Análisis de coincidencias con R-Shiny

M. Escobar, L. Martínez-Uribe, F. Martínez y J.L. A. Berrocal

Universidad de Salamanca y Fundación Juan March

VII Jornadas de Usuarios de R - Comunidad R Hispano

Salamanca, 5 de Noviembre

Índice

Análisis de coincidencias

Definición

El análisis de coincidencias es un conjunto de técnicas cuyo objeto consiste en detectar y representar qué sucesos, objetos o sujetos tienden a aparecer al mismo tiempo en unos espacios delimitados.

- Estos N espacios delimitados (i) se denominan escenarios y pueden considerarse unidades de análisis(registros).
- En cada uno de estos escenarios (campos) un conjunto de J sucesos (x_{ij}) pueden estar presentes (1) o ausentes (0).
- Un conjunto de escenarios forman una matriz binaria de incidencias ($\mathbf{X}$) con dimensiones ($N \times J$).
- Estos escenarios pueden agruparse en H subconjuntos para poderlos comparar.

Material de análisis

Matriz de incidencias (Aparición o ausencia de 8 sucesos in 4 escenarios)

El material de análisis en el análisis de coincidencias es una matriz **X** construida con *i* filas, que representan los escenarios, y *j* columnas, que representan los sucesos

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

Matriz de coincidencias

Definición

- A partir de la matriz de incidencias ($\mathbf{X}$), se puede obtener la matriz de coincidencias ($\mathbf{F}$) mediante la siguiente operación
  - donde cada elemento f_{jk} representa el número de escenarios en los que x_{ij} y x_{ik} tienen el valor de 1, lo que equivale a decir que coinciden.
- Los elementos diagonales de la matriz (f_{jj}) equivalen al número de x_{ij} en los N escenarios.

Ejemplo de matriz de coincidencias

Coapariciones o coocurrencias en los escenarios)

La matriz simétrica **F** está compuesta por *i* filas and *j* columnas y representa las incidencias (diagonal) y coincidencias de los sucesos:

$$\mathbf{F} = \begin{bmatrix} 3 & & & & & & & \\ 3 & 4 & & & & & & \\ 3 & 4 & 4 & & & & & \\ 2 & 3 & 3 & 3 & & & & \\ 2 & 3 & 3 & 3 & 3 & & & \\ 1 & 2 & 2 & 2 & 2 & 2 & & \\ 1 & 2 & 2 & 2 & 2 & 2 & 2 & \\ 1 & 1 & 1 & 1 & 1 & 0 & 0 & 1 \end{bmatrix}$$

7 grados de coincidencias

Clasificación

Las coincidencias entre sucesos pueden graduarse del modo siguiente:

- Sin coincidencia (sucesos mutuamente excluyentes)
- Mera coincidencia (al menos coinciden en un escenario)
- Probable ($\mathrm{p}(\mathbf{x}_j|\mathbf{x}_k) > 0.5$)
- Estadísticamente probable ($\mathrm{P}(\mathrm{p}(\mathbf{x}_j|\mathbf{x}_k) \leq 0.5) < c$)
- Condicional ($\mathrm{p}(\mathbf{x}_j) < \mathrm{p}(\mathbf{x}_j|\mathbf{x}_k)$)
- Estadísticamente condicional ($\mathrm{P}(\mathrm{p}(\mathbf{x}_j) - \mathrm{p}(\mathbf{x}_j|\mathbf{x}_k) \leq 0) < c$)
- Total (siempre ocurren en los mismos escenarios)

Gráfico condicional de coincidencias

Grados de coincidencia (a).

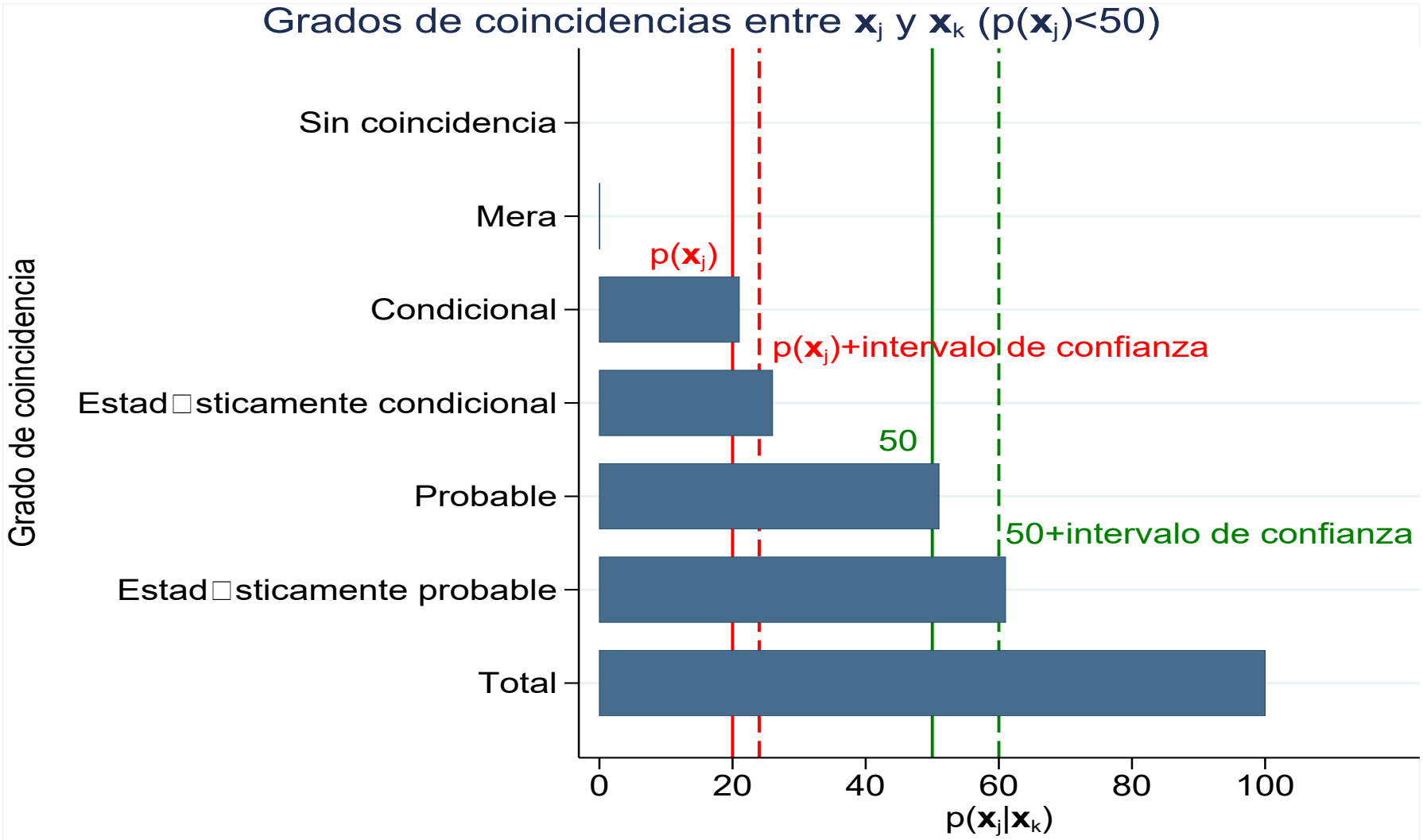
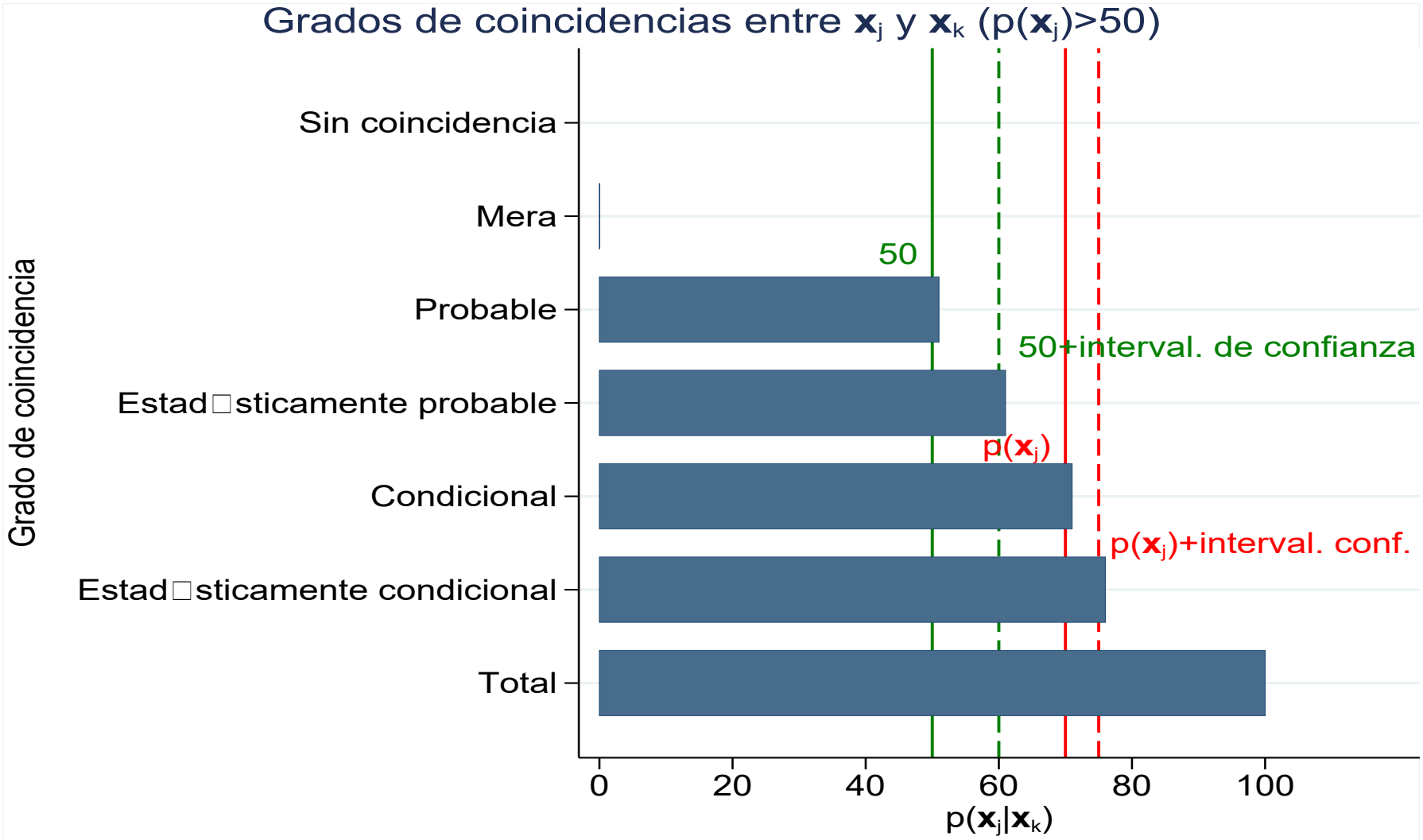


Gráfico condicional de coincidencias

Grados de coincidencia (b)



Dependencia estadística
Medición

- Se pueden emplear los residuos de Haberman (r_{jk}) para estimar la significación estadística de la coincidencia entre dos sucesos.

$$r_{jk} = \frac{f_{jk} - \frac{f_{jj} f_{kk}}{n}}{\sqrt{\frac{1-f_{jj}}{n} \frac{1-f_{kk}}{n}}}$$

Adyacencias

Definición operacional de sucesos adyacentes (coincidentes).

- Dos sucesos j y k pueden ser considerados adyacentes si siguen las siguiente norma:

$$A[j,k] = 1 \Leftrightarrow [\mathrm{P}(r_{jk} \leq 0) < c] \wedge j \neq k$$

- Por tanto, se puede construir una $J \times J$ matriz **A** con elementos igual a 1 en el caso en que r_{jk} sea significativo a un determinado nivel ($c.$) e iguales a 0 en el resto de elementos incluidos los diagonales.
- A partir de **A** puede calcularse la $J \times J$ matriz de distancias geodésicas (distancia mínima entre elementos conectados) **D** y elaborar un grafo en el que los vértices o nodos sean los sucesos y los vínculos o aristas indiquen las coincidencias entre ellos.

Adyacencias (cont.)

Otras operacionalizaciones de adyacencias

- Por extensión, se pueden elaborar otras matrices de adyacencias con otros criterios:
  - Criterio de la mera adyacencia

$$A[j,k] = 1 \Leftrightarrow f_{jk} \geq 1$$

  - Criterio de la adyacencia probable

$$A[j,k] = 1 \Leftrightarrow [\mathrm{P}(r_{jk} \leq 0) < 0.5] \wedge j \neq k$$

Representaciones gráficas de las coincidencias

Tipos de gráficos

- Gráficos de barras
- Dendrograms
- Grafos
  - Geométricos
    - Circular
    - Estrella
    - Rejilla
  - Físicos
    - Fruchterman-Reingold
    - Kamada-Kawai
  - Estadísticos
    - mds (escalas multidimensionales
    - pca (análisis de componentes principales)
    - ca (análisis de correspondencias
    - biplot
- Comunidades
- Bloques recursivos

Índice

Implementación del análisis de coincidencias

Necesidades, requisitos, funciones y librerías

- El algoritmo de análisis coin se desarrolló inicialmente en Stata.
- Se consideró interesante hacerlo en otras plataformas de análisis estadístico como R.
- Necesitades y requisitos
  - Realizar cálculos estadísticos y matriciales.
  - Visualizaciones con gráficos.
  - Interfaz web interactiva capaz de mostrar e interactuar con diversos ejemplos.
- Funciones propias para los cálculos.
  - gcoin
  - dichotomize
- Librerías para visualizaciones e interacción web.
  - igraph
  - R-Shiny

Función gcoin()

Disponible en github.

- Dos subfunciones importantes:
  - **Haberman**: cálculo de los residuos para estimar significación estadística.
  - **Adjacency**: genera la matriz de adyacencias.

Función dichotomize()

Preparación de la matriz

- Convierte datos con una columna de escenarios (Exposición) y otra de sucesos (Artistas) en una matriz de incidencias con tantas columnas como sucesos haya habido en el conjunto de escenarios.

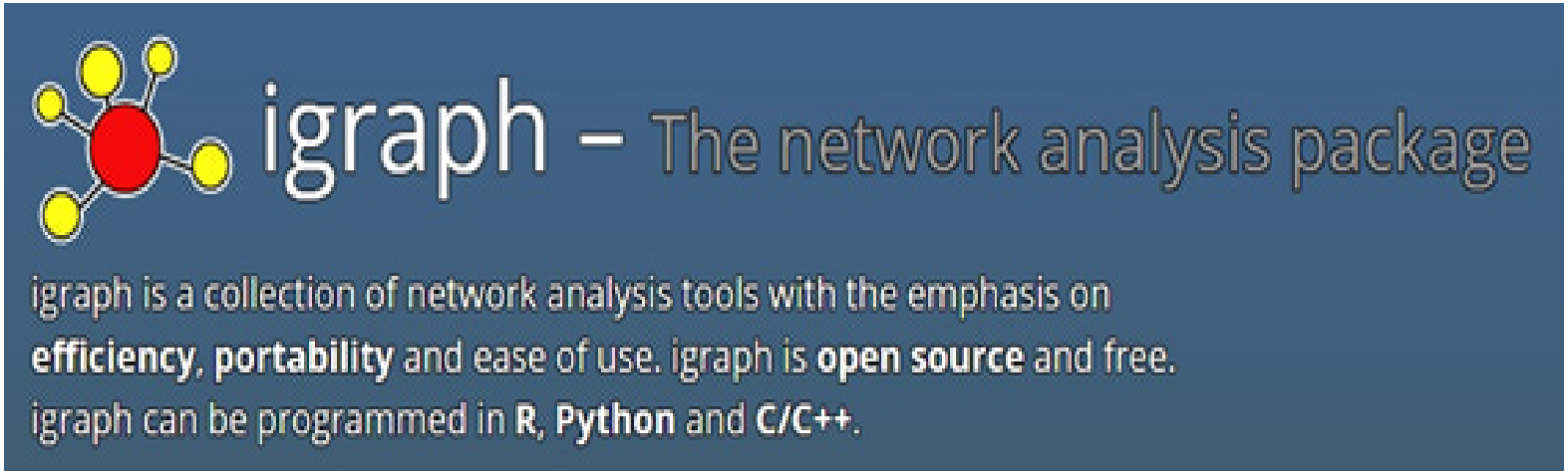
	Exposición	Artistas
1	Vladímir Lébedev (1891–1967)	Lébedev, Vladímir 1891–1967
2	Pablo Palazuelo	Palazuelo, Pablo 1916–2007
3	The American Landscapes of Asher B. Durand (1796–...	Durand, Asher B. 1796–1886
4	Andy Warhol	Warhol, Andy 1928–1987
5	Georges Braque	Braque, Georges 1882–1963
6	Magritte	Magritte, René 1898–1967
7	Malevich	Malevich, Kazimir Severinovich 1878–1935
8	Mark Rothko	Rothko, Mark 1903–1970
9	Max Ernst	Ernst, Max 1891–1976
10	Medio siglo de escultura (1900–1945)	Archipenko, Alexander 1887–1964;Arp, Jean 1887–1...
11	Monet en Giverny	Monet, Claude 1840–1926
12	Rauschenberg	Rauschenberg, Robert 1925–2008
13	Vieira da Silva	Vieira da Silva, Maria Helena 1908–1992
14	Zero, un movimiento europeo	Arman 1928–2005;Bury, Pol 1922–2005;Dorazio, Pie...
15	David Hockney	Hockney, David 1937–
16	Edward Hopper	Hopper, Edward 1882–1967

	Titulo	Kandinsky, Wassily 1866–1944	Picasso, Pablo 1881–1973	Arp, Jean 1887–1966	Albers, Josef 1888–1975
V1	Vladímir Lébedev (1891–1967)	0	0	0	0
V2	Pablo Palazuelo	0	0	0	0
V3	The American Landscapes of Asher B. Durand (1796–...	0	0	0	0
V4	Andy Warhol	0	0	0	0
V5	Georges Braque	0	0	0	0
V6	Magritte	0	0	0	0
V7	Malevich	0	0	0	0
V8	Mark Rothko	0	0	0	0
V9	Max Ernst	0	0	0	0
V10	Medio siglo de escultura (1900–1945)	0	1	1	0
V11	Monet en Giverny	0	0	0	0
V12	Rauschenberg	0	0	0	0
V13	Vieira da Silva	0	0	0	0
V14	Zero, un movimiento europeo	0	0	0	0
V15	David Hockney	0	0	0	0
V16	Edward Hopper	0	0	0	0

igraph
Paquete para el análisis de redes

Proporciona un conjunto de tipos de datos y funciones para:

1. Generación de gráficos simples de redes
2. Administración de grandes gráficos con miles de vértices y aristas
3. Visualización gráfica



Shiny
Entorno web para programas de R

- Permite crear aplicaciones web interactivas con R.
- Vincula entradas y salidas de manera automática y reactiva.
- Posee una extensa colección de widgets pre-construidos para elaborar aplicaciones fácilmente.

Índice

Primer ejemplo de Shiny

Compositores en Londres ($n > 15$) (a): Color

Coincidence Analysis (Bachtrack concerts reviewed 2009-2015)

Coincidence type: Probable

Layout: Fruchterman-Reingold

City: London

Node minimum frequency: 1 — 16 — 60

Zoom: None

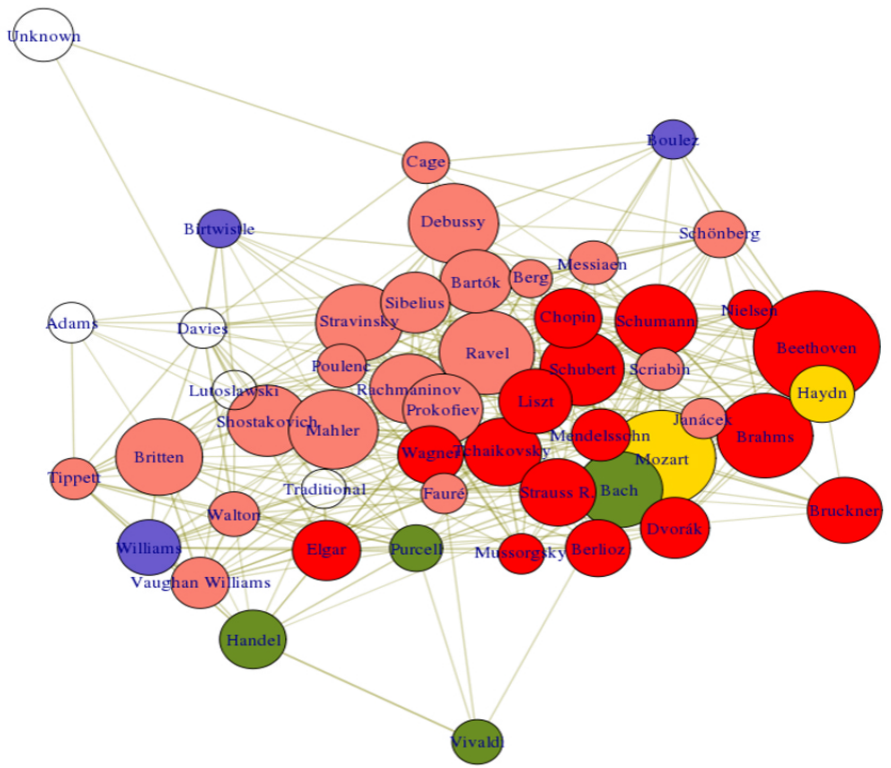
Zoom scope: 1

Color: Periodo

- None
- Periodo

Node size: 1 — 50 — 100

Edge width: 1 — 5 — 20



Primer ejemplo de Shiny

Compositores en New York ($n > 4$) (b): Comunidades óptimas.

Coincidence Analysis (Bachtrack concerts reviewed 2009-2015)

Coincidence type:
Probable

Layout:
Fruchterman-Reingold

City:
New York

Node minimum frequency:
1 5 60
1 7 13 19 25 31 37 43 49 55 60

Zoom:
None

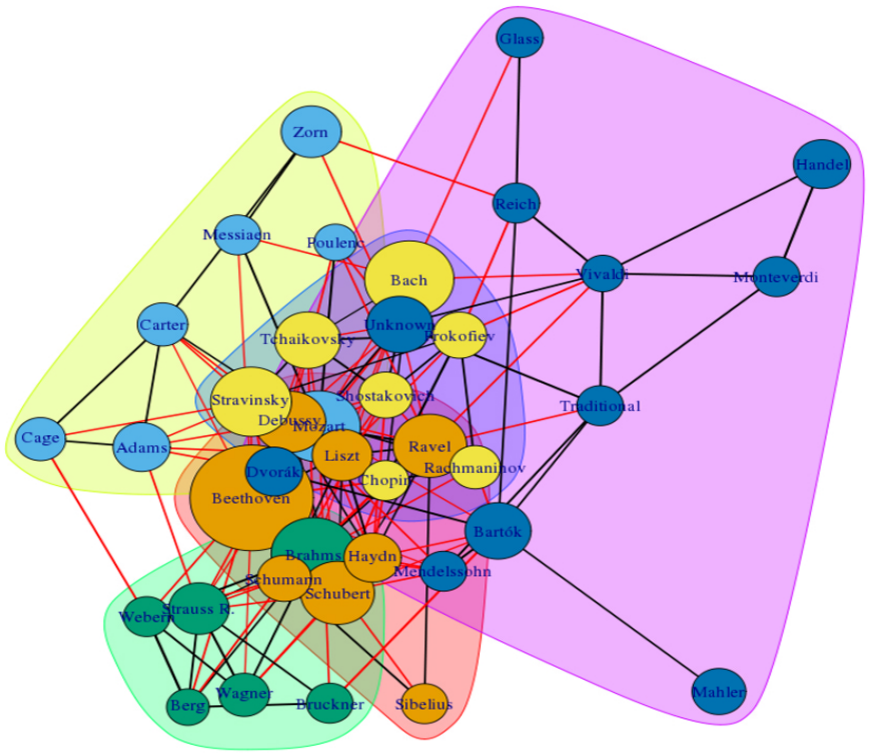
Zoom scope:
1

Color:
Periodo

Blocks/communities:
Optimal communities
None
Blocks
Optimal communities
1 11 21 31 41 51 61 71 81 91 100

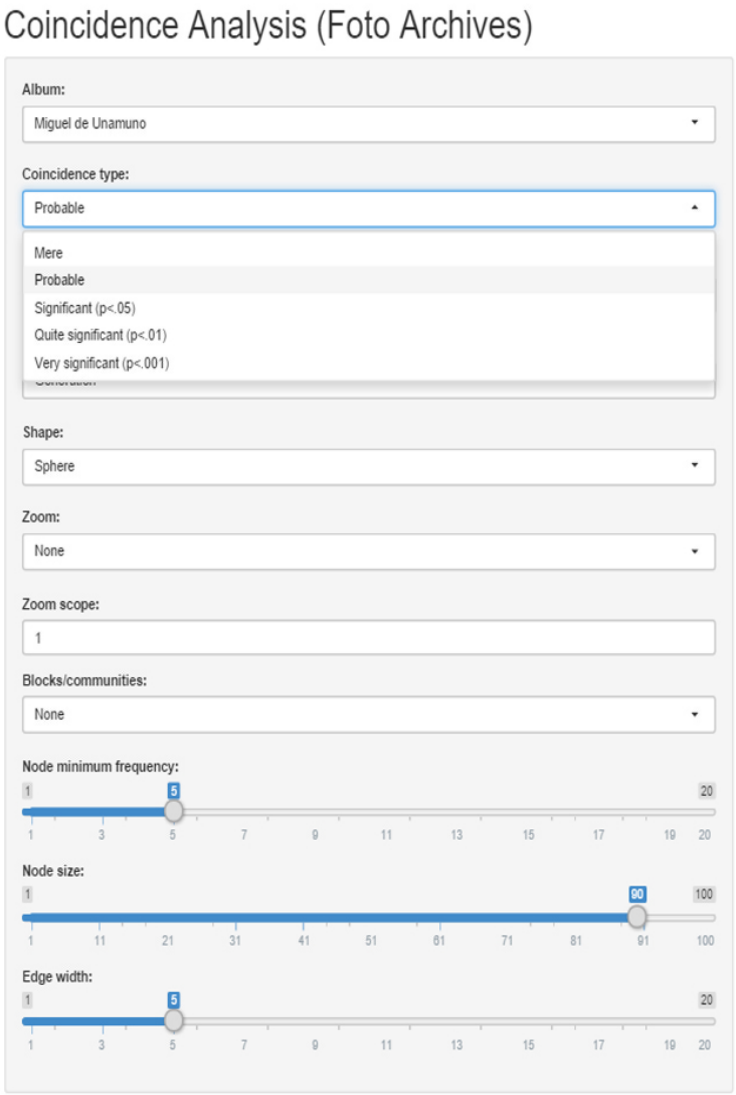
Edge width:
1 5 20
1 3 5 7 9 11 13 15 17 19 20

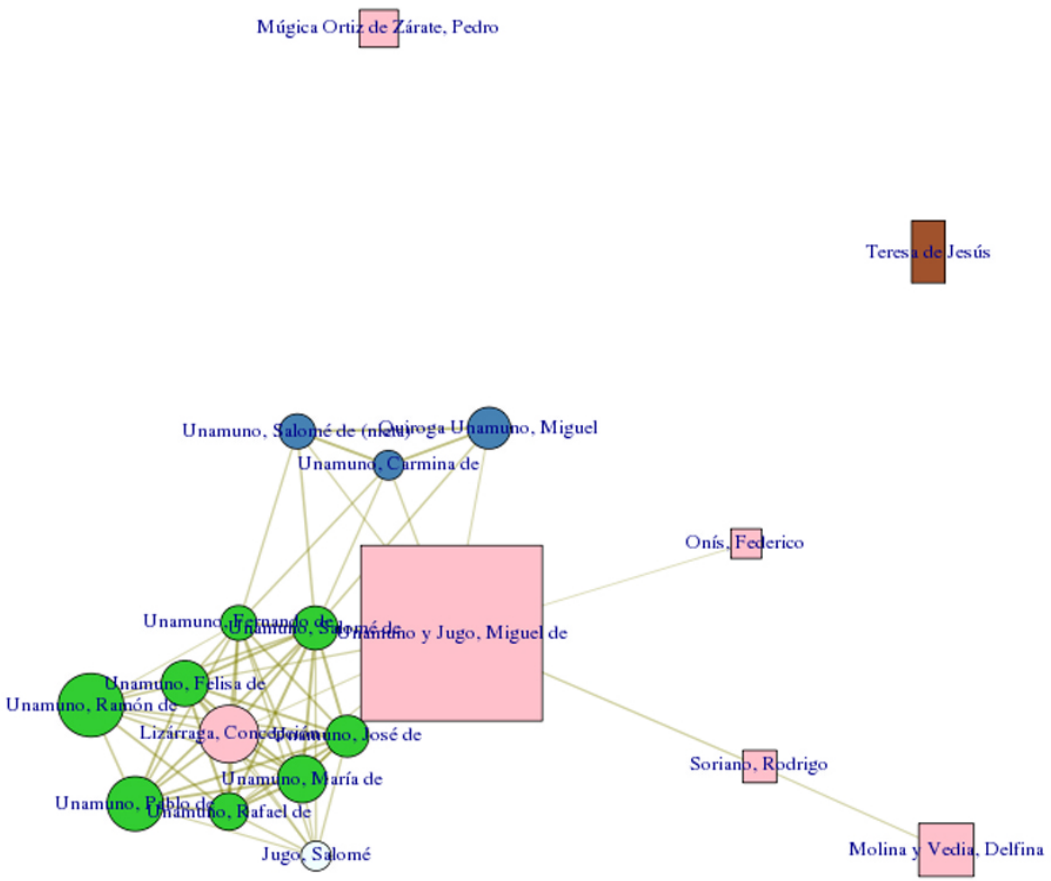
Optimal communities



Segundo ejemplo de Shiny

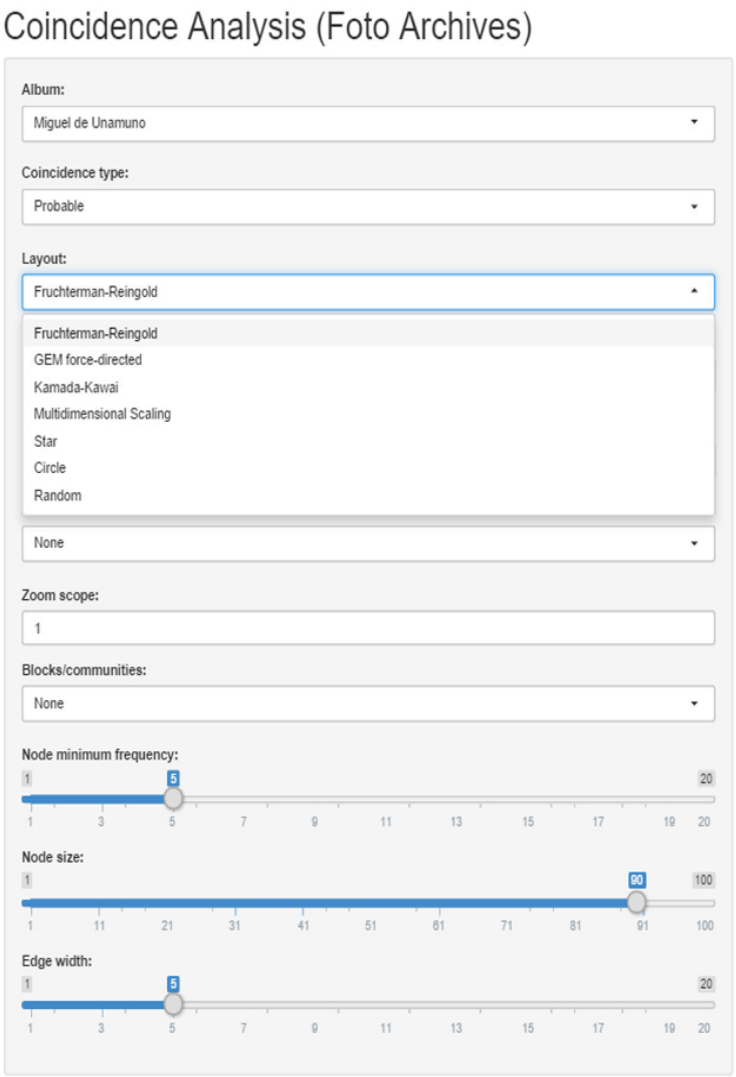
Unamuno (a): Tipo de coincidencias

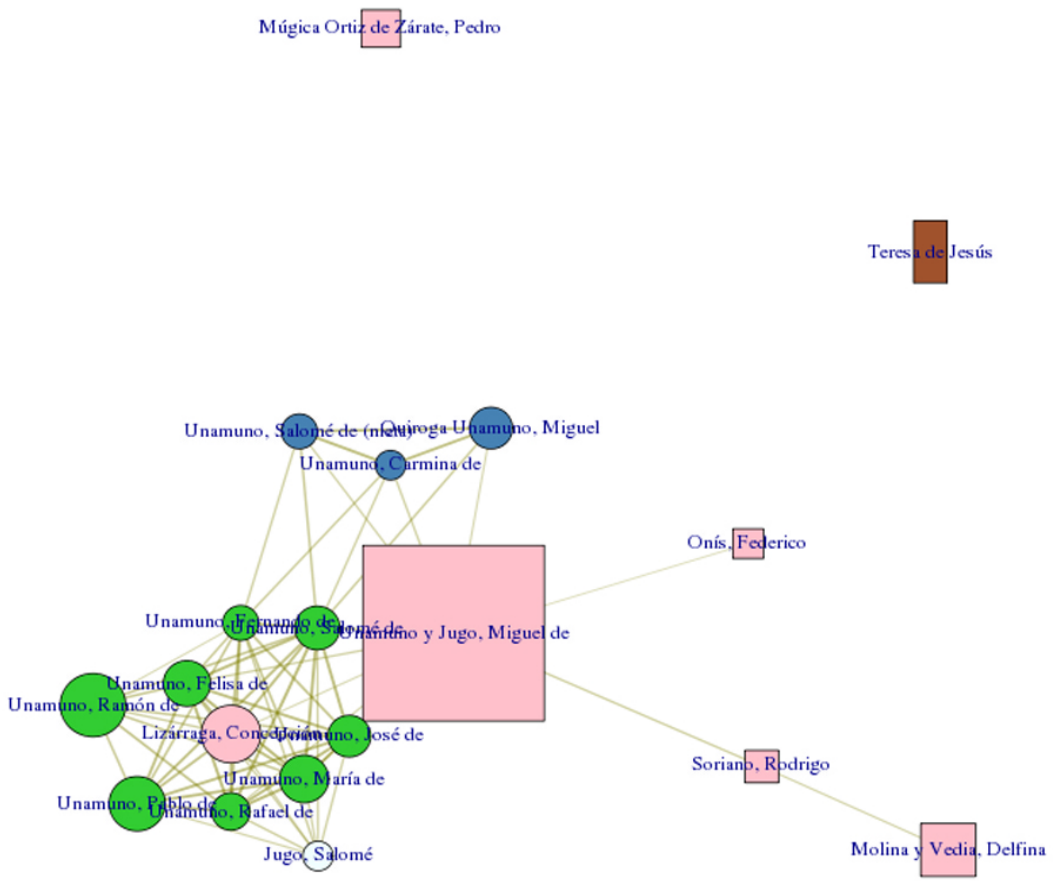




Segundo ejemplo de Shiny

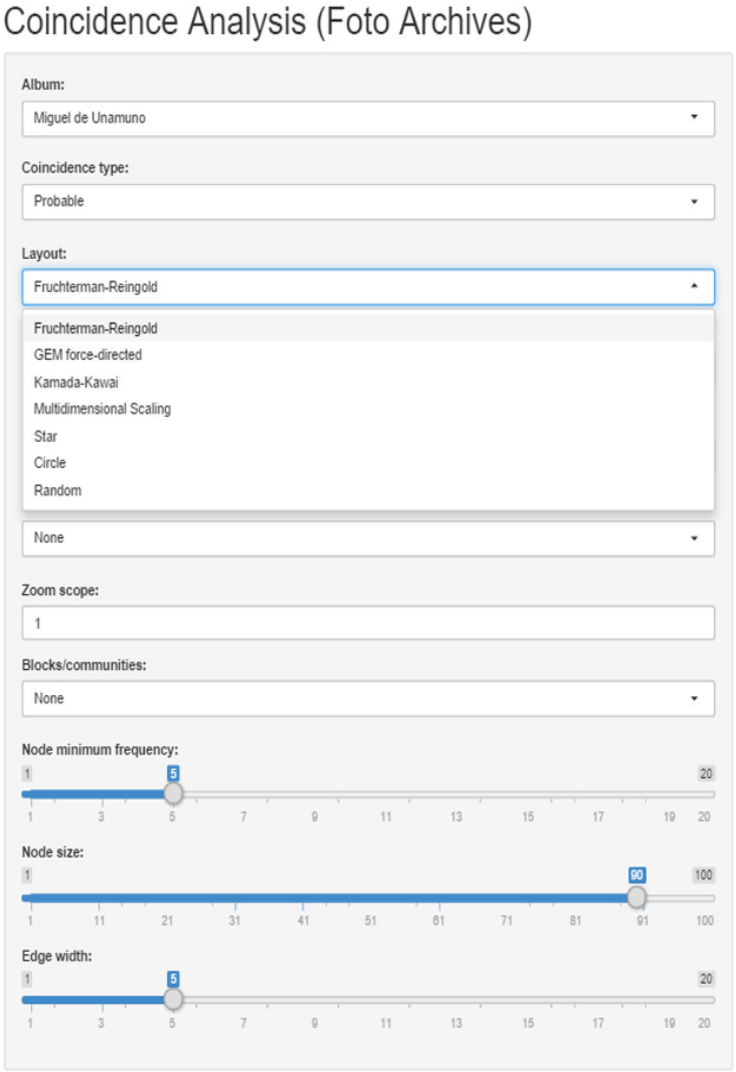
Unamuno (b): Disposición de los nodos

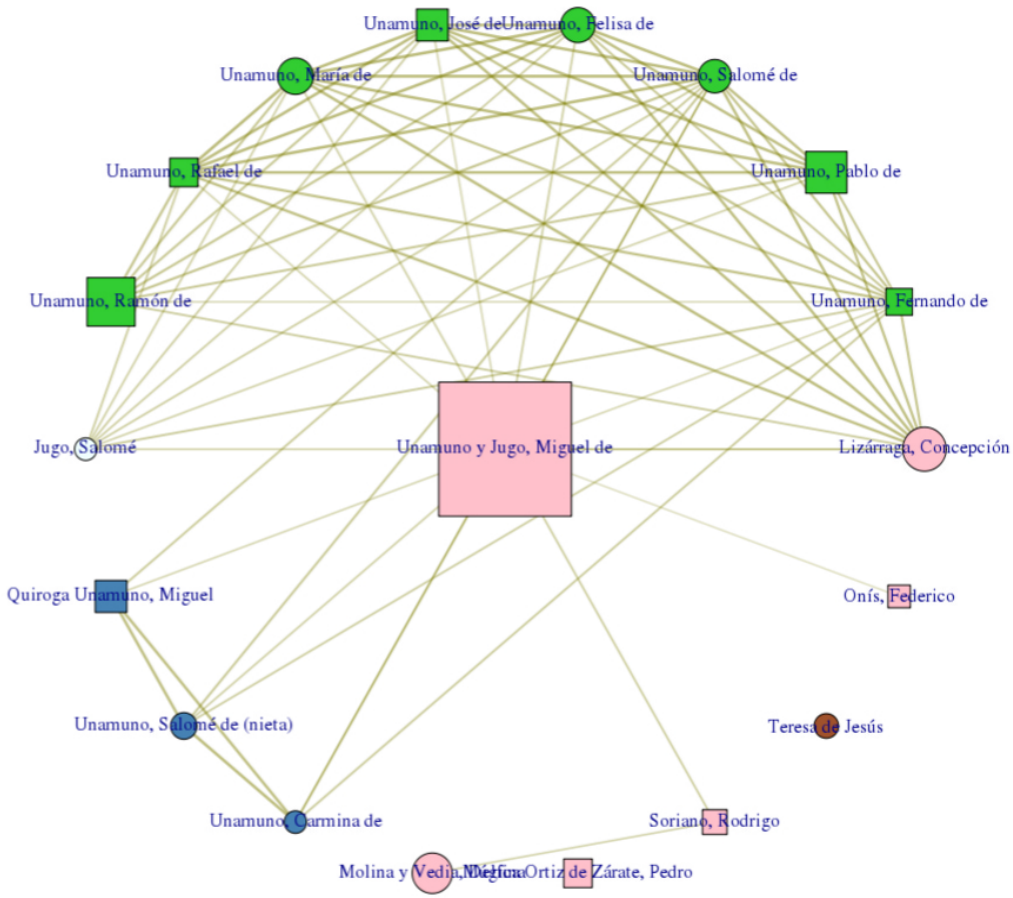




Segundo ejemplo de Shiny

Unamuno (c): Disposición estelar.





Tercer ejemplo de Shiny

L'Oreal (a): Colores por categorías, formas por tipos

Análisis de coincidencias (El poder de lo auténtico: L'Oreal)

Escenario:
Con género y medio

Grado de coincidencia:
Probable

Representación:
Fruchterman-Reingold

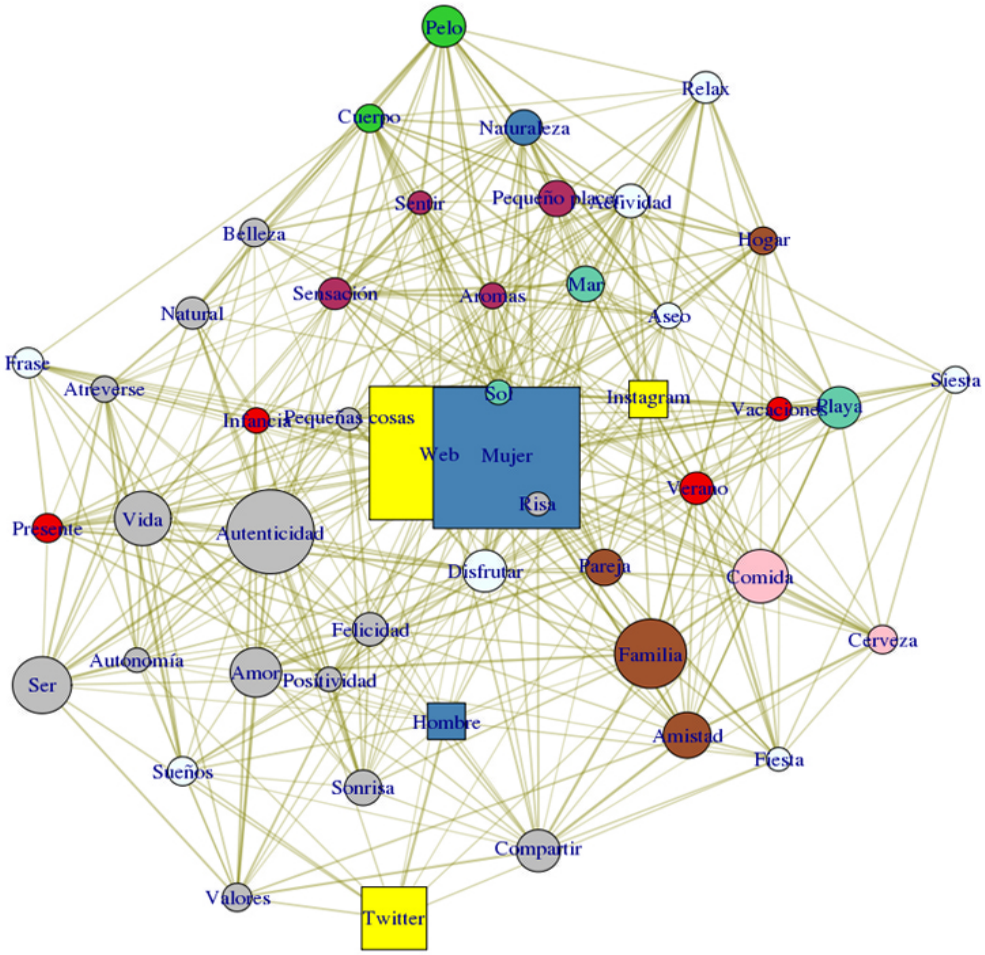
Color:
Categoría

Forma:
Tipo
Ninguno
Tipo

Frecuencia mínima categoría:
10 25 50
10 14 18 22 26 30 34 38 42 46 50

Tamaño de la categoría:
1 50 100
1 11 21 31 41 51 61 71 81 91 100

Grosor del vínculo:
1 5 20
1 3 5 7 9 11 13 15 17 19 20



Tercer ejemplo de Shiny

L'Oreal (b): Bloques recursivos

Análisis de coincidencias (El poder de lo auténtico: L'Oreal)

Escenario:
Con género y medio

Grado de coincidencia:
Probable

Representación:
Fruchterman-Reingold

Color:
Categoría

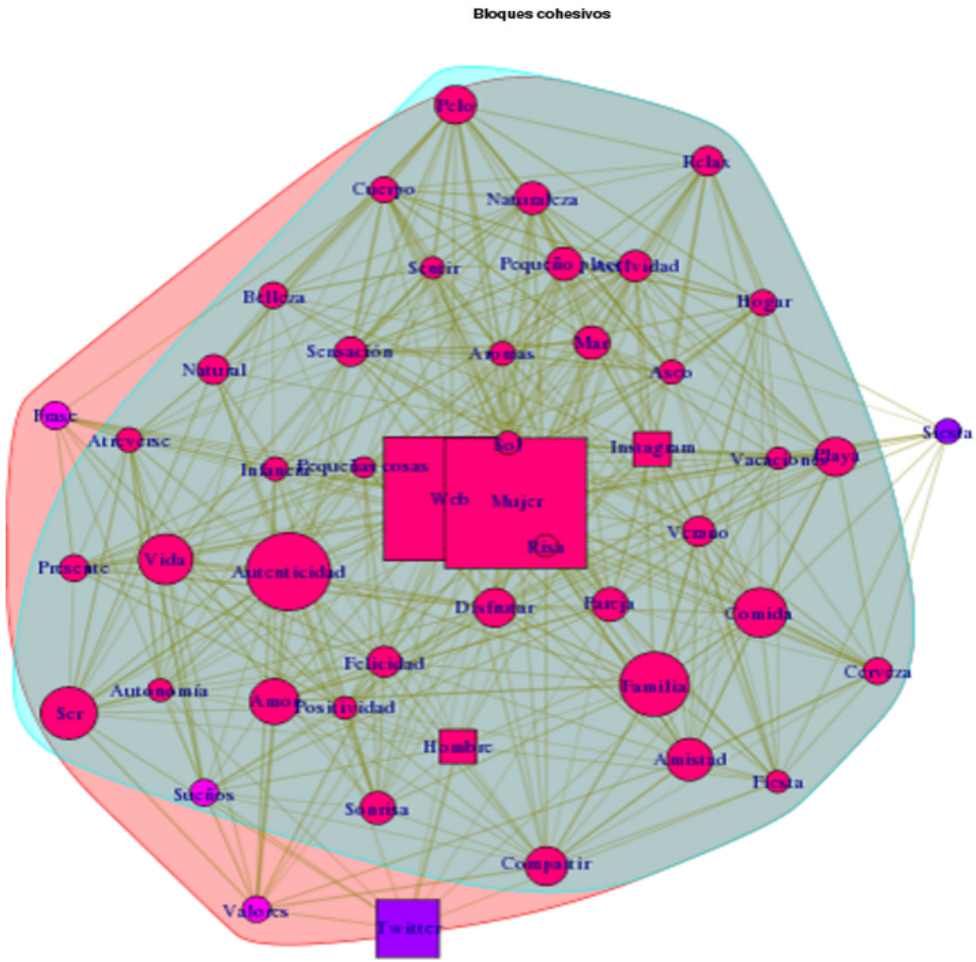
Forma:
Tipo

Bloques/comunidades:
Comunidades óptimas

- Ninguno
- Bloques
- Comunidades óptimas

Tamaño de la categoría:
1 50 100
1 11 21 31 41 51 61 71 81 91 100

Grosor del vínculo:
1 5 20
1 3 5 7 9 11 13 15 17 19 20



Tercer ejemplo de Shiny

L'Oreal (c): Comunidades

Análisis de coincidencias (El poder de lo auténtico: L'Oreal)

Escenario:
Con género y medio

Grado de coincidencia:
Probable

Representación:
Fruchterman-Reingold

Color:
Categoría

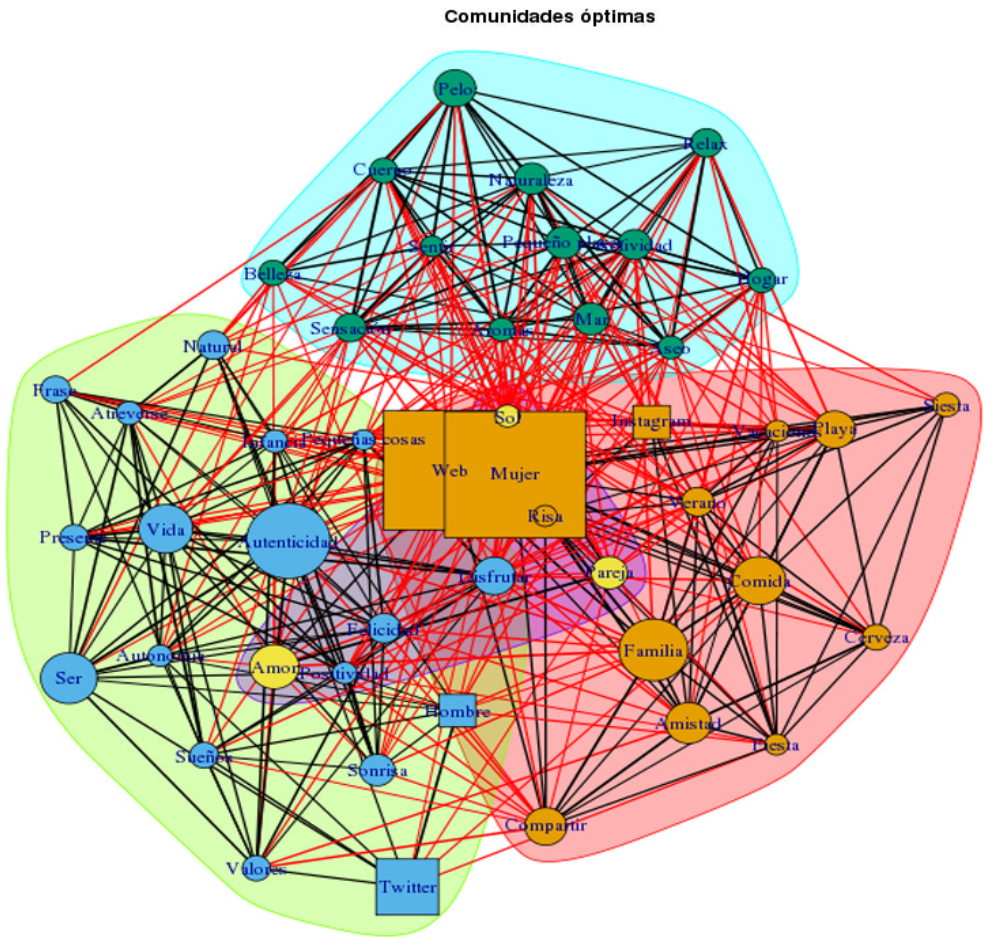
Forma:
Tipo

Bloques/comunidades:
Comunidades óptimas

- Ninguno
- Bloques
- Comunidades óptimas

Tamaño de la categoría:
1 50 100
1 11 21 31 41 51 61 71 81 91 100

Grosor del vínculo:
1 5 20
1 3 5 7 9 11 13 15 17 19 20



Índice

Próximos pasos

- Modelizar las coincidencias
  - Modelos log-lineales para estimar las frecuencias.
  - Modelos QAP y ERG para descubrir los determinantes de las coincidencias.
- Relacionar el análisis de coincidencias con otros modelos
  - Análisis comparado cualitativo.
  - Reglas de asociación.
  - Otros modelos de aprendizaje automático.
- Divulgación
  - Construcción y divulgación de un paquete para analizar coincidencias.
  - Mejorar la interatividad con Shiny.
  - Aplicar Shiny a otros problemas para divulgar los análisis estadísticos.

VNIVERSIDAD D SALAMANCA

¡Muchísimas gracias!

modesto@usal.es
lmartinez@march.es
martinez@march.es
berrocal@usal.es

Código de gcoin.

```
gcoin<-function(Data, variables, Characteristics=NULL, color="",
                shape="",minimum=5, p=.5, Bonferroni=FALSE, size=30, lwidth=5)
  require(igraph)
  D <- subset(Data, select=variables)
  Q=data.matrix(D)
  Q<-Q[,colSums(Q)>=minimum]
  N<-Haberman(Q)
  ifelse(Bonferroni,b<-ncol(N)*(ncol(N)-1)/2,b<-1)
  A<-Adjacency(N, p, b, nrow(D))
  G<-graph.adjacency(A, weighted=T, mode="undirected")
  G<-simplify(G)
  if (is.data.frame(Characteristics))
     V(G)$shape<-as.character(Characteristics[V(G),shape])
     V(G)$label<-as.character(Characteristics[V(G),"label"])
     V(G)$color<-as.character(Characteristics[V(G),color])

  V(G)$size<-colSums(data.matrix(Q))/max(colSums(data.matrix(Q)))*size
  egam<-log(E(G)$weight+1)/max(log(E(G)$weight+1))
  E(G)$color <- rgb(0.5, 0.5, 0, egam)
  E(G)$width <- egam*lwidth
  plot(G)
  return(G)
```

Código de Haberman y Adjacency.

```
Haberman<-function (Q, minimum=5)
  L<-colSums(Q)
  M<-crossprod(Q)
  n=nrow(Q)
  E<-tcrossprod(L)/n
  N<-((M-E)/sqrt(E))/sqrt(tcrossprod(1-(L/n)))
  return(N)
```

```
Adjacency<- function (N, p, b, n)
  A<-N
  A[(1-pt(N,n))>=(p/b)]<-0
  diag(A)<-0
  return(A)
```