

Reporte de trabajo: Análisis de Redes con datos del Tamizaje para la detección de riesgos a la Salud Mental por Covid-19

Elena Villalobos Nolasco

Enero, 2022.

El presente documento tiene los avances realizados en el proyecto de **Análisis de Redes como alternativa de análisis para estudiar psicopatologías**¹ y tiene como base principal el tutorial de Epskamp et al. (2018). En la primera parte del documento hablaremos de los conceptos básicos para entender el análisis de redes, así como de la descripción de los gráficos principales para estudiar las características de las redes. En la segunda parte, presentaremos algunos resultados preliminares obtenidos de la base de datos del Tamizaje para la detección de riesgos a la Salud Mental por Covid-19.

1. Conceptos básicos para entender el análisis de redes

El análisis de redes, basado en la teoría de grafos, es un método analítico que tiene como objetivo estudiar las interrelaciones existentes entre entidades. En éste, los nodos representan entidades (aeropuestos, personas, etc.), y las conexiones, también conocidas como aristas, son observadas, medidas y conocidas (número de vuelos entre aeropuestos, amistades, etc.). Además, en el análisis de redes, las aristas pueden ser dirigidas o no dirigidas, en las primeras hay un nodo de origen y un nodo destino, en las segundas una arista representa una conexión simétrica entre dos nodos. Para este estudio se utilizarán redes no dirigidas que representan una relación entre los nodos.

De manera similar pero estructuralmente diferente al análisis de redes, las redes psicológicas consisten en nodos que representan las variables observadas que están interconectados por aristas que representan relaciones estadísticas, las cuales se infieren a partir de los mismos datos. Por ejemplo, en un cuestionario psicométrico las respuestas a las preguntas de presencia/ausencia de síntomas son los nodos, y lo que se infiere son interrelaciones entre los nodos o síntomas. Como dichas interrelaciones se estiman a partir de los datos, es muy recomendable tener más datos que ayuden a hacer cálculos más confiables.

¹Todo el código e información asociada a este proyecto se encuentra en el siguiente repositorio: https://github.com/ElenaVillano/mental_networks

A grandes rasgos, el análisis de redes psicológicas tiene dos pasos involucrados:

1. Estimar el modelo estadístico sobre los datos, es decir, la estimación de la red con sus pesos correspondientes entre las variables observadas.
2. Analizar si las estimaciones de los pesos de la red son adecuadas, a partir de medidas desarrolladas en teoría de grafos.

De manera específica, en las redes psicológicas **los nodos representan las variables observadas y las aristas representan coeficientes de correlación parcial entre variables (i.e. pesos de las aristas), condicionadas sobre las otras variables observadas.**

En toda red estimada, se tiene que valorar la importancia o conectividad de los nodos dentro de la red, para lo que se computan las medidas de centralidad que son:

Fuerza del nodo: que tan bien el nodo está **directamente** conectado con otros nodos. Una manera de interpretarlo, es que tan importante es el nodo dentro de la red conectando directamente con otros nodos.

Cercanía o closeness: cuantificar que tan bien el nodo es indirectamente conectado a otros nodos. Es una especie de promedio que nos dice qué tan lejos están los otros nodos de la red.

Intermediación o betweenness: que tan importante es un nodo en el camino promedio entre dos nodos. Otra manera de decirlo es que tan único es el nodo para conectar con otros pares de nodos de la red.

Dentro del tutorial de Epskamp et al. (2018), simularon una red con un modelo gráfico gaussiano del cual obtuvieron las medidas de centralidad ya descritas, ambos presentados en la Figura 1. La red la estimaron utilizando una base de datos donde cada uno de los 17 reactivos, hablaban de síntomas de estrés post-traumático. En ésta se puede observar de lado izquierdo que las conexiones más fuertes positivas son entre los nodos o síntomas 16 y 17, 3 y 4, y 11 y 5, pues entre más concentración de color y grosor en la arista, más correlación positiva parcial hay entre esos nodos. De la misma manera, se puede observar una relación negativa delgada entre los nodos 10 y 12, representada por la presencia del color rojo. También, la ausencia de conexión entre los nodos es indicio de que hay independencia estadística (no hay relación) entre los nodos cuando se condiciona sobre los otros síntomas. En cuanto a las medidas de centralidad en el panel derecho, el nodo 17 y el 3 salen como los más importantes para la red, tanto para la fuerza del nodo como para *betweenness*.

A partir de los parámetros obtenidos del modelo anterior se simuló datos de 500 sujetos más. Se observó que la red que se generó era muy similar a la original. Sin embargo, se encontró que las medidas de centralidad difieren mucho entre las simulaciones y el modelo original. Por lo que destacaron la importancia de evaluar la precisión de las estructuras de las redes psicológicas, y para ello sugieren los siguientes tres pasos:

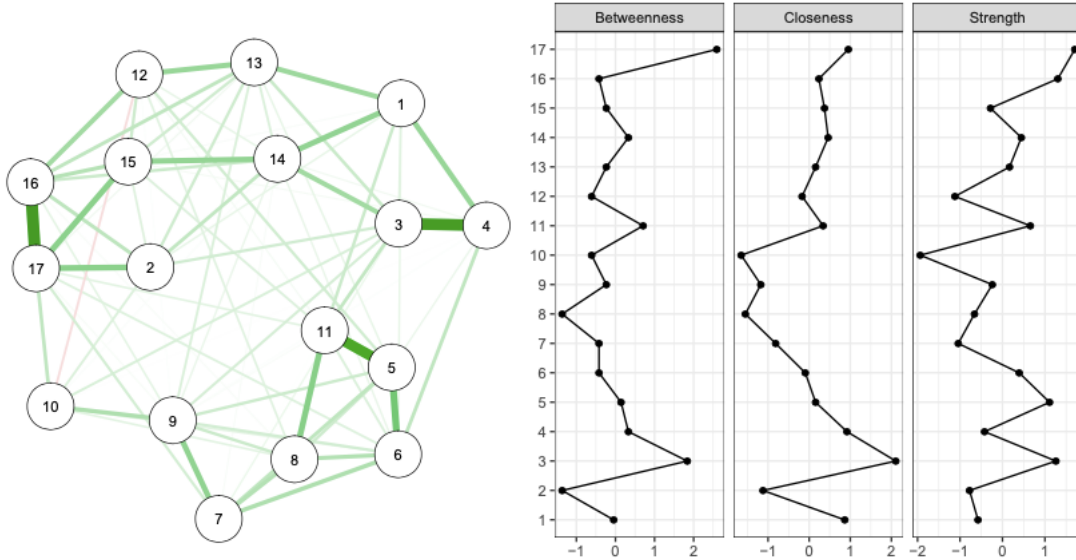


Figura 1: Red de modelo gráfico gaussiano del cuestionario PTSD con 17 reactivos (panel izquierdo) y sus respectivos índices de centralidad (panel derecho), presentados en el tutorial de Epskamp et al. (2018)

1. Estimar la precisión de los pesos de las aristas, utilizando Intervalos de Confianza con bootstrap.
2. Evaluar la estabilidad de los índices de centralidad observados sobre sub-conjuntos de datos.
3. Realizar tests de diferencias significativas entre los pesos de las aristas y los índices de centralidad.

Antes de presentar la evaluación de la precisión de las redes, en la siguiente sección destacaremos algunas partes importantes sobre el análisis de redes.

Especificaciones de estimaciones de redes psicológicas.

Un modelo de redes que se usa popularmente para estimar redes psicológicas es un Campo Aleatorio de Markov por Pares (Pairwise Markov Random Field, PMRF). En este modelo, los nodos representan variables, y se conectan a través de aristas no-direccionadas, indicando dependencia condicional entre dos variables; dos variables que no están conectadas son independientes después de condicionar sobre las otras variables. Cuando los datos son normales multivariados, esta independencia condicional correspondería a una correlación parcial igual a cero. De manera específica, los parámetros de los pesos de las aristas se interpretan como la fuerza de asociaciones únicas entre variables, que **pueden** indicar potenciales relaciones causales.

Cuando los datos son binarios, el modelo PRMF a utilizar es el modelo Ising. Mientras que si los datos siguen una densidad normal multivariada, el modelo apropiado de PRMF es el modelo gráfico

gausiano (GGM), en el que las aristas pueden ser directamente interpretadas como coeficientes de correlación parcial. El GGM requiere un estimado de la matriz de covarianza como entrada, que en caso de que los datos sean ordinales se pueden utilizar correlaciones policóricas (i.e. correlación para variables ordinales de variables latentes).

El modelo PMRF tiene el problema de que el número de parámetros a estimar crece mucho con el tamaño de la red y en generalmente en el área de psicología no se tienen tantos datos que compensen esta sobre-estimación. Para tratar dicho problema se utiliza la forma de regularización de LASSO (Least Absolute Shrinkage and Selection Operator), técnica penaliza el uso de muchos parámetros, limitando la suma de valores paramétricos absolutos. Por lo que LASSO regresa un modelo de red más conservador, que en otras palabras, sólo un pequeño número de aristas se utilizan para explicar la covariación de la estructura en los datos².

1.1. Precisión de los pesos de las aristas

Para evaluar la variabilidad de los pesos de las aristas se pueden estimar Intervalos de Confianza³ con la técnica de re-muestreo bootstrap. Se aconseja que para datos ordinales, se utilice bootstrap no paramétrico.

Es importante mencionar que los resultados del bootstrap no deberían ser utilizados para hacer test de significancia diferente de cero, pues la regularización de LASSO ayuda a quitar los pesos que no son importantes en la red. Esto significa que los pesos que aparecen dentro de la red, ya pueden ser significativamente diferentes de cero, por lo que conviene considerarlos como importantes.

Entonces la interpretación de los Intervalos de Confianza con bootstrap para los pesos de las aristas no se deben asumir diferentes de cero, si no que sólo muestran la precisión de esos pesos y lo que se debe de hacer es comparar los valores de las aristas entre ellos. Cuando un intervalo es ancho, es difícil hacer la interpretación de la fuerza del nodo, por lo que probablemente resultará en poca precisión para las otras medidas de centralidad.

La Figura 2, son los Intervalos de Confianza para el modelo presentado al inicio de este documento (figura 1). En este gráfico cada línea horizontal representa una arista de la red, ordenadas desde la arista con el peso más alto, hasta la arista con el peso más bajo. Es decir, las conexiones colocadas en la parte superior del gráfico son las más fuertes, que son entre los nodos 17 y 16, 3 y 4, y 11 y 5. Se podría considerar que son, de manera confiable, las tres aristas con más fuerza de conexión debido a que sus Intervalos de Confianza no se superponen con los intervalos de ninguna otra arista.

²El paquete de `qgraph` utiliza lasso en combinación con la selección de modelos EBIC para estimar un GGM regularizado.

³Para construir Intervalos de Confianza se necesita saber la *distribución muestral* del estadístico de interés. Sin embargo, saber dicha distribución para medidas de centralidad, es difícil por su misma complejidad de cómputo. Por lo que se utiliza bootstrap, que es una técnica que implica estimar repetidamente un modelo con datos muestreados o simulados para obtener el estadístico de interés. Se puede hacer bootstrap de dos maneras, paramétrico y no paramétrico. En el *no paramétrico*, las observaciones de los datos se re-muestran con reemplazo para crear un nuevo conjunto de datos, mientras que el *paramétrico* muestrea nuevas observaciones del modelo paramétrico que fue estimado a partir de los datos originales, lo que crea una serie de valores que pueden ser utilizados para estimar la distribución muestral.

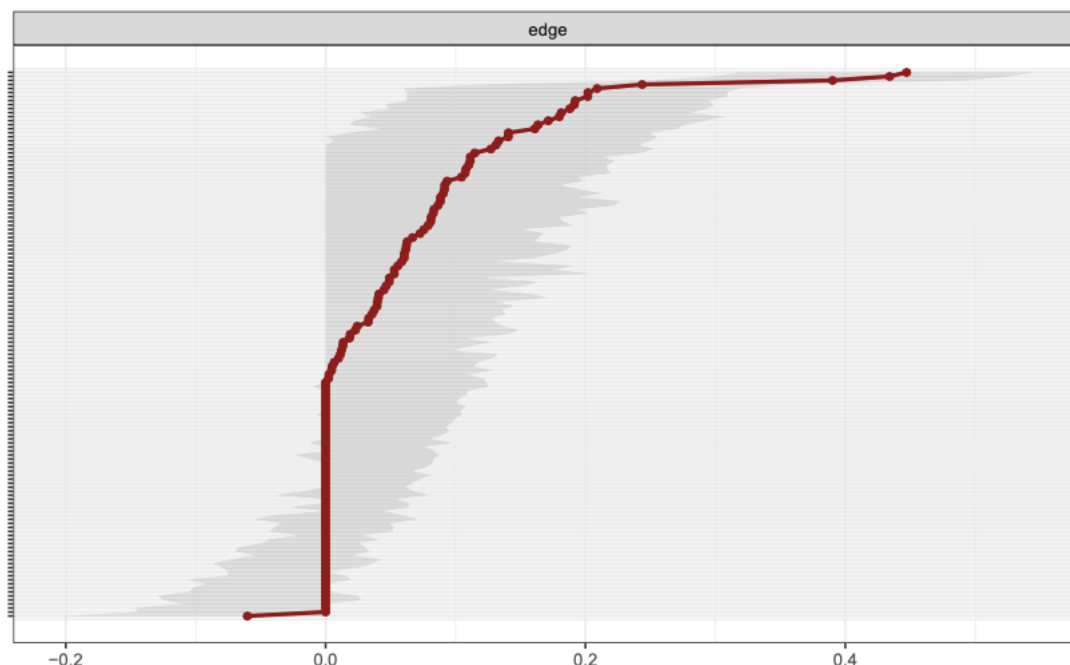


Figura 2: Intervalos de Confianza bootstrap de Epskamp et al. (2018).

1.2. Estabilidad de centralidad

Para saber la estabilidad de las medidas de centralidad no se puede utilizar bootstrap⁴, por lo que se propone estudiar dichas medidas con subconjuntos de datos. Se dice que los índices de centralidad son estables cuando el orden de los índices es el mismo en diferentes submuestras del mismo conjunto de datos. Lo que se hace es que se aplica la técnica de re-muestreo (el bootstrap regular) para diversas proporciones de los mismos participantes (o variables) y se evalúa si la correlación entre los índices de centralidad originales y aquellos obtenidos en las submuestras, son estables. Si la correlación cambia completamente después de quitar 10 % de la muestra, por ejemplo, entonces las interpretaciones de centralidad pueden ser erróneas. A esto se le llama bootstrap de subconjuntos con *case-dropping*. Para cuantificar dicha estabilidad se propone la medida de coeficiente de estabilidad de correlación (CS-coefficient), que se recomienda sea mayor a 0.7.

La Figura 3, muestra la estabilidad de las medidas de centralidad del modelo que se ha venido presentando. Éste debe de permanecer estable y tener promedio de correlación similar entre los diferentes porcentajes de muestra. Por lo tanto, se espera que las líneas se mantengan rectas y en la parte superior a lo largo de los porcentajes que están en el eje x. Para este caso, la medida de fuerza parece ser la más estable, mientras que *betweenness* y *closeness*, tienen mayor disminución conforme se reduce el porcentaje de muestra.

⁴Esto es porque se generan distribuciones muestrales sesgadas.

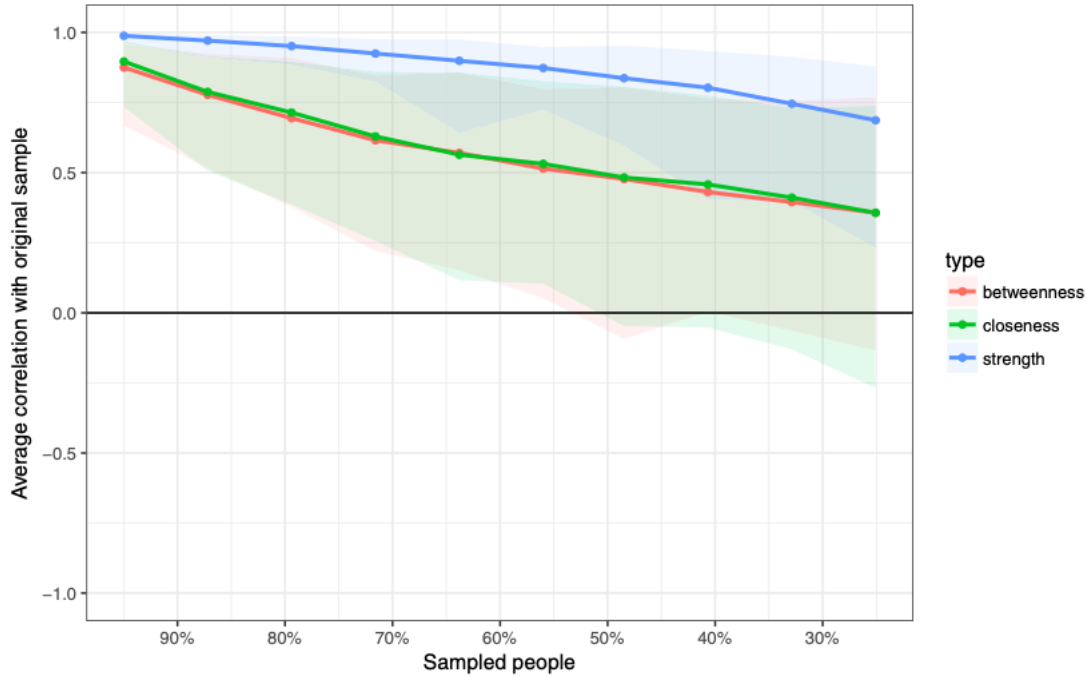


Figura 3: Estabilidad de las medidas de centralidad del ejemplo Epskamp et al. (2018).

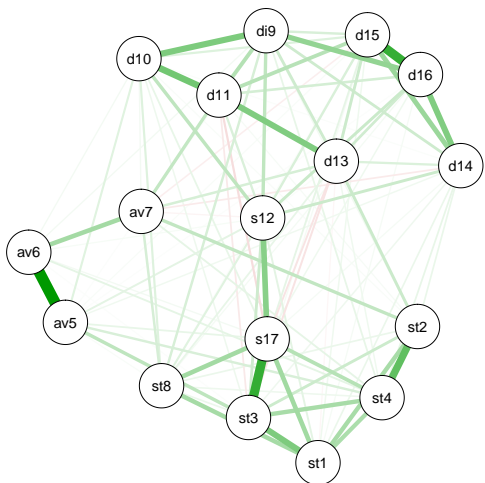
1.3. Test para diferencias significativas

El test de diferencias significativas, que también se hace con bootstrap, se utiliza para saber si una arista es significativamente distinta de otra, o si las medidas de centralidad son significativamente distintas de otras. En este test se recomienda, tener cuidado en la interpretación debido a que el no rechazar la hipótesis nula, no es evidencia de que la hipótesis alternativa sea verdadera⁵. El tipo de gráficos que se presentan para estos test son aquellos donde se presenta una matriz de cuadros, los que son negros son significativos entre los pares que se indican en los ejes x y y, y los grises no son significativos. En la siguiente sección presentaremos un gráfico como el que se menciona.

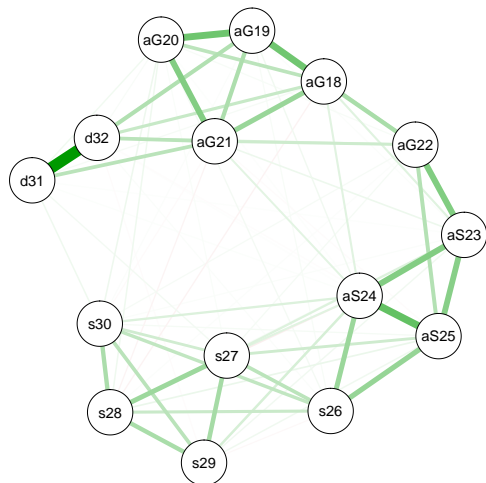
⁵Hipótesis nula: no hay diferencias.

2. Resultados preliminares

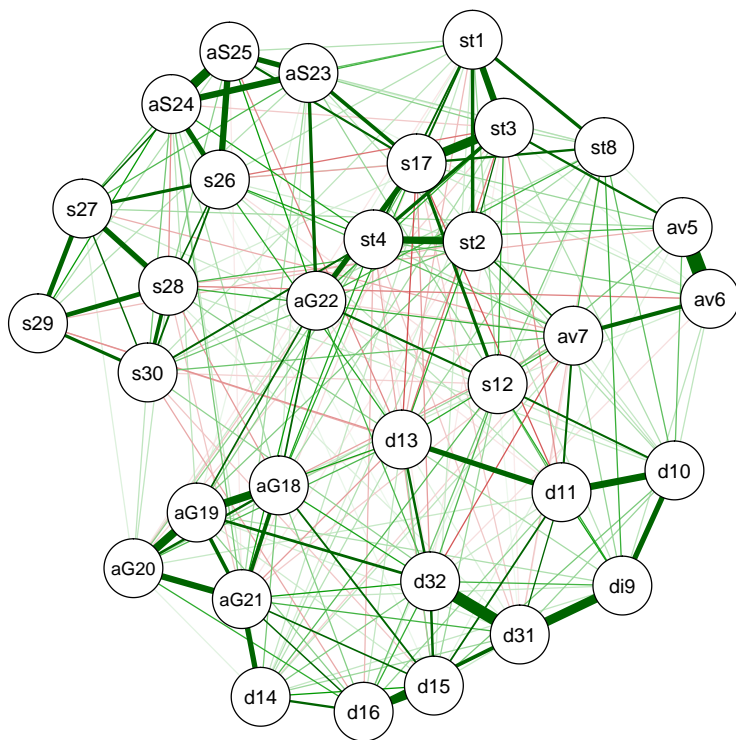
Los siguientes gráficos se presentan a partir de los datos obtenidos en el Tamizaje para la detección de riesgos a la Salud Mental por Covid-19. La base de datos actual, sólo utiliza las 32 preguntas relacionadas a los constructos de depresión, ansiedad y estrés. Se tienen un total de 102,209 participantes.



(a) Red PCL5.



(b) Red DEAN.



(c) Red de todos los reactivos.

Figura 4: Redes evaluadas.

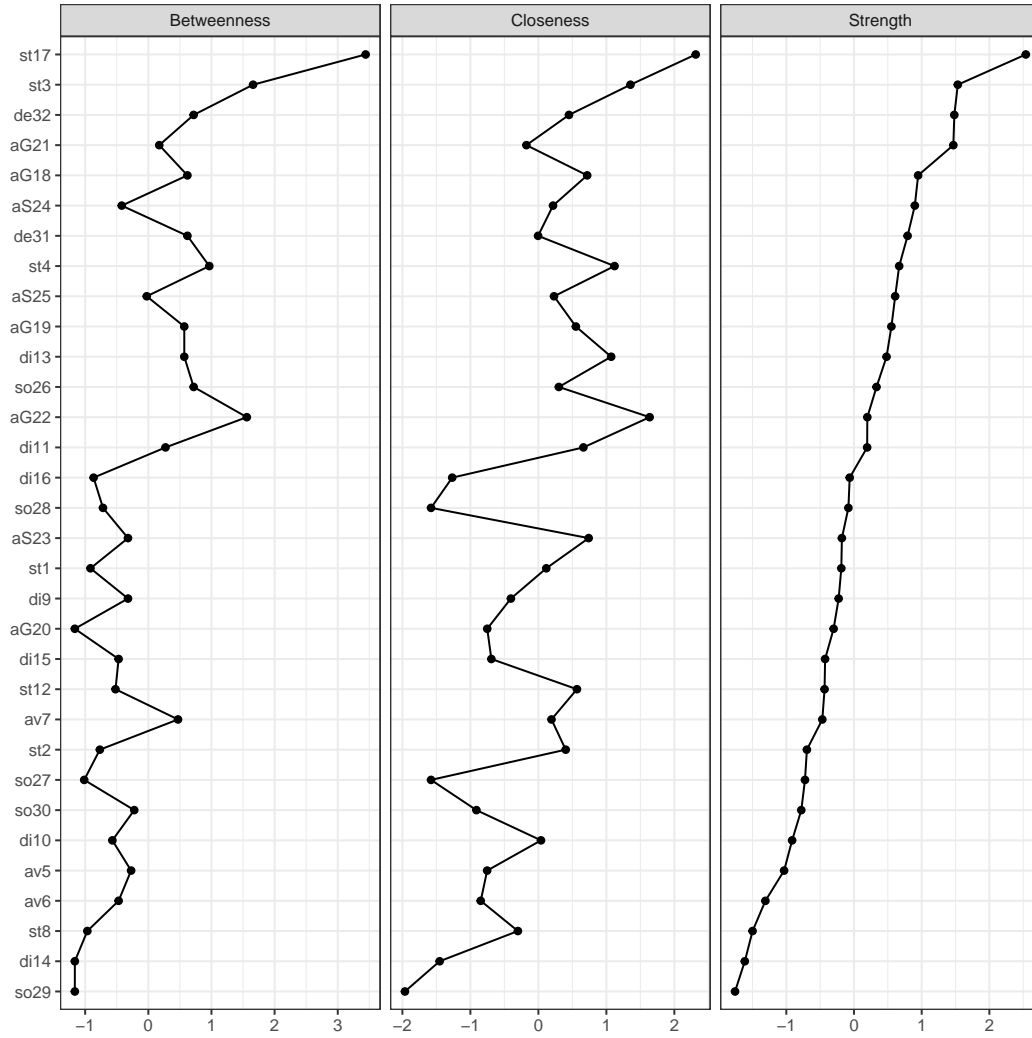


Figura 5: Medidas de centralidad del todo

Referencias

Epskamp, S., Borsboom, D., and Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods*, 50(1):195–212.

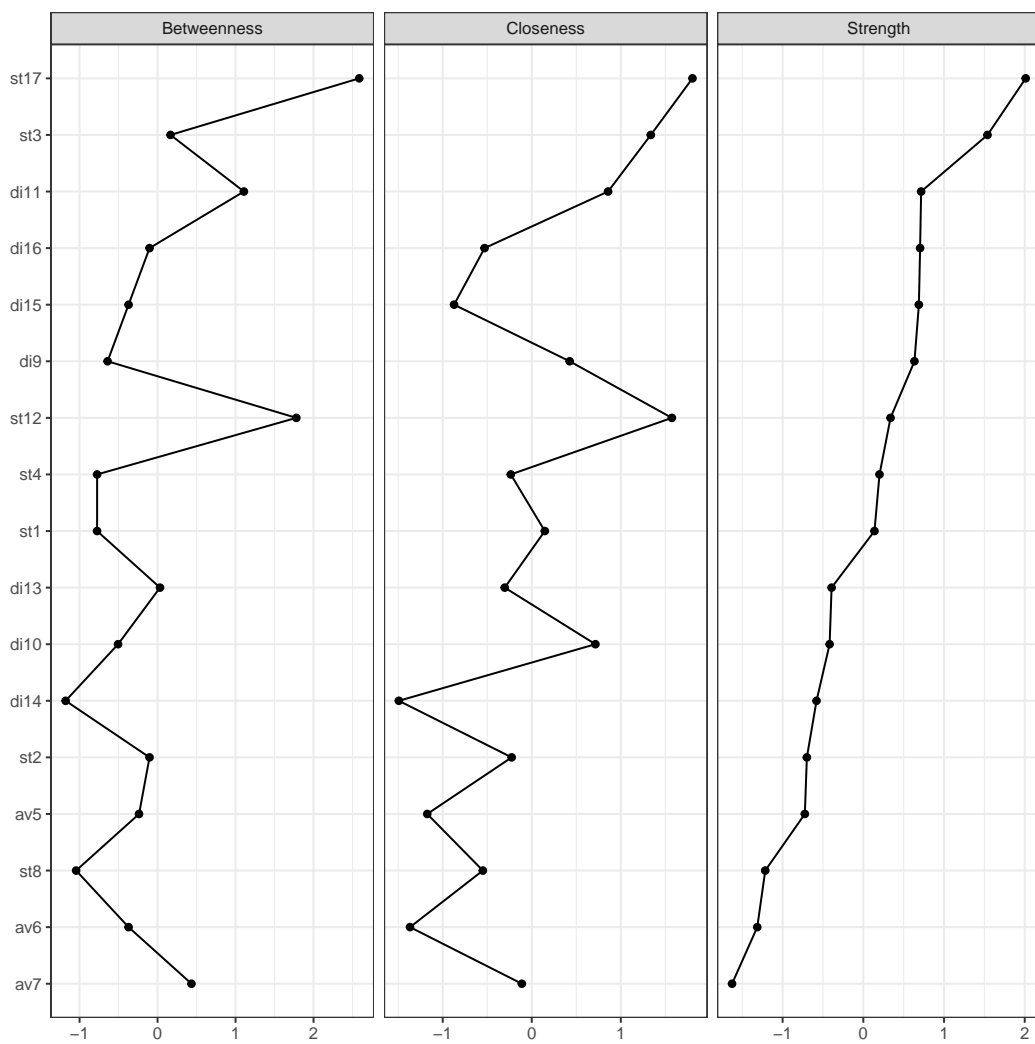


Figura 6: Medidas de centralidad del pcl5

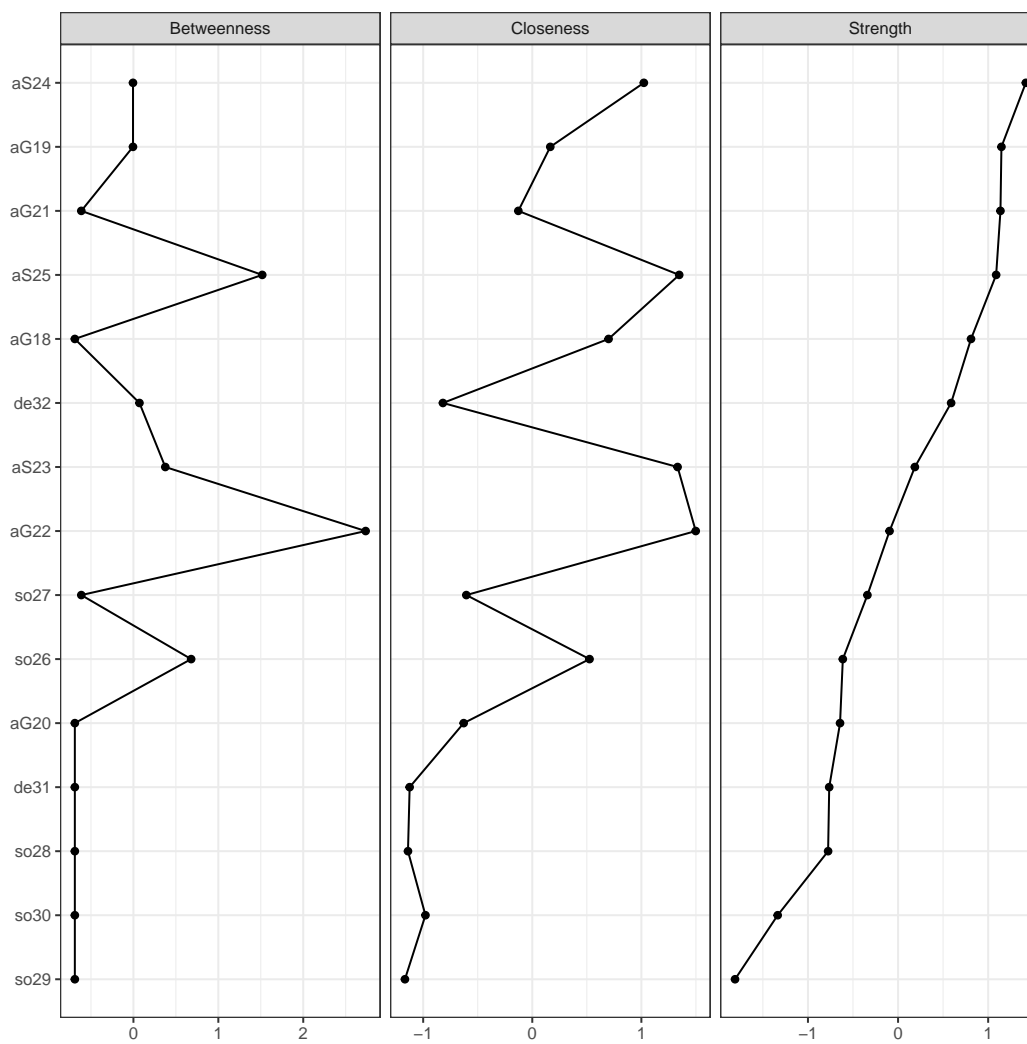


Figura 7: Medidas de centralidad del dean

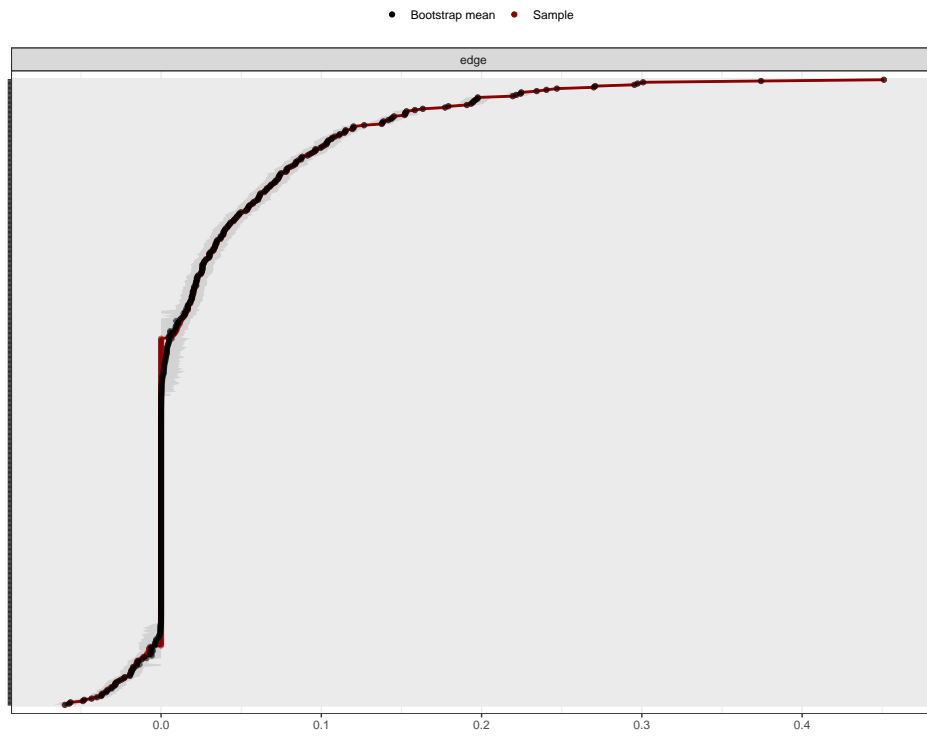


Figura 8: Interval red todo

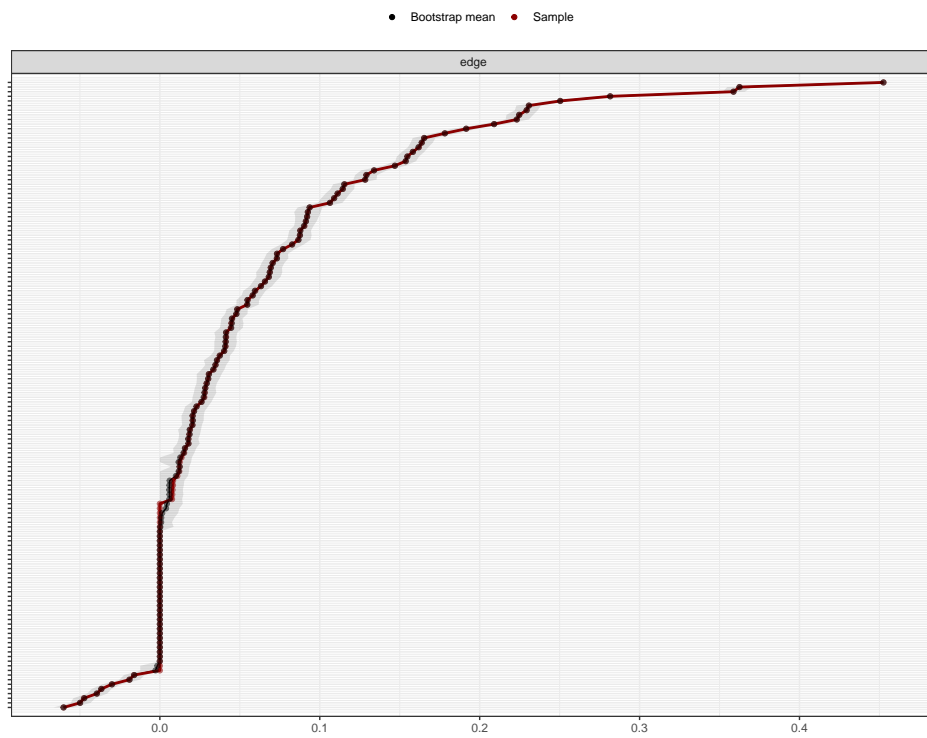


Figura 9: Interval red pcl5

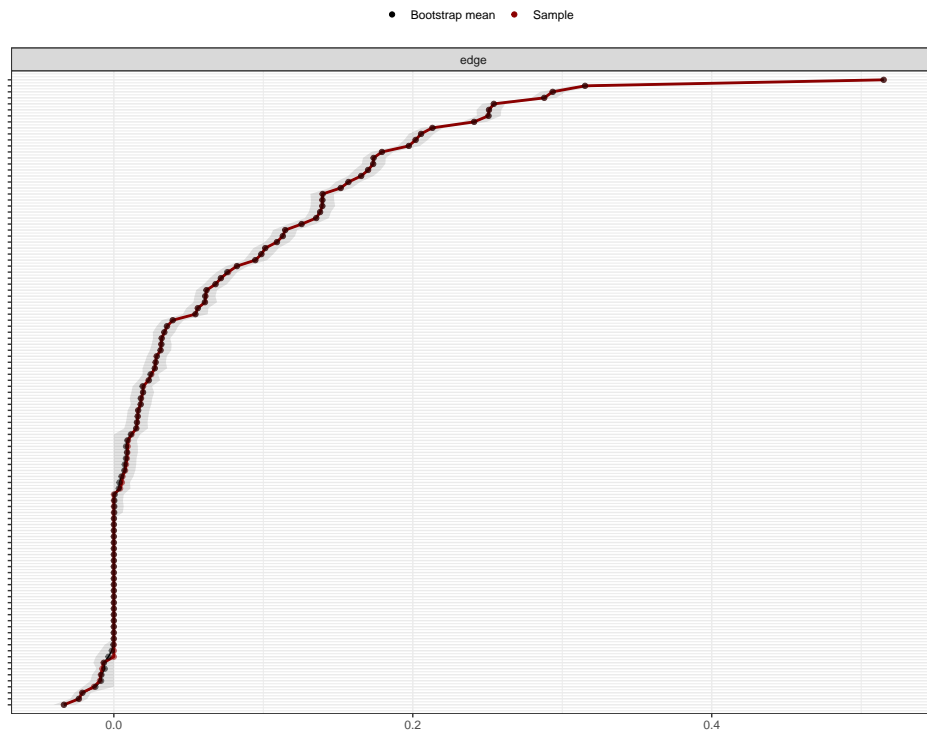


Figura 10: Interval red dean

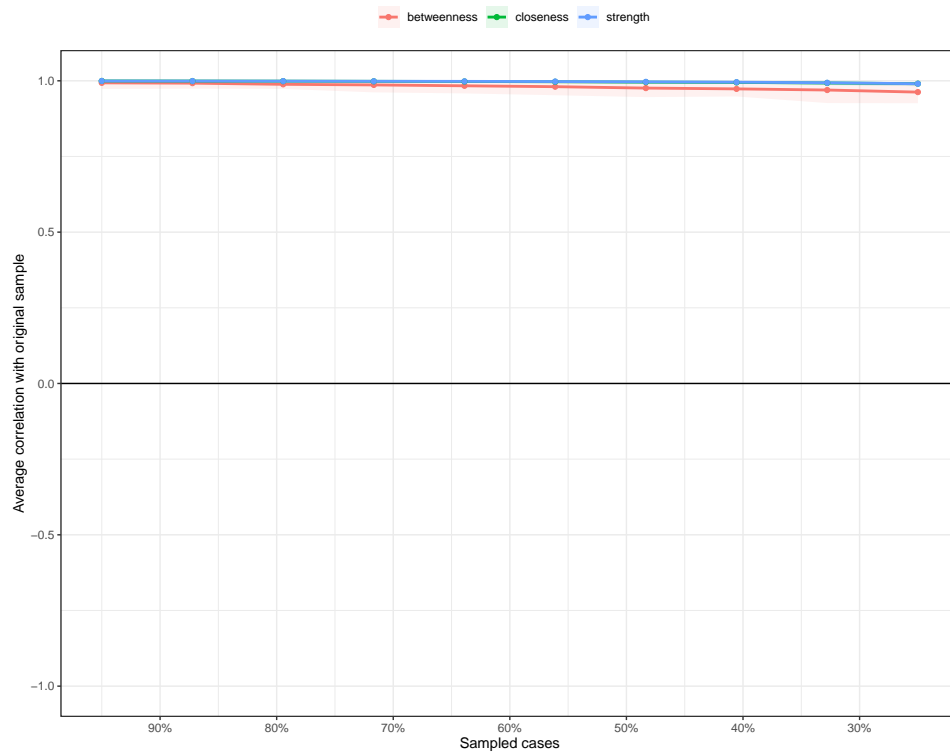


Figura 11: Estabilidad red todo

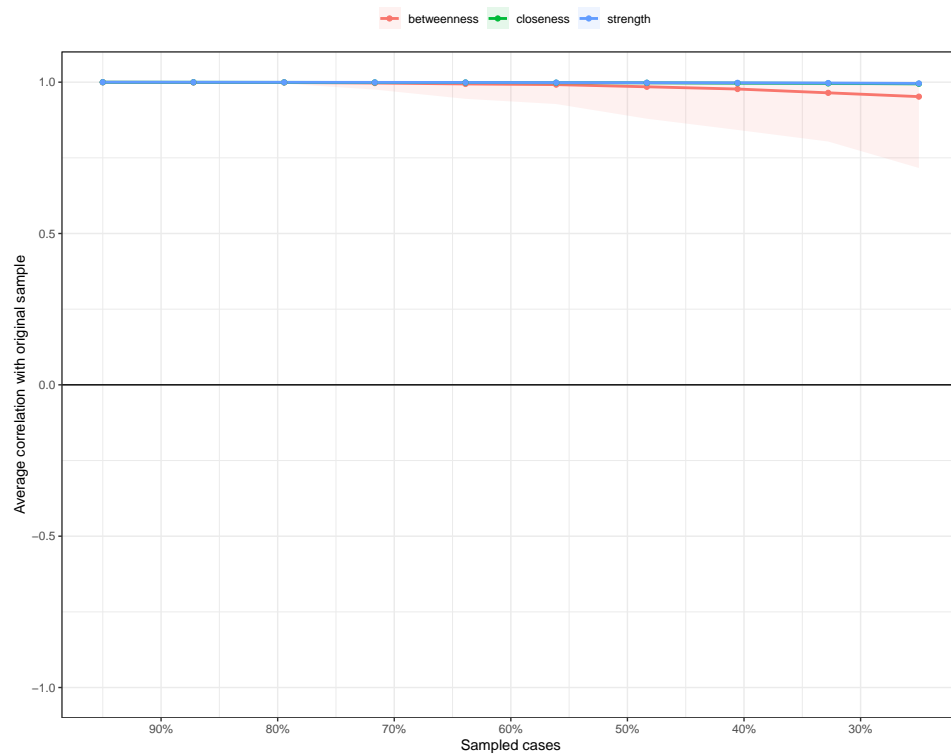


Figura 12: Estabilidad red dean

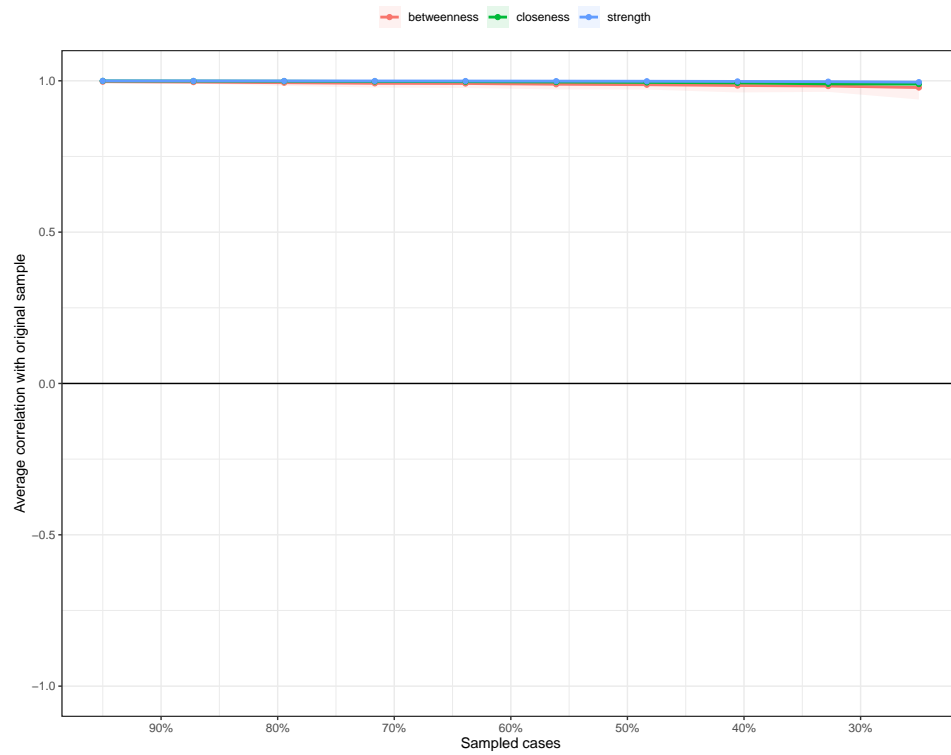


Figura 13: Estabilidad red pcl5

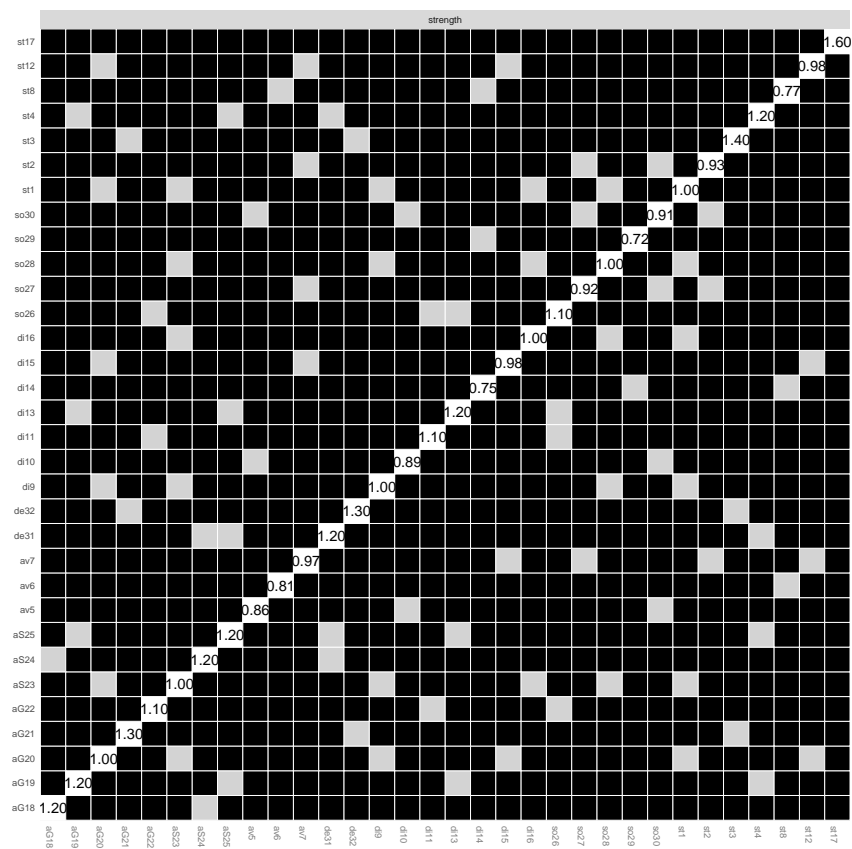


Figura 14: Significancia