1. Criterios de normalización

A continuación se presentan diferentes criterios de normalización a usar con la distancia espectral.

1) **Normalización Min-Max**

Sea D el conjunto de todas las distancias espectrales entre los modelos teóricos y los modelos empíricos a estudiar. Definimos la distancia espectral normalizada $d_{norm}(G, G')$ entre los grafos G y G' de la siguiente manera:

$$d_{norm}(G, G') = \frac{d(G, G') - \text{mín}(D)}{\text{máx}(D) - \text{mín}(D)}$$

Donde $d(G, G')$ denota la distancia espectral.

Esta medida es dependiente del conjunto de datos de grafos del que se dispone, puede ser interesante para posteriormente hacer clustering.

2) **Normalización por cotas**

a) Por el teorema sobre la cota superior del radio espectral [Francesco Bullo], podemos afirmar que $\lambda_i^A \leqslant \text{máx}(A\mathbb{1}_n)$. Entonces, sabemos que el sumatorio empleado para calcular la distancia espectral $\sqrt{\sum_{i=1}^{n}(\lambda_i^A - \lambda_i^{A'})^2} \leqslant \sqrt{\sum_{i=1}^{n}\left(\text{máx}(A\mathbb{1}_n) - \text{mín}(A'\mathbb{1}_n)\right)^2}$. Por tanto, podemos normalizar los valores de las distancias espectrales en forma de tasa de represantividad de distancia mediante:

$$d_{norm}(G, G')_A = \frac{d(G, G')}{\sqrt{\sum_{i=1}^{n}\left(\text{máx}(A\mathbb{1}_n) - \text{mín}(A'\mathbb{1}_n)\right)^2}}$$

Este ejemplo representa el calculo de esta tasa con respecto a la matriz de adyacencia.

b) Para la matriz laplaciana, el teorema de los ceros de Geršgorin nos permite obviar la cota inferior por ser cero. No obstante, para la cota superior, por la definición variacional, tendríamos que:

$$\lambda_n = \sup_{\substack{x \in R^n \\ x \perp v_1, \ldots, v_{n-1}}} \frac{x^T L x}{x^T x} = \inf_{\substack{x \in R^n \\ x \neq 0}} \frac{x^T L x}{x^T x}$$

Entonces esta tasa quedaría como:

$$d_{norm}(G, G')_L = \frac{d(G, G')}{\sqrt{\sum_{i=1}^{n}\left(\inf_{\substack{x \in R^n \\ x \neq 0}} \frac{x^T L x}{x^T x}\right)^2}} = \frac{d(G, G')}{\sqrt{n\left(\inf_{\substack{x \in R^n \\ x \neq 0}} \frac{x^T L x}{x^T x}\right)^2}} = \frac{d(G, G')}{\sqrt{n\,(\lambda_n)^2}}$$

Sin embargo, lo interesante de esta normalización, es que al dividir por el autovalor mayor, realmente lo que estamos haciendo es dividir por el máximo valor del cociente de Rayleigh. Esta división es típica en análisis espectral, ya que nos permite hacer la distancia espectral independiente de la magnitud de escala de los autovalores crecientes de la matriz Laplaciana normalizada.

En caso de no querer mantener esta normalización entre 0 y 1, simplemente se podría dividir por el radio espectral o por la diferencia entre los radios espectrales de ambos grafos, permitiendo obviar la diferencia de escala en la resta.

$$\frac{d(G,G')}{\left(\lambda_n^L - \lambda_n^{L'}\right)^2}$$

c) Para la matriz laplaciana normalizada, si el grafo adjunto es bipartito y conexo, entonces podemos afirmar que el mayor de los autovalores cumple $\lambda_n = 2$ [Spectral Stanford].

Por tanto, si en la expresión anterior tenemos en cuenta este resultado y el teorema de los ceros de Geršgorin, se tiene que:

$$d_{norm}(G,G')_{\overline{L}} = \frac{d(G,G')}{\sqrt{\sum_{i=1}^{n} 4}} = \frac{d(G,G')}{\sqrt{4n}}$$

En caso de no serlo, se puede recurrir a las técnicas anteriormente descritas.

Al trabajar con la matriz laplaciana normalizada, la ventaja que ofrece esta forma de trabajo es permitir una comparación en la variación de los grados de los nodos entre dos redes. Esto supone que analizando la matriz laplaciana normalizada y comparándola con la correspondiente del modelo ER, se puede ofrecer una comparativa de proximidad del grado de aleatoriedad entre las distribuciones de los grafos. [Esto se debe a que si los autovalores más grandes están próximos a ceros se espera una homogeneidad en la distribución de los grados]

3) **Normalización por IQR**

Aunque sea una forma simple de trabajo, la normalización propuesta en algunos papers trabaja con un criterio Z-score. No obstante, esto siempre implica asumir normalidad. Una variación más robusta que nos permita trabajar con la distancia espectral podría ser el rango interquartílico.

De esta manera, podemos coger todas las distancias espectrales calculadas y calcular el primer y tercer cuartil. La normalización quedaría de la siguiente manera:

$$IQR = Q_3 - Q_1$$

$$d_{norm}(G,G') = \frac{d(G,G') - Q_1}{IQR}$$

4) **Normalización vía índices**

a) Una manera de normalizar los valores obtenidos por la distancia espectral es mediante el uso de índices. Estos índices han sido estudiados en profundidad y nos permiten atajar características de los grafos bipartitos que nos pueden servir como indicadores de patrones o de discrepancias o de similitud entre diferentes redes.

Como el objetivo es comparar grafos bipartitos de diferentes escalas, es necesario utilizar índices que sean independientes de la escala del grafo. Algunos propuestos incluyen los siguientes:

1) Índice de anidamiento por temperatura (NTI)
2) Conectancia
3) Regularidad de las interacciones (IE)
4) Índice de Jaccard

Sería interesante ver qué índices pueden ser de mayor aportación bajo juicio experto en el tipo de redes que estamos analizando (polinizadores, dispersores y parásitos).

Asimismo, se podría analizar mediante técnicas de machine learning algún tipo de regresión que fuera capaz de registrar la relevancia de los índices mediante ponderación. Una vez obtenida, esta se utilizaría para normalizar la distancia mediante la siguiente notación:

$$d_{norm}(G, G') = \frac{d(G, G')}{\alpha_1 I_1 + \alpha_2 I_2 + ... + \alpha_n I_n}$$

Este ejemplo está basado en los índices propuestos en el paper [Indices, Graphs and Null Models: Analyzing Bipartite Ecological Networks].

Si se obtuviera una configuración aceptable, mantendríamos el problema de la escala resuelto y tendríamos una forma de comparar distancias espectrales en base a las correlaciones entre los valores obtenidos por cada red su distancia espectral sin normalizar y sus índices.

5) **Normalización vía matriz de covarianza**

a) Asimismo, se podrían considerar los autovalores de la matriz de covarianza para los datos obtenidos y dividir la distancia espectral por la suma de los k autovalores. Esto se obtendría de la siguiente manera:

Sea

$$\Sigma = \begin{bmatrix} \mathrm{Var}(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_n) \\ \mathrm{Cov}(X_2, X_1) & \mathrm{Var}(X_2) & \cdots & \mathrm{Cov}(X_2, X_n) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_n, X_1) & \mathrm{Cov}(X_n, X_2) & \cdots & \mathrm{Var}(X_n) \end{bmatrix}$$

$$\det(\Sigma - \lambda I) = (\lambda_{I_1}^{\Sigma} - k_1)(\lambda_{I_2}^{\Sigma} - k_2) \cdots (\lambda_{I_n}^{\Sigma} - k_n)$$

Es importante destacar que cada variable aleatoria debe seguir estrictamente su definición para que esta matriz aporte valor. Cada X_i debe representar un modelo nulo del que se extraen redes generadas a través de los métodos comentados anteriormente.

La distancia normalizada quedaría de la siguiente manera:

$$d_{norm}(G, G') = \frac{d(G, G')}{\lambda_{I_1}^{\Sigma} + \lambda_{I_2}^{\Sigma} + ... + \lambda_{I_n}^{\Sigma}}$$

Como esta escala no se ajusta al rango $[0, 1]$, podemos aplicar la siguiente transformación:

$$z_{norm}(G, G') = 1 - \exp\left(-d_{norm}(G, G')\right)$$

De esta manera, la normalización de esta distancia tiene en cuenta los siguientes factores:

1) La magnitud de la escala una vez normalizada es una magnitud relativa que calcula la importancia del valor original con respecto a la variabilidad capturada por los autovalores. Es decir, un valor grande en escala representaría que esa distancia espectral está muy lejos de la variabilidad total de los modelos nulos y, por tanto, es estadísticamente lejana de la red real.
2) Esta normalización relativiza la escala pero conserva la idea principal de medir diferencias entre modelos nulos y sus respectivas redes reales. En propocionalidad, un valor cercano a 1 indica que esa distancia espectral es un contribuidor principal de la variabilidad de los modelos nulos.
3) Conserva la adimensionalidad en la medida
4) La escala se mantiene entre 0 y 1
5) Al ser la transformación no lineal, existe un factor de amortiguamiento que es el término exponencial.
6) Esta transformación evita que los valores tiendan a 1 de forma significativa y solo ocurrirá cuando la distancia es sustancialmente mayor que la variabilidad.
7) En este calculo se tiene en cuenta la correlación entre los diferentes modelos nulos y pondera la distancia espectral para que sea comparable en términos estadísticos y no geométricos.