

Teoría de Algoritmos II

Parcialito 2

Autor: Alejo Tomás Mariño
Padrón: 105645

Consigna:

1.

- a. Obtener una visualización de las comunidades presentes en dicha red (indicando el algoritmo utilizado).
- b. Considerando lo que respondiste en el parcialito 1 (ejercicio 2):
 - i. Si mencionaste que había homofilia, ¿corresponde por el mismo tipo que mencionaste anteriormente? ¿por qué?
 - ii. Si mencionaste que no había homofilia (o bien no realizaste el ejercicio), ¿qué tipo de homofilia se puede ver presente?
- c. Obtener los nodos correspondientes a una de las subredes (con al menos 20% de los nodos), y realizar una visualización de las sub-comunidades presentes.

2.

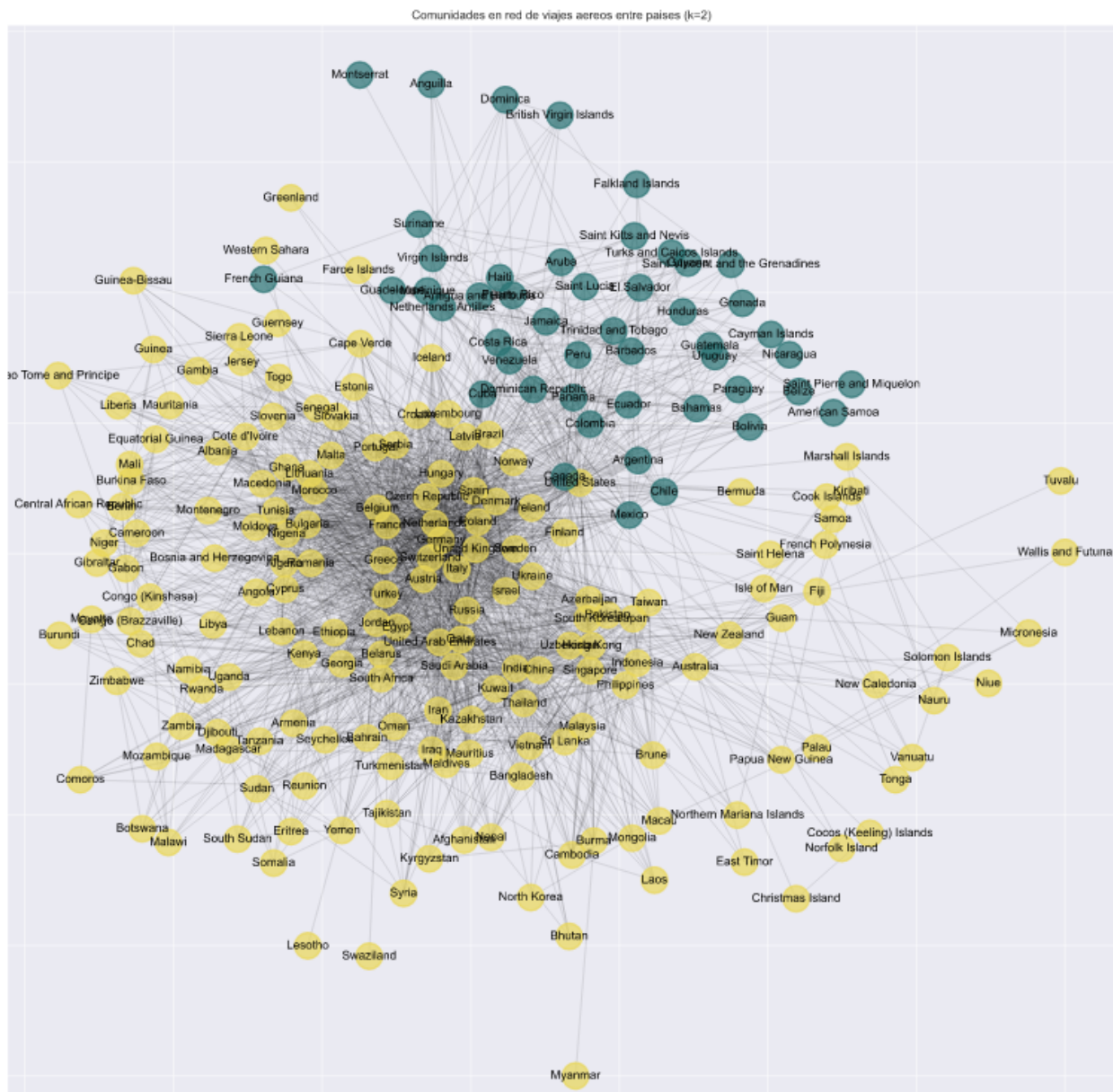
- a. Calcular los motifs de hasta 5 nodos de la subred definida en el punto 1.c.
- b. Calcular el promedio y desvío estándar de los motifs de una red de baseline. Calcular el significant profile de la red, y hacer un gráfico.
- c. Intentar dar con una explicación del resultado obtenido en el punto anterior.

3. Detectar los roles en dicha red utilizando el algoritmo RolX, explicando el resultado obtenido.

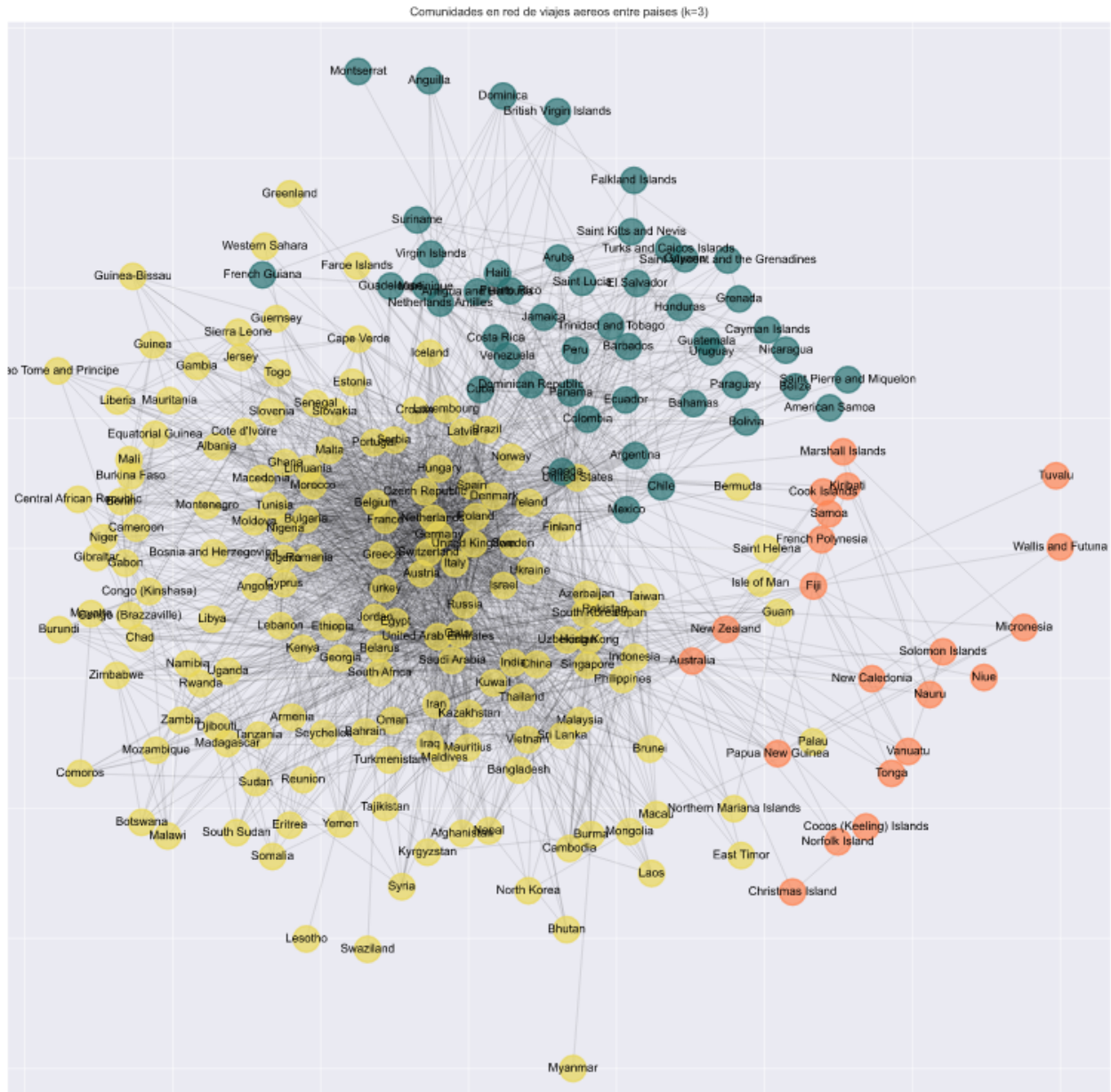
1. a.

Para obtener una visualización por comunidades de la red intente usar la biblioteca con el set de funciones proporcionada (notebook jupyter de Star Wars), pero al usar la función 'get_number_communities(grafo)' con grafo el grafo de viajes a analizar, obtuve que había una única comunidad. Tras esto, decidí usar la función de clustering espectral 'cluster_and_plot()' pero asumí distintos valores de n_clusters, realice gráficos para n_clusters igual a 2, 3 y 4 (los .html estarán dentro del .zip en la entrega para una mejor visualización). Nota: luego de hacer todo un compañero por slack recomendó el algoritmo de Louvain y este si que me dio resultado luego de intentar un rato así que agrego el resultado.

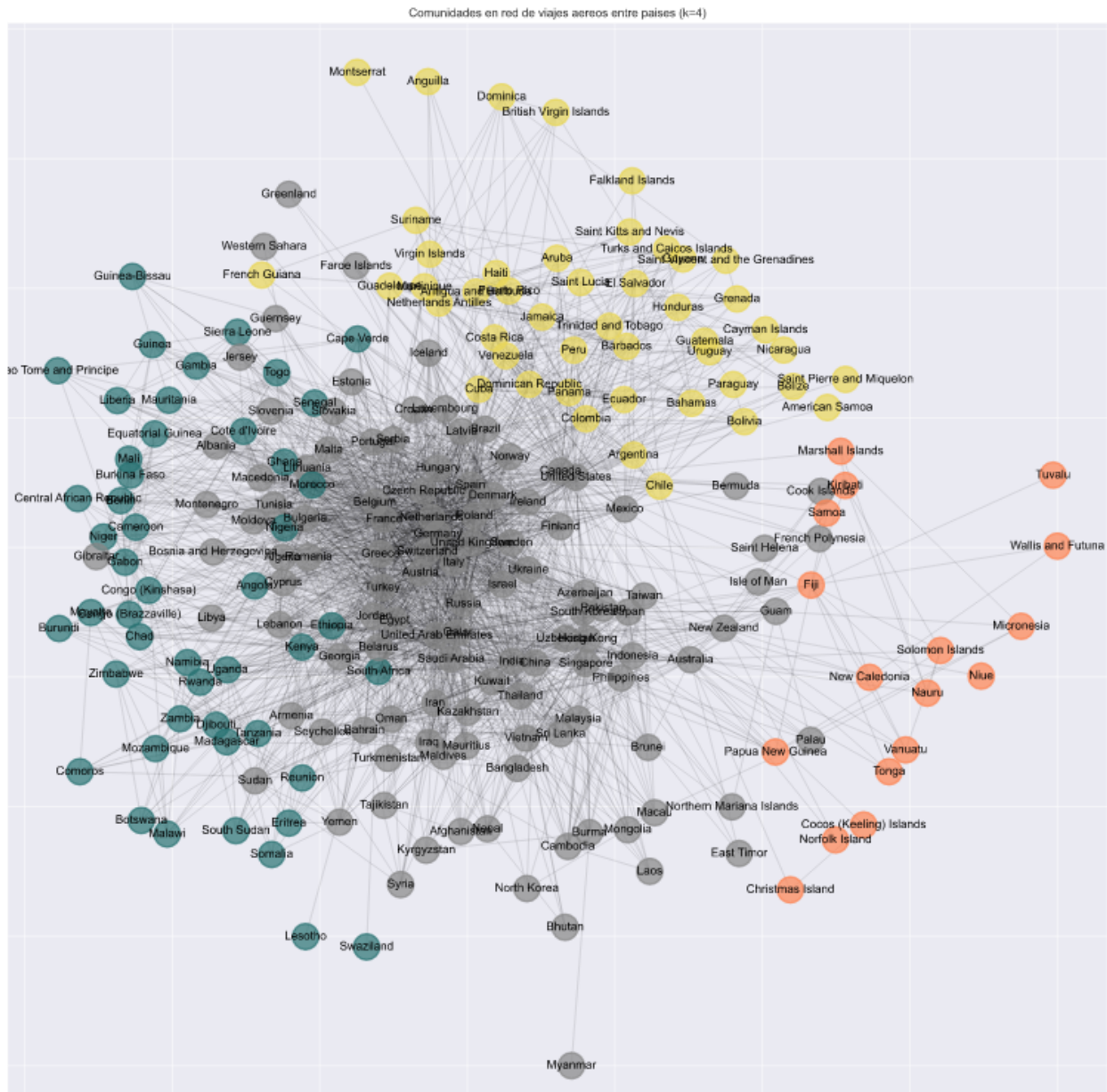
k=2:



k=3:



k=4:



Louvain:

empobrecidos, solo hay un $\sim 0.87\%$ de los países ricos del mundo (recordando del parcialito 1, por la métrica dada el total de nodos de países ricos era de alrededor $\sim 16,15$). En la otra comunidad se encuentra el resto (los 2 restantes) de los países ricos. Muestra que por más que no haya homofilia para este k de clustering espectral parecería que la métrica elegida estaba basada en alguna idea semi coherente (parece mantenerse segregados ambos tipos pobres y ricos para $k=3$ y $k=4$ también). Para $k=3$ parecería agregarse una comunidad que se desprende de la más grande en $k=2$. Una curiosidad a destacar es que parece contener a países los cuales parecen ser islas o partes de archipiélagos (podríamos considerar a australia una gran isla). Para $k=4$ parece que se agrega otra comunidad la cual también se desprende de la más grande de $k=3$ y es posible ver cómo básicamente todos sus nodos pertenecen a países africanos, quedando los países 'islandeses' en la componente que surgió en $k=3$ y países del medio oriente, europa y asia mayoritariamente en la última componente. Analizando lo obtenido por el algoritmo de Louvain obtenemos 4 comunidades (cabe destacar que corriendo el algoritmo a veces varía entre 4 y 5 comunidades), lo obtenido por este algoritmo parece predecir la ubicación continental de cada uno de los países con escalofriante exactitud, dejando en la comunidad amarilla países mayoritariamente asiáticos y oceánicos (probablemente no sea una palabra, pero me refiero a de oceanía) en la comunidad gris países mayoritariamente americanos, en la comunidad naranja países mayoritariamente europeos (aunque se cuelan algunos del medio oriente) y en la última comunidad países americanos. Este algoritmo evidencia una homofilia por continentes bastante fácil de visualizar.

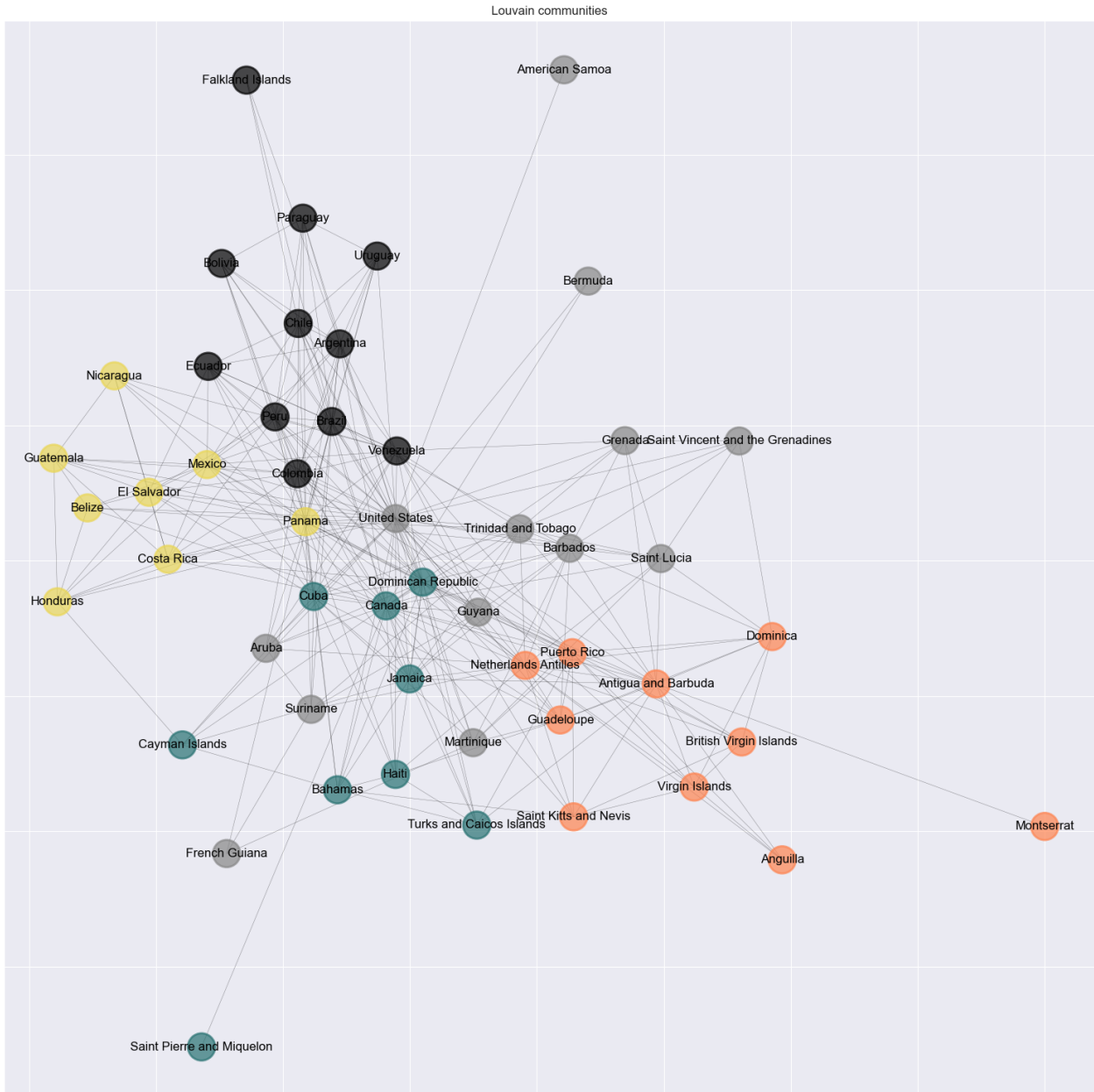
Porcentaje de países ricos en cada comunidad para cada una de las representaciones (el código de todo estará adjunto en el .zip entregado, las últimas 4 comunidades sin header en la imagen son las de Louvain):

```

===== k = 2 =====
Percentage of rich countries in 0 community: 15.283842794759824
Percentage of rich countries in 1 community: 0.8733624454148471
===== k = 3 =====
Percentage of rich countries in 1 community: 0.8733624454148471
Percentage of rich countries in 0 community: 14.41048034934498
Percentage of rich countries in 2 community: 0.8733624454148471
===== k = 4 =====
Percentage of rich countries in 1 community: 0.0
Percentage of rich countries in 0 community: 15.72052401746725
Percentage of rich countries in 2 community: 0.43668122270742354
Percentage of rich countries in 3 community: 0.0
=====
Disclaimer: it should be remarked that in the previous study, with the elected metrics rich countries were about  $\sim 16.15$  of the nodes of the network.
Percentage of rich countries in 0 community: 5.676855895196507
Percentage of rich countries in 2 community: 8.296943231441048
Percentage of rich countries in 3 community: 1.7467248908296942
Percentage of rich countries in 1 community: 0.43668122270742354

```

C.



Usando el algoritmo de Louvain obtengo esta subcomunidad habiendo agarrado la comunidad gris de la representación gráfica obtenida por Louvain del punto 1.a. Lo que veo que cabe destacar para esta subcomunidad es que el algoritmo claramente sigue siendo eficiente para obtener comunidades aún para comunidades pequeñas. Claramente se ve que la comunidad negra son países sudamericanos (Argentina, Falkland Islands, Paraguay, Bolivia, Uruguay, Chile, Ecuador, Perú, Brazil, Colombia, Venezuela), también se puede claramente distinguir países centroamericanos en la componente amarilla (Mexico y todos países cercanos a Mexico en si), no encuentre una buena distinción entre las otras 3 comunidades pero puede seguramente atribuirse a mi pobre conocimiento geográfico ya que estoy seguro que por ejemplo los países de la

componente naranja deben tener algo en comun ademas de ser todos islas/archipiélagos.

2. a.

(Para el punto 2 use colab.research.google.com para realizar los cálculos, ya que el paralelismo daba ciertos problemas en Windows y en vez de pasarme a Linux para ejecutar el código decidí aprender esta herramienta nueva, aun así en resolucion.py está todo el código que cargue en colab.)

```
✓ 18m ▶ MAX_NODOS_MOTIFS = 5
motifs_graph = calcular_motifs(indexed_american_graph, MAX_NODOS_MOTIFS)
print(motifs_graph)

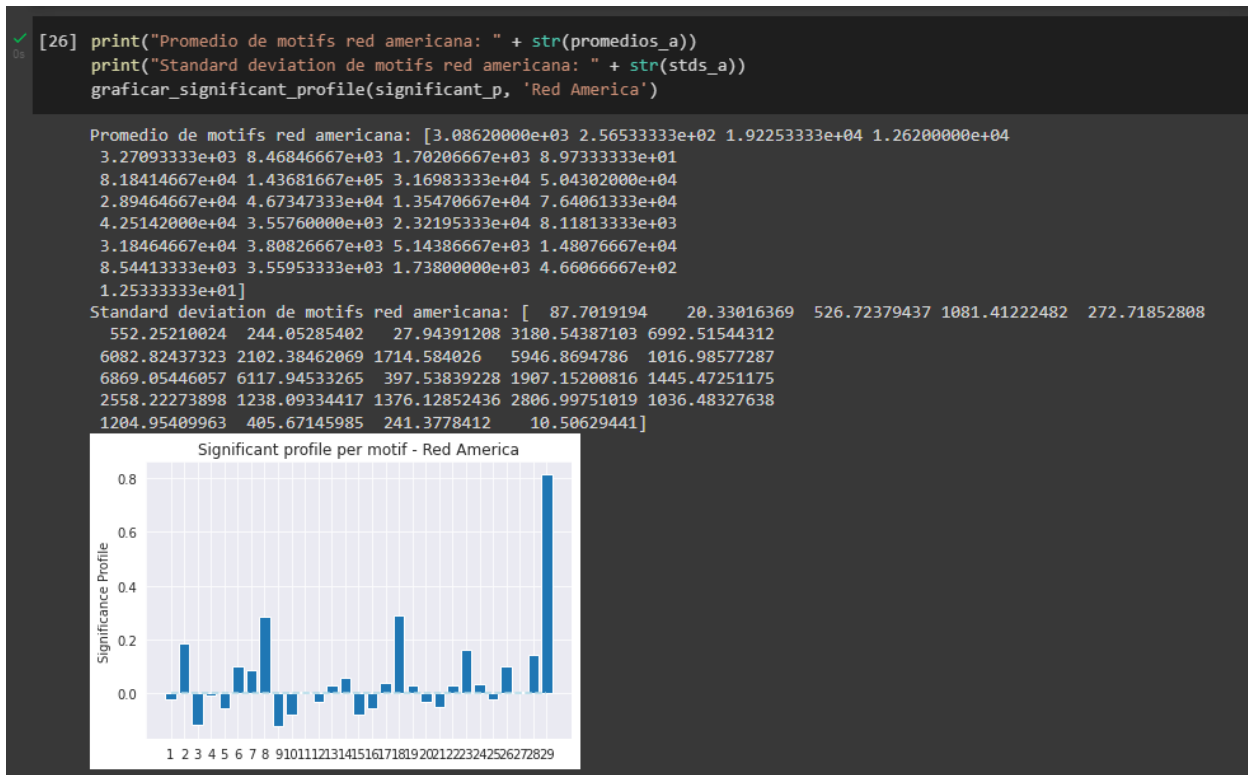
[ 2801  820 10195 11235  1120 16894  4789  1273 25389 61485 36377 40679
 37030 98476  1702 19151 79815 20672 31279  1527 12752  9139 38170 30207
 4996 21666  2203  5560  1272]
```

b.

(Tuve que correr con 15 iteraciones y no más por el hecho de que cuando lo hice con 20 y vi que la primera me tardó >29 minutos me di cuenta de que si hacía 20 no iba a terminar antes de irme a dormir)

```
✓ 3m [23] distribucion = distribucion_grados(indexed_american_graph)
promedios_a, stds_a = motif_grafo_eleatorios(lambda: configuration_model(distribucion), MAX_NODOS_MOTIFS, iters=15)
significant_p = significance_profile(motifs_graph, promedios_a, stds_a)
print("Significant Profile red americana: ", significant_p)

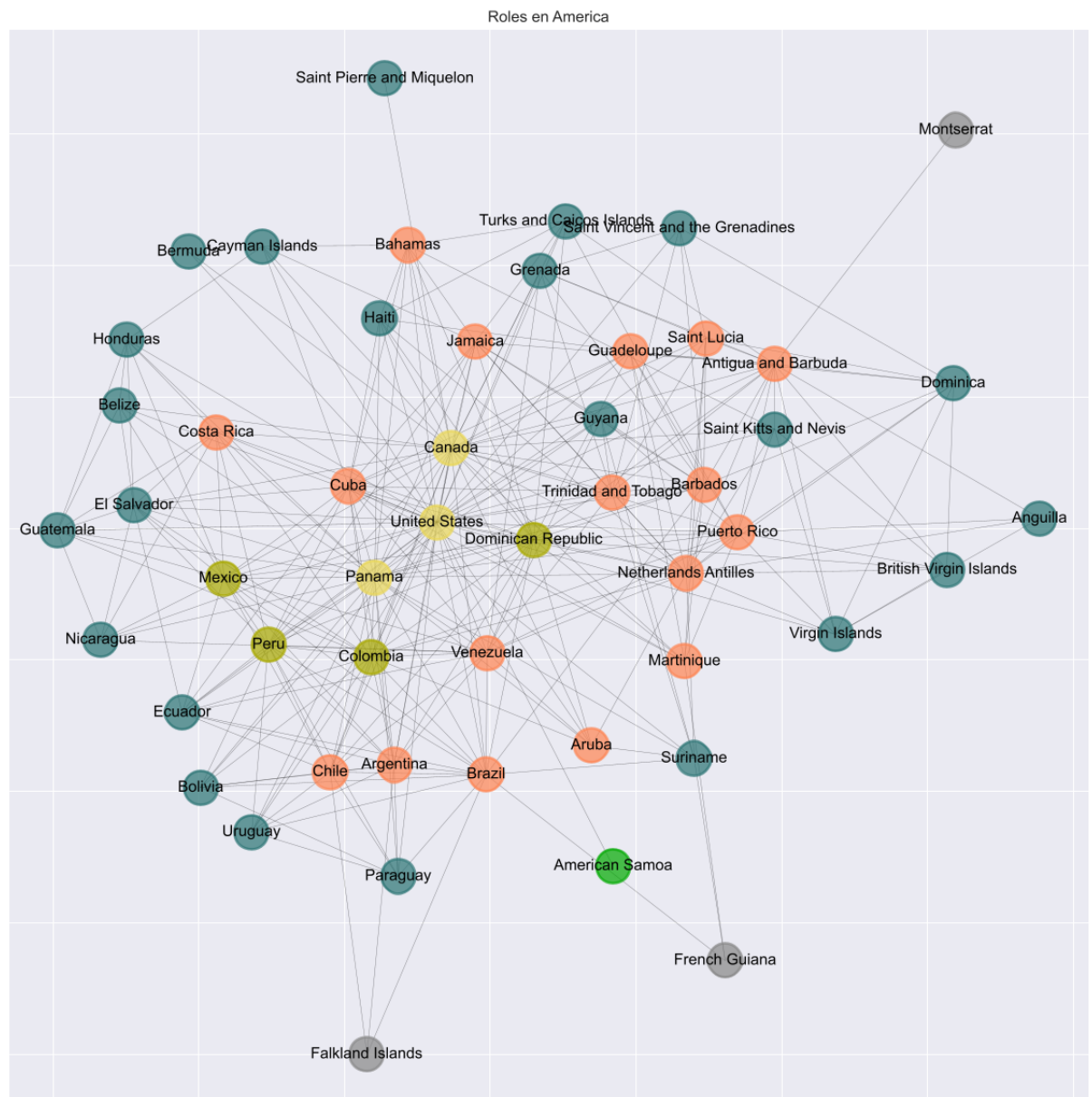
Iteracion 1
Iteracion 2; anterior: 1317.54 segs
Iteracion 3; anterior: 1444.38 segs
Iteracion 4; anterior: 1351.84 segs
Iteracion 5; anterior: 1262.87 segs
Iteracion 6; anterior: 1533.63 segs
Iteracion 7; anterior: 1566.90 segs
Iteracion 8; anterior: 1364.41 segs
Iteracion 9; anterior: 1311.28 segs
Iteracion 10; anterior: 1526.54 segs
Iteracion 11; anterior: 1362.37 segs
Iteracion 12; anterior: 1425.50 segs
Iteracion 13; anterior: 1494.82 segs
Iteracion 14; anterior: 1320.27 segs
Iteracion 15; anterior: 1380.13 segs
Significant Profile red americana: [-0.02204377  0.18786952 -0.11621718 -0.00868177 -0.05346395  0.1034212
 0.08574166  0.28703193 -0.12031831 -0.07968398  0.00521395 -0.03144098
 0.03195895  0.0589792  -0.07895371 -0.05650251  0.04132975  0.29183128
 0.02864647 -0.03091009 -0.05059637  0.0291866  0.16268546  0.03718862
-0.02320532  0.10186223  0.00777012  0.14305544  0.81254254]
```



c. Lo principal que puedo concluir del punto 2.b es que, al ver ese gran pico en 29, ese motif representa al motif de $n = 5$ en el cual todos están unidos con todos, algo bastante característico de varias redes sociales. En nuestra red dada (subred de la red original que tiene los nodos de países americanos) podemos extrapolar la conclusión de que todos los países americanos parecen estar altamente conectados entre sí. De hecho, en el parcialito 1 había mencionado algo por el estilo, debido al alto nivel de globalización y de facilidad de viaje en muchos países del mundo actual, básicamente cualquier país civilizado tiene un aeropuerto capaz de recibir vuelos internacionales por lo que tiene sentido que justamente el motif que es un k -clique de $k=5$ este tan sobrerrepresentado.

3.

(el archivo con la imagen está en la carpeta del .zip)



Analizando un poco el gráfico obtenido por roles podemos distinguir como países “primarios” (aquellos que casi todos los países de america tienen conexión hacia ellos) Estados Unidos, Canada y Panama. Si uno lo fuera a lo mejor pensar como un viajante que tiene que ir de un país A que no tenga un rol tan principal en la red a un país B,

solo pudiendo viajar entre países americanos, es muy probable que viajar a alguno de estos 3 países con este rol primero (viaje de escala) y luego viajar a B sea básicamente siempre posible.

Los países con el rol azul oscuro son, en su mayoría, países cuyo terreno son mayoritariamente islas o países centroamericanos los cuales parecen tener comunicaciones a una cantidad bastante restringida de otros países.

No pude deducir mucho más allá los roles devueltos por el algoritmo, no se si es por no tener el ojo apropiadamente entrenado para leer la data o si simplemente no hay mucho más que deducir por los roles.