### Máster en Business Analytics y Big Data Edición 2016 / 2017



Asignatura: Análisis de Redes Sociales

Trabajo: Movimientos migratorios internos en España durante 2015

Autor: Juan Antonio garcía Cuevas

Fecha: Diciembre 2016

# Índice

1. Introducción	3
2. Obtención de los datos	4
3. Limpieza y preparación de los datos	5
4. Importación en Guephi	6
5. Configuración del Layout	7
6. Análisis estadísticos	8
7. Laboratorio de datos	12
8. Cuestiones	13
9. Conclusión	14

### 1. Introducción

El Análisis de Redes Sociales es una técnica de investigación orientada a identificar y comparar las relaciones entre individuos, grupos y sistemas, con el objetivo de modelar las interacciones humanas que son el centro de los procesos de conocimiento y aprendizaje organizacional.

Gephi es una poderosa herramienta para el análisis de redes sociales.

Este documento constituye el entregable principal de la práctica de la asignatura, y trata de los movimientos migratorios interprovinciales en España durante el año 2015.

### 2. Obtención de los datos

El **Instituto Nacional de Estadística** de España (**INE**) ha creado un sitio web para facilitar a los usuarios el acceso la información estadística y otros contenidos divulgativos.

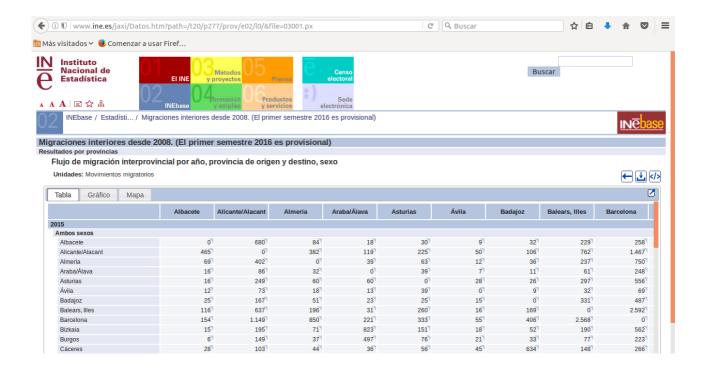
#### http://www.ine.es/inebmenu/indice.htm

Este sitio web incluye la **Estadística de Migraciones**, una aproximación estadística de los flujos migratorios de España, de cada comunidad autónoma y de cada provincia con el extranjero, así como de las migraciones interautonómicas e interprovinciales.

Sus resultados están desagregados por sexo, año de nacimiento, edad, país de nacionalidad y país de nacimiento del migrante, así como por origen y destino de la migración, y guardan completa consistencia con las Cifras de Población y las Estadísticas del Movimiento Natural de la Población.

El sitio web permite realizar operaciones de filtrado pre-establecidas para obtener la información deseada antes de exportarlo a un fichero en formato CSV, Excel, etc.

Para nuestra práctica obtendremos información de las *migraciones de población entre provincias* (*y ciudades autónomas*) *españolas en el año 2015*, y la descargaremos en un fichero CSV.



Consulta en la web del INE

El fichero CSV resultante contiene una tabla de datos con la lista de nombres de provincias y ciudades autónomas tanto en la primera fila como en la primera columna, presentando en las intersecciones el número de migrantes desde una provincia a otra.

Fichero CSV original

### 3. Limpieza y preparación de los datos

Pero el fichero CSV contiene además algunas filas extra de texto, espacios en blanco, diferencias de codificación de caracteres, decimales, etc. que deberemos limpiar antes de procesar los datos. Para ello usamos, en primer ligar, un editor de texto plano.

```
| Albacete; Alicante; Almeria; Alava; Asturias; Avila; Badajoz; Baleares; Barcelona; Vizcaya; Burgos; Cáceres; Cádiz; Cantabria; Castellón; Ciudad Real; Córdoba; Coruña; Cuenca; Guipizcoa; Gerona; Granada; Guadala jara; Huelva; Huesca; Jaén; León; Lérida; Lugo; Madrid; Málaga; Murcia; Navarra; Orense; Palencia; Las Palmas; Pontevedra; La Rioja; Salamaca; S. C. Tenerife; Segovia; Sevilla; Soria; Tarragona; Teruel; Toledo; Valencia; Valladolid; Zamora; Zarragoza; Ceuta; Mellila Rioja; Salamaca; S. C. Tenerife; Segovia; Sevilla; Soria; Tarragona; Teruel; Toledo; Valencia; Valladolid; Zamora; Zarragoza; Ceuta; Mellila Rioja; Salamaca; S. C. Tenerife; Segovia; Sevilla; Soria; Tarragona; Teruel; Toledo; Valencia; Valladolid; Zamora; Zarragoza; Ceuta; Mellila Rioja; Valla; Salamaca; S. C. Tenerife; Segovia; Sevilla; Soria; Tarragona; Teruel; Toledo; Valencia; Valladolid; Zamora; Zarragoza; Ceuta; Mellila Rioja; Valla; V
```

Fichero CSV limpiado

Después creamos un script Python para leer el fichero CSV y generar dos nuevos ficheros CSV, uno con los 52 nodos correspondientes a provincias y ciudades autónomas, y otro con las 2.637 aristas que relacionan cada par de provincias para las que ha habido movimiento migratorio.

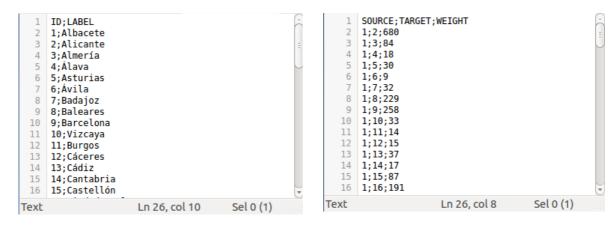
Script Python

El fichero CSV de nodos contiene dos columnas:

- ID: identificador de provincia (numérico secuencial).
- LABEL: nombre de provincia.

El fichero CSV de aristas contiene 3 columnas:

- SOURCE: identificador numérico del nodo origen (provincia de origen).
- TARGET: identificador numérico del nodo destino (provincia de destino).
- WEIGHT: número de migrantes.



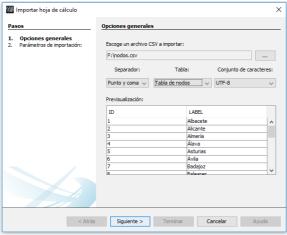
Fichero CSV de nodos

Fichero CSV de aristas

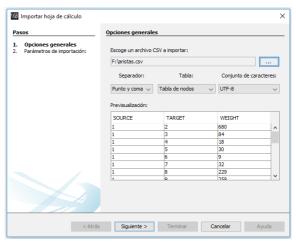
### 4. Importación en Gephi

Para importar el grafo en Gephi hemos de importar primero el fichero de nodos y después el de aristas.

La importación de los ficheros se realiza mediante la opción "Importar hoja de cálculo" del menú "Archivo".

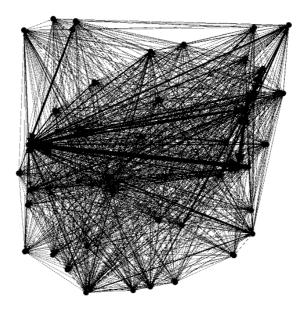


Importar fichero CSV de nodos



Importar fichero CSV de aristas

El grafo se muestra inicialmente de la siguiente manera:

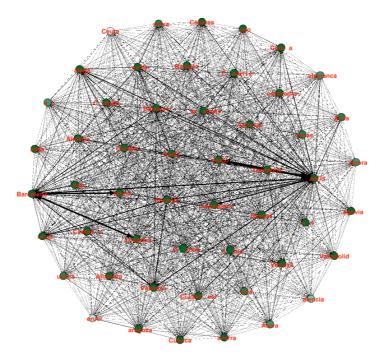


Aspecto del grafo por defecto

## 5. Configuración del Layout

A la izquierda de la interfaz de Gephi se encuentra el panel inferior algoritmos de distribución para componer la forma gráfica.

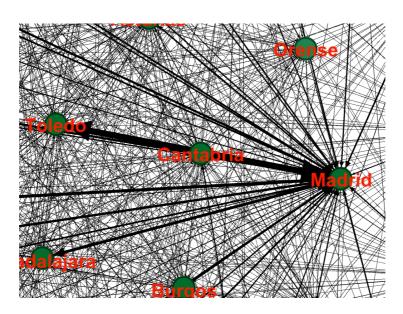
Elegimos Fruchterman Reingold y pulsamos Ejecutar. Activamos las etiquetas y modificamos el color de éstas y de los nodos. El resultado es el siguiente:

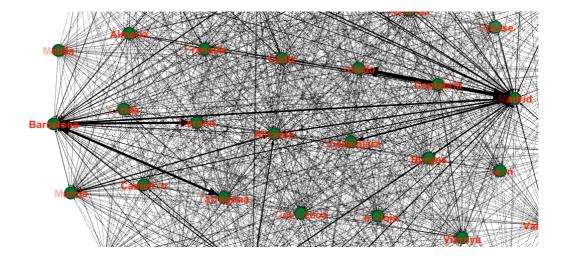


Aspecto del grafo con el algoritmo Fruchterman Reingold

Podemos ver en la nueva representación del grafo dos zonas en las que las líneas son especialmente gruesas. Estas indican dónde se produjeron los mayores movimientos migratorios en 2015.

Si ampliamos las zonas (fácilmente visibles), podemos ver que los movimientos migratorios más numerosos se dan entre Madrid y Toledo, mayor con origen en Toledo y destino en Madrid. También entre Barcelona, Tarragona y Gerona, mayor hacia Barcelona.





### 6. Análisis estadísticos

Existen diversas métricas y análisis estadísticos:

#### ■ Métricas de Centralidad:

- Centralidad por grado: Número de enlaces que posee un nodo con los demás.
- **Centralidad por cercanía**: Promedio de las distancias más cortas desde un nodo hacia todos los demás.

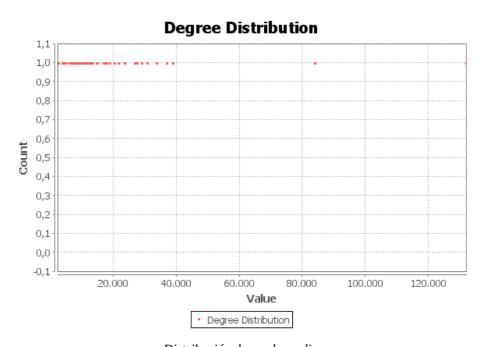
- **Centralidad por intermediación**: Frecuencia o número de veces que un nodo actúa como un puente a lo largo del camino más corto entre otros dos nodos.
- **PageRank**: El cálculo del PageRank de Google, utilizado para medir la relevancia de páginas web en Internet, es una variante de la medida Centralidad de vector propio.

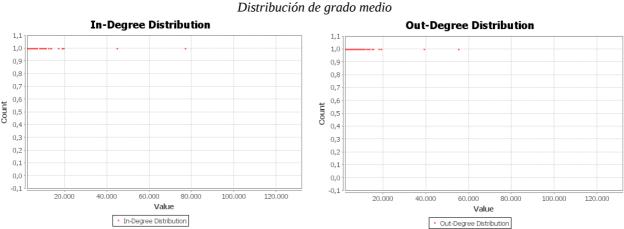
### ■ Métricas de componentes conectados:

- **Componentes fuertemente conectados**: Cada nodo dentro del componente se puede alcanzar desde cada otro nodo del componente siguiendo los enlaces dirigidos.
- Componentes débilmente conectados: cada nodo puede ser alcanzado desde cualquier otro nodo siguiendo los enlaces en cualquier dirección. En componentes no dirigidos en conversaciones simplemente sobre componentes conectados.
- **Componente gigante**: si el componente más grande abarca una fracción significativa del gráfico, se llama componente gigante.

Gephi nos proporciona en el panel situado a la derecha de la pantalla herramientas para obtener algunos estadísticos:

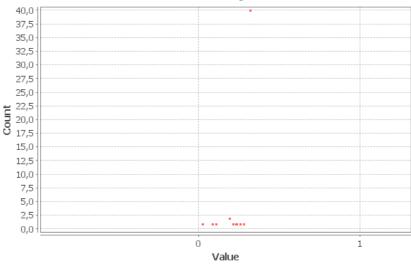
### **Degree Distribution**



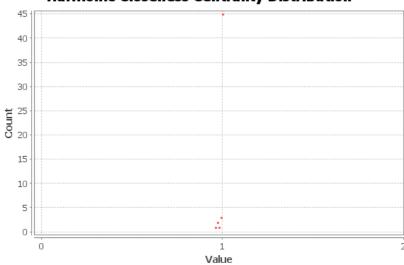


### **Betweenness, Closeness, Eccentricity**

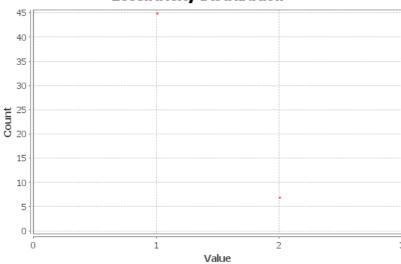




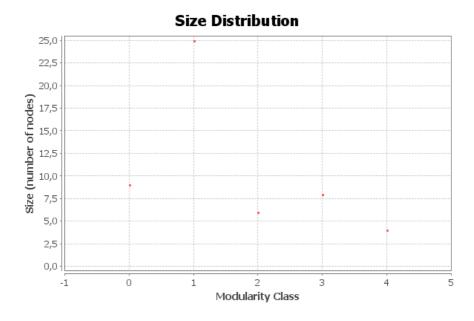
### **Harmonic Closeness Centrality Distribution**



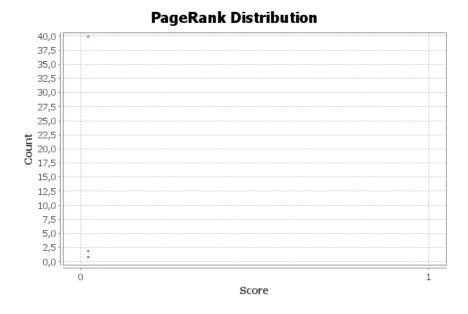
### **Eccentricity Distribution**



### Modularity

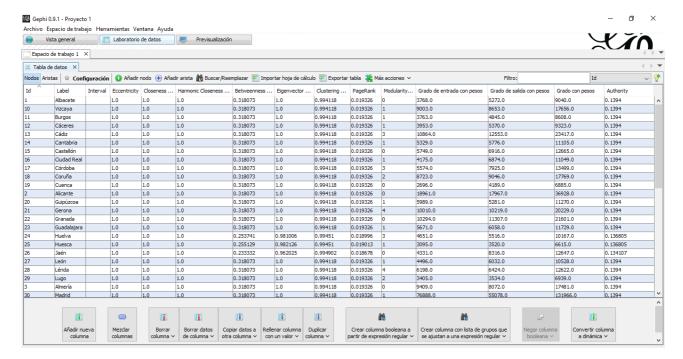


### **PageRank**

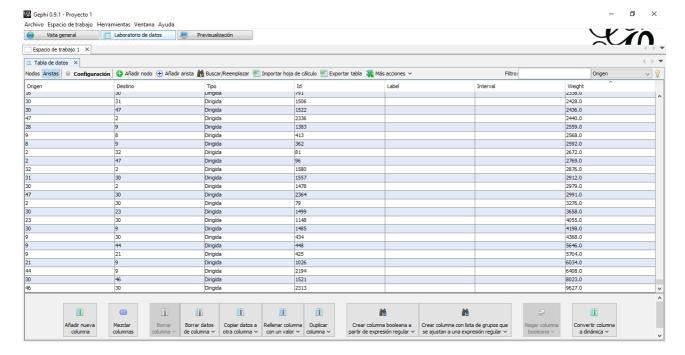


### 7. Laboratorio de datos

Gephi cuenta con una herramienta en forma de tabla denominada Laboratorio de Datos, que nos muestra los cálculos de todos los algoritmos tanto sobre nodos como sobre aristas, y nos permite jugar con ellos para detectar y resolver cuestiones.



Laboratorio de datos - Nodos



Laboratorio de datos - Aristas

### 8. Cuestiones

### Inspeccionando visualmente el grafo

#### ¿Qué provincias reciben mayores movimientos migratorios?

Según el grafo, las provincias españolas que reciben mayores flujos migratorios son Madrid en primer lugar y Barcelona en segundo lugar.

#### ¿Qué provincias presentan mayor flujo migratorio con Madrid?

Según el grafo, la provincia de Toledo es la que mayor flujo migratorio aporta a Madrid. Otras provincias destacadas son Ciudad Real y Guadalajara, de la comunidad castellano manchega, y las provincias mediterráneas de Barcelona y Alicante.

#### ¿Qué provincias presentan mayor flujo migratorio con Barcelona?

Según el grafo, la provincia de Barcelona recibe mayor flujo migratorio de las provincias de Tarragona y Gerona, ambas de la misma comunidad autónoma. Curiosamente no recibe mucha migración de Lérida, la otra provincia catalana. Sigue Madrid como provincia que más intercambia flujos migratorios con Barcelona.

#### Consultando el laboratorio de datos

#### ¿Qué provincias soportan mayor flujo migratorio?

Según el grado con peso, Madrid se encuentra en primer lugar, con unos 132.000, le sigue Barcelona con casi 84.000, y un poco más lejos Valencia, Alicante, Málaga y Baleares con valores entre 30.000 y 38.000.

#### ¿Qué provincias soportan menor flujo migratorio?

La provincia con menor flujo migratorio es Soria, con unos 2.500 aproximadamente. Se siguen Palencia, Ceuta, Teruel, Zamora y Melilla con valores entre 3.700 y 4.500.

### • ¿Cual es el flujo migratorio más importante?

La información del laboratorio de datos corroboran lo que veíamos en el grafo: el mayor flujo migratorio en 2015 fue de 9.627 individuos desde la provincia de Toledo a la provincia de Madrid, seguida por el flujo inverso (de Madrid a Toledo), con 8.025 individuos.

### ¿En qué provincias no son significativos, numéricamente hablando, los movimientos migratorios?

Si en el laboratorio de datos eliminamos las aristas con peso inferior a 1.000, podemos ver los nodos que quedan con grado 0. Estos corresponden a las provincias con movimientos migratorios inferiores a 1.000 individuos: Teruel, Álava, Guipúzcoa, La Rioja, Palencia, Burgos, Orense, Zamora, Soria, Albacete y Ceuta.

# • ¿Con qué otras provincias mantienen mayores flujos migratorios provincias como Córdoba, Jaén, Badajoz y Cáceres?

Utilizando el grafo modificado anteriormente, podemos ver que Córdoba y Jaén mantienen flujos migratorios principalmente con Madrid y Málaga, mientras que las provincias extremeñas, Badajoz y Cáceres, con Madrid.

#### • ¿Cuáles son las cifras más relevantes de los flujos migratorios?

Utilizando el grafo modificado anteriormente, podemos ver que, al eliminar las aristas con menos de 1.000 individuos, el número de aristas restantes en el grafo ha bajado drásticamente desde 2.637 a sólo 93. Por tanto, el 96% de los movimientos migratorios son de menos de 1.000 individuos, y sólo el 4% son de un número más relevante de individuos.

### 9. Conclusión

La red importada está constituida por 52 nodos correspondientes a las provincias y ciudades autónomas de origen/destino y 2637 aristas que corresponden a los flujos migratorios entre provincias españolas durante el año 2015.

Todos estos datos componen un grafo dirigido en el que cada arista indica el número de migrantes y el sentido de esa migración. En el nodo de salida el flujo de migrantes se denomina emigración y en el de llegada inmigración.

Algunos de los datos más relevantes descubiertos durante la exploración de la red son:

- Las provincias con mayor flujo migratorio son Madrid y Barcelona.
- Madrid recibe migración de todas las provincias, pero principalmente de Toledo, y en segundo lugar de Ciudad Real, Guadalajara y Barcelona.
- Barcelona recibe inmigración principalmente de dos provincias de su misma comunidad autónoma, Tarragona y Gerona, pero curiosamente muy alejado de la otra provincia de la comunidad, Lérida.
- La provincia que soporta menor flujo migratorio es Soria, seguida por Palencia, Ceuta, Teruel, Zamora y Melilla.