

Grau en Estadística

Títol: Anàlisi diferencial de xarxes des del punt de vista estadístic

Autor: Núria Serra Pons

Director: Antonio Miñarro i Esteban Vegas

Departament: Genètica, Microbiologia i Estadística.
Facultat de Biologia. Universitat de Barcelona

Convocatòria: Juny 2024



UNIVERSITAT DE
BARCELONA



UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Facultat de Matemàtiques i Estadística

AGRAÏMENTS

M'agradaria agraïr la dedicació dels meus tutors d'aquest treball, Antonio Miñarro i Esteban Vegas. Gràcies pel vostre temps i per donar-me suport amb aquesta feina, proporcionant-me l'assistència necessària durant el procés d'elaboració i ajudant-me amb qualsevol dubte que pogués sorgir.

També vull agraïr el suport de la meva família, que sap com ha sigut de complex aquest camí. En especial a tu, Nico, pel teu suport constant i per estar sempre al meu costat.

Gràcies.

RESUM

L'anàlisi diferencial de xarxes és una eina per comparar xarxes gràfiques. S'ha aplicat en una àmplia gamma de camps i ha demostrat ser útil per identificar diferències en l'estructura i la funció de les xarxes construïdes sobre els mateixos nodes, és a dir, les mateixes variables, però en condicions diferents. Per exemple, es compararan xarxes d'associacions entre individus sans i aquells que pateixen una determinada patologia.

Per analitzar aquest tipus de xarxes compararem les matrius d'adjacència que determinen la xarxa, però, també hi ha altres mètodes per seleccionar les mètriques que serveixin per capturar les diferències més importants entre aquestes.

Per tant, el principal objectiu és utilitzar aquesta anàlisi diferencial de manera eficaç i realitzar una primera exploració per obtenir uns resultats preliminars.

PARAULES CLAU: Anàlisi diferencial de xarxes, Modelització gràfica, Estadística d'alta dimensió, Inferència de xarxes.

ABSTRACT

Title: Differential network analysis from a statistical perspective

Differential network analysis is a tool for comparing graphical networks. It has been applied in a wide range of fields and has proven useful for identifying differences in the structure and function of networks built on the same nodes, i.e., the same variables, but under different conditions. For example, it can compare networks of associations between healthy individuals and those suffering from a specific pathology.

To analyze this type of network, we will compare the adjacency matrices that define the network. However, there are also other methods to select the metrics that capture the most important differences between them.

Therefore, the main goal is to use this differential analysis effectively and perform an initial scan to obtain some preliminary results.

KEY WORDS: Differential network analysis, graphical modeling, high-dimensional statistics, network inference.

CLASSIFICACIÓ AMS

Classificació AMS (MSC2010):

05Cxx Graph theory

05C82 Small world graphs, complex networks

SUMARI

AGRAÏMENTS	2
RESUM	3
ABSTRACT	3
CLASSIFICACIÓ AMS	4
1. INTRODUCCIÓ	7
2. METODOLOGÍA	8
3. TEORIA DELS GRAFS	9
3.1 Introducció: Euler i el problema dels ponts de Königsberg	9
3.2 Definició bàsica	11
3.3 Propietats	11
3.4 Tipus de grafs	13
3.5 Representació	14
4. TIPUS D'ASSOCIACIONS PER LA CREACIÓ DE GRAFS	20
4.1 Associacions marginals	20
4.2 Associacions condicionals	21
4.3 Models gràfics gaussians	21
4.4 Càlcul directe de correlacions parcials	22
5. COMPARACIÓ DE XARXES	23
5.1 Mesures de diferència entre xarxes	23
5.2 Prova global de diferències de xarxa	23
5.3 Estadístics avaluats pel contrast d'hipòtesi de diferència de xarxes	24
6. APLICACIÓ AMB DADES REALS	26
6.1 Introducció	26
6.2 Comparació de Xarxes	30
6.3 Estadístics de prova	35
6.4 Discussió de resultats	39
7. EL PAQUET Differentialnetworkanalysis	40
7.1 Creació del paquet	40
7.2 Objectiu i contingut del paquet	40
8. CONCLUSIONS	41

BIBLIOGRAFIA	42
ANNEX	44

1. INTRODUCCIÓ

En els darrers anys, l'anàlisi de xarxes ha emergit com una disciplina transversal que abasta una àmplia gamma de camps, des de la biologia fins a la sociologia, passant per la informàtica i l'estadística. Les xarxes, que es poden conceptualitzar com a conjunts de nodes interconnectats per enllaços o connexions, ofereixen una representació poderosa i visual de l'estructura i les interaccions dins de sistemes complexos. Aquest enfocament permet entendre millor la dinàmica i la configuració de diferents sistemes, des de les relacions socials fins a les interaccions biològiques.

L'anàlisi diferencial de xarxes, una extensió de l'anàlisi de xarxes convencional, ha guanyat prominència com un enfocament innovador per a estudiar com canvien les propietats estructurals i dinàmiques de les xarxes en resposta a diferents condicions, intervencions o factors contextuais. Aquesta metodologia, que combina tècniques d'estadística, matemàtiques i computacionals, permet identificar canvis significatius en la topologia i la funció de les xarxes. Una característica atractiva d'aquest enfocament és la capacitat de crear una simplificació visual del patró observat d'associacions. En aquest context, els enfocaments ingenus basats en correlacions van ser i encara són àmpliament utilitzats.

Es proposa que mitjançant l'aplicació de mètodes estadístics avançats es poden identificar i quantificar diferències significatives entre diverses xarxes, oferint així una comprensió més profunda de les dinàmiques subjacents a les dades analitzades. A més, aquests mètodes poden ser aplicables a una varietat de camps, demostrant la seva versatilitat i utilitat. Els objectius principals d'aquest treball són: implementar i aplicar tècniques estadístiques per a l'anàlisi diferencial de xarxes, identificar i quantificar les diferències entre diverses xarxes, aplicar aquests mètodes a casos pràctics en biologia per demostrar la seva eficàcia.

La metodologia utilitzada en aquest estudi es divideix en dues parts: una part teòrica i una part pràctica. En la part teòrica, es va dur a terme una revisió de la literatura existent sobre anàlisi diferencial de xarxes. Això inclou l'estudi de diferents mètodes estadístics utilitzats en l'anàlisi de xarxes.

En la part pràctica, s'han recopilat unes dades biològiques, on a partir d'aquestes, s'han construït xarxes. I, finalment, s'ha procedit a la interpretació dels resultats, destacant les diferències identificades i discutint les seves implicacions en el context específic de les xarxes analitzades.

2. METODOLOGÍA

La metodologia emprada en aquest estudi s'ha dividit en dues parts principals: teòrica i pràctica.

En la part teòrica, s'ha dut a terme una revisió de la literatura actual sobre l'anàlisi diferencial de xarxes. Aquesta revisió ha inclòs l'estudi de diferents mètodes estadístics utilitzats en l'anàlisi de xarxes.

En la part pràctica, les dades biològiques utilitzades en aquest estudi procedeixen del conjunt de dades de diabetis dels Pima Indians, disponible a la web de Kaggle (*Paul Mooney, 2018*). Aquest conjunt de dades inclou variables clíniques com els embarassos, la glucosa en sang, la pressió arterial diastòlica, el gruix de la pell, el nivell d'insulina, l'índex de massa corporal (*BMI*), la funció Pedigree de la diabetis i l'edat. La base de dades consta de 768 observacions, amb 500 casos de no diabetis i 268 casos de diabetis. A partir de les dades recopilades, s'han construït xarxes on els nodes representen les característiques clíniques i les arestes representen les interaccions o correlacions entre aquestes característiques.

S'han utilitzat diverses tècniques estadístiques per a l'anàlisi de les xarxes. La correlació parcial s'ha aplicat per aïllar l'efecte directe d'una variable sobre una altra, controlant per l'impacte de totes les altres variables en el conjunt de dades. També s'han utilitzat tècniques de regularització com el lasso i el ridge regression per obtenir solucions més estables i interpretables, especialment en casos amb alta dimensionalitat. Per avaluar la significança de les diferències observades entre les xarxes de pacients amb diabetis i sense, s'han utilitzat tests de permutació.

Per comparar les dues xarxes, s'ha calculat la diferència entre els valors de les matrius de correlació parcial de cada grup per identificar discrepàncies directes en les correlacions entre les variables clíniques. Això ha permès analitzar les diferències en valors, signes i estructura de xarxes per identificar patrons específics de connectivitat que varien entre els dos grups estudiats.

Per a la implementació dels mètodes estadístics i la construcció de les xarxes, s'ha utilitzat el programa estadístic R versió 4.4.0. (*The R Project for Statistical Computing, 2024*). Aquest programari ha permès la manipulació de les dades, la realització dels anàlisis estadístics i la visualització de les xarxes construïdes.

3. TEORIA DELS GRAFS

3.1 Introducció: Euler i el problema dels ponts de Königsberg

La teoria de grafs es va originar amb el problema matemàtic dels set ponts de Königsberg. La ciutat russa, Kaliningrad, és una ciutat per la qual passa el riu Pregolya. Enmig del riu hi havia dues grans illes on estaven connectades entre elles, mitjançant una estructura de set ponts en total.

Els habitants de la ciutat es van preguntar si era possible dibuixar un camí que recorregués els set ponts de manera que només es passés un cop per cada pont.

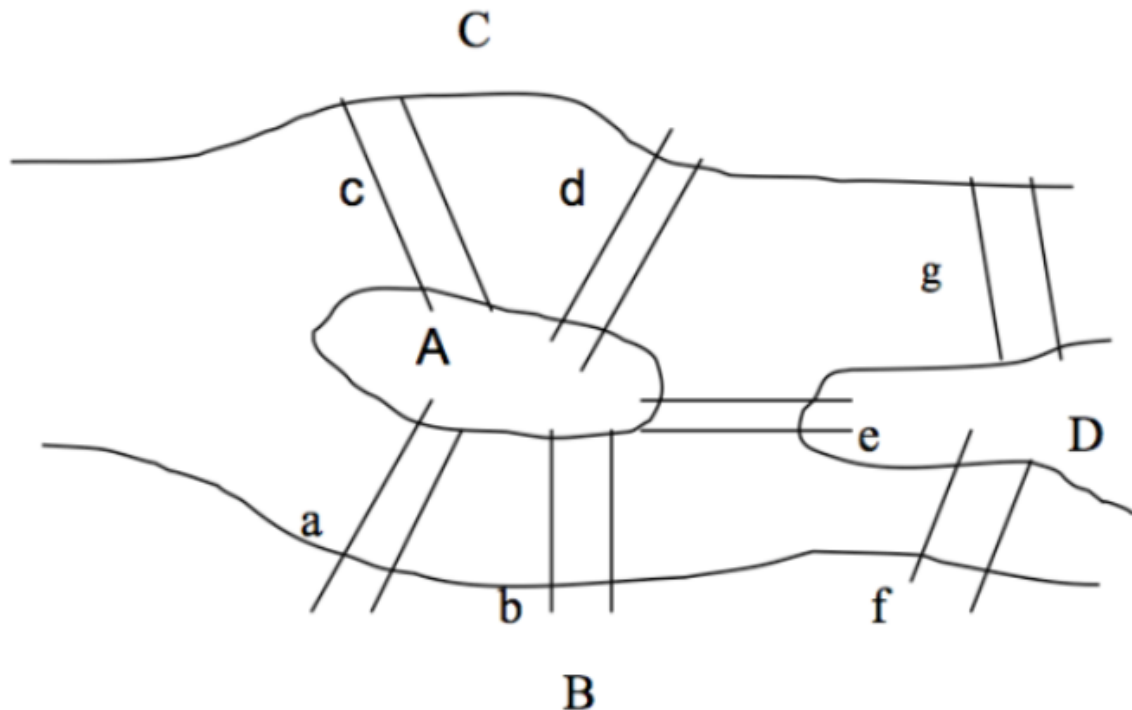


Figure 1: Camí dels set ponts

El matemàtic Leonhard Euler el 1736, va demostrar que no era possible. Euler va trobar la solució representant de forma “abstracta” els elements més importants del problema: va substituir les illes i riberes per punts (nodes), i els ponts per camins (arestes) que els enllaçaven. El resultat forma l'estructura que s'anomena avui en dia xarxa o graf.

Euler va veure que l'única forma de resoldre el problema era mitjançant grau dels vèrtexs. El grau d'un vèrtex és el nombre d'arestes que el connecten amb altres vèrtexs. En el problema dels set ponts, veiem que tres vèrtexs tenen grau 3, i un té grau 5. Euler va demostrar que el circuit que els ciutadans descrivien era possible si, i només si, hi havia dos vèrtexs de grau senar o si no n'hi havia cap. Un camí d'aquest tipus s'anomena camí eulerià. A més, si dos vèrtexs tenen grau senar, aquests han de ser el punt d'inici i final del camí eulerià.

Per tant, com que el graf corresponent als ponts de Königsberg té quatre vèrtexs de grau senar, no pot tenir un camí eulerià.

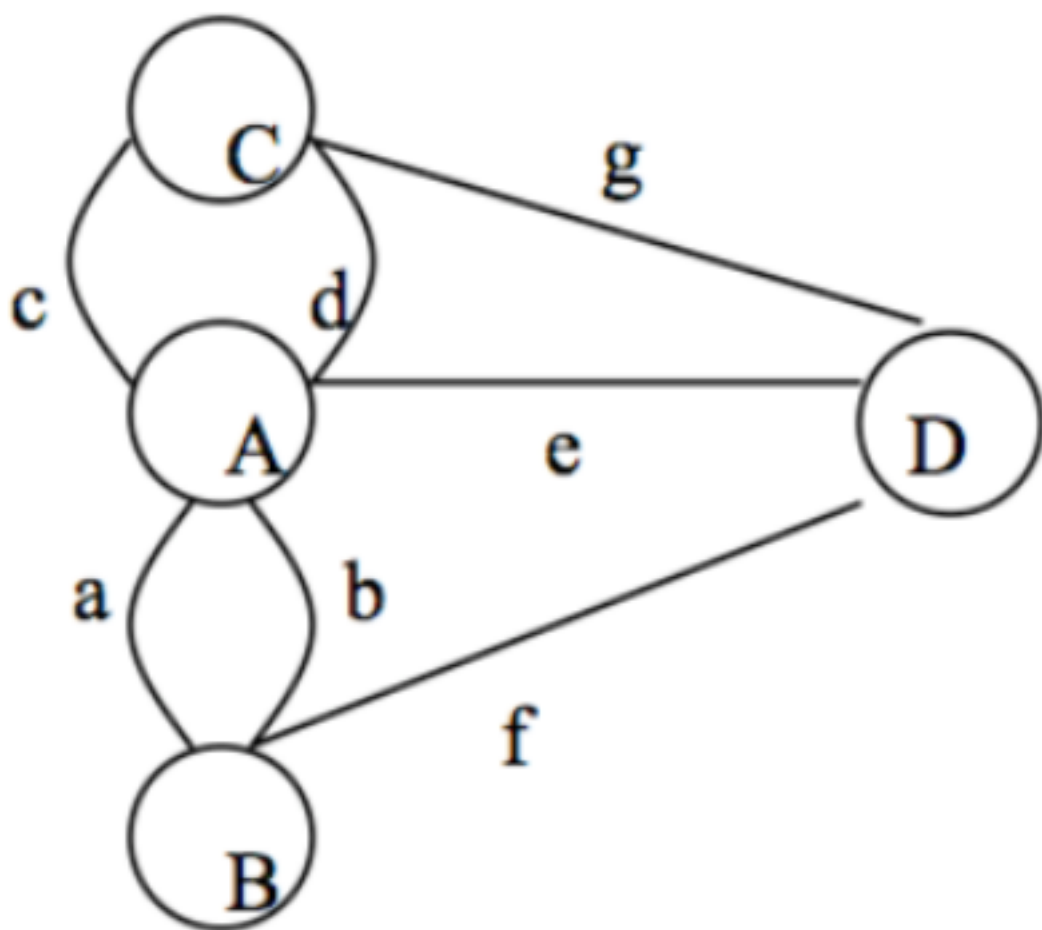


Figure 2: Graf dels ponts

Una altra forma de resoldre el mateix problema era fer un camí que passi per tots els ponts i també tingui el mateix punt d'inici i final. Un circuit així s'anomena un circuit eulerià. Existeix un circuit eulerià si, i només si, no hi ha vèrtexs de grau senar. Es pot veure que tots els circuits eulerians són també camins eulerians.

3.2 Definició bàsica

Per poder entendre la utilitat de la teoria de grafs primer hem d'entendre què és un graf. Primer donarem una definició general de graf i introduïrem la terminologia amb la qual hem d'estar familiaritzats abans d'endinsar-nos en la teoria de grafs. El concepte de graf es definirà dues formes diferents, geomètricament i algebraicament.

Segons la Real Academia Española un graf és un diagrama que representa mitjançant punts i línies les relacions entre parells d'elements i que s'usa per a resoldre problemes lògics, topològics i de càlcul combinatori.

En matemàtiques i ciències de la computació, un graf és un conjunt d'objectes anomenats vèrtexs o nodes connectats per arestes (també dites vores o enllaços), que permeten representar relacions binàries entre elements d'un conjunt. Per tant, diem que un graf es representa gràficament com un conjunt de punts (nodes) units per línies (arestes).

Des d'un punt de vista pràctic, els grafs permeten estudiar les interrelacions entre unitats que interactuen les unes amb les altres. Pràcticament qualsevol problema pot representar-se mitjançant un graf, i el seu estudi transcendeix a les diverses àrees com en matemàtiques discretes, ciències de la computació i teoria de xarxes s'utilitzen per a representar relacions entre diferents objectes.

Acabem de veure que un graf G és un conjunt de punts i de línies connectant algunes parelles de punts. Si volem formalitzar aquesta definició recorrerem a l'àlgebra. Així, algebraicament, un graf G es defineix com un parell ordenat:

$$G = (V, E) = (V(G), E(G))$$

on V és un conjunt no buit de punts de l'espai topològic, també anomenats vèrtexs, i E és un conjunt de parells no ordenats d'elements diferents a V , anomenats arestes. Que G sigui un parell ordenat significa que és una parella d'objectes matemàtics en la qual es distingeix un element de l'altre, i en el qual l'ordre d'aquests elements també forma part de la seva definició.

3.3 Propietats

Els grafs posseeixen una sèrie de propietats que poden ser útils per a comprendre la seva estructura i comportament. Seguidament, s'expliquen algunes de les propietats més importants.

Si dos vèrtexs i i j estan units per una mateixa aresta, diem que són adjacents. Representarem aquesta aresta com a (i, j) o com a (j, i) . Podem dir que els vèrtexs i i j són vèrtexs veïns, que són els extrems del mateix costat o que són incidents a l'aresta (i, j) . Dues arestes són adjacents quan tenen almenys un vèrtex comú.

Els grafs són molt utilitzats per representar xarxes de comunicació o transport; per tant, és important conèixer l'existència de camins que recorren tots els vèrtexs de forma més econòmica possible.

Un camí en un graf G és una seqüència en la qual s'alternen vèrtexs i arestes de G :

$$v_0 \rightarrow (v_0, v_1) \rightarrow v_1 \rightarrow (v_1, v_2) \rightarrow \cdots \rightarrow (v_{N-1}, v_N) \rightarrow v_N$$

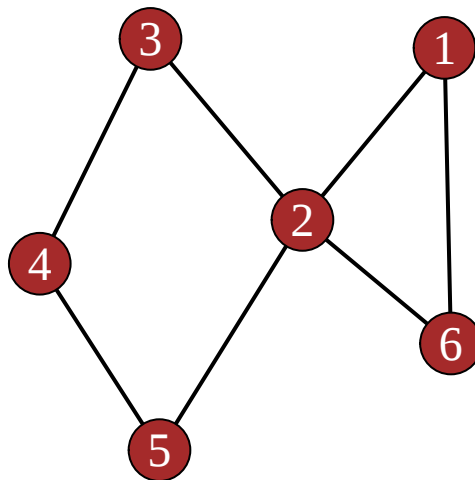
on cada costat té per extrems els vèrtexs anteriors i posteriors a la seqüència.

Per tant, pot ser representada d'aquesta manera sense perdre informació:

$$v_0 \rightarrow v_1 \rightarrow v_2 \rightarrow \cdots \rightarrow v_N$$

Els vèrtexs v_0 i v_N són els extrems del camí. La longitud d'un camí és el nombre d'arestes que conté. Una propietat del camí és que dos costats consecutius han de ser per força adjacents o bé idèntics, si estem retrocedint pel mateix costat. Per saber més sobre els camins ens fixarem en el següent graf:

```
plot(g, vertex.size = 30, vertex.label.cex = 1.5, vertex.label.color = "white",  
     vertex.color = "brown", edge.color = "black", edge.width = 2)
```



Diem que un camí és tancat quan comença i acaba al mateix vèrtex. Per exemple, el camí $4 \rightarrow 3 \rightarrow 2 \rightarrow 5 \rightarrow 4$. Si no té aquesta propietat, el camí és obert.

Un circuit és un camí tancat en què tots els costats (tot i que no necessàriament tots els vèrtexs) són diferents. Per exemple, el camí $2 \rightarrow 5 \rightarrow 4 \rightarrow 3 \rightarrow 2 \rightarrow 1 \rightarrow 6 \rightarrow 2$.

Un cicle és un camí tancat en el qual tots els vèrtexs (excepte l'inicial i el final) són diferents. Per tant, els costats també han de ser diferents. Per exemple, el camí $1 \rightarrow 2 \rightarrow 6 \rightarrow 1$.

La longitud de la via més curta entre un punt i i un punt j en un graf G es coneix com la distància entre els dos vèrtexs i es representa per $d(i, j)$. Aquesta distància és una quantitat positiva dins els nombres enters i té les següents propietats:

$$d(i, j) = 0 \iff i = j$$

És a dir, la distància entre els vèrtexs i i j és igual a zero només quan el vèrtex i i el vèrtex j són el mateix vèrtex.

$$d(i, j) = d(j, i), \quad i, j \in V(G)$$

És a dir, la distància entre els vèrtexs i i j és igual a la distància entre j i i per a tots els valors de i i de j , ja que només és la mateixa aresta llegida a la inversa.

$$d(i, k) \leq d(i, j) + d(j, k), \quad i, j, k \in V(G)$$

És a dir, la distància entre i i k sempre és més petita (o igual si els punts formen una línia recta) que la suma de les distàncies entre i i j i entre j i k per a tots els valors de i , j i k .

$$d(i, j) = 1 \iff (i, j) \in E(G)$$

És a dir, la distància entre el vèrtex i i el vèrtex j és igual a 1 només quan formen una aresta del graf, ja que la distància més curta que els uneix és aquesta aresta.

3.4 Tipus de grafs

Els grafs poden classificar-se en diferents tipus segons unes certes característiques estructurals o propietats. Alguns dels tipus més comuns de grafs són:

3.4.1 Graf simple i multigraf La definició de graf que hem donat fins ara correspon al que anomenem graf simple. Un graf simple és aquell en el qual no hi ha arestes múltiples entre els mateixos parells de nodes i no hi ha bucles, és a dir, és genera quan un conjunt no buit de vèrtex està unit a través d'una o més arestes. Com s'ha dit, els elements d' E són parells no ordenats d'elements diferents a V . Per tant, al no permetre que un element aparegui repetit en un mateix parell, no es permet l'existència de loops. Els loops són arestes que comencen i acaben al mateix vèrtex.

Diem que E és un conjunt d'elements i no una família. Per aquest motiu, entenem per conjunt una col·lecció d'elements diferents; una família, en canvi, és una col·lecció d'elements que es poden repetir. Així, $(1, 2, 3, 4, 5, 6)$ és un conjunt i una família d'elements, mentre que $(1, 1, 2, 3, 4, 4, 5, 6)$ és només una família.

Tenint en compte això, diem que un graf no simple, multigraf, és una composició lleugerament més complexa en la qual un conjunt de nodes es connecta amb altres conjunts de nodes a través d'arestes múltiples o paral·leles. Per tant, en la definició formal de multigraf es permet que $E(G)$ sigui una família i no només un conjunt, o sigui que es permet que alguns elements es repeteixin.

3.4.2 Graf dirigit o dígraf Després de veure el concepte de la distància introduïrem el de direcció. En un graf dirigit, les arestes tenen una direcció específica, cosa que significa que la connexió entre dos nodes té un sentit definit.

D'altra banda, en un graf no dirigit, les arestes no tenen direcció, la qual cosa implica que la connexió entre dos nodes és bidireccional. En altres paraules, en un graf dirigit, es pot anar d'un node a un altre seguint la direcció de les arestes, mentre que en un graf no dirigit, la relació entre dos nodes és simètrica i no té un sentit definit.

3.4.3 Graf ponderat i no ponderat Un graf ponderat o etiquetat és aquell en el qual a cada aresta se li assigna un pes (nombre enter generalment) que representa alguna mesura de distància, cost, capacitat, etc. En canvi, el no ponderat és aquell en el qual totes les arestes tenen el mateix pes o no tenen pes assignat.

3.4.4 Graf connex i inconnex Un graf connex és aquell en el qual hi ha un camí entre qualsevol parell de nodes, és a dir, no hi ha nodes aïllats. Per contra, un graf no connex o inconnex té almenys dos nodes que no estan connectats per cap camí.

De forma algebraica, G és connex si $u, v : \exists \mu \leq \langle u, v \rangle$. Sent μ un camí entre els nodes v i u .

3.4.5 Graf complet i incomplet Un graf és complet si existeixen arestes unint totes les possibles parelles de vèrtexs. És a dir, tota parella de vèrtexs v, u ha de tenir una aresta (v, u) que els uneixi. El conjunt dels grafs complets és anomenat usualment K , sent K_n el graf complet de n vèrtexs.

Un K_n té exactament $\frac{n(n-1)}{2}$ arestes.

3.5 Representació

Existeixen diferents maneres de representar un graf simple, a més de la geomètrica i molts mètodes per a emmagatzemar-los en una computadora. L'estructura de dades usada depèn de les característiques del graf i l'algorisme usat per a manipular-lo. Entre les estructures més senzilles i usades es troben les matrius. Aquestes matrius proveeixen accés ràpid, però poden consumir grans quantitats de memòria.

3.5.1 Matriu d'adjacència És una matriu booleana que representa les connexions entre parells de vèrtexs. La matriu d'adjacència d'un graf és simètrica. Si un vèrtex és aïllat llavors la corresponent fila (columna) aquesta composta només per zeros. Si el graf és simple llavors la matriu d'adjacència conté sols zeros i uns (matriu binària) i la diagonal aquesta composta només per zeros.

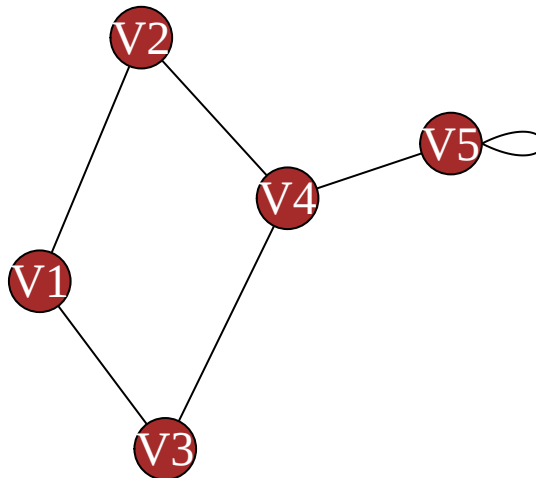
La matriu d'adjacència d'un dígraf no és simètrica. És una matriu binària. El número d'uns que apareixen en una fila és igual al grau de sortida del corresponent vèrtex i el número d'uns que apareixen en una determinada columna és igual al grau d'entrada del corresponent vèrtex.

Seguidament, s'explica l'algorisme per a construir la matriu d'adjacència d'un graf:

- Pas 1: Se li assignen un ordre arbitrari als vèrtexs.
- Pas 2: Es construeix una matriu de dimensió $n \times n$, cardinalitat (# vèrtexs) per cardinalitat (# vèrtexs).
- Pas 3: En la posició $j - \text{ésima}$ es col·loca 1 si el vèrtex i és adjacent al vèrtex j ; 0 en l'altre cas.

Veiem un exemple, donat el següent graf, es construeix la matriu d'adjacència (MA)

```
plot(g, vertex.label = c("V1", "V2", "V3", "V4", "V5"),  
     vertex.size = 30, vertex.label.cex = 1.5, vertex.label.color = "white",  
     vertex.color = "brown", edge.color = "black")
```



```
print(adj_matrix)
```

```
##      V1 V2 V3 V4 V5
## V1  0  1  1  0  0
## V2  1  0  0  1  0
## V3  1  0  0  1  0
## V4  0  1  1  0  1
## V5  0  0  0  1  1
```

S'obté la valència de cadascun dels vèrtexs si se suma la fila o columna corresponent a aquest vèrtex. Diem que el grau o valència d'un graf és el nombre d'arestes que incideixen en un vèrtex. En els grafs no dirigits, és el nombre d'arestes connectades a un vèrtex. En grafs dirigits, es distingeix entre grau entrant (nombre d'arestes que entren en un vèrtex) i grau sortint (nombre d'arestes que surten d'un vèrtex).

```
adj_matrix_extended
```

```
##      V1 V2 V3 V4 V5
## V1  0  1  1  0  0 2
## V2  1  0  0  1  0 2
## V3  1  0  0  1  0 2
## V4  0  1  1  0  1 3
## V5  0  0  0  1  1 2
##      2  2  2  3  2 0
```

Amb la matriu d'adjacència no es poden representar els costats paral·lels. En el cas que utilitzem un graf no dirigit, en la matriu d'adjacència la diagonal principal conté zeros. La representació per mitjà de matriu es prefereix per a algorismes on el nombre d'arcs és gran en proporció al número de vèrtexs. Si succeís el contrari es prefereix un altre mètode anomenat llistes d'adjacència.

```
print("MA^2:")
```

```
## [1] "MA^2:"
```

```
print(MA_squared)
```

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    2    0    0    2    0
## [2,]    0    2    2    0    1
## [3,]    0    2    2    0    1
## [4,]    2    0    0    3    1
## [5,]    0    1    1    1    2
```

En la diagonal de la matriu quadrada s'obté el grau o la valència dels vèrtexs, i els altres elements d'aquesta matriu quadrada indiquen el nombre de costats de longitud 2.

$$MA^2[1, 4] = 2 \quad (v_1, v_2) (v_2, v_4) \quad i \quad (v_1, v_3) (v_3, v_4)$$

$$MA^2[2, 3] = 2 \quad (v_2, v_1) (v_1, v_3) \quad i \quad (v_2, v_4) (v_4, v_3)$$

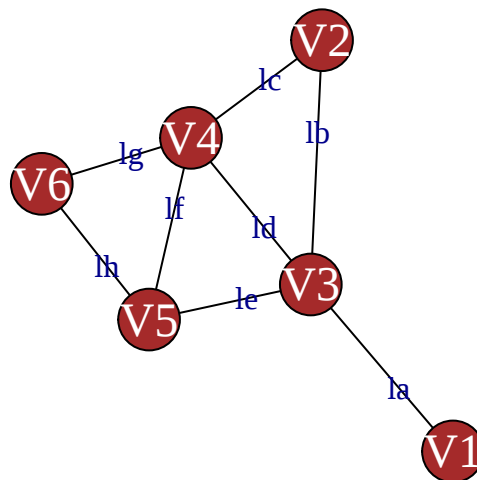
$$MA^2[2, 5] = 1 \quad (v_2, v_4) (v_4, v_5)$$

Aquest mètode té una sèrie de desavantatges, per exemple és poc eficient si el nombre de vèrtexs varia al llarg del temps de vida del graf al igual que si el graf té pocs arcs, és a dir, si la matriu és dispersa. També, pot donar-se el cas que el número de vèrtexs sigui major del número inicial.

La matriu d'adjacència d'un graf ponderat es pot utilitzar per emmagatzemar els pesos de les arestes. Si a una aresta li falta un valor especial, potser un valor negatiu, zero o un valor gran per a representar l'infinit, indica aquest fet.

3.5.2 Matriu d'incidència La matriu d'incidència només conté zeros i uns, per tant, diríem que es tracta d'una matriu binària. Com cada aresta incideix exactament en dos vèrtexs, cada columna té exactament dos uns. El número d'uns que apareix en cada fila és igual al grau del vèrtex corresponent. Diem que una fila composta només per zeros correspon a un vèrtex aïllat. Ho veiem a la següent representació:

```
plot(g, vertex.label = V(g)$name, edge.label = E(g)$name, vertex.size = 30,
     vertex.label.cex = 1.5, vertex.label.color = "white", vertex.color = "brown",
     edge.color = "black")
```



```
print(incidence_matrix)
```

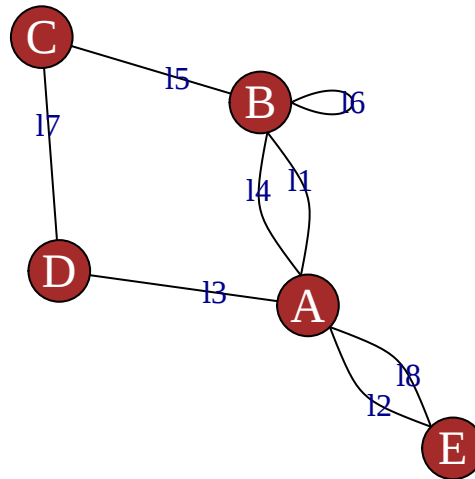
```
##      la lb lc ld le lf lg lh
## V1   1  0  0  0  0  0  0  0
## V2   1  1  1  0  0  0  0  0
## V3   0  1  0  1  1  0  0  1
## V4   0  0  1  1  0  1  1  0
## V5   0  0  0  0  1  1  0  0
## V6   0  0  0  0  0  0  1  1
```

La matriu d'adjacència d'un dígraf no és simètrica, sinó que es tracta d'una matriu binària. El número d'uns que apareixen en una fila és igual al grau de sortida del corresponent vèrtex i el número d'uns que apareixen en una determinada columna és igual al grau d'entrada del corresponent vèrtex.

Seguidament, s'explica l'algorisme per a construir la matriu d'incidència d'un graf:

- Pas 1: Se li assigna un ordre arbitrari a cadascun dels vèrtexs i costats del graf.
- Pas 2: Es construeix la matriu d'incidència de dimensió $n_{vertex} \times n_{arestes}$, cardinalitat (nombre de vèrtexs) per cardinalitat (nombre de costats o arestes).
- Pas 3: S'assigna un 1 a l'element i, j si el costat j és incident al vèrtex i , 0 en un altre cas.

```
plot(g, vertex.label = V(g)$name, edge.label = E(g)$name, vertex.size = 30,
     vertex.label.cex = 1.5, vertex.label.color = "white", vertex.color = "brown",
     edge.color = "black")
```



```
print(incidence_matrix)
```

```
##    l1 l2 l3 l4 l5 l6 l7 l8
## A   1  1  1  0  0  0  0  0
## B   0  0  1  1  1  0  0  1
## C   0  1  0  0  1  1  1  0
## D   0  0  0  0  1  1  0  1
## E   1  0  0  1  0  0  0  0
```

Es poden observar una sèrie de incisos, com que la suma de la fila en una matriu d'incidència representa el nombre de costats incidents en aquest vèrtex. Per tant, si la suma de la columna és un en la matriu d'incidència llavors aquest costat s'anomena llaç. Aquesta matriu ens permet representar grafs amb costats paral·lels i llaços simultàniament.

4. TIPUS D'ASSOCIACIONS PER LA CREACIÓ DE GRAFS

Les interaccions en els sistemes biològics i els canvis en aquestes interaccions es dedueixen comunament de les mesures en els nodes. Aquesta disciplina busca detectar variacions en el conjunt d'arestes, V , inferits a partir de n observacions en cada node $j \in \mathbb{E}$

Per fer-ho ens centrem en l'anàlisi diferencial de xarxes no dirigides. Els mètodes per a l'aprenentatge de l'estructura de xarxes no dirigides es poden classificar en dos grans grups: basats en associacions marginals i en associacions condicionals

4.1 Associacions marginals

Els mètodes d'inferència basats en associacions marginals detecten una aresta no dirigida entre dues variables X_j i X_k si i només si són dependents entre elles, és a dir, les associacions marginals es refereixen a la dependència entre dues variables sense considerar la influència d'altres variables en el sistema. Aquest enfocament és útil perquè proporciona una primera aproximació a les connexions directes entre parells de nodes en una xarxa.

Per identificar aquestes dependències, s'utilitzen diverses mesures d'associació. A continuació, es descriuen algunes de les més comunes:

- La Correlació de Pearson mesura la força i la direcció de la relació lineal entre dues variables contínues. Es calcula com el coeficient de correlació ρ_{jk} entre les variables X_j i X_k , definit com:

$$\rho_{jk} = \frac{\text{Cov}(X_j, X_k)}{\sigma_{X_j} \sigma_{X_k}}$$

on $\text{Cov}(X_j, X_k)$ és la covariància i σ_{X_j} i σ_{X_k} són les desviacions estàndard de X_j i X_k

- Les Correlacions basades en el rang són mesures com la correlació de Spearman o la correlació de Kendall, s'utilitzen quan les dades no són normalment distribuïdes o contenen valors extrems. Es basa en els rangs de les dades en lloc dels seus valors originals. Per exemple, la correlació de Spearman es calcula com la correlació de Pearson dels rangs de les dades
- La informació mútua mesura la dependència entre dues variables discretes o contínues, capturant tant relacions lineals com no lineals. Es defineix com:

$$I(X_j; X_k) = \sum_{x_j} \sum_{x_k} p(x_j, x_k) \log \left(\frac{p(x_j, x_k)}{p(x_j)p(x_k)} \right)$$

on $p(x_j; x_k)$ és la distribució conjunta de X_j i X_k , i $p(x_j)$ i $p(x_k)$ són les seves distribucions marginals.

- Les Mesures de dependència basades en el nucli on utilitzen mètodes no paramètrics per capturar dependències complexes entre variables. Un exemple és la Mesura de Dependència basada en el Nucli (Kernel Dependency Measure, KDM). S'utilitza un nucli (kernel) per projectar les dades en un espai de característiques d'alta dimensionalitat, on les dependències poden ser més fàcilment detectades.

En aquest context, dos nodes j i k estan connectats a la xarxa G (és a dir, $j - k \in E$) si la mesura d'associació marginal $(X_j, X_k) > 0$. Per provar la independència de cada parell de variables X_j i X_k , es poden fer servir tests estadístics que generen valors p . Un llindar per a aquests valors p actua com a paràmetre d'ajust per controlar l'esparsitat de la xarxa resultant.

4.2 Associacions condicionals

Les associacions marginals només mesuren les relacions entre parells de variables, sense diferenciar entre efectes directes i indirectes. Això significa que dues variables poden estar correlacionades indirectament a través d'una tercera variable.

Per abordar aquesta limitació, les associacions condicionals es centren en la dependència condicional. En un graf d'independència condicional ($C|G$), hi ha una aresta entre j i k , és a dir, X_j i $X_k \in E$ si i només si X_j i X_k són condicionalment dependents donades totes les altres variables.

En el cas de variables multivariades Gaussians, la independència condicional equival a què la correlació parcial entre les variables sigui zero. Així, per inferir la presència d'una aresta entre j i k en la xarxa, es comprova la condició de dependència condicional entre X_j i X_k mentre es controlen les altres variables del sistema.

4.3 Models gràfics gaussians

Els procediments existents per a l'anàlisi de xarxes diferenciades es basen principalment en models gràfics paramètrics. Els Models Gràfics Gaussians (GGM) assumeixen que les variables segueixen una distribució normal multivariada.

Per a dades normals multivariades $X = (X_1, \dots, X_p) \sim \mathcal{N}(\mu, \Sigma)$, la correlació parcial $\rho(x_i, x_j | \text{rest})$ entre X_j i X_k donat que totes les variables restants estan relacionades amb la matriu de precisió $\Omega = (w_{ij}) = \Sigma^{-1}$ on $\rho(x_i, x_j | \text{rest}) = -\frac{w_{ij}}{\sqrt{w_{ii}w_{jj}}}$ on:

- w_{ij} és l'element de la matriu de precisió Ω corresponent a la relació entre les variables X_i i X_j
- w_{ii} i w_{jj} són els elements diagonals de la matriu de precisió Ω corresponents a les variàncies de X_i i X_j , respectivament.
- *rest* es refereix a la resta de les variables del conjunt de dades que s'estan considerant per calcular aquesta correlació parcial.

Per tant, diem que aquesta correlació parcial mesura la força i la direcció de la relació lineal entre X_i i X_j després de controlar per l'efecte de les altres variables en el conjunt de dades (les "restants" variables).

Una entrada de $w_{ij} = 0$ indica independència condicional, corresponent a una correlació parcial nul·la. Una entrada no nul·la $w_{ij} \neq 0$ correspon a una correlació parcial no nul·la. Aquestes correlacions parcials són essencials per comprendre les relacions directes entre parelles de variables dins del conjunt de dades, eliminant l'efecte de la resta de variables.

Una estratègia consisteix a calcular la matriu de covariància empírica, S , i estimar les entrades no nul·les de S^{-1} aplicant un llinard o utilitzant un procediment d'inferència. Aquest procés permet identificar les connexions significatives dins de la xarxa, proporcionant una visió més precisa de l'estructura de dependència subjacent.

4.4 Càlcul directe de correlacions parcials

La inversa de la matriu empírica de covariància S^{-1} , pot no estar ben condicionada fins i tot quan $n > m$ i la inversa no existeix quan $m > n$ (on n és el nombre d'observacions i m és el nombre de variables). Això pot generar problemes en l'estimació de les correlacions parcials i les independències condicionals. En aquests casos, la matriu empírica de covariància no és suficient per obtenir resultats fiables, especialment quan hi ha més variables que observacions.

La correlació parcial entre X_j i X_k es pot calcular mitjançant la regressió de cadascuna d'elles en les altres variables i després calculant la correlació entre els residus d'aquestes dues regressions. Aquest mètode és útil perquè permet aïllar l'efecte directe entre les dues variables d'interès, eliminant la influència de les altres variables.

Amb variables centrades i escalades, considerem m regressions lineals $X_j = \sum_{k \neq j} \beta_{jk} X_k + \delta_j$, $j = 1, \dots, m$. Aquí β_{jk} representa la correlació parcial entre X_j i X_k . No obstant això, donada una mostra de mida n , les independències condicionals estimades entre X_j i X_k basades en β_{kj} i β_{jk} poden no coincidir. Això, pot ser degut a la variabilitat en l'estimació dels paràmetres de regressió, especialment en mostres petites o amb alta col·linealitat entre variables.

A més, la interpretació de les correlacions parcials pot ser complicada si les variables tenen relacions no lineals o si existeixen interaccions complexes. Per millorar l'estimació de les independències condicionals, sovint es recorre a tècniques de regularització com el lasso o el ridge regression, que ajuden a obtenir solucions més estables i interpretables.

Un altre enfocament és l'ús de mètodes bayesians, que incorporen informació a priori i permeten gestionar millor la incertesa en les estimacions. Aquests mètodes són particularment útils quan es treballa amb mostres petites o amb alta dimensionalitat.

En resum, calcular la inversa de la matriu de covariància empírica i les correlacions parcials requereix tenir en compte diverses consideracions estadístiques i tècniques per assegurar-se que les estimacions siguin fiables i interpretables en el context d'anàlisi de xarxes diferenciades.

5. COMPARACIÓ DE XARXES

5.1 Mesures de diferència entre xarxes

Ens centrarem en la comparació de dues xarxes G^1 i G^2 que tenen el mateix conjunt de nodes V i conjunts d'arestes E^1 i E^2 , o, les seves matrius d'adjacència A^1 i A^2 . Les arestes E^1 i E^2 poden haver estat observades directament, obtingudes d'experiments, o inferides a partir d'observacions sobre els nodes mitjançant mètodes de modelització gràfica.

5.1.1 Diferències globals La pregunta fonamental aquí és si A^1 és igual a A^2 . Per avaluar si les dues matrius d'adjacència són iguals, es poden utilitzar diferents normes o mesures de distància.

Per exemple, es pot calcular $A^1 - A^2$ utilitzant a alguna norma matricial específica. En el context de Models Gràfics Gaussians (*GGM*), això es tradueix en mesures com la distància estructural entre A^1 i A^2 , que compta el nombre total de diferències entre les dues xarxes. Aquesta distància estructural proporciona una mesura qualitativa de les diferències, mentre que la norma matricial proporciona una mesura quantitativa. A més, la topologia de les xarxes permet considerar mesures addicionals de diferències, com ara la mida i/o nombre de clústers, la connectivitat mitjana i la distribució del grau dels nodes.

5.1.2 Diferències locals En moltes aplicacions, les diferències locals entre les dues xarxes també poden ser d'interès. Això, inclou diferències en arestes individuals, veïnats o subxarxes. Aquest tipus d'anàlisi és especialment rellevant en aplicacions biològiques, on els biomarcadors basats en xarxes es poden utilitzar per investigar els mecanismes d'iniciació i progressió de malalties. Identificar les diferències locals és també útil després d'una prova global de diferència entre les dues xarxes. Aquestes diferències es poden avaluar tant qualitativament com quantitativament. Per exemple, es pot considerar la diferència entre els pesos de les arestes β_{jk1} i β_{jk2} , o identificar si una arista existeix en G_1 però no en G_2 i viceversa.

5.2 Prova global de diferències de xarxa

La hipòtesi nul·la global de no diferència entre dues xarxes de models gràfics gaussians (*GGM*), expressada formalment com $H_0 : E_1 = E_2$, es pot comprovar examinant si les matrius de correlació o les matrius de correlació parcial de les dues poblacions són diferents. En altres paraules, aquesta prova busca determinar si les estructures de dependència entre les variables dins de cada població són idèntiques.

Formalment, dues xarxes *GGM* es consideren iguals si es compleix la hipòtesi nul·la $H_0 : \Omega_1 = \Omega_2$, on Ω representa les matrius de precisió (l'inversa de les matrius de covariància) per a les dues poblacions. La igualtat de les matrius de precisió implica que les estructures de dependència condicional entre les variables són idèntiques en ambdues poblacions.

Els mètodes clàssics d'anàlisi multivariant poden tenir problemes de sensibilitat o baixa potència, especialment en situacions en què les dades són escasses. Aquestes limitacions fan que sigui crucial aplicar tècniques de regularització adequades per obtenir estimacions fiables de les matrius de precisió Ω , i així avaluar correctament la hipòtesi de no diferència entre les dues xarxes de models gràfics gaussians.

Per comparar dues xarxes i determinar si hi ha diferències significatives entre elles, es segueix un procés específic que inclou el càlcul d'un estadístic de prova i del p – *valor* associat. Primer, es defineix una hipòtesi nul·la (H_0) que estableix que no hi ha diferències entre les xarxes. A continuació, es calcula un estadístic de prova que mesura la diferència entre les xarxes, com ara la màxima diferència absoluta entre els pesos de les arestes.

Per avaluar la significança d'aquesta diferència, es realitzen múltiples permutacions dels nodes o de les dades associades als nodes per generar noves xarxes sota la hipòtesi nul·la. En cada permutació, es recalcula l'estadístic de prova per obtenir una distribució nul·la d'aquest estadístic. El p – *valor* es calcula com la proporció de permutacions en què l'estadístic de prova és igual o més extrem que l'estadístic observat.

Finalment, es compara l'estadístic de prova observat amb un valor crític obtingut de la distribució nul·la per a un nivell de significança determinat (per exemple, $\alpha = 0.05$). Si l'estadístic observat és més gran que el valor crític, es rebutja la hipòtesi nul·la, indicant que hi ha diferències significatives entre les xarxes comparades.

5.3 Estadístics avaluats pel contrast d'hipòtesi de diferència de xarxes

Com ja s'ha dit, Ω , omega, representa la matriu de precisió, que descriu les connexions directes entre nodes en una xarxa. Aquesta matriu és essencial en els models gràfics gaussians (GGM) perquè codifica tant la presència com les connexions de les arestes entre variables.

5.3.1 Estructura de Xarxa Invariant La prova d'invariància de xarxa avalua la hipòtesi nul·la que totes les arestes són iguals entre dues xarxes, formalment expressada com $H_0 : \Omega_1 = \Omega_2$. Això, implica que no hi ha diferències significatives en les connexions entre les dues xarxes, suggerint que les estructures de dependència condicional entre les variables són idèntiques en ambdues poblacions.

Una de les estadístiques de prova utilitzades per a aquesta avaluació és l'estadística màxima. Aquesta estadística proporciona el valor més gran, en termes absoluts, entre les diferències de pes de les arestes individuals de la xarxa 1 i la xarxa 2. L'estadística màxima es defineix com:

$$M(\omega^1, \omega^2) = \max_{ij} |\omega_{ij}^1 - \omega_{ij}^2|$$

On Ω_1 i Ω_2 representen els elements individuals de les matrius d'adjacència de les dues xarxes respectivament. Aquesta fórmula calcula la diferència absoluta més gran entre les corresponents arestes de les dues xarxes.

5.3.2 Força Global Invariant La tercera hipòtesi d'invariància afirma que el nivell general de connectivitat és el mateix entre les subpoblacions. La connectivitat global es pot resumir per la força global, que es defineix com la suma absoluta ponderada de totes les arestes de la xarxa. Aquesta mesura proporciona una visió agregada de la connectivitat total de la xarxa. La comparació de la força global seria rellevant si algú està interessat en les diferències en el nivell de connectivitat global, en lloc de diferències en arestes específiques.

No obstant això, la manca de diferències en la força global no implica necessàriament que les xarxes siguin similars en tots els aspectes. És possible que les xarxes tinguin el mateix nivell de connectivitat global, però difereixin significativament en la distribució de les connexions específiques. Observem la fórmula:

$$S(\omega^1, \omega^2) = \left| \sum_{i=1}^p \sum_{j>p} (|\omega_{ij}^1| - |\omega_{ij}^2|) \right|$$

En aquest context, p indica quantes variables o nodes hi ha a la xarxa i per tant quantes connexions possibles s'estan considerant en el càlcul de la força global.

5.3.3 Força de Vora Invariant La hipòtesi nul·la de la prova d'invariància de força de vora és $H_0 : w_{1ij} = w_{2ij}$, on w_{1ij} i w_{2ij} són els pesos de les arestes entre els nodes i i j en les dues xarxes respectivament.

$$E(\omega_{ij}^1, \omega_{ij}^2) = |\omega_{ij}^1 - \omega_{ij}^2|$$

Aquesta prova és útil en dos escenaris principals: Quan un investigador té una hipòtesi específica sobre una o més arestes concretes i en el cas que no tingui una hipòtesi específica, però la prova d'invariància de xarxa òmnibus és significativa, suggereix diferències generals en la xarxa. En el primer cas, la prova d'invariància de força de vora permet avaluar diferències precises en les connexions d'interès. I, en el segon cas, aquesta prova ajuda a identificar quines connexions específiques contribueixen a les diferències observades en l'estructura global de la xarxa.

En resum, aquestes estadístiques de prova ofereixen eines poderoses per a l'anàlisi comparativa de xarxes, permetent als investigadors avaluar tant les diferències globals com les específiques entre les xarxes de dades.

6. APLICACIÓ AMB DADES REALS

La diabetis mellitus (*DM*), comunament coneguda com a diabetis, és un grup de trastorns metabòlics en què hi ha nivells elevats de sucre en sang durant un període prolongat. La diabetis tipus 1 és el resultat de la manca de producció d'insulina del pàncrees. La diabetis tipus 2 comença amb la resistència a la insulina, una condició en què les cèl·lules no responen correctament a la insulina. A partir del 2015, s'estima que 415 milions de persones tenien diabetis a tot el món, i la diabetis tipus 2 representava al voltant del 90% dels casos. Això representa el 8,3% de la població adulta.

Aquest conjunt de dades de diabetis de Pima Indians es pot utilitzar per predir si un pacient determinat té diabetis a partir dels registres mèdics. Conté mesures relacionades amb els embarassos, la glucosa en sang, la pressió arterial diastòlica, el gruix de la pell, el nivell d'insulina, l'Índex de Massa Corporal (*BMI*), la funció Pedigree de la diabetis i l'edat. Finalment, hi ha la variable objectiu *Outcome*, on ens diu si té o no diabetis. Aquesta base de dades consta de 9 variables i 768 observacions, d'aquestes 500 pertanyen al grup de no diabetis i 268 als pacients que sí que tenen la malaltia.

En resum, l'anàlisi de dades clíniques es una eina poderosa per comprendre millor com la diabetis afecta a diferents variables de la salut. En aquest estudi, utilitzem dades de pacients amb i sense diabetis per explorar les relacions entre diverses variables clíniques. La tècnica de correlació parcial ens permet aïllar l'efecte directe d'una variable sobre un altre, controlant l'impacte de totes les altres variables en el conjunt de dades. Aquest estudi té com a objectiu comparar les relacions entre diverses variables, utilitzant les matrius de correlació parcial per controlar per altres variables i determinar les relacions directes entre parells de variables. A més, avaluem la significança d'aquestes diferències utilitzant un test de permutacions. Aquest enfocament ens permetrà identificar diferències claus en les interaccions entre variables clíniques que podrien estar influenciades per la presència de diabetis.

És important tenir en compte que cada persona és única, i la combinació de factors de risc pot variar en importància per a cada individu. Per tant, és essencial avaluar completa i personalitzada de la diabetis. Ja que la detecció i el maneig precoç de la diabetis són crucials per prevenir complicacions a llarg termini.

6.1 Introducció

En primer lloc, es procedeix a carregar les llibreries necessàries per a aquest estudi, així com les dades corresponents. Aquestes dades es divideixen en dos subconjunts diferents basats en la variable *Outcome*, creant així una base de dades per als pacients que tenen diabetis i una altra per als que no en tenen.

Un cop realitzats els passos inicials, s'assigna un nivell de significació per a l'estudi, que en aquest cas s'ha establert en 0,05. Aquest nivell de significació servirà com a llindar per determinar quines correlacions són estadísticament significatives.

A continuació, es procedeix a calcular les matrius de correlació parcial per a cadascun dels subconjunts de dades. Les matrius de correlació parcial permeten aïllar l'efecte directe d'una variable sobre una altra, controlant l'impacte de totes les altres variables en el conjunt de dades. Això és especialment útil per obtenir una visió més precisa de les relacions entre les variables clíniques.

Durant el càlcul de les matrius de correlació parcial, cada correlació obtinguda tindrà un *p* — *valor* associat, que indica la probabilitat que la correlació observada sigui deguda a l'atzar. Si aquest *p* — *valor* és superior

al nivell de significació establert, es considera que la correlació no és estadísticament significativa. En aquests casos, el valor de la correlació en la matriu es reemplaça per 0, indicant que no hi ha una relació significativa entre les dues variables corresponents.

Aquest procés assegura que només les correlacions significatives es mantinguin en les matrius de correlació parcial, proporcionant una representació més clara i precisa de les relacions entre les variables clíniques per als pacients amb diabetis i sense.

Observem cada una de les matrius de correlació parcial i el seus grafs corresponents. Els nodes dels grafs representen les diferents característiques mesurades en les dades de diabetis. Aquests nodes inclouen Prg (Embarassos), Age (Edat), Glc (Glucosa), BIP (Pressió Arterial Diastòlica), SkT (Gruix de la Pell), Ins (Insulina), BMI (Índex de Massa Corporal) i DPF (Funció Pedigree de la Diabetis).

Les arestes del graf representen les relacions entre les diferents característiques i el gruix d'aquestes indiquen la força de la correlació entre les variables. El color, també, és important, ja que si són de color verd representen correlacions positives i les vermelles negatives. Per tant, direm que les arestes més gruixudes i de color més intens suggereixen una correlació més forta.

Per aquest motiu, aquest tipus de gràfic és útil per identificar relacions significatives entre variables en el grup de persones sense diabetis i amb diabetis. Això pot ajudar a entendre quines variables estan més fortament relacionades i com poden influir entre elles.

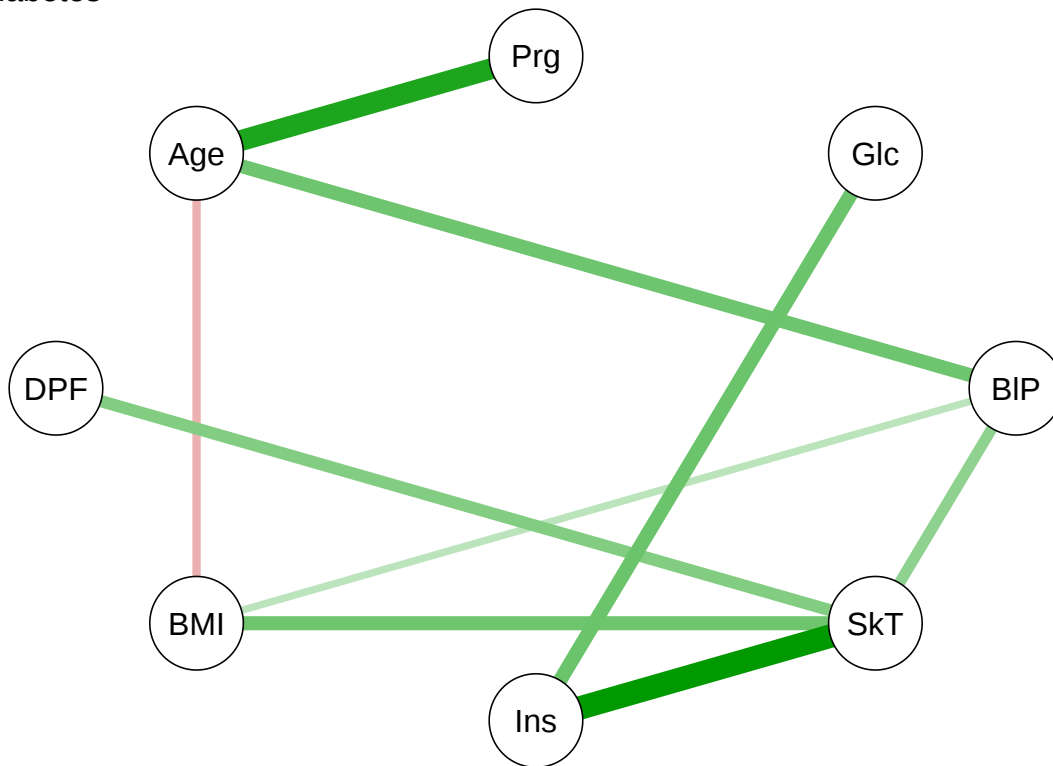
Matrius i gràfics

```
matriz_corr_si_diabetes
```

##	Pregnancies	Glucose	BloodPressure	SkinThickness
## Pregnancies	0.000	0.00	0.000	0.000
## Glucose	0.000	0.00	0.000	0.000
## BloodPressure	0.000	0.00	0.000	0.204
## SkinThickness	0.000	0.00	0.204	0.000
## Insulin	0.000	0.27	0.000	0.465
## BMI	0.000	0.00	0.122	0.265
## DiabetesPedigreeFunction	0.000	0.00	0.000	0.226
## Age	0.412	0.00	0.263	0.000
##	Insulin	BMI	DiabetesPedigreeFunction	Age
## Pregnancies	0.000	0.000	0.000	0.412
## Glucose	0.270	0.000	0.000	0.000
## BloodPressure	0.000	0.122	0.000	0.263
## SkinThickness	0.465	0.265	0.226	0.000
## Insulin	0.000	0.000	0.000	0.000
## BMI	0.000	0.000	0.000	-0.143
## DiabetesPedigreeFunction	0.000	0.000	0.000	0.000
## Age	0.000	-0.143	0.000	0.000

```
qgraph(matriz_corr_si_diabetes, title="Si diabetes")
```

Si diabetes



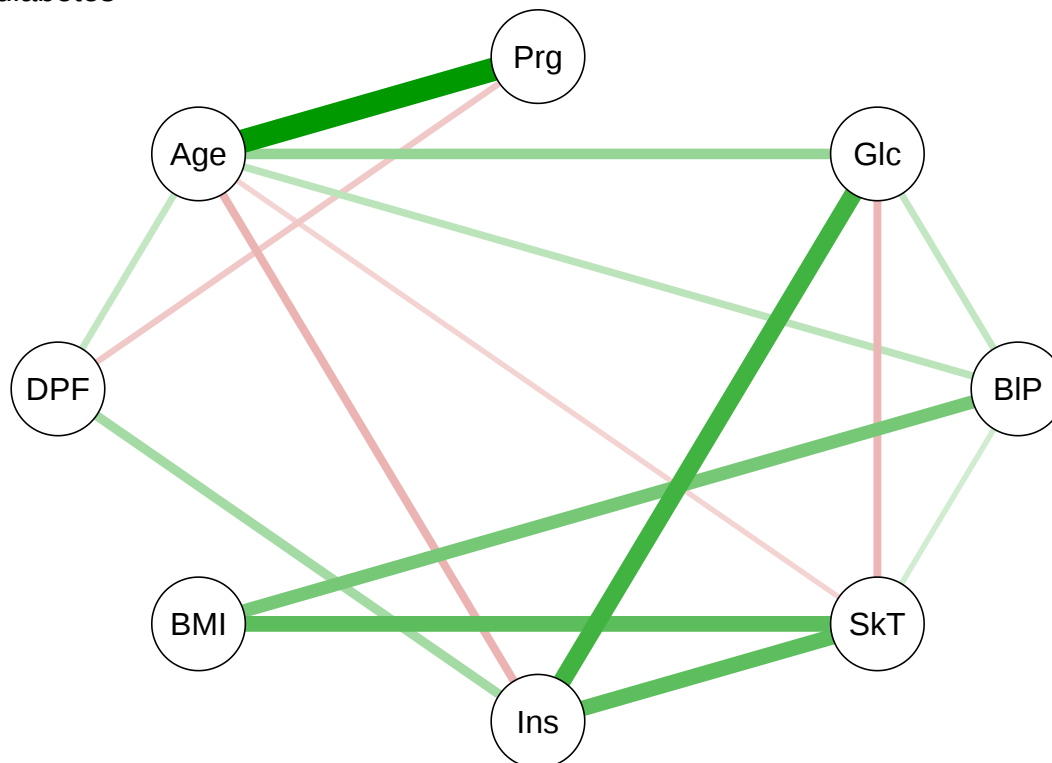
```
matriz_corr_no_diabetes
```

##	Pregnancies	Glucose	BloodPressure	SkinThickness
## Pregnancies	0.000	0.000	0.000	0.000
## Glucose	0.000	0.000	0.130	-0.156
## BloodPressure	0.000	0.130	0.000	0.098
## SkinThickness	0.000	-0.156	0.098	0.000
## Insulin	0.000	0.404	0.000	0.348
## BMI	0.000	0.000	0.291	0.350
## DiabetesPedigreeFunction	-0.117	0.000	0.000	0.000
## Age	0.543	0.222	0.145	-0.093
##	Insulin	BMI	DiabetesPedigreeFunction	Age
## Pregnancies	0.000	0.000	-0.117	0.543
## Glucose	0.404	0.000	0.000	0.222
## BloodPressure	0.000	0.291	0.000	0.145
## SkinThickness	0.348	0.350	0.000	-0.093
## Insulin	0.000	0.000	0.193	-0.161
## BMI	0.000	0.000	0.000	0.000
## DiabetesPedigreeFunction	0.193	0.000	0.000	0.128

```
## Age -0.161 0.000 0.128 0.000
```

```
qgraph(matriz_corr_no_diabetes, title="No diabetes")
```

No diabetes



En el graf “Si Diabetes”, es poden observar relacions fortes entre *Ins* (Insulina) i *SkinThickness* (Gruix de la Pell), entre *Pregnancies* (Embarassos) i *Age* (Edat). També es poden veure relacions moderades entre *BloodPressure* (Pressió Arterial Diastòlica) i *SkinThickness* (Gruix de la Pell), *BMI* (Índex de Massa Corporal) i *SkinThickness* (Gruix de la Pell), *BloodPressure* (Pressió Arterial Diastòlica) i *Age* (Edat), i *DiabetesPedigreeFunction* (Funció Pedigree de la Diabetis) i *SkinThickness* (Gruix de la Pell). Les relacions més febles es troben entre altres característiques, on les arestes són més fines i clares.

En el segon graf, també es poden veure relacions fortes, moderades i febles entre les diferents característiques. Les relacions fortes es troben entre *Pregnancies* (Embarassos) i *Age* (Edat), i entre *Glucose* (Glucosa) i *Insulin* (Insulina). També es poden veure relacions moderades entre *Insulin* (Insulina) i *SkinThickness* (Gruix de la Pell), i entre *BMI* (Índex de Massa Corporal) i *SkinThickness* (Gruix de la Pell). Tanmateix, hi ha diferències en les correlacions més febles en comparació amb el graf “Si Diabetes”

Com s’observa, ambdós grafos mostren relacions fortes entre *Pregnancies* (Embarassos) i *Age* (Edat), però en el graf “Si Diabetes” les relacions fortes inclouen també *Insulin* (Insulina) i *SkinThickness* (Gruix de la Pell), mentre que en el graf “No Diabetes” les relacions fortes inclouen *Glucose* (Glucosa) i *Insulin* (Insulina).

Les principals diferències que veiem fent la comparació de les dues xarxes anteriors són que en el graf “Si Diabetes”, la relació entre *BloodPressure* (Pressió Arterial Diastòlica) i *SkinThickness* (Gruix de la Pell) és

moderada, mentre que en el graf “No Diabetes” sembla menys destacada. Algunes relacions febles que es troben en el graf “Si Diabetes” són diferents en el graf “No Diabetes”, indicant canvis en les correlacions entre les característiques quan es compara amb diabetis o sense.

6.2 Comparació de Xarxes

6.2.1 Diferència de valors Per comparar les dues xarxes, s'utilitzen diversos enfocaments que reflecteixin i quantifiquin les diferències entre elles. No és una llista exhaustiva però són els més immediats i intuitius.

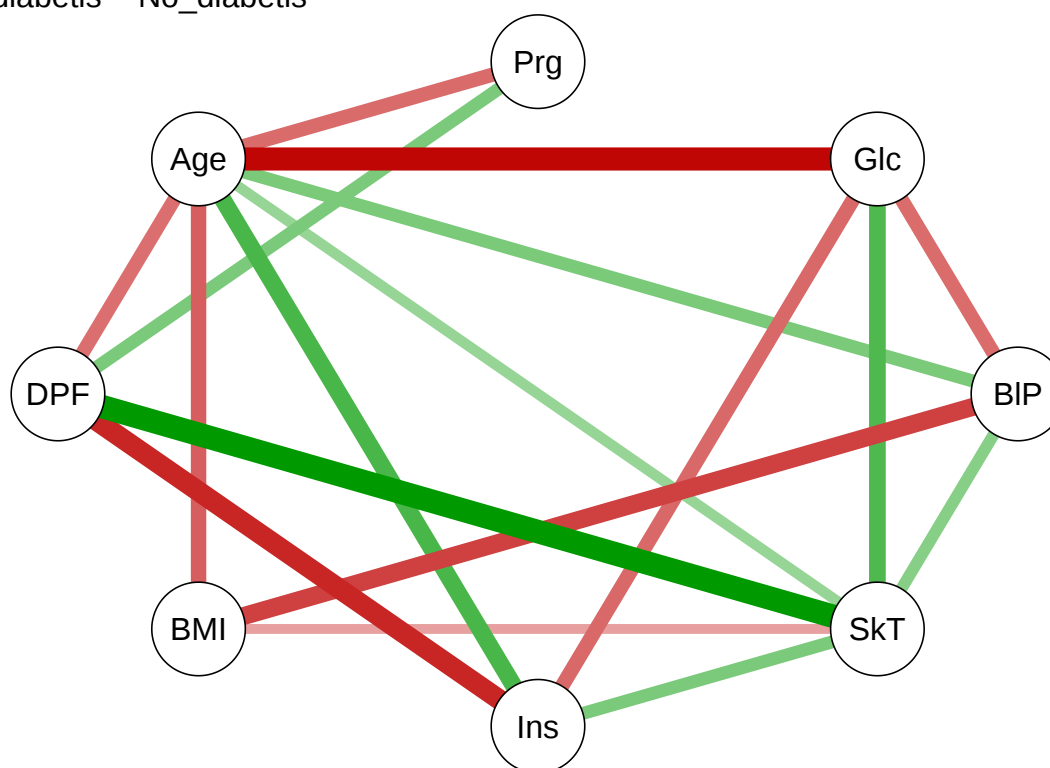
El primer pas consisteix en calcular la diferència entre els valors corresponents de les matrius de correlació parcial de cada xarxa. Aquest enfocament permet identificar les discrepàncies directes en les correlacions entre les variables clíniques dels dos grups estudiats.

Aquest procés implica restar els valors de la matriu de correlació parcial dels pacients sense diabetis dels valors corresponents de la matriu de correlació parcial dels pacients amb diabetis. La resultant matriu de diferències ens diu quines correlacions són més fortes o més febles en cada grup. Els valors positius en aquesta matriu indiquen que la correlació és més forta en el grup de pacients amb diabetis, mentre que els valors negatius indiquen una correlació més forta en el grup de pacients sense diabetis.

Aquest mètode proporciona una manera clara i directa de visualitzar i quantificar les diferències en les relacions entre variables clíniques entre els dos grups. Això pot ajudar a identificar patrons específics que podrien ser rellevants per a la comprensió i el tractament de la diabetis.

```
# Gràfica de la resta  
qgraph(matriu_diferencia, title="Si_diabetis - No_diabetis")
```

Si_diabetis – No_diabetis



El graf de la “Matriu diferència de valors” mostra les diferències en les correlacions entre les característiques per als grups “Si Diabetes” i “No Diabetes”.

En el graf, es poden observar diverses relacions més fortes en el grup “Si Diabetes”. Per exemple, la correlació entre DiabetesPedigreeFunction (Funció Pedigree de la Diabetes) i SkinThickness (Gruix de la Pell) és clarament més forta en el grup “Si Diabetes”. També es destaca una correlació més forta entre Age (Edat) i Insulin (Insulina) en el mateix grup. A més, la correlació entre SkinThickness (Gruix de la Pell) i Glucose (Glucosa) és més significativa en el grup amb diabetes.

Pel que fa a les relacions més fortes en el grup “No Diabetes”, es pot veure que la correlació entre Age (Edat) i Glucose (Glucosa) és més pronunciada en aquest grup. A més, la relació entre DiabetesPedigreeFunction (Funció Pedigree de la Diabetes) i Insulin (Insulina) també és més forta en el grup sense diabetes. De la mateixa manera, les correlacions entre BMI (Índex de Massa Corporal) i BloodPressure (Pressió Arterial Diastòlica), i entre Age (Edat) i BMI (Índex de Massa Corporal) són més destacades en el grup “No Diabetes”.

Finalment, els canvis en les correlacions són evidents entre els dos grups. Algunes relacions que eren fortes en un grup poden ser febles o fins i tot negatives en l'altre grup. Això suggereix que la presència de diabetes pot influir significativament en com es correlacionen aquestes característiques entre si, proporcionant informació valuosa sobre l'impacte de la diabetes en diferents mesures corporals.

En conclusió, diem que aquest graf de diferències ens permet identificar com varien les correlacions entre les característiques dels dos grups d'estudi. Les arestes verdes i gruixudes ens indiquen correlacions més fortes en el grup “Si Diabetes”, mentre que les arestes vermelles i gruixudes indiquen correlacions més fortes en el grup “No Diabetes”.

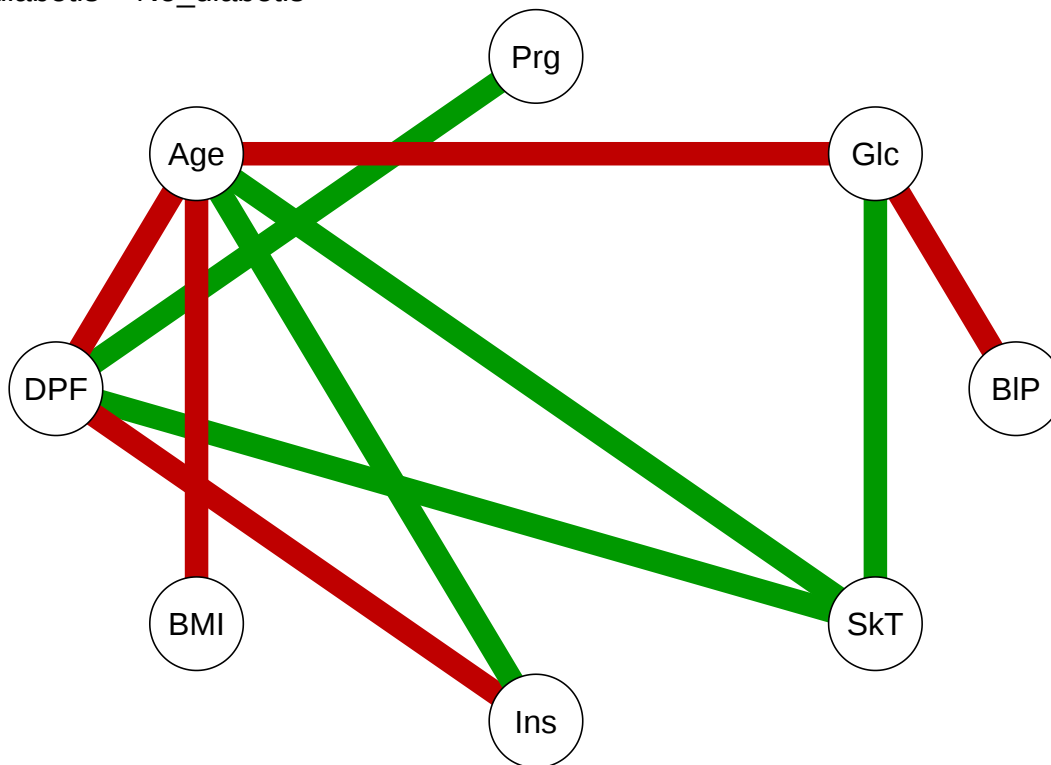
6.2.2 Diferència de signes La següent metodologia que utilitzarem per comparar les xarxes es basa en l'anàlisi dels signes de les correlacions.

Aquest mètode implica examinar si les correlacions entre les variables són positives o negatives en cada grup. Comparar els signes de les correlacions ens permet identificar si una relació determinada canvia de signe entre els dos grups.

Aquesta comparació es realitza assignant un valor positiu o negatiu a cada correlació en funció del seu signe. Després es resten aquests valors entre les dues matrius de correlació, resultant en una matriu de diferències de signe. Els valors d'aquesta matriu indicaran les diferències en el signe de les correlacions entre els dos grups.

```
# Gràfica del signe  
qgraph(matriu_signe, title="Si_diabetis - No_diabetis")
```

Si_diabetis - No_diabetis



En el graf, es poden observar diverses diferències en els signes de les correlacions entre les característiques en els dos grups.

En primer lloc, la correlació entre Age (Edat) i Insulin (Insulina) és positiva en el grup "Si Diabetes" i negativa en el grup "No Diabetes", com es veu en l'aresta verda gruixuda que connecta aquests dos nodes. Aquesta mateixa observació es fa entre SkinThickness (Gruix de la Pell) i Glucose (Glucosa), Pregnancies (Embarassos) i DiabetesPedigreeFunction (Funció Pedigree de la Diabetes), Age (Edat) i SkinThickness (Gruix de la Pell), i DiabetesPedigreeFunction (Funció Pedigree de la Diabetes) i SkinThickness (Gruix de la Pell), on la correlació és positiva en el grup "Si Diabetes" i negativa en el grup "No Diabetes".

Una altra observació important és la correlació entre `DiabetesPedigreeFunction` (Funció Pedigree de la Diabetis) i diverses característiques. Per exemple, entre `DiabetesPedigreeFunction` i `Insulin` (Insulina) i entre `DiabetesPedigreeFunction` i `Age` (Edat), les correlacions són positives en el grup “No Diabetes” però negatives en el grup “Si Diabetes”, representades per les arestes vermelles gruixudes.

Finalment, la correlació entre `Glucose` (Glucosa) i `BloodPressure` (Pressió Arterial Diastòlica) és positiva en el grup “No Diabetes” i negativa en el grup “Si Diabetes”, com es veu en l'aresta vermella entre aquests dos nodes. També, la correlació entre `Age` (Edat) i `Glucose` (Glucosa) segueix el mateix patró, amb una aresta vermella gruixuda que indica una correlació positiva en “No Diabetes” i negativa en “Si Diabetes”.

En resum, aquest graf de diferències en signes de les correlacions ens permet identificar canvis significatius en les relacions entre les característiques dels dos grups d'estudi. Les arestes verdes indiquen correlacions positives en el grup “Si Diabetes” i negatives en el grup “No Diabetes”, mentre que les arestes vermelles indiquen el contrari.

6.2.3 Diferència estructural Finalment, compararem les xarxes mitjançant l'anàlisi de les diferències en les seves estructures.

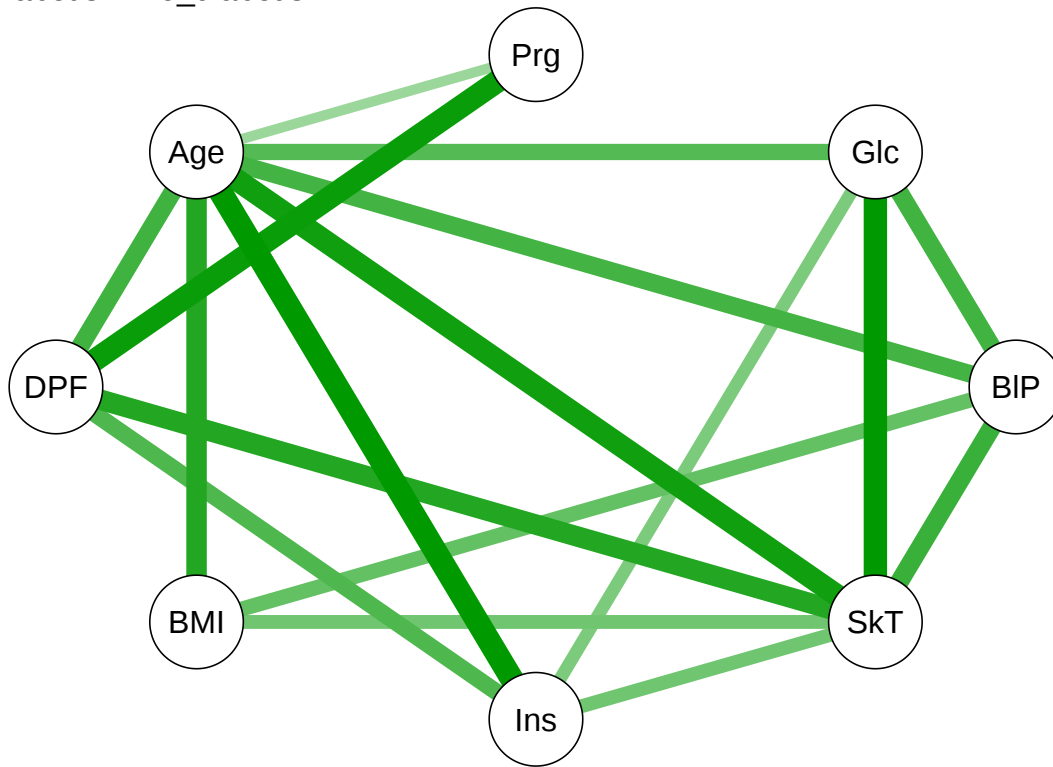
Aquest mètode consisteix a examinar les diferències en la connectivitat i l'organització global de les xarxes de correlació entre els dos grups. L'objectiu és identificar quines relacions entre les variables existeixen en un grup però no en l'altre.

Per fer això, es crea una matriu binària per a cada grup, on cada cel·la es converteix en 1 si hi ha una correlació significativa (diferent de 0) i en 0 si no hi ha correlació significativa. La matriu resultant de restar les dues matrius de correlació no és binària. Conté les diferències absolutes dels valors de correlació significatius entre els dos grups. Això vol dir que si hi ha una connexió significativa en un grup però no en l'altre, el valor absolut d'aquesta diferència es reflectirà en la matriu resultant.

Aquesta anàlisi permet identificar les diferències estructurals en les xarxes de correlació, destacant quines relacions entre variables són úniques per a un grup.

```
# Gràfica de l'estructura
qgraph(difnet, title="Si_diabetis - No_diabetis")
```

Si_diabetis – No_diabetis



En aquest gràfic es poden observar diverses diferències significatives en les estructures de correlació entre les característiques dels grups “Si Diabetes” i “No Diabetes”.

En primer lloc, les relacions fortes es destaquen amb arestes verdes i gruixudes. Això indica que moltes de les correlacions són més fortes o només existeixen en el grup “Si Diabetes”. Per exemple, la correlació entre Age (Edat) i Insulin (Insulina) és molt més forta en el grup “Si Diabetes”. Aquest patró es repeteix en altres parelles de nodes, com SkinThickness (Gruix de la Pell) i Glucose (Glucosa), i Pregnancies (Embarassos) i DiabetesPedigreeFunction (Funció Pedigree de la Diabetis).

Pel que fa a les relacions moderades, les arestes més fines i clares indiquen diferències menys significatives en les connexions entre els dos grups. Això suggereix que, tot i que hi ha diferències en les correlacions, aquestes no són tan pronunciades com en altres parelles de nodes.

Un aspecte destacable és la presència de múltiples correlacions fortes entre SkT (Gruix de la Pell) i altres característiques, com Age (Edat), Glc (Glucosa), i DPF (Funció Pedigree de la Diabetis) en el grup “Si Diabetes”. Això indica que el gruix de la pell està fortament relacionat amb aquestes altres mesures en persones amb diabetis, mostrant una connexió única o més pronunciada en aquest grup.

En conclusió, veiem com les arestes verdes i gruixudes indiquen correlacions més fortes en el grup “Si Diabetes”, destacant les connexions úniques o més prominents en aquest grup.

6.3 Estadístics de prova

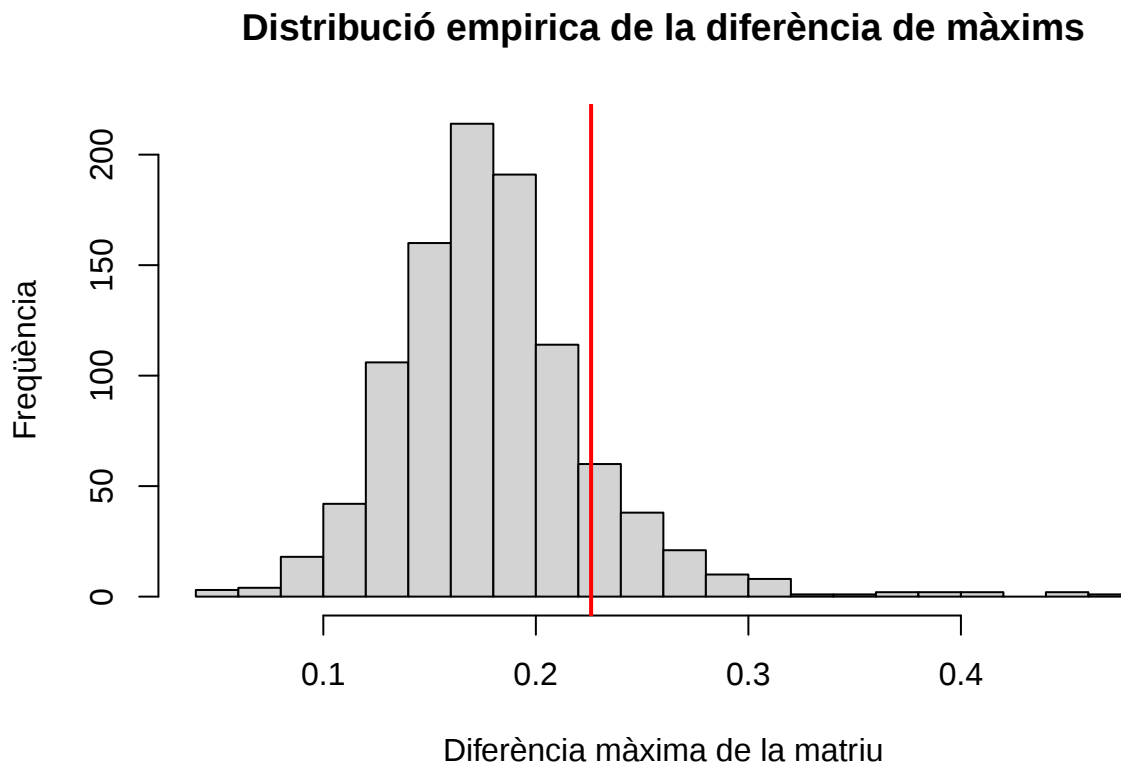
6.3.1 Màxima diferència Com ja s'ha esmentat, un cop sabem com comparar les matrius, utilitzarem diferents estadístics per analitzar les dades. Per fer-ho, primerament necessitarem calcular l'estadístic de màxima diferència en valor absolut, que ens servirà de base per a les permutacions.

A més, es pot visualitzar aquesta anàlisi en un histograma que mostra tots els valors dels màxims de les dades permutades, amb el valor màxim de les dades originals destacat en vermell. Aquesta representació visual ajuda a il·lustrar com es compara la diferència màxima original amb les distribucions de les diferències màximes permutades.

```
# P-valor
p_val <- sum(resultado_perm >= maxabsdif) /length(resultado_perm)
p_val
```

```
## [1] 0.129
```

```
# Distribució empírica
hist(resultado_perm, breaks=20, main="Distribució empírica de la diferència de màxims",
      xlab="Diferència màxima de la matriu", ylab="Freqüència")
abline(v=maxabsdif, col="red", lwd=2)
```



En el nostre cas, observem que el p — *valor* obtingut (0.112) és més gran que el nivell de significació establert. Per tant, podem concloure que no hi ha diferències significatives entre les xarxes. Per comprovar-ho plenament, realitzarem una altra prova utilitzant un estadístic diferent de la màxima diferència en valor absolut.

6.3.2 Nivell connectivitat Aquest nou estadístic representarà la diferència en el nivell de connectivitat entre les xarxes. Es calcularà sumant, per a cada aresta, la diferència dels valors absoluts en cada condició. Aquesta metodologia ens permetrà avaluar les diferències en la connectivitat general entre les xarxes en lloc de centrar-nos només en les diferències màximes individuals.

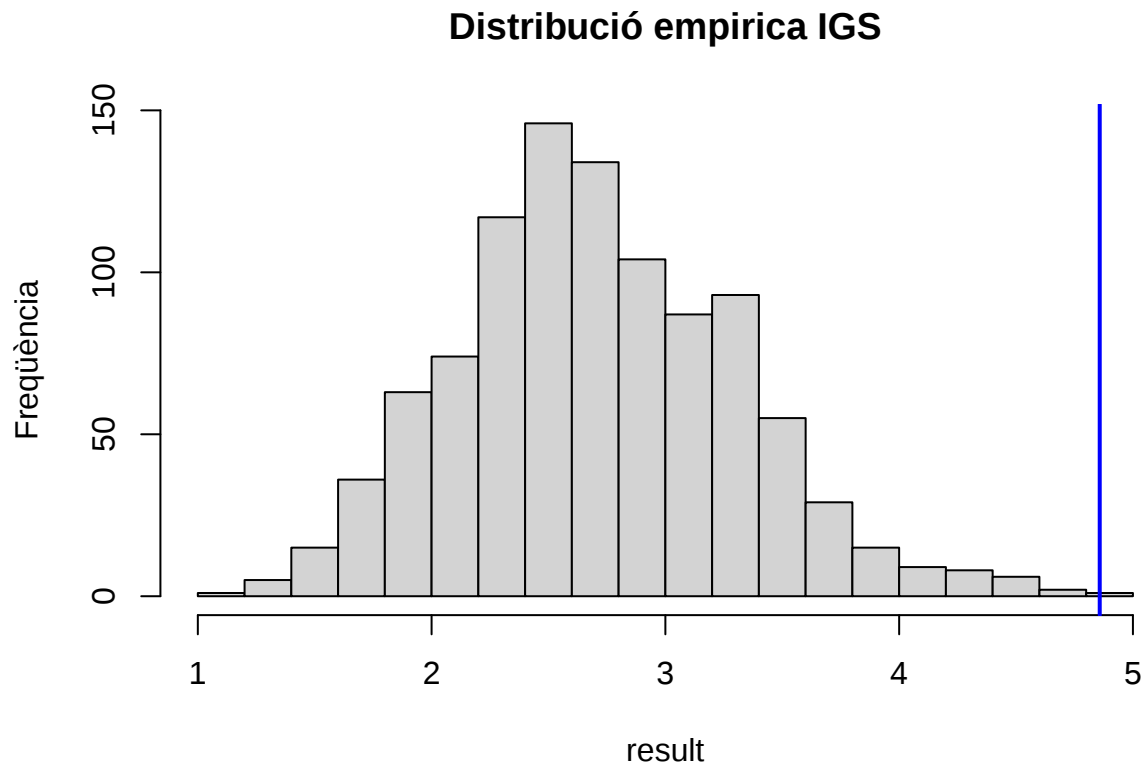
Després, es calcularà el p — *valor* del test de permutacions. Observem que aquest p -valor és més petit que el nivell de significació establert. Per tant, podem concloure que sí hi ha diferències significatives entre les xarxes.

A més, es pot visualitzar aquesta anàlisi en un histograma que mostra tots els valors de la diferència del nivell de connectivitat de les dades permutades, amb el valor corresponent a les dades originals destacat en blau. Aquesta representació visual ajuda a il·lustrar com es compara la diferència en el nivell de connectivitat de les dades originals amb les distribucions de les diferències obtingudes a partir de les permutacions.

```
# P-valor
p_val <- sum(result >= diferencia_cor) /length(result)
p_val
```

```
## [1] 0
```

```
# Distribució empírica
hist(result, breaks=20, main="Distribució empírica IGS", ylab="Freqüència")
abline(v=diferencia_cor, col="blue", lwd=2)
```



6.3.3 Força dels pesos de les arestes També podem analitzar l'estadístic de la força dels pesos de les arestes entre els nodes en les dues xarxes.

Aquest mètode implica comparar els pesos de les arestes, que representen la força de les correlacions entre les variables, en les dues xarxes (pacients amb diabetis i pacients sense diabetis). La força del pes d'una aresta indica la intensitat de la relació entre dos nodes (variables).

Per fer aquesta anàlisi, es calcula la suma dels pesos de les arestes per a cada node en ambdues xarxes. Això ens permet veure com de fortament connectades estan les variables dins de cada grup. Posteriorment, es comparen aquestes sumes entre les dues xarxes per identificar diferències en la força de les connexions.

Aquesta comparació de la força dels pesos de les arestes proporciona una visió detallada de com les relacions entre variables poden variar en intensitat depenent de la presència de diabetis, oferint així informació addicional sobre les diferències clíniques significatives entre els dos grups.

```
# Definim el node
nodo <- "SkinThickness"

# Calcul diferència mitjana i p-valor
resultat <- calculo(si_diabetes, no_diabetes, nodo)

# Resultat
```

```
print(resultat)
```

```
##          Nodo Diferencia_Media P_valor
## 1 SkinThickness          2.5    0.049
```

Segons els resultats de l'anàlisi estadística realitzada per comparar el node "SkinThickness" entre dos grups, "si_diabetes" i "no_diabetes", s'observa una diferència mitjana positiva de 2.50. Això significa que, de mitjana, les persones del grup "si_diabetes" tenen un valor de "SkinThickness" aproximadament 2.5 unitats superior en comparació amb les persones del grup "no_diabetes". Aquesta diferència suggereix una tendència cap a un augment del "SkinThickness" en individus amb diabetis.

A més a més, el valor del p-valor obtingut és de 0.049, està just en el llindar típic de significança estadística del 0.05. Aquest resultat implica que hi ha prou evidència estadística per rebutjar la hipòtesi nul·la, que assumeix que no hi ha diferència en "SkinThickness" entre els dos grups. Per tant, amb un nivell de confiança del 95%, podem concloure que la diferència observada en "SkinThickness" entre les persones amb i sense diabetis és estadísticament significativa. Això ens permet inferir que la diabetis pot estar associada amb un augment en els valors de "SkinThickness".

```
# Definim el node
nodo <- "Glucose"

# Calcul diferència mitjana i p-valor
resultat <- calculo(si_diabetes, no_diabetes, nodo)

# Resultat
print(resultat)
```

```
##          Nodo Diferencia_Media P_valor
## 1 Glucose          31.277      0
```

Segons els resultats de l'anàlisi estadística realitzada per comparar el node "Glucose" entre dos grups, "si_diabetes" i "no_diabetes", s'observa una diferència mitjana positiva de 31.277. Això significa que, de mitjana, les persones del grup "si_diabetes" tenen un nivell de glucosa aproximadament 31.28 unitats superior en comparació amb les persones del grup "no_diabetes". Aquesta diferència significativa indica que els individus amb diabetis tendeixen a tenir nivells de glucosa més alts.

A més a més, el valor p obtingut de 2.644e-36 és extremadament petit, molt inferior al llindar típic de significança estadística de 0.05. Aquest resultat implica que hi ha una fortíssima evidència estadística per rebutjar la hipòtesi nul·la, que assumeix que no hi ha diferència en els nivells de glucosa entre els dos grups. Amb un nivell de confiança altíssim, podem concloure que la diferència observada en els nivells de glucosa entre les persones amb i sense diabetis és estadísticament significativa. Això reforça la idea que la diabetis està fortament associada amb un augment en els nivells de glucosa.

6.4 Discussió de resultats

L'anàlisi diferencial de xarxes ha permès identificar diferències significatives en l'estructura i connectivitat de les variables entre pacients amb i sense diabetis. Les xarxes de pacients amb diabetis mostren una major correlació entre l'edat i la insulina, així com entre el gruix de la pell i la glucosa, i entre els embarassos i la funció pedigree de la diabetis. Això suggereix que la diabetis afecta no sols els nivells de glucosa sinó també altres variables clíniques clau, com la resistència a la insulina i les característiques corporals.

En els pacients sense diabetis, les correlacions més fortes són entre l'edat i la glucosa, i entre la funció pedigree de la diabetis i la insulina. Això indica una interacció diferent i més equilibrada entre les variables en comparació amb els pacients amb diabetis. A més, en els pacients amb diabetis s'observa una correlació forta entre la insulina i el gruix de la pell, la qual cosa podria reflectir resistència a la insulina, característica de la diabetis tipus 2. La correlació entre el gruix de la pell i la glucosa és més notable en els pacients amb diabetis, possiblement a causa dels seus efectes a llarg termini.

Les diferències trobades subratllen la necessitat d'una vigilància dels pacients amb diabetis, considerant no sols els nivells de glucosa sinó també altres variables interrelacionades. Això pot ajudar a prevenir i manejar millor les complicacions de la diabetis. A més, aquestes troballes suggereixen que un enfocament integral en el tractament de la diabetis, que inclogui gestió de pes, control de la pressió arterial i monitoratge de la insulina, podria ser més efectiu.

Encara que els resultats són significatius, l'estudi pot tenir limitacions, com la utilització de dades d'una població específica, la qual cosa limita la generalització dels resultats. Futures recerques haurien d'ampliar l'anàlisi a altres poblacions i malalties per a validar aquestes troballes i explorar la incorporació de diferents tipus de dades.

7. EL PAQUET `Differentialnetworkanalysis`

7.1 Creació del paquet

El primer pas en la creació del paquet va ser establir l'estructura de directoris necessària. Es van crear els directoris `R`, `data` i `man`. El directori `R` conté els scripts de funcions, el directori `data` emmagatzema les dades necessàries per als anàlisis i el directori `man` guarda la documentació generada automàticament per a cada funció i conjunt de dades del paquet.

Es va crear el fitxer `DESCRIPTION` per definir les metainformacions del paquet. Aquest fitxer inclou detalls crucials com el nom del paquet, la versió, l'autor, una breu descripció, la llicència i les dependències d'altres paquets de R necessaris per al correcte funcionament del paquet. A més, es va especificar la codificació UTF-8 per assegurar la correcta interpretació dels caràcters.

El fitxer `NAMESPACE` es va configurar per exportar les funcions i dades que estaran disponibles per als usuaris del paquet. Això inclou totes les funcions desenvolupades i el conjunt de dades utilitzat en l'anàlisi.

7.2 Objectiu i contingut del paquet

L'objectiu de desenvolupar el paquet `Differentialnetworkanalysis` va ser facilitar l'anàlisi de xarxes diferencials en estudis de dades complexes, específicament en el context de la diabetis. El paquet es va dissenyar per proporcionar eines que permetin realitzar anàlisis de correlacions parcials i tests de permutacions per identificar diferències significatives entre grups de dades.

Al fitxer `funciones.R`, es van definir diverses funcions clau necessàries per a l'anàlisi de xarxes diferencials. Aquestes funcions es poden trobar detallades a l'annex. Inclouen:

- Divisió de Dades i Eliminació de Variables de Resposta; Funció per dividir les dades i eliminar la variable de resposta, la qual cosa és útil per preparar les dades per a anàlisis més específics.
- Càlcul de la Matriu de Correlació Parcial; Funció per calcular la matriu de correlació parcial, que permet entendre les relacions entre variables mentre es controla per l'efecte d'altres variables.
- Tests de Permutacions; Funcions per realitzar tests de permutacions sobre les diferències en matrius de correlació, tant per a diferències màximes com totals, la qual cosa és crucial per determinar la significança estadística de les diferències observades entre grups.
- Càlcul de Diferències Estadístiques per Node; Funció per calcular diferències estadístiques per node entre dues matrius de correlació, que ajuda a identificar nodes (variables) específics que mostren diferències significatives entre els grups.

Per documentar adequadament el conjunt de dades `diabetes`, es va crear un fitxer `datos.R` al directori `R`. Aquest fitxer proporciona una descripció detallada del conjunt de dades, incloent-hi el format de les dades i una descripció de cada variable.

El desenvolupament del paquet `Differentialnetworkanalysis` no només facilita l'anàlisi de xarxes diferencials en estudis de dades complexes, sinó que també proporciona un marc estructurat i reproducible per dur a terme aquests anàlisis. En encapsular les funcions i dades necessàries en un paquet, s'assegura que els mètodes i resultats siguin consistents i que es puguin compartir fàcilment amb la comunitat de recerca.

8. CONCLUSIONS

Els resultats obtinguts mitjançant la comparació de xarxes amb dades reals han revelat diferències significatives en les relacions entre variables clíniques dels pacients amb i sense diabetis, suggerint que la diabetis influeix notablement en la correlació d'aquestes característiques. La metodologia utilitzada, basada en tests de permutacions i en l'anàlisi de diferències de valors, signes i estructura de xarxes, ha permès identificar patrons específics de connectivitat que varien entre els dos grups estudiats. A més, l'aplicació d'aquests mètodes en casos pràctics, com l'anàlisi de dades biològiques, ha demostrat no només la seva utilitat, sinó també la seva versatilitat en diversos camps d'estudi, subratllant la importància de continuar refinant aquests mètodes per maximitzar el seu potencial en la investigació científica.

Un dels objectius del treball, i que s'ha assolit al llarg del mateix, ha estat demostrar que es pot plantejar la comparació entre dues xarxes construïdes amb els mateixos nodes en diferents situacions, a través de tècniques estadístiques inferencials (test de permutacions) anant més enllà d'una comparació descriptiva. Això proporciona una eina poderosa per a l'anàlisi diferencial de xarxes, permetent identificar i quantificar diferències significatives.

Aquest estudi presenta certes limitacions que cal tenir en compte. En primer lloc, els resultats obtinguts amb dades de pacients Pima Indians poden no ser aplicables a altres poblacions, per la qual cosa és necessari validar els resultats amb conjunts de dades més diversos. A més, les tècniques estadístiques utilitzades tenen certes limitacions i assumpcions que poden influir en els resultats, i arribar a tenir biaixos.

Pel que fa als futurs treballs, seria beneficiós ampliar l'anàlisi a altres malalties per validar la utilitat de l'anàlisi diferencial de xarxes en altres àmbits de la medicina. Incorporar altres tipus de dades podria proporcionar una visió més completa de les interaccions biològiques. Finalment, realitzar estudis de validació clínica col·laborant amb professionals de la salut per implementar les troballes en pràctiques clíniques reals, seria crucial per acabar de validar les conclusions obtingudes en aquest treball.

BIBLIOGRAFIA

1. **Thomas Cormen and Devin Balkcom**, “Representar grafos,” 2016. Disponible en: <https://es.khanacademy.org/computing/computer-science/algorithms/graph-representation/a/representing-graphs>.
2. **INAOE**, “Ciencias computacionales,” 2024. Disponible en: https://posgrados.inaoep.mx/archivos/PosCsComputacionales/Curso_Propedeutico/Matematicas_Discretas/Capitulo_4_Grafos.pdf.
3. **Seongho Kim**, “ppcor: Partial and Semi-Partial (Part) Correlation,” 2015. Disponible en: <https://cran.r-project.org/web/packages/ppcor/index.html>.
4. **Antoni Miñarro**, “Differential Network Analysis,” 2024. Disponible en: DNA.pdf.
5. “Package ‘NetworkComparisonTest’,” 2023. Disponible en: <https://cran.r-project.org/web/packages/NetworkComparisonTest/NetworkComparisonTest.pdf>.
6. **Real Academia Española**, “Definición de grafo,” 2024. Disponible en: <https://dle.rae.es/graf%C3%B3?m=form>.
7. **Universidad de Valladolid**, “TFG: Título del Trabajo Final de Grado,” 2021. Disponible en: <https://uvadoc.uva.es/handle/10324/50020>.
8. **R Bloggers**, “Estimating Mixed Graphical Models,” 2015. Disponible en: <https://www.r-bloggers.com/2015/11/estimating-mixed-graphical-models/>.
9. **Springer**, “Graphical Models with R,” 2012. Disponible en: <https://extras.springer.com/?query=978-1-4614-2298-3>.
10. **Journal of Statistical Software**, “mgm: Estimating Time-Varying Mixed Graphical Models in High-Dimensional Data,” 2020. Disponible en: <https://www.jstatsoft.org/article/view/v093i08>.
11. **American Mathematical Society**, “Mathematics Subject Classification (MSC2020),” 2020. Disponible en: <https://mathscinet.ams.org/msc/msc2020.html>.
12. **Wikipedia**, “Grafo,” 2024. Disponible en: <https://es.wikipedia.org/wiki/Grafo>.
13. **Universidad Autónoma del Estado de Hidalgo**, “Tipos de grafos,” 2022. Disponible en: http://cidecame.uaeh.edu.mx/lcc/mapa/PROYECTO/libro5/42_tipos_de_grafos.html.
14. **Graph Everywhere**, “Tipos de grafos,” Disponible en: <https://www.grapheverywhere.com/tipos-de-grafos/>.
15. **Paul Mooney**, “Predict Diabetes from Medical Records,” 2018. Disponible en: <https://www.kaggle.com/code/paultimothymooney/predict-diabetes-from-medical-records/notebook>.
16. **Helena U. Zacharias**, “Gaussian and Mixed Graphical Models as (multi-)omics data analysis tools,” 2019. Disponible en: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7166149/>.
17. **Jonas Haslbeck**, “Estimating Group Differences in Network Models using Moderation,” 2020. Disponible en: <https://jonashaslbeck.com/Groupdifferences-via-Moderation/>.
18. **Adrián Alacreu Crespo**, “Network analysis workshop,” 2022. Disponible en: <https://rpubs.com/aalacreu/network>.
19. **Centre for Complex Systems Studies**, “MGM Mixed Graphical Models,” Disponible en: <https://www.uu.nl/en/research/centre-for-complex-systems-studies-ccss/research/complexity-lab-utrecht-clue/software/data-analysis/mgm-mixed-graphical-models>.
20. **Seungjun Ahn and Somnath Datta**, “Differential network connectivity analysis for microbiome data adjusted for clinical covariates using jackknife pseudo-values,” 2024. Disponible en: <https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-024-05689-7>.

21. **Jonas Haslbeck**, “*Package ‘mgm’*,” 2023. Disponible en: <https://cran.r-project.org/web/packages/mgm/mgm.pdf>.
22. **The R Project for Statistical Computing**. Disponible en: <https://www.r-project.org/>.
23. **Mariana Montes**, “*Academic Writing in R Markdown II*,” Disponible en: <https://www.marianamontes.me/post/academic-writing-in-r-markdown-ii/>.
24. **Ulrike Lyngs**, “*Oxforddown*,” Disponible en: <https://ulyngs.github.io/oxforddown/index.html>.
25. **Our Coding Club**, “*RMarkdown Dissertation Tutorial*,” Disponible en: <https://ourcodingclub.github.io/tutorials/rmarkdown-dissertation/>.
26. **The Comprehensive R Archive Network**, “*dineR Package*,” Disponible en: <https://cran.r-project.org/web/packages/dineR/index.html>.
27. **BioRxiv**, “*Differential network analysis for high-dimensional, heterogeneous data*,” Disponible en: <https://www.biorxiv.org/content/10.1101/2021.04.10.439301v2.full>.
28. **ScienceDirect**, “*Differential network analysis of high-dimensional time series*,” Disponible en: <https://www.sciencedirect.com/science/article/pii/S0378111912014497>.
29. **BioNetStat**, “*Differential Network Analysis*,” Disponible en: <https://rdr.io/bioc/BioNetStat/man/diffNetAnalysis.html>.
30. **Statnet**, “*Introduction to SNA Tools*,” Disponible en: <https://statnet.org/workshop-intro-sna-tools/>.
31. **Statnet**, “*Home Page*,” Disponible en: <https://statnet.org/>.

ANNEX

Funcions del paquet Differentialnetworkanalysis:

```
# Divideix i elimina la variable resposta
divide_y_elimina <- function(df, respuesta_var) {
  data_1 <- df[df[[respuesta_var]] == 1, -which(names(df) == respuesta_var)]
  data_0 <- df[df[[respuesta_var]] == 0, -which(names(df) == respuesta_var)]
  return(list(data_1 = data_1, data_0 = data_0))
}

# Càlcul de la matriu de correlació parcial
matriz_correlacion <- function(df, nivel_significacion = 0.05) {
  correlaciones <- pcor(df)
  mat_cor <- correlaciones$estimate
  p_values <- correlaciones$p.value
  mat_cor[p_values > nivel_significacion] <- 0
  for (i in 1:ncol(mat_cor)) mat_cor[i,i] <- 0
  return(mat_cor)
}

# Test de permutacions per la diferència màxima de matriu de correlació
test_permutaciones <- function(df, respuesta_var, n_permutaciones = 1000,
                               nivel_significacion = 0.05) {
  resultados_maximos <- numeric(n_permutaciones)
  for (i in 1:n_permutaciones) {
    df_permutado <- df
    df_permutado[[respuesta_var]] <- sample(df[[respuesta_var]])
    division <- divide_y_elimina(df_permutado, respuesta_var)
    matriz_cor_1 <- matriz_correlacion(division$data_1, nivel_significacion)
    matriz_cor_0 <- matriz_correlacion(division$data_0, nivel_significacion)
    diferencia_maxima <- max(matriz_cor_1 - matriz_cor_0)
    resultados_maximos[i] <- diferencia_maxima
  }
  return(resultados_maximos)
}

# Test de permutacions per la diferència total de matrius de correlació
test <- function(df, respuesta_var, n_permutaciones = 1000, nivel_significacion
                 = 0.05) {
  resultados_IGS <- numeric(n_permutaciones)
  for (i in 1:n_permutaciones) {
    df_permutado <- df
    df_permutado[[respuesta_var]] <- sample(df[[respuesta_var]])
```

```

    division <- divide_y_elimina(df_permutado, respuesta_var)
    matriz_cor_1 <- matriz_correlacion(division$data_1, nivel_significacion)
    matriz_cor_0 <- matriz_correlacion(division$data_0, nivel_significacion)
    diferencia_cor <- sum(abs(matriz_cor_1 - matriz_cor_0))
    resultados_IGS[i] <- diferencia_cor
  }
  return(resultados_IGS)
}

# Càlcul estadístic per node entre matrius de correlació
calculo <- function(matriz1, matriz2, nodo) {
  if (!(nodo %in% colnames(matriz1)))
    stop("El nodo no existe en la matriz1")
  if (!(nodo %in% colnames(matriz2)))
    stop("El nodo no existe en la matriz2")
  valores1 <- matriz1[, nodo]
  valores2 <- matriz2[, nodo]
  test <- t.test(valores1, valores2)
  p_valor <- test$p.value
  diferencia_media <- mean(valores1) - mean(valores2)
  results <- data.frame(Nodo = nodo, Diferencia_Media = round(diferencia_media,
    3), P_valor = round(p_valor, 3))
  return(results)
}

```

Codi per l'anàlisi de les dades:

```

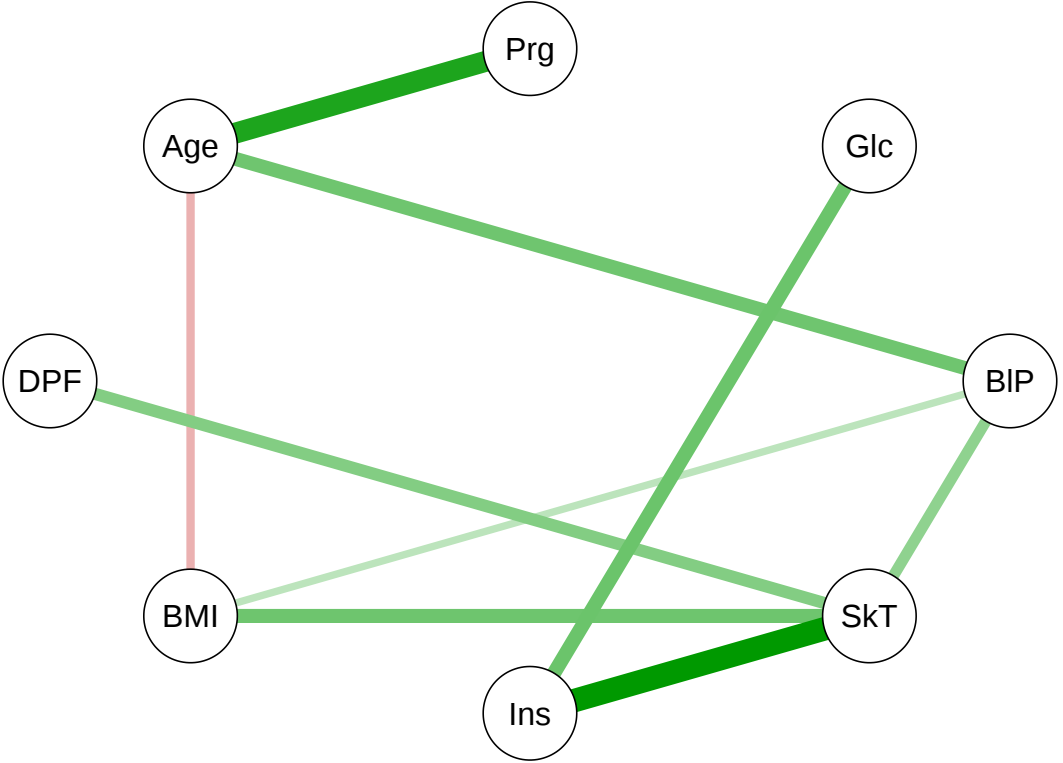
# Dades
library(Differentialnetworkanalysis)
data("diabetes")
result <- divide_y_elimina(diabetes, "Outcome")
si_diabetes <- as.data.frame(result$data_1)
no_diabetes <- as.data.frame(result$data_0)

# Matrius de correlació
matriz_corr_no_diabetes <- round(matriz_correlacion(no_diabetes, nivel_significacion = 0.05), 3)
matriz_corr_si_diabetes <- round(matriz_correlacion(si_diabetes, nivel_significacion = 0.05), 3)

# Gràfics de les matrius
qgraph(matriz_corr_si_diabetes, title="Si diabetes")

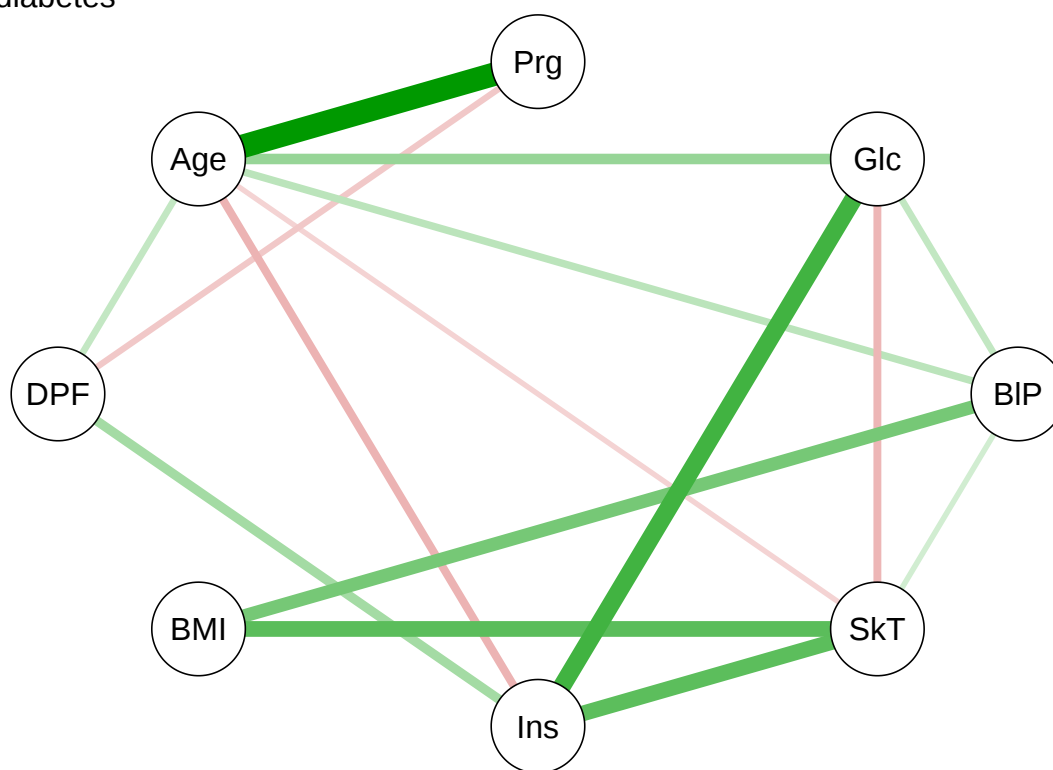
```

Si diabetes



```
qgraph(matriz_corr_no_diabetes, title="No diabetes")
```

No diabetes

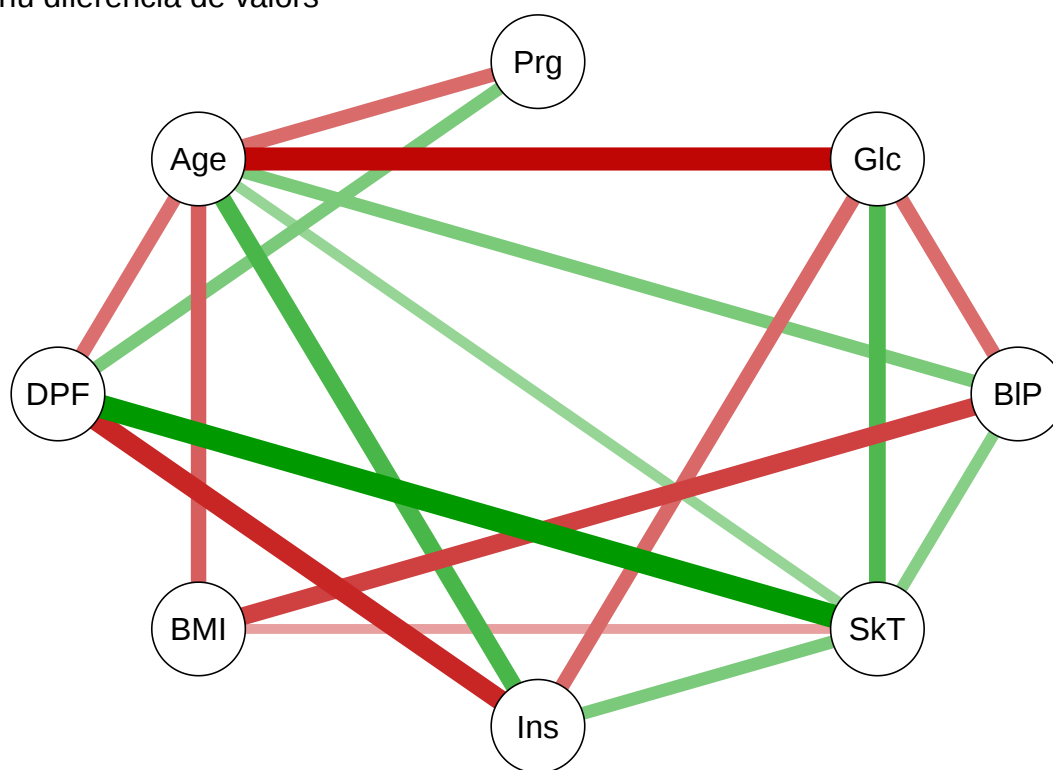


```
# Diferència de valors
```

```
matriu_diferencia <- matriz_corr_si_diabetes - matriz_corr_no_diabetes
```

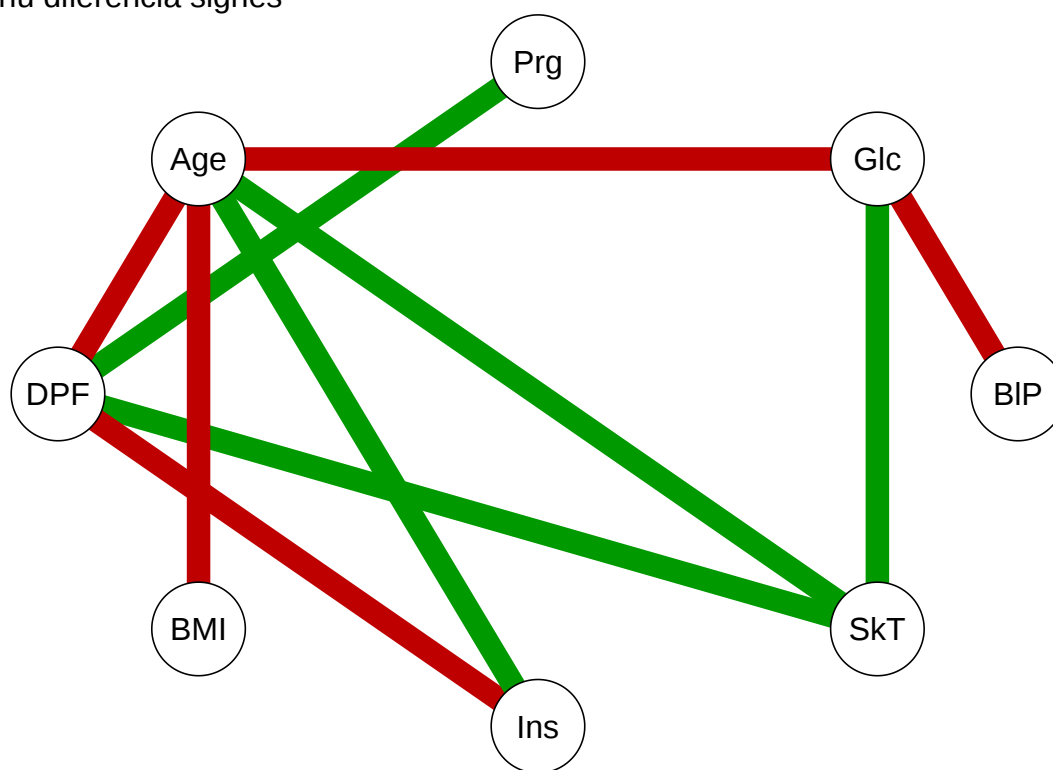
```
qgraph(matriu_diferencia, title="Matriu diferència de valors")
```

Matriu diferència de valors



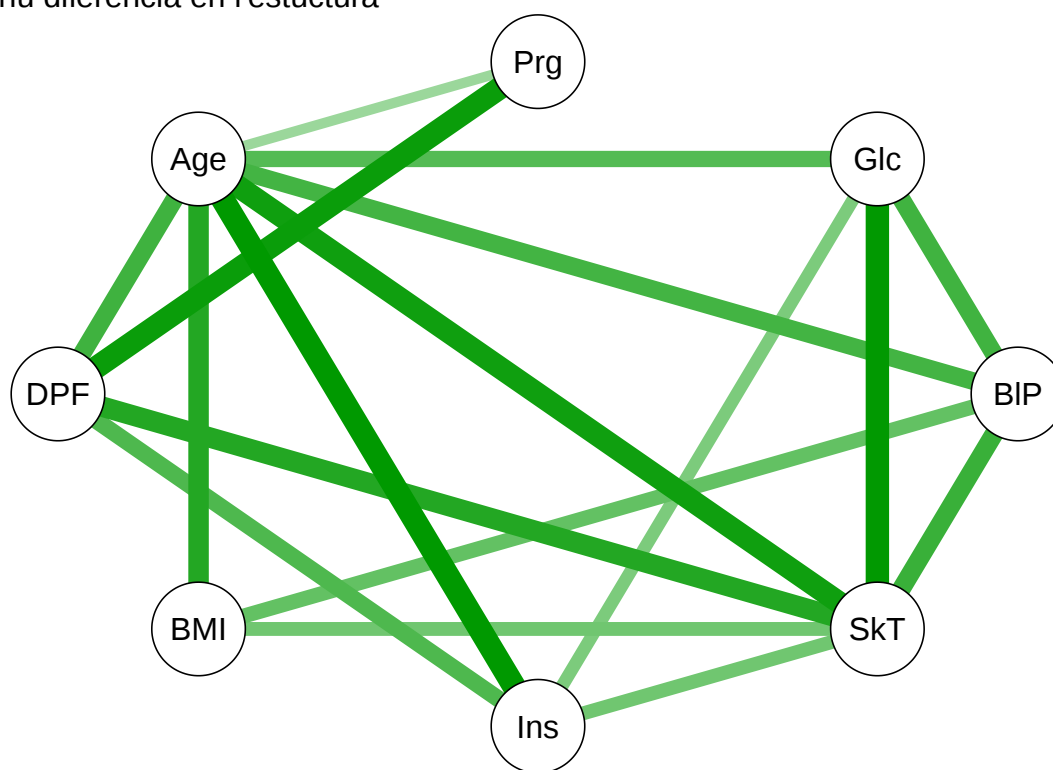
```
# Diferència del signe
signe_si <- sign(matriz_corr_si_diabetes)
signe_no <- sign(matriz_corr_no_diabetes)
matriu_signe <- signe_si - signe_no
qgraph(matriu_signe, title="Matriu diferència signes")
```

Matriu diferència signes



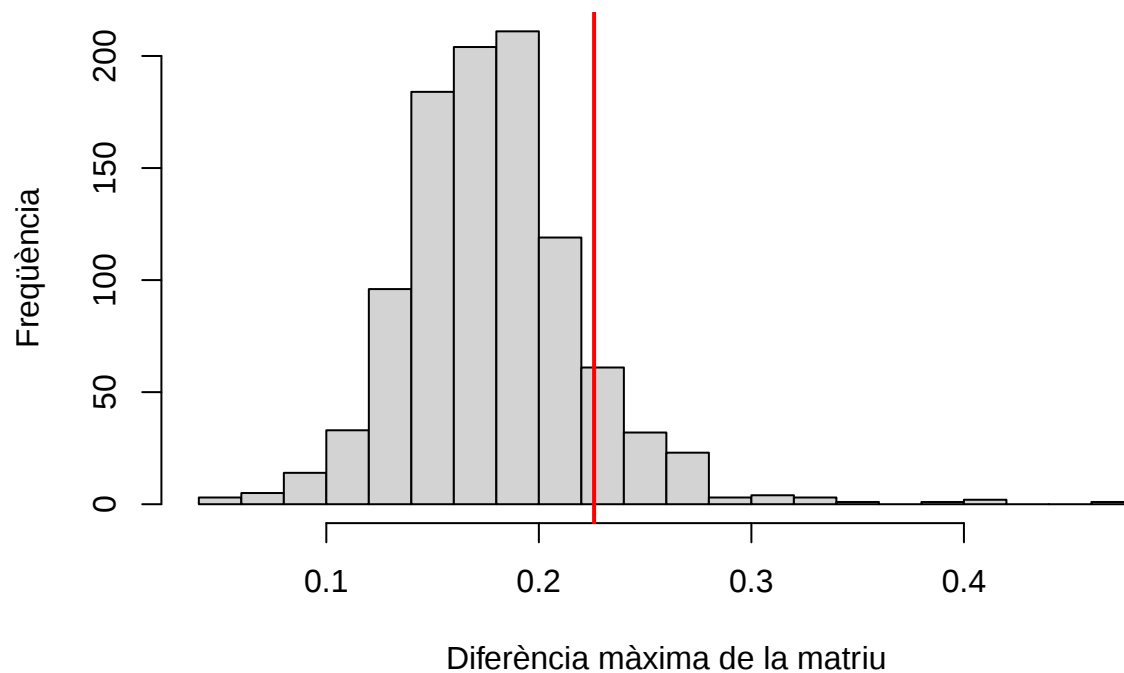
```
# Diferència en l'estructura
netSi <- matriz_corr_si_diabetes
netSi[netSi!=0] <- 1
netNo <- matriz_corr_no_diabetes
netSi[netNo!=0] <- 1
difnet <- abs(netSi-netNo)
qgraph(difnet, title="Matriu diferència en l'estructura")
```

Matriu diferència en l'estuctura



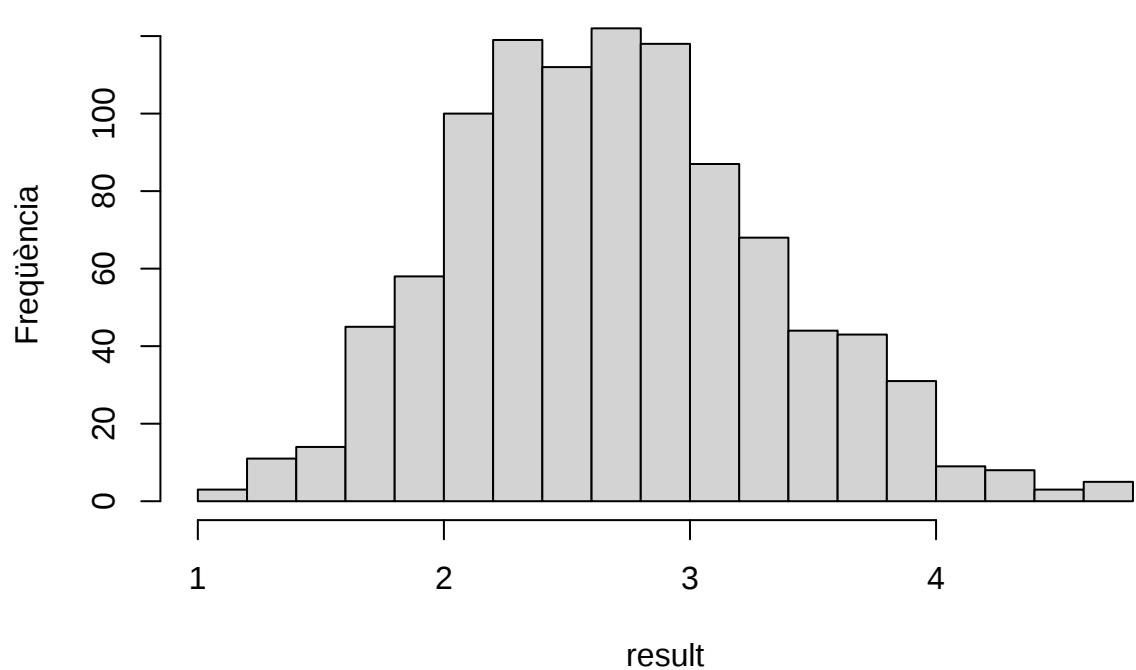
```
# Estadístic: Màxima diferència absoluta entre arestes
maxabsdif <- max(matriu_diferencia) #Valor dades originals
resultado_perm <- test_permutaciones(diabetes, respuesta_var = "Outcome", n_permutaciones = 1000)
p_val <- sum(resultado_perm >= maxabsdif) /length(resultado_perm)
hist(resultado_perm, breaks=20, main="Distribució empírica de la diferència de màxims", xlab="Diferència")
abline(v=maxabsdif, col="red", lwd=2)
```

Distribució empírica de la diferència de màxims



```
# Estadístic: Diferència en el nivell de connectivitat
diferencia_cor <- sum(abs(matriz_corr_si_diabetes-matriz_corr_no_diabetes))
result <- test(diabetes, respuesta_var = "Outcome", n_permutaciones = 1000)
p_val <- sum(result >= diferencia_cor) /length(result)
hist(result, breaks=20, main="Distribució empírica IGS", ylab="Freqüència") # Distribució empírica
abline(v=diferencia_cor, col="blue", lwd=2)
```

Distribució empírica IGS



```
# Estadístic pel node "SkinThickness"
nodo <- "SkinThickness"
resultat <- calculo(si_diabetes, no_diabetes, nodo) # Calcul diferència mitjana i p-valor
print(resultat)
```

```
##          Nodo Diferencia_Media P_valor
## 1 SkinThickness          2.5    0.049
```

```
# Estadístic pel node "Glucose"
nodo <- "Glucose"
resultat <- calculo(si_diabetes, no_diabetes, nodo) # Calcul diferència mitjana i p-valor
print(resultat)
```

```
##          Nodo Diferencia_Media P_valor
## 1 Glucose          31.277      0
```