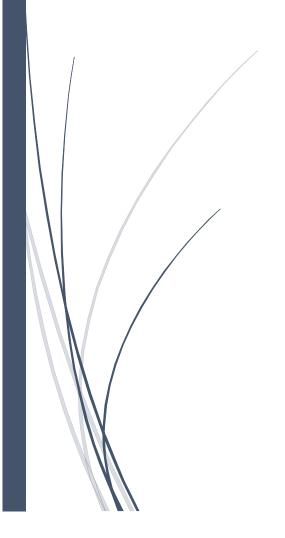
Redes y Sistemas Complejos

Análisis y detección de comunidades en la red de idiomas mundial

Sergio Carrasco Márquez

CORREO: sergiocmarq@gmail.com



Contenido

1 Cuestión planteada	3
2 Estructura de la red	3
2.1 Red de idiomas	3
2.2 Red de migrantes	3
3 Métricas de la red	4
3.1 Red de idiomas	4
3.2 Red de migrantes	5
4 Propiedades de la red	6
5 Valores propios de una red social	7
6 Comunidades de la red	9
6 Visualización de la red	13
6.1 Red de idiomas	13
6.2 Red de migrantes	15
7 Discusión de los resultados	15
8 Comparación con otra red de la literatura	20
Tabla 1: Métricas de la red de idiomas	5 8 8 9
Ilustración 1: Esquema de la red de idiomas	5
Ilustración 2: Distribución de grados de la red de idiomas	6
Ilustración 3: Distribución de centralidad de cercanía de la red de idiomas	
Ilustración 4:Coeficiente de clustering de la red de idiomas	
Ilustración 6: Distribución de la clase de modularidad en la componente gigante de la red de	9
idiomas	10
Ilustración 7: Comparación entre la componente gigante y toda la red en la detección de	40
comunidades	
gigante de la red de idiomas	
Ilustración 9: Distribución de la clase de modularidad con resolución a 2.8 en la componente	
gigante de la red de idiomas	11

Ilustración 12: Visualización de la componente gigante de la red de idiomas	13
Ilustración 13: Comunidad hispana y portuguesa	14
Ilustración 14: Visualización de la red de migración	15
Ilustración 15:Comunidad de habla inglesa	16
Ilustración 16: Comunidad de habla francesa	17
Ilustración 17: Comunidad de habla hispana y portuguesa	18
Ilustración 18: Grupo de países de cultura árabe	19
Ilustración 19: Red de migración, destacando los países periféricos	20

1 Cuestión planteada

A lo largo de la historia la cultura propia de cada país se ve afectada por la acción propia del ser humano. Desde la época colonial, en la que diversas potencias europeas colonizaban naciones a lo largo de todo el mundo descubierto, hasta la época de la globalización, en la que los avances técnicos han hecho que el flujo de personas a lo largo del planeta sea mucho más fácil. A lo largo de este documento se va a analizar este fenómeno. Para ello, el análisis se va a sustentar en dos redes. La primera red, que es la que será analizada en profundidad, es una red que refleja los grupos culturales a lo largo del planeta basándose en un elemento esencial para una cultura como es el idioma hablado, hermanando así a los países que compartan un idioma. La segunda red es una red que refleja el flujo migratorio entre los años 1990 y 1995. En base a esa red se analizará el tránsito de las culturas actuales sobre cada país. En cuanto a la red que relaciona a los países por su idioma se han usado un conjunto de 195 países y todos los idiomas que se hablan dicho país. **Estos** datos se han recaudado https://www.worldclasslearning.com/general-knowledge/list-countries-capital-currencieslanguages.html. En cuanto a la red migratoria se ha usado un conjunto de 196 países y los datos se han obtenido de un artículo de la revista Science. Cabe destacar que el conjunto de datos no es una medida real, sino una estimación realizada. Los datos están disponibles en http://science.sciencemag.org/highwire/filestream/595376/field highwire adjunct files/1/Ab el-Database-s2.xlsx

2 Estructura de la red

Como se ha mencionado en el apartado anterior, se van a usar dos redes distintas para el análisis.

2.1 Red de idiomas

En cuanto a la red de idiomas cada enlace representa un país, de un total de 195. Dos países tienen un enlace por cada idioma que compartan entre ellos. La red no es dirigida y todos los enlaces tienen un peso unitario.

2.2 Red de migrantes

La red de migrantes tiene un nodo por cada país estudiado. Los países se relacionan entre si en función del flujo de migrantes existente, de forma que el peso de cada enlace dependerá del número de personas que migren hacia el país destino. En este caso la red si debe ser dirigida, ya que se tiene en cuenta el país de origen y el país destino, de forma que la flecha que representa los enlaces va desde el origen hacia el destino.

3 Métricas de la red

3.1 Red de idiomas

En cuanto a la red que relaciona a los países en función del idioma se obtienen las siguientes métricas.

Nodos	195.0000
Enlaces	2784.0000
Densidad	0.1470
Grado medio	28.5540
Distancia media	1.8344
Distancia media de una red aleatoria	1.8715
Coeficiente de clusterring medio	0.9270
Coeficiente de clusterring medio de una red aleatoria	0.1440
Componentes conexas	44.0000
Enlaces componente gigante	2746.0000
Nodos componente gigante	136.0000

Tabla 1: Métricas de la red de idiomas

La red muestra una conectividad muy alta entre nodos dentro de una misma componente. La distancia media no llega a ser 2, pero se acerca bastante. Esta característica se suele presentar en grafos con forma de estrella, que para ir de un nodo a otro solo necesitan pasar por el nodo central y desde este ir al nodo destino, llegando a cualquier nodo desde cualquier nodo en 2 pasos. Si existiese algún enlace más entre nodos periféricos la distancia media bajaría de 2. En este caso podemos ver que las componentes conexas no tienen forma de estrella, sino que se asemejan más a grafos completos. Esta conclusión se extrae en base al coeficiente medio de clustering, cercano a 1 y muy lejano al esperado en una red aleatoria de las mismas características y al esperado en una red con forma de estrella. Dicho coeficiente indica que la mayoría de los nodos de una componente están conectados entre si. Pero si eso fuera totalmente cierto la distancia media se acercaría más a 1 que a 2. Este fenómeno puede explicarse de la siguiente forma. Existen grupos de nodos que están totalmente conectados entre si, y entre cada grupo de nodos dispuesto de esa forma existe un número bajo de nodos que une a dichos grupos. Más adelante se analizará qué significan dichas estructuras y por qué se producen.

Otro dato a tener en cuenta es el elevado número de nodos que pertenecen a la componente gigante y la cantidad de componentes conexas que existen. Al tener una componente gigante de 136 nodos, quedan fuera de dicha componente 59 nodos repartidos entre 43 componentes conexas. Esta información refleja que existe una componente gigante que abarca la gran mayoría de los nodos y luego algunos nodos sueltos que no están unidos con casi ningún otro.

De todas las conclusiones anteriores se puede concluir que la red original tiene una estructura similar a la mostrada en la siguiente imagen.

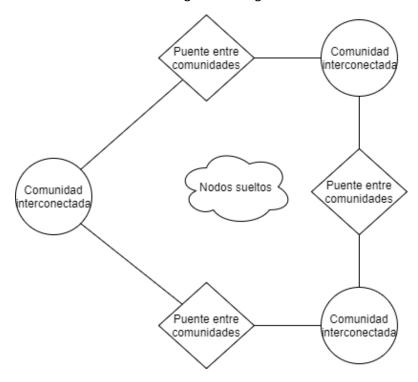


Ilustración 1: Esquema de la red de idiomas

3.2 Red de migrantes

La siguiente tabla muestra los resultados obtenidos en la red de migrantes. No se va a profundizar demasiado en analizar esta red en concreto, ya que como se ha mencionado al principio, esta red es un complemento a la red anterior para extraer algunas conclusiones extra en el análisis cultural.

Nodos	196.0000
Enlaces	11655.0000
Densidad	0.3050
Grado medio	59464.0000
Distancia media	1.7611
Distancia media de una red aleatoria	1.6894
Coeficiente de clusterring medio	0.4790
Coeficiente de clusterring medio de una red aleatoria	0.3100
Componentes conexas débiles	1.0000
Componentes conexas fuertes	19.0000
Enlaces componente gigante	11655.0000
Nodos componente gigante	196.0000

Tabla 2: Resultados de la red de migrantes

La densidad en este caso es bastante mayor que la densidad estudiada en la red anterior y se sitúa en un 0.305. La distancia media también es menor y se asemeja a la de la red aleatoria generada. Si nos fijamos en el clustering medio el valor es similar al obtenido en la aleatoria, por lo que no existe una gran interconexión entre vecinos como en la red anterior. La componente gigante muestra que abarca todos los nodos de la red, por lo que no existen nodos que estén

aislados totalmente de la red. Sí que se encuentran 19 componentes conexas fuertes, por lo que si suponemos que en la red solo se puede navegar en el sentido de las relaciones, entonces encontramos que no se puede alcanzar a todos los nodos desde cualquier nodo.

En esta red en particular no encontramos ningún patrón destacable, como el descrito en la red anterior, sino que se asemeja a una red aleatoria con las mismas características de nodos y densidad.

4 Propiedades de la red

A continuación, se muestran algunas de las gráficas obtenidas en la red que relaciona países en función del idioma.

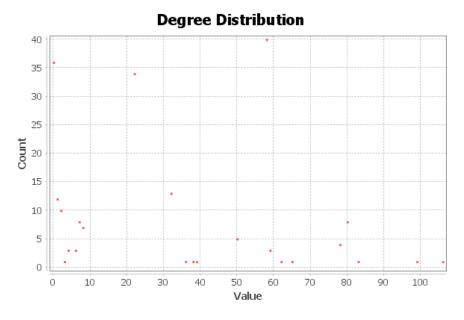


Ilustración 2: Distribución de grados de la red de idiomas

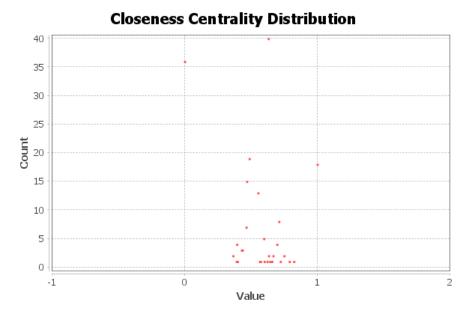


Ilustración 3: Distribución de centralidad de cercanía de la red de idiomas

Clustering Coefficient Distribution

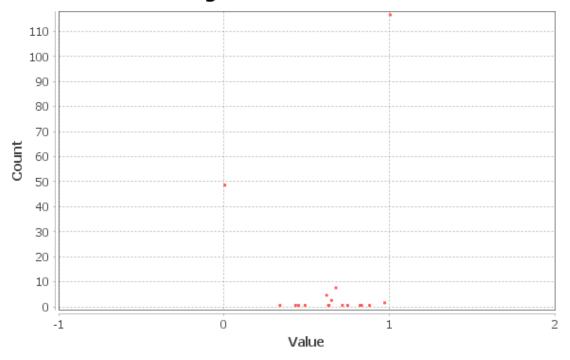


Ilustración 4:Coeficiente de clustering de la red de idiomas

En base a las gráficas obtenidas se puede concluir que la red no es libre de escala, ya que no sigue la distribución de grados esperada en una red de dichas características. En esta red la distribución de grados no sigue ninguna tendencia clara, existen muchos nodos que se encuentran desconectados, lo que hace que la distribución comience con valores altos. Al tener esta red grupos que están altamente interconectados eso hace que la distribución de grados aumente mucho al medir la distribución de grados de los nodos pertenecientes a dichos grupos.

En cuanto a las gráficas de distribución de coeficiente de clustering y centralidad de cercanía, muestran que cumple las propiedades de una red de mundo pequeño, ya que el coeficiente de clustering muestra una gran interconexión entre nodos. Aunque existen bastantes nodos desconectados en la red, existen muchos nodos que consiguen una interconexión muy alta entre sus vecinos. La distribución de centralidad de cercanía, muestra que la centralidad de los nodos es bastante alta, es decir, que la distancia entre unos nodos y otros de la red es bastante pequeña. Al igual que antes los nodos aislados muestran una centralidad de cercanía similar a 0, pero el número de nodos que tienen una centralidad alta es muy elevado.

5 Valores propios de una red social

De acuerdo con la definición de red social, las relaciones lingüísticas entre países si son una red social, ya que surgen da la interacción humana.

Para encontrar los nodos más relevantes en la red, una de las medidas a tener en cuenta es el grado. A continuación, la tabla muestra la centralidad de grado de los 5 nodos con un mayor valor en dicha medida.

País	Grado
Guinea	106
Dominica	99
Monaco	83
Canada	80
Cameroon	80
Seychelles	80

Tabla 3:Centralidad de grado de la red de idiomas

En este caso el nodo con el grado más alto es el que representa a Guinea. A priori se puede pensar que en la estructura de la red que se ha descrito anteriormente, formada por grupos de nodos interconectados entre si, el nodo con mayor grado debe estar en la comunidad con más nodos, y debe estar empatado con todos los demás nodos de dicha comunidad. Lo cierto es que los nodos que van a tener mayor grado son los que hacen de puente entre comunidades, ya que están unidos a todos los elementos de cada una de las comunidades. Guinea es el país que representa una mayor centralidad de grado, seguido de Dominica y Mónaco. Si ahora se analiza la intermediación en la siguiente tabla, Guinea sigue en primer puesto y Dominica en el segundo.

País	Grado	Intermediación
Guinea	106	1403.730879
Dominica	99	760.1154946
Singapore	62	528
Equitorial	59	525.5132784
Guinea	59	525.5132784
United		
States of	78	431.2443277
America		

Tabla 4: Centralidad de intermediación en la red de idiomas

Con estos datos se puede intuir que los nodos con mayor grado son aquellos que sirven de conexión entre las dos mayores comunidades. Esto no sería así si existiera una comunidad de un tamaño mucho mayor al resto de comunidades totalmente aislada, pero en este caso ninguna comunidad de tamaño considerable está aislada del resto.

La cercanía es otro elemento a tener en cuenta a la hora de medir la importancia de los nodos en la red. En este caso es necesario escoger solo la componente gigante de la red, ya que los nodos sueltos van a tener un índice de 1.

País	Grado	Cercanía	Intermediación
Guinea	106	0.82317073	1403.730879
Dominica	99	0.78947368	760.1154946
Monaco	83	0.72192513	385.0101832
Canada	80	0.71052632	105.7464469
Niger	80	0.71052632	105.7464469

Tabla 5: Centralidad de cercanía de la red de idiomas

De nuevo Guinea y Dominica siguen apareciendo en los primeros puestos. En este caso, al unir comunidades tan grandes, la distancia entre cualquiera de los nodos de ambas comunidades es muy baja, lo que hace que la media de distancia sea menor.

La centralidad de vector propio se refleja en la siguiente tabla.

País	Grado	Vector propio	Cercanía	Intermediación
Guinea	106	1	0.82317073	1403.730879
Dominica	99	0.995004363	0.78947368	760.1154946
Monaco	83	0.959823853	0.72192513	385.0101832
Canada	80	0.958120871	0.71052632	105.7464469
Niger	80	0.958120871	0.71052632	105.7464469

Tabla 6: Centralidad de vector propio de la red de idiomas

En cuanto a la centralidad de vector propio, era de esperar que Guinea y Dominica vuelvan a ser los más relevantes, ya que al estar unidos a grandes comunidades, la centralidad de sus vecinos es bastante alta debido a la interconexión existente dentro de una comunidad.

A la luz de todas estas medidas, se puede afirmar que en una red de estas características los elementos con más influencia son los que interconectan a las comunidades de nodos. En este caso esto se refleja en la multiculturalidad y el tamaño de las comunidades que comparten idioma con un nodo.

6 Comunidades de la red

Para la red de idiomas se ha obtenido una modularidad distribuida de la siguiente forma.

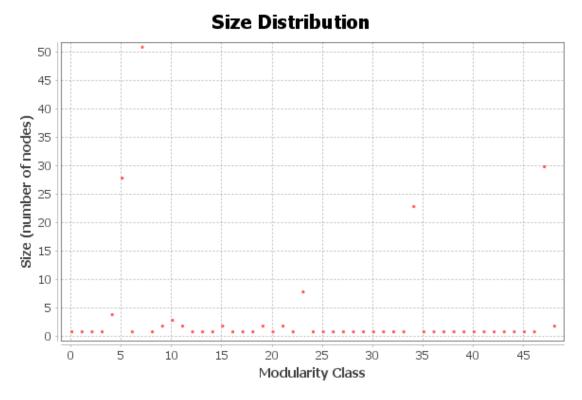


Ilustración 5: Distribución de la clase de modularidad de la red de idiomas

Según esta distribución se pueden apreciar 4 comunidades con un gran número de nodos, mientras que la mayoría de estas comunidades son nodos aislados de la red. Más adelante se analizará que representa cada una de las comunidades detectadas en la red de idiomas.

Un enfoque interesante sería utilizar solamente la componente gigante de la red para eliminar del análisis todos los nodos aislados para ver si en la componente gigante siguen quedando las mismas comunidades.

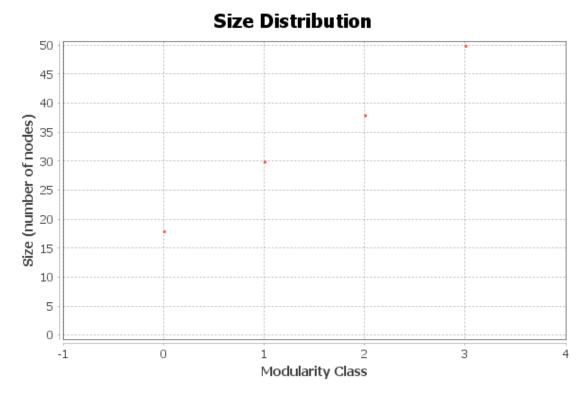


Ilustración 6: Distribución de la clase de modularidad en la componente gigante de la red de idiomas

Sigue habiendo 4 comunidades, pero el tamaño de las mismas ha variado ligeramente. A pesar de las variaciones de los tamaños de las comunidades, las comunidades detectadas dentro de la componente gigante parecen ser las mismas. Para comprobarlo, a continuación se muestran las comunidades de la componente gigante contando con todos los nodos, o solo con los nodos que forman parte de la misma.

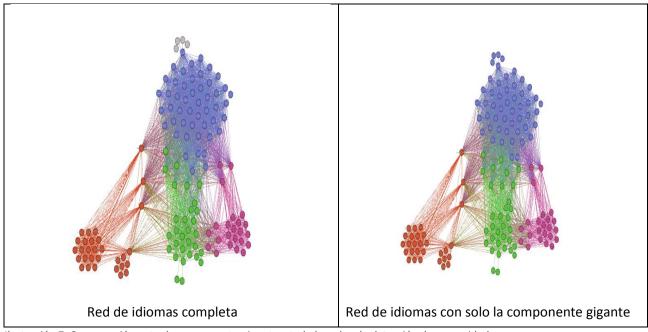


Ilustración 7: Comparación entre la componente gigante y toda la red en la detección de comunidades

En la imagen superior se ve con más claridad que nodos son los que cambian de comunidad, y como a grandes rasgos las comunidades detectadas son las mismas que en el caso anterior. El parámetro de la resolución en las comunidades permite forzar que se encuentren más comunidades o menos. A continuación, se muestran los resultados para distintos valores de resolución, menor que 1 para forzar a encontrar más comunidades y superior a 1 para que encuentre menos.

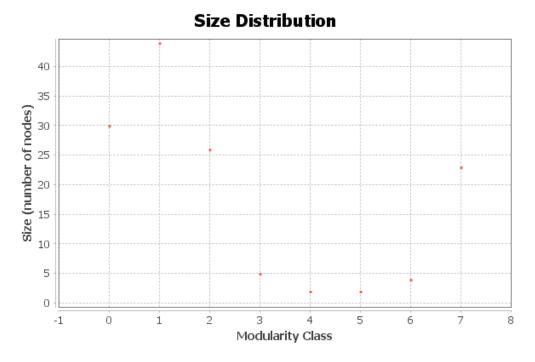


Ilustración 8: Distribución de la clase de modularidad con resolución a 0.7 en la componente gigante de la red de idiomas

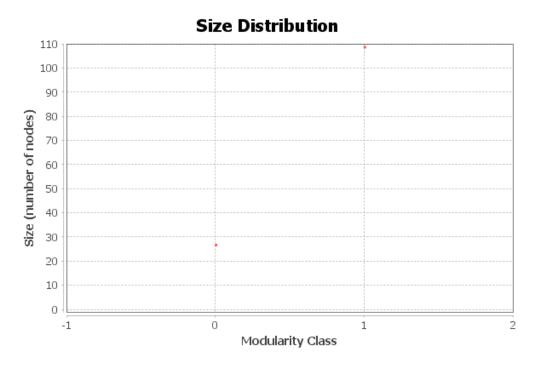


Ilustración 9: Distribución de la clase de modularidad con resolución a 2.8 en la componente gigante de la red de idiomas

Con dichas comunidades se obtienen las siguientes visualizaciones.

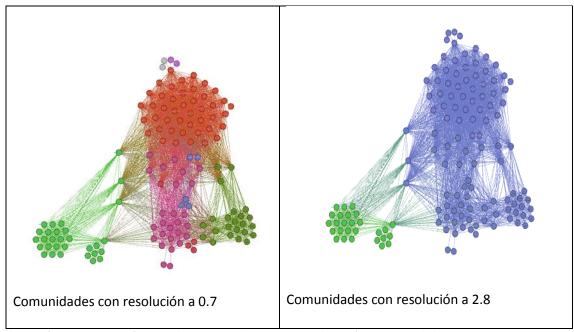


Ilustración 10: Comparación entre distintas resoluciones para la detección de comunidades

Al forzar la detección de muchas comunidades, algunos nodos forman comunidades muy pequeñas para estar en una red tan interconectada. Si por el contrario se generaliza demasiado la detección de comunidades, se observa como la comunidad que de la parte superior abarca a las comunidades detectadas en la parte inferior derecha. Dichas comunidades si habían sido distinguidas con una resolución igual a 1.

Una vez que se obtienen todas estas comparaciones, cabe preguntarse qué índice de resolución detecta mejor las comunidades. Anteriormente se ha analizado que se esperaba encontrar comunidades fuertemente enlazadas y nodos que sirven de puente entre dichas comunidades. Con una resolución de 2.8 se está metiendo en la misma comunidad a países que no comparten en absoluto ningún idioma. Lo ideal sería encontrar comunidades de forma que el número de nodos que están comunicados por un camino mínimo de longitud 2 sea el menor posible y que exista una comunidad por cada grupo de tamaño máximo de nodos que estén interconectados totalmente. En cuanto a los resultados obtenidos con la resolución de 0.7, el número de comunidades bastante alto, pero no aglomera grupos interconectados en la misma comunidad, ni tampoco produce una separación de dichos grupos. Cuando la resolución vale 1, sucede igual que cuando vale 0.7, salvo que aglomera en una comunidad a nodos que no están totalmente interconectados. A la vista de lo anterior, valores de resolución cercanos a 1, pero algo inferiores, son los que producen una detección de comunidades similares a la esperada.

6 Visualización de la red

6.1 Red de idiomas



Ilustración 11:Visualización de la red de idiomas

En la visualización anterior se observan los nodos aislados y una componente gigante que abarca los grupos interconectados que ya se han mencionado con anterioridad. La visualización de la componente gigante es la siguiente.

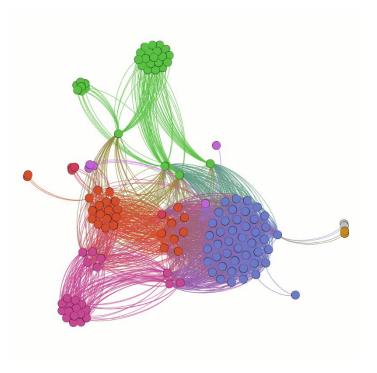


Ilustración 12:Visualización de la componente gigante de la red de idiomas

En esta visualización si se ven con claridad cuales son los grupos interconectados, que en este caso representan comunidades de idiomas, y que nodos son puentes entre comunidades. En este caso los nodos puente son los que comparten idioma con dos comunidades. Un detalle a tener en cuenta es que los nodos que pertenecen a una comunidad comparten color, pero algunas comunidades se han aglomerado dentro de otras, por lo que la visualización con el algoritmo kamada-kawai los grupos interconectados aparecen separados en el espacio y son fácilmente distinguibles. Un ejemplo de este suceso se encuentra en la comunidad representada con color verde. Dicha comunidad representa los países hispano-hablantes y los países que hablan portugués.

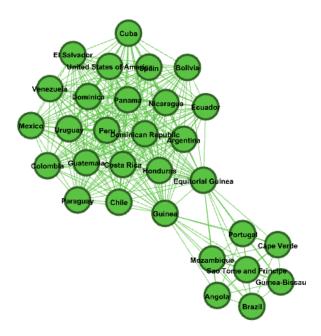


Ilustración 13: Comunidad hispana y portuguesa

6.2 Red de migrantes

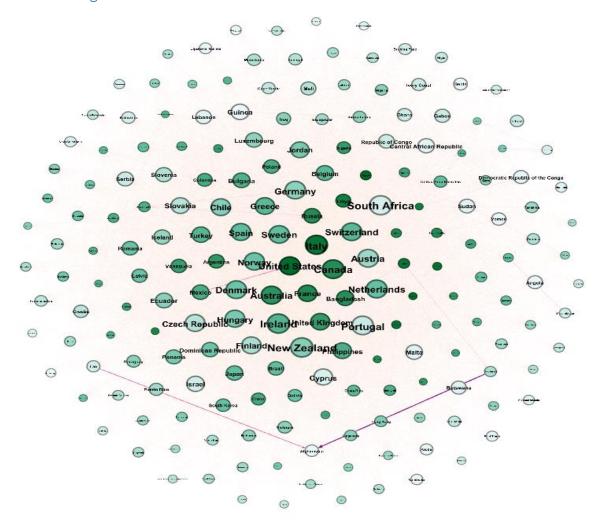


Ilustración 14: Visualización de la red de migración

Para que la visualización sea representativa se usa el algoritmo de Fruchterman-Reingold. De esta forma los nodos con mayor grado quedan centrados. También se ha usado un código de color para representar el grado de salida, desde el más claro que indica un menor grado, hasta el más oscuro. Las aristas son más oscuras si el peso asociado es mayor. El tamaño del nodo representa el grado de entrada.

Con esta información se pueden ver qué países tienen un mayor grado de entrada, de salida y entre qué países se producen los principales flujos migratorios.

Al observar la red se observa a simple vista que los países que más personas reciben, también tienen un grado de salida alto.

7 Discusión de los resultados

En cuanto a los resultados obtenidos, en ambas redes se puede hacer un análisis cultural a nivel mundial. Gracias a la primera red se detectan los grandes grupos coloniales del siglo XIV y XVIII. En la ilustración 12 se observan un grupo de color azul de mayor tamaño que los demás grupos, si se analiza ese grupo con algo más de detalle y se ve qué países forman parte del mismo se

pueden extraer algunas conclusiones. A continuación, se muestra la parte del grafo correspondiente solo a esa zona.

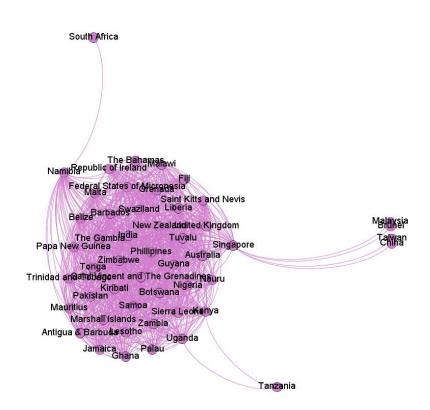


Ilustración 15:Comunidad de habla inglesa

Dentro del grupo totalmente conectado se encuentran países de habla principalmente inglesa, como Reino Unido o Irlanda. Dentro de la componente totalmente conexa, todos los países comparten el inglés como idioma. Esta comunidad refleja la existencia del Imperio Británico y como este se expandía y exportaba su cultura hacia países tanto del Atlántico, América y África.

Otra comunidad que llama la atención es la marcada de color rojo en la ilustración 12. Dicha comunidad tiene los siguientes nodos.

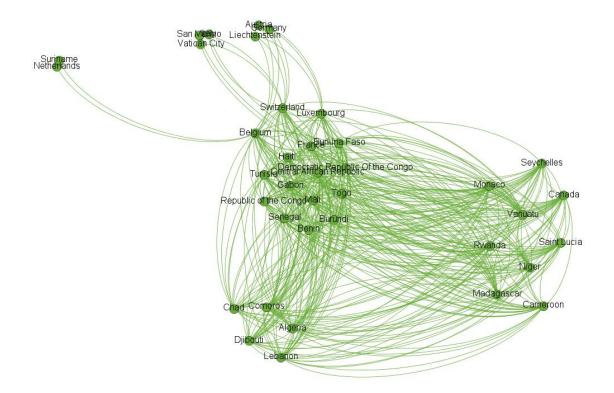


Ilustración 16: Comunidad de habla francesa

En esta imagen se ve que los países de la comunidad interconectada son de habla francesa. Esta comunidad existente es la huella de todas las colonias francesas. En comparación con el Imperio Británico, Francia no ha expandido su cultura a tantos países. Si se aprecian las colonias africanas que Francia llegó a tener y las colonias que Francia tenía en América en la zona de la actual Canadá.

Uno de los imperios coloniales que destacaron a lo largo de la historia fue el Español que duró hasta finales del siglo XIX. La influencia de España en la cultura mundial también se aprecia en el grafo. En la imagen 12 la comunidad hispana se encuentra marcada de color verde y es la comunidad que está más a la derecha, ya que la otra comunidad es la de habla portuguesa. Abajo se muestra con algo más de detalle.

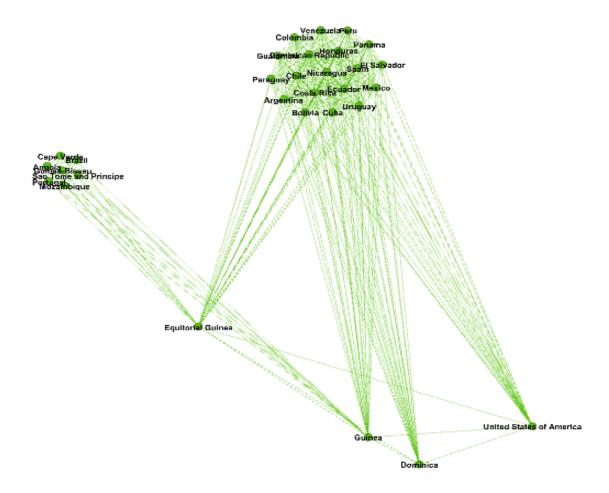


Ilustración 17: Comunidad de habla hispana y portuguesa

Se observa la influencia de España en todo el territorio de Latinoamérica. Destaca la presencia de los Estados Unidos. Esta presencia se debe a que Puerto Rico forma parte de EEUU y se habla español en dicha isla, por lo que es uno de los idiomas que se hablan de forma oficial en territorio estadounidense.

Junto al grupo hispanohablante, aparece el grupo que está bajo la influencia cultural de Portugal, que también formó colonias en algunas islas de Pacífico y en la actual zona de Brasil.

Hasta ahora, solo observando los grupos formados se han podido reconocer comunidades culturales a lo largo del mundo y las principales potencias coloniales en Europa. La comunidad que aparece de color rosado en la imagen 12 está formada por países árabes de la zona de la península arábiga.

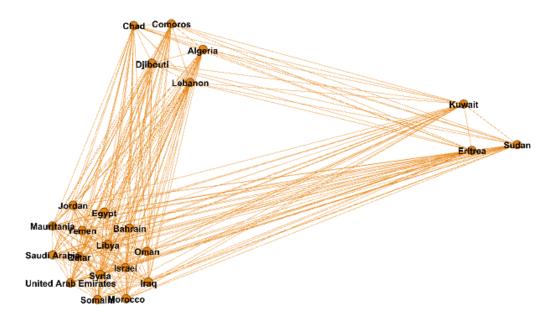


Ilustración 18: Grupo de países de cultura árabe

Otra conclusión que se puede extraer de la estructura de esta red se basa no en lo que se ve, sino en lo que no se ve. La Unión de Repúblicas Socialistas Soviéticas no puede apreciarse dentro de la red. Una pregunta razonable es por qué no aparece un grupo de países tan numeroso que coexistió bajo un mismo gobierno durante casi un siglo y llegó a ocupar países como Alemania. Se puede atribuir la ausencia de este grupo de países a que el asentamiento de un idioma y una cultura no se produce de forma rápida, no ocurre de un día a otro. Estos procesos se aceleran bajo un proceso de colonización, que tiende a imponer su cultura. Este no es el caso de la URSS, ya que por ejemplo Alemania, que estuvo bajo influencia Soviética hasta 1989, no tiene ningún enlace con Rusia.

Otra cuestión a tener en cuenta es cómo viaja la cultura hoy en día en un mundo globalizado. Para responder a esta cuestión la red de migración arroja algo de luz. En la imagen 14 se ve qué países tienen mayor recepción cultural. Llama la atención como EEUU recibe a mucha gente de Méjico, por lo que es de esperar que en dicho país surjan comunidades de habla hispana. En el centro de la visualización aparecen principalmente países europeos y norteamericanos. Viendo esto se ve que son los países del primer mundo los que muestran una tendencia mayor a la hora de importar y exportar cultura. De hecho, a grandes rasgos los países primermundistas comparten muchos rasgos culturales, aunque no se hable un mismo idioma o existan costumbres iguales, el modo de vida existente es bastante similar. Los países en vías de desarrollo aparecen en la periferia de la red, por lo que no tienen un intercambio cultural alto y se mantienen más aislados del mundo.

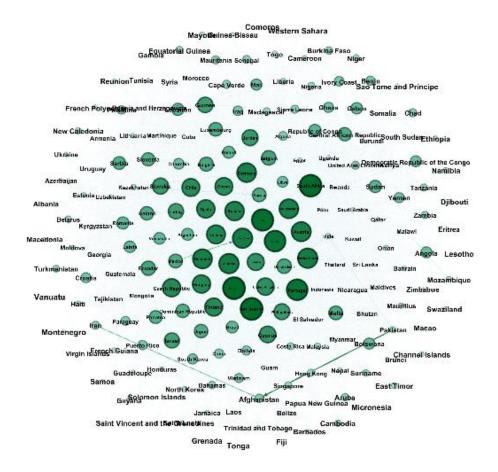


Ilustración 19: Red de migración, destacando los países periféricos

8 Comparación con otra red de la literatura

La red de idiomas analizada tiene una estructura bastante particular que no se encuentra en demasiadas redes en la literatura. Un enfoque interesante es saber si con simplemente las métricas se puede observar la particularidad de esta red.

La red de Silwood Park food servirá como punto de partida en esta comparación por su similitud en cuanto al tamaño. Dicha red ya ha sido analizada y los datos pueden consultarse en el artículo de *Statistical mechanics of complex networks* de Réka Albert y Albert-Lászlo Barabási

Red	Tamaño	Grado medio	Longitud media del	Coeficiente de
			camino	clustering
Silwood Park food	154	4.75	3.4	0.15
Idiomas	195	28.554	1.8344	0.927

Tabla 7: Comparación entre la red de idiomas y Silwood Park food

En la tabla se aprecia que la principal diferencia reside tanto en el grado medio como en el coeficiente de clustering debido a la interconexión que presenta la red de idiomas. La red de Silwood Park food no presenta dicha interconexión, pero aún sin presentar tantos enlaces la longitud media del camino no dista tanto con la red de idiomas.

Si bien se escoge una red más parecida en cuanto al grado medio, como la red de MEDLINE coauthorship, encontramos diferencias en otras medidas.

Red	Tamaño	Grado medio	Longitud media	Coeficiente	de
			del camino	clustering	
Idiomas	195	28.554	1.8344	0.927	
MEDLINE co-authorship	1520251	18.1	4.6	0.066	

Tabla 8:Comparación entre la red de idiomas y la red MEDLINE co-authorship

Tanto el tamaño como el coeficiente de clustering presentan grandes diferencias entre las dos redes.