

Aprendizaje Automático para Datos en Grafos

Laboratorio 3

Graciana Castro
4.808.848-2
gcastro@fing.edu.uy

Julian O'Flaherty
6.285.986-9
julian.o.flaherty@fing.edu.uy

1. Introducción

2. Grafos Erdös-Rényi

Los grafos *Erdös-Rényi*[1] (ER) son grafos aleatorios con un algoritmo de generación muy simple, donde a cada par de nodos se le asigna una arista con una probabilidad p . Pese a la simplicidad del algoritmo, los grafos ER tienen propiedades interesantes. Una de estas propiedades es que si

$$p > \frac{(1 + \epsilon)\ln(n)}{n} \quad \epsilon > 0 \quad (1)$$

entonces la probabilidad de que el grafo sea conexo es prácticamente 1. Analogamente, si

$$p < \frac{(1 - \epsilon)\ln(n)}{n} \quad \epsilon > 0 \quad (2)$$

entonces la probabilidad de que el grafo sea conexo es prácticamente 0.

En la figura 1 hacemos una validación empírica de esta propiedad, generando nueve grafos con el algoritmo de ER con 200 nodos, variando la probabilidad p alrededor del umbral de conectividad $p \approx \ln(n)/n \approx 0,02649$. Para cada valor de p se generan tres grafos distintos.

La primera observación es que para valores de p menores que el umbral de conectividad, el grafo es siempre desconexo, mientras que para valores de p mayores que el umbral de conectividad, el grafo es siempre conexo. Cabe recalcar que las condiciones de las cotas (1) y (2) son probabilísticas, por lo que puede existir una realización del grafo que no cumpla con la condición, solo que esto es altamente improbable. En el caso que p es igual al umbral de conectividad, no podemos afirmar ningún comportamiento, y se ve reflejado en que algunos de los grafos son conexos y otros no.

Otra característica a destacar, es que cuando el grafo ER no es conexo, las componentes que no pertenecen a la componente conexa más grande, son nodos aislados. Esto está relacionado a la tendencia de los grafos ER a tener una componente gigante cuando $np > 1$, condición que cumplen las 3 probabilidades elegidas.

3. Grafos SBM

Los grafos *Stochastic Block Model* (SBM) son una extensión natural de los grafos ER, que busca solventar algunas de las limitaciones de los mismos. En particular, los grafos ER no son capaces de generar comunidades, que es una característica importante en muchos grafos reales. Los grafos SBM solucionan este problema trabajando con una matriz de probabilidad Q simétrica, donde Q_{ij} es la probabilidad de que un nodo del grupo i se conecte con un nodo del grupo j . La cantidad de nodos en cada grupo son parámetros del modelo n_i .

Analizaremos el impacto de los vectores propios de la matriz Q en el grafo generado. Por simplicidad, trabajaremos con dos comunidades de igual tamaño $n_1 = n_2 = 50$. En la figura 2 se

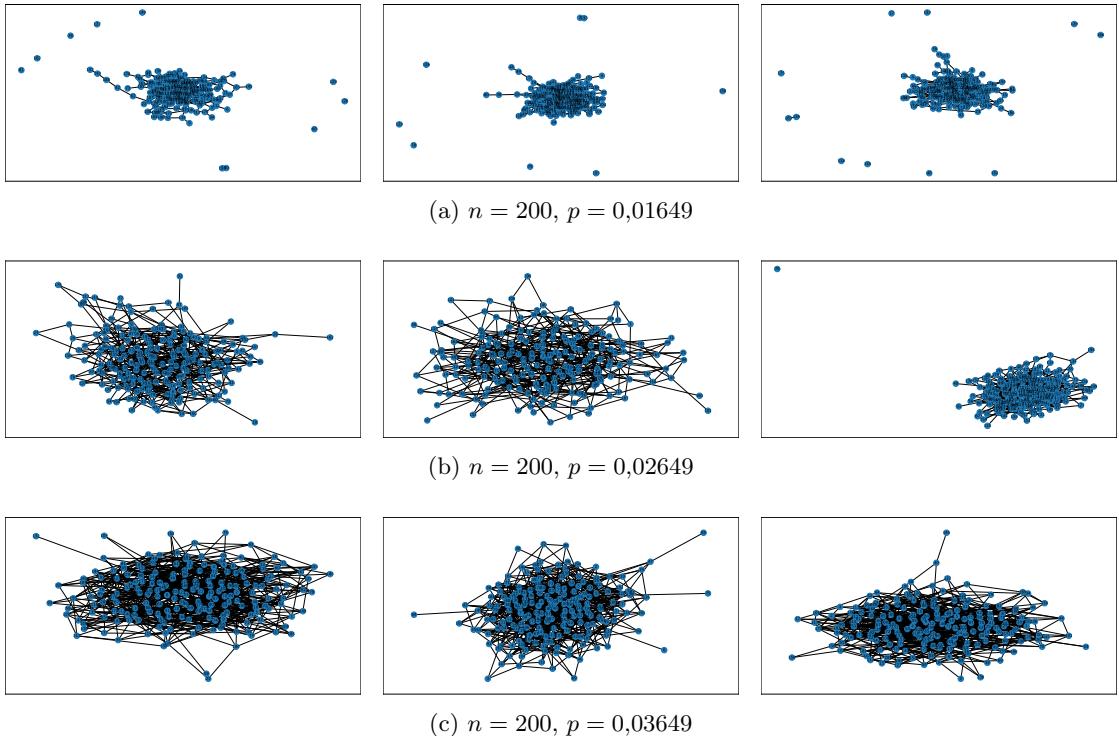


Figura 1: Nueve realizaciones de grafos ER con $n = 200$ y tres valores de p alrededor del umbral de conectividad $p \approx \ln()/n$.

muestran tres realizaciones de un grafo SBM, con las matrices:

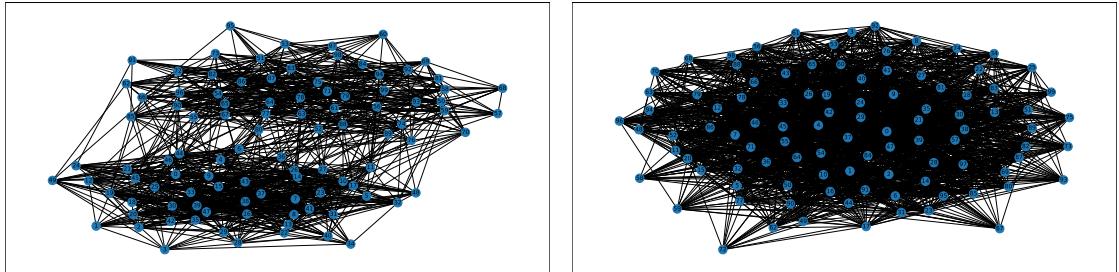
$$Q_1 = \begin{pmatrix} 0,4 & 0,05 \\ 0,05 & 0,3 \end{pmatrix} \quad Q_2 = \begin{pmatrix} 0,2 & 0,5 \\ 0,5 & 0,1 \end{pmatrix} \quad Q_3 = \begin{pmatrix} 0,8 & 0,05 \\ 0,05 & 0,8 \end{pmatrix} \quad (3)$$

Intuitivamente, los valores de la diagonal indican que tan densamente conectada esta una comunidad, mientras que los valores fuera indican que tanto se conectan entre comunidades. Viendo la figura 2b vemos el resultado de la matriz Q_2 , donde el valor de intraconexión de una comunidad es similar al valor de interconexión entre comunidades. Esto deriva en un valor propio negativo, y en un grafo con una sola comunidad.

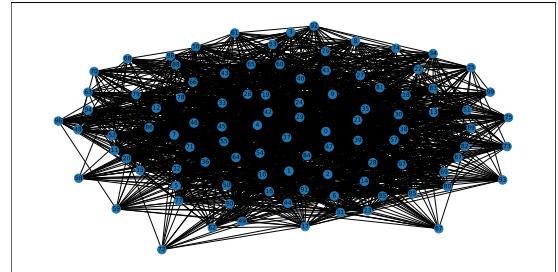
En los otros dos casos, tenemos matrices Q con valores propios positivos. Podemos notar como las comunidades de la figura 2c son muchas densas que en la figura 2a, lo cual es esperable dada la matriz Q asociada. Si observamos los valores propios, vemos que son mayores para Q_3 , de lo que podemos concluir que hay una correlación entre la densidad de la comunidad y el valor propio asociado.

En resumen, los valores propios de la matriz Q son indicadores de la cantidad y densidad de las comunidades del grafo generado, donde la cantidad de valores propios positivos es igual a la cantidad de comunidades. Apliquemos este resultado a una matriz Q con $n = [45, 5, 45, 5]$:

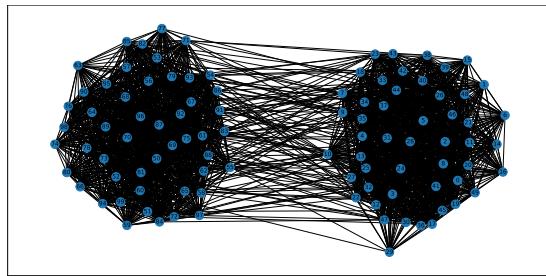
$$Q = \begin{pmatrix} 0,05 & 0,9 & 0,0 & 0,0 \\ 0,9 & 0,8 & 0,0 & 0,5 \\ 0,0 & 0,0 & 0,05 & 0,9 \\ 0,0 & 0,5 & 0,9 & 0,9 \end{pmatrix} \quad (4)$$



(a) SBM con valores propios 0,421 y 0,279, asociado a la matriz Q_1 (3).



(b) SBM con valores propios 0,652 y -0,352, asociado a la matriz Q_2 (3).



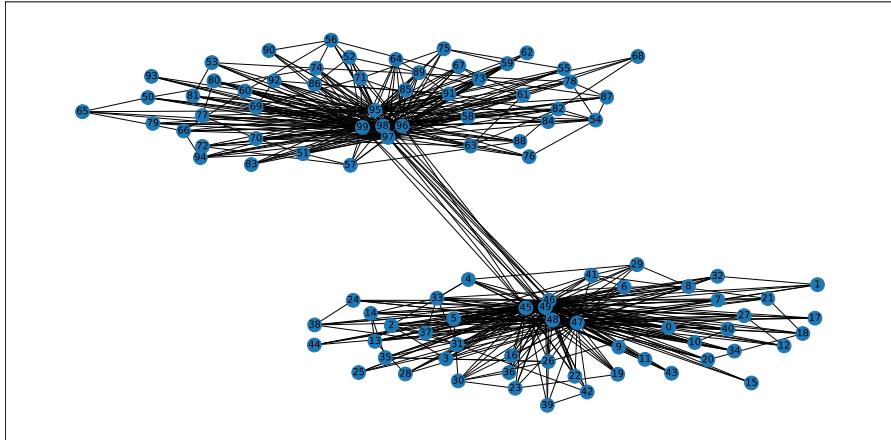
(c) SBM con valores propios 0,85 y 0,75, asociado a la matriz Q_3 (3).

Figura 2: Tres realizaciones de un grafo SBM con dos comunidades de igual tamaño, y dos valores distintos de Q .

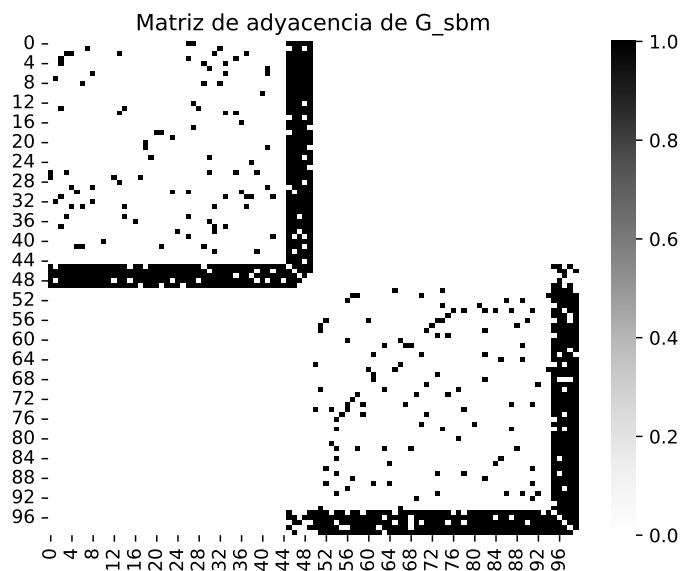
Antes que nada, intuitivamente podemos ver que hay comunidades que no van a existir, puesto que su valor en la diagonal es menor que los valores fuera de la diagonal. Calculando los valores propios de la matriz Q , obtenemos:

$$\lambda_1 = 1,81, \quad \lambda_2 = 1,11, \quad \lambda_3 = -0,71, \quad \lambda_4 = -0,41 \quad (5)$$

Podemos ver que hay dos valores propios positivos, y dos negativos. Por lo tanto, deberíamos obtener un grafo con dos comunidades. La figura 3a muestra el resultado de la generación, donde efectivamente observamos que el grafo resultante tiene dos comunidades. En la figura 3b podemos ver la matriz de adyacencia del grafo resultante, donde se entiende la intuición que planteamos inicialmente: la conexión intracomunidad (valor de la diagonal) tiene que ser suficientemente mayor que la conexión intercomunidad (valor fuera de la diagonal) para que se forme una comunidad.



(a) Realización del grafo SBM.



(b) Matriz de adyacencia del grafo SBM.

Figura 3: SBM con valores propios (5), asociado a la matriz Q (4).

4. Grafos Random Dot Product Graph (RDPG)

Los Random Dot Product Graphs (RDPG) representan una generalización tanto de los grafos Erdős-Rényi como de los Stochastic Block Models, introduciendo el concepto de *posiciones latentes*. En este modelo, cada nodo $i = 1, \dots, n$ se asocia con un vector $\mathbf{x}_i \in \mathbb{R}^d$ en un espacio latente de dimensión d , donde la probabilidad de conexión entre dos nodos i y j está determinada por el producto interno de sus respectivas posiciones latentes: $P(A_{ij} = 1) = \langle \mathbf{x}_i, \mathbf{x}_j \rangle$.

4.1. Relación con modelos previos

Para verificar la generalidad del modelo RDPG, analizamos cómo los modelos ER y SBM pueden ser expresados como casos particulares. En el caso de un grafo Erdős-Rényi con probabilidad de conexión p , la matriz de probabilidades es $\mathbf{P} = p\mathbf{1}_{n \times n}$. Para representar esto en el modelo RDPG, consideramos $d = 1$ y tomamos $\mathbf{x}_i = \sqrt{p}$ para todos los nodos, de forma que $\langle \mathbf{x}_i, \mathbf{x}_j \rangle = p$ para cualquier par de nodos.

Para el caso SBM, trabajamos con la matriz de probabilidades entre comunidades $\mathbf{Q} \in \mathbb{R}^{C \times C}$ y una matriz $\mathbf{Z} \in \mathbb{R}^{N \times C}$, donde $Z_{ij} = 1$ si el nodo i pertenece a la comunidad j , y $Z_{ij} = 0$ en caso contrario. La matriz de probabilidades entre nodos se expresa como $\mathbf{P} = \mathbf{Z}\mathbf{Q}\mathbf{Z}^T$. Para representar esto como RDPG, necesitamos encontrar una descomposición $\mathbf{XX}^T = \mathbf{Z}\mathbf{Q}\mathbf{Z}^T$.

Sin embargo, no todos los grafos SBM pueden ser representados exactamente por un RDPG. Esto se debe a que la matriz $\mathbf{P} = \mathbf{XX}^T$ debe ser semidefinida positiva por construcción (al ser una matriz de Gram), lo que impone restricciones sobre los valores propios de \mathbf{Q} . Cuando \mathbf{Q} tiene valores propios negativos no es posible la descomposición exacta bajo el modelo RDPG estándar.

4.2. Inferencia de posiciones latentes

Dado un grafo G con matriz de adyacencia \mathbf{A} , el objetivo es estimar la matriz de posiciones latentes \mathbf{X} . Dado que $\mathbb{E}[\mathbf{A}] = \mathbf{P} = \mathbf{XX}^T$, buscamos la mejor aproximación de rango d de la matriz \mathbf{A} . La estimación se obtiene resolviendo:

$$\hat{\mathbf{X}} = \underset{\mathbf{X}: \text{rango}(\mathbf{XX}^T)=d}{\operatorname{argmin}} \|\mathbf{A} - \mathbf{XX}^T\|_F^2 \quad (6)$$

La solución utiliza la descomposición espectral de $\mathbf{A} = \mathbf{Q}\Lambda\mathbf{Q}^T$. Ordenando los valores propios por magnitud y seleccionando los d mayores, obtenemos $\hat{\mathbf{X}} = \hat{\mathbf{Q}}\sqrt{\hat{\Lambda}}$, donde $\hat{\mathbf{Q}}$ y $\hat{\Lambda}$ corresponden al “recorte” apropiado.

Para la implementación utilizamos la biblioteca *graspologic*, específicamente la clase `AdjacencySpectralEmbed`. En los experimentos con grafos ER, realizamos la gráfica de la figura 4, que muestra la varianza de las estimaciones \mathbf{x}_i y de $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ para distintos valores de n y p . En ambos casos, al aumentar n , la varianza disminuye, reflejando una mayor estabilidad en las estimaciones. Sin embargo, al aumentar p , la varianza de \mathbf{x}_i disminuye, mientras que la varianza de $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ se mantiene prácticamente constante. Esto sugiere que el producto $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ es más robusto frente a variaciones en p , probablemente debido al efecto de promediado en el producto matricial.

La selección de la dimensión d se realiza mediante el análisis de los valores propios de la matriz de adyacencia. Al graficar los valores propios ordenados por magnitud, buscamos un “codo” en la curva que indique una separación clara entre los valores propios señal y ruido. La biblioteca *graspologic* automatiza este proceso mediante el parámetro `n_elbows`. Se muestra en la figura 5 como el codo se hace presente en $d = 1$ tal como se determinó en el código igualando `n_elbows` a 1.

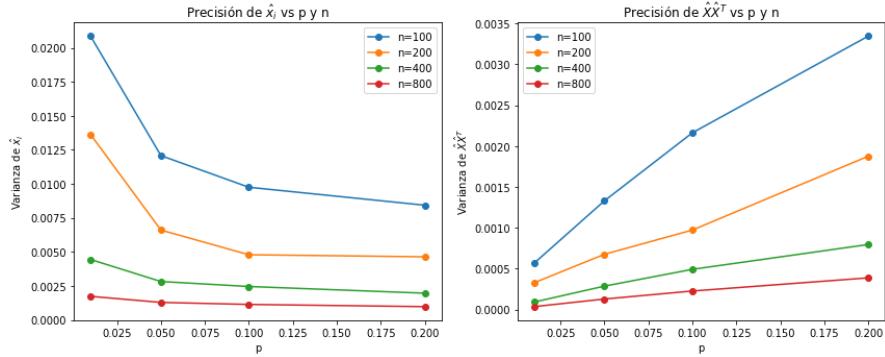


Figura 4: Varianza de las estimaciones \mathbf{x}_i y $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$ para distintos valores de n y p en grafos ER.

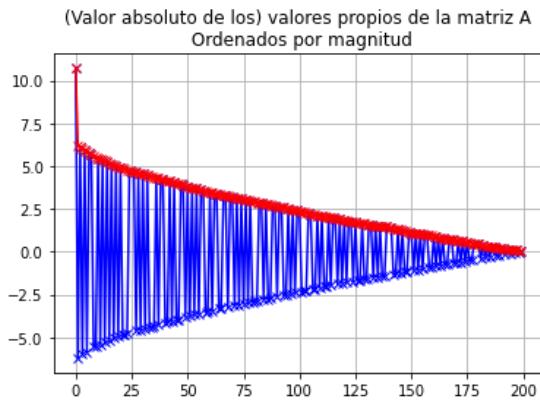


Figura 5: Valor absoluto de los valores propios de \mathbf{A} ordenados por magnitud.

4.3. Análisis de embeddings en grafos SBM

Para grafos SBM, analizamos la interpretación geométrica de las posiciones latentes estimadas. En experimentos con grafos de dos comunidades, representamos los embeddings en coordenadas polares, donde observamos que el ángulo entre vectores \mathbf{x}_i y \mathbf{x}_j está relacionado con la probabilidad de conexión: ángulos pequeños corresponden a alta probabilidad (nodos de la misma comunidad), mientras que ángulos grandes indican baja probabilidad de conexión (nodos de comunidades diferentes).

La magnitud $\|\mathbf{x}_i\|$ está relacionada con la probabilidad de conexión intra-comunidad, aproximadamente $\sqrt{p_c}$ donde p_c es la probabilidad de conexión dentro de la comunidad correspondiente. Esta interpretación geométrica facilita la comprensión de cómo el modelo RDPG captura la estructura de comunidades. Se observa en la figura 6 la distinción clara de dos comunidades, donde una se localiza sobre el eje horizontal y otra sobre el eje vertical. Si medimos los ángulos entre las componentes de una misma comunidad, obtenemos valores pequeños que tienden a cero, mientras que entre comunidades los ángulos son prácticamente ángulos rectos.



Figura 6: Representación en coordenadas polares de los embeddings latentes estimados para un grafo SBM con dos comunidades.

4.4. Detección de comunidades mediante clustering

Una vez obtenidas las posiciones latentes, aplicamos algoritmos de clustering sobre los vectores \mathbf{x}_i para detectar comunidades. Si bien lo clásico es utilizar k-means, los clusters observados muestran que utilizar Gaussian Mixture Models (GMM) puede ser una mejor opción.

Para la detección de comunidades a partir de los embeddings latentes, empleamos la clase `AutoGMMCluster` de la biblioteca *graspologic*, que permite ajustar automáticamente los hiperparámetros del modelo de mezclas gaussianas (GMM). En la figura 7 se ilustran los resultados obtenidos para diferentes valores de Q y n . Se observa que la tarea de identificar comunidades se vuelve más desafiante cuando la matriz Q presenta valores similares en la diagonal y la antidiagonal, ya que en estos casos la probabilidad de pertenencia a una u otra comunidad es comparable, dificultando la separación clara entre grupos.

5. Ejemplo real

Aplicaremos lo discutido en la Sección 4 a un grafo real. Contamos con datos históricos de partidos de fútbol entre equipos nacionales, desde 1872 hasta 2016. Con estos datos, podemos construirnos un grafo no dirigido, donde los nodos son los equipos (países) y las aristas denotan si jugaron una partida entre ellos. Opcionalmente, podemos agregar la cantidad de partidos disputados como el peso de la arista. Trataremos de detectar las comunidades en el grafo y analizar si se corresponde con lo que esperamos ver.

Analizaremos los datos para tres años:

- 1950: año de mundial pero con pocos datos, por lo que podemos observar las comunidades formadas con detalle.
- 2004: año donde se disputaron: Copa América, Eurocopa, Copa Asiática, Copa Africana de Naciones, y la OFC nations cup (oceania). Es decir, la copa de todas las confederaciones excepto CONCACAF.
- 2010: año de mundial, por lo que deberíamos tener mucha actividad entre algunos países de cada confederación.



Figura 7: Resultados de la detección de comunidades mediante clustering de los embeddings latentes en grafos SBM para distintos valores de Q y n .

5.1. Año 1950

En el 1950 se jugo el mundial de Brazil, donde Uruguay se proclamo campeón. El dataset no contiene muchos datos de ese año, contando solo con 141 partidos entre 67 equipos registrados, por lo que nos permite graficar los clusters para analizar con más detalle.

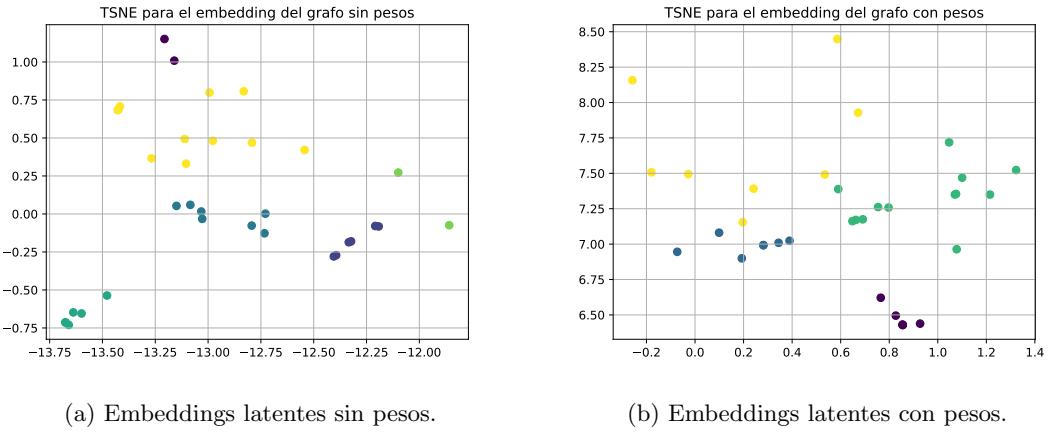
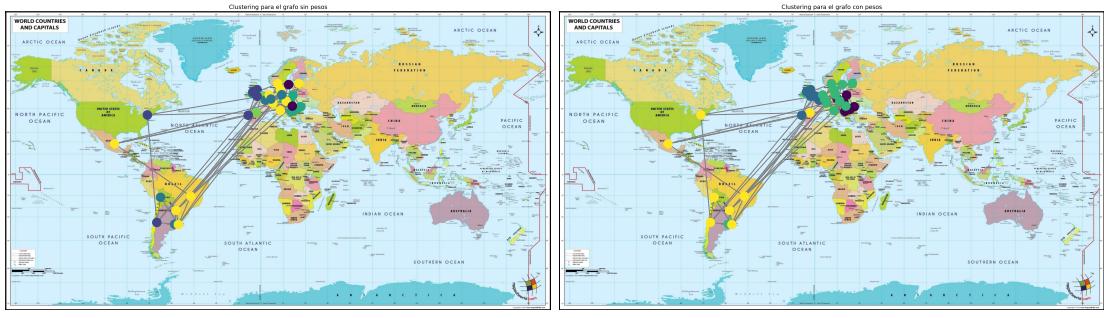


Figura 8: Resultado de la extracción de embeddings latentes para el año 1950 proyectado con TSNE. El color de los nodos indica la comunidad asignada por el algoritmo de clustering.

En la figura 8 se observan los embedding latentes obtenidos para cada país, para el caso con y sin pesos, así como las comunidades asignadas por el algoritmo de clustering. La principal diferencia que vemos, es que cuando incluimos los pesos desaparece una comunidad.

En la figura 9 observamos el resultado de la figura 8 proyectado en el mapa, lo que nos da una noción geográfica de las comunidades. Recordando que 1950 fue un año de mundial, cuando no



(a) Sin pesos.

(b) Con pesos.

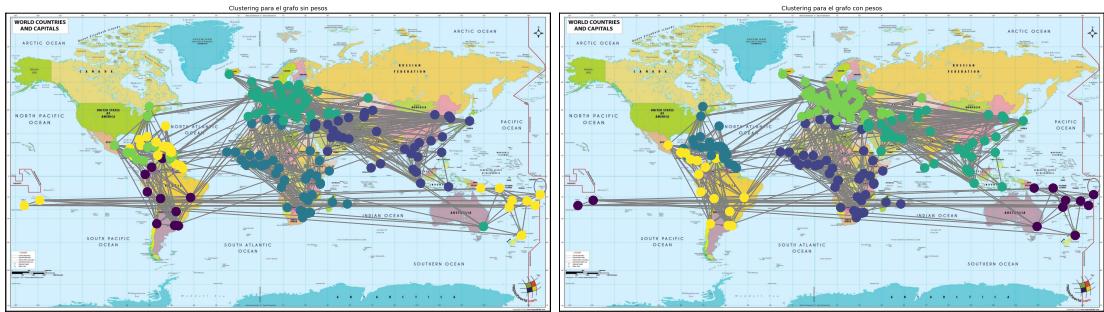
Figura 9: Mapa de los países con sus comunidades asignadas.

incluimos pesos vemos que las confederaciones no se distinguen claramente, viendo que américa del sur queda dividida en varias comunidades.

Cuando agregamos los pesos, notamos más coherencia en las comunidades obtenidas. Primero, vemos que los países de américa pertenecen todos a la misma comunidad salvo Argentina¹, que en ese año no participo del mundial (el resto de equipos americanos si lo hicieron). En europa, vemos tres comunidades, que se corresponden con europa del oeste (reino unido, francia, bélégica,...), europa central (alemania, suzia, checoslovaquia, polonia, etc...) y europa del este (yugoslavia, albania, etc...).

5.2. Año 2004

En el año 2004 tenemos muchos más datos, con 1089 partidos entre 189 países registrados. Como mencionamos anteriormente, en este año sucedieron todas las copas de las confederaciones excepto la de CONCACAF, por lo que las relaciones intra confederación deberían ser muy fuertes cuando incluimos pesos.



(a) Sin pesos.

(b) Con pesos.

Figura 10: Mapa de los países que disputaron partidos en 2004 con sus comunidades asignadas.

La primer observación es que en la figura 10b, el incluir pesos hace que la división de las comunidades sea muy clara y acertada. En el caso de América, vemos la inclusión de México y los países de América central en la misma comunidad que los países del sur, pero es consistente con la participación de algunos de ellos en la copa américa como invitados.

¹Argentina solo disputo un partido en 1950 contra Paraguay

Un detalle relevante, es que logra detectar a los países que juegan en una confederación distinta a la de su continente, cómo por ejemplo Armenia o Kazajstán.

Cuando no incluimos los pesos, las grandes comunidades se mantienen (Europa, África, Asia y América del sur), pero notamos algunos nodos como Australia, o el caribe y oceanía perteneciendo a la misma comunidad.

5.3. Año 2010

En el año 2010 se disputó el mundial de sudáfrica, por lo que deberíamos tener mucha actividad inter confederación, dado que además del mundial se suelen jugar muchos amistosos entre equipos no clasificados. Contamos con un total de 865 partidos disputados entre 198 países.

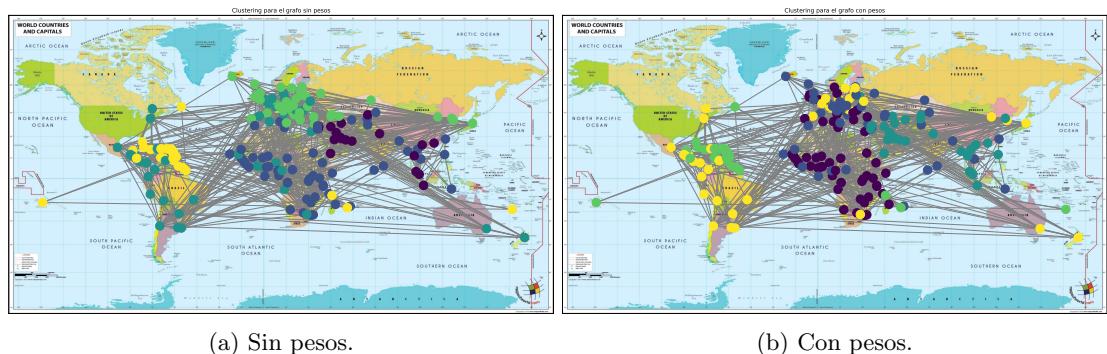


Figura 11: Mapa de los países con sus comunidades asignadas.

A diferencia del año 2004, en la figura 11 vemos que en el 2010 el resultado del clustering no es tan claro. Tanto cuando se incluyen pesos como cuando no, África, Europa y Asia tiene a sus integrantes repartidos entre una o más comunidades. En el caso de África, notamos que los países “mal clasificados” coinciden o son vecinos de los países que participaron del mundial (ej: Sudáfrica, Ghana, Costa de Marfil).

Por otro lado, cercano a un mundial se celebran los partidos de despedida, que suelen ser amistosos entre un país clasificado y uno que no clasificó, generalmente de confederaciones distintas. Esto genera conexiones incluso más aleatorias, lo que deja un grafo bastante ruidoso si el objetivo es la detección de comunidades.

6. Conclusiones

Referencias

- [1] Paul Erdős and Alfréd Rényi. On random graphs. i. *Publicationes Mathematicae*, 6(3-4):290–297, 1959.