estudiantes.

Los planes de futuro también incluyen que la herramienta de visualización se logre instalar y conectar a la instancia de Open edX Selene para que los docentes puedan a tiempo real revisar las secuencias que han estado haciendo sus estudiantes. Esto también abre otras posibilidades de trabajo futuro para evaluar cómo de efectiva es la herramienta, así como hacer casos de estudio con docentes utilizándola como parte de su diseño institucional. Otra posibilidad que se abre es que los estudiantes la puedan usar para fomentar comportamientos reflexivos que favorezcan la autoregulación del aprendizaje.

Teniendo en cuenta que la educación online, es cada vez más común en educación superior, los profesores necesitan de herramientas para poder monitorizar de forma más correcta cual es el avance de sus estudiantes con los contenidos. En este sentido, Moockly ejemplariza el potencial de las secuencias de acciones de los estudiantes para que los docentes puedan entender su proceso de aprendizaje.

Bibliografía

[1] L. Deslauriers, E. Schelew, and C. Wieman, "Improved Learning in a Large-Enrollment Physics Class," *Science*, vol. 332, no. 6031, pp. 862–864, 2011. [Online]. Available: http://science.sciencemag.org/content/332/6031/862

[2] S. Palmer, D. Holt, and S. Bray, "Does the discussion help? The impact of a formally assessed online discussion on final student results," *British Journal of Educational Technology*, vol. 39, no. 5, pp. 847–858, 2008. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8535.2007.00780.x

[3] D. Yang, M. Wen, A. Kumar, E. P. Xing, and C. P. Rosé, "Towards an integration of text and graph clustering methods as a lens for studying social interaction in MOOCs," *International Review of Research in Open and Distance Learning*, vol. 15, no. 5, pp. 215–234, 2014.

[4] R. F. Kizilcec, C. Piech, and E. Schneider, "Deconstructing Disengagement : Analyzing Learner Subpopulations in Massive Open Online Courses," *Lak '13*, p. 10, 2013.

[5] A. Anderson, D. Huttenlocher, J. Kleinberg, and J. Leskovec, "Engaging with Massive Online Courses," 2014. [Online]. Available: http://arxiv.org/abs/1403.3100

[6] N. Gillani and R. Eynon, "Communication patterns in massively open online courses," *Internet and Higher Education*, vol. 23, pp. 18–26, 2014. [Online]. Available: http://dx.doi.org/10.1016/j.iheduc.2014.05.004

[7] Z. Xu, D. Goldwasser, B. B. Bederson, and J. Lin, "Visual analytics of MOOCs at Maryland," in *1st ACM Conference on Learning at Scale, L@S 2014*. Com-

puter Science, Univ. of Maryland, United States: Association for Computing Machinery, 2014, pp. 195–196.

[8] N. N. Sonwalkar, "The First Adaptive MOOC: A Case Study on Pedagogy Framework and Scalable Cloud Architecture—Part I," *MOOCs FORUM*, vol. 1, no. P, pp. 22–29, 2016.

[9] J. J. Maldonado, R. Palta, J. Vázquez, J. L. Bermeo, M. Pérez-sanagustín, and J. Munoz-gama, "MOOCs Based on Self-Regulated Learning and Learning Styles," 2016.

[10] D. Cormier, B. Stewart, G. Siemens, and A. McAuley, "What is a mooc?" 2010.

[11] P. Hyman, "In the year of disruptive education," *Communications of the ACM*, vol. 55, no. 12, pp. 20–22, 2012.

[12] L. Pappano, "The Year of the MOOC," *The New York Times*, pp. 1–7, 2012. [Online]. Available: http://www.edinaschools.org/cms/lib07/MN01909547/Centricity/Domain/272/The Year of the MOOC NY Times.pdf

[13] D. Shah, "By the Numbers: MOOCS in 2016." [Online]. Available: https://www.classcentral.com/report/moocs-stats-and-trends-2021, year = 2021

[14] D. Jaramillo-Morillo, M. S. Sarasty, G. R. González, and M. Pérez-Sanagustín, "Follow-up of learning activities in open edx: A case study at the university of cauca," in *European Conference on Massive Open Online Courses*. Springer, 2017, pp. 217–222.

[15] P. H. Winston, "Artifical intelligence. 3-rd ed," 1992.

[16] N. Sonwalkar, "Adaptive Learning Technologies :," *EDUCAUSE,Center for Applied Research*, vol. 2005, no. 7, pp. 1–11, 2005. [Online]. Available: www.educause.edu/ecar/

[17] A. Cohen, U. Shimony, R. Nachmias, and T. Soffer, "Active learners' characterization in mooc forums and their generated knowledge," *British Journal of Educational Technology*, vol. 50, no. 1, pp. 177–198, 2019.

[18] J. Davies and M. Graff, "Student grades," *British Journal of Educational Technology*, vol. 36, no. 4, pp. 657–663, 2005.

[19] P. Riehmann, M. Hanfler, and B. Froehlich, "Interactive Sankey diagrams," in *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, 2005, pp. 233–240.

[20] S. Rizvi, B. Rienties, and S. A. Khoja, "The role of demographics in online learning; a decision tree based approach," *Computers Education*, vol. 137, pp. 32–47, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360131519300818

[21] I. E. Allen and J. Seaman, "Changing course: ten years of tracking online education in the United States," *Nursing standard (Royal College of Nursing (Great Britain) : 1987)*, vol. 26, p. 47, 2013. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/23494723

[22] T. Anderson and R. McGreal, "Disruptive Pedagogies and Technologies in Universities," *Journal of Educational Technology & Society*, vol. 15, no. 4, pp. 380–389, 2012. [Online]. Available: http://www.jstor.org/stable/jeductechsoci.15.4.380

[23] B. Stephen, "Back to the Future with MOOCs," *Icicte*, vol. 77, no. 11, pp. 237–246, 2013. [Online]. Available: http://www.icicte.org/Proceedings2013/Papers%5Cn2013/06-3-Brown.pdf

[24] S. Porter, *To MOOC or Not to MOOC: how can online learning help to build the future of higher education?* Chandos Publishing, 2015.

[25] G. Chen, D. Davis, C. Hauff, and G.-J. Houben, "On the Impact of Personality in Massive Open Online Learning," pp. 121–130, 2016.

[26] P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. D. Kloos, "Sentiment analysis in moocs: A case study," in *2018 IEEE Global Engineering Education Conference (EDUCON)*, April 2018, pp. 1489–1496.

[27] S. J. Daniel, E. V. Cano, and M. G. Cervera, "Art:10.7238/Rusc.V12I1.2475," no. January, 2015.

[28] B. Dietz-Uhler and J. E. Hurn, "Using learning analytics to predict (and improve) student success: A faculty perspective," *Journal of interactive online learning*, vol. 12, no. 1, pp. 17–26, 2013.

[29] D. F. O. Onah and J. Sinclair, "Massive Open Online Courses – an Adaptive Learning Framework," *The University of Warwick*, vol. INTED2015, no. March, pp. 1258–1266, 2015.

[30] A. Teixeira, A. Garcia-Cabot, E. Garcia-Lopez, J. Mota, and L. De-Marcos, "a New Competence-Based Approach for Personalizing Moocs in a Mobile Collaborative and Networked Environment," *RIED. Revista Iberoamericana de Educación a Distancia*, vol. 19, no. 1, pp. 143–160, 2015. [Online]. Available: http://revistas.uned.es/index.php/ried/article/view/14578

[31] K. Thulasiraman and M. N. Swamy, *Graphs: theory and algorithms.* John Wiley & Sons, 2011.

[32] M. A. Rodriguez and P. Neubauer, "The graph traversal pattern," 2010.

[33] R. King, "Semantic Database Modeling: Survey, Applications, and Research," *Computing*, vol. 19, no. 3, 1987.

[34] A. Silvescu, D. Caragea, and A. Atramentov, "Abstract Graph Databases."

[35] A. L. Montgomery, "to the Internet," *Interfaces*, pp. 90–108, 2001.

[36] J. B. Kruskal and J. M. Landwehr, "Icicle Plots: Better Displays for Hierarchical Clustering," *The American Statistician*, vol. 37, no. 2, pp. 162–168, 1983. [Online]. Available: https://amstat.tandfonline.com/doi/abs/10.1080/00031305.1983.10482733

[37] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman, "Temporal event sequence simplification," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2227–2236, 2013.

[38] J. Zhao, A. Wilson, Z. Liu, A. Hertzmann, and M. Dontcheva, "MatrixWave," pp. 259–268, 2015.

[39] F. van Ham, H. van de Wetering, and J. J. van Wijk, "Interactive visualization of state transition systems," *IEEE Trans. Vis. Comput. Graph.*, vol. 8, pp. 319–329, 2002.

[40] E. Maguire, P. Rocca-Serra, S. A. Sansone, J. Davies, and M. Chen, "Visual compression of workflow visualizations with automated detection of macro motifs," *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2576–2585, 2013.

[41] A. Perer and F. Wang, "Frequence: Interactive Mining and Visualization of Temporal Frequent Event Sequences," *Proceedings of the 19th International Conference on Intelligent User Interfaces*, pp. 153–162, 2014. [Online]. Available: http://dl.acm.org/citation.cfm?id=2557500.2557508

[42] K. Wongsuphasawat and D. Gotz, "Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2659–2668, 2012.

[43] J. F. Burnham, "Scopus database: a review," *Biomedical digital libraries*, vol. 3, no. 1, p. 1, 2006.

[44] T. Reuters, "Web of science." 2012.

[45] S. Jiang, A. E. Williams, K. Schenke, M. Warschauer, and D. O. Dowd, "Predicting MOOC Performance with Week 1 Behavior," *Proceedings of the 7th International Conference on Educational Data Mining (EDM)*, pp. 273–275, 2014.

[46] J. Bravo-Agapito, S. J. Romero, and S. Pamplona, "Early prediction of undergraduate student's academic performance in completely online learning: A five-year study," *Computers in Human Behavior*, vol. 115, p. 106595, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563220303423

[47] G. Balakrishnan, "Predicting student retention in massive open online courses using hidden markov models," Master's thesis, EECS Department, University of California, Berkeley, May 2013. [Online]. Available: http://www2.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.html

[48] A. F. Gkontzis, S. Kotsiantis, C. T. Panagiotakopoulos, and V. S. Verykios, "A predictive analytics framework as a countermeasure for attrition of students," *Interactive Learning Environments*, vol. 30, no. 6, pp. 1028–1043, 2022. [Online]. Available: https://doi.org/10.1080/10494820.2019.1709209

[49] P. I. N. Assessment, "Breslow et al - edx's first mooc (1)," no. March 2012, pp. 13–25, 2013.

[50] S. Vonderwell and S. Zachariah, "Factors that influence participation in online learning," *Journal of Research on Technology in Education*, vol. 38, no. 2, pp. 213–230, 2005.

[51] Y. Goda and C. Hayward, "Relationship between learning time in an online course and learning behavior and outcomes," in *EdMedia+ Innovate Learning*. Association for the Advancement of Computing in Education (AACE), 2022, pp. 917–925.

[52] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson, "Patterns and Sequences: Interactive Exploration of Clickstreams to Understand Common Visitor Paths," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 1, pp. 321–330, 2017.

[53] H. Jeong and G. Biswas, "Mining student behavior models in learning-by-teaching environments," in *Educational data mining 2008*, 2008.

[54] M. Köck and A. Paramythis, "Activity sequence modelling and dynamic clustering for personalized e-learning," *User Modeling and User-Adapted Interaction*, vol. 21, no. 1-2, pp. 51–97, 2011.

[55] M. Kock and A. Paramythis, "Towards adaptive learning support on the basis of behavioural patterns in learning activity sequences," in *2010 International Conference on Intelligent Networking and Collaborative Systems*. IEEE, 2010, pp. 100–107.

[56] D. H. Shanabrook, D. G. Cooper, B. P. Woolf, and I. Arroyo, "Identifying high-level student behavior using sequence-based motif discovery," in *Educational Data Mining 2010*, 2010.

[57] P. J. Guo and K. Reinecke, "Demographic differences in how students navigate through moocs," in *Proceedings of the first ACM conference on Learning@ scale conference*, 2014, pp. 21–30.

[58] P. Mukala, J. Buijs, M. Leemans, and W. Van Der Aalst, "Learning analytics on coursera event data: A proceb mining approach," in *5th International Symposium on Data-Driven Process Discovery and Analysis, SIMPDA 2015*, R.-M. S. and C. P., Eds., vol. 1527. Department of Mathematics and Computer Science, Eindhoven University of Technology, Eindhoven, Netherlands: CEUR-WS, 2015, pp. 18–32.

[59] M. Wen and C. P. Rosé, "Identifying latent study habits by mining learner behavior patterns in massive open online courses," in *Proceedings of the 23rd ACM international conference on conference on information and knowledge management*, 2014, pp. 1983–1986.

[60] M. A. Mercado-Varela, A. García-Holgado, F. J. García-Peñalvo, and M. S. Ramírez-Montoya, "Analyzing navigation logs in MOOC: A case study," in *4th International Conference on Technological Ecosystem for Enhancing Multiculturality, TEEM 2016*, G.-P. F.J., Ed., vol. 02-04-Nove, Instituto de Investigación Y Desarrollo Educativo, Universidad Autónoma de Baja California, Ensenada, Mexico, 2016, pp. 873–880.

[61] C. G. Brinton, S. Buccapatnam, M. Chiang, and H. V. Poor, "Mining MOOC Clickstreams: Video-Watching Behavior vs. In-Video Quiz Performance," *IEEE Transactions on Signal Processing*, vol. 64, no. 14, pp. 3677–3692, 2016.

[62] S. Ardchir, M. A. Talhaoui, and M. Azzouazi, "Towards an adaptive learning framework for MOOCs," Faculty of Sciences Ben M'Sik, Hassan II University of Casablanca, Casablanca, Morocco, pp. 236–251, 2017.

[63] D. Davis, G. Chen, C. Hauff, and G. J. Houben, "Gauging MOOC learners' adherence to the designed learning path," *Proceedings of the 9th International Conference on Educational Data Mining, EDM 2016*, pp. 54–61, 2016.

[64] T. Sinha, P. Jermann, N. Li, and P. Dillenbourg, "Your click decides your fate: Inferring Information Processing and Attrition Behavior from MOOC Video Clickstream Interactions," 2014. [Online]. Available: http://arxiv.org/abs/1407.7131

[65] F. Anacona, "Descubrimiento de patrones de navegación en open edx – una aproximación arquitectónica."

[66] J. Maldonado-Mahauad, M. Pérez-Sanagustín, P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. Delgado-Kloos, "Predicting Learners' Success in a Self-paced MOOC Through Sequence Patterns of Self-regulated Learning," in *Lifelong Technology-Enhanced Learning*, V. Pammer-Schindler, M. Pérez-Sanagustín, H. Drachsler, R. Elferink, and M. Scheffel, Eds. Cham: Springer International Publishing, 2018, pp. 355–369.

[67] R. F. Kizilcec, J. Munoz-Gama, J. Maldonado-Mahauad, N. Morales, and M. Pérez-Sanagustín, "Mining theory-based patterns from Big data: Identifying self-regulated learning strategies in Massive Open Online Courses," *Computers in Human Behavior*, vol. 80, pp. 179–196, 2018.

[68] D. Yildirim and Y. Usluel, "Interrelated analysis of interaction, sequential patterns and academic achievement in online learning," *Australasian Journal of Educational Technology*, vol. 38, no. 2, p. 181–200, Apr. 2022. [Online]. Available: https://ajet.org.au/index.php/AJET/article/view/7360

[69] A. Çebi, R. D. Araújo, and P. Brusilovsky, "Do individual characteristics affect online learning behaviors? an analysis of learners sequential patterns," *Journal of Research on Technology in Education*, vol. 0, no. 0, pp. 1–21, 2022. [Online]. Available: https://doi.org/10.1080/15391523.2022.2027301

[70] J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, D. Leony, and C. Delgado Kloos, "ALAS-KA: A learning analytics extension for better understanding the lear-

ning process in the Khan Academy platform," *Computers in Human Behavior*, vol. 47, pp. 139–148, 2015.

[71] D. Leony, A. Pardo, L. de la Fuente Valentín, D. S. de Castro, and C. D. Kloos, "Glass: a learning analytics visualization tool," in *Proceedings of the 2nd international conference on learning analytics and knowledge*, 2012, pp. 162–163.

[72] A. A. Mubarak, S. A. Ahmed, and H. Cao, "Mooc-asv: Analytical statistical visual model of learners' interaction in videos of mooc courses," *Interactive Learning Environments*, pp. 1–16, 2021.

[73] S. García-Molina, C. Alario-Hoyos, P. M. Moreno-Marcos, P. J. Muñoz-Merino, I. Estévez-Ayres, and C. Delgado Kloos, "An algorithm and a tool for the automatic grading of mooc learners from their contributions in the discussion forum," *Applied Sciences*, vol. 11, no. 1, p. 95, 2020.

[74] S. Liu, S. Liu, Z. Liu, X. Peng, and Z. Yang, "Automated detection of emotional and cognitive engagement in mooc discussions to predict learning achievement," *Computers & Education*, vol. 181, p. 104461, 2022.

[75] T. Rohloff, D. Sauer, and C. Meinel, "Student perception of a learner dashboard in moocs to encourage self-regulated learning," in *2019 IEEE international conference on engineering, technology and education (TALE)*. IEEE, 2019, pp. 1–8.

[76] J. A. Ruipérez-Valiente, P. J. Muñoz-Merino, J. A. Gascón-Pinedo, and C. D. Kloos, "Scaling to massiveness with analyse: A learning analytics tool for open edx," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 6, pp. 909–914, 2017.

[77] A. Dipace, B. Fazlagic, and T. Minerva, "The design of a learning analytics dashboard: Eduopen mooc platform redefinition procedures," *Journal of E-Learning and Knowledge Society*, vol. 15, no. 3, pp. 29–47, 2019.

[78] O. Standars and PMI, *PMBOK GUIDE Sexth Edition*, 2018. [Online]. Available: https://www.pmi.org.pe/pmbok6/

[79] J. Maldonado-Mahauad, M. Pérez-Sanagustín, P. M. Moreno-Marcos, C. Alario-Hoyos, P. J. Muñoz-Merino, and C. Delgado-Kloos, "Predicting learners' success in a self-paced mooc through sequence patterns of self-regulated learning," in *European conference on technology enhanced learning*. Springer, 2018, pp. 355–369.

[80] S. P. Borgatti and M. G. Everett, "A graph-theoretic perspective on centrality," *Social networks*, vol. 28, no. 4, pp. 466–484, 2006.

[81] L. Yujian and L. Bo, "A normalized levenshtein distance metric," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 1091–1095, 2007.