Universidad de Buenos Aires

Facultad de Ciencias Exactas y Naturales

Maestría en Explotación de Datos y Descubrimiento del Conocimiento

Tesis de Maestría

Comunidades en los cursos virtuales

Análisis de grafos y distinción de usuarios

Autor: Esp. Ivo Rusconi

Director: Dr. Marcelo Soria

2018

ÍNDICE

AGRADECIMIENTOS

Debo agradecer a todas aquellas personas que han aportado tiempo e ideas a la confección del presente trabajo.

En primer lugar, agradezco la tutoría del Profesor de Data Mining en Ciencia y Tecnología y actual Director de la Maestría, Dr. Marcelo Soria, quien mediante las correcciones efectuadas y consiguientes sugerencias en la confección del desarrollo, ha realizado aportes significativos para la realización del trabajo.

Agradezco además la participación de Mariela Verónica Velasco, quien colaboró en materia de confección y cumplimientos formales para la presentación del trabajo, leyendo los numerosos borradores efectuados hasta obtener la versión definitiva.

Agradezco a Dawoon Choi por haberme acompañado en los dos arduos años de cursada, por haber compartido gran parte de tiempo, trabajo y esfuerzo, sin el cual posiblemente no hubiera llegado hoy a este punto.

Asimismo, agradezco al Instituto Nacional del Cáncer, quien me proveyó desinteresadamente de las bases de datos con buena parte de su información, sin las cuales este trabajo no hubiera sido posible.

Finalmente, debo agradecer a la Universidad de Buenos Aires, por brindarme la oportunidad de seguir capacitándome y formándome dentro del ámbito de la educación pública y nacional.

INTRODUCCIÓN

ANÁLISIS DEL PROBLEMA

En el presente trabajo se realiza un análisis sobre distintos cursos virtuales brindados por el Instituto Nacional del Cáncer (INC), gestionados a través de la plataforma Moodle (https://moodle.org). En dicha plataforma existe información de variada naturaleza (cursos, exámenes, foros) que será objeto de análisis.

La participación supera los cinco mil alumnos y una decena de docentes presentes en aproximadamente cuarenta cursos dictados entre marzo de 2016 y julio de 2017.

Dentro de todo curso, sea cual fuere el ámbito de aplicación o grado de este, existe una realidad latente y es que no todos los alumnos se gradúan del mismo. Existen distintas razones que justifican este comportamiento. Muchas veces dependerá de la situación personal que atraviesa cada individuo durante el transcurso de la cursada, sin embargo, también influye seguramente la dinámica del curso, del docente a cargo y de la currícula en general que darán una mayor o menor motivación al alumnado para finalizar la cursada y, eventualmente, graduarse o certificarse.

MARCO TEÓRICO

El presente trabajo involucra el análisis del dictado de cursos en forma virtual, siendo dichos análisis realizados a partir de la modelización de los usuarios y sus interacciones en forma de grafos.

Mientras en [17] se hacen análisis de grafos de usuarios de Facebook, una plataforma extremadamente más grande que la que se presentará en este trabajo, existe cierta cantidad de literatura dedicada al análisis de cursos tanto presenciales como online. Por ejemplo, en [25] se compara la efectividad de cursos de tipo presencial y online, analizando las características de uno y otro tipo, así como los factores que pueden afectar diferencialmente a cada uno. Estos trabajos, al igual que algunos otros que serán presentados a continuación, sirven tanto de marco teórico para la tesis actual como de fundamento de la necesidad de ésta. Todos este material, al igual que otro que fue de utilidad para el presente trabajo y no será explícitamente citado durante la presente tesis, puede ser consultado en el apartado de “Bibliografía”.

Instituto Nacional de Cáncer

La página web oficial del INC (https://www.argentina.gob.ar/salud/inc) define en forma detallada su historia y objetivo:

El Instituto Nacional del Cáncer de la Argentina (INC) es un ente dependiente del Ministerio de Salud de la Nación.

Creado el 9 de septiembre de 2010 por Decreto Presidencial 1286, es el responsable del desarrollo e implementación de políticas de salud, así como de la coordinación de acciones integradas para la prevención y control del cáncer.

Su principal objetivo es disminuir la incidencia y mortalidad por cáncer en Argentina, a la vez que mejorar la calidad de vida de las personas afectadas.

Entre sus funciones, el INC es el encargado de coordinar acciones de promoción y prevención, detección temprana, tratamiento y rehabilitación, así como también la investigación del cáncer en Argentina y la formación de recursos humanos. Sus actividades incluyen el desarrollo de normativas para la asistencia integral de los pacientes con cáncer; la promoción de la salud y reducción de los factores de riesgo; la definición de estrategias para la prevención y detección temprana; la formación de profesionales especializados y el establecimiento de un sistema de vigilancia y análisis epidemiológico.

La creación del INC en Argentina significó colocar al cáncer en un lugar de suma relevancia dentro de la agenda sanitaria de gobierno. Su existencia implica la coordinación de políticas públicas específicas a nivel nacional y el trabajo en consonancia con las instituciones pares de la región y del mundo, mediante el intercambio de experiencias, capacidades y esfuerzos.

En el mismo sitio se señalan su visión, misión y algunos de sus objetivos específicos, los cuales dan utilidad al presente trabajo, buscando mejorar el impacto de las actividades de dicho Instituto.

Visión

Ser el organismo referente en materia del cáncer en la República Argentina, dedicado a coordinar la prevención, detección temprana, tratamiento y rehabilitación y a promover la formación de recursos humanos y la investigación en cáncer.

Misión

El INC es el responsable del desarrollo e implementación de políticas de salud, así como de la coordinación de acciones integradas para la prevención y el control del cáncer en Argentina.

Objetivo general

Contribuir a la disminución de la incidencia y la morbi mortalidad por cáncer en la Argentina.

Objetivos específicos

- *Promover la inclusión de programas de prevención y control de cáncer en las diferentes jurisdicciones y fortalecer los recursos necesarios para su implementación.*
- *Propiciar un marco a través del cual se facilite y promueva la producción científica en materia oncológica.*
- *Articular esfuerzos frente al cáncer en el marco internacional, mediante el intercambio de experiencias, capacidades y esfuerzos destinados al control del cáncer.*

- *Promover la formación de recursos humanos especializados.*
- *Incentivar la investigación en cáncer.*

Explotación de datos

Explotación de datos y descubrimiento de conocimiento

Es menester hacer un breve resumen del proceso de explotación de datos y descubrimiento de conocimiento, objeto principal y motivo de la presente maestría. A propósito de esto, se citan a continuación algunos pasajes relevantes extraídos de [9].

> *"El data mining puede ser visto como el resultado de la evolución natural de la tecnología de la información."*

"La rápidamente creciente cantidad de datos, recopilados y almacenados en grandes y numerosos repositorios, ha superado por lejos nuestra capacidad humana de comprensión sin el uso de herramientas eficaces."

"Mucha gente usa explotación de datos como sinónimo de otro término popularme usado, descubrimiento de conocimiento o KDD (por sus siglas en inglés, Knowledge Discovery from Data), mientras que otros ven a la explotación de datos simplemente como un paso esencial en el proceso de descubrimiento de conocimiento. El proceso de descubrimiento de conocimiento es mostrado como una secuencia iterativa de los siguientes pasos:

1. *Data cleaning (Limpieza de datos: remover ruido e inconsistencias en los datos)*
2. *Data integration (Integración de datos: donde multiples fuentes de datos son combinadas)*
3. *Data selection (Selección de datos: donde los datos relevantes a la tarea de análisis son extraidos de las bases de datos)*
4. *Data transformation (Transformación de datos: donde los datos son transformados y consolidados en formas apropiadas para la explotación realizando operaciones de sumarización o agregación)*
5. *Data mining (Explotación de datos: un proceso esencial donde se aplican métodos inteligentes para extraer patrones de los datos)*
6. *Pattern evaluation (Evaluación de patrones: identificar patrones realmente interesantes que representen conocimiento basado en medidas de interés)*
7. *Knowledge presentation (Presentación de conocimiento: donde la visualización y representación de técnicas de conocimiento son usadas para presentar el conocimiento obtenido a los usuarios)"*

"Explotación de datos es el proceso de descubrir patrones interesantes y conocimiento a partir de grandes volúmenes de datos."

El presente trabajo se vale de diferentes técnicas que pueden englobarse dentro del proceso de descubrimiento de conocimiento, y puntualmente dentro de la explotación de datos, para obtener información de utilidad. Los pasos mencionados previamente, en distinta medida, abarcan la totalidad del presente trabajo.

Explotación de datos en plataformas de e-learning

Existen numerosos antecedentes de trabajos de explotación de datos asociados a plataformas de e-learning (aprendizaje en contextos virtuales) e incluso utilizando la plataforma Moodle. A continuación, serán citados aquellos trabajos que sirvieron tanto de puntapié como de inspiración para el desarrollo de la tesis actual. Los mismos son relevantes no sólo por estar relacionados con plataformas de e-learning, sino también por estar relacionados, en varios casos, a ámbitos de aplicación similares a los abarcados en la presente tesis de maestría.

Posiblemente uno de los trabajos más relevantes en este contexto y con relación al presente trabajo es [15]. En dicho paper se menciona la importancia de realizar análisis de redes sociales para extraer datos estructurales con aplicabilidad a sistemas educativos. El mismo se centra en analizar estas redes sociales dadas por la interacción de usuarios en un ambiente de e-learning.

Por otro lado, trabajos como [6] son de importancia al estar relacionados a un área de aplicación relevante para la presente tesis. El mismo detalla el uso de equipos particulares de tecnología de la información en la salud, comparando dos métodos de entrenamiento de trabajadores de la salud: uno presencial con uno mixto.

Dentro de los sistemas de e-learning existe una variedad diferente llamada MOOC (por sus siglas en inglés) que también es objeto de análisis en papers como [19]. Los cursos MOOC son, por definición, de tipo online abiertos y masivos. Los mismos tienen un menor seguimiento y una dinámica diferente respecto a los brindados por el INC, que sigue un plazo y una currícula definida en forma asimilable a un curso presencial. En dicho paper se definen las principales características y motivos para que un Instituto decida emprender este tipo de cursos, así como para que un estudiante los elija. Estos cursos no son comparables en su funcionamiento y en sus características a otros cursos de e-learning, pero su comprensión es útil al momento de hacer análisis y definir particularidades.

Como últimos ejemplos, es importante señalar los trabajos [11] y [5], ambos dedicados al análisis de información presente específicamente dentro de la plataforma Moodle. En [11] se hace una revisión de herramientas para explotación de datos de actividades educativas, y se propone una nueva herramienta desarrollada por sus autores para realizar análisis sobre Moodle. En forma complementaria, en [5] se hace una revisión de distintos tipos de herramientas para analizar datos de Moodle y se comentan casos de uso de cada uno, con sus ventajas y desventajas. Una de las herramientas más significativas mencionadas en [5] es GraphFES, servicio web que conecta los archivos con sesiones de Moodle y extrae información de todos los tableros de mensajes de cada curso. Toda esta información es luego procesada en distintos grafos que pueden ser analizados utilizando herramientas como Gephi. Esta metodología, enfoque de trabajo y herramienta (Gephi) fueron de utilidad para el desarrollo de la tesis actual.

DEFINICIÓN DE OBJETIVOS Y JUSTIFICACIÓN DEL TRABAJO

El objetivo principal del trabajo es comprender las diferencias que existen entre aquellos alumnos que logran finalizar su curso y obtener el correspondiente certificado de aprobación, y aquellos que no lo logran. Se analizarán las redes de contactos de usuarios, buscando entender cuáles son las principales características que diferencian usuarios con distinta cantidad de contactos.

Para ello, se analizará el comportamiento de cada grupo de usuarios, buscando entender sus diferencias subyacentes con la finalidad de aumentar la tasa de alumnos certificados. Se entiende que es deseable que la proporción de alumnos certificados sea lo más cercana posible al total de alumnos inscriptos por curso. Será parte del trabajo el identificar características de los usuarios y diferenciar así grupos con rasgos similares dentro de cada grupo y diferentes del resto de los grupos. Para esto se usarán diferentes métricas y modelos específicos para el análisis de grafos, de los usuarios que los componen y las relaciones que existen entre ellos.

Una vez comprendidas las diferencias entre grafos, analizando en forma diferencial sus características como pueden ser su rol o el poseer un certificado de aprobación, se buscarán modelar las circunstancias que llevan a un usuario a poseer una determinada red de contactos. Se parte de la hipótesis de que la sociabilidad de cada usuario está íntimamente relacionada con sus resultados académicos. Se buscarán armar modelos capaces de predecir la probabilidad de poseer una determinada cantidad de contactos en base a sus características y comportamientos, con la finalidad de realizar acciones tendientes a apuntalar aquellos que poseen una baja probabilidad de certificarse, así como buscar el mayor desempeño posible de los que poseen una alta probabilidad de hacerlo.

Este trabajo es necesario por cuanto todo curso con vocación profesional busca no sólo brindar conocimiento al alumnado, sino otorgarle un certificado que avale dicho conocimiento y mejore por tanto su práctica. En este sentido, es posible mencionar la evidencia experimental existente de cómo una mayor y mejor gestión de las plataformas online pueden aumentar el rendimiento de los alumnos. En [4] se identifica que la mayor participación en foros virtuales contribuye con un efecto facilitador hacia la materia, asociado a mejores resultados a lo largo del curso, incluso si el docente invierte una cantidad mínima de su tiempo en la participación dentro del foro.

Siendo el INC una institución pública dependiente del Ministerio de Salud de la Nación, el objetivo previamente planteado cobra mayor relevancia. Como referente a nivel nacional de la investigación y lucha contra el cáncer, será objeto de estudio y mejora la calidad de los cursos que brinde. Esta mejora en la calidad deberá redundar por un lado en la formación de profesionales cada vez más capaces, pero a su vez también – y en este sentido se concentrará el presente trabajo – en el aumento de la tasa de alumnos certificados por curso. Entendiendo que una mayor cantidad de alumnos con certificado redundará en un beneficio no sólo para el propio alumnado, sino para la sociedad en su conjunto.

Los resultados y conclusiones a las que se arriben deberán servir para mejorar el funcionamiento de los cursos y del Instituto en su totalidad. El mayor conocimiento de las interacciones de sus alumnos, su participación y los resultados obtenidos deben ser

de utilidad para reconocer los procesos y tareas que se realizan correctamente, así como realizar cambios o mejoras en aquellas que presenten oportunidades. La determinación de oportunidades de mejora dentro del dictado de los cursos escapa al presente trabajo, sin embargo, es deseable que sirva como disparador para el análisis de alumnos, cursos y certificaciones.

Este trabajo debe funcionar como puntapié inicial de un proceso de mejora constante dentro del área académica del Instituto, que sirva para aggiornar tanto los cursos virtuales que se dictan actualmente, como la totalidad de la oferta de formación que está disponible.

MATERIALES Y MÉTODOS

MATERIALES

El INC proveyó, en forma voluntaria, la base de datos que permitió la realización del presente trabajo. La misma fue una extracción hecha de la plataforma Moodle (https://moodle.org), en donde el INC posee toda la información relativa a los cursos virtuales que brinda. Esta información incluye detalle de los alumnos, personal docente y no docente, cursos con exámenes y resultados, certificaciones, foros y participación en los mismos, así como distintos datos de logueos y tráfico en el sitio. Por razones de seguridad y confidencialidad toda la información relacionada a los usuarios fue anonimizada, eliminando todo tipo de información personal por fuera de lo relacionado estrictamente con los cursos y la participación en la plataforma, resultando así imposible la identificación de los usuarios.

El total de la base de datos ocupa 6,2 gigabytes en disco. El modelo de datos completo consta de 336 entidades, si bien tan solo 149 de éstas poseen al menos un registro insertado. Los datos constan de aproximadamente cuarenta cursos dictados entre marzo de 2016 y julio de 2017, en los cuales existen más de cinco mil alumnos, una decena de docentes y personal no docente (administradores de la plataforma).

La extracción de la base de datos a la que se tuvo acceso se encontraba en un archivo .sql que fue extraído utilizando MySQL Workbench 6.3 CE (https://mysql.com/products/workbench/). Las consultas SQL fueron realizadas con el mismo software, esto incluye no sólo el armado de los datatsets que serán utilizados luego para realizar análisis, sino también la obtención de ciertas métricas básicas, más generales y descriptivas.

Como primer paso se analizó cada una de las tablas, se separó aquellas que no poseían ningún registro, mientras que se procedió a analizar el resto en forma manual de forma tal de comprender la utilidad de estas. Fue necesario realizar un análisis exploratorio con el fin de comprender la definición y utilidad de cada tabla y variable. Se realizaron consultas básicas en SQL de modo de visualizar el contenido general de cada tabla y separar variables que pudieran aportar valor de aquellas variables cuyo valor es una constante. En la propia base de datos se encuentra documentación que fue utilizada, existiendo una pequeña descripción de cada tabla y de cada variable, así como también se consultó la documentación pública oficial de Moodle (https://docs.moodle.org/all/es/) en la que se detalla tanto el esquema general del modelo de datos y de la herramienta en general, como algunas consultas de ejemplo de forma tal de poder obtener rápidamente métricas generales y comprender la naturaleza de la información. Finalmente, es de destacar que, si bien la labor exploratoria implica una cantidad de tiempo y trabajo no menor, no se ahondará en detalles sobre la misma entendiendo que escapa a la finalidad del presente trabajo.

A modo de resumen, las entidades dentro del modelo de datos de Moodle se pueden diferenciar en las siguientes áreas temáticas que serán desarrolladas a lo largo del presente trabajo:

- Usuarios
    - Incluyen al universo de todas aquellas personas que acceden a la aplicación e interactúan con la misma.
    - Están identificados a través de una clave de identificación (id) única dentro de todo el modelo de datos, como se mencionó previamente, en forma anonimizada.
    - Se pueden diferenciar según el rol que poseen y de esta forma su perfil de interacción con los cursos (alumno, docente, administrador).
    - Sobre éstos se podrá medir su grado de interacción con la aplicación, cursos a los que está asociado, o interacción tanto pública (foros) como privada (mensajes personales) dentro de la aplicación.
- Cursos
    - Abarca la información tanto de cursos dictados, así como de exámenes, preguntas, respuestas, puntajes y certificados.
    - Sobre éstos se puede relacionar a alumnos y docentes, así como también foros asociados.
    - Es importante destacar que los cursos se pueden diferenciar principalmente en dos grupos: aquellos que proveen únicamente un certificado de aprobación, y aquellos que otorgan un certificado de aprobación (y por ende asociados a la aprobación de un examen previo).
- Interacciones
    - Incluyen foros, posteos y mensajes entre usuarios.
    - Los posteos están asociados a foros, así como los foros están asociados cursos.
    - Los mensajes ocurren por fuera de los foros y se dan únicamente entre pares de usuarios.
    - Con el objetivo de preservar la privacidad y confidencialidad de los usuarios, si bien es posible obtener un registro de interacción por mensajes personales entre usuarios, el contenido de estos mensajes no está presente dentro del modelo de datos, protegiendo así la privacidad de estos.
- Métricas del sitio
    - Abarcan todas las estadísticas de logueo y paginado dentro del sitio.
    - Existen también tablas que acumulan en forma periódica (diaria, semanal, mensual) dichas estadísticas a nivel usuario.

Por último, los análisis se realizaron utilizando el software RStudio (version 1.2.1335) (https://rstudio.com/), bajo el motor de R (versión 3.5.3) (https://www.r-project.org/). Las tablas y gráficos se crearon con el mismo software, así como con la ayuda de Microsoft Excel (versión dentro de Microsoft Office 365) (https://products.office.com/es-ar/excel). Los grafos y modelos predictivos se simularon utilizando tanto paquetes dentro de R (por ejemplo, “igraph”), así como las visualizaciones fueron realizadas principalmente con Gephi (versión 0.9.2) (https://gephi.org/).

Un mayor detalle de resultados y scripts serán explayados dentro del Anexo.

CONCEPTOS Y TERMINOLOGÍA

A continuación, se presentará la definición de grafo, los elementos que lo componen, al igual que las variantes más importantes que existen y que serán abarcadas en el presente trabajo. Se definirán, a su vez, las métricas utilizadas para analizarlos y describirlos. Se definirá el concepto de red de mundo pequeño, tipo de grafo que será objeto de análisis y de utilidad para la caracterización de los grafos de usuarios presentados. Por último, se detallarán los modelos estadísticos aplicados a grafos que serán utilizados sobre el final del desarrollo. Dichos modelos servirán para definir mejor el comportamiento de cada grafo, así como entender las causas que llevan a que un usuario posea contacto con otros. De esta manera, se identificarán las variables más relevantes al momento de definir una mayor asociación entre grafos, y se podrá cuantificar la importancia o peso de cada una.

Introducción al dominio de grafos

Definición de grafo

En matemáticas y ciencias de la computación, un grafo G es un conjunto de objetos llamados vértices o nodos V unidos por enlaces llamados aristas o arcos E, que permiten representar relaciones binarias entre elementos de un conjunto. Es decir, un grafo en su forma más base es un par de conjuntos de nodos y aristas (pares de nodos). El grado de un nodo está dado por el número de aristas que están unidos a él.

$$G = \{V, E\}$$

Ecuación 1. Definición general de grafo.

Tipos de nodos y grafos

Existe un número importante de variantes de grafos y de características que los definen. Se indican a continuación los tipos de grafos más relevantes y que se usarán en esta tesis. Para una descripción más exhaustiva consultar [16].

Grafo simple y grafo múltiple

Dado un grafo G, si más de una arista une a dos nodos entonces se trata de un nodo múltiple. En el caso de que un nodo esté unido por una arista a sí mismo se llama a esto un bucle. Un grafo que no posee nodos múltiples ni bucles es llamado grafo simple.

Grafo dirigido y grafo no dirigido

En el caso de que una arista posea una dirección, es decir que un nodo es el inicio de la arista y el otro su destino, se llama entonces a éste un grafo dirigido. Se llama paso a la vía que se lleva de un nodo a otro dentro de un grafo. En el presente trabajo se analizarán únicamente grafos no dirigidos.

Grafo conectado y grafo desconectado

Los grafos se pueden diferencias en grafos conectados, cuando todos los nodos poseen al menos una arista que los conecte. Mientras que un grafo desconectado es aquel que posee por lo menos un nodo o grupo de nodos que no está conectado al resto.

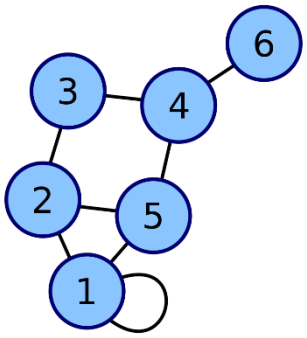

Ilustración 1. Ejemplo de grafo conectado, no dirigido, con nodos múltiples (1, 2, 3, 4, 5), y con un bucle sobre el nodo 1 (https://en.wikipedia.org/wiki/Loop_(graph_theory)).

Métricas utilizadas para el análisis de grafos

A continuación, un listado de las métricas más relevantes para el análisis de grafos. Para más información es posible consultar la bibliografía recomendada [13] [18].

Diámetro

El diámetro de un grafo está definido por la longitud de la geodésica más larga, es decir, por la mayor distancia posible entre dos nodos v y u.

$$d = \max_{v \in V} \left(\max_{u \in V} d(v,u) \right)$$

Ecuación 2. Diámetro.

Densidad

La densidad D de un grafo se define como la relación entre el número de aristas e y la cantidad de aristas posibles, dados los vértices v presentes en el grafo.

$$D = \frac{2|e|}{|v|(|v|-1)}$$

Ecuación 3. Densidad.

Transitividad

La transitividad, o coeficiente de clustering, mide la probabilidad de que los vértices adyacentes a un vértice dado estén conectados. En el presente trabajo, el cálculo se consideró a partir de la obtención del coeficiente de clustering global, el mismo se obtiene a partir de la proporción entre triángulos y triples conectados en el grafo.

Es posible definir, por lo tanto, como coeficiente de clustering global $\bar{C}$:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

Ecuación 4. Coeficiente de clustering global.

Donde cada coeficiente de clustering local C_i, dado un grafo no dirigido, se define como:

$$C_i = \frac{2\left|\{e_{jk}: v_j, v_k \in N_i, e_{jk} \in E\}\right|}{k_i(k_i - 1)}$$

Ecuación 5. Coeficiente de clustering local.

Siendo e_{ij} la arista que conecta los vértices v_i y v_j. Mientras que se definen k_i cantidad de vértices para cada vecindario N_i. Es posible definir cada vecindario N_i como:

$$N_i = \{v_j: e_{ij} \in E \vee e_{ij} \in E\}$$

Ecuación 6

Asortatividad

El coeficiente de asortatividad mide el nivel de homofilia del grafo en función de etiquetas o de valores asociados a los vértices. Si el coeficiente es grande, esto significa que los vértices conectados tienden a tener las mismas etiquetas o valores asociados similares. Es una métrica asimilable al coeficiente de correlación de Pearson.

Dentro del presente trabajo se utilizará la asortatividad basada en valores asociados a los vértices definida como se detalla a continuación:

$$r = \sum_{jk} jk\left(e_{jk} - q_j q_k\right) \Big/ \sigma_q^2$$

Ecuación 7. Asortatividad.

Donde e_{jk} es la fracción de aristas que conectan los vértices de grado j y k. El valor q_j está definidos como el “exceso de grado” calculado como uno menos que el grado de cada vértice correspondiente. Esto se debe a que en muchos casos estamos interesados en el número de aristas unidas a cada vértice excluyendo la arista particular que se está analizando en ese momento.

Siendo p_k la probabilidad de que al elegir una arista al azar tendrá grado k, entonces el “exceso de grado se define se define como a continuación:

$$q_k = \frac{(k+1)p_{k+1}}{\sum_{j\geq 1} j p_j}$$

Ecuación 8

Y por lo tanto se establecen las siguientes relaciones:

$$\sum_{jk} e_{jk} = 1$$

Ecuación 9

$$\sum_{j} e_{jk} = q_k$$

Ecuación 10

Construcción de grafos

Finalmente, es importante mencionar cómo se construye un grafo y cómo fue efectivamente construido a partir de la información brindada por el INC en base a lo extraído de la plataforma Moodle.

Al extraer la información, las relaciones se encuentran generalmente dadas por matrices en forma de lista donde en dos variables o columnas se listan todos los pares de nodos conectados entre sí. De esta forma, el usuario A por ejemplo aparecerá en la primera columna tantas veces como conexiones tenga con otros usuarios, listados éstos en la segunda columna o variable. De esta forma, se repetirá cada nodo varias veces y cada tupla tendrá su correlación con una arista distinta dentro del grafo.

En ciertos casos es necesario transformar estas matrices en forma de lista, en matrices de adyancencia. Las matrices de adyacencia se construyen como matrices cuadradas donde cada fila representa un nodo, en forma ordenada, y tiene correspondencia con cada columna. El valor de cada elemento de la matriz referencia la presencia, o no, de una arista entre el par de nodos correspondiente a los respectivos valores de fila/columna. Es importante destacar que, para grafos no dirigidos, la matriz de adyancencia es simétrica.

Redes de mundo pequeño

Se llama red de mundo pequeño a un conjunto de redes en el que la distancia geodésica media (es decir, la ruta más corta) entre nodos aumenta lo suficientemente lentamente como una función del número de nodos en la red. A modo de ejemplo, se puede definir que dicha distancia L entre dos nodos tomados al azar crece proporcionalmente como el logaritmo de la cantidad de vértices N dentro de la red.

$$L \propto \log N$$

Ecuación 11

Se trata así de un tipo de grafo para el que la mayoría de los nodos no son vecinos entre sí, y sin embargo la mayoría de los nodos pueden ser alcanzados desde cualquier nodo origen a través de un número relativamente corto de saltos entre ellos.

Para establecer si un grafo es asimilable a una red de mundo pequeño existen diferentes métodos. Un primer método es ajustar el grafo mediante una distribución de potencias y verificar, entre otros valores, la prueba de significancia de Kolmogorov-Smirnov. La significancia de la prueba determinará si se trata o no de una red de mundo pequeño. Es igualmente necesario revisar los valores de alfa y xmin correspondientes.

El valor de alpha debe ser superior a la unidad, y el valor de xmin relativamente pequeño, para estar éstos dentro de parámetros aceptables para pertenecer a una ley de potencias.

Un segundo método es utilizar tanto el modelo de Barabási-Albert, como el modelo de Erdos-Renyi. A partir de la utilización de estos modelos se pueden generar nuevos grafos asimilables a redes de mundo pequeño. Por lo tanto, luego de la generación de estos nuevos grafos se puede utilizar el coeficiente de clustering para comparar y descartar, o no, si la red analizada se trata de una asimilable a una red de mundo pequeño.

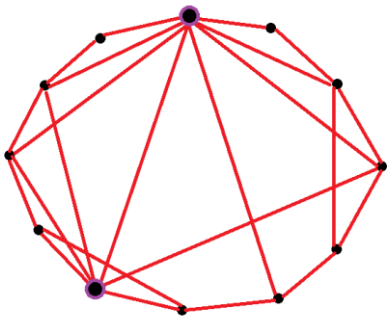

Ilustración 2. Ejemplo de red de mundo pequeño (https://en.wikipedia.org/wiki/Small-world_network).

Prueba de Kolmogorov-Smirnov

Se trata de una prueba, o test, no paramétrico de igualdad de dos distribuciones de probabilidad continuas y unidimensionales. Se puede usar para comparar una muestra con una distribución de probabilidad determinada, o comparar dos muestras entre sí.

El estadístico de Kolmogorov-Smirnov mide la distancia entre las dos distribuciones dadas. El mismo se calcula en forma asimilable a un test de hipótesis en el cual la hipótesis nula sostiene que ambas distribuciones son similares.

$$D_{nm} = \sup_x \left| F_{1,n}(x) - F_{2,n}(x) \right|$$

Ecuación 12. Estadístico de Kolmogorov-Smirnov.

Donde $F_{1,n}$ y $F_{2,n}$ son las funciones de distribución empíricas y sup es el supremo del módulo de la diferencia, es decir, la mínima cota superior de este.

Dentro del presente trabajo, se calculará el p-valor de la prueba de Kolmogorov-Smirnov con el objeto de determinar si se comprueba, o no, la hipótesis nula de que la distribución original del grafo analizado se corresponde con una distribución de leyes de potencia. Valores pequeños del p-valor indicarán que la prueba rechaza dicha hipótesis nula y, por ende, la distribución del grafo no se corresponde con una ley de potencias. Valores altos del p-valor no rechazarán dicha hipótesis.

Modelo de Barabási-Albert

Se denomina modelo de Barabási-Albert al modelo capaz de generar redes libres de escala empleando un mecanismo llamado de conexión preferencial [1]. Se define como red libre de escala a aquella red en la cual algunos nodos están altamente conectados, es decir, poseen un gran número de enlaces a otros nodos, mientras el grado de conexión de la mayoría de los nodos es relativamente bajo.

Para definir una red de este tipo se debe comenzar por un conjunto de m_0 nodos conectados aleatoriamente, donde $m_0 \geq 2$ y el grado de cada nodo en la red inicial debe ser al menos uno. Se añaden los nuevos nodos uno a uno, donde cada nodo es conectado a m nodos de la red con una probabilidad proporcional al número de enlaces que posee los nodos de la red. Formalmente, la probabilidad p_i de que un nuevo nodo se conecte con i, dado un grado k_i para cada nodo i, es:

$$p_i = \frac{k_i}{\sum_j k_j}$$

Ecuación 13

Por ende, los nodos con un alto grado tienden a acumular rápidamente más conexiones, mientras que los que poseen un grado bajo rara vez son el origen de nuevas conexiones. Los nuevos nodos según este algoritmo se dice que poseen una "preferencia" a ser conectados con los nodos de mayor grado. Se menciona, por lo tanto, que la distribución de este modelo sigue una ley de potencias.

Este modelo fue propuesto a partir del análisis de la World Wide Web, en donde se vio que los nuevos sitios tenían una preferencia por estar conectados con sitios con un alto grado de conexiones, como por ejemplo lo es Google, antes que por otros sitios menos populares. Por lo tanto, la probabilidad de seleccionar un sitio al azar es proporcional con el grado de conexiones que posee dicho sitio.

Modelo de Erdos-Renyi

El modelo de Erdos-Renyi es un modelo capaz de generar grafos aleatorios [7]. A diferencia del modelo de Barabási-Albert, en el modelo de Erdos-Renyi todos los grafos con una cantidad fija de aristas y de vértices son igualmente probables, por lo cual su distribución no sigue una ley de potencias.

Se define un grafo G tal que posea n vértices y m aristas, donde las m aristas se eligen en forma aleatoria y uniforme sobre el total de aristas posibles.

$$G(n, m)$$

Ecuación 14

Modelos estadísticos aplicados a grafos

A continuación, se desarrollarán distintos modelos estadísticos relevantes aplicados a grafos. El desarrollo e implementación tanto teórica como práctica de estos modelos en R seguirá los lineamientos dados en [10], lo que a su vez podrá servir de consulta para un mayor detalle y entendimiento de estos.

Modelos exponenciales de grafos aleatorios

Los modelos exponenciales de grafos aleatorios (ERGMs, por sus siglas en inglés) son asimilables a los clásicos modelos lineales generalizados (GLMs, por sus siglas en inglés). Están formulados de manera tal que intentan facilitar la adaptación y extensión de principios estadísticos y métodos para la construcción, ajuste, y comparación de los modelos. Sin embargo, mucha de la infraestructura de inferencia estándar disponible para GLMs, que descansa en aproximaciones asintóticas a distribuciones de chi-

cuadrado, tiene que ser todavía formalmente justificada en el caso de ERGMs. Por lo tanto, si bien estos modelos poseen un amplio potencial, deben ser usados con cuidado en la práctica sobre bases pertenecientes a grafos.

Formulación general

Dado un grafo aleatorio $G = (V, E)$; siendo $Y_{ij} = Y_{ji}$ una variable binaria aleatoria que indica la presencia o ausencia de una arista $e \in E$ entre dos vértices i y j en V. La matriz $\boldsymbol{Y} = [Y_{ij}]$ es entonces una matriz (aleatoria) de adyacencia para G; siendo $\boldsymbol{y} = [y_{ij}]$ un caso particular de $\boldsymbol{Y}$. Un modelo exponencial de grafo aleatorio es un modelo especificado en forma de la familia exponencial para la distribución conjunta de los elementos de $\boldsymbol{Y}$. La especificación básica para un modelo ERGM es de la forma siguiente:

$$P_\theta(\boldsymbol{Y} = \boldsymbol{y}) = \left(\frac{1}{k}\right) exp\left\{\sum_H \theta_H g_H(\boldsymbol{y})\right\}$$

Ecuación 15

Donde:

i. Cada H es una configuración definida como un conjunto de posibles aristas dentro de un subconjunto de vértices en G;
ii. $g_H(\boldsymbol{y}) = \prod_{y_{ij}} \in H\ y_{ij}$, y es por lo tanto uno si la configuración de H ocurre en $\boldsymbol{y}$, o cero en caso contrario;
iii. un valor distinto de cero para θ_H significa que las Y_{ij} son dependientes de todos los pares de vértices $\{i, j\}$ en H, condicional al resto del grafo; y
iv. $k = k(\theta)$ es una constante de normalización

$$k(\theta) = \sum_{\boldsymbol{y}} exp\left\{\sum_H \theta_H g_H(\boldsymbol{y})\right\}$$

Ecuación 16

Especificación de un modelo

Dentro de la familia de modelos ERGM, es posible definir un caso particular equivalente a un modelo de grafo aleatorio de Bernoulli. Para esto, es necesario definir que, para cualquier par de vértices, la presencia o ausencia de una arista entre dicho par es independiente del estado de posibles aristas entre cualquier otro par de vértices. Por lo tanto $\theta_H = 0$ para toda configuración de H que involucre tres o más vértices. Como resultado, el modelo ERGM se reduce a:

$$P_\theta(\boldsymbol{Y} = \boldsymbol{y}) = \left(\frac{1}{k}\right) exp\left\{\sum_{i,j} \theta_{ij} y_{ij}\right\}$$

Ecuación 17

Si además se asume que los coeficientes θ_{ij} son iguales a un valor común θ (típicamente señalado como una asunción de homogeneidad a través de la red) entonces la ecuación 17 se reduce a:

$$P_\theta(\boldsymbol{Y} = \boldsymbol{y}) = \left(\frac{1}{k}\right) exp\{\theta L(\boldsymbol{y})\}$$

Ecuación 18

Donde $L(y) = \sum_{i,j} y_{ij} = N_e$ es el número de aristas en el grafo. El resultado es equivalente a un modelo de grafo aleatorio de Bernoulli, con $p = exp(\theta)/[1 + exp(\theta)]$.

Este modelo posee una amplia variedad de estadísticos a ser introducidos. Sin embargo, se presentará un modelo que contemple los efectos de la transitividad, es decir, la relación de cada elemento con sus contactos de distinto grado, así como los efectos de las variables a nivel usuario detalladas previamente, consideradas como variables independientes.

Un primer estadístico está relacionado con la proporción de aristas que puede haber en una red, definida dentro del modelo en la ecuación 18. Es posible incluir distintos estadísticos de distintos niveles dentro de la estructura de la red, contados como *k-estrellas* $S_k(\boldsymbol{y})$ (donde $S_1(\boldsymbol{y}) = N_e$ es igual al número de aristas), y triángulos $T(\boldsymbol{y})$.

$$P_\theta(\boldsymbol{Y} = \boldsymbol{y}) = \left(\frac{1}{k}\right) exp\left\{\sum_{k=1}^{N_v - 1} \theta_k S_k(\boldsymbol{y}) + \theta_\tau T(\boldsymbol{y})\right\}$$

Ecuación 19

Normalmente, este tipo de modelos incluyen un recuento de estrellas S_k no superior a $k = 2$, o $k = 3$ como mucho.

El efecto de transitividad se puede modelar a través de la inclusión de un parámetro de recuento de grados ponderados geométricamente, definido como:

$$GWD_\gamma(\boldsymbol{y}) = \sum_{d=0}^{N_v - 1} e^{-\gamma d} N_d(\boldsymbol{y})$$

Ecuación 20

Donde $N_d(\boldsymbol{y})$ es el número de vértices con grado d y $\gamma > 0$ está relacionado con λ a través de la expresión $\gamma = log[\lambda/(\lambda - 1)]$.

Por otro lado, es importante modelar estadísticos que no sean sólo función de la red $\boldsymbol{y}$, es decir, no únicamente controlar efectos endógenos. Sino esperar que la probabilidad de que una arista conecte dos vértices dependa no sólo en el estado (presencia o ausencia) de aristas entre otros pares de vértices, sino también de atributos de los propios vértices (permitiendo la presencia de efectos exógenos). Es posible incluirlos en forma de estadísticos adicionales dentro del término exponencial de la ecuación 15, con la constante de normalización k modificada de acuerdo con la ecuación 16.

Dicho estadístico puede definirse de la siguiente manera:

$$g(\boldsymbol{y}, \boldsymbol{x}) = \sum_{1 \le i \le j \le N_v} y_{ij} h(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

Ecuación 21

Donde h es una función simétrica de $\boldsymbol{x}_i$ y $\boldsymbol{x}_j$, y $\boldsymbol{x}_i$ (o $\boldsymbol{x}_j$) es un vector de atributos observados para el i-ésimo (o j-ésimo) vértice. De manera intuitiva, si h es cierta medida de "similaridad" en atributos, entonces el estadístico de la ecuación 21 valora la similaridad total entre vecinos de la red.

Dos elecciones comunes de h producen en forma análoga "efectos primarios" y "efectos de segundo orden" en ciertos atributos. Los efectos primarios sobre un atributo particular x están definidos en forma aditiva:

$$h(x_i, x_j) = x_i + x_j$$

Ecuación 22

Por otro lado, los efectos de segundo orden están definidos usando un indicador de equivalencia respecto al atributo en cada par de vertices, representando la similaridad u homofilia en cada par de nodos.

$$h(x_i, x_j) = I\{x_i = x_j\}$$

Ecuación 23

Ajuste del modelo

Para ajustar modelos exponenciales como el presentado anteriormente, asumiendo variables independientes e idénticamente distribuídas, se usa el método de máxima verosimilitud. Dado el modelo ERGM de la ecuación 15, el estimador máximo verosímil (MLE, por sus siglas en inglés) $\hat{\theta}_H$ del parámetro θ_H está bien definido pero su cálculo no es trivial.

El MLE del vector $\theta = (\theta_H)$ está definido como $\hat{\theta} = \arg max_\theta l(\theta)$, donde $l(\theta)$ es el logaritmo de la verosimilitud, que tiene la forma simple común a las familias exponenciales:

$$l(\theta) = \theta^T \boldsymbol{g}(\boldsymbol{y}) - \psi(\theta)$$

Ecuación 24

Donde $\boldsymbol{g}$ se refiere al vector de funciones g_H y $\psi(\theta) = \log k(\theta)$. En forma alternativa, aplicando derivadas a cada lado de la ecuación y usando el hecho de que $E_\theta[\boldsymbol{g}(\boldsymbol{Y})] = \partial\psi(\theta)/\partial\theta$, el MLE puede también expresarse como la solución del siguiente sistema de ecuaciones:

$$E_{\hat{\theta}}[\boldsymbol{g}(\boldsymbol{Y})] = \boldsymbol{g}(\boldsymbol{y})$$

Ecuación 25

Sin embargo, la función $\psi(\theta)$ que aparece tanto en la ecuación 24 como en la 25, no puede ser evaluada en forma explícita en ninguno salvo los casos más triviales, ya que implica la suma de la ecuación 16 sobre $2^{\binom{N_v}{2}}$ posibles valores de $\boldsymbol{y}$, para cada

candidato de θ. Por lo tanto, es necesario usar métodos numéricos para calcular valores aproximados de $\hat{\theta}$.

A partir de utilizar el paquete ergm en R, los modelos de ajustan a partir de la función *ergm*, que implementa una versión de estimación de máxima verosimilitud sobre un modelo de Monte Carlo utilizando cadenas de Markov.

Es posible a su vez, realizar un análisis de la varianza (ANOVA), para comprender si existe evidencia de que las variables utilizadas en el modelo explican la variación en la conectividad de la red.

De igual manera, se puede analizar la contribución individual de cada variable al modelo. Para la interpretación de cada coeficiente, es útil pensar en términos de probabilidad de que determinado par de vértices tenga una arista, condicional al estado de las aristas entre todos los otros pares. Definiendo $\boldsymbol{Y}_{(-ij)}$ como todos los elementos de $\boldsymbol{Y}$ excepto Y_{ij}, la distribución de Y_{ij} condicional a $Y_{(-ij)}$ es una Bernoulli y satisface la siguiente expresión:

$$\log\left[\frac{P_\theta(Y_{ij}=1|\boldsymbol{Y}_{(-ij)}=\boldsymbol{y}_{(-ij)})}{P_\theta(Y_{ij}=0|\boldsymbol{Y}_{(-ij)}=\boldsymbol{y}_{(-ij)})}\right]=\theta^T\Delta_{ij}(\boldsymbol{y})$$

Ecuación 26

Donde $\Delta_{ij}(\boldsymbol{y})$ es el *cambio estadístico*, dado por la diferencia entre $\boldsymbol{g}(\boldsymbol{y})$ cuando $y_{ij}=1$ y cuando $y_{ij}=0$.

Bondad de ajuste

Para modelos como el planteado previamente, la práctica común es primero simular varios grafos aleatorios sobre el modelo ajustado para luego comparar características de alto nivel de estos grafos con los grafos observados originalmente. Ejemplos de estas características incluyen la distribución de cualquier número de los distintos resúmenes de estructuras de redes como grado, centralidad, o distancia geodésica. Si las características de los grafos de redes observados no coinciden con los valores típicos surgidos del modelo de grafos aleatorios, se puede interpretar que existen diferencias sistémicas entre la clase de modelos especificados y los datos y, por lo tanto, hay una ausencia de bondad en el ajuste del modelo.

Modelos de bloque

Se observó que los modelos ERGM son asimilables a modelos de regresión estándar, donde la presencia o ausencia de aristas (esto es, de Y_{ij}) es tomada como la variable de respuesta, mientras que el rol de las variables predictoras es tomado por una combinación de estadísticos de la red (variables endógenas) y funciones sobre vértices y atributos de las aristas (variables exógenas). En esta sección, se verá que los modelos de bloque de redes son análogos a modelos de mezcla. Esto es, dada una variable aleatoria X sigue una distribución mixta de Q-clases si la función de densidad de probabilidad es de la forma $f(x)=\sum_{q=1}^{Q}\alpha_k f_q(x)$, para una densidad específica de clase f_q, donde la mezcla de los pesos α_q son no negativos y suman uno.

A continuación, se hará una breve mención de la funcionalidad y utilidad de este tipo de modelos. Es importante describir este tipo de modelos ya que los modelos de red latente, que serán descriptos a continuación, son una extensión de los modelos de bloque de redes.

Especificación del modelo

Suponiendo que cada vértice $i \in V$ de un grafo $G = (V, E)$ puede pertenecer a una de las Q clases, es decir $C_1, \dots, C_Q$. Y, además, suponiendo que se sabe la etiqueta de la clase $q = q(i)$ para cada vértice i. Un modelo de bloque para G especifica que cada elemento Y_{ij} de la matriz de adyacencia $\boldsymbol{Y}$ es, condicional a la etiqueta de clases q y r de los vértices i y j respectivamente, una variable aleatoria independiente de Bernoulli, con probabilidad π_{qr}. Para un grafo no dirigido entonces $\pi_{qr} = \pi_{rq}$.

Por lo tanto, el modelo de bloque es una variante de un modelo de grafo aleatorio de Bernoulli, donde las probabilidades de una arista están restringidas a ser una de sólo Q^2 valores posibles π_{qr}. Adicionalmente, en forma análoga a la ecuación 18, este modelo puede ser representado en la forma de un ERGM:

$$P_\theta(\boldsymbol{Y} = \boldsymbol{y}) = \left(\frac{1}{k}\right) exp\left\{\sum_{q,r} \theta_{qr} L_{qr}(\boldsymbol{y})\right\}$$

Ecuación 27

Donde $L_{qr}(\boldsymbol{y})$ es el número de aristas en el grafo observado $\boldsymbol{y}$**,** conectando pares de vértices de clases q y r.

Sin embargo, no es posible asumir que siempre se podrá saber la clase de cada vértice del grafo, o que la clase real $C_1, \dots, C_Q$ del grafo haya sido correctamente especificada. Por lo tanto, se utiliza más comúnmente lo que se denomina un modelo de bloque estocástico (SBM, por sus siglas en inglés). Este modelo especifica que sólo existen Q clases, para algún Q, pero no especifica la naturaleza de estas clases ni tampoco la pertenencia de vértices individuales a estas clases. Más bien, determina que la pertenencia a una clase en cada vértice i está determinada de forma independiente de acuerdo con una distribución común $\{1, \dots, Q\}$.

Formalmente, siendo $Z_{iq} = 1$ si el vértice i es de clase q, y cero en caso contrario. En un modelo de bloque estocástico, los vectores $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iQ})$ están determinados independientemente, donde $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iQ})$ y $\boldsymbol{Z}_i = (Z_{i1}, \dots, Z_{iQ})$. Entonces, condicional a los valores $\{Z_i\}$, las entradas Y_{ij} son modeladas como variables aleatorias independientes de Bernoulli, con probabilidad π_{qr}, como en el modelo de bloque no estocástico.

Ajuste del modelo

Un modelo de bloque no estocástico se puede ajustar de manera sencilla. Los únicos parámetros que estimar son la probabilidad de cada vértice π_{qr}, y los estimadores de máxima verosimilitud corresponden simplemente con las frecuencias empíricas.

En el caso de modelos SBM, tanto la probabilidad de cada vértice π_{qr} y la probabilidad de pertenencia a una clase α_q tienen que ser estimadas. Si bien no parece un gran cambio respecto a modelos de bloque ordinarios, la tarea de ajuste se vuelve más compleja. Esto se debe a que para obtener la verosimilitud del modelo es necesario utilizar métodos intensivos computacionalmente.

Una solución es usar el algoritmo de maximización de la expectativa (EM, expectation-maximization en inglés). Dado un estimador del vértice π_{qr}, se calculan los valores esperados $\boldsymbol{Z}_i$, condicional a $\boldsymbol{Y} = \boldsymbol{y}$. Estos valores son a su vez utilizados para calcular nuevos estimadores del vértice π_{qr}, usando principios de máxima verosimilitud (condicional). Se repiten estos pasos hasta lograr una convergencia en dichos valores. Para esto, se proponen una serie de métodos que aproximan o alteran el problema de máxima verosimilitud original. Uno de ellos es el llamado enfoque variacional, que optimiza el límite inferior de verosimilitud de datos observados.

La salida de un modelo SBM ajustado permite además la asignación de vértices a clases. Por lo tanto, pueden usarse estos modelos como una forma de particionar grafos. Específicamente, produciendo estimadores de los parámetros π_{qr} y α_q, algoritmos de este tipo calculan necesariamente estimadores de los valores esperados para $\boldsymbol{Z}_i$, condicional a $\boldsymbol{Y} = \boldsymbol{y}$. Esto es, calculan estimadores de la probabilidad a posteriori de membresía a una clase, lo que a su vez puede usarse para determinar la asignación a dicha clase.

Bondad de ajuste

Por último, la bondad de ajuste de los modelos SBM se calcula utilizando los mismos tipos de métodos de simulación empleados para el análisis de los modelos ERGM. Sin embargo, la forma particular de estos modelos lleva a utilizar ciertos métodos específicos de modelado. A modo de ejemplo se pueden mencionar la probabilidad de clasificación integrada (ICL, por sus siglas en inglés), la reorganización de la matriz de adyacencia en función de las clases, la distribución de grado del grafo, así como la comparación entre las clases reales y las obtenidas a partir de las probabilidades dadas por el modelo. Todas estas métricas son de utilidad para determinar la presencia, o no, de bondad de ajuste en el modelo seleccionado.

Modelos de red latente

Finalmente, es posible presentar los modelos de red latente como una extensión de los modelos SBM. Este tipo de modelos incorpora variables latentes, esto es, variables no observadas pero que son importantes para determinar la probabilidad de que un par de vértices tengan incidencia entre sí.

Formulación general

En líneas generales, y en la ausencia de cualquier información sobre covariables, es posible asumir que los vértices $v \in V$ son intercambiables, y por lo tanto que cada elemento Y_{ij} de la matriz de adyacencia $\boldsymbol{Y}$ puede ser expresado en la forma:

$$Y_{ij} = h(\mu, u_i, u_j, \varepsilon_{ij})$$

Ecuación 28

Donde μ es una constante, u_i y u_j son variables latentes independientes e idénticamente distribuídas, ε_{ij} son efectos específicos de pares independientes e idénticamente distribuídos, y la función h es simétrica en su segundo y tercer argumento. En resumen, bajo intercambiabilidad, cualquier matriz aleatoria de adyacencia $\boldsymbol{Y}$ puede ser escrita como función de variables latentes.

Dada la ecuación 28, es posible formular distintos tipos de modelos de red latente. Es necesario especificar que (i) los efectos ε_{ij} están distribuidos como variables normales estándar aleatorias; (ii) las variables latentes u_i y u_j entran dentro de la función *h* sólo como una función simétrica; y (iii) la función *h* es un simple indicador de si el argumento es o no positivo, y si adicionalmente se aumenta el parámetro μ para incluir una combinación lineal de covariables de pares específicos $\boldsymbol{x}_{ij}^T\beta$, se obtiene una versión en redes de un modelo probit. Bajo este modelo, las Y_{ij} son condicionalmente independientes, con distribución:

$$P(Y_{ij} = 1|\boldsymbol{X}_{ij} = \boldsymbol{x}_{ij}) = \Phi\left(\mu + \boldsymbol{x}_{ij}^T\beta + \alpha(u_i, u_j)\right)$$

Ecuación 29

Donde Φ es la función de distribución acumulada de una variable aleatoria normal estándar.

Si se señalan las probabilidades en la ecuación 29 como p_{ij}, entonces el modelo condicional para $\boldsymbol{Y}$ toma la forma:

$$P(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{X}, u_1, \dots, u_{N_v}) = \prod_{i<j} p_{ij}^{y_{ij}}(1 - p_{ij})^{1-y_{ij}}$$

Ecuación 30

Condicional a las covariables $\boldsymbol{X}$ y las variables latentes $u_1, \dots, u_{N_v}$, este modelo para G tiene la forma de un modelo de grafo aleatorio de Bernoulli, con probabilidad p_{ij} para cada par de vértices i, j.

Especificación de los efectos latentes

Es posible señalar, a modo de ejemplo, tres tipos de variables latentes posibles de ser modeladas dada una variable latente u con probabilidad de que existe una arista entre un par de vértices bajo la función de forma $\alpha(.,.)$.

En primer lugar, en forma análoga a los modelos SBM, se puede señalar un modelo utilizando clases latentes que son codificadas bajo el principio de equivalencia estructural dentro de la teoría de redes sociales. La equivalencia estructural se puede resumir como el grado en que dos nodos están conectados con otros nodos iguales, es decir, tienen el mismo entorno social. Esto se define especificando que u_i toma valores

dentro de $\{1, \dots, Q\}$, y $\alpha(u_i, u_j) = m_{u_i,u_j}$, para una matriz simétrica $\boldsymbol{M} = [m_{q,r}]$ con valores reales de entrada $m_{q,r}$.

En segundo lugar, el principio de homofilia (tendencia a que individuos similares se asocien entre sí) puede ser modelado basado en el concepto de distancia en el espacio latente. Este tipo de modelos también son llamados modelos de distancia latente. Dicha distancia calculada como la diferencia entre características de cada par de nodos. Las variables latentes u_i son vectores $(u_{i1}, \dots, u_{iQ})^T$ de números reales, interpretados como vértices importantes, pero con características desconocidas que influencian la existencia de aristas entre ellos. Vértices con mayor similitud en características tendrán una mayor probabilidad de que estén conectados. Los efectos latentes son especificados como $\alpha(u_i, u_j) = -|u_i - u_j|$, para una métrica de distancia $|\cdot|$.

Como tercera posibilidad de especificar efectos latentes existe una combinación de las dos anteriores, basada en el principio de "eigen-analysis" (análisis propio). Este tipo de modelos puede verse como una generalización de los anteriores y, por lo tanto, su uso permite la combinación de principios de equivalencia estructural y de homofilia. La forma en la cual estos dos principios son combinados puede determinarse durante el proceso de ajuste del modelo. Los u_i son vectores aleatorios Q-largos, pero los efectos latentes están dados por la forma $\alpha(u_i, u_j) = u_i^T \Lambda u_j$, donde Λ es una matriz diagonal QxQ. Siendo las variables latentes u modeladas como vectores aleatorios independientes e idénticamente distribuídos provenientes de la misma distribución, entonces la correlación entre cada par de u_i y u_j es cero. Juntando los u_i en una matriz $\boldsymbol{U} = [u_1, \dots, u_Q]$, el producto $\boldsymbol{U}\Lambda\boldsymbol{U}^T$ puede entonces pensarse como una descomposición propia de la matriz para todos los pares de efectos latentes $\alpha(u_i, u_j)$.

Ajuste del modelo

El ajuste del modelo se utiliza realizando una aproximación bayesiana para la inferencia de los valores a calcular. La aproximación bayesiana es obtenida a partir de técnicas de simulación de Monte Carlo para cadenas de Markov (MCMC, por sus siglas en inglés). Son de particular interés el parámetro β (describiendo los efectos de pares específicos de covariables $\boldsymbol{x}_{ij}$), los elementos de la matriz diagonal Λ (sumarizando la importancia relativa de cada vector latente u_i), y los vectores latentes. Ya que los vectores latentes inferidos $\hat{u}_i$ no son ortogonales, son útiles en la interpretación de la salida del modelo para usar en su lugar los vectores propios de la matriz $\widehat{\boldsymbol{U}}\widehat{\Lambda}\widehat{\boldsymbol{U}}^T$.

Bondad de ajuste

Finalmente, para obtener la bondad de ajuste de los modelos de red latente es posible utilizar métodos de simulación similares a los utilizados para los modelos ERGM. Adicionalmente, es posible utilizar principios de validación cruzada. Una práctica común en modelado de redes es evaluar la precisión en la cual, a partir de ajustar el modelo a una parte del grafo, es posible predecir la parte restante de la red. Esto es, partir el grafo en K-subpartes, entrenar con cada una de dichas K-partes y realizar la predicción sobre las partes restantes en cada caso, para luego evaluar el desempeño del modelo en cada iteración.

Los resultados de las predicciones generadas pueden ser evaluados a partir de la observación de la curva ROC (por sus siglas en inglés. En castellano Característica Operativa del Receptor. La misma representa la ratio de verdaderos positivos frente al ratio de falsos positivos según se varía el umbral de discriminación). La forma más común de evaluar dicha curva es a partir del cálculo del área por debajo de la misma, también llamada AUC (por sus siglas en inglés). La misma se puede interpretar como la probabilidad de que un clasificador ordene o puntúe una instancia positiva elegida aleatoriamente más alta que una negativa. El modelo que tenga un mayor valor de AUC será el mejor valorado y el que mayor capacidad discriminante posea.

PRE-PROCESAMIENTO

Análisis del modelo de datos

Si bien la base de datos posee un total de 336 entidades, sólo 149 de éstas poseen al menos un registro insertado. El archivo .sql que concentra todo el modelo de datos extraído ocupa cerca de 6,2 gigabytes en disco. Se está en presencia de un total de más de 5.200 usuarios (más de cinco mil alumnos, una decena de docentes y personal no docente), asociados a cerca de cuarenta cursos, en un lapso de 16 meses de historia entre marzo de 2016 y julio de 2017.

En un primer lugar, se analizaron individualmente cada una de estas 149 entidades para establecer si la información que contenía podría ser o no útil a los fines de esta investigación. Muchas entidades son puramente transaccionales, otras poseen pocos registros y no aportan valor. Existen también entidades que conforman agrupamientos de una tercera categoría que, si bien pueden ser de utilidad al facilitar el trabajo de agrupamiento para el armado de reportes, no fueron consideradas en el presente trabajo. Es posible mencionar a modo de ejemplo las tablas: mdl_stats_user_daily, mdl_stats_user_weekly, mdl_stats_user_monthly. Estas tablas acumulan a nivel diario, semanal, y mensual, respectivamente, ciertas estadísticas a nivel usuario.

Si bien puede verse en el Anexo un detalle minucioso tanto de todas las entidades como de las consultas utilizadas, se puede resumir que la investigación se concentra en el uso de una decena de entidades con información relevante. Las mismas serán detalladas en la sección siguiente.

Creación de variables agrupadas, variables derivadas y nuevas entidades

Luego del análisis del modelo de datos completo, se prosiguió a la creación de entidades con un mayor nivel de agregación, principalmente desde el punto de vista de las descripciones.

A modo de ejemplo: existe una tabla (mdl_user) donde se encuentran listados todos los usuarios; otra tabla (mdl_role_assignments) donde se ve el rol (alumno, docente, administrador) que tiene cada usuario en forma numérica; y otra tabla (mdl_role) donde se decodifica dicho rol numérico en una descripción legible.

Son numerosos los casos en los que se debe recurrir a realizar varios cruces por razones similares a las mencionadas en el ejemplo. Estos cruces entre tablas y agregaciones se encuentran detalladas dentro del Anexo.

En una segunda instancia se seleccionaron las entidades necesarias para hacer las selecciones de usuarios y relaciones que serían finalmente objeto del desarrollo del presente trabajo.

A continuación, las entidades más relevantes para la obtención de los usuarios y armado de las relaciones:

- mdl_user
  - Entidad que posee el listado de todos los usuarios.
- mdl_role
  - Entidad que detalla todos los roles posibles dentro de la plataforma.
- mdl_ role_assignments
  - Entidad que identifica el rol de cada uno de los usuarios.
- mdl_message_contacts
  - Entidad que señala la relación entre usuarios y contactos que se enviaron mensajes.
- mdl_course
  - Entidad donde se encuentran los cursos. La misma fue necesario cruzarla con las entidades mdl_context y mdl_role_assignments para obtener los usuarios asociados a cada curso.
- mdl_context
  - Entidad que identifica los cursos asociados a cada uno de los usuarios.
- mdl_certificate
  - Entidad donde se encuentran todos los certificados posibles aplicables a cada uno de los cursos, con una descripción de cada uno de ellos.
- mdl_certificate_issues
  - Entidad donde se encuentran todos los certificados emitidos. La misma fue necesario cruzarla con la entidad mdl_certificate para obtener la descripción de cada certificado. Se descartaron todos los cursos y certificados señalados como de “prueba”, así como los certificados señalados como de participación. De esta forma, se diferenciaron manualmente los certificados asociados a la aprobación del curso a través de un examen o evaluación, de aquellos que implican únicamente la participación en el mismo.

Luego de obtenidas las entidades derivadas en forma de lista, fue necesario convertirlas en matrices de adyacencia a nivel usuario para poder trabajar con esta información, generar los respectivos grafos, y realizar los distintos análisis que se detallarán a lo largo del presente trabajo.

Las matrices de adyacencia se construyen como matrices cuadradas donde cada fila representa un nodo, en forma ordenada, y tiene correspondencia con cada columna. El valor de cada elemento de la matriz referencia la presencia, o no, de una arista entre el par de nodos correspondiente a los respectivos valores de fila/columna. Según el caso

y a modo de ejemplo, la comunicación entre cada usuario a partir de un mensaje privado, o el haber compartido un determinado curso o foro.

Se podrán apreciar en detalle en el Anexo las consultas SQL, así como el script de R utilizados. En el caso puntual de cursos, se consideró que todo par de alumnos que hayan compartido un mismo curso posee una relación, y por ende existe una arista entre el par de nodos que estos usuarios representan.

DESARROLLO

ESTADÍSTICA DESCRIPTIVA Y PRIMEROS RESULTADOS

A continuación, se presenta un breve resumen descriptivo del dataset completo.

Usuarios y roles

Se trabaja sobre un total de 5.694 usuarios que se dividen según lo que se observa en el gráfico 1.

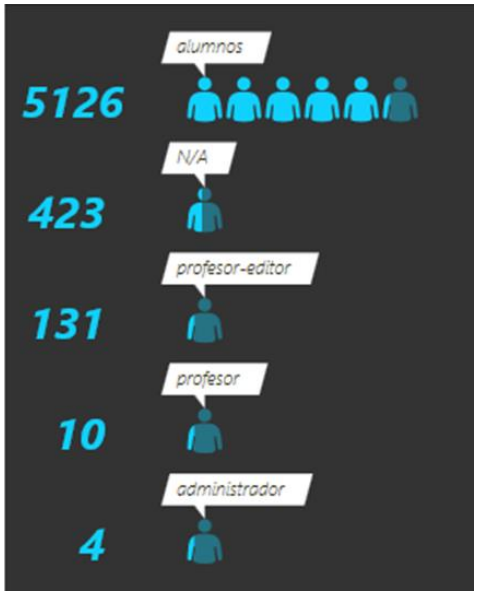

Gráfico 1. Distribución de usuarios por rol asignado.

Se puede visualizar que la mayor parte (5.126 usuarios, el 90% del total de la base) son alumnos. Existe un total de 141 usuarios que podrían ser catalogados como docentes (señalados como teacher y editingteacher en el modelo de datos), con distinto grado de privilegio en su rol, tan sólo 4 están catalogados como administradores, mientras que 423 usuarios están presentes dentro de la tabla de usuarios (mdl_user) pero no poseen un rol asignado como usuario (mdl_role_assignments). Éstos últimos fueron excluídos del estudio en su completo ya que no pudieron ser identificados, tomando como hipótesis más probable que puedan tratarse en gran parte de usuarios de prueba o que quedaron truncos al momento de la extracción.

Cursos y alumnos

La disponibilidad mensual de cursos oscila entre uno y siete, dependiendo del mes calendario. En septiembre 2016 y mayo 2017 se presentan la mayor cantidad de cursos disponibles.

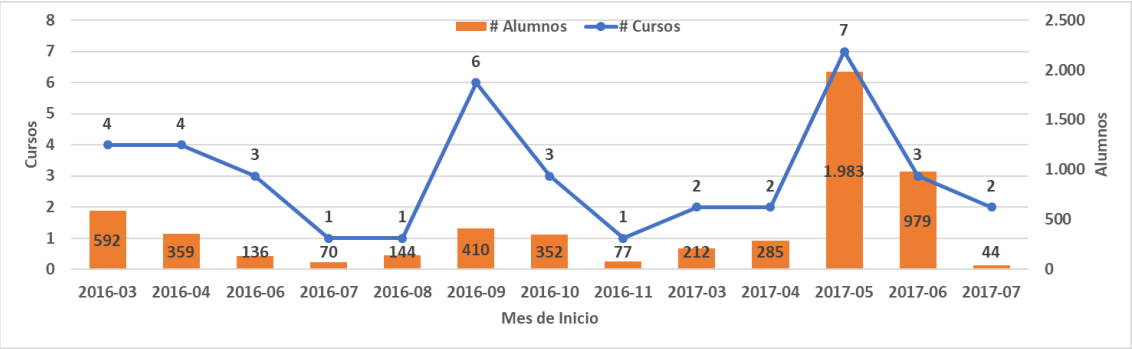

Gráfico 2. Cursos iniciados y evolución de alumnos – mensual –

A su vez, durante el periodo marzo 2016 a abril 2017 la cantidad de alumnos registrados por curso se mantiene relativamente estable. En el mes de mayo 2017 se observa un alza significativa, alcanzando a 1.983 alumnos para el total de cursos. Sin embargo, para los meses sucesivos, desciende considerablemente hasta los 44 alumnos en julio 2017.

El comportamiento cíclico de cantidad de cursos iniciados y alumnos por mes es esperable teniendo en cuenta las fechas normales del calendario lectivo argentino. Existe una correlación positiva entre inicio de cursos y aumento de la cantidad de alumnos que también es esperable teniendo en cuenta que al comenzar un curso es natural que aumente la cantidad de alumnos inscriptos nuevos para el sistema. Sin embargo, se destaca un pico de cursos iniciados en septiembre 2016 pero que no se condice con este comportamiendo viendo una baja cantidad de usuarios anotados en dichos cursos. Este caso pueda tal vez explicarse por ser principalmente el comienzo de cursos adicionales sobre alumnos que ya venían realizando algún curso previo.

Por último, es de destacar la presencia de algunas aparentes anomalías dentro de la información observada. Por ejemplo, no se presenta información de cursos que hayan iniciado en mayo 2016. Así como tampoco existe información para los meses de diciembre 2016 a febrero 2017, lo que es razonable teniendo en cuenta que coinciden con los meses de receso de verano en Argentina.

Certificados por curso

Es importante destacar que existen principalmente dos tipos de certificados. Aquellos otorgados meramente por la participación en el mismo, y aquellos otros que implican una aprobación por parte del alumno al curso en cuestión.

Se puede señalar que la totalidad de cursos oscila entre dos y nueve, dependiendo del mes calendario. En noviembre y diciembre 2016 se presentan la mayor cantidad de cursos vigentes (nueve), con un pico de siete en junio 2017. El resto de los meses posee variaciones siendo cercano a cinco la cantidad de cursos vigentes promedio por mes.

La cantidad de certificados otorgados presenta variaciones importantes según el mes que se observe. Existen dos picos con una alta cantidad de certificados otorgados en noviembre 2016 (232 certificados totales, 230 de aprobación) y junio 2017 (218 certificados totales, en coincidencia con certificados de aprobación). En el mes de

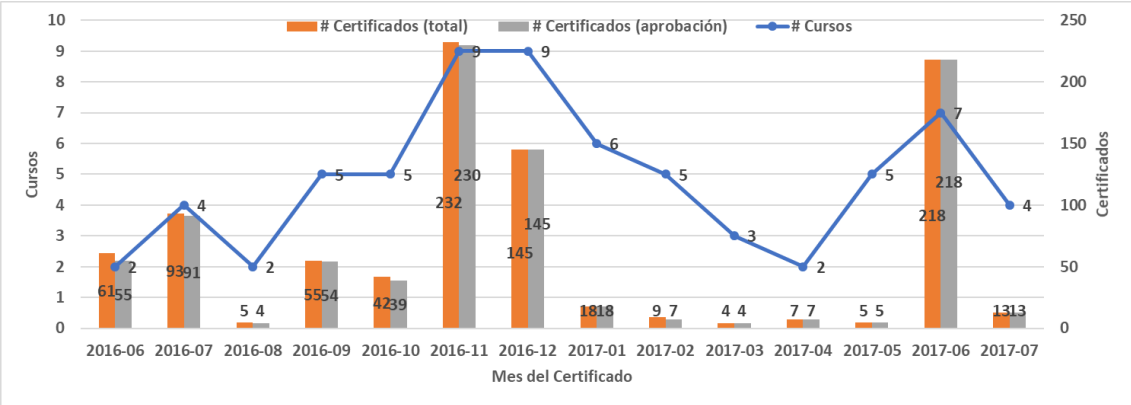

Gráfico 3. Cantidad de cursos y certificados otorgados – mensual –

diciembre 2016 se observa un valor relativamente alto de otorgamiento de certificados (145), contrarrestando la baja de los meses sucesivos. Por otro lado, no existe prácticamente diferencia entre cantidad de certificados totales otorgados y de certificados de aprobación.

Existe una correlación positiva entre cursos vigentes y aumento de la cantidad de alumnos certificados que también es esperable teniendo en cuenta que luego de la certificación disminuirá temporalmente la cantidad de cursos vigentes, para luego aumentar con el nuevo comienzo de ciclo de clases.

Contactos entre usuarios

Al agregar la información durante la totalidad del periodo de información analizado, es posible obtener la cantidad de usuarios y de contactos, diferenciados entre usuarios que contactaron a otros, así como aquellos que bloquearon a al menos otro usuario. Esta información da noción del volumen de interconexión que hay entre usuarios, dados por los contactos que tuvieron entre ellos en forma privada dentro de la plataforma.

	Usuarios	Contactos
Contactados	231	1.446
Bloqueados	18	24
Total	249	1.470

Tabla 4. Usuarios totales y contactos entre usuarios.

Se puede mencionar que:

- 231 usuarios contactaron al menos otro.
- 18 usuarios bloquearon por lo menos a otro.
- 1.446 son los contactos totales entre usuarios.
- 24 fueron los usuarios bloqueados en total.

Se puede concluir que solamente el 4,5% (231 de los más de 5 mil usuarios totales) contacta a otro usuario, lo que da cuenta de la pasividad de una gran proporción de usuarios al no generar vínculos con otros. Esta información se desprende de la tabla mdl_messages_contacts en la cual se puede ver de manera resumida la relación que existe entre los usuarios. Este número de contactos entre usuarios está en línea con lo

observado en otros trabajos [5]. Según lo ya visto en [4] una mayor interacción y participación sería beneficiosa para un mayor desempeño de los alumnos.

COMPARACIÓN DE GRAFOS

A partir de la información de contactos entre usuarios se construyeron distintos grafos que permitieran analizar la relación entre dichos usuarios a través de los contactos que poseen entre sí, así como la participación en igual curso.

Usuarios y mensajes

Características topológicas del grafo

En primer lugar, se verán los resultados que se obtienen al analizar el grafo de usuarios que se contactan a través de mensajes.

Si bien la tabla original presenta los datos como una lista con usuario y contacto, se modificó la misma de forma tal de convertirla en una matriz de adyacencia no dirigida y con pesos siempre iguales a uno. Esto se debe a que se está priorizando el contacto entre usuarios sin importar la relación origen-destino, siendo especialmente pocos los usuarios presentes.

Componente	# Nodos	% Nodos
1	613	89,9%
2	30	4,4%
3	8	1,2%
4	7	1,0%
5	4	0,6%
6	3	0,4%
7	3	0,4%
8	2	0,3%
9	2	0,3%
10	2	0,3%
11	2	0,3%
12	2	0,3%
13	2	0,3%
14	2	0,3%
Total	682	100,0%

Tabla 5. Detalle de tamaño por componente.

Se puede destacar que se está hablando de un total de 682 usuarios que contactan o son contactados. Es decir, cerca de un 12% del total de usuarios envía o recibe algún mensaje. De todos estos, 613 usarios que representan casi el 90% del total del grafo, pertenecen al componente mayor.

A simple vista se puede mencionar que existen ciertos grupos o subgrafos con su propia importancia si bien se encuentran generalmente intercontectados formando el componente principal. No obstante, no se trata de una red completamente conectada. Existen pequeños componentes sin conexión entre ellos o con el componente principal, el cual será objeto de análisis.

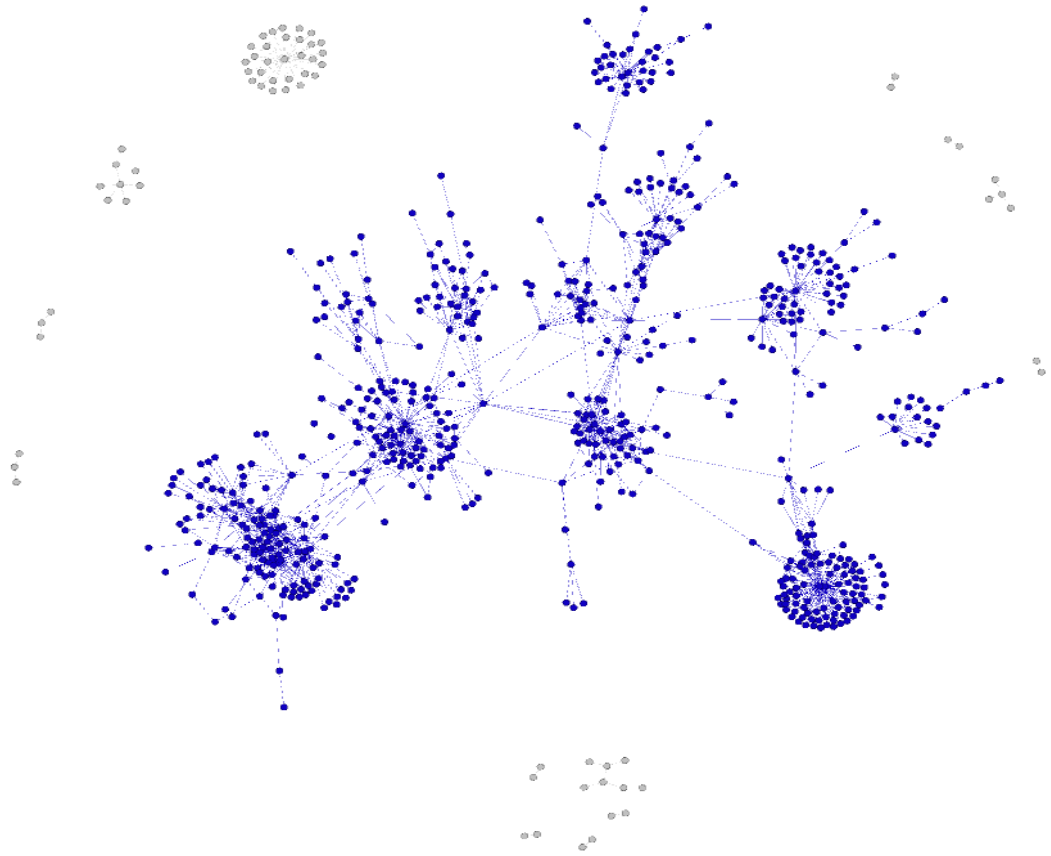

Gráfico 6. Grafo de usuarios conectados por al menos un mensaje enviado/recibido (en azul el componente mayor).

Analizando en detalle el componente principal se observan que, de los 682 nodos que componen al grafo completo, 613 nodos representan el componente mayor que a su vez están conectados con 1.387 aristas. Dicho componente posee un diámetro de

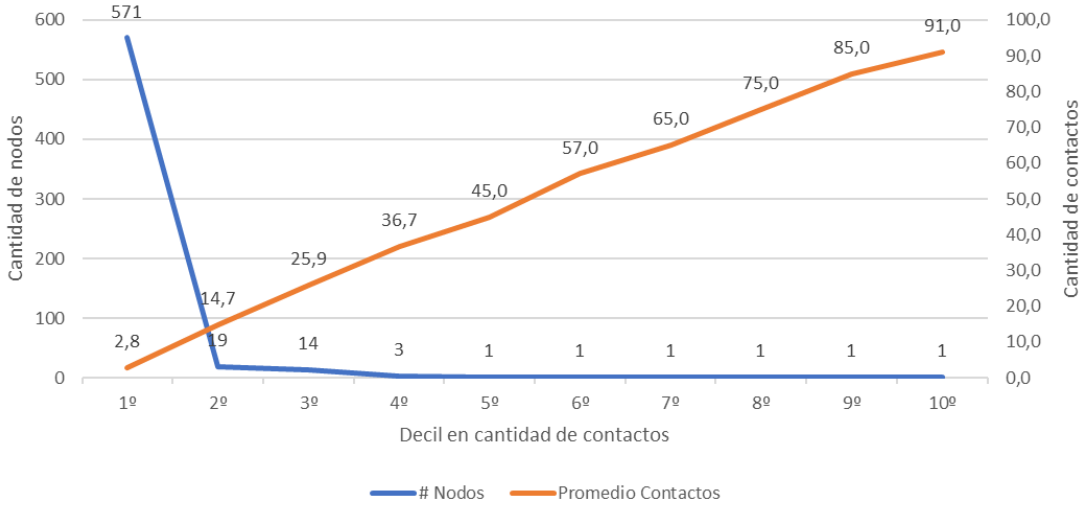

Gráfico 7. Distribución de nodos y contactos por decil de contactos.

14, un largo promedio de caminos de 5.44 pasos y una densidad de 0.0074. Mientras que el coeficiente de clustering global es de 0.0784.

Es importante a su vez observar la distribución de nodos por decil en cantidad de contactos. En el gráfico 7 se puede apreciar que la mayoría de los nodos se concentran en el primer decil y poseen muy pocos contactos (2,8 en promedio). Esto se refleja en un quiebre abrupto en el gráfico a partir del segundo decil, con una cantidad de nodos muy inferior y un promedio de contactos superior al primer decil. A partir del quinto decil los deciles pasan a estar conformados por un nodo únicamente, siempre con una creciente cantidad de contactos.

De esta forma se puede afirmar que la mayoría de los nodos poseen muy pocos contactos entre sí, mientras que una pequeña cantidad de nodos concentran una gran cantidad de contactos. Se puede afirmar que hay una preferencia de los nodos a conectarse hacia nodos con un alto grado de contactos.

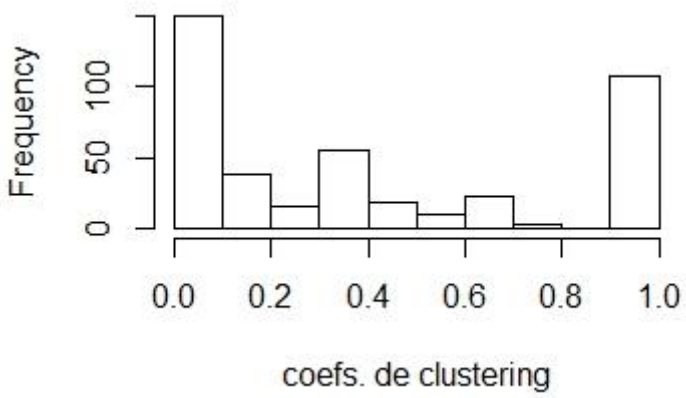

Gráfico 8. Histograma de coeficientes de clústering y frecuencia.

Es útil a su vez, analizar ciertas métricas que permitan corroborar los resultados presentados. A continuación, algunas de las métricas planteadas. Si se observa el gráfico 8 se distingue bastante dispersión en las observaciones, con una gran cantidad acumulada en el primer decil, pocos valores en deciles bajos, y otro pico menor en el último decil de la distribución. Este gráfico es coherente y está en línea con lo mencionado previamente.

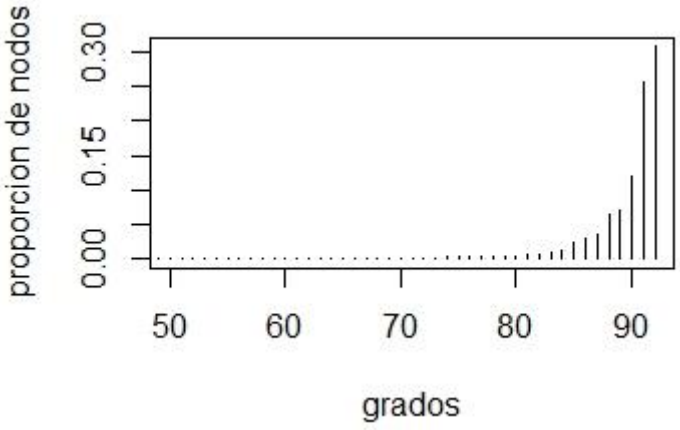

Gráfico 9. Gráfico de distribución por grado.

Del mismo modo, podemos observar la distribución entre grados y proporción de nodos. En el gráfico 9 se ve claramente una distribución con un sesgo pronunciado, habiendo un grupo de nodos con un grado alto y una mayor dispersión en grados inferiores a 90.

Finalmente, en el gráfico 10, se ve la misma distribución en un diagrama de densidad en el que se señala la proporción de nodos con un grado superior a X. En todos estos casos, se pueden obtener iguales conclusiones a las ya mencionadas y por lo tanto se corrobora la distribución y conclusiones observadas en el gráfico 7.

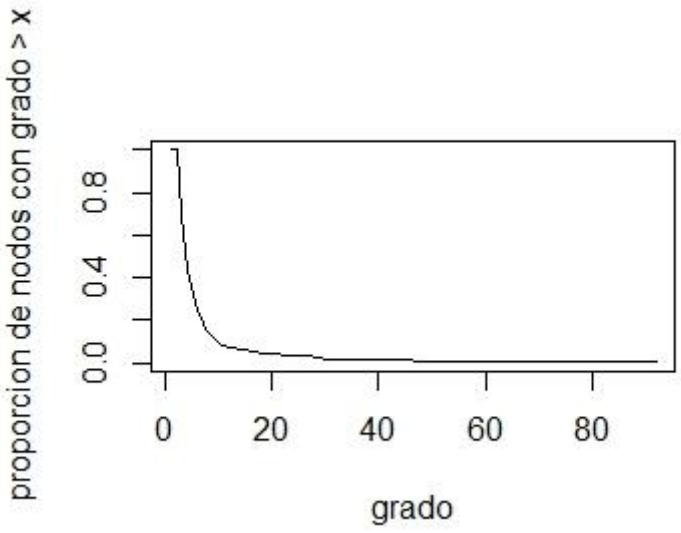

Gráfico 10. Diagrama de densidad con distribución por grado.

Redes de mundo pequeño

Dados los resultados observados previamente, y viendo que la distribución del grafo podría asimilarse a la de una ley de potencias, se comprueba si el mismo se trata de una red de mundo pequeño. El definir qué tipo de red se trata puede ser de utilidad no sólo para caracterizarla, sino para comprender con mayor profundidad su funcionamiento y distribución. Para evaluar este punto se utilizarán dos métodos diferentes, los cuales fueron explicados en detalle en el apartado de "Materiales y métodos". A continuación, se comentarán únicamente los resultados observados y las conclusiones obtenidas a partir de los mismos.

Las primeras métricas son consistentes con los parámetros observados en distribuciones de ley de potencias. Se obtiene el valor de alfa igual a 2.4286; por lo que al ser mayor a la unidad es correcto. La prueba de ajuste de Kolmogorov-Smirnov es de 0.9238 por lo que no es significativo y no puede rechazarse la hipótesis nula. Por último, el xmin es de 5, lo que indica que es un valor aceptable siendo el mismo relativamente pequeño. Todos estos indicadores son consistentes y contribuyen con la asociación del grafo analizado a una red de mundo pequeño.

Luego, se prueba que los coeficientes de clustering de la red son mayores que grafos al azar con características topológicas similares. Para esto se simulan 1000 redes al azar, tomando como prueba sus interacciones. Se crean grafos al azar que siguen el modelo de Barabási–Albert, luego se generan grafos de acuerdo con el modelo de

Erdos-Renyi. En el primer caso se utilizan la cantidad de nodos y el alfa obtenido como parámetros del modelo, así como se itera el valor de m hasta obtener un número de aristas similar al grafo original que fue dado por el m igual a 2. En el segundo caso el argumento es el número de nodos y de aristas directamente.

En forma visual, se puede apreciar en el gráfico 11 del lado izquierdo que el coeficente de clustering del grafo (línea roja) se encuentra dentro del rango de valores de grafos simulados siguiendo un modelo de red libre de escala.

En el mismo gráfico, del lado derecho, se observa que, por el contrario, el promedio de los coeficientes de clustering del grafo es significativamente mayor que los de grafos que siguen modelos al azar, tal como se espera para un grafo de mundo pequeño.

Por lo tanto, basados en el ajuste a una ley de potencias y la distribución de los valores de coeficientes de clustering no podemos descartar la hipótesis de que se trate de un grafo de mundo pequeño.

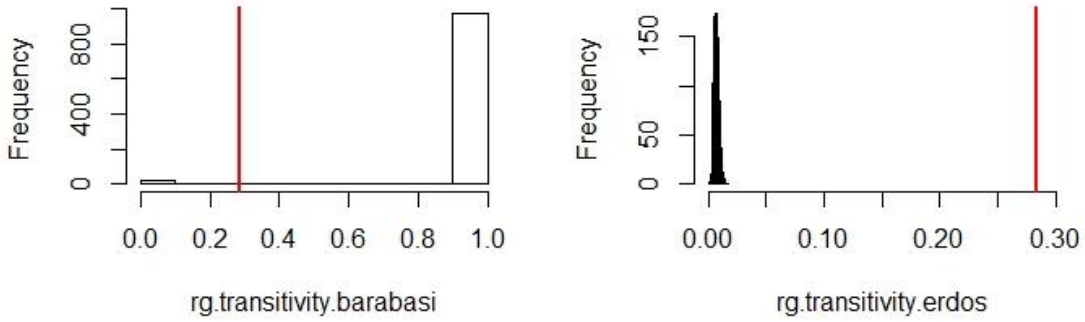

Gráfico 11. Izquierda: Grafo de Barabási-Albert y coeficiente de clústering (línea roja). Derecha: Modelo de grafo al azar y coeficiente de clústering (línea roja).

Por último, se obtuvo el coeficiente de asortatividad de grado sobre el componente gigante con el objeto de comprender mejor la conectividad del grafo. El mismo dio como resultado -0.3745; lo que indica que si bien no se trata de una red asortativa hay cierto grado de disortatividad. Es decir, en cierta medida los nodos de alto grado, en promedio, están conectados con otros nodos de grado bajo mientras que nodos de grado bajo están, en promedio, conectados con nodos de grado alto.

Atributos del grafo

Resulta interesante señalar algunas métricas y características asociadas al componente principal del grafo mencionado. Debiendo analizar cada componente por separado, se consideró únicamente el componente principal al poseer la mayor cantidad de nodos asociados, y por ende la mayor capacidad descriptiva del mismo. Los componentes menores no asociados al principal deberían verse por separado individualmente, pero, al ser pocos y de tamaño pequeño, escapan al presente trabajo.

Analizando algunas métricas asociadas a la cantidad de contactos dados a partir de cada nodo por rol, es posible mencionar que los usuarios con el rol de “Profesor-Editor” son quienes poseen, en promedio, la mayor cantidad de contactos. Si bien, por otro lado, también poseen una alta variabilidad, teniendo este grupo el mayor valor de desvío estándar. El grupo de “Alumnos”, siendo el mayoritario, posee una cantidad promedio y desvío estándar de contactos cercana a la media total. Sin embargo, la variabilidad total

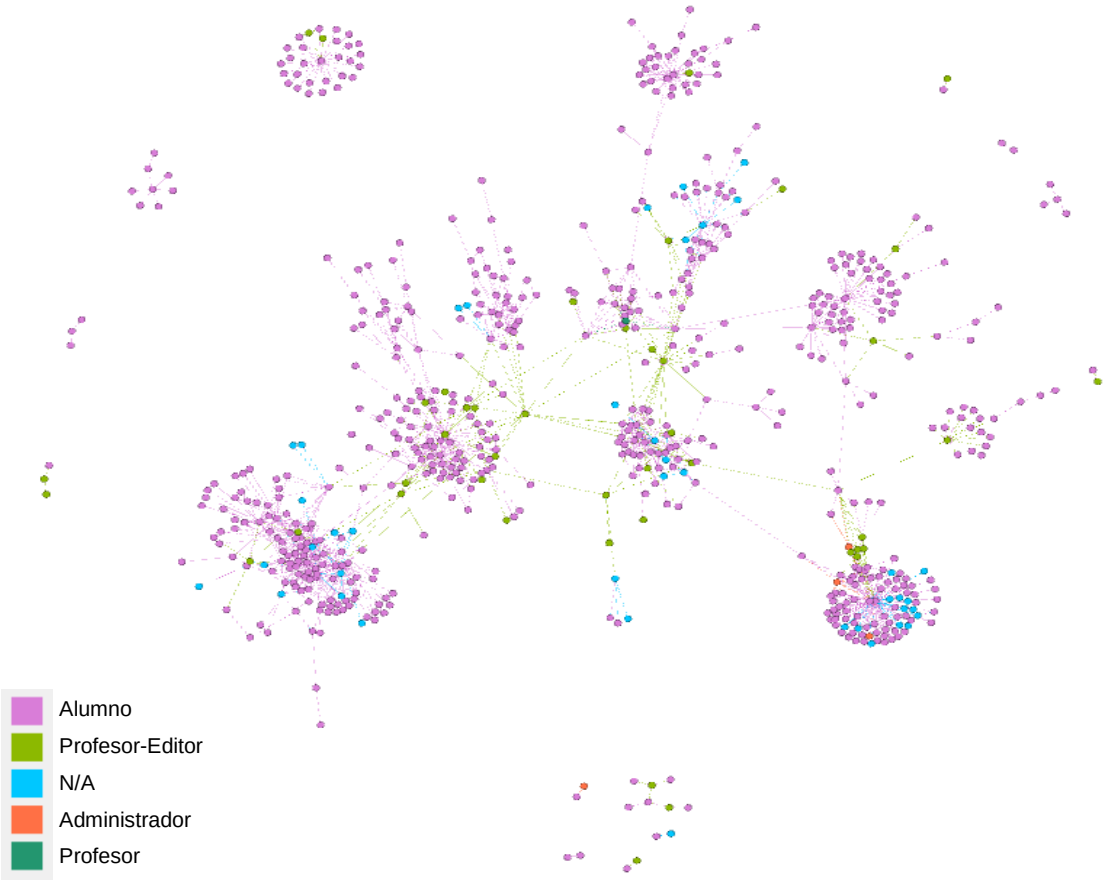

Gráfico 12. Grafo de usuarios conectados por al menos un mensaje enviado/recibido (en distinto color cada rol de usuario).

de este grupo es mayor ya que existe al menos un usuario que posee hasta 91 contactos en total. Los grupos de "Administradores" y "Profesor" son muy minoritarios, mientras que existen 40 usuarios dentro del grafo sobre los que no se pudo identificar el rol.

Rol	# Nodos	Prom. Contactos	Desv. Contactos	Mín. Contactos	Máx. Contactos
Administrador	3	5,3	4,16	2	10
Alumno	535	4,4	8,35	1	91
N/A	40	3,2	4,26	1	27
Profesor	1	3,0	-	3	3
Profesor-Editor	34	7,6	11,68	1	65
Total general	**613**	**4,5**	**8,37**	**1**	**91**

Tabla 13. Nodos y métricas asociadas a contactos por rol de usuario (en orden: promedio de contactos, desvío estándar de contactos, mínima cantidad de contactos, máxima cantidad de contactos).

Es posible analizar además la desagregación de métricas por roles en aquellos que poseen al menos un certificado de aprobación, y aquellos que no poseen ningún certificado. Como primera observación es de destacar que tres de los 34 usuarios con el rol de "Profesor-Editor" poseen además al menos un certificado de aprobación. Se trata seguramente de docentes de algún curso que realizaron a su vez la certificación de otro. Es importante señalar a su vez que dicho grupo es el que, en promedio, posee la mayor cantidad de contactos, incluso por encima de quienes ostentan el mismo rol pero que no obtuvieron certificados. Una posible explicación tiene que ver con el hecho

de contar con contactos tanto como docente y como alumno. El desvío estándar de este grupo es relativamente más bajo comparado tanto con usuarios con certificado como con los dos más grandes grupos de usuarios sin certificado de aprobación (“Alumno” y “Profesor-Editor).

Los usuarios con el rol de “Alumno” que obtuvieron al menos un certificado poseen un promedio de contactos superior a los que no obtuvieron ninguno, con un desvío estándar menor que este grupo. Ambos grupos contienen una cantidad similar de nodos (263 y 272 nodos respectivamente), si bien la cantidad máxima de contactos en aquellos que no obtuvieron certificados es sensiblemente mayor respecto a los que sí obtuvieron al menos un certificado de aprobación (un máximo de 91 contactos en el primer grupo, contra un máximo de 57 en el segundo).

Rol	# Nodos	Prom. Contactos	Desv. Contactos	Mín. Contactos	Máx. Contactos
Con Certificado	**266**	**5,1**	**7,49**	**1**	**57**
Alumno	263	5,0	7,49	1	57
Profesor-Editor	3	10,7	7,23	6	19
Sin Certificado	**347**	**4,1**	**8,97**	**1**	**91**
Alumno	272	3,9	9,08	1	91
N/A	40	3,2	4,26	1	27
Profesor-Editor	31	7,3	12,06	1	65
Administrador	3	5,3	4,16	2	10
Profesor	1	3,0	-	3	3
Total general	**613**	**4,5**	**8,37**	**1**	**91**

Tabla 14. Nodos y métricas asociadas a contactos diferenciados con y sin certificado de aprobación, y por rol de usuario (en orden: promedio de contactos, desvío estándar de contactos, mínima cantidad de contactos, máxima cantidad de contactos).

Usuarios y cursos

Características topológicas del grafo

En este apartado se observarán los resultados que se obtienen al analizar el grafo de usuarios conectados a través de haber realizado el mismo curso. Es decir, si dos usuarios cualesquiera realizaron al menos un curso en común, entonces estarán conectados, si no poseen ningún curso en común no lo estarán. Para esta definición no se tomó en consideración el resultado al que arribaron dentro de cada curso. Es decir, no fue tenido en cuenta ni se hizo diferenciación entre usuarios que obtuvieron, o no, algún tipo de certificación por el curso que los uniera. Podría darse el caso de dos usuarios que participaron del mismo curso donde uno haya aprobado y obtenido el certificado de aprobación, mientras el otro tan solo haya asistido a unas pocas clases. A los fines de este análisis, estos usuarios estarán igualmente conectados y no se otorgó ningún tipo de peso especial a las aristas que los conectan con el fin de simplificar los diferentes análisis a los que fueron sometidos. De igual manera, se definió a este grafo como no dirigido.

Es de destacar que el grafo conformado posee un total de 5.270 nodos que están conectados a través de 2.375.264 aristas. Se observa, por tanto, un grafo

significativamente más grande que el anterior ya que cubre prácticamente a la totalidad de los usuarios presentes dentro de la base de datos.

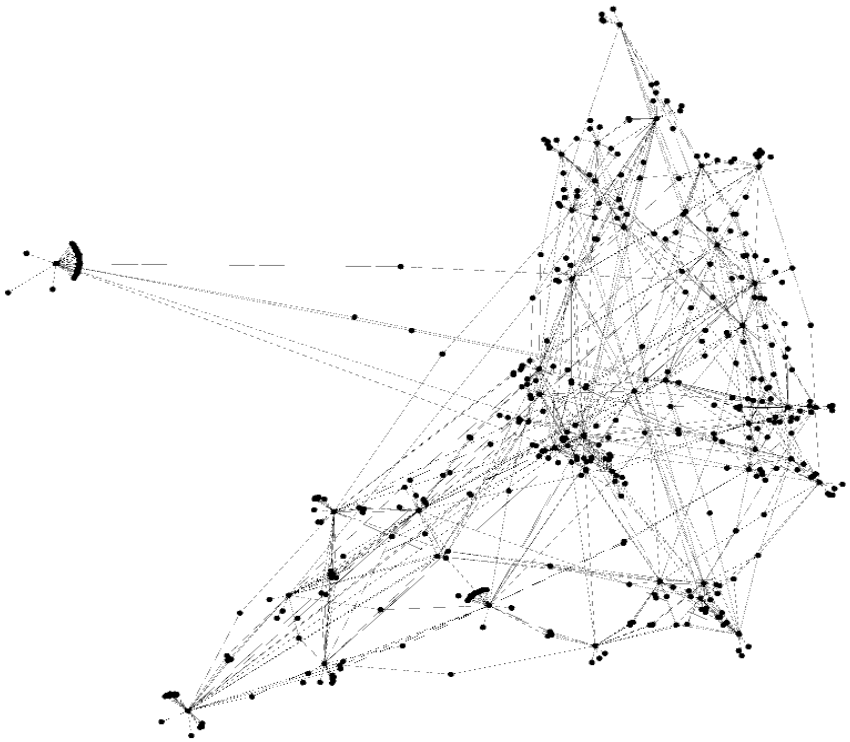

Gráfico 15. Grafo de usuarios conectados por cursos que compartieron.

Se trata en este caso de una red completamente conectada. Esto implica que, si bien todos los usuarios que comparten un mismo curso están conectados, también es cierto que existe siempre al menos un usuario que conecte a usuarios de distintos cursos.

En cuanto a sus características más duras, es posible mencionar que este grafo posee un diámetro de cuatro, un largo promedio de caminos de 2.54 pasos y una densidad de 0.0855. Mientras que el coeficiente de clustering global es igual a 0.9790.

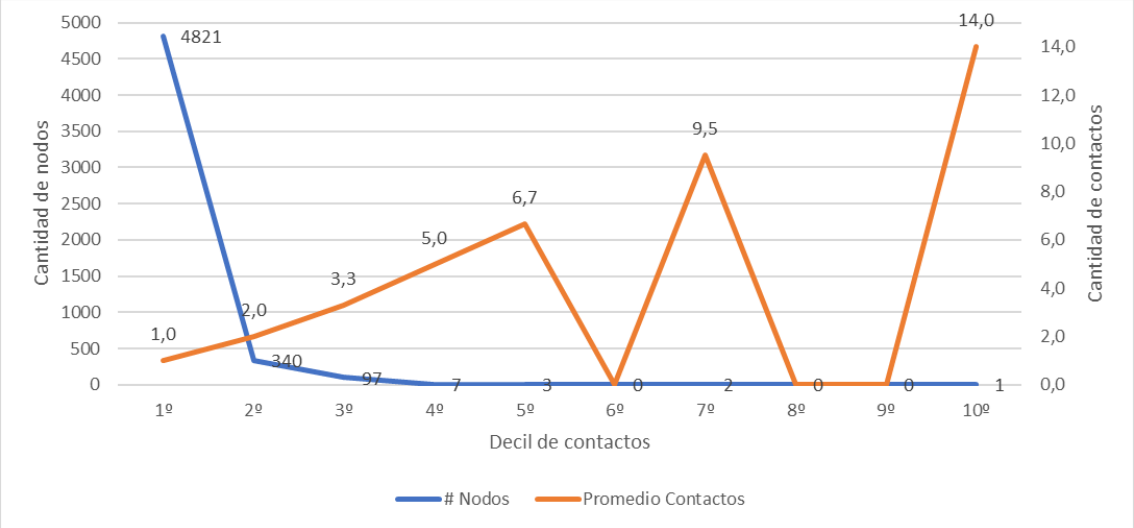

Gráfico 16. Distribución de nodos y contactos por decil de contactos.

Es importante a su vez observar la distribución de nodos por decil en cantidad de contactos. En el gráfico 16 se puede apreciar que la mayoría de los nodos se concentran en el primer decil y poseen muy pocos contactos (1 en promedio). Esto se refleja en un quiebre abrupto en el gráfico a partir del segundo decil, con una cantidad de nodos muy inferior y un promedio de contactos superior al primer decil. A partir del cuarto, los deciles pasan a estar conformados por muy pocos nodos, habiendo saltos entre deciles con grupos sin ningún nodo, siempre con una creciente cantidad de contactos. De esta forma es posible afirmar que la mayoría de los nodos poseen muy pocos contactos entre sí, mientras que una muy pequeña cantidad de nodos poseen una mayor cantidad de contactos (la mayoría entre 2 y 3 contactos, y muy pocos nodos con 5 o más contactos).

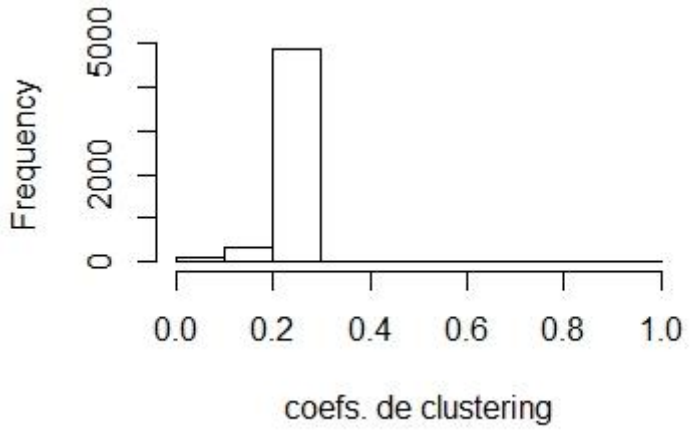

Gráfico 17. Histograma de coeficientes de clústering y frecuencia.

Es útil a su vez, analizar ciertas métricas que permitan corroborar los resultados presentados. A continuación, algunas de las métricas planteadas. Si se observa el gráfico 17 se puede distinguir que prácticamente la totalidad de los valores están concentrados en el tercer decil de la distribución, con muy pocas observaciones en los primeros dos deciles y prácticamente ninguno en el resto hacia la derecha. Esto indica un valor general bajo de concentramiento en la conexión de los vértices dentro del grafo.

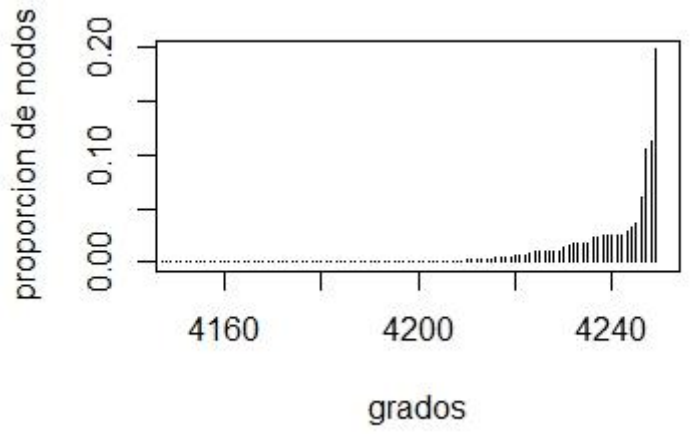

Gráfico 18. Gráfico de distribución por grado.

Del mismo modo, es posible apreciar la distribución entre grados y proporción de nodos con el objeto de evaluar la consistencia con los resultados ya mencionados. En

el gráfico 18 se ve una distribución con un sesgo extremadamente pronunciado y valores de grados sumamente altos. Este resultado está en línea con lo observado previamente y vuelve a validar la idea de que existe una gran cantidad de nodos con bajo grado de conexión, y una muy pequeña cantidad de nodos con un grado muy alto.

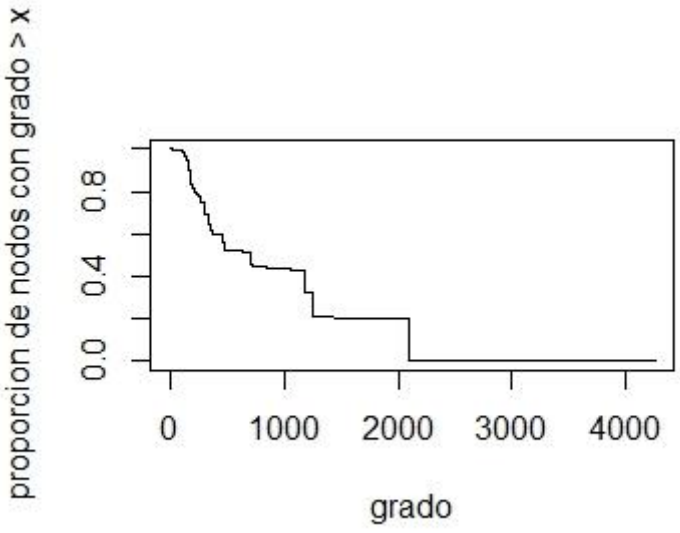

Gráfico 19. Diagrama de densidad con distribución por grado.

Por último, en el gráfico 19 se ve la misma distribución en un diagrama de densidad en el que se señala la proporción de nodos con un grado superior a X. Para este grado la variación en la curva es mucho menos pronunciada que para el grafo visto anteriormente, el cual marcaba una abrupta pendiente en grados bajos. Se aprecia también que los grados de los nodos son significativamente más altos que lo visto en el grafo anterior.

Estos resultados sirven para corroborar que la mayor proporción de nodos poseen un grado muy bajo, pero sin embargo existe unos pocos nodos con un grado extremadamente alto. El gran tamaño del grafo hace que el grado se incremente en forma muy rápida para nodos con una alta conectividad hacia el resto de los nodos.

Redes de mundo pequeño

Dados los resultados observados previamente, y viendo que la distribución del grafo podría asimilarse a la de una ley de potencias, se comprueba si el mismo se trata de una red de mundo pequeño. El definir qué tipo de red se trata puede ser de utilidad no sólo para caracterizarla, sino para comprender con mayor profundidad su funcionamiento y distribución. Para evaluar este punto se utilizarán dos métodos diferentes, los cuales fueron explicados en detalle en el apartado de “Materiales y métodos”. A continuación, se comentarán únicamente los resultados observados y las conclusiones obtenidas a partir de los mismos.

Las primeras métricas son consistentes con los parámetros observados en distribuciones de ley de potencias. Se obtiene el valor de alfa igual a 11.2737; por lo que al ser mayor a la unidad es correcto. El test de ajuste de Kolmogorov-Smirnov es igual a 0.7863, por lo que no es significativo y no puede rechazarse la hipótesis nula. Por último, el xmin es de 2.428 siendo éste un valor extremadamente elevado, acorde a los grados vistos en el grafo. Si bien este último indicador es bastante elevado, en valores

relativos al tamaño total del grafo representa una proporción relativamente pequeña (0.1%). Por otro lado, los otros dos indicadores son consistentes y contribuyen con la asociación del grafo analizado a una red de mundo pequeño.

Luego, se prueba que los coeficientes de clustering de la red son mayores que grafos al azar con características topológicas similares. Para esto se simulan 1000 redes al azar, tomando como prueba sus interacciones. Se crean grafos al azar que siguen el modelo de Barabási–Albert, luego se generan grafos de acuerdo con el modelo de Erdos-Renyi. En el primer caso se utilizan la cantidad de nodos y el alfa obtenido como parámetros del modelo, así como se itera el valor de m hasta obtener un número de aristas similar al grafo original que fue dado por el m igual a 470. En el segundo caso el argumento está dado por el número de nodos y de aristas directamente.

En forma visual, se puede apreciar en el gráfico 20 del lado izquierdo, que el coeficente de clustering del grafo (línea roja) se encuentra dentro del rango de valores de grafos simulados siguiendo un modelo de red libre de escala.

En el mismo gráfico, del lado derecho, se observa que, por el contrario, el promedio de los coeficientes de clustering del grafo es significativamente mayor que los de grafos que siguen modelos al azar, tal como se espera para un grafo de mundo pequeño.

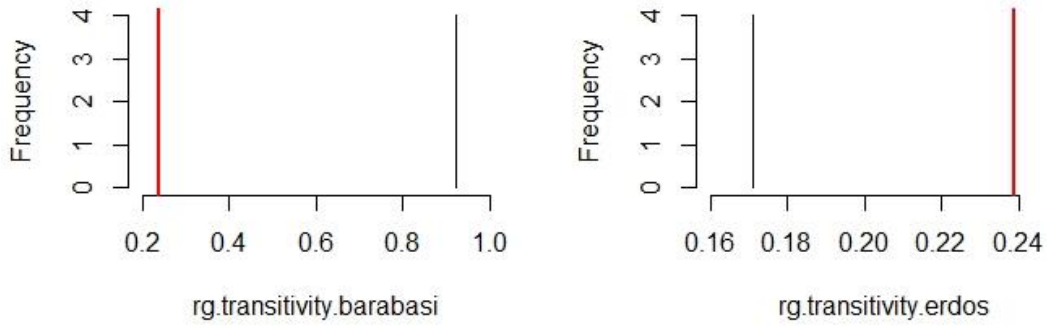

Gráfico 20. Izquierda: Grafo de Barabási-Albert y coeficiente de clústering (línea roja). Derecha: Modelo de grafo al azar y coeficiente de clústering (línea roja).

Por lo tanto, basados en el ajuste a una ley de potencias y la distribución de los valores de coeficientes de clustering no podemos descartar la hipótesis de que se trate de un grafo de mundo pequeño.

Por último, se obtuvo el coeficiente de asortatividad de grado con el objeto de comprender mejor la conectividad del grafo. El mismo dio como resultado 0.9082; lo que indica que si bien se trata de una red asortativa. Es decir, los nodos de alto grado, en promedio, están conectados con otros nodos de grado alto mientras que nodos de grado bajo están, en promedio, conectados con nodos de grado bajo.

Atributos del grafo

Analizando algunas métricas asociadas a la cantidad de contactos diferenciados a partir de cada nodo por rol, es posible ver que los usuarios con el rol de “Administrador” son quienes poseen, en promedio, la mayor cantidad de contactos. Si bien también poseen una alta variabilidad, teniendo este grupo el mayor valor de desvío estándar luego del grupo con el rol de “Profesor-Editor”. Sin embargo, este último grupo es mucho

más numeroso, mientras que hay sólo 4 nodos identificados con el rol de "Administrador".

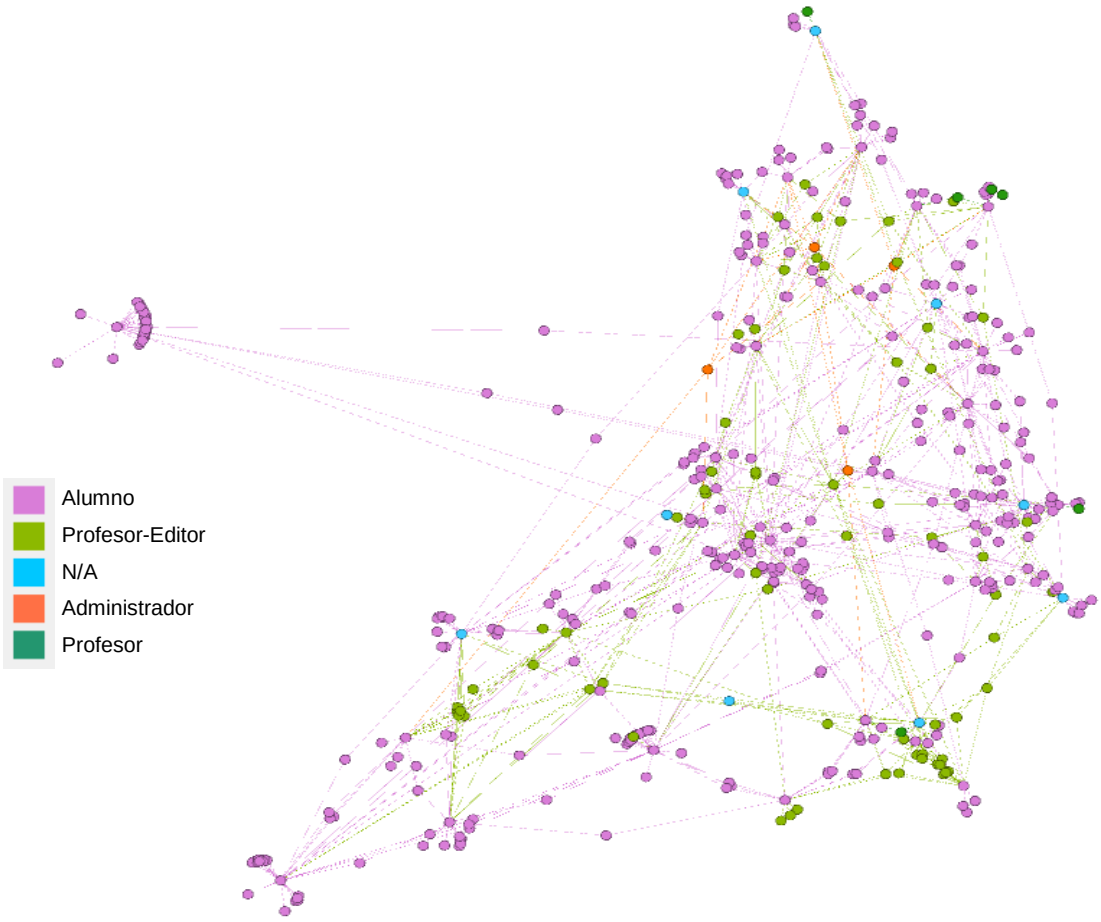

Gráfico 21. Grafo de usuarios conectados por cursos que compartieron (en distinto color cada rol de usuario).

El grupo de "Alumnos", siendo el mayoritario, posee una cantidad promedio de contactos cercana a la media total, con un desvío estándar por debajo de la media total. La variabilidad total de este grupo es relativamente baja.

Rol	# Nodos	Prom. Contactos	Desv. Contactos	Mín. Contactos	Máx. Contactos
Administrador	4	5,8	1,50	4	7
Alumno	5126	1,1	0,31	1	4
Profesor	10	1,3	0,48	1	2
Profesor-Editor	131	2,5	1,90	1	14
Total general	**5271**	**1,1**	**0,50**	**1**	**14**

Tabla 22. Nodos y métricas asociadas a contactos por rol de usuario (en orden: promedio de contactos, desvío estándar de contactos, mínima cantidad de contactos, máxima cantidad de contactos).

El grupo "Profesor-Editor" es uno de los más numerosos y posee la más alta dispersión en la cantidad de contactos, con valores de dispersión cercanos al grupo de "Administradores".

Todos estos resultados son consistentes con una gran cantidad de cursos en los que los docentes pueden participar en muchos de ellos, pero los alumnos están presentes sólo en algunos pocos.

Es posible analizar además la desagregación de métricas por roles en aquellos que poseen al menos un certificado de aprobación, y aquellos que no poseen ningún certificado. Como primera observación es de destacar la presencia de tres usuarios con el rol de “Profesor-Editor” que poseen al menos un certificado de aprobación. Dicho grupo es el que, en promedio, posee la mayor cantidad de contactos, incluso por encima de quienes ostentan el mismo rol pero que no obtuvieron certificados. Una posible explicación tiene que ver con el hecho de contar con contactos tanto como docente y como alumno. El desvío estándar de este grupo es también el más alto de todos los grupos con y sin certificado de aprobación.

Rol	# Nodos	Prom. Contactos	Desv. Contactos	Mín. Contactos	Máx. Contactos
Con Certificado	**828**	**1,3**	**0,67**	**1**	**10**
Alumno	825	1,2	0,52	1	4
Profesor-Editor	3	7,7	3,21	4	10
Sin Certificado	**4443**	**1,1**	**0,46**	**1**	**14**
Alumno	4301	1,1	0,25	1	3
Profesor-Editor	128	2,4	1,70	1	14
Profesor	10	1,3	0,48	1	2
Administrador	4	5,8	1,50	4	7
Total general	**5271**	**1,1**	**0,50**	**1**	**14**

Tabla 23. Nodos y métricas asociadas a contactos diferenciados con y sin certificado de aprobación, y por rol de usuario (en orden: promedio de contactos, desvío estándar de contactos, mínima cantidad de contactos, máxima cantidad de contactos).

Los usuarios con el rol de “Alumno” que obtuvieron al menos un certificado poseen un promedio de contactos levemente superior a los que no obtuvieron ninguno, con un desvío estándar superior que este último grupo. Por otro lado, quienes poseen un certificado representan una cantidad mucho menor de usuarios, cercano al 15% del total de “Alumnos” (825 alumnos con certificado y 4301 alumnos sin certificado).

No existen usuarios “Profesor” o “Administrador” con algún certificado otorgado. Estos dos grupos son de los menos numerosos y poseen una cantidad promedio de contactos diferente entre sí. Mientras que los “Administradores” son el segundo grupo con mayor cantidad promedio de contactos, los “Profesores” poseen una cantidad de contactos promedio bastante cercana al promedio de los “Alumnos”. Estos resultados son razonables entendiendo que los “Administradores” estarán vinculados a más cursos en su rol de administrador, mientras que es sensato que los “Profesores” tengan uno o pocos cursos asociados en su rol docente.

MODELADO ESTADÍSTICO

Hasta este punto, todos los análisis se focalizaron en características descriptivas de cada uno de los grafos. A continuación, se presentarán distintos modelos estadísticos diseñados específicamente para el análisis de grafos. Se hará la presentación y desarrollo de análisis y modelos implementados en R siguiendo lo desarrollado en [10].

En la sección de “Materiales y métodos” se encuentra en forma detallada el desarrollo de cada modelo planteado a continuación.

El objetivo de utilizar estos modelos estadísticos radica en, por un lado, encontrar modelos capaces de ajustar los grafos presentados y así encontrar variables que expliquen su forma y características. Por otro lado, estas variables y características encontradas que los definen pueden dar un indicio de qué palancas se pueden ajustar, modificando así la composición final del grafo, logrando los objetivos buscados.

Para el armado de las entidades sobre las que se aplicarán los modelos se utilizó la base de usuarios conectados a través de mensajes. Se utilizó la matriz de adyacencia dada por los usuarios conectados a través de al menos un mensaje perteneciente al componente principal del grafo, así como características a nivel usuario. Estas características fueron recopiladas de distintas tablas a fin de constituir un perfilado único por usuario.

Se buscó predecir la distribución del grafo y las conexiones dadas a partir de las siguientes variables a nivel usuario:

- Meses desde la creación del usuario.
- Rol del usuario.
- Cantidad de cursos en los que está inscripto el usuario.
- Cantidad de certificados de aprobación que posee el usuario.

Por lo tanto, será posible identificar qué variables son las que mejor predicen una mayor asociación entre usuarios y, de esta manera, preveer acciones para que haya una mayor conectividad entre los mismos con el fin de lograr un mayor compromiso con los cursos del INC. Por otro lado, en caso de que esto no fuera posible, se logrará de todas formas una mejor caracterización de los grafos y usuarios, entendiendo de esta manera si los mismos poseen un comportamiento comparable con el de otros estudios en condiciones similares.

A modo de resumen, se utilizarán a continuación dos tipos de modelos distintos, que corresponden a tipos de modelos estadísticos conocidos para bases de datos no pertenecientes a grafos. En primer lugar, se utilizará un modelo de la familia de modelos exponenciales de grafos aleatorios. Dichos modelos son análogos a los modelos de regresión estándar, y en particular, a modelos lineales generalizados. Por otro lado, se utilizarán modelos de red latente. Los modelos de red latente son una variante basada en redes del uso de variables observadas y no observadas (latentes) en el modelado de un resultado (en este caso, la presencia o ausencia de vértices en la red).

Modelos exponenciales de grafos aleatorios

Especificación de un modelo

A partir del modelo ya definido en “Materiales y métodos”, se utilizó la implementación del modelo ERGM a través de la librería *ergm* en R.

Dentro de dicho paquete, se representó la transitividad mediante la utilización del término *gwesp*. Dicho término puede traducirse como “distribución de socios

compartidos de borde geométrico ponderado" y utiliza un factor de decaimiento para modelar el efecto de la transitividad entre nodos dentro del grafo. Por otro lado, se representaron los efectos primarios y los efectos de segundo orden a partir de los términos *nodemain* y *match*, respectivamente.

A continuación, la fórmula del modelo utilizado expresado en R. El mismo permite controlar tanto la densidad de la red (estadístico *edges*), así como efectos de transitividad (estadístico *gwesp*). Este modelo permite evaluar el efecto en la formación de lazos colaborativos por antigüedad en meses de la creación del usuario (estadístico *nodemain* sobre la variable *meses_creacion*), por rol (estadístico *nodemain* sobre la variable *roleid*), por cantidad de cursos a los que está asociado (estadístico *nodemain* sobre la variable *cursos*), por cantidad de certificados de aprobación obtenidos (estadístico *nodemain* sobre la variable *certificados*); así como también el poseer el mismo rol en común entre pares de nodos (estadístico *match* sobre la variable *roleid*).

```
Formula:
us_us_2_adj.s ~ edges + gwesp(1, fixed = TRUE) + nodemain("meses_creacion")
+ nodemain("roleid") + nodemain("cursos") + nodemain("certificados")
 + match("roleid")
```

Ajuste del modelo

Es necesario a continuación, evaluar si el modelo planteado presenta un buen ajuste a la distribución real de los datos. Para esto, se realiza un análisis de la varianza (ANOVA).

```
Model 1: us_us_2_adj.s ~ edges + gwesp(1, fixed = TRUE) + nodemain("meses_c
reacion") + nodemain("roleid") + nodemain("cursos") + nodemain("certificado
s") + match("roleid")
```

A partir del análisis de la varianza se puede observar que existe una fuerte evidencia que las variables utilizadas en el modelo explican la variación de la conectividad de la red, con un cambio en la desviación de 241 con sólo siete variables de entrada (variables explicadas en la especificación del modelo).

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>|Chisq|)
NULL			187,6	260	
Model 1	7	241,2	187,6	18,8	< 2.2e-16 ***
*Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*					

Tabla 24. Salida del Análisis de la Varianza (ANOVA) sobre el Modelo 1.

En forma análoga es posible examinar la contribución relativa de cada variable dentro del modelo a partir del análisis de ajuste de este. Es posible interpretar el coeficiente estimado de cada estadístico como el log-odds ratio condicional de contacto entre usuarios. Esta interpretación se debe hacer en forma marginal y suponiendo que no existen cambios en el valor de otras variables.

Se puede observar que, por ejemplo, la antigüedad en meses de creación del usuario aumenta las posibilidades de que haya contacto entre usuarios en un 0.42%, la cantidad de cursos a los que está asociado el usuario reduce las posibilidades de que haya contacto en un 0.5%, y la cantidad de certificados de aprobación del usuario aumenta la posibilidad de que haya contacto en un 4.2%. En forma similar, el poseer el mismo rol disminuye la posibilidad de que haya contacto entre usuarios en un 29.49%.

Es importante destacar que todos los estadísticos utilizados fueron señalados como significativos para el modelo, a excepción del estadístico relacionado a los cursos asociados al usuario.

En forma similar, la magnitud del coeficiente del estadístico *gwesp* (2.0882) y el relativamente bajo error estándar correspondiente indican que hay evidencia de un efecto de transitividad no trivial. Se debe destacar que con la inclusión de estadísticos de "segundo orden" en el modelo, la cuantificación de este efecto controla naturalmente la similaridad básica de dichos atributos. Por lo tanto, existe seguramente algún otro atributo además de similaridad de rol que no fue controlado, o un posible proceso social de formación de grupos.

	Estimate	Std. Error	MCMC %	z value	Pr(>|z|)
edges	-9.2448	0.1795	1	-51.51	<1e-04 ***
gwesp.fixed.1	2.0882	0.0560	1	37.29	<1e-04 ***
nodecov.meses_creacion	0.0042	0.0010	4	4.21	<1e-04 ***
nodecov.roleid	0.0722	0.0082	4	8.76	<1e-04 ***
nodecov.cursos	-0.0050	0.0060	3	-0.83	0.404
nodecov.certificados	0.0412	0.0092	3	4.49	<1e-04 ***
nodematch.roleid	-0.3494	0.0342	5	-10.21	<1e-04 ***
*Signif. Codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1*					
Null Deviance: 260038 on 187578 degrees of freedom Residual Deviance: 18776 on 187571 degrees of freedom					
AIC: 18790 BIC: 18861					

Tabla 25. Salida del ajuste del modelo luego de 20 iteraciones. Resultados obtenidos a partir de la estimación máximo verosímil de Monte Carlo.

Bondad de ajuste

En cualquier problema de modelado, el mejor ajuste elegido dentro de una clase de modelos no es necesariamente un buen ajuste a los datos si la clase de modelo elegido no contiene un conjunto lo suficientemente rico de modelos sobre los cuales elegir. El concepto de bondad de ajuste es, por lo tanto, de suma importancia. Sin embargo, si

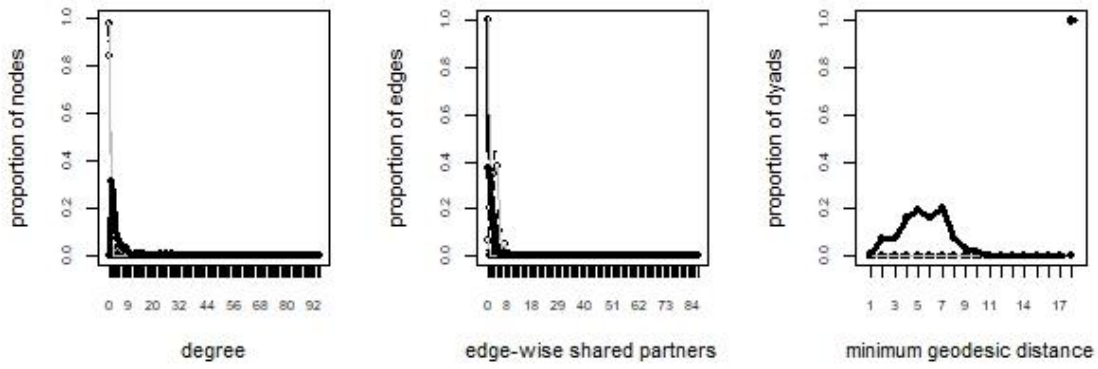

Gráfico 26. Bondad de ajuste comparada entre la red de contactos y los datos obtenidos por Monte Carlo sobre dicho modelo. Las comparaciones se basan en la distribución por grado (degree), cantidad de contactos compartidos por vértice (edge-wise shared partners), y longitud geodésica mínima (minium geodesic distance); representados por gráficos de cajas y bigotes. Los valores correspondientes a la red de contactos están graficados con líneas gruesas. En la distribución de distancias geodésicas entre pares, el gráfico de más a la derecha está separado y corresponde a la proporción de pares no alcanzables.

bien este concepto está bastante desarrollado en contextos estándar de modelado como por ejemplo son los modelos lineales, se encuentre posiblemente en sus inicios en lo concerniente a modelos sobre grafos de redes, según lo que se señala en [10].

Dado el modelo mencionado, para obtener la bondad de ajuste de éste se utilizará la función *gof* presente en el paquete *ergm*. Dicha función ejecuta simulaciones de Monte Carlo y calcula las comparaciones con el grafo original en términos de grado, longitud geodésica, y cantidad de contactos compartidos por cada par de vértices.

A partir de la se observación de los resultados de la bondad de ajuste del modelo, y dada la gran cantidad de nodos existententes dentro del grafo y la distribución de éstos, pueden observarse distribuciones que se encuentran relativamente sesgadas con proporciones altas de nodos y vértices únicamente para valores bajos en grado y en contactos compartidos por vértice, respectivamente. A modo de resumen, se puede mencionar que el ajuste del modelo es relativamente bueno.

Modelos de red latente

Especificación y ajuste del modelo

Según lo definido en “Materiales y métodos”, los moledos de red latente permiten capturar los efectos de homofilia o similaridad, así como de distancia. Se utiliza, al igual que para el modelo ERGM, el grafo de usuarios donde el contacto está dado por el envío de mensajes entre ellos. A partir del mismo es posible suponer que el contacto entre usuarios está estimulado, al menos en parte, por la simililaridad en el rol que posee cada usuario (“Alumno”, “Profesor”, “Administrador”). Por otro lado, también se podría suponer que el contacto está dado por la antigüedad dada entre usuarios. Es decir, pensar que dos usuarios tendrán mayor posibilidad de estar conectados si ambos poseen una antigüedad similar, mientras que, si la diferencia en antigüedad entre la creación de los usuarios es mayor, dicha probabilidad disminiurá. De igual manera, otras hipótesis serían que existe mayor conexión si ambos usuarios están asociados a una cantidad similar de cursos, o que existe mayor conexión si ambos obtuvieron una cantidad similar de certificados de aprobación en ellos.

Para la especificación y ajuste del modelo se utiliza el paquete *eigenmodel* de R. Se definen a continuación cinco modelos distintos, que serán luego ajustados y analizados de manera individual.

Para todos los modelos se utilizó un espacio de dos dimensiones ($R = 2$), mil cien muestras de la cadena de Markov ($S = 1100$), de las cuales mil serán las que se eliminarán a partir de las exploraciones iniciales de la cadena de Markov ($burn = 1000$). La definición de estos parámetros, y principalmente de la cantidad de dimensiones, se hizo para mantener la simplicidad del modelo y no agregar dimensiones y complejidad sin un justificativo específico, facilitando luego tanto su interpretación como su exposición. A su vez, se buscó que sea un proceso potente, pero computacionalmente no tan costoso, por lo cual se decidió limitar el tamaño tanto de las muestras generadas como de las que serían posteriormente descartadas.

1. Modelo sin covariables específicas para cada par de nodos.

```
us_us_2_adj.leig.fit1 <- eigenmodel_mcmc(A, R=2, S=1100, burn=1000)
```

2. Modelo con covariable para rol en común ($same.role$).

```
us_us_2_adj.leig.fit2 <- eigenmodel_mcmc(A, same.role, R=2, S=1100, burn=1000)
```

3. Modelo con covariable para similar antigüedad como usuario en la plataforma ($same.ant$).

```
us_us_2_adj.leig.fit3 <- eigenmodel_mcmc(A, same.ant, R=2, S=1100, burn=1000)
```

4. Modelo con covariable para similar cantidad de cursos a los que está asociado el usuario ($same.cur$).

```
us_us_2_adj.leig.fit4 <- eigenmodel_mcmc(A, same.cur, R=2, S=1100, burn=1000)
```

5. Modelo con covariable para similar cantidad de certificados de aprobación obtenidos ($same.cer$).

```
us_us_2_adj.leig.fit5 <- eigenmodel_mcmc(A, same.cer, R=2, S=1100, burn=1000)
```

Para el armado de estos modelos, y con el objeto de incluir los distintos efectos mencionados, fue necesario crear vectores con dicha información y ajustar cada modelo con cada argumento en forma adicional, según cada caso mencionado. A modo de ejemplo: el poseer un rol en común en el segundo modelo; similaridad en la antigüedad como usuario en la plataforma para el modelo tres; similaridad en la cantidad de cursos a los que el usuario está asociado para el modelo cuatro; o similaridad en la cantidad de certificados de aprobación que posee el usuario para el modelo cinco.

Con el objeto de comparar la representación de las redes en cada una de las dos dimensiones latentes para estos modelos, se extraen los vectores propios de cada modelo ajustado.

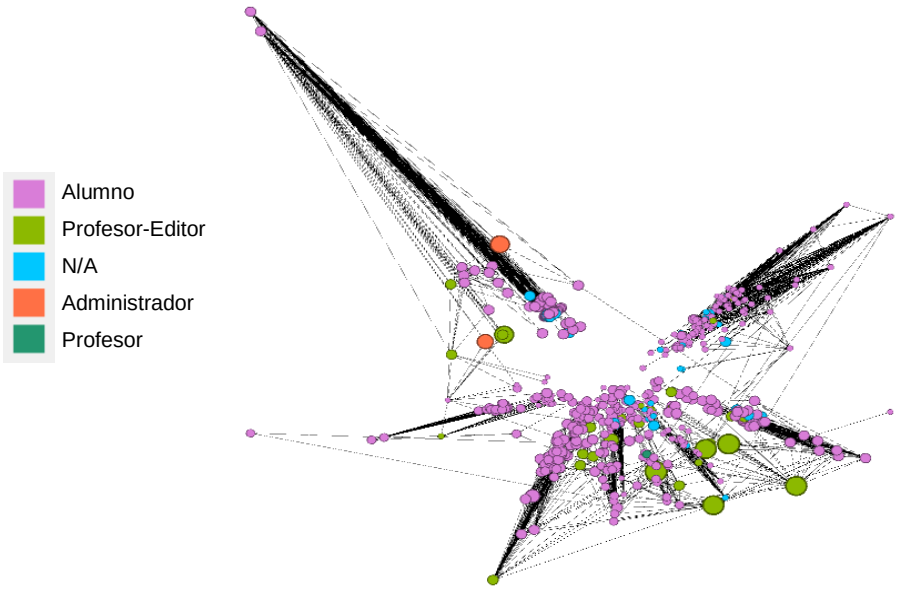

Gráfico 27. (Modelo 1) Grafo de usuarios conectados por al menos un mensaje enviado/recibido. La posición en el plano está dada por las dos dimensiones del primer modelo planteado. En distinto color cada rol de usuario, en distinto tamaño según antigüedad en meses de creación del usuario (a mayor antigüedad mayor tamaño del nodo).

En los gráficos 27 a 31 se puede observar el grafo obtenido según cada uno de los cinco modelos planteados, donde las coordenadas de cada nodo están dadas por los valores propios obtenidos en cada modelo. Para la representación gráfica de los grafos fue necesario hacer un reescalado de los valores obtenidos para cada dimensión, eliminando valores aberrantes en caso de que los hubiere para algún par de coordenadas.

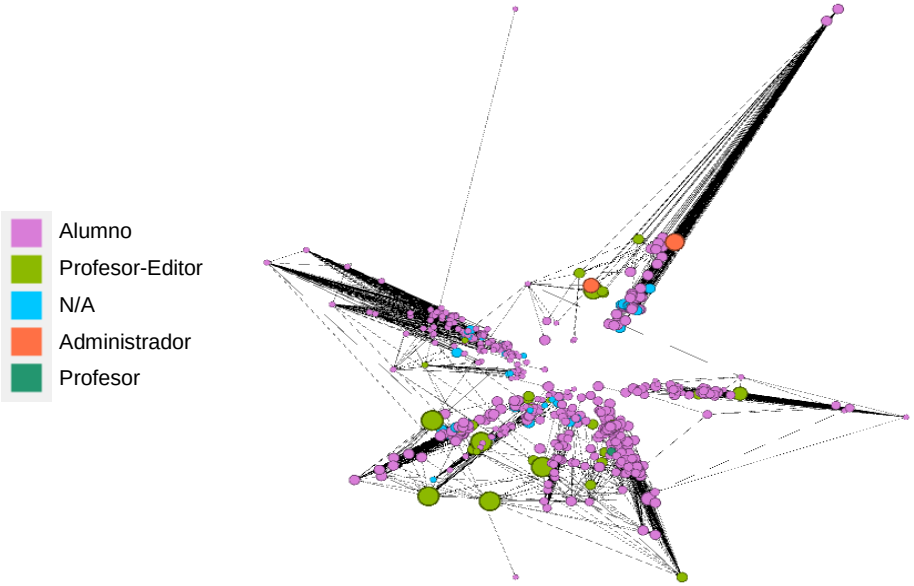

Gráfico 28. (Modelo 2) Grafo de usuarios conectados por al menos un mensaje enviado/recibido. La posición en el plano está dada por las dos dimensiones del segundo modelo planteado. En distinto color cada rol de usuario, en distinto tamaño según antigüedad en meses de creación del usuario (a mayor antigüedad mayor tamaño del nodo).

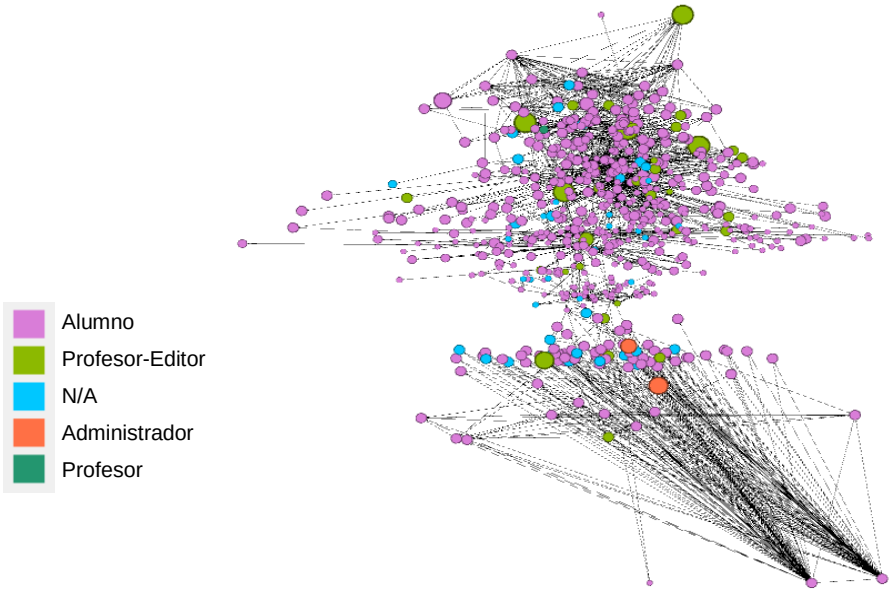

Gráfico 29. (Modelo 3) Grafo de usuarios conectados por al menos un mensaje enviado/recibido. La posición en el plano está dada por las dos dimensiones del tercer modelo planteado. En distinto color cada rol de usuario, en distinto tamaño según antigüedad en meses de creación del usuario (a mayor antigüedad mayor tamaño del nodo).

A partir de la observación de los cinco grafos, es posible afirmar que las primeras dos representaciones poseen una similitud mucho mayor entre sí respecto a la tercera representación. La cuarta y quinta representación son similares entre sí, y en menor

medida también a las primeras dos. Estas cuatro representaciones poseen una forma asimilable a una estrella, con ciertos nodos más alejados del resto, siendo la última representación la más diferente. El tercer modelo representado se encuentra distribuido en forma más homogénea similar a una gran nube, con un pequeño grupo de nodos agrupados en forma mucho más cercana y de manera alargada en la parte inferior de la distribución con conexiones a ciertos nodos más dispersos.

Como principal conclusión, y dado que todas las representaciones exceptuando la tercera son muy similares entre sí, se debe destacar que la antigüedad de creación del usuario explica, comparativamente, más sobre el grafo que el resto de las variables

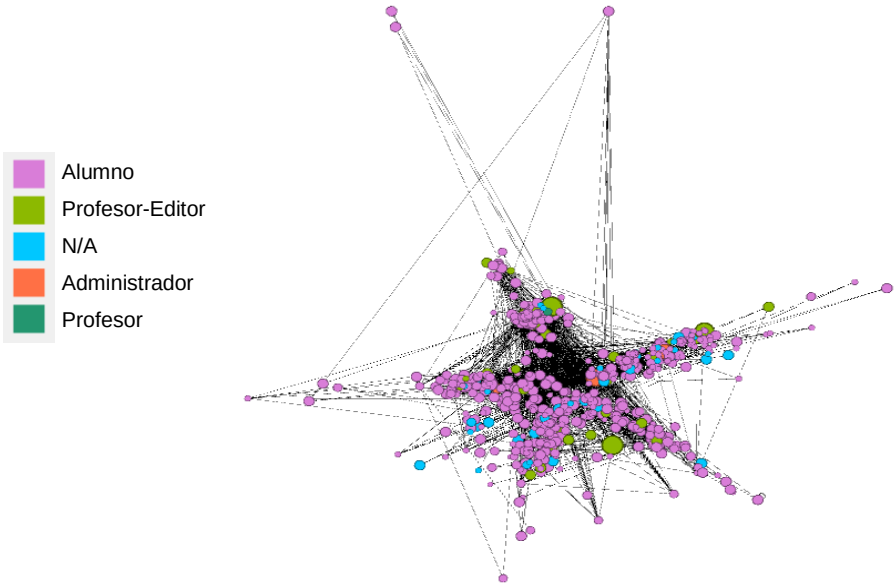

Gráfico 30. (Modelo 4) Grafo de usuarios conectados por al menos un mensaje enviado/recibido. La posición en el plano está dada por las dos dimensiones del cuarto modelo planteado. En distinto color cada rol de usuario, en distinto tamaño según antigüedad en meses de creación del usuario (a mayor antigüedad mayor tamaño del nodo).

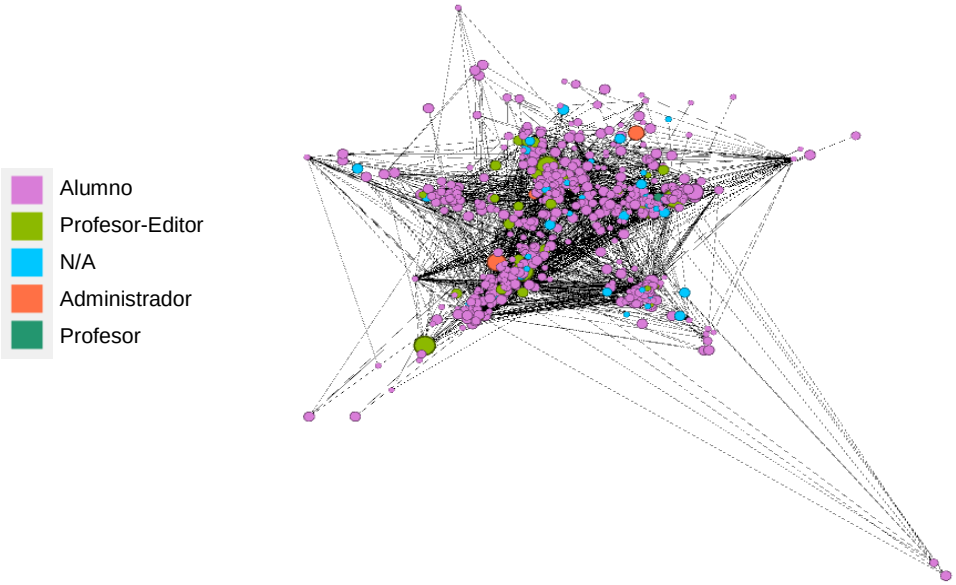

Gráfico 31. (Modelo 5) Grafo de usuarios conectados por al menos un mensaje enviado/recibido. La posición en el plano está dada por las dos dimensiones del quinto modelo planteado. En distinto color cada rol de usuario, en distinto tamaño según antigüedad en meses de creación del usuario (a mayor antigüedad mayor tamaño del nodo).

modeladas (rol del usuario, cantidad de cursos a los que está asociado, y cantidad de certificados de aprobación que posee el usuario).

Esta conclusión se encuentra en línea con lo observado a partir de las medias de los elementos de cada vector de valores propios. En la tabla 32 se puede apreciar que los resultados para los primeros dos modelos son extremadamente similares en ambos valores propios. En menor medida son estos resultados comparables con los relacionados a los modelos cuatro y cinco, similares éstos últimos entre sí. Mientras que para el tercer modelo los valores son muy diferentes, dominado principalmente por uno de ellos.

	Modelo 1	Modelo 2	Modelo 3	Modelo 4	Modelo 5
Media Vector 1	0,6547	0,6467	0,8756	0,5166	0,5122
Media Vector 2	0,6181	0,6405	-0,2654	0,5905	0,5552

Tabla 32. Valores medios de cada vector propio para cada uno de los cinco modelos basados en el grafo de usuarios conectados por al menos un mensaje enviado/recibido. (1) modelo sin covariables específicas para cada par de nodos; (2) modelo con covariable para rol en común; (3) modelo con covariable para similar antigüedad como usuario en la plataforma; (4) modelo con covariable para similar cantidad de cursos a los que está asociado el usuario; (5) modelo con covariable para similar cantidad de certificados de aprobación obtenidos.

Bondad de ajuste

Para evaluar la bondad de ajuste de los cinco modelos presentados previamente se utilizó el método de validación cruzada. Los resultados de las predicciones generadas

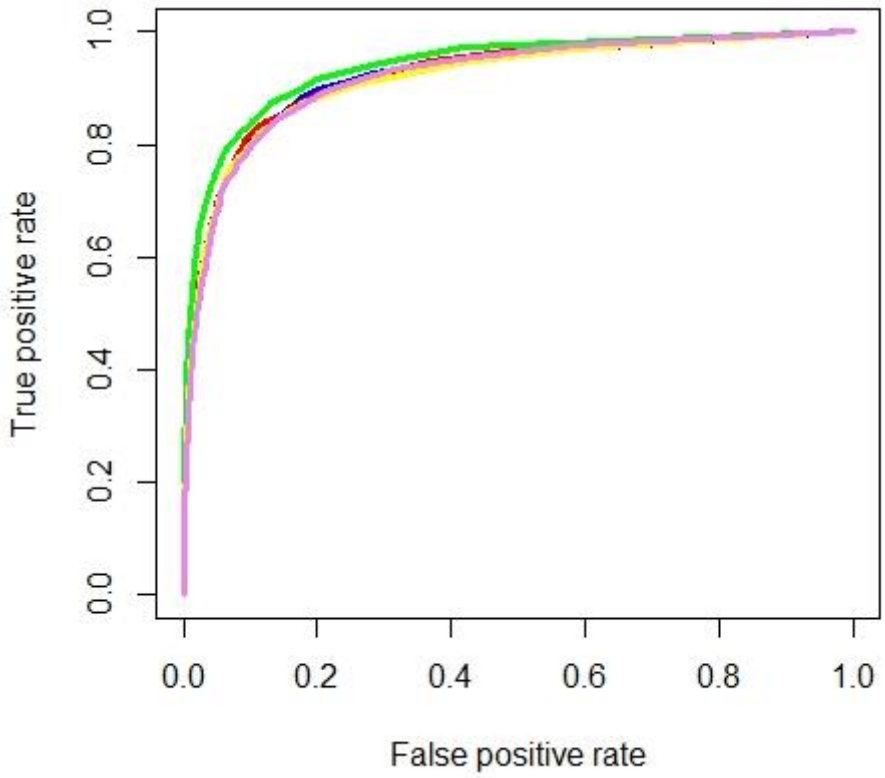

Gráfico 33. Curvas ROC comparando la bondad de ajuste de los tres modelos basados en el grafo de usuarios conectados por al menos un mensaje enviado/recibido. (1) modelo sin covariables específicas para cada par de nodos (azul); (2) modelo con covariable para rol en común (rojo); (3) modelo con covariable para similar antigüedad como usuario en la plataforma (verde); (4) modelo con covariable para similar cantidad de cursos a los que está asociado el usuario (amarillo); (5) modelo con covariable para similar cantidad de certificados de aprobación obtenidos (violeta).

para cada uno de los modelos son evaluados a partir de la observación de la curva ROC obtenida. Se utilizó, de manera arbitraria, un valor de K = 5, es decir, cinco subpartes.

A partir de la observación del gráfico 31, que muestra las cinco curvas ROC para cada uno de los cinco modelos de red latente planteados previamente, y de la tabla 34, que muestra los valores de AUC obtenidos para cada uno de estos modelos, es posible afirmar que los cinco modelos son comparables en rendimiento, presentando todos buenos resultados con valores de área bajo la curv AUC cercanos al 93%. El alto valor de AUC puede explicarse por el gran tamaño del grafo, lo que a su vez se interpreta como una alta homogeneidad de este. Es decir, con sólo una pequeña porción del grafo es posible explicar el comportamiento de nodos no vistos durante la etapa de entrenamiento del modelo con una muy alta precisión.

A modo de conclusión, el tercer modelo es el que presenta un mayor valor de AUC, mientras que el cuarto modelo es el que presenta el menor valor de AUC. De todas formas, y como se mencionó previamente, todos los cinco modelos presentan resultados muy similares.

	Modelo 1	**Modelo 2**	**Modelo 3**	**Modelo 4**	**Modelo 5**
AUC	0,9218	0,9240	0,9387	0,9179	0,9209

Tabla 34. Valores de AUC comparando la bondad de ajuste de los tres modelos basados en el grafo de usuarios conectados por al menos un mensaje enviado/recibido. (1) modelo sin covariables específicas para cada par de nodos; (2) modelo con covariable para rol en común; (3) modelo con covariable para similar antigüedad como usuario en la plataforma; (4) modelo con covariable para similar cantidad de cursos a los que está asociado el usuario; (5) modelo con covariable para similar cantidad de certificados de aprobación obtenidos.

DISCUSIÓN

Mediante el presente trabajo como tesis de maestría, se buscó implementar dentro del campo de la explotación de datos, procesos y técnicas aplicadas sobre una base de datos de cursos virtuales otorgados por el Instituto Nacional del Cáncer.

Se analizaron numerosos trabajos y papers que sostienen la importancia del presente trabajo. Por ejemplo, en el realizar análisis de redes sociales para extraer datos estructurales con aplicabilidad a sistemas educativos [15], así como en comprender la efectividad de los cursos brindados, analizando las características y factores que pueden afectarlos [25].

Primeramente, se realizó un preprocesamiento de las bases de datos provistas por el INC. Se analizaron individualmente cada una de las entidades para establecer si la información que contenía podría ser útil a los fines de esta investigación. De las casi 150 entidades con datos cargados se utilizaron finalmente tan sólo poco más de una decena para el desarrollo del trabajo.

Una vez obtenidas todas las entidades procesadas se prosiguió con la realización de los primeros análisis descriptivos. Se hizo un resumen de métricas a nivel usuario, dado por roles, cursos, alumnos, certificados, y contactos. Se identificaron algunos valores atípicos y datos faltantes que deben servir para mejorar la calidad de la información, así como se extrajeron métricas y resultados para una mejor comprensión del universo analizado. Estos primeros resultados permitieron entender la complejidad de la información y del área en general.

Se observó la distribución de cursos comenzados y vigentes, así como la cantidad de usuarios inscriptos y certificados otorgados en cada mes dentro del periodo de análisis. En septiembre 2016 y mayo 2017 se presentan la mayor cantidad de cursos disponibles. Asimismo, existen dos picos con una alta cantidad de certificados otorgados en noviembre 2016 y en junio 2017.

Adicionalmente, se identificó la cantidad de usuarios y contactos que existen entre ellos. Se pudo determinar que sólo el 4,5% tiene un contacto con otro usuario, número que está en línea con lo observado en otros trabajos [5]. Sin embargo, y según lo ya visto en [4], una mayor interacción y participación sería beneficiosa para un mayor desempeño de los alumnos.

Luego del análisis del modelo de datos completo, se prosiguió a la creación de entidades con un mayor nivel de agregación. Se seleccionaron las entidades necesarias para hacer las selecciones de usuarios y relaciones que serían finalmente objeto del desarrollo del trabajo. Luego de obtenidas las entidades derivadas fue necesario convertirlas en matrices de adyacencia a nivel usuario. En todos los casos se buscó analizar las relaciones entre usuarios armando matrices rectangulares. En este sentido se unió a usuarios en base a los contactos que poseen entre ellos, el haber cursado una misma materia o el haber obtenido un mismo certificado de participación o de aprobación.

A continuación, se generaron distintos grafos sobre los cuales se realizaron pruebas y análisis para entender su composición y características. Se realizaron análisis topológicos para comprender su composición, características y distribución. Se validó a

su vez si se trataban, en cada caso o no, de redes de mundo pequeño o de grafos al azar. Se planteó este análisis debido a la composición propia de cada grafo y a la información que es posible obtener de la caracterización de estos.

Se trabajó siempre sobre grafos de usuarios conectados según distintos criterios: usuarios que enviaron mensajes a otros usuarios; usuarios que compartieron un mismo curso. En ambos casos, se considera que dichos usuarios poseen una conexión. Los grafos se conformaron como no dirigidos, se caracterizaron topológicamente, fue de utilidad comprobar si se trataban de grafos asimilables a redes de mundo pequeño, y por último se analizaron los atributos de cada grafo.

El grafo de usuarios conectados a través de mensajes es un grafo no conectado, pero con una componente principal muy grande la cual fue analizada más en detalle y sobre la cual se obtuvieron todas las métricas. Se observó que un pequeño grupo de usuarios poseen una gran cantidad de contactos, mientras que la gran mayoría posee una cantidad muy pequeña de estos. No se pudo descartar la hipótesis de que se trate de un grafo de mundo pequeño. Se observó, además, que los alumnos que poseen un certificado de aprobación conforman una red más concentrada, con mayor cantidad de interconexiones, que aquellos que no lo poseen. También se pudo distinguir que los usuarios con el rol de “Profesor-Editor” son los que poseen la mayor cantidad de contactos.

Por otro lado, el grafo de usuarios conectados a través de cursos en común es un grafo conectado con una enorme cantidad de nodos. Al igual que en el caso anterior, se observó que un pequeño grupo de usuarios poseen una gran cantidad de contactos, mientras que la gran mayoría posee una cantidad muy pequeña de estos. No se pudo descartar la hipótesis de que se trate de un grafo de mundo pequeño. Se observó, además, que los “Administradores” poseen la mayor cantidad promedio de contactos, resultado esperable dado a que estarán vinculados a más cursos por causa de su rol. Existen usuarios con el rol de “Profesor-Editor” que poseen al menos un certificado de aprobación, siendo este grupo el que posee la mayor cantidad de contactos. Mientras que los usuarios con el rol de “Alumno” que obtuvieron al menos un certificado poseen un promedio de contactos levemente superior a los que no obtuvieron ninguno, resultado en línea con lo observado en [4].

Finalmente, se trabajó sobre el análisis de los usuarios de forma individual observando sus características. Se presentaron distintos modelos estadísticos diseñados específicamente para el análisis de grafos. Se utilizó la base de usuarios conectados a través de mensajes para aplicar dichos modelos. Se utilizó la matriz de adyacencia dada por los usuarios conectados a través de al menos un mensaje perteneciente al componente principal del grafo, así como características a nivel usuario. Se buscó entender la interacción entre los distintos usuarios y evaluar la importancia de las variables insertadas en cada caso, comprender procesos de homofilia o similitud, así como de transitividad, dentro de la conformación del grafo y el establecimiento de conexiones entre los usuarios.

De esta forma, y según lo explicado, se puso en práctica la realización del proceso de descubrimiento de conocimiento. A continuación, se resumen los pasos que integran este proceso y las tareas que fueron realizadas dentro del trabajo, en línea con el mismo.

1. Data cleaning: necesario durante la extracción y primer procesamiento de la información obtenida del INC.
2. Data integration: combinación de las entidades presentes dentro de la base de datos.
3. Data selection: selección de datos relevantes para los fines del presente trabajo (usuarios, roles, contactos, cursos, certificados).
4. Data transformation: transformación y consolidación de la información para poder realizar análisis de grafos, o explotar los modelos estadísticos planteados.
5. Data mining: aplicación de modelos estadísticos para entender la importancia de las variables dentro del grafo, y la interacción de estas.
6. Pattern evaluation: identificar patrones o variables relevantes que permitan explicar el comportamiento de los grafos y la distribución de estos.
7. Knowledge presentation: presentación de los resultados en el presente trabajo, resumiendo lo analizado en forma tanto escrita como gráfica.

Recordando que el objetivo principal del trabajo era “comprender las diferencias que existen entre aquellos alumnos que logran finalizar su curso y obtener el correspondiente certificado de aprobación, y aquellos que no lo logran”; se puede validar el cumplimiento del objetivo propuesto.

Para el armado de la base de datos a ser evaluada utilizando los modelos estadísticos se consideraron las siguientes variables:

- Meses desde la creación del usuario.
- Rol del usuario.
- Cantidad de cursos en los que está inscripto el usuario.
- Cantidad de certificados de aprobación que posee el usuario.

Se plantearon dos tipos de modelos estadísticos. Modelos exponenciales de grafos aleatorios; y modelos de red latente.

A partir del armado de los modelos exponenciales de grafos aleatorios se analizó la importancia de cada una de las variables planteadas. Se creó un modelo que luego fue ajustado, y sobre el cual por último se analizó la bondad de su ajuste.

Mediante el mismo se pudo identificar que el que un par de usuarios posea el mismo rol disminuye significativamente (en casi un 30%) la posibilidad de que haya contacto entre los mismos, siendo ésta la variable con mayor importancia del modelo. En segundo lugar de importancia, se observó que la cantidad de certificados de aprobación del usuario aumenta las posibilidad de que haya contacto en un 4.2%. La antigüedad en meses de creación del usuario aumenta las posibilidades de que haya contacto entre usuarios en un 0.42%. La cantidad de cursos a los que está asociado el usuario reduce las posibilidades de que haya contacto en un 0.5%, sin embargo, este estadístico fue señalado como no significativo para el modelo. Por último, se encontró evidencia de un efecto de transitividad no trivial dentro del grafo, por lo que existe ya sea algún otro atributo que no fue controlado, o un posible proceso social de formación de grupos.

La bondad de ajuste de dicho modelo fue buena. Se observó que dada la gran cantidad de nodos existententes dentro del grafo y la distribución de éstos, las distribuciones se encontraban relativamente sesgadas con proporciones altas de nodos

y vértices únicamente para valores bajos en grado y en contactos compartidos por vértice, respectivamente.

En segundo lugar, se utilizaron moledos de red latente ya que permiten capturar los efectos de homofilia o similaridad, así como de distancia. Por lo tanto, se definieron cinco modelos distintos con diferentes covariables para luego analizar el comportamiento de cada uno:

1. Modelo sin covariables específicas para cada par de nodos.
2. Modelo con covariable para rol en común.
3. Modelo con covariable para similar antigüedad como usuario en la plataforma.
4. Modelo con covariable para similar cantidad de cursos a los que está asociado el usuario.
5. Modelo con covariable para similar cantidad de certificados de aprobación obtenidos.

Como principal conclusión se identificó que la antigüedad de creación del usuario explica, comparativamente, más sobre el grafo que el resto de las variables modeladas (rol del usuario, cantidad de cursos a los que está asociado, y cantidad de certificados de aprobación que posee el usuario). Los resultados para los primeros dos modelos fueron extremadamente similares. En menor medida fueron estos resultados comparables con los relacionados a los modelos cuatro y cinco, similares éstos últimos entre sí. Mientras que el tercer modelo fue el más distinto.

En cuanto a la bondad de ajuste de estos cinco modelos, todos obtuvieron altos valores de AUC, muy similares entre sí y cercanos al 93%. El tercer modelo es el que presentó el mayor valor de AUC, mientras que el cuarto modelo es el que presentó el menor valor de AUC.

Por último, si se comparan los resultados obtenidos comparados entre el modelo exponencial de grafos aleatorios y los modelos de red latente, es posible señalar que existen algunas diferencias destacables. La principal diferencia apunta a la importancia de la variable de similaridad en la antigüedad como usuario en la plataforma para el modelo de red latente, observándose que aportaba resultados significativamente distintos respecto al modelo sin covariables, mientras que dentro del modelo exponencial de grafos aleatorios si bien fue una variable significativa, no obtuvo tanta relevancia. Por el contrario, el rol del usuario se destacó como la variable más importante del modelo exponencial de grafos aleatorios, pero sin embargo no presentó una gran diferencia respecto del modelo sin covariables dentro de los modelos de red latente.

Estos resultados, si bien pueden resultar contradictorios, se basan en la utilización de dos modelos distintos con diferente funcionamiento. Los modelos exponenciales de grafos aleatorios evalúan la interacción de todas las variables en su conjunto, determinando la importancia y significancia para el modelo. Los modelos de red latente fueron ajustados individualmente evaluando la importancia y significancia de cada modelo con su covariable individual, en contraposición de un modelo general sin covariables específicas para cada par de nodos. Por lo tanto, si bien los modelos de red latente pueden ser útiles para la identificación de variables relevantes en forma individual, no capturan la interacción de éstas entre sí.

SUGERENCIAS PARA FUTUROS DESARROLLOS

A partir de los resultados obtenidos y las conclusiones extraídas, es importante destacar algunos puntos de importancia de cara a pensar próximas acciones. Como se mencionó al comienzo del trabajo, este trabajo debe funcionar como puntapié inicial de un proceso de mejora constante dentro del área académica del Instituto, que sirva para aggiornar tanto los cursos virtuales que se dictan actualmente, como la totalidad de la oferta de formación que está disponible.

En vistas de continuar con la línea de trabajo ahora planteada, es deseable en primera instancia contar con mayor historia en los datos. Se deberían validar los resultados observados analizando una ventana más amplia de tiempo, de forma tal de capturar mayor variabilidad y, especialmente, mayores períodos de inicio y fin de curso, y con esto de certificación de alumnos. El no contar con dicha información puede llevar a un sobreajuste sobre el período analizado, con resultados observados demasiado apegados al ciclo observado, pero poco explicativos del comportamiento general habitual de la cursada. Una mayor cantidad de información permitirá no sólo combatir el sobreajuste, sino también validar los resultados y conclusiones ya obtenidos.

En segundo lugar, será deseable focalizarse realizando un análisis más específico sobre ciertos grupos de variables, en particular, de las participaciones de los alumnos dentro de los foros. Siendo ésta una tarea altamente demandante desde el punto de vista de la interpretación del lenguaje y que puede derivar en todo un proyecto en sí mismo, con aplicación en este ámbito como en otros contextos.

Por último, no hay que dejar de resaltar que toda la información presente provista por el INC debe ser de utilidad para los fines de dicho instituto. Por tanto, es menester la implicación de personal y autoridades del Instituto, buscando con este trabajo generar información útil para la gestión de los cursos por ellos provista. El poder enfocar los esfuerzos a la realización de análisis de utilidad para el Instituto debiera ser el objetivo principal de cualquier nuevo análisis que desee ser abordado.

BIBLIOGRAFÍA

[1] Barabási, A.L.; Albert R. (1999) “Emergence of scaling in random networks”. Science.

[2] Barbera, Marco V.; Epasto, Alessandro; Mei, Alessandro; Perta, Vasile C.; Stefa, Julinda. Department of Computer Science, Sapienza University of Rome, Italy. (2013). “Signals from the Crowd: Uncovering Social Relationships through Smartphone Probes”.

[3] Calvó-Armengol, Antoni; Patacchini, Eleonora; Zenou, Yves. (2009). “Peer Effects and Social Networks in Education”.

[4] Cheng, Cho Kin; Paré, Dwayne E.; Collimore, Lisa-Marie; Joordens, Steve. (2010). “Assessing the effectiveness of a voluntary online discussion forum on improving students’ course performance”.

[5] Conde, Miguel Ángel; Hernandez-García, Ángel; García-Peñalvo, Francisco J.; Séin-Echaluce, María Luisa. Department of Mechanics, Computer Science and Aerospace Engineering, University of León, Spain; Departamento de Ingeniería de Organización, Administración de Empresas y Estadística, Universidad Politécnica de Madrid, Spain; Department of Computer Science, Faculty of Science, University of Salamanca, Spain; Department of Applied Mathematics, School of Engineering and Architecture, University of Zaragoza, Spain. (2015) “Exploring Student Interactions: Learning Analytics Tools for Student Tracking”

[6] Edwards, Gina; Kitzmiller, Rebecca R.; Breckenridge-Sproat, Sara. (2012). “Innovative Health Information Technology Training. Exploring Blended Learning”.

[7] Erdos, P; Renyi A. (1959) “On random graphs I”.

[8] Faraway, Julian J. (2006). “Extending the Linear Model with R”.

[9] Han, Jiawei; Kamber, Micheline; Pei, Jian. (2012). “Data Mining: Concepts and Techniques, 3rd Edition”.

[10] Kolaczyk, Eric D.; Csárdi, Gábor. (2014). “Statistical Analysis of Network Data with R”.

[11] Luna, J. M.; Castro, C.; Romero, C. Department of Computer Science and Numerical Analysis, University of Córdoba, Spain. (2016). “MDM Tool: A Data Mining Framework Integrated Into Moodle”.

[12] Myers, Seth A.; Sharma, Aneesh; Gupta, Pankaj; Lin, Jimmy. Twitter Inc. (2014). “Information Network or Social Network? The Structure of the Twitter Follow Graph”.

[13] Newman, M. E. E. (2013). “Mixing patterns in networks”.

[14] Noldus, Rogier; Van Mieghem, Piet (2015). “Assortativity in Complex Networks”.

[15] Rabbany k., Reihaneh; Takaffoli, Mansoureh; Zaïane, Osmar R. Department of Computing Science, University of Alberta, Canada. “Social Network Analysis and Mining to Support the Assessment of On-line Student Participation”.

[16] Trudeau, Richard J. (1993). “Introduction to Graph Theory”.

[17] Ugander, Johan; Karrer, Brian; Backstrom, Lars; Marlow, Cameron. (2011): “The Anatomy of the Facebook Social Graph”.

[18] Wasserman, S.; Faust, K. (1994). Cambridge: Cambridge University Press. “Social Network Analysis: Methods and Applications”.

[19] White, Su; Davis, Hugh; Dickens, Kate; León, Manuel; Sánchez-Vara, Ma Mar. Centre for Innovation in Technologies and Education, University of Southampton, UK. Department of Didactics and School Organisation, University of Murcia, Spain. (2015). “MOOCs: What Motivates the Producers and Participants?”

[20] Wilson, Christo; Boe, Bryce; Sala, Alessandra; Puttaswamy, Krishna P.N.; Zhao, Ben Y. Computer Science Department, University of California at Santa Barbara. (2009). “User Interactions in Social Networks and their Implications”.

[21] Wilson, Marcella; Nicholas, Charles. (2008). “Topological Analysis of an Online Social Network for Older Adults”.

[22] Wilson, Robin J. (1972). “Introduction to Graph Theory, 4th Edition”.

[23] Witten, Ian H.; Frank, Eibe. (2005). “Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition”.

[24] Xu, Jin; Gao, Yongqin; Christley, Scott; Madey, Gregory. Dept. of Computer Science and Engineering, University of Notre Dame. (2005). “A Topological Analysis of the Open Source Software Development Community”.

[25] Ya Ni, Anna. (2013). “Comparing the Effectiveness of Classroom and Online Learning: Teaching Research Methods”.

[26] Zou, Kelly H.; O’Malley, A. James; Mauri, Laura. (2007). “Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models”.