Universidad de la República

Facultad de Ingeniería

Aprendizaje Automático para Datos en Grafos

Laboratorio 3 | Introducción a modelos generativos

Autores:
Federico Bello
Gonzalo Chiarlone

28 de septiembre de 2025

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

1.

Para verificar lo pedido alcanza con utilizar el método *is_connected* para distintos valores de n y p. La idea detrás es que cuanto mayor sea la probabilidad de conexión entre dos nodos, mas probabilidad debería tener el grafo de ser conexo. A su vez, a medida que la cantidad de nodos crece uno esperaría el mismo efecto, ya que se están realizando una mayor cantidad de sorteos con la misma probabilidad. Esto mencionado se observa en la tabla 1, donde se ve claramente que a mayor valor de p o n mayor es la probabilidad de que el grafo sea conexo.

n \ p	00002	0.002	0.02	0.2
10	False	False	False	False
100	False	False	False	True
1000	False	False	True	True
10000	False	True	True	True

Cuadro 1: Conexidad de G según p y n

2.

Cuando los valores propios son negativos se tiende a formar un grafo bipartito, mientras que cuando son positivos ocurre lo contrario, se forman comunidades muy intraconectadas[1] pero muy poco interconectadas. Es decir, si los valores propios son positivos se estará frente un grafo con *assorative mixing*, mientras que si son negativos es un grafo con *disassorative mixing*.

3.

Al utilizar los parámetros dados se obtiene el grafo de la figura 1. A priori, parecen haber dos comunidades distintas bien marcadas. Sin embargo, al analizar un poco mejor el grafo y los valores de la matriz de probabilidad, se ve como cada comunidad a su vez cuenta con un *núcleo*, en este caso de 5 nodos, los cuales son los responsables de las aristas entre cada grupo grande. Por lo tanto, si bien viendo los parámetros de la matriz de probabilidad el grafo debería tener 4 comunidades, se podría argumentar que cuenta solamente con 2.

4.

Para generar un Erdös-Rényi con probabilidad p es necesario obtener una matriz de probabilidad $\mathbf{P}$ de valor constante p. Es decir, se precisa que: $< \mathbf{x}_i, \mathbf{x}_j >= p \quad \forall \ \ i, j = 1 \dots n$. Tomando $\mathbf{x}_i = \mathbf{x}_j$ se llega a que:

$$\langle \mathbf{x}_i, \mathbf{x}_j \rangle = \langle \mathbf{x}_i, \mathbf{x}_i \rangle = ||\mathbf{x}_i||^2 = p \Leftrightarrow ||\mathbf{x}_i|| = \sqrt{p} \quad \forall \ \ i = 1 \dots n$$

[1] No existe la palabra, pero el prefijo indica conexión *dentro de* o *en el interior de*

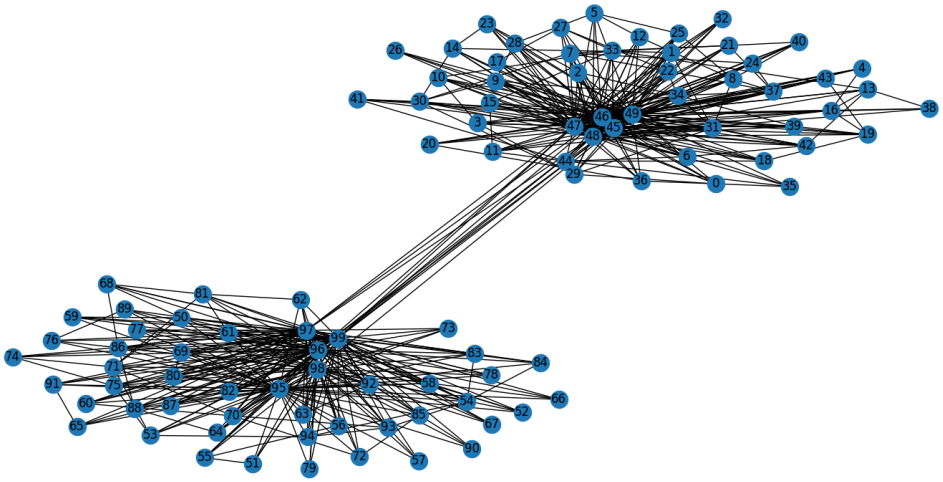

Figura 1: Grafo resultante de utilizar los parámetros dados

5.

En este caso, es necesario encontrar variables latentes tales que su multiplicación formen la matriz de probabilidades del SBM. Es decir, se debe buscar que $\pi_{qr} = \mathbf{x}_q \mathbf{x}_r$, donde π_{qr} son los valores de la matriz del SBM. Para esto, basta con definir un vector para cada uno de los clusters que cumpla con la igualdad anterior. A su vez, para definir la probabilidad de pertenencia a cada grupo es necesario que se cumpla: $P(\mathbf{X} = \mathbf{x}_q) = \alpha_q$

6.

Para poder utilizar el modelo RDPG, la matriz $\mathbf{P}$ debe tener todos los valores positivos. Esto se debe a que por un lado, la matriz es igual a $\mathbf{P} = \mathbf{X}\mathbf{X}^T$. Además, se tiene que la misma es simétrica y real, por lo tanto, es diagonalizable. Juntando los dos resultados anteriores se tiene que:

$$\mathbf{P} = \mathbf{Q}\Lambda\mathbf{Q^T} = \mathbf{Q}\sqrt{\Lambda}\sqrt{\Lambda}\mathbf{Q^T} = \mathbf{Q}\sqrt{\Lambda}(\mathbf{Q}\sqrt{\Lambda})^T = \mathbf{X}\mathbf{X^T}$$

donde se llega al resultado deseado recordando que $\mathbf{X}$ tiene entradas reales.

7.

En la figura 2 se ven los histograma de los valores del $\mathbf{X^2}$ estimado. La razón de tomarlo al cuadrado es para poder comprarlo mas fácilmente con la probabilidad, como se vio en la sección 4. Se observa entonces como tanto al ir aumentando el valor de n como el de p la aproximación mejora, centrándose cada vez mas en el valor correcto de p y disminuyendo su varianza.

8.

Luego, al analizar el producto $\hat{\mathbf{X}}\hat{\mathbf{X}^T}$, donde $\hat{\mathbf{X}}$ es la estimación. En este caso, idealmente se esperaría ver una matriz de valor constante p. En la figura 3 se observa como a medida que aumenta el valor de n y de p

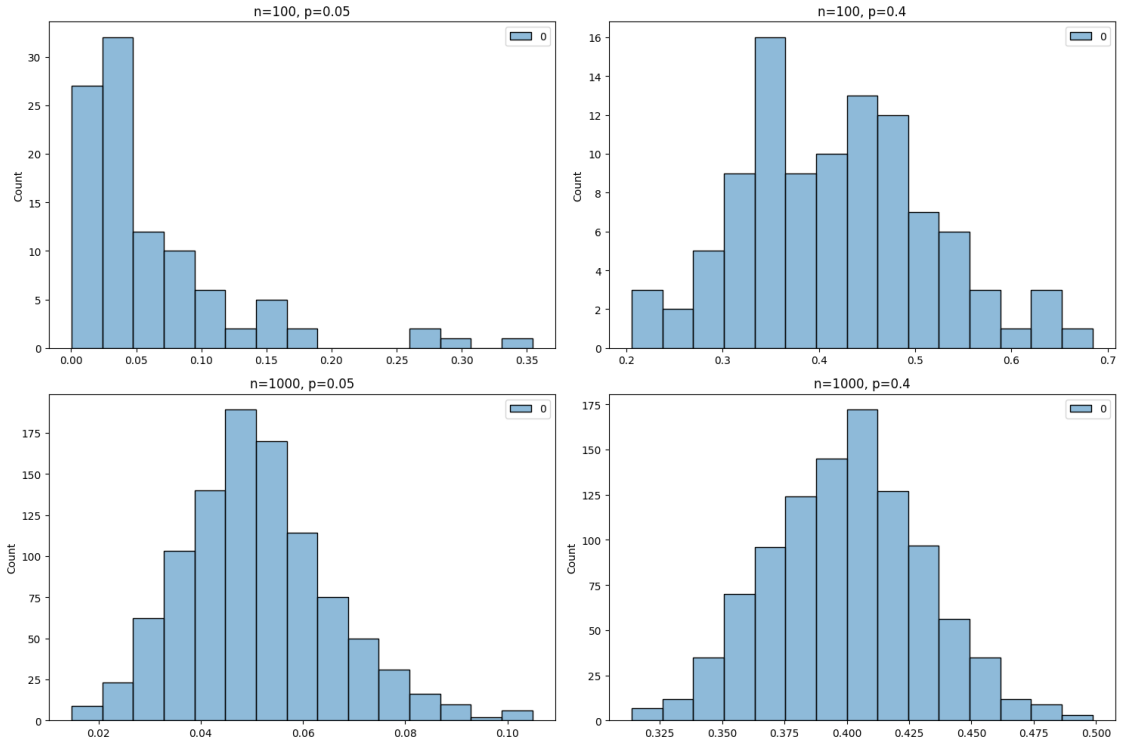

Figura 2: Estimación de los $\mathbf{x}_i^2$

la aproximación es cada vez mas precisa, al igual que el caso anterior.

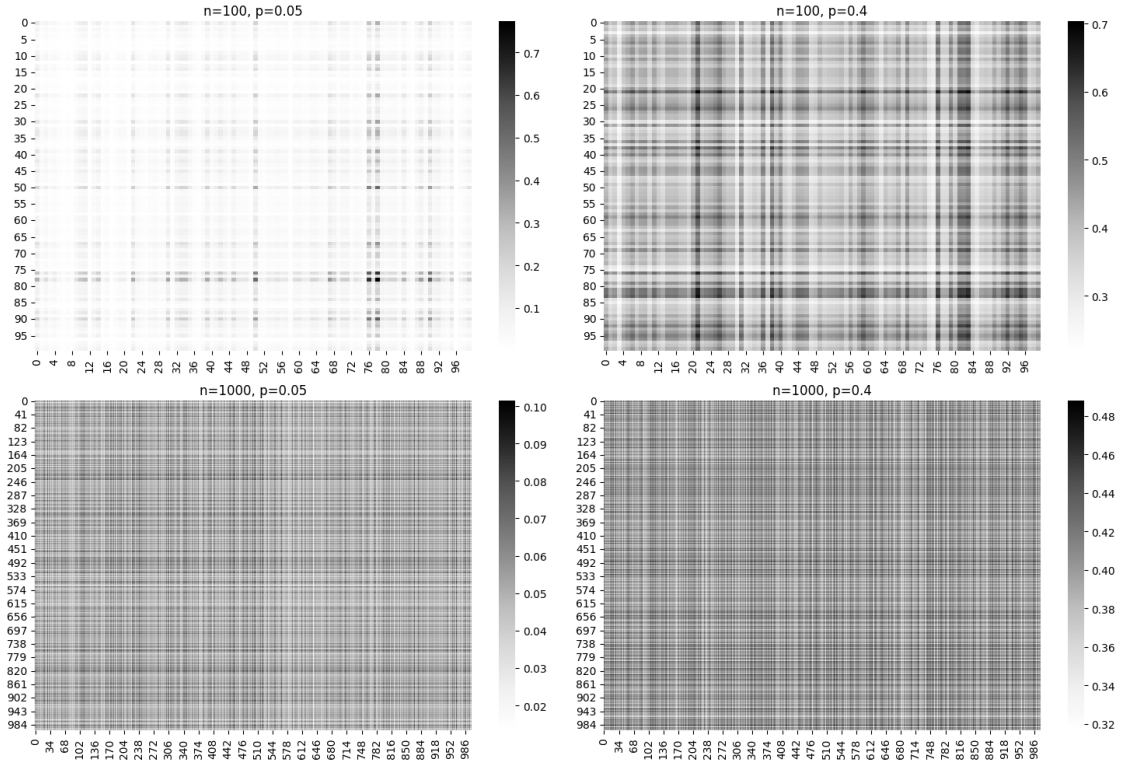

Figura 3: Estimación de $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$

9.

De manera genérica, para estimar la cantidad de valores propios a utilizar por la aproximación (d), alcanza con graficarlos y buscar el “codo"de la gráfica. Esto es para buscar un compromiso entre la cantidad de valores propios (y por lo tanto un menor costo computacional) y conseguir una representación fiel de la matriz $\mathbf{A}$. Luego del ‘’codo"de la gráfica, los valores propios disminuyen considerablemente en valor absoluto, indicando que conllevan menos información de la matriz. En el caso de un modelo ER se tiene que el valor óptimo de d es 1, lo cual se puede verificar fácilmente en la gráfica de la figura 4

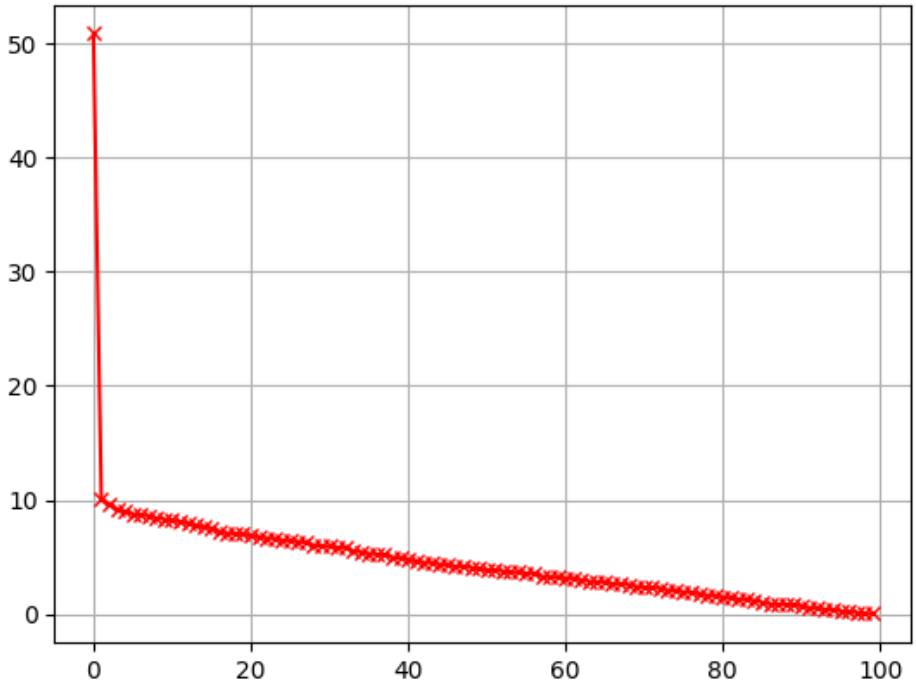

Figura 4: Valor absoluto de los valores propios de la matriz A, ordenados por magnitud

10.

Al hacer la gráfica polar de los valores del $\hat{\mathbf{X}}$, el ángulo entre dos nodos dados indica directamente su probabilidad de conexión. Si el ángulo tiene a ser ortogonal, su probabilidad de conexión es baja, ya que $< \mathbf{x_i}, \mathbf{x_j} > \approx 0$, mientras que la misma se maximiza si el ángulo que forman es nulo.

11.

Por otro lado, la magnitud de cada $\mathbf{x_i}$ indica que tan propenso es ese nodo a formar aristas, por lo que a mayor magnitud mayor grado del nodo.

Teniendo estos últimos dos comentarios a consideración, es fácil entonces detectar comunidades utilizando la representación polar. Viendo la figura 5 se diferencian claramente dos comunidades, las cuales están muy poco conectadas entre si (ya que los nodos forman un ángulo cercano a $\frac{\pi}{2}$) mientras que a su vez en cada comunidad están muy conectados (ya que son *casi* colineales).

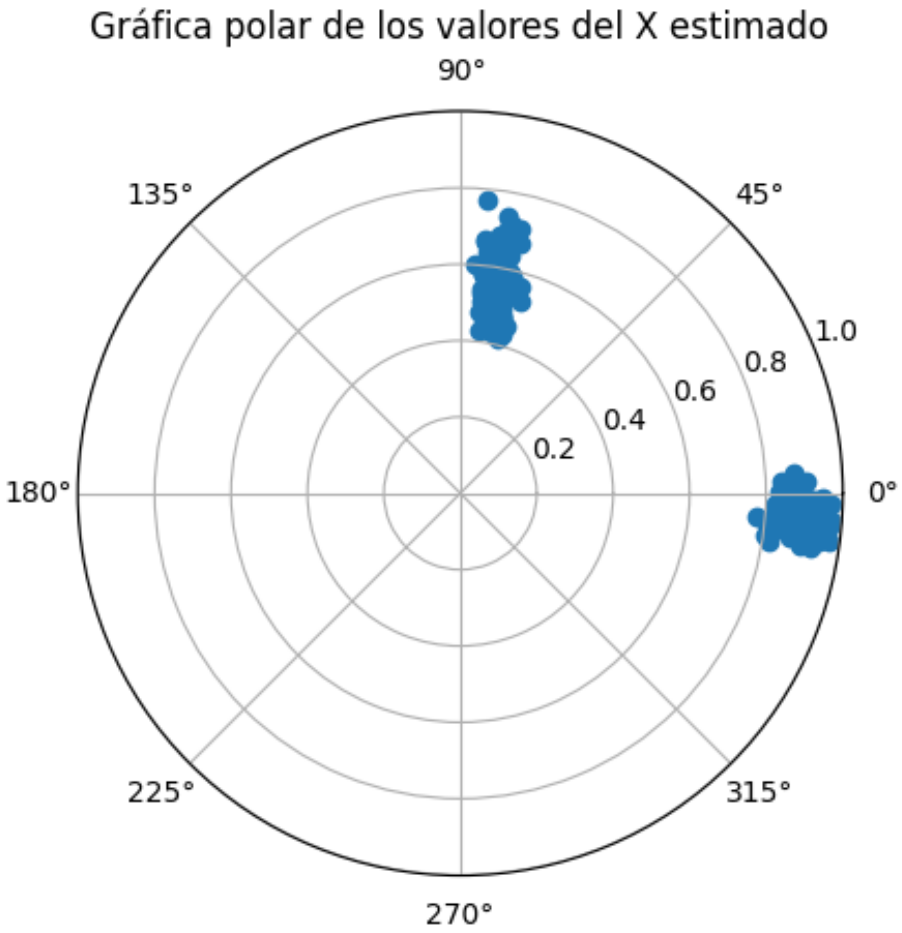

Figura 5: Gráfica polar de los valores del X estimado

12.

Por un lado, es muy fácil arruinar la búsqueda de comunidades cambiando los parámetros *min_ components* y *max_ components*. En el caso de *min_ components* sea mayor a la cantidad de comunidades reales o *max_ components* sea menor es claro ya que permite forzar un valor de comunidades incorrecto sobre el algoritmo. Por otro lado, si el intervalo entre *min_ components* y *max_ components* es muy grande, el algoritmo no suele dar buenos resultados, en particular, suele marcar una mayor cantidad de comunidades de las que realmente hay. A modo de ejemplo, al poner 1 y 99 como valores de cada parámetro respectivamente se ve una mayor cantidad de las 2 o 4 comunidades esperadas en el grafo de la pregunta 3 (figura 6). Por otro lado, también se vuelve difícil la detección de comunidades si las probabilidades de los vértices dentro de un mismo *cluster* son bajas. Esto tiene sentido ya que los grafos que generados con una matriz como la que se menciono tienden a tener pocas conexiones dentro de un mismo grupo. Además, la detección también se vuelve difícil para grafos con poca cantidad de nodos. En particular, para menos de 30 nodos se la clusterización tiende a ser mala.

13.

Al considerar los partidos de jugados entre selecciones es esperable que las comunidades sean cada continente. Esto se debe a que hay una gran cantidad de competencias o amistosos entre países vecinos, mientras

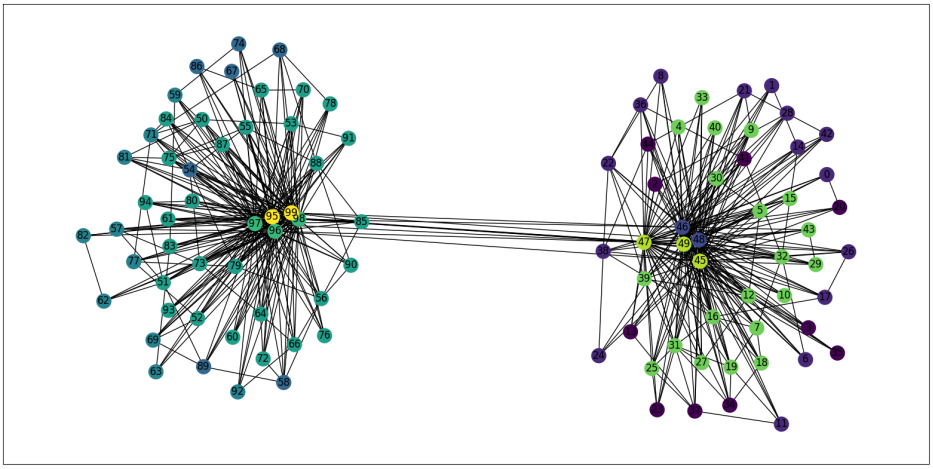

Figura 6: Estimación de $\hat{\mathbf{X}}\hat{\mathbf{X}}^T$

que las competencias mundiales como los mundiales o las olimpiadas son las menos. Además, también es esperable que al tomar el grafo con pesos las comunidades estén mejor divididas, ya que se esta le esta agregando información al problema, no es lo mismo jugar un ocasional partido contra una selección de otro continente en comparación a la gran cantidad de partidos que se juegan entre países cercanos.