75.19 Teoría de la Comunicación

Pablo Notari - 88548

Año 2017

Índice

Justificación

Las redes complejas toman más relevancia en los últimos años con el crecimiento de las redes sociales y el comercio electrónico como grandes generadores de datos. Lo que da pie a otra rama de la informática dedicada a resolver problemas de almacenamiento y análisis de datos llamada “big data”.

Las soluciones a problemas de “big data” distan en tecnología y en técnicas (algoritmos, modelos, etc) de los problemas tradicionales de tratamiento de los datos. Las redes complejas ocupan un gran lugar dentro de esta rama ya que son de gran ayuda para conceptual izar y analizar grandes volúmenes de datos y dan un soporte solido al modelado de problemas de “big data” permitiendo analizar y comparar los datos desde los distintos indicadores que provee una red compleja.

Fuente de datos

Los datos a analizar tienen que ver con distintos indicadores de países, tales como: PBI, esperanza de vida, gasto en salud per capita, etc. toda estos datos están centralizados y disponibles en el proyecto LN Data del diario La Nación [LN]. Pero la fuente original de datos son organismos internacionales y serán citados a medida que se mencionen.

Tareas

- Análisis de tema a desarrollar
  - Bibliografía
  - Papers
- Análisis de fuentes de datos
  - LN Data
  - Network Repository [NR]
- Selección de herramientas
  - Modelado de Grafos → Gephi
  - Depuración → Java y MySQL
- Depuración de datos
- Pruebas de concepto

Análisis de Fuentes de Datos

Trabajando con los datos provistos por Network Repository se llegaron a buenos resultados en las pruebas de concepto relacionadas con los tipos de distribución que presentan las distintas categorías de redes, por ejemplo analizando redes sociales en la

mayoría de los casos se obtenía una distribución “power-law” (gráfico 1). Es una fuente de datos muy poderosa pero requiere algo de trabajo entender que tipos de relaciones se están modelando, más que nada en las relaciones de tipo biológicas, es algo a considerar para futuras investigaciones en las que se busque profundidad en el análisis de distribuciones y cualquier otra característica de rede complejas.

Los sets provistos por La Nación Data son más sencillos de analizar ya que dan datos sobre dominios ya conocidos, lo que da la posibilidad de enfocarse en la construcción y análisis de la red y no tanto en la comprensión .

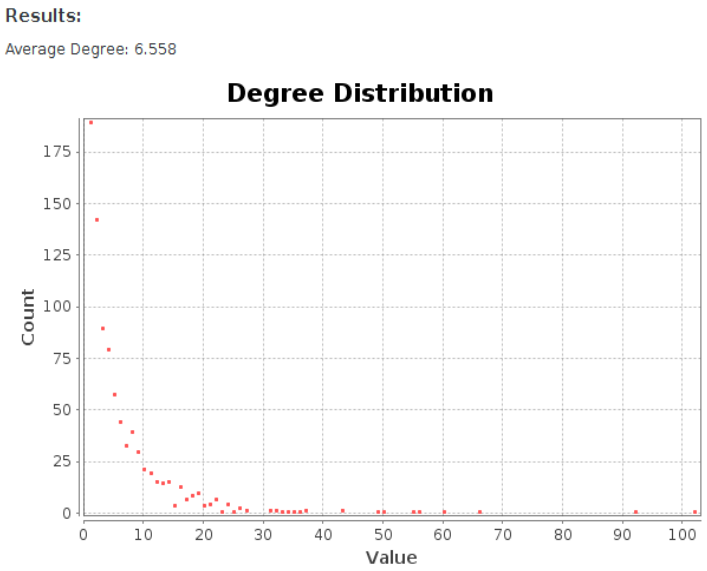

Gráfico 1, Wikipedia votes [NR].

Descripción

En este trabajo practico tomaremos como central el clustering de una red compleja creada a partir de datos crudos, con el fin mostrar con ejemplos como el clustering en una red compleja da información de como se relacionan los componentes de esa red, mostrando cómo el clustering relaciona nodos que en principio no son tan evidentes cuando se analizan varios indicadores. En redes sociales pueden modelar grupos de interés, desde una comunidad educativa hasta filiación política. En los ejemplos que se desarrollan en este trabajo se pueden observar no solo los agrupamientos que se generan sino también quienes son los nexos entre distintos agrupamientos.

Convenciones

En el desarrollo de este trabajo se toman las siguientes convenciones

- Los grafos son todos no dirigidos.
- Las aristas no tienen peso o el peso es 1 para todas.

- No se consideran los bucles.

Modelado

El trabajo se desarrolla con el objetivo de mostrar como se arman los clusters (agrupamientos) en redes complejas donde las aristas se generan por dos o más Indicadores entre dos nodos, en este caso países.

Depuración de Datos

Los datos estaban desnormalizados, esto significa que los nombres de los países estaban en cada set de datos sin ningún código de referencia y mencionados de distinto modo, la tarea de normalizar los nombres y crear una tabla para países se hizo manualmente.

Por otro lado fue necesario asociar cada país a un continente, para esta tarea se hizo uso de la api que provee Rest Countries [RC] y un pequeño desarrollo en Java para obtener los continentes de todos los países, de todos modos posterior a esto fue necesario un trabajo manual para los países que no tenían los nombres completos o eran incorrectos.

Por último se normalizaron los números de cada set de datos ya que estaban expresados de distinto modo, algunos con coma otros con punto, etc.

Con estas tareas se llego a una base de datos de la que se puede extraer información de varios indicadores y varios años, tales como: PISA – Ciencia, Índice de fertilidad, Libertad Económica, Nivel de Inglés, Población, Ranking FIFA, PBI per capita, Progreso Social, Esperanza de Vida, Competitividad y varios más.

Metodología

Para el desarrollo del primer caso de estudio se consideran dos variables de las que se cuentan con datos del año 2015:

- PBI per Capita [TWB]
- Esperanza de Vida [WHO]

Cada variable tiene un máximo (M) y un mínimo(m) y tiene asignado un porcentaje de variación de manera arbitraria (p). Ademas cada país cuenta con un score (s) para una variable determinada. Dados dos países, p1 y p2, entonces p1 está relacionado con p2 para la variable v, si:

$$p1.s - (v.M - v.m)/100*v.p <= p2.s <= p1.s + (v.M - v.m)/100*v.p$$

Por ejemplo para la variable PBI per capita:

Cais	Score
Argentina	13431,88
Letonia	13648,55

Variable	Mínimo	Máximo	%
PBI per Capita	100,3	9999,09	10

$$argentina.s - (pbi.M - pbi.m)/100*pbi.p <= letonia.s <= argentina.s + (pbi.M - pbi.m)/100*pbi.p$$

$$13431,88 - (100,3 - 9999,09)/100*10 <= 13648,55 <= 13431,88 + (100,3 - 9999,09)/100*10$$

$$13431,88 - 216,67 <= 13648,55 <= 13431,88 + 216,67$$

$$13215,21 <= 13648,55 <= 13648,55$$

Por lo tanto Argentina esta relacionado con Letonia. Con esta información se construye la matriz de adyasencias que luego se procesa con Gephi se aplica un layuot “Fruchterman Reingold”[FR] para que la visualización de los cluster quede en evidencia (gráfico 2).

Una segunda red se generó con las variables también de año 2015:

- PBI per Capita [TWB]
- Esperanza de Vida [TWB]

Se utilizó el mismo método para la visualización de los clusters, se puede ver en el gráfico 3.

Una tercera red se generó con las variables también de año 2014:

- Gasto en salud per Capita [TWB]
- Igualdad de Genero [WEF]

Se utilizó el mismo método para la visualización de los clusters, se puede ver en el gráfico 4.

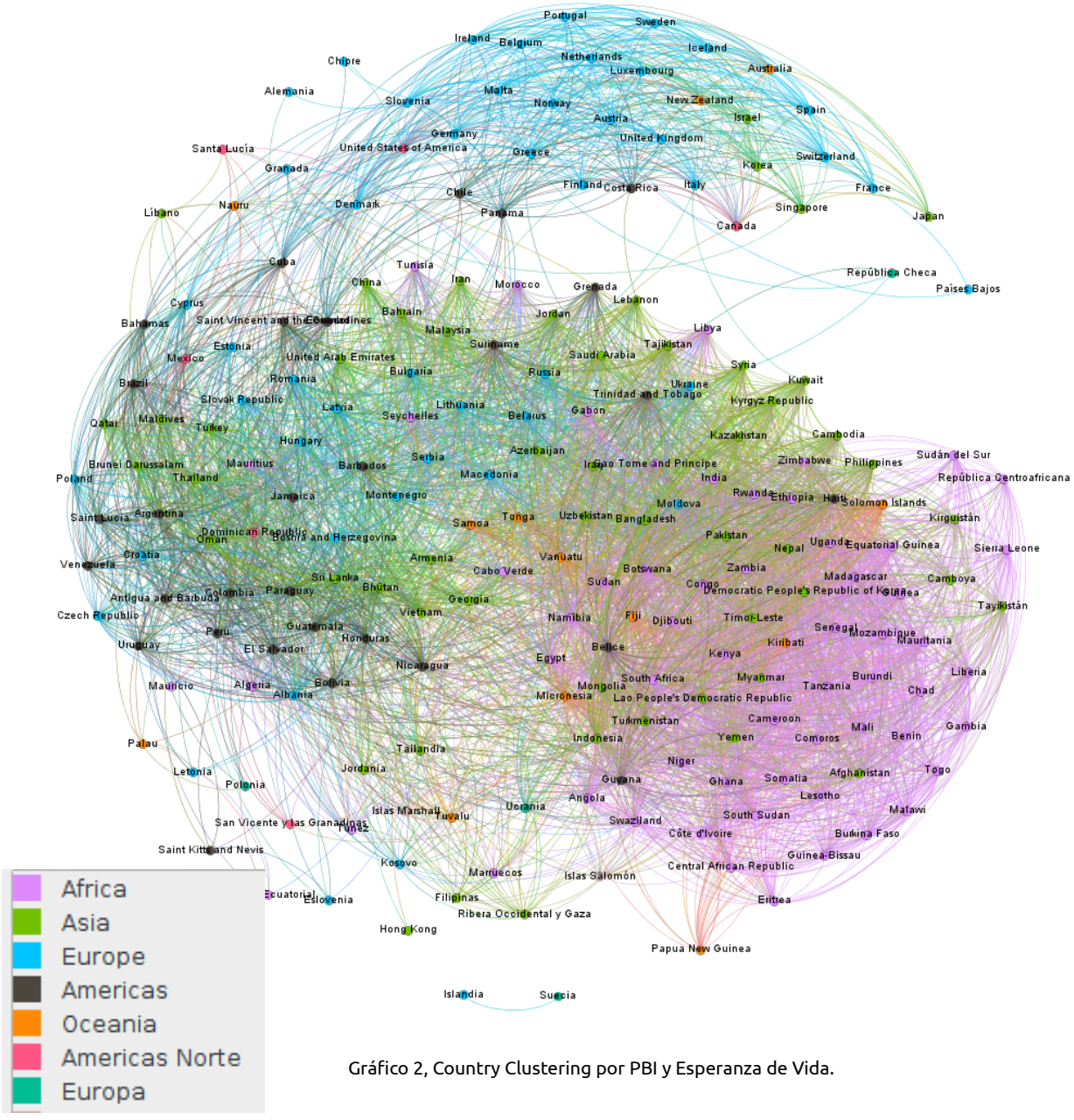

Gráfico 2, Country Clustering por PBI y Esperanza de Vida.

DataSet	#Nodes
PBI per Cápita	180
Ezperanza de Vida	181

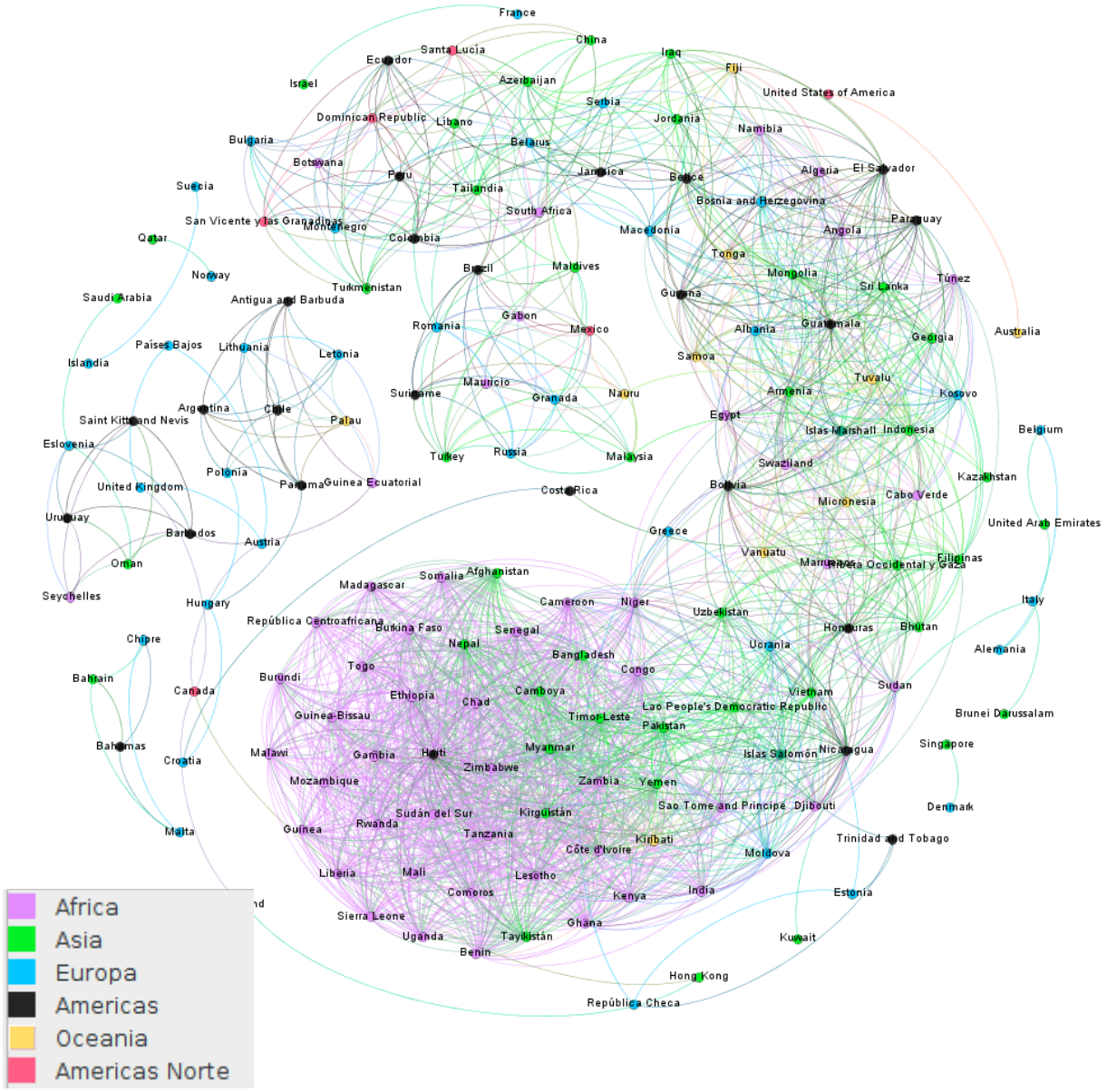

Gráfico 3, Country Clustering por PBI y Población.

DataSet	#Nodes
PBI per Cápita	180
Población	211

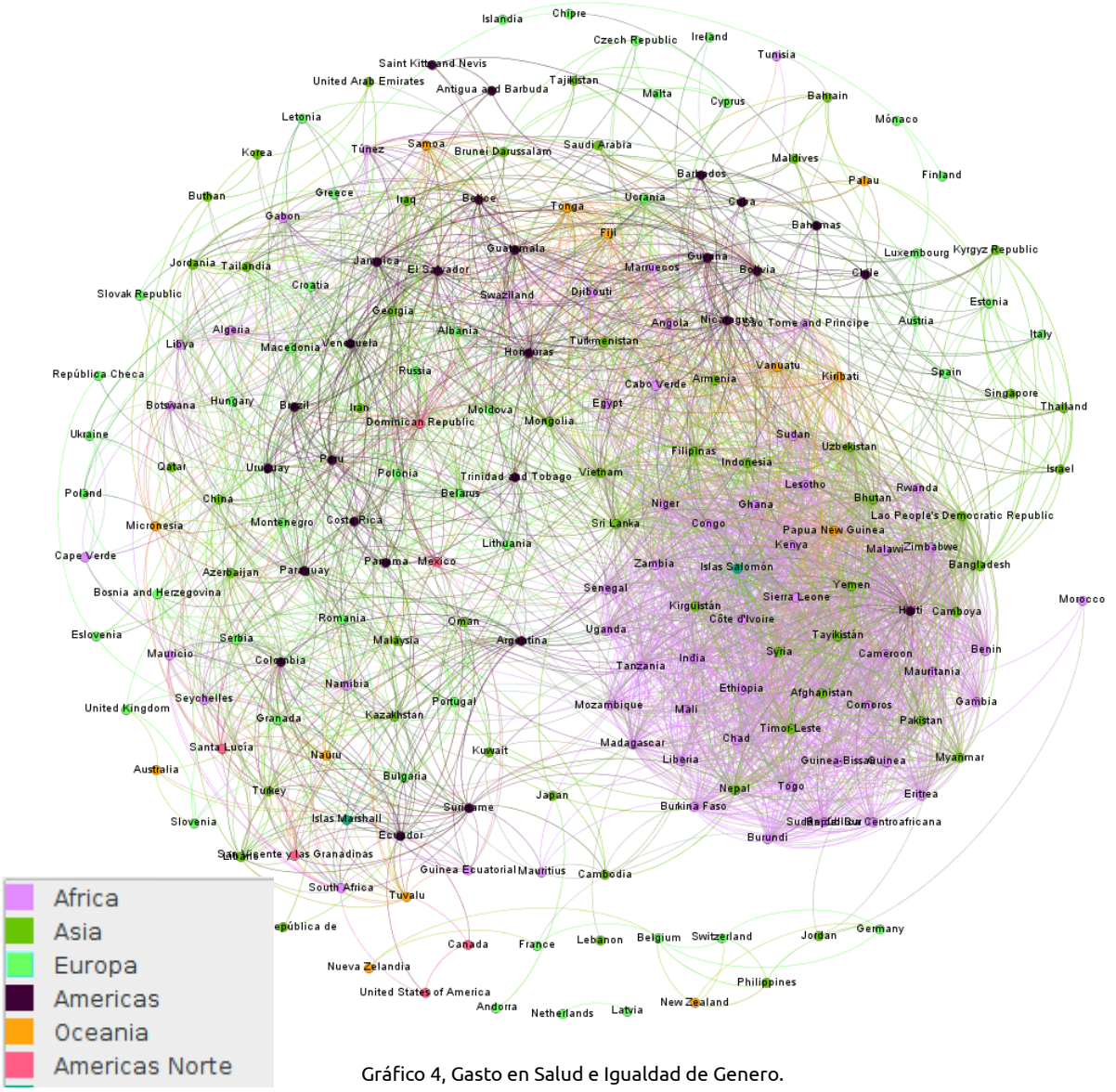

Gráfico 4, Gasto en Salud e Igualdad de Genero.

DataSet	#Nodes
Gasto en Salud per Capita	186
Igualdad de genero	142

Conclusiones

Se puede observar en los distintos gráficos presentados como utilizando distintas variables para generar las adyasencias las regiones en algunos casos se diferencian de forma marcada y en otras se entremezclan, un caso excepcional es el continente africano, que en todos los casos se agrupa casi uniformemente, esto da una aproximación, en principio, a las características particulares de dicho continente y da una

muestra de cómo el análisis de redes complejas se puede utilizar para genera información a partir de datos.

Bibliografía/Referencias

[FR] Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph Drawing by Force-Directed Placement.

[LN] LN Data: http://www.lanacion.com.ar/data

[NR] Network Repository: http://networkrepository.com/

[TWB] http://www.bancomundial.org/

[RC] Rest Countries: https://restcountries.eu/

[WEF] World Economic Forum: https://www.weforum.org

[WHO] Organización Mundial de la Salud: http://www.who.int/es/