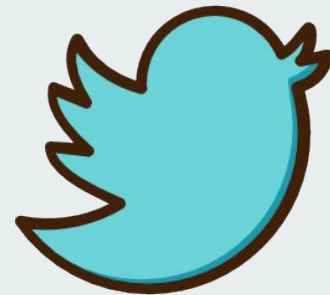




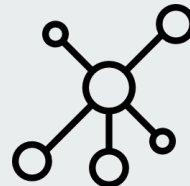
# The Twitter Alliance

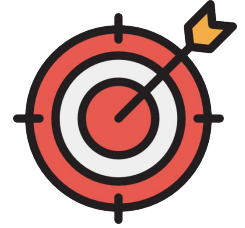


Descubriendo las relaciones entre los usuarios más seguidos de Twitter

David Arroyo Segovia  
Adrián Camacho Pérez  
Paula Muñoz Lago  
Carla Paola Peñarrieta Uribe

**Facultad de Informática - Universidad Complutense de Madrid**



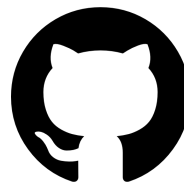


# Objetivos

- **Encontrar insiders.** Descubrir que cuentas son seguidas por personas clave en la red social Twitter. Estas cuentas suelen pertenecer a líderes de opinión que no tienen porqué contar con el mayor número de seguidores.
- **Desvelar las cuentas favoritas de un grupo determinado.** Qué cuentas son las más seguidas por una comunidad de usuarios, por ejemplo, el top 100 de España, políticos o periodistas.
- **Identificar nodos influyentes.** Determinar miembros clave de la comunidad que a menudo son desconocidos para los outsiders.

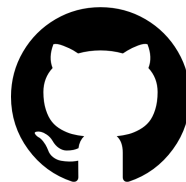
---

# Primeras pruebas: T-Hoarder, Twecoll, Scrapy y Tweepy



# T-Hoarder

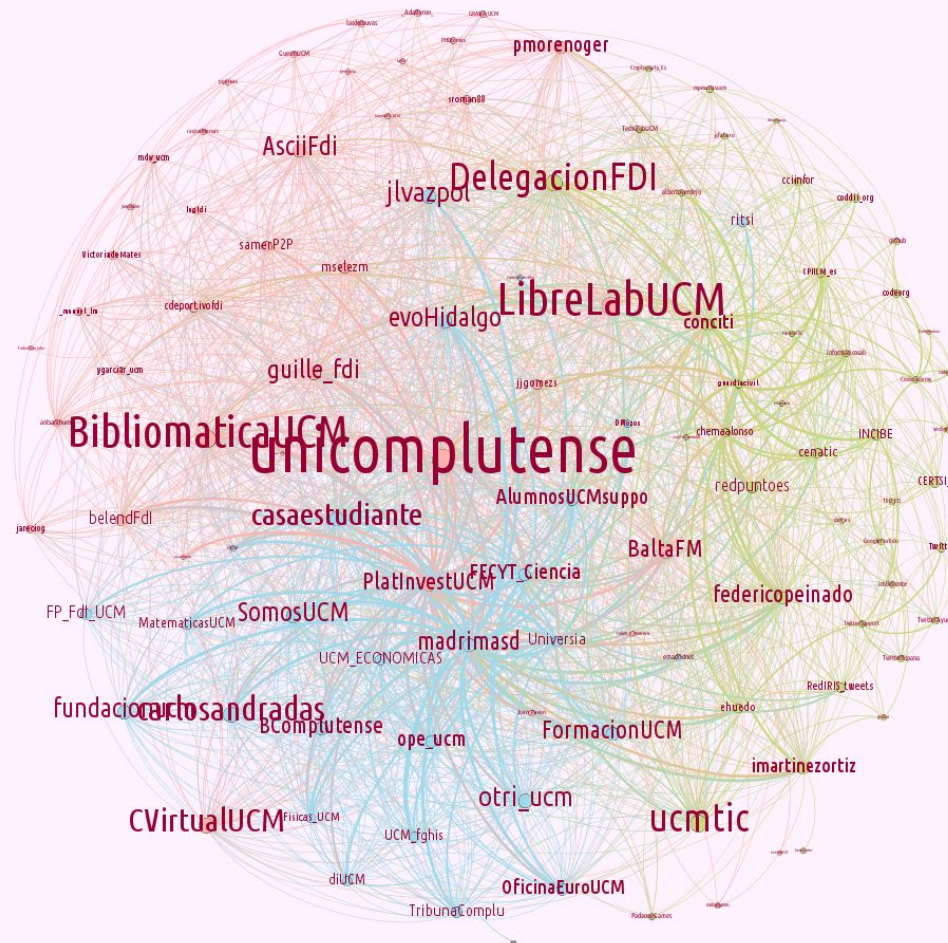
- T-Hoarder es una plataforma desarrollada por M. Luz Congosto (investigadora de la UC3M) para **extraer información de Twitter a largo plazo**. Esta información es procesada y visualizada automáticamente.
- Es una herramienta muy potente y con posibilidad de generar grafos.
- Sin embargo, no nos proporciona la suficiente flexibilidad para extraer la información que necesitamos.
- GitHub: [https://github.com/congosto/t-hoarder\\_kit](https://github.com/congosto/t-hoarder_kit)



# Twecoll: Twitter Collection Tool

- Es una herramienta desarrollada en **Python** que se utiliza desde la línea de comandos para extraer datos de Twitter.
- También da la posibilidad de **generar grafos** a partir de la información extraída.
- Genera grafos de tus **relaciones de primer y segundo grado** en Twitter.
- **Toma mucho tiempo** debido a los API throttling. Una de las razones por las cuales no nos ha sido útil.
- Extrae todos los seguidores de una cuenta y los de todos los seguidores de los anteriores (ignorando los que no se encuentren relacionados con el primero).
- GitHub: <https://github.com/jdevoo/twecoll>

# Followings: 114  
# Followers: 1629



Red generada con **Twecoll** a partir de la cuenta de @informaticaucm

# La solución

## Scrapy



- Framework para scraping y web crawling desarrollado en Python.
- Librería Open Source que nos permite extraer información importante que se encuentra en internet. Es posible leer el contenido de una web y movernos en su estructura **HTML** para **analizar sus datos y explotarlos** según sean nuestras necesidades.
- Lo utilizamos para **automatizar la extracción** de las **listas de cuentas más seguidas** de cada país.
- <https://scrapy.org/>

## Tweepy



- Es una librería escrita en Python que te permite acceder a todos los métodos que proporciona la **API REST de Twitter**.
- Proporciona una interfaz muy sencilla encapsulando algunas partes que nos permiten **agilizar el trabajo**.
- Permite **realizar esperas cuando se ha superado el *rate limit*** que impone Twitter para no detener la ejecución del script.
- <http://www.tweepy.org/>

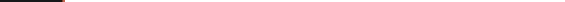
```
david@parrot-server:~/Code/The-Twitter-Alliance/Code/TwCounterSpider$ python3 main.py --country spain
```

[illegible]

```

2018-01-14 21:58:51 [scrapy.utils.log] INFO: Scrapy 1.5.0 started (bot: TwCrawlerSpider)
2018-01-14 21:58:51 [scrapy.utils.log] INFO: Versions: lxml 4.1.1.0, libxml2 2.9.7, cssselect 1.3.1, w3lib 1.18.0, Twisted 17.9.0, Python 3.5.2 (default, Nov 23 2017, 16:37:01) - [64bit], pyOpenSSL 17.5.0 (OpenSSL 1.1.0g 2 Nov 2017), cryptography 2.1.4, Platform Linux-4.4.0-164-ubuntu-Ubuntu-16.04-xenial
2018-01-14 21:58:51 [scrapy.crawler] INFO: Overridden settings: {'SPIDER_MODULES': ['TwCrawlerSpider'], 'BOT_NAME': 'TwCrawlerSpider', 'NEWSPIDER_MODULE': 'TwCrawlerSpider.spiders'}
2018-01-14 21:58:51 [scrapy.middleware] INFO: Enabled extensions:
'scrapy.extensions.memusage.MemoryUsage',
'scrapy.extensions.telnet.TelnetConsole',
'scrapy.extensions.corestats.CoreStats',
'scrapy.extensions.logstats.LogStats']
2018-01-14 21:58:51 [scrapy.middleware] INFO: Enabled downloader middlewares:
'scrapy.downloadermiddlewares.httpauth.HttpAuthMiddleware',
'scrapy.downloadermiddlewares.downloadtimeout.DownloadTimeoutMiddleware',
'scrapy.downloadermiddlewares.defaultheaders.DefaultHeadersMiddleware',
'scrapy.downloadermiddlewares.useragent.UserAgentMiddleware'

```



```
{'filename': 'Top100_spain.txt',
 'user_accounts': ['@realmadrid'
```

```
@realmadrid',
'@FCBarcelona',
'@andresiniesta8',
'@AlejandroSanz',
'@3gerardpique',
'@Pontifex_es',
'@RafaelNadal',
'@enriqueiglesias',
'@SergioRamos',
'@FCBarcelona_es',
'@XabiAlonso',
'@Carles5puvol',
```

# Lución

Extracción del top 100 de Twitter España de TwitterCounter.com





## Algoritmo: descubrir amistades

01100  
10110  
11110

- De cada una de las cuentas anteriormente extraídas se comprueba si existe un **seguimiento mutuo** con todas las demás.
- Si existe un seguimiento mutuo decimos que hay una **relación de amistad**.
- Se intentan hacer las **menos peticiones a la API** posibles: cacheo de peticiones pasadas, optimización del código, etc.
- El resultado se exporta a un fichero de texto plano que es posteriormente parseado con un **script en C++** que genera la relación de nodos y aristas.
- El resultado final es un **grafo no dirigido**.

## Algoritmo: all follows

01100  
10110  
11110

- A diferencia del procedimiento anterior, se **extraen los follows** que existen entre los usuarios de la lista aunque **no sean correspondidos** (sin ser mutuos).
- Se generan muchas más aristas que en el caso anterior.
- El resultado final es un **grafo dirigido**.



```

david@parrot-server: ~/Code/The-Twitter-Alliance/Code/TwitterAlliance
2018-01-15 11:44:36,939:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:37,133:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:37,133:INFO:PARAMS: {'target_screen_name': 'b@instagram', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:37,144:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:37,488:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:37,488:INFO:PARAMS: {'target_screen_name': 'b@marcmarquez93', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:37,419:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:37,627:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:37,634:INFO:PARAMS: {'target_screen_name': 'b@MesutOzil1088', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:37,639:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:37,841:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:37,841:INFO:PARAMS: {'target_screen_name': 'b@TownGamePlay', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:37,846:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:38,086:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:38,086:INFO:PARAMS: {'target_screen_name': 'b@rihanna', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:38,093:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:38,290:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:38,290:INFO:CURRENT SOURCE USER: @MarcBartra
2018-01-15 11:44:38,290:INFO:PARAMS: {'target_screen_name': 'b@MelendiOficial', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:38,297:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:38,545:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:38,545:INFO:PARAMS: {'target_screen_name': 'b@mario.casas_', 'source_screen_name': 'b@elmundoes'}
2018-01-15 11:44:38,550:INFO:Starting new HTTPS connection (1): api.twitter.com
2018-01-15 11:44:38,770:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:38,770:INFO:Friendship @MarcBartra - @lorenzo99 detected!
2018-01-15 11:44:38,770:INFO:TOTAL COMPLETED: 29%
2018-01-15 11:44:38,770:INFO:PARAMS: {'target_screen_name': 'b@zaynmalik', 'source_screen_name': 'b@elmundoes'}

```

```

david@parrot-server: ~/Code/The-Twitter-Alliance/Code/TwitterAlliance
2018-01-15 13:14:57,203:INFO:Friendship @AnnaSimonMari - @danimartinezweb detected!
2018-01-15 13:14:59,727:INFO:Friendship @AnnaSimonMari - @3gerardpique detected!
2018-01-15 13:15:08,224:INFO:Friendship @AnnaSimonMari - @floodezz detected!
2018-01-15 13:15:14,729:INFO:Friendship @AnnaSimonMari - @El_Hormiguero detected!
2018-01-15 13:30:00,276:INFO:Friendship @Reuters - @TheEconomist detected!
2018-01-15 13:44:56,389:INFO:Friendship @andresiniesta8 - @lorenzo99 detected!
2018-01-15 13:44:57,349:INFO:Friendship @andresiniesta8 - @Guaje7Villa detected!
2018-01-15 13:44:58,813:INFO:Friendship @andresiniesta8 - @ristomejide detected!
2018-01-15 13:44:58,813:INFO:Friendship @andresiniesta8 - @FCBarcelona_cat detected!
2018-01-15 13:45:01,610:INFO:Friendship @andresiniesta8 - @JordiAlba detected!
2018-01-15 13:45:01,852:INFO:Friendship @andresiniesta8 - @SSantiagoasegura detected!
2018-01-15 13:45:05,328:INFO:Friendship @andresiniesta8 - @llorentefer19 detected!
2018-01-15 13:45:05,560:INFO:Friendship @andresiniesta8 - @FCBarcelona_es detected!
2018-01-15 13:45:05,858:INFO:Friendship @andresiniesta8 - @CarlesSpuyol detected!
2018-01-15 13:45:07,237:INFO:Friendship @andresiniesta8 - @MarcBartra detected!
2018-01-15 13:45:13,825:INFO:Friendship @andresiniesta8 - @aarbeloa17 detected!
2018-01-15 13:45:14,451:INFO:Friendship @andresiniesta8 - @paugasol detected!
2018-01-15 13:45:15,685:INFO:Friendship @andresiniesta8 - @3gerardpique detected!
2018-01-15 13:45:18,112:INFO:Friendship @andresiniesta8 - @D_DeGea detected!
2018-01-15 13:45:18,384:INFO:Friendship @andresiniesta8 - @SeFutbol detected!
2018-01-15 13:45:19,792:INFO:Friendship @andresiniesta8 - @ctello91 detected!
2018-01-15 13:45:21,543:INFO:Friendship @andresiniesta8 - @Thiago6 detected!
2018-01-15 13:45:21,759:INFO:Friendship @andresiniesta8 - @DANIROVIRA detected!
2018-01-15 13:45:22,670:INFO:Friendship @andresiniesta8 - @jordievolle detected!
2018-01-15 13:45:26,896:INFO:Friendship @andresiniesta8 - @Buenafuente detected!
2018-01-15 13:45:27,122:INFO:Friendship @andresiniesta8 - @SergioRamos detected!
2018-01-15 13:45:31,179:INFO:Friendship @andresiniesta8 - @Pedro17_ detected!
2018-01-15 14:00:00,209:INFO:Friendship @andresiniesta8 - @marcmarquez93 detected!
david@parrot-server:~/Code/The-Twitter-Alliance/Code/TwitterAlliance$

```

Tiempo ejecución script ≈ 13 horas

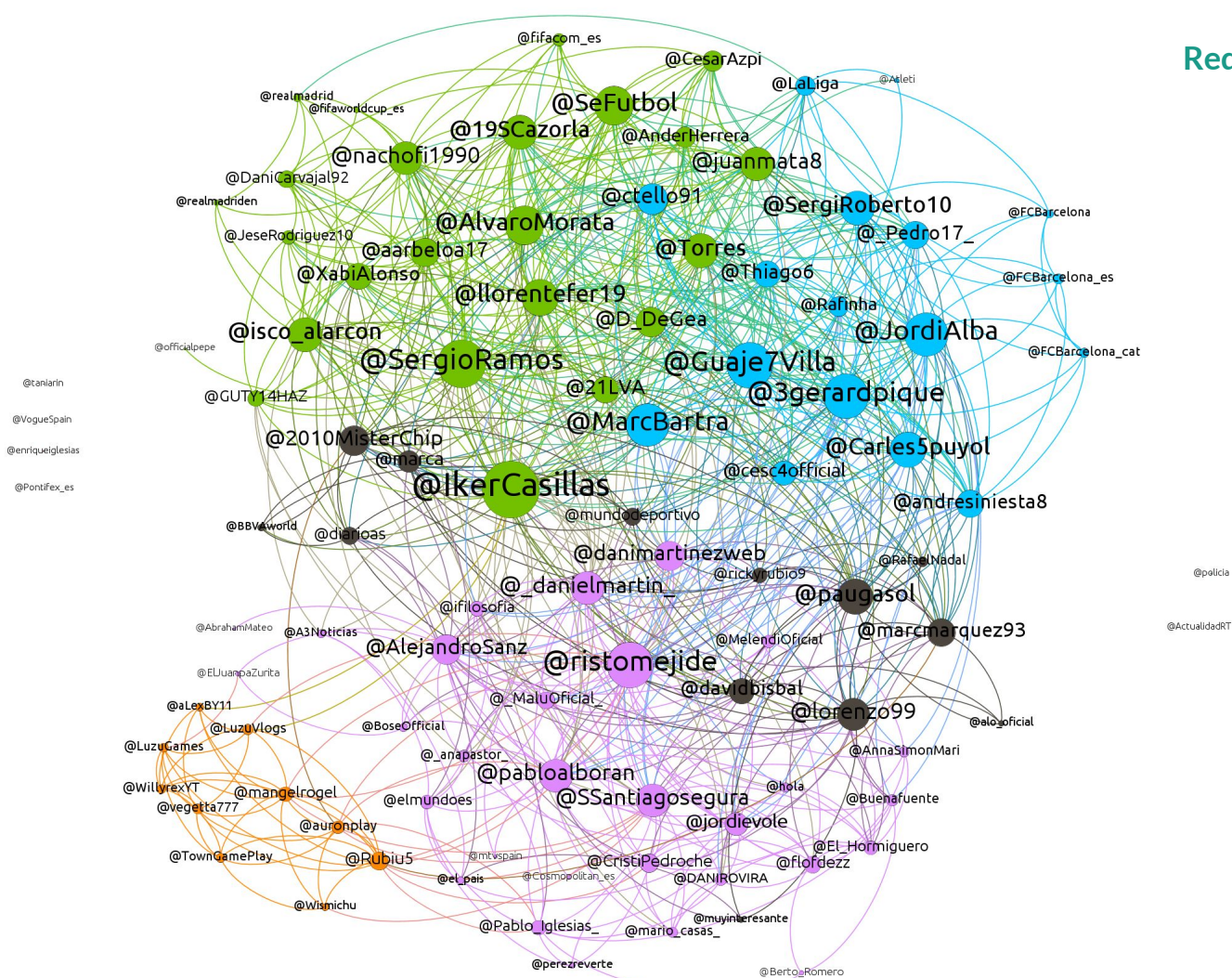
Logs generados al ejecutar el algoritmo sobre el top 100 de Twitter España



---

# Resultados preliminares: primeras redes

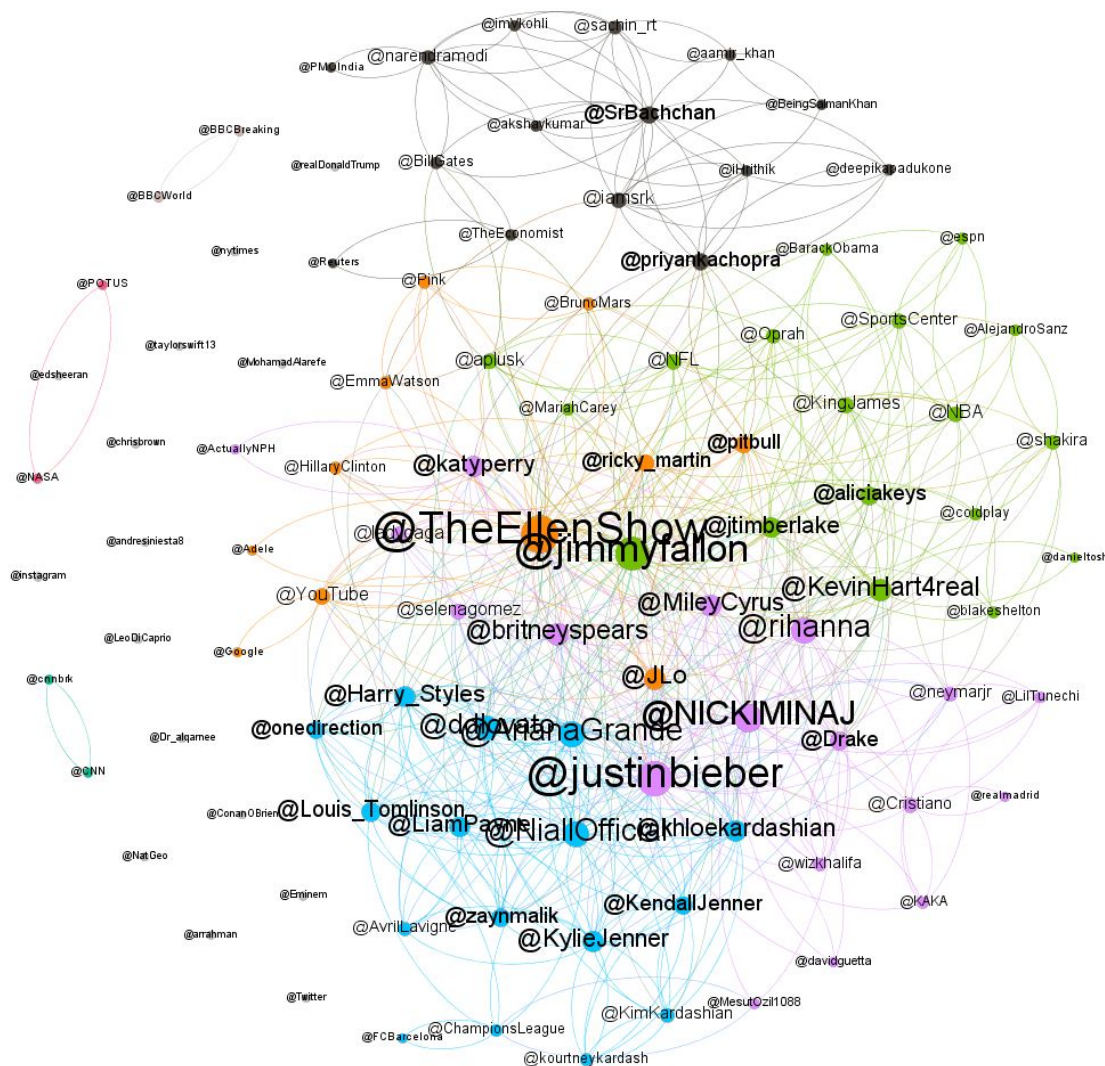
## Red del top 100 de Twitter España







## Red del top 100 de Twitter









# Trabajo por hacer

- Determinar sobre qué **más grupos o comunidades** realizar el análisis.
- **Analizar** los datos y **comprender** las relaciones.
- **Interpretar** y **comparar** las redes obtenidas.

