Minería de datos: Práctica 1. Análisis y visualización básica de una red social con Gephi.

Carlos Manuel Sequí Sánchez
DNI: 20 48 69 26 K
e-mail: sequi96@corre.ugr.es

April 2019

1 Análisis básico de la red.

1.1 Conjunto de datos utilizado.

Los datos escogidos contienen la red de los partidos de futbol americano entre las universidades de la división IA en la temporada de otoño del año 2000 según M. Girvan y M. Newman. Los nodos contienen valores para indicar a que conferencias pertenecen. Dichos valores son los siguientes:
0 = Atlantic Coast
1 = Big East
2 = Big Ten
3 = Big Twelve
4 = Conference USA
5 = Independents
6 = Mid-American
7 = Mountain West
8 = Pacific Ten
9 = Southeastern
10 = Sun Belt
11 = Western Athletic

Para la correcta observación del grafo he utilizado los parámetros de ajuste Force Atlas 2 y Expansión de la ventana de Distribución con el fin de obtener el siguiente grafo:

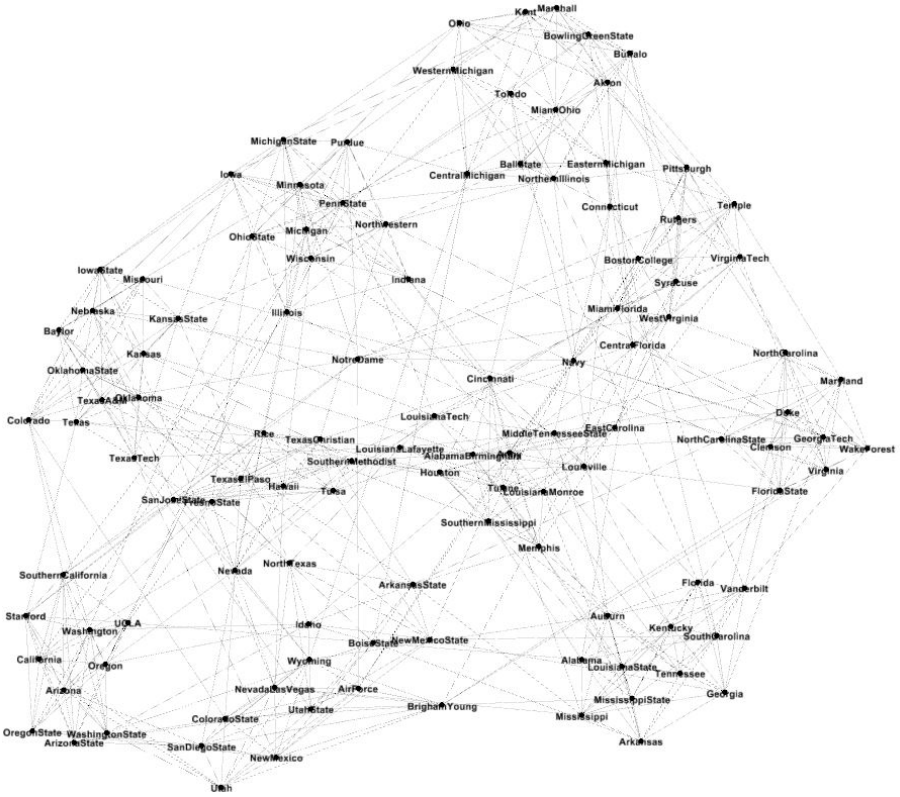


Figure 1: Grafo inicial ajustado.

1.2 Medidas globales básicas

- Número de nodos (L): 115
- Número de enlaces: 613
- Número máximo de enlaces (Lmax): 115*(114)/2 = 6555
- Grado medio (k): 10.661
- Densidad de grafo (L/Lmax): 0.094
- Coeficiente medio de clustering (C): 0.403

Como vemos, la media de partidos jugados por cada equipo (grado medio) en la temporada de otoño del año 2000 es de 10.661, así como la cantidad de partidos disputados frente a los que se podrían haber disputado si todos los equipos hubiesen jugado con todos (densidad) es de 0.094 y, finalmente, podemos ver como el coeficiente de clustering medio es bastante alto, lo que nos indica un grado significativo de clustering local (ya que cada equipo ha jugado con casi todos los equipos de su region) y, además, en la gráfica, podemos ver como este valor de coeficiente de clustering está "centrado" en el sentido de que podemos observar cómo no existen esos hubs con un coeficiente de clustering mucho más bajo en comparación con el resto como ocurriría si existiesen esos hubs.

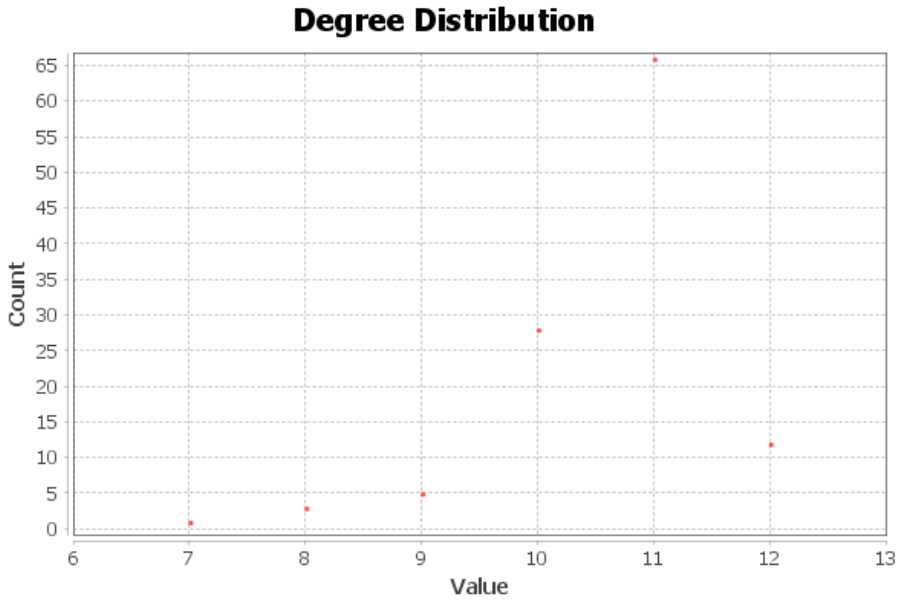


Figure 2: Degree distribution.

Como podemos observar, la gráfica obtenida no es una distribución en potencia, ya que la mayor parte de los nodos (equipos) han disputado un total de 11 partidos, otros cuantos un total de 10 partidos, y despúes hay excepciones en equipos que han efectuado 7, 8, 9 y 12 partidos. Debido a esto, sabemos que en esta red no se cumple la propiedad libre de escala, ya que no existen unos pocos equipos relacionados con muchos más equipos comparados con el resto (hubs).

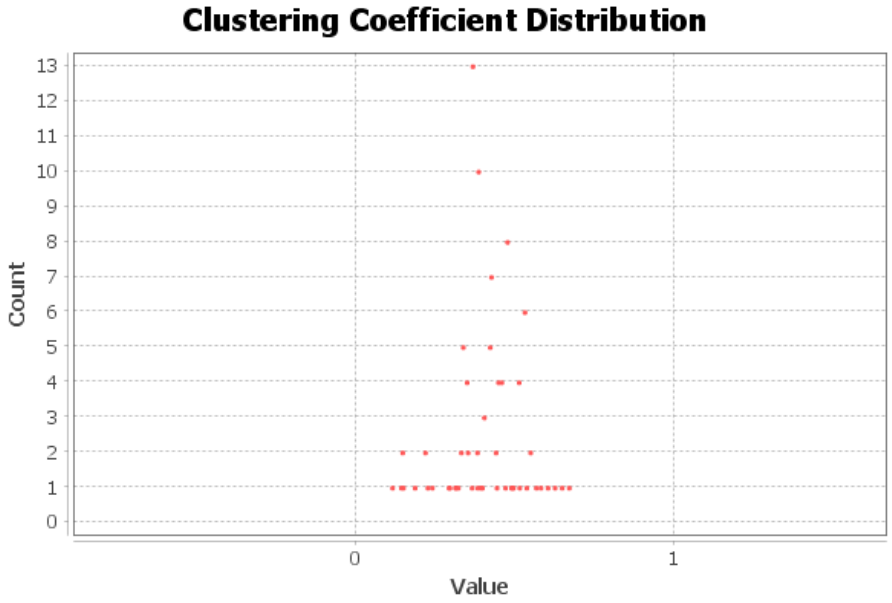


Figure 3: Clustering distribution.

1.3 Conectividad de la red.

- Componenters conexos: 1
- Componente gigante: 100% de los nodos

El grafo presenta una sola componente conexa debido a que todos los equipos han participado en la misma competición, por lo que el grafo se trata de una única componente conexa y, así mismo, la componente gigante engloba al 100% de los equipos del grafo. A esto se debe la forma de la gráfica "size distribution", donde aparece un único punto con el 100% de los nodos.

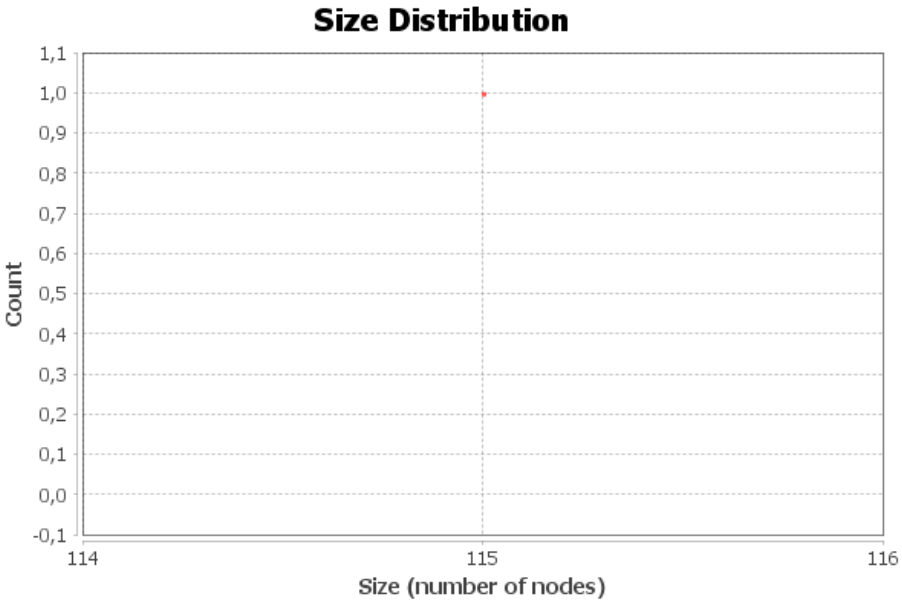


Figure 4: Size distribution.

1.4 Medidas globales.

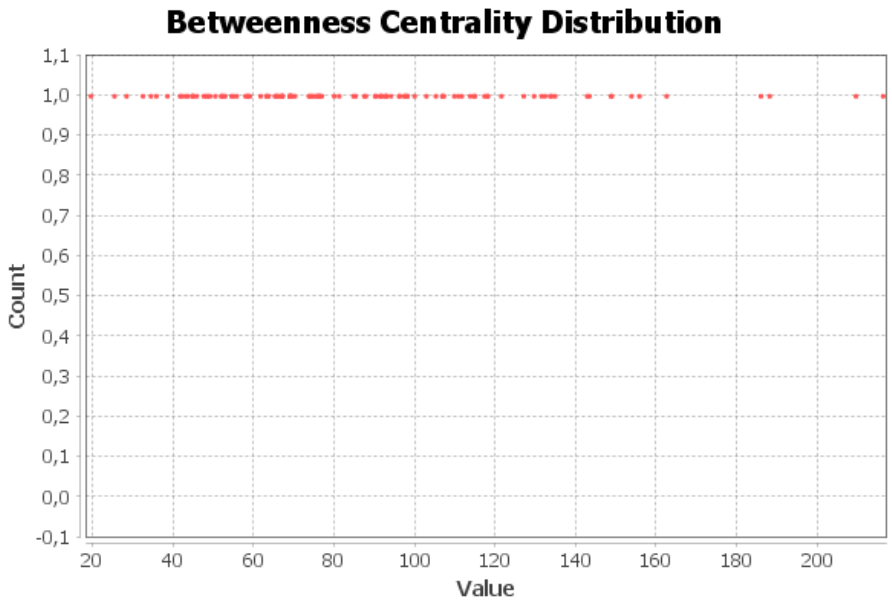


Figure 5: Betweennes centrality distribution.

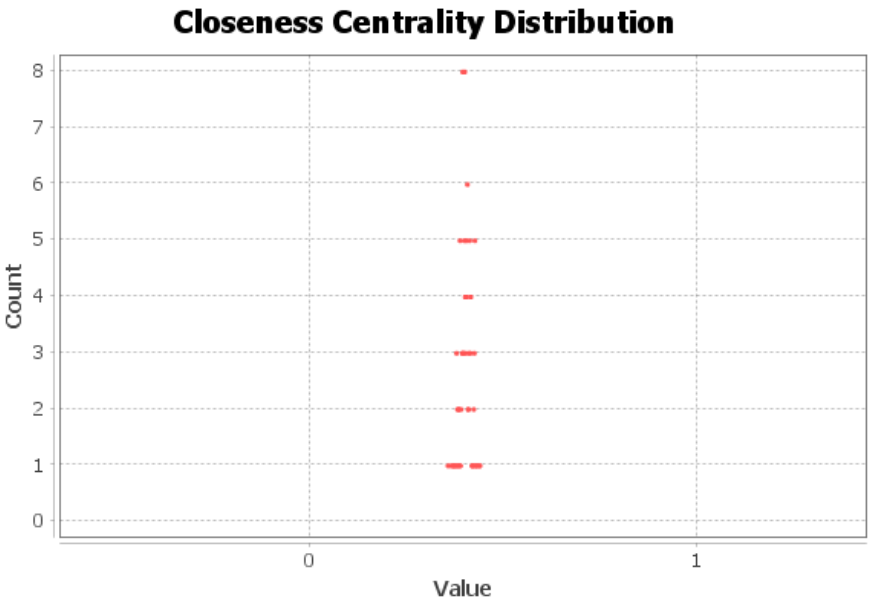
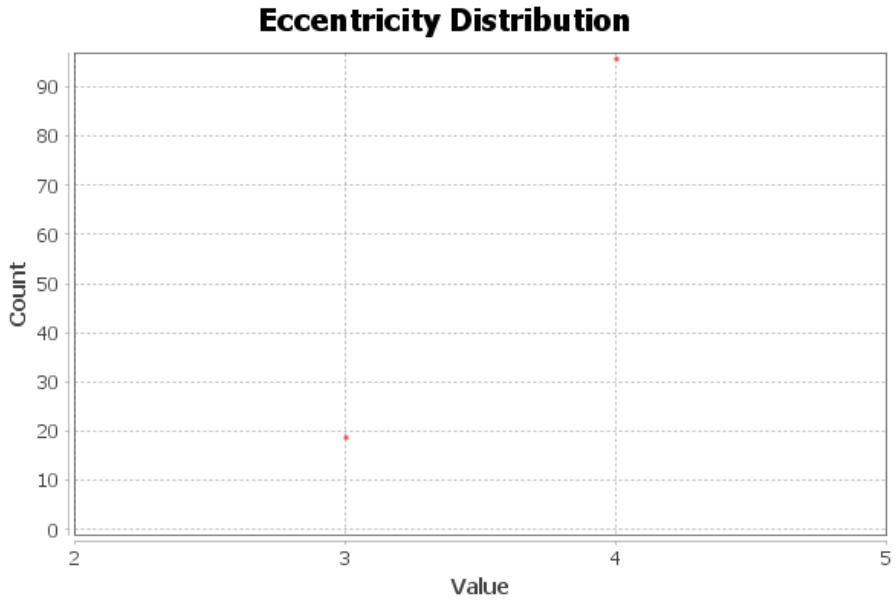


Figure 6: Closeness centrality distribution.



**

Figure 7: Eccentricity distribution.

- Diámetro (dmax, distancia más larga entre los dos nodos mas alejados): 4
- Radio: 3
- Distancia media (d): 2.508
- Intermediación(betweenness): frecuencia con la que un nodo aparece en el camino más corto entre nodos de la red. Como observamos en la gráfica, hay bastantes equipos con cierto nivel de intermediación, debido a que todos los equipos de todas las regiones al menos compiten con un equipo de fuera de su región, por lo que eso genera un grado de intermediación en cada nodo al hacer de intermediario entre los equipos de su región con al menos un equipo de otra distinta.
- Cercanía: Distancia media desde un nodo inicial a todos los demás nodos de la red. Tal como vemos en la gráfica no existen distancias grandes, de hecho hay poca variación entre estas, lo cual es indicativo de la propiedad de mundos pequeños.
- Excentricidad: distancia desde un nodo al nodo más alejado de él a la red. Según la gráfica, y tal como vemos en los valores de diámetro y radio del grafo, los valores de excentricidad existentes en todos los nodos del grafo son de 3 y de 4.

2 Estudio de la centralidad de los actores

Centralidades			
Grado	Intermediación	Cercanía	Vector propio
Tulsa (12)	NotreDame (215.98)	LouisianaTech (0.4367)	Nevada (1)
PennState (12)	BrighamYoung (209.26)	Navy (0.4351)	SouthernMethodist (0.9804)
BrighamYoung (12)	Navy (187.82)	Tulsa (0.4301)	Tulsa (0.9662)
Wisconsin (12)	LouisianaTech (185.64)	Indiana (0.4269)	Iowa (0.9626)
NevadaLasVegas (12)	CentralMichigan	PennState (0.4253)	Wisconsin (0.9592)

A partir de la tabla de centralidades de los actores en función de las medidas sabemos que:

- Los equipos de Tulsa, PennState, BrighamYoung, Wisconsin y NevadaLasVegas son los más céntricos a nivel de grado (partidos jugados contra otros equipos). Aún así hay unos cuantos mas con el mismo número de partidos disputados que estos, 12.

- Los equipos de NotreDame, BrighamYoung, Navy, LouisianaTech y CentralMichigan son los que poseen mayor centralidad en cuanto a la medida de intermadiación, debido a que son los equipos que más han salido a jugar a distintas regiones, por lo que son los equipos que mayormente conectan a equipos de distintas regiones.

- Los equipos de LouisianaTech, Navy, Tulsa, Indiana y Pennstate son los mas centrales con respecto a la cercanía, pues son los que más se concentran en la distancia geodésica de muchos actores con todos los demás, es decir, se encuentran muy próximos a las zonas de correduría.

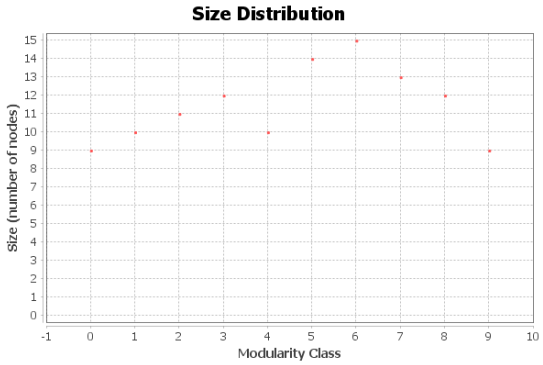
- Los equipos de Nevada, SouthernMehodist, Tulsa, Iowa y Wisconsin son los que mayor centralidad poseen (mayor ego) gracias a la medida de centralidad de vector propio calculada de forma recursiva por los vecinos de cada nodo, sin centrarse exclusivamente en la medida de grado propia de cada individuo.

Figure 8: Equipos centrales por medida de vector propio.

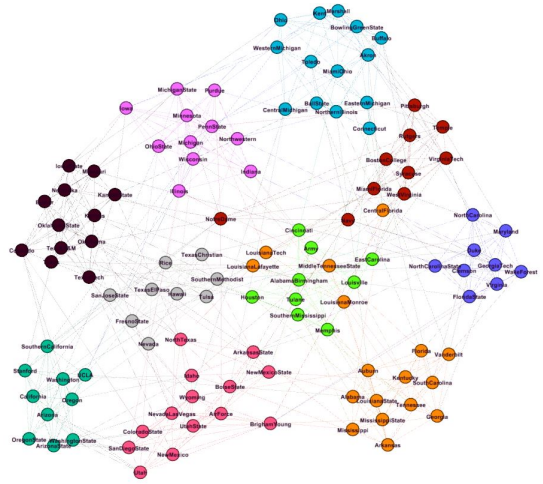
Figure 9: Equipos centrales por medida de intermediación.

3 Detección de comunidades.

3.1 Resolución = 1. Modularidad = 0.604. 10 comunidades.

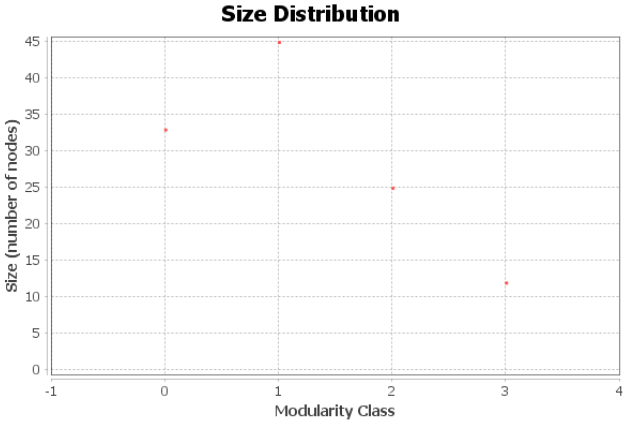


(a) Distribución de comunidades.

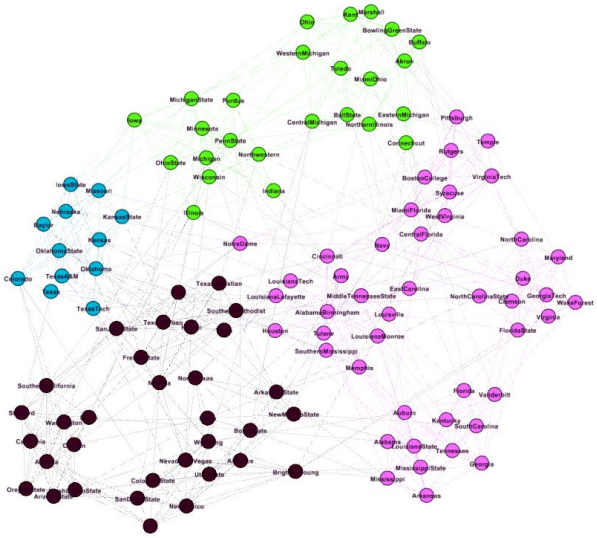


(b) Vista del grafo.

3.2 Resolución = 2. Modularidad = 0.534. 4 comunidades.

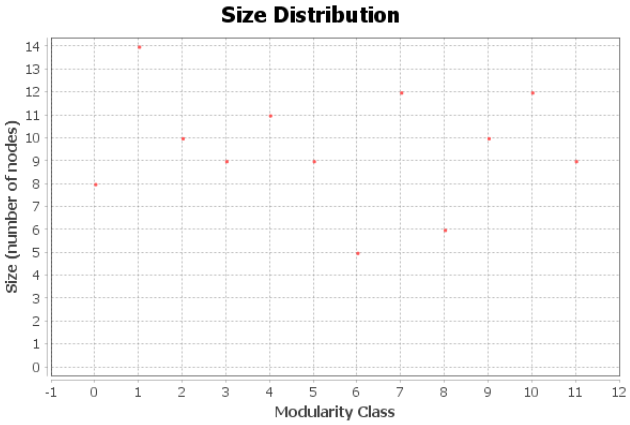


(a) Distribución de comunidades.

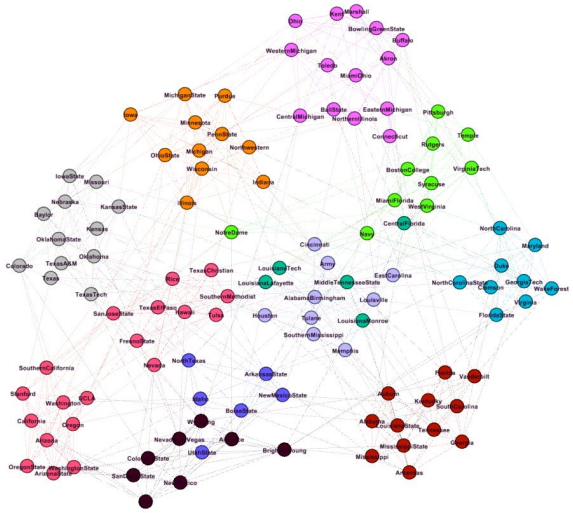


(b) Vista del grafo.

3.3 Resolución = 0.5. Modularidad = 0.601. 12 comunidades.



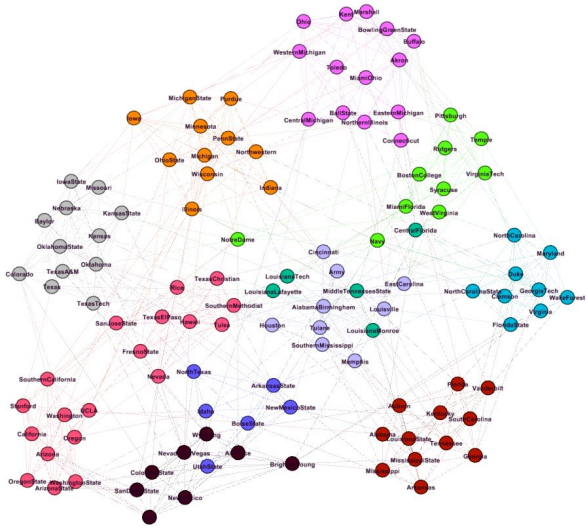
(a) Distribución de comunidades.



(b) Vista del grafo.

3.4 Resolución = 0.75. Modularidad = 0.603. 11 comunidades

(a) Distribución de comunidades.



(b) Vista del grafo.

Tal como indiqué al principio de la documentación, los nodos ya están clasificados por regiones, ya que cada nodo tiene asociado un valor indicando la región a la que pertenece, por tanto procedemos ahora a realizar una pequeña comparación de los agrupamientos creados mediante el método Lovaina sobre ésta última medida de resolución (0.75), la cual extrae exactamente el mismo número de comunidades que de regiones contiene el dataset original (11 en total):

Label	value	Modularity Class
FresnoState	11	10
Rice	11	10
SouthernMethodist	11	10
Nevada	11	10
SanJoseState	11	10
TexasElPaso	11	10
Tulsa	11	10
TexasChristian	4	10
Hawaii	11	10
KansasState	3	9
TexasTech	3	9
Baylor	3	9
Colorado	3	9
Kansas	3	9
IowaState	3	9
Nebraska	3	9
TexasA&M	3	9
Oklahoma	3	9
Texas	3	9
Missouri	3	9
OklahomaState	3	9
FloridaState	0	8
NorthCarolinaState	0	8
Virginia	0	8
GeorgiaTech	0	8
Duke	0	8
NorthCarolina	0	8
Clemson	0	8
WakeForest	0	8
Maryland	0	8

Figure 14: Valores reales asociados a las comunidades 8, 9 y 10.

Como bien podemos observar, el método de Lovaina ha generado poco error al establecer cada uno de los equipos a una comunidad concreta. La columna "value" es la región a la que pertenece cada equipo de forma original, mientras que la columna "Modularity Class" es la comunidad asociada a cada equipo por el método de Lovaina. Tan sólo se ha ejecutado sobre las comunidades 8, 9 y 10 para no extender demasiado el tamaño de la imagen de comprobación del método Lovaina.