UNIVERSIDAD DE GRANADA
E.T.S.I. INFORMÁTICA Y TELECOMUNICACIÓN

UNIVERSIDAD
DE GRANADA

Departamento de Ciencias de la
Computación e Inteligencia Artificial

Minería de Medios Sociales

Guión de Prácticas Bloque I.1:

Análisis de una Red Social con *Gephi*

Curso 2024-2025

Máster en Ciencia de Datos e Ingeniería de Computadores

Práctica Bloque I.1

Análisis de una Red Social con *Gephi*

1. Objetivos

El objetivo de esta práctica es doble. Por un lado, familiarizarse con los procedimientos de análisis de redes y con las medidas habitualmente consideradas para esta tarea. Por otro, aprender el manejo de una herramienta estándar de análisis y visualización de redes como *Gephi* [1], disponible para su descarga en https://gephi.org/users/download/.

Para ello, se requerirá que el/la estudiante escoja una red social, la cargue en la herramienta, la visualice y calcule los valores de una serie de medidas estándar de análisis de redes para estudiar las características principales de la misma así como la influencia de los distintos actores que la componen y su posible estructura de comunidades.

La práctica se evalúa sobre un total de **3 puntos**. La fecha límite de entrega será el **viernes 11 de abril de 2025** antes de las 23:55 horas. La entrega de la práctica se realizará en el espacio de la asignatura en la plataforma Prado.

2. Trabajo a Realizar

En esta práctica, la red a analizar será una red escogida de entre las disponibles en la literatura [2]. Estas páginas web incluyen repositorios de redes que se pueden emplear, así como cualquier otra propuesta por el/la estudiante y comunicada con anterioridad al profesor para su aceptación:

Web	*Descripción*	*Formatos*
http://www-personal.umich.edu/~mejn/netdata/	Repositorio de *Mark Newman*. Contiene 16 redes de distintos dominios.	`GML`
https://networkrepository.com/index.php	*Network Repository*. Contiene varios miles de redes de 33 dominios distintos.	Ficheros de texto con listas de enlaces (*comprimidos*)
http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:nets:old http://vladowiki.fmf.uni-lj.si/doku.php?id=pajek:data:pajek:index	Repositorios de *Pajek*. Contiene una gran cantidad de distintos dominios, incluyendo redes de relaciones sociales en colegios.	Principalmente `NET`

[1] Aunque se recomienda el uso de *Gephi* y este guión de prácticas está personalizado para esa herramienta, el alumno puede optar por realizarla con cualquier otra de las herramientas de análisis y visualización de redes existentes.

[2] Como alternativa, también se permite que el/la estudiante analice una red obtenida de Twitter (o de cualquier otra red social online) mediante el plugin de *Gephi* o usando cualquier otro *scraper*, que deberá ser comunicada y aceptada con anterioridad por el profesor.

https://github.com/gephi/gephi/wiki/Datasets	Repositorio de *Gephi*. Contiene unas 30 redes de distintos dominios, la mayoría redes sociales.	Varios: `GEFX`, `GML`, `NET`, `GraphML`, …
http://snap.stanford.edu/data/	Repositorio *SNAP* de la *Universidad de Stanford*. Contiene unas 100 redes de distintos dominios, todas de gran tamaño y varias de redes sociales y redes sociales online.	Ficheros de texto con listas de enlaces (*comprimidos*), también algún `CSV` y `JSON`
http://vlado.fmf.uni-lj.si/pub/networks/data/UciNet/UciData.htm	Repositorio *UCINet*. Contiene unas 20 redes sociales de interacción humana. Son de tamaño pequeño.	Principalmente, ficheros de texto con matrices de adyacencia completas o listas de enlaces. Alguna en formato `NET`
https://marketplace.sshopencloud.eu/dataset/cwXY70 https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/T4HBA3	Repositorio *Moviegalaxies*. Contiene 773 redes sociales de películas famosas realizadas entre 1915 y 2012. Los nombres de las películas y las medidas de la red correspondiente están en el enlace *Network_ metadata.tab* y se pueden ver pulsando en el icono del ojo. Las redes se descargan pulsando en el enlace *Gexf.zip* en la parte de abajo de la pantalla (o en el icono del ojo) y eligiendo la deseada.	`GEFX`
https://bansallab.github.io/asnr/data.html	Repositorio *Animal Network Data*. Contiene 770 redes sociales de interacción de 69 especies animales.	`GraphML`
http://konect.cc/	Repositorio del proyecto *KONECT*. Contiene 1326 redes sociales de 24 dominios, disponibles en el enlace *Categories*.	Ficheros de texto con listas de enlaces (*comprimidos*)
http://www.sociopatterns.org/datasets/	Repositorio del proyecto *SocioPatterns*. Contiene distintas redes sociales de 15 dominios de interacción humana (varias por cada dominio). En formato csv, JASON y GEXF	Varios: Ficheros de texto con listas de enlaces, `GEFX`, `CSV`, `JSON`, …
http://www.martingrandjean.ch/	Portal de *Digital Humanities* | *Data Visualisation* | *Network Analysis* de *Martin Grandjean*. Aunque realmente no es un repositorio de redes, incluye *posts* de trabajos de investigación muy completos sobre análisis de redes sociales en distintos campos de las Humanidades Digitales. La mayoría incluyen las redes	Varios: `GEFX`, `CSV`, `JSON`, …
https://labs.polsys.net/playground/spotify/	Portal de la Digital Methods Initiative (Bernhard Rieder, University of Amsterdam) que permite generar la red de artistas y grupos musicales relacionados con el artista/grupo semilla elegido.	`GDF`
https://labs.polsys.net/tools/tumblr/	Portal de la Digital Methods Initiative (Bernhard Rieder, University of Amsterdam) que permite generar la red de tags relacionados de la red social Tumblr en posts sobre un tema (tag) semilla elegido.	`GDF`

El/la estudiante puede escoger la red que desee de entre las disponibles en estos repositorios o en cualquier otro, pero deberá comunicarlo al profesor para que le confirme la asignación antes de comenzar a realizar la práctica. Esto se hace con objeto de que no se escojan las mismas redes en las prácticas de varios/as estudiantes. **Para establecer una comunicación fluida en las dos direcciones (comunicación con el profesor y conocimiento de las redes que han escogido los compañeros), se ha habilitado un wiki en el espacio de Prado de la asignatura**. **Se valorará positivamente el uso de redes con un tamaño razonable, al menos de unos pocos cientos de nodos**.

2.1. Análisis Básico de la Red

Una vez generada la red, se cargará en *Gephi* y se realizarán tareas básicas de análisis y visualización. Si la red presenta más de una componente conexa, se recomienda usar *Force Atlas 2* como algoritmo de *layout (*en la ventana *Distribución)*. Para evitar que las componentes conexas queden fuera de la vista principal que muestra la componente gigante, fijar el valor del parámetro *Gravedad* en *Puesta a punto* a un valor entre 10 y 20. Si todo queda demasiado amontonado, se puede probar a marcar la opción *Disuadir Hubs* y/o *Evitar el solapamiento*. Los aspectos estéticos de la visualización se dejan al parecer del estudiantado, que puede probar las distintas variantes de algoritmos de *layout* implementados en *Gephi* y distintos valores de parámetros para determinar cuál le proporciona la distribución que más le guste.

Para los primeros pasos del análisis, comenzaremos por anotar los valores de las **medidas globales** básicas: número de nodos N y número de enlaces L, que aparecen directamente en la ventana *Contexto*, además de calcular manualmente el número máximo de enlaces L_{max}. Posteriormente, calcularemos otra medida global, el grado medio $<k>$, ejecutando la opción correspondiente en la ventana *Estadísticas*. Al realizar el cálculo del grado medio, obtendremos también la distribución de grados de la red completa, que debemos grabar (*Gephi* lo guarda en una carpeta con una imagen *png* y un fichero *html*).

La opción *Densidad de grafo* nos mide la relación entre número de enlaces L y el número máximo de enlaces L_{max}. La ejecutaremos y anotaremos el valor.

Posteriormente, ejecutaremos la opción *Coeficiente medio de clustering* para obtener la medida del mismo nombre, $<C>$. Dicha opción nos proporcionará también la distribución de coeficientes de clustering de la red, que guardaremos [3].

Ahora pasaremos a analizar la **conectividad de la red**. En primer lugar, obtendremos el número de componentes conexas ejecutando la opción *Componentes conexos* y lo anotaremos. Luego nos centraremos en la componente gigante y calcularemos su número de nodos. Para ello, iremos a *Filtros*, seleccionaremos *Topología→Componente gigante* y arrastramos el filtro a la ventana de abajo llamada *Consultas* donde pone *Arrastrar filtro aquí*. Entonces pulsaremos en el botón *Filtrar* con la flecha verde en la esquina inferior izquierda de la pantalla. La visualización cambiará y sólo mostrará la componente gigante. La ventana *Contexto* en la esquina superior izquierda nos mostrará el número de nodos y enlaces de dicha componente y sus porcentajes con respecto a la red total, los cuales anotaremos.

Finalmente, calcularemos las restantes **medidas globales** (diámetro d_{max} y distancia media d) sobre la componente gigante de la red ejecutando la opción correspondiente al *Diámetro de la red* en la ventana *Estadísticas*. El cálculo del diámetro nos proporciona también el valor de la distancia media, que anotaremos, así como el de tres medidas de Centralidad (**intermediación**, **cercanía** y **excentricidad**), que emplearemos en la siguiente sección de la práctica **y no deben incluirse en esta sección**.

[3] Hay veces que *Gephi* falla y devuelve una gráfica de coeficiente de clustering vacía. Esto ocurre cuando la red es no dirigida. En ese caso, habrá que generarla a mano usando *Excel*. Para ello, basta con entrar en la pestaña *Laboratorio de datos* de *Gephi*, exportar los datos correspondientes en formato *csv* e importarlos en *Excel* para generar la gráfica correspondiente.

La última tarea para realizar será escribir un pequeño análisis de la red estudiada a partir de los valores de medidas y de las gráficas de distribución de grados, etc. obtenidas. Será un análisis igual al que se realiza para las redes de proteínas de la levadura y de amistad de Facebook del profesor en las transparencias de la Sesión I.1 del curso. No se trata de escribir mucho sino de hacer un análisis razonable considerando los conocimientos limitados que tenemos sobre el análisis de redes.

2.2. Estudio de la Centralidad de los Actores

El/la estudiante realizará un pequeño análisis de redes sociales sobre la red basado en medidas de Centralidad. Determinará los 5 actores principales de la misma mediante las medidas de **grado**, **intermediación**, **cercanía** y **vector propio**.

El valor de tres de estas medidas ya está calculado con los pasos que hemos realizado en la sección anterior. La centralidad de grado (no normalizada) se generó al calcular el *Grado medio* en la ventana *Estadísticas*. Las de intermediación y cercanía se generaron con la opción *Diámetro de la red*. En este caso, sí que es posible especificar si se desean obtener normalizadas o no normalizadas con el *checkbox Normalizar centralidades en el rango [0,1]*. Finalmente, la *Centralidad de vector propio* se calcula en la opción del menú *Estadísticas* del mismo nombre.

En el caso en que nuestra red presente más de una componte conexa, se recomienda calcular la Centralidad de cercanía sobre la componente gigante de la misma. Es decir, aplicar primero el filtro de Componente gigante estudiado en la primera práctica y luego calcular el valor de esta medida. Esto se hace para evitar que nodos pertenecientes a componentes conexas de tamaño pequeño obtengan valores altos en la Centralidad de cercanía y falseen el análisis de ésta.

Los valores de centralidad de cada nodo pueden visualizarse en la tabla *Nodos* de la pestaña *Laboratorio de datos*, junto con el resto de la información asociada a cada nodo. Cada vez que se calcula una nueva medida usando las opciones de *Gephi*, aparece una nueva columna en esta tabla con sus valores. Se pueden ordenar los nodos por columnas simplemente pulsando sobre ellas. El/la estudiante anotará los nombres de los 5 actores con mejor valor para cada una de las cuatro medidas anteriores, así como el valor de dichas medidas y los almacenará en una tabla como la siguiente:

Centralidad de Grado	Centralidad de Intermediación	Centralidad de Cercanía	Centralidad de Vector propio
Nombre 1er actor: valor 1er actor	*Nombre 1er actor: valor 1er actor*	*Nombre 1er actor: valor 1er actor*	*Nombre 1er actor: valor 1er actor*
Nombre 2o actor: valor 2o actor	*Nombre 2o actor: valor 2o actor*	*Nombre 2o actor: valor 2o actor*	*Nombre 2o actor: valor 2o actor*
Nombre 3er actor: valor 3er actor	*Nombre 3er actor: valor 3er actor*	*Nombre 3er actor: valor 3er actor*	*Nombre 3er actor: valor 3er actor*
Nombre 4o actor: valor 4o actor	*Nombre 4o actor: valor 4o actor*	*Nombre 4o actor: valor 4o actor*	*Nombre 4o actor: valor 4o actor*
Nombre 5o actor: valor 5o actor	*Nombre 5o actor: valor 5o actor*	*Nombre 5o actor: valor 5o actor*	*Nombre 5o actor: valor 5o actor*

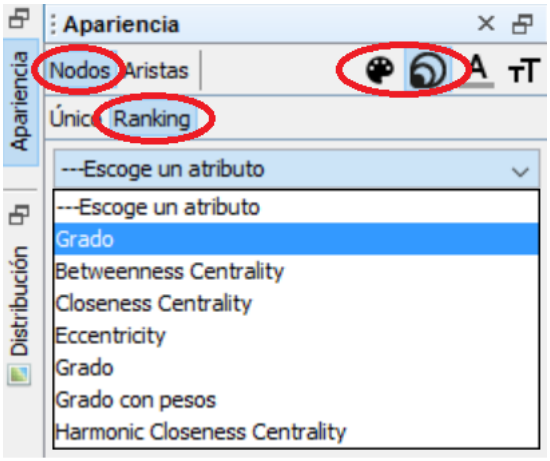
Tabla X: Valores de las distintas medidas de Centralidad en los actores principales de la red

Se piden **al menos cinco actores** para realizar el análisis, pero el/la estudiante puede añadir más en alguna o todas las medidas de Centralidad. Como vimos en el Seminario asociado al cálculo de medidas de Centralidad en *Gephi*, esto dependerá de la distribución concreta de los valores de cada medida. Por ejemplo, si tenemos siete actores con los valores más altos de una medida y luego encontramos un salto en el que el valor baja mucho en el octavo actor, incluiríamos esos siete. Si hay tres actores con un valor destacado y luego el cuarto y el quinto presentan un valor significativamente más bajo, entonces bastará con incluir esos cinco.

Finalmente, realizará un pequeño análisis de los actores más importantes de la red desde una perspectiva global en función de los valores de estas medidas y el conocimiento adquirido en la Sesión I.2 del curso.

Se valorará adicionalmente la realización de gráficas extra tales como:

- Representaciones de la red en las que se visualicen dos de las medidas anteriores (por ejemplo, la intermediación en el tamaño de los nodos y la centralidad de vector propio en el color de estos) como las mostradas en las transparencias de la Sesión I.2 del curso. Estas visualizaciones pueden realizarse directamente en *Gephi*, usando las opciones *Nodos* y *Ranking* en la ventana *Apariencia*. Los dos iconos con la paleta y las bolas de distinto tamaño de la parte superior derecha de la pantalla permiten escoger qué valor de medida se desea emplear para definir el color y el tamaño de los nodos en la visualización, respectivamente:



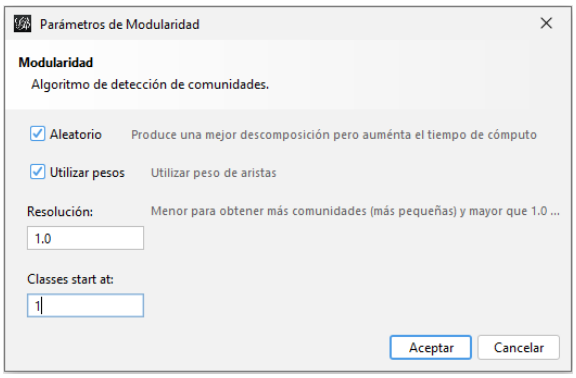
- Gráficos que representen los valores de dos de las medidas para todos los actores de la red en ejes de coordenadas como los estudiados en la Sesión I.2. Para realizarlos, puede exportar los valores de la Tabla de datos de la red en *csv* con la opción *Exportar tabla* y generarlos fácilmente usando Excel, R o cualquier otro software estadístico.

El/la estudiante escogerá las medidas que desee visualizar y justificará su elección. Las visualizaciones nos permitirán localizar las posiciones de los nodos más centrales de la red según cada medida. **Se analizará cada visualización y gráfico presentados**.

2.3. Detección de Comunidades

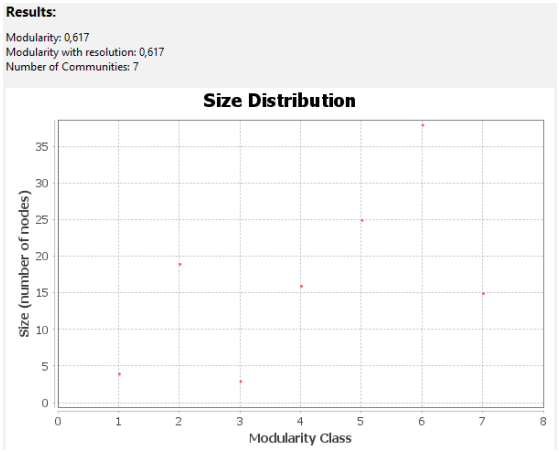
Se aplicarán al menos dos métodos de detección de comunidades sobre la red estudiada para determinar la estructura modular de la red. Antes de realizar ninguna otra tarea, el/la estudiante comprobará el número y la composición de las componentes conexas de la red escogida. Si la red tiene una única componente conexa o varias componentes conexas de gran tamaño, se trabajará con la red completa. En cambio, **si la red social escogida tiene muchas componentes conexas de pequeño tamaño es recomendable aplicar los algoritmos de detección de comunidades sobre la componente gigante**, aplicando primero el filtro de *Componente gigante*. En caso de no hacerlo así y ejecutar los algoritmos sobre la red completa, se obtendrá un número alto de comunidades pequeñas, pudiendo incluso dejar la componente gigante como una única comunidad, lo que falsearía la partición de comunidades.

El primer método para ejecutar será el de Lovaina, disponible en *Gephi*, que se aplicará ejecutando la opción *Modularidad* en la ventana *Estadísticas*:



El/la estudiante escogerá **al menos cinco valores** para el parámetro *Resolución*, que determina el número de comunidades de la partición devuelta por el algoritmo, recordando que un valor más alto del parámetro genera un número menor de comunidades de mayor tamaño y viceversa. El valor intermedio escogido será el 1.0 y después se emplearán al menos dos valores por encima y dos valores por debajo (como es lógico, esto puede depender de la estructura de la red). Si por ejemplo un valor inmediatamente superior a 1.0 ya genera un número mínimo de 1 o 2 comunidades, no tiene sentido generar otra partición con un valor mayor). Asimismo, **se podrán generar distintas particiones con el mismo valor de modularidad haciendo uso del *checkbox Aleatorio***. En cualquier caso, **el objetivo final será la obtención de una buena partición (valor alto de modularidad) con un número razonable de comunidades que permita realizar un buen análisis experto de la estructura de comunidades obtenida**.

El/la estudiante recopilará los valores de la medida de modularidad y el número de actores y enlaces de cada comunidad para cada particionamiento realizado. Estos datos se muestran en la información que proporciona *Gephi* después de cada ejecución del método de detección:



mientras que la composición de las comunidades en sí (la asignación de cada nodo a cada comunidad) pueden consultarse en la columna *Modularity Class* de la pestaña *Laboratorio de datos*. Deberá perseguirse la obtención de un número razonable de comunidades que permita realizar un buen análisis de la estructura de comunidades obtenida. Se mostrarán los valores de la medida de modularidad asociados a cada particionamiento realizado y se analizará la composición de las comunidades generadas para determinar si tienen algún tipo de influencia en la estructura de la red.

Una vez calculadas las medidas asociadas a las distintas particiones obtenidas, el/la estudiante construirá **una tabla de robustez de la modularidad de la red** como la siguiente:

Resolución	Modularidad	Número de comunidades
Valor Z menor que Y	*XXX*	*XXX*
Valor Y menor que 1.0	*XXX*	*XXX*
1.0	*XXX*	*XXX*
Valor K mayor que 1.0	*XXX*	*XXX*
Valor J mayor que K	*XXX*	*XXX*

Tabla Y: Propiedades de las distintas particiones obtenidas * *

** Si se han considerado más valores de resolución, se añadirán las filas necesarias*

*** Si se han realizado varias ejecuciones con el mismo valor de resolución, se añadirán las filas necesarias*

Además, seleccionará una partición, la de mayor modularidad, la de mejor equilibrio entre el número de comunidades y el valor de modularidad, o la que crea conveniente aportando una justificación, y generará **una tabla de cohesión estructural** que recoja las medidas asociadas:

Número de comunidad	Número de nodos	Porcentaje de nodos	Número de enlaces	Porcentaje de enlaces
Comunidad 0	*XXX*	*XXX*	*XXX*	*XXX*
Comunidad 1	*XXX*	*XXX*	*XXX*	*XXX*
…	*XXX*	*XXX*	*XXX*	*XXX*
Comunidad X	*XXX*	*XXX*	*XXX*	*XXX*

Tabla Z: Propiedades de las distintas comunidades de la partición X

Para el segundo método, se podrán emplear los disponibles como plugins de *Gephi*, como el de Girvan-Newman o el de Leiden; los existentes en la herramienta *DetCom* proporcionada; o cualquier otra implementación disponible. Se mostrarán los valores de la medida de modularidad asociados a cada particionamiento obtenido y se escogerá una partición y se mostrará la cohesión estructural de las comunidades generadas como para el caso de Lovaina.

Se comparará el comportamiento de los dos métodos ejecutados para determinar si tienen algún tipo de influencia en la estructura de la red. Se tendrá en cuenta que la estructura de comunidades obtenida no tiene por qué ser significativa ya que no todas las redes presentan una estructura modular. Se valorará el considerar distintos valores de parámetros de los algoritmos (si existieran) para obtener detectar distintas estructuras de comunidades.

El/la estudiante realizará también dos o más visualizaciones de las particiones más significativas obtenidas por cada método usando las opciones *Nodos* y *Partition*/*Ranking* en la ventana *Apariencia* para colorear los nodos en función de la comunidad a la que pertenezcan.

Finalmente, efectuará un **pequeño análisis de cohesión estructural y análisis semántico de las comunidades obtenidas en la partición escogida para cada método ejecutado**. El objetivo es aprender a analizar la cohesión estructural de una partición de comunidades, haciendo uso del valor obtenido en la medida de modularidad y de los enlaces internos de cada comunidad, y su consistencia semántica considerando información experta sobre el dominio para determinar las relaciones semánticas que dan lugar a que cada comunidad sea realmente un grupo cohesivo de actores.

3. Documentación y Ficheros a Entregar

El/la estudiante guardará el proyecto desde *Gephi* nombrándolo con sus apellidos y su nombre propio. Luego almacenará todos los valores obtenidos en la tabla incluida en el fichero Excel disponible en el espacio de la asignatura en la plataforma, llamado *MedidasRedesPracticaMMS-I-1.xls*, renombrando el fichero de la misma forma.

La **documentación** de la práctica será un fichero *pdf* que deberá incluir, al menos, el siguiente contenido:

a) Portada con el título de la práctica, el curso académico y el nombre, DNI y dirección e-mail del/de la estudiante.

b) Una sección que incluya:
   - Una imagen de la red completa y otra de la componente gigante con una visualización lo más estética posible.
   - La tabla Excel con los valores de las medidas estudiadas incrustada.
   - Los gráficos de las distribuciones de grado, etc.
   - El análisis de la red en función de dichos datos.

c) Una sección que describa el análisis de la centralidad de los actores de la red desarrollado en la Sección 2.2, incluyendo los gráficos y visualizaciones realizadas.

d) Una sección que describa el estudio de las comunidades extraídas de la red en la Sección 2.3, incluyendo los gráficos y visualizaciones realizadas.

e) Una sección con las visualizaciones y gráficos adicionales (**en caso de haberlos realizado**). También se puede optar por incluirlos en la sección anterior que corresponda.

f) Referencias bibliográficas u otro tipo de material distinto del proporcionado en la asignatura que se haya consultado para realizar la práctica (en caso de haberlo hecho).

Aunque lo esencial es el contenido, también debe cuidarse la presentación y la redacción. El fichero *pdf* de la documentación, el fichero original de la red, el fichero del proyecto *Gephi* y el fichero Excel con los valores de las medidas se comprimirán conjuntamente en un fichero *zip* etiquetado con los apellidos y nombre del/de la estudiante (Ej. Pérez Pérez Manuel.zip). Este fichero será entregado por internet a través de la plataforma PRADO (http://prado.ugr.es/).