

# El análisis de las relaciones delimita el contexto para el NLP

Mariluz Congosto Martínez (@congosto)

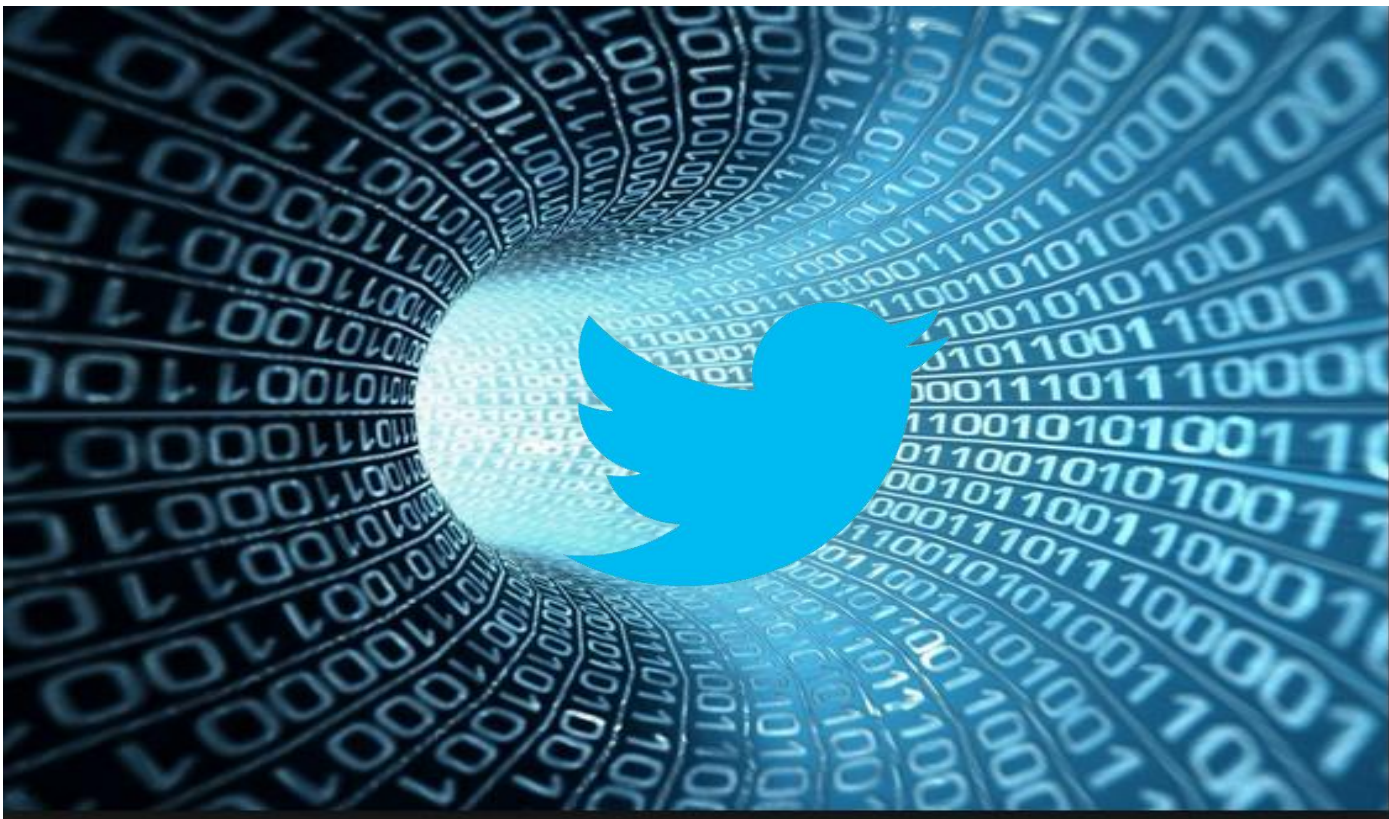
28 de noviembre de 2019




# ÍNDICE

1. Las redes sociales como fuentes de datos
2. Metodologías mixtas
3. Ejemplos

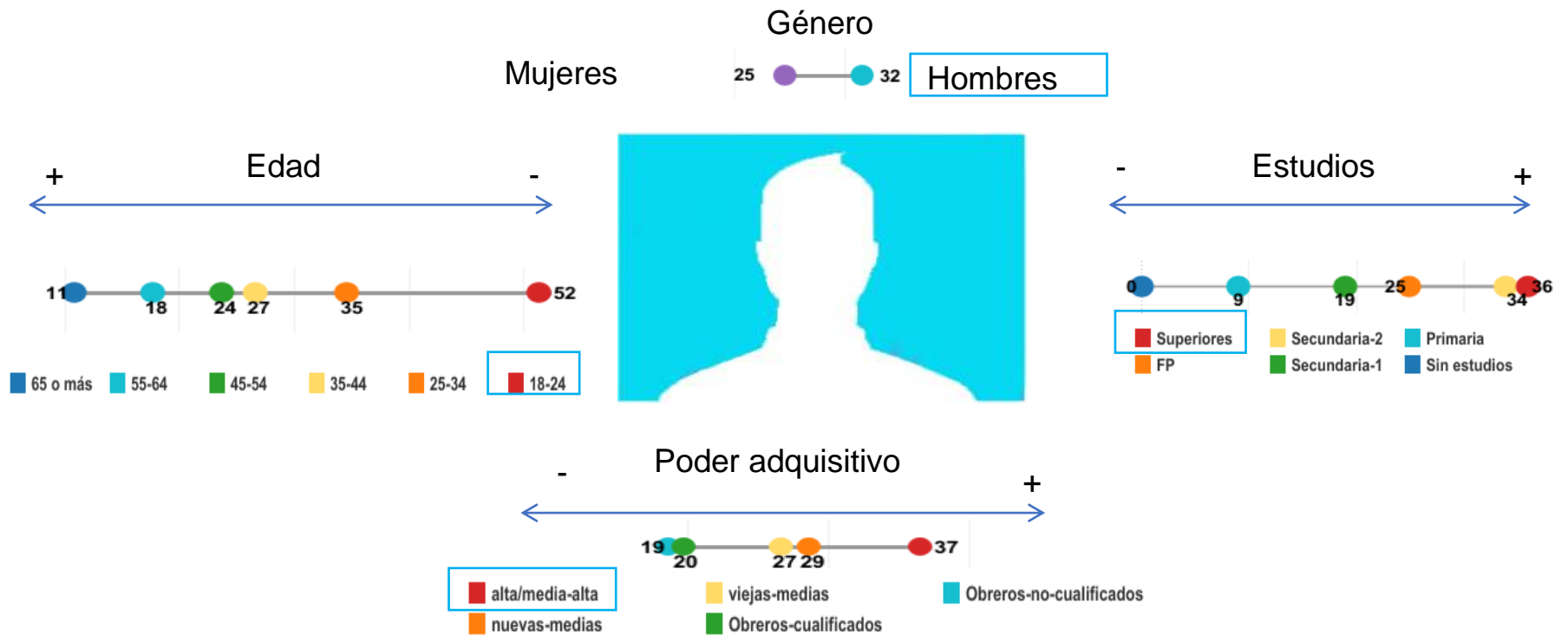
# 1. LAS REDES SOCIALES COMO FUENTE DE DATOS



# Pros y contras de los datos de Twitter

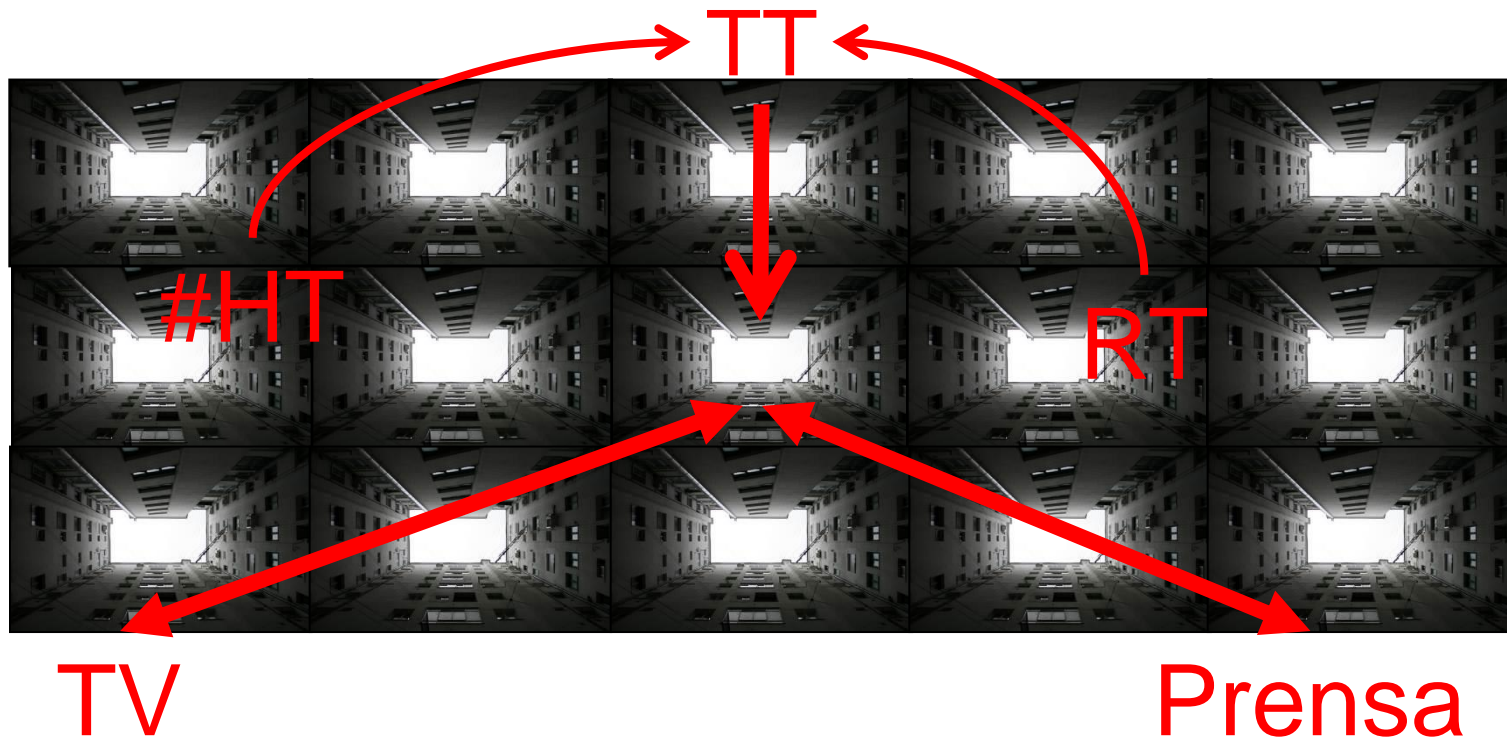
1	Los mensajes son públicos		1	Los usuarios no están segmentados
2	Es un medio utilizado para compartir y difundir opiniones		2	Anonimato: no todos son personas, hay personajes ficticios.
3	Tiempo real. Mide el pulso de la actualidad		3	Ruido: muchos trending topics están organizados
4	Tiene una API accesible		4	Hay opinión oculta y sobrerrepresentación

# Quién participa en Twitter

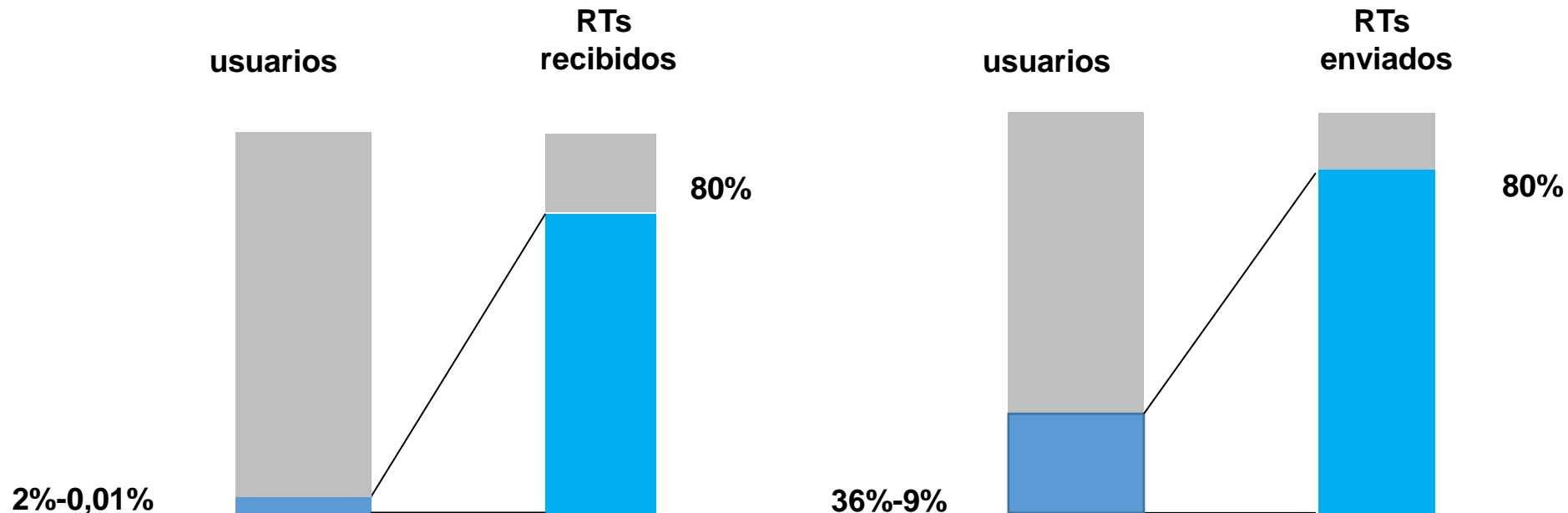


Fuente: Postelectoral elecciones. Generales 2019. Variables sociodemográficas. [http://datos.cis.es/pdf/Es3248sd\\_A.pdf](http://datos.cis.es/pdf/Es3248sd_A.pdf)

# Twitter de lo local a lo global



# Asimetría en la difusión



Fuente: Tesis Caracterización de usuarios y propagación de mensajes en Twitter en el entorno de temas sociales. <https://e-archivo.uc3m.es/handle/10016/22826>

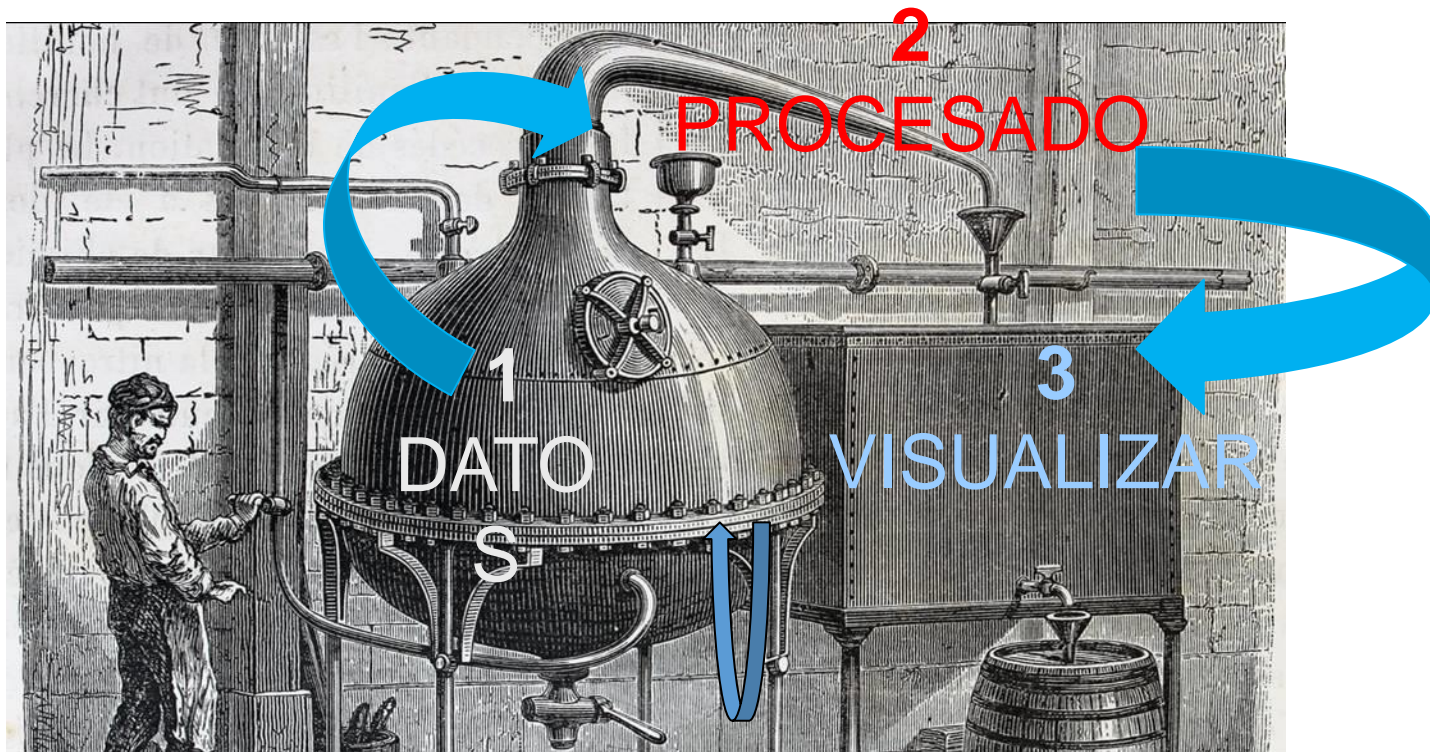


# Para analizar datos de Twitter hay que tener en cuenta...

- ✓ No hay una representación demográfica homogénea -Los estudios estadísticos tendrían mucho sesgo muestral
- ✓ Múltiples fuentes, muchas veces opacas -> Necesidad de caracterizar las fuentes
- ✓ Mensajes muy fragmentados -> Dificultad para analizar texto (discurso, análisis de sentimiento, etc..)
- ✓ Enormes cantidades de información -> Es imprescindible el análisis cuantitativo
- ✓ Dinámicas de comunicación -> Un mundo por explorar

# Metodologías mixtas

# Proceso de Análisis



# Captura de datos



- Se puede obtener todo lo publicado en a Web
- No hay rate limit
- Estructura de datos:
  - [HTML](#), [XHTML](#), [validated](#)
  - [JSON](#), [JSONP](#), [XML](#), [Microdata](#)

- Se puede obtener lo que ofrezca la API, a veces más, otras veces menos que en la Web
- Hay rate limit
- Menos volumen de información
- Estructura de datos:
  - [JSON](#)

# Uso de la API de Twitter: software



Python 2.7.12

## Tweepy

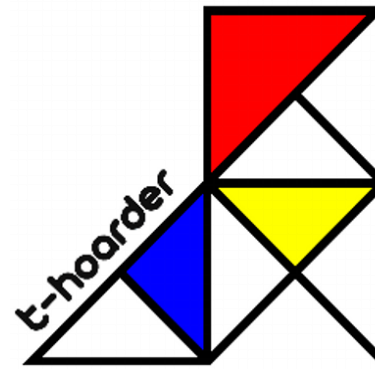
An easy-to-use Python library for accessing the Twitter API.

Fork 766 Star 2,144

[oauthlib](#)

[request\\_oauthlib](#)

[six](#)



[https://github.com/congosto/t-hoarder\\_kit](https://github.com/congosto/t-hoarder_kit)

# Máquina virtual



[https://www.dropbox.com/s/j0p26bmgmct3vll/como\\_instalar\\_VM\\_taller\\_datos\\_twitter.pdf?dl=0](https://www.dropbox.com/s/j0p26bmgmct3vll/como_instalar_VM_taller_datos_twitter.pdf?dl=0)

# Descarga de datos



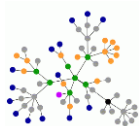
Autenticar



Obtener información de usuarios



Obtener tuits



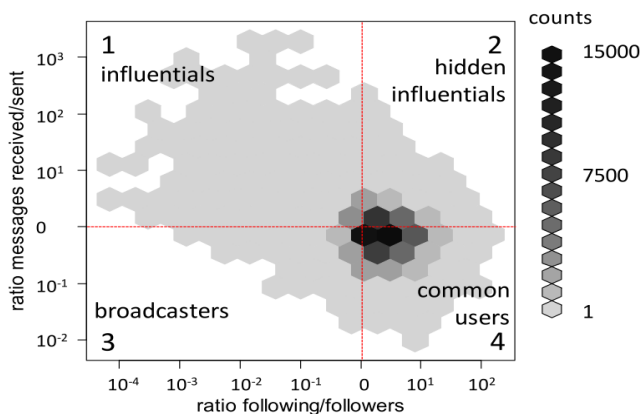
Obtener relaciones

# Concepto de Rol

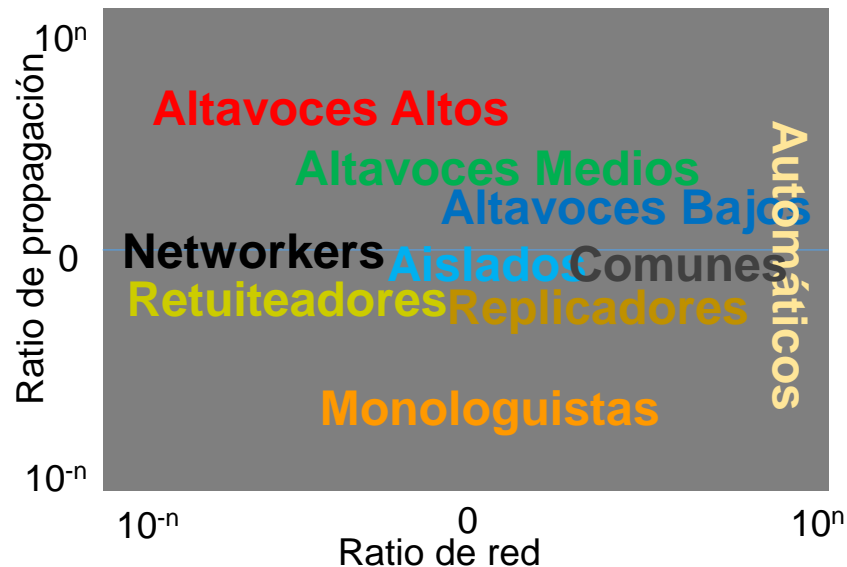
## Inspiración

Clasificación de los usuarios que participaron en la movilización social del 15M.

(González-Bailón, Borge-Holthoefer, & Moreno, 2012).



## Adaptación





# Clasificación por Roles

- Altavoces altos
- Altavoces medios
- Altavoces bajos

Capacidad de ser propagados

- Networkers
- Retuiteadores
- Replicadores

Alta interacción

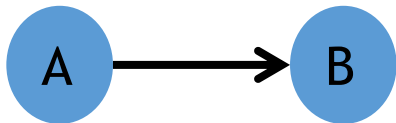
- Monologuistas
- Aislados
- Automáticos
- Comunes

Baja interacción

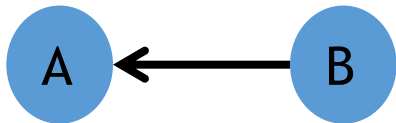
# Obtener relaciones

## Relaciones declaradas

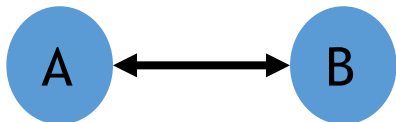
A sigue a B



A es seguido por B

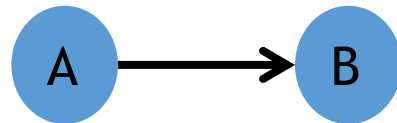


A y B se siguen mutuamente

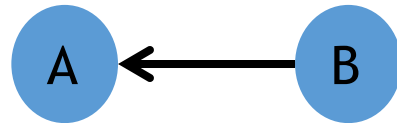


## Relaciones dinámicas

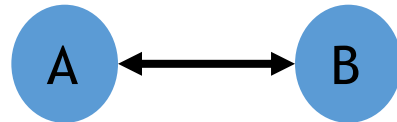
A retuitea a B



A es retuiteado por B



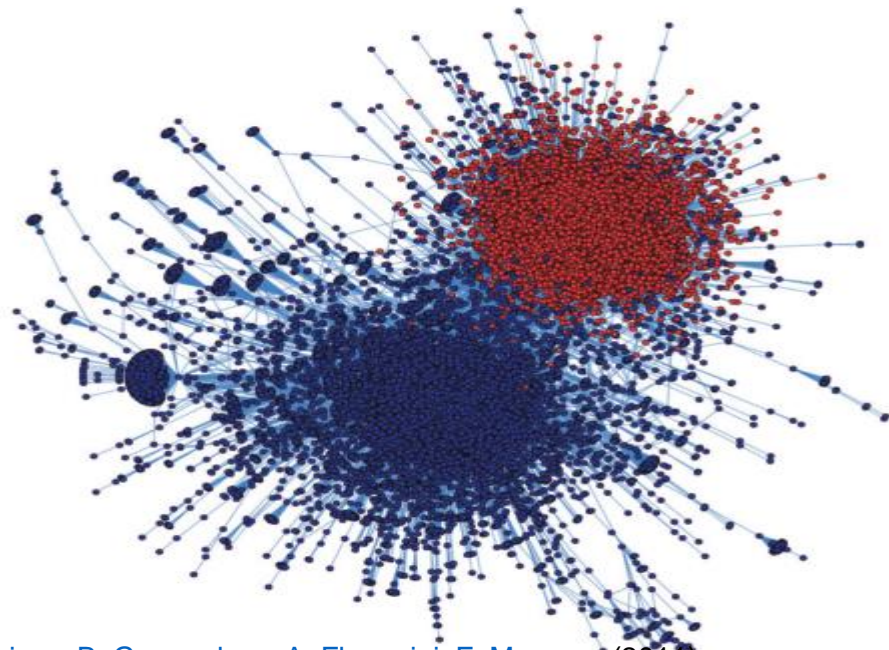
A y B se retuitean mutuamente



# Cámaras de eco

2010 U.S. congressional midterm elections

RT actúa como voto positivo y en las campañas electorales aflora comunidades de afinidad ideológica



[Political Polarization on Twitter](#) M. D. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, A. Flammini, F. Menczer (2011)

# Análisis de relaciones

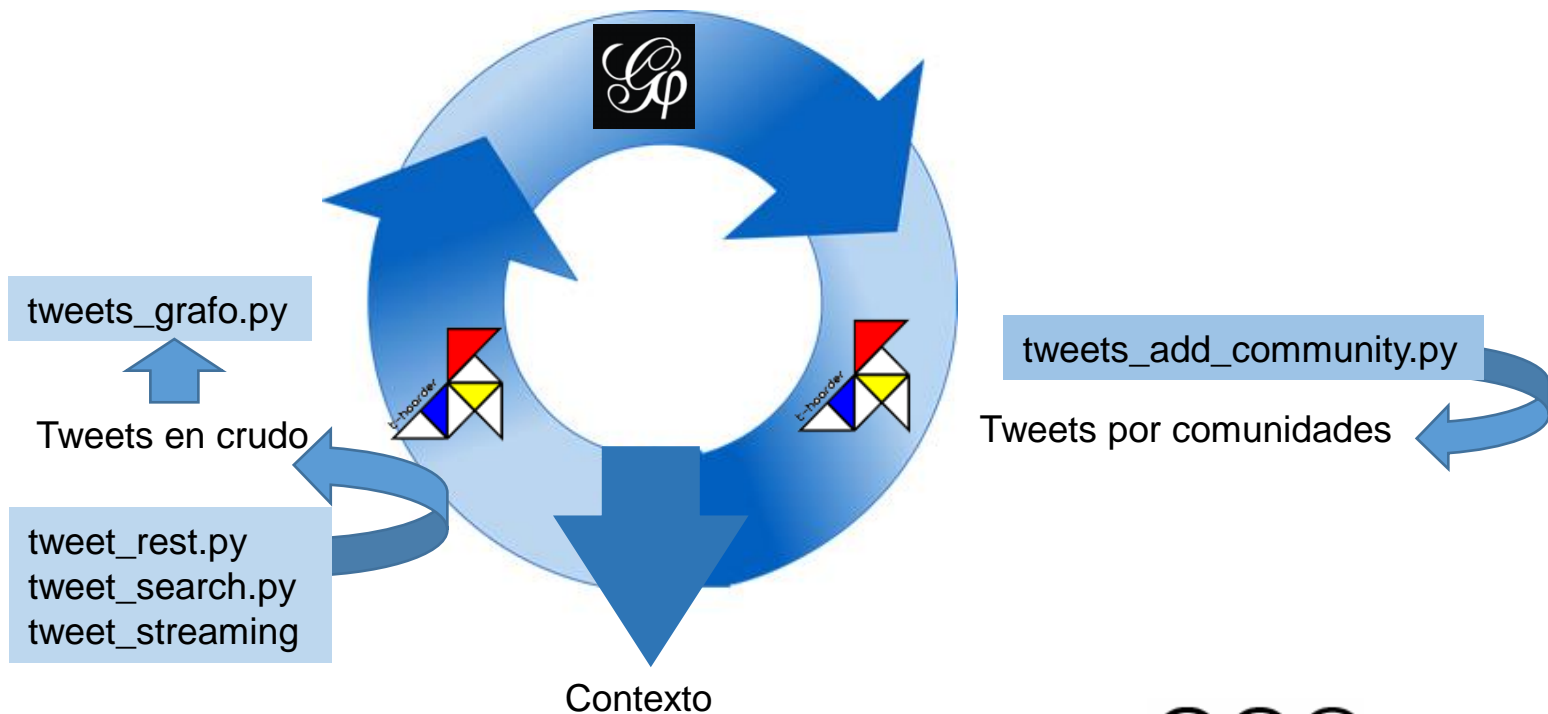


Gephi

- ✓ Calcula los parámetros globales de la red y de los nodos
- ✓ Permite filtrar nodos o conexiones
- ✓ Calcula la modularidad (comunidades)
- ✓ Da forma con distintos layouts
- ✓ Modela los nodos en tamaño y color según unos atributos
- ✓ Renderiza el grafo generando salida en pdf, png y svg

# Metodología

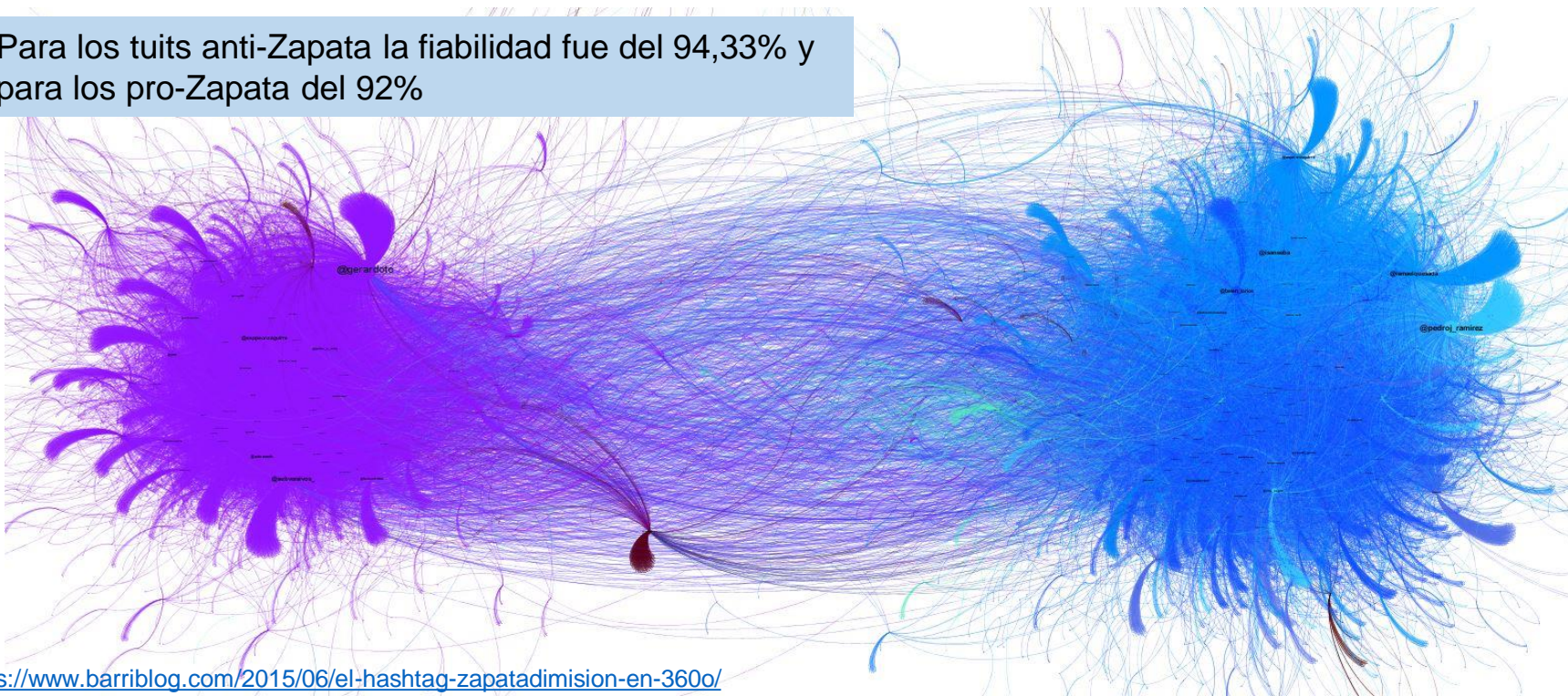
Análisis de relaciones



# Ejemplos

# Zapata dimisión (2015)

Para los tuits anti-Zapata la fiabilidad fue del 94,33% y para los pro-Zapata del 92%

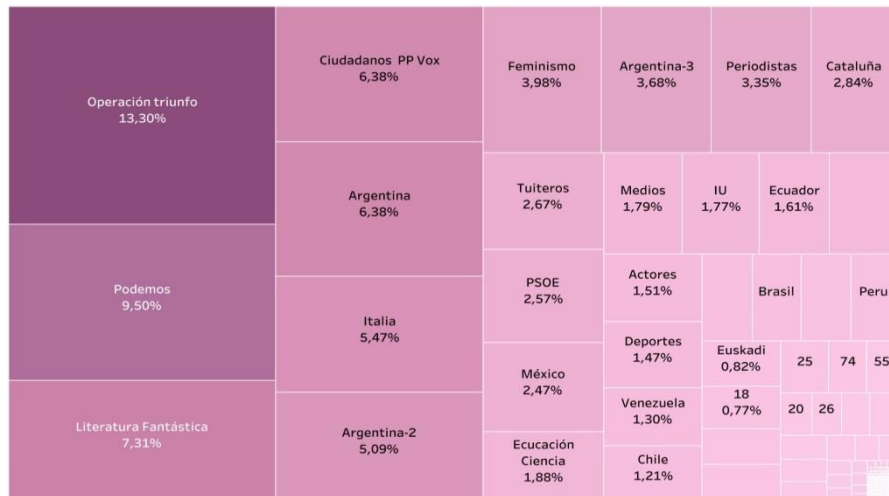
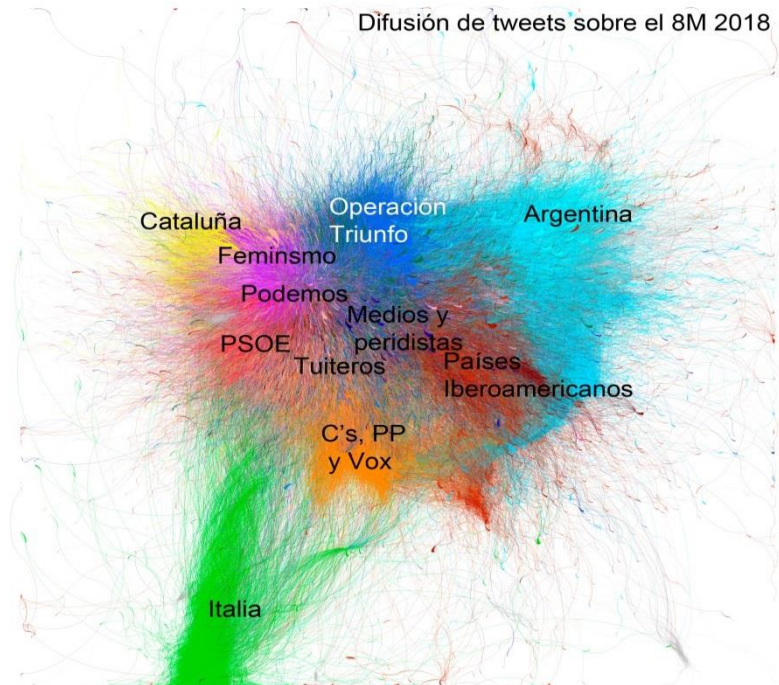


<https://www.barriblog.com/2015/06/el-hashtag-zapatadimision-en-360o/>



8M-2018

## Análisis de redes

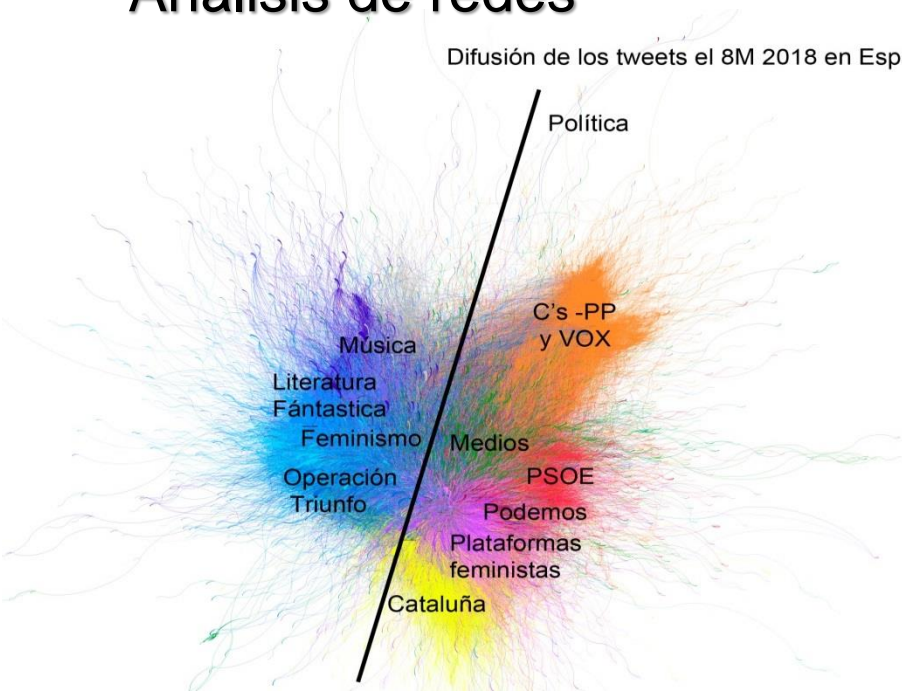




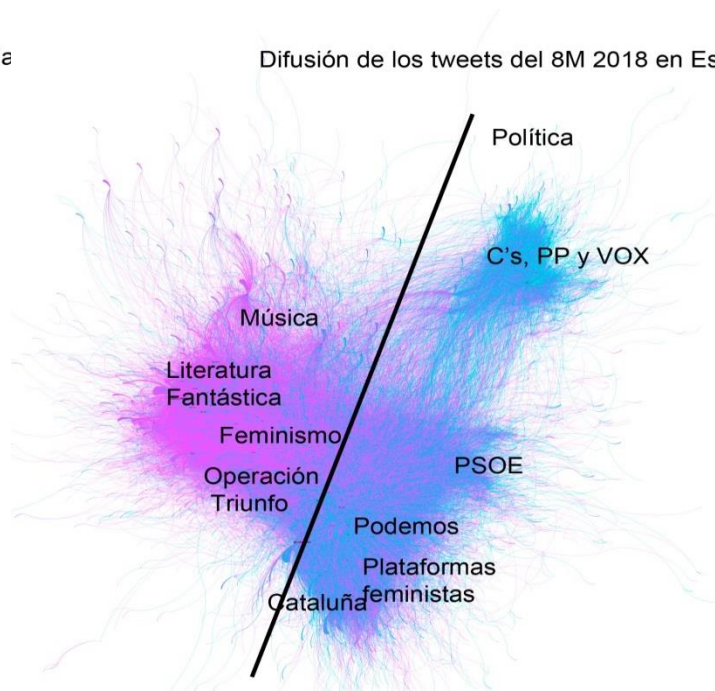
8M-2018

## Análisis de redes

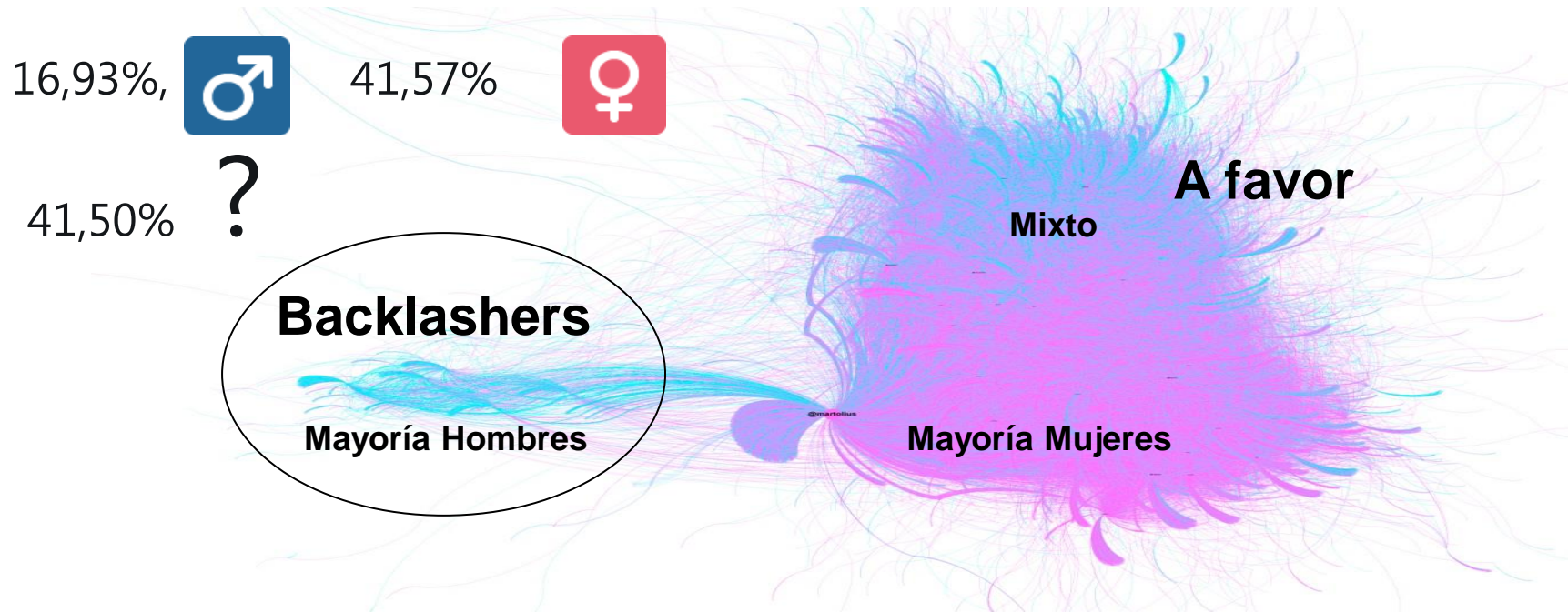
Difusión de los tweets el 8M 2018 en España



Difusión de los tweets del 8M 2018 en España por género

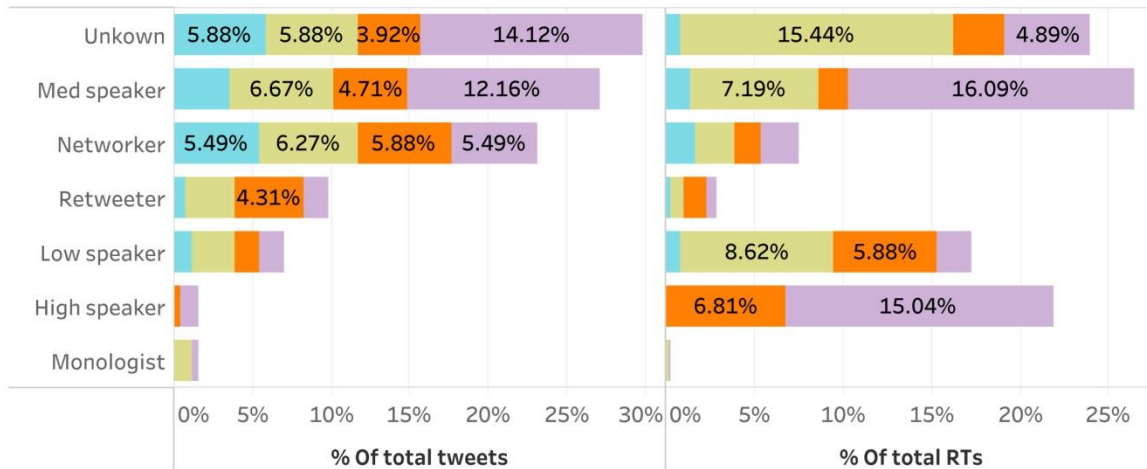


# Propagación de #cuéntalo (Género)



# Temas por role de los backlashers

Themes by role (tweets and retweets)



## Theme

MEN (men being abused)

MAN (La Manada and court sentences)

LEF (biased left)

FEM (Tyrannical feminism)

# Resumiendo

Las redes sociales son una **fuentes continua** de datos para analizar

Cuando se difunden, se citan o se comentan los mensajes, se le incorpora **la relación**

La actividad y el impacto caracterizan a los usuarios en **roles**

El análisis de red nos va a detectar las **cámaras de eco**. Nos definen mejor las relaciones que los contenidos

Dar **contexto** a los datos ayuda, complementa y enriquece el NLP