# Interactive Graphs with Stata

M. Escobar (modesto@usal.es)  P. Cabrera (pablocal@usal.es)

C. Prieto (cprietos@usal.es)  D. Barrios (metal@usal.es)

University of Salamanca

2019 Spanish Stata Users Group meeting

Madrid, 17th October



# Presentation

## Aims

The aims of this presentation are:

- To show *network coincidence analysis*, which is a statistical framework to study concurrence of events.
- To present `coin`, an ado program that is able to perform this analysis.
- To show interactive graphs with Stata with the command `netcoin`.
- As an example, an analysis of people in the picture albums of an eminent character in the early $20^{\text{th}}$ century will be presented.
- This kind of representations can also be applied to
  - Social media analysis.
  - Content analysis of media and textbooks.
  - Multiresponse, glm and sem analysis in questionnaires.
  - Historical representation of eminent figures.

# Coincidence analysis

## Definition

- Coincidence analysis is a set of techniques whose object is to detect which people, subjects, objects, attributes or events tend to appear at the same time in different delimited spaces.
- These delimited spaces are called $n$ scenarios, and are considered as units of analysis ($i$).
- In each scenario a number of $J$ events $X_j$ may occur (1) or may not (0) occur.
- We call incidence matrix ($\mathbf{X}$) an $n \times J$ matrix composed by 0 and 1, according to the incidence or not of every event $X_j$.
- In order to make comparative analysis of coincidences, these scenarios may be classified in $H$ sets

# An example of incidences matrix

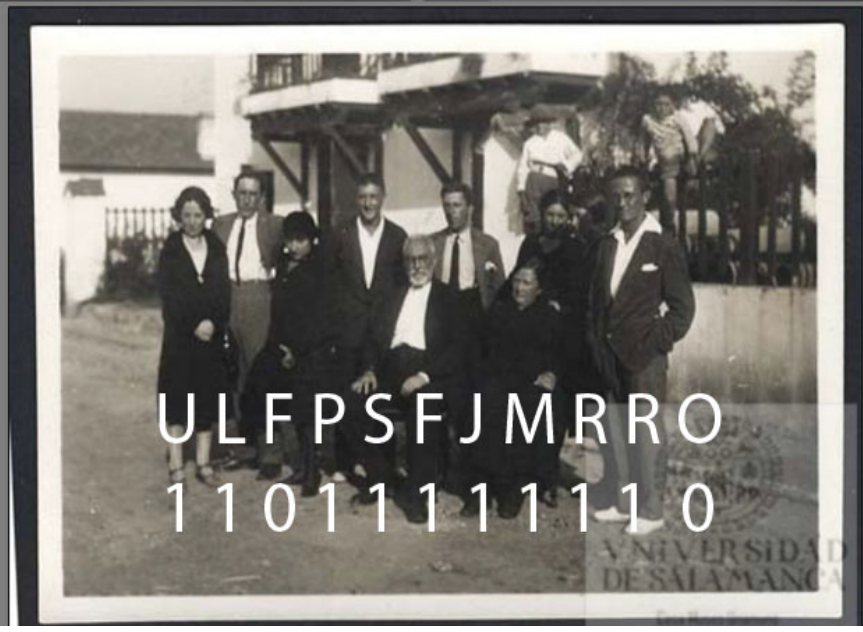
## Meeting the people



# An example of incidences matrix

## Coding the people



# Input of the analyses

## Incidences matrix (appearance or not appearance of 8 events in 4 scenarios)

The input of the analysis is a $\mathbf{X}$ matrix constructed with $i$ rows representing scenarios, and the $j$ columns representing events:

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

# Coincidences matrix

## Definition

- From the incidence matrix ($\mathbf{X}$), the coincidences matrix ($\mathbf{F}$) can be obtained by

$$\mathbf{F} = \mathbf{X}'\mathbf{X}$$

  - where each element $f_{jk}$ represents the number of scenarios where $X_j$ and $X_k$ are both 1, that is to say, the two events coincide.

- As may be imagined, there are special elements ($f_{jj}$) in the diagonal, which represent the number of incidences of $X_j$ in the $n$ scenarios.

# Example of coincidences matrix

## Coincidences matrix (co-appearances in the pictures)

The symmetric **F** matrix is compose by $i$ rows and $j$ columns representing incidences (diagonal) and coincidences of events:

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

# 3 grades of coincidence

## Mere and probable events

- Two events ( $X_j$ and $X_k$) are defined as 1) **merely** coincident if they occur in the same scenario at least once:

$$[\exists_i (x_{ij} = 1 \wedge x_{ik} = 1)] \vee f_{jk} \geq 1$$

- Additionally, two events ($X_j$ and $X_k$) are defined as 2) **conditionally** coincident if they occur more frequently than if they are independent:

$$f_{jk} > \frac{f_{jj} f_{kk}}{n}$$

# 3 grades of coincidence (cont.)

## Statistically probable events

- And two events are 3) **statistically conditional** if the joint frequency of their events meets one of the following inequalities:

$$\begin{aligned} \mathrm{P}(r_{jk} \leq 0) &< c \\ \mathrm{P}(\theta_{jk} \leq 1) &< c \\ \mathrm{P}(\mathrm{p}(X_j) - \mathrm{p}(X_j|X_k) \leq 0) &< c \end{aligned}$$

  - where $r_{jk}$ is the Haberman residual, $\theta_{jk}$ is the odd ratio, and the third equation represents a one tailed Fisher exact test. Furthermore, $c$ is the selected level of significance, normally 0.05)

- Haberman residuals ($r_{jk}$) with normal distribution may be used to assess statistically conditional events:

$$r_{jk} = \frac{f_{jk} - \frac{f_{jj} f_{kk}}{n}}{\sqrt{\frac{f_{jj} f_{kk} (n - f_{jj})(n - f_{kk})}{n^3}}}$$

# Graph

## Definition



- "A graph $\mathcal{G}$ consist of two sets of information: a set of Nodes (events), $\mathcal{N} = \{n_1, n_2, ..., n_g\}$, and a set of lines (adjacencies), $\mathcal{L} = \{l_1, l_2, ..., l_L\}$ between pair of nodes ". (Wasserman and Faust 1994).

# Adjacencies

## Elaboration of the adjacency matrices

- From the residual matrix, an adjacency $J \times J$ matrix $\mathbf{A}$ may be elaborated with all the elements equal to 0, but 1 in the case where $r_{jk}$ is significantly below the level $c$.

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

- By extension, other adjacency matrices can be elaborated following
  - The mere coincidence criterion

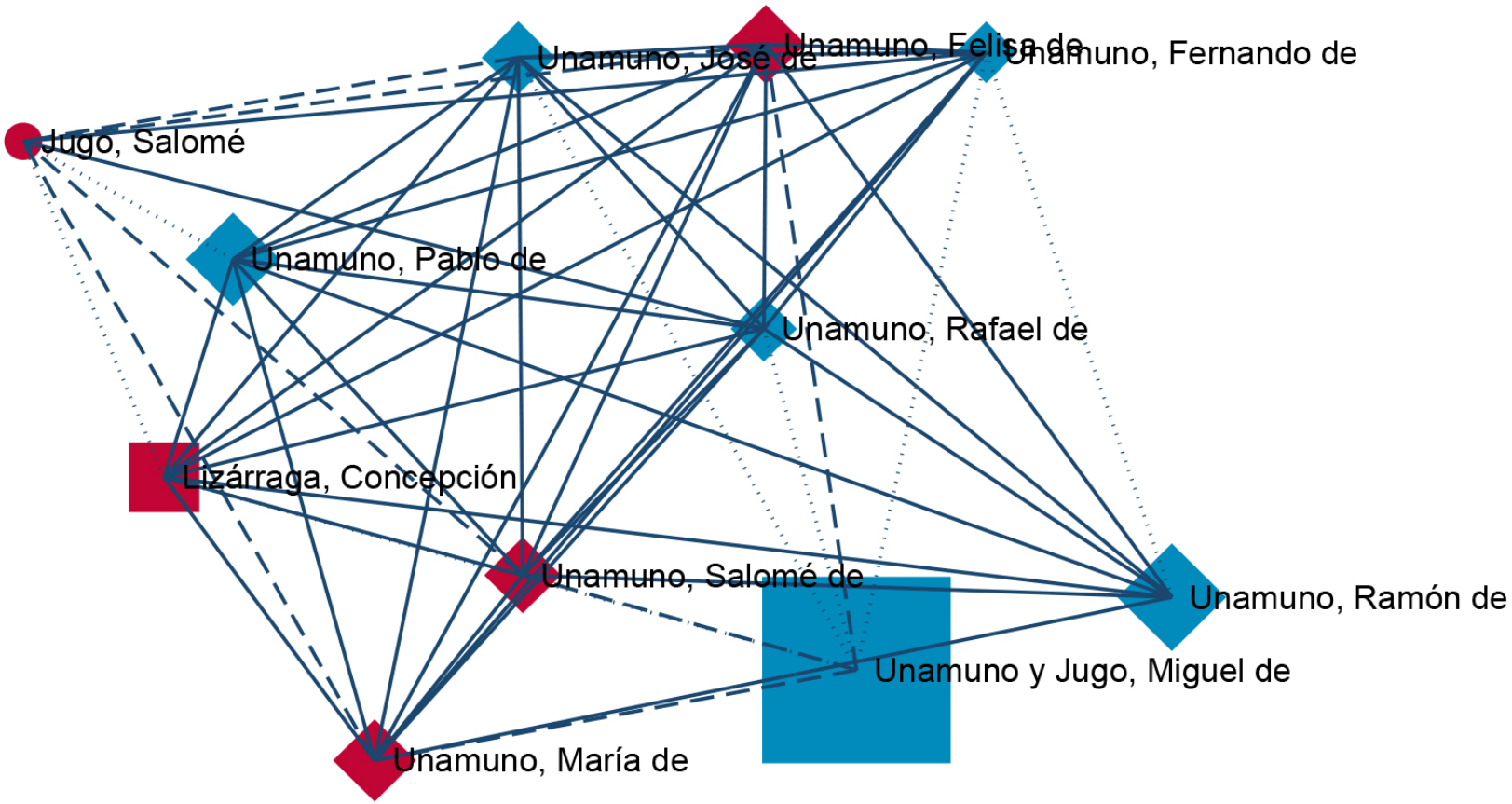
$$\mathbf{A}[j,k] = 1 \Leftrightarrow f_{jk} \geq 1$$

  - Or the conditional coincidence criterion

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

# Graph representation

## Fruchterman-Reingold layout




Fruchterman-Reingold coordinates

# Social network programs

## Stata program

- Stata has no tools for SNA.
- However, some advanced users have begun to write some routines. I wish to highlight the following works from which I have obtained insights:
  - Corten (2010) wrote a routine to visualize social networks [`netplot`].
  - Mihura (2012) created routines (SGL) to calculate networks centrality measures, including two Stata commands [`netsis` and `netsummarize`].
  - Afterwards, White (2013) presented a suite [`network`] of Stata programs for meta-analysis which includes the network graphs of Anna Chaimani in the UK. users group meeting.
  - And Grund (2013-2018, forthcoming) have presented a collection of programs to plot and analyze social networks [`nwcommands`].

# coin

## What is it?

- `coin` is an ado program in its development phase, which is capable of performing coincidence analysis.
- Its input is a dataset with scenarios as rows and events as columns.
- Its outputs are:
  - Different matrices (frequencies, percentages, residuals (3), distances, adjacencies and edges).
  - Several bar graphs, network graphs (circle, mds, pca, ca, biplot) and dendrograms (single, average, waverage, complete, wards, median, centroid).
  - Measures of centrality (degree, closeness, betweenness, information) (eigenvector and power)
  - Options to export to excel and .csv files.
- Its syntax is simple, but flexible. Many options such as output, bonferroni, p value, minimum, special event, graph controls, ...

# Command

## coin

`coin` *varlist* [ *if* ] [ *in* ] [ *weight* ] [ , *options* ]

Options can be classified into the following groups:

- **Outputs**: `f`, `g`, `v`, `h`, `e`, `r`, `s`, `n`, `ph`, `o`, `po`, `pf`, `t`, `a`, `d` , `l`, `c`, `all`, `x`, `xy`.
- **Controls**: head(*varlist*), variable(*varname*), ascending, descending, minimum (#), support(#), pvalue(#), levels(# # #), bonferroni, lminimum(#), iterations(#).
- **Plots**
  - Bar: bar, cbar(*varname*)
  - Graph: plot(circle|mds|ca|pca|biplot)
  - Dendrograms: dendrogram(single|complete|average|wards)

# Data examples

## Coincidences matrix of Unamuno's nuclear family

```
. coin Unamuno-Jugo, f
329 scenarios. 51 probable coincidences amongst 11 events. Density: 0.93. Components: 1.
11 events(n>=5): Unamuno Lizarraga Fernando Pablo Salome Felisa Jose Maria Rafael Ramon Jugo
```

| Frequencies | Una~o | Liz~a | Fer~o | Pablo | Sal~e | Fel~a | Jose | Maria | Raf~l | Ramon | Jugo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unamuno y Jugo, Migu~e | 176 | | | | | | | | | | |
| Lizárraga, Concepción | 12 | 19 | | | | | | | | | |
| Unamuno, Fernando de | 5 | 4 | 7 | | | | | | | | |
| Unamuno, Pablo de | 9 | 8 | 3 | 17 | | | | | | | |
| Unamuno, Salomé de | 9 | 8 | 3 | 7 | 11 | | | | | | |
| Unamuno, Felisa de | 10 | 9 | 2 | 8 | 8 | 12 | | | | | |
| Unamuno, José de | 7 | 8 | 3 | 8 | 7 | 7 | 10 | | | | |
| Unamuno, María de | 10 | 10 | 3 | 10 | 9 | 10 | 8 | 13 | | | |
| Unamuno, Rafael de | 6 | 6 | 3 | 7 | 7 | 7 | 6 | 8 | 8 | | |
| Unamuno, Ramon de | 5 | 4 | 1 | 5 | 5 | 5 | 4 | 5 | 5 | 23 | |
| Jugo, Salomé | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 5 |

# Data examples

## Haberman's residuals matrix of Unamuno's nuclear family

```
. coin Unamuno-Jugo, normalized
329 scenarios. 51 probable coincidences amongst 11 events. Density: 0.93. Components: 1.
11 events(n>=5): Unamuno Lizarraga Fernando Pablo Salome Felisa Jose Maria Rafael Ramon Jugo
```

| Haberman residuals | Una~o | Liz~a | Fer~o | Pablo | Sal~e | Fel~a | Jose | Maria | Raf~l | Ramon | Jugo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unamuno y Jugo, Migu~e | 18.1 | | | | | | | | | | |
| Lizárraga, Concepción | 0.9 | 18.1 | | | | | | | | | |
| Unamuno, Fernando de | 1.0 | 5.9 | 18.1 | | | | | | | | |
| Unamuno, Pablo de | -0.0 | 7.5 | 4.6 | 18.1 | | | | | | | |
| Unamuno, Salomé de | 1.9 | 9.7 | 5.9 | 8.9 | 18.1 | | | | | | |
| Unamuno, Felisa de | 2.1 | 10.5 | 3.6 | 9.8 | 12.4 | 18.1 | | | | | |
| Unamuno, José de | 1.1 | 10.2 | 6.2 | 10.9 | 11.9 | 11.4 | 18.1 | | | | |
| Unamuno, María de | 1.7 | 11.2 | 5.3 | 11.9 | 13.5 | 14.4 | 12.5 | 18.1 | | | |
| Unamuno, Rafael de | 1.2 | 8.5 | 7.0 | 10.7 | 13.4 | 12.8 | 12.0 | 14.1 | 18.1 | | |
| Unamuno, Ramón de | -3.2 | 2.5 | 0.8 | 3.7 | 5.1 | 4.8 | 4.2 | 4.5 | 6.2 | 18.1 | |
| Jugo, Salomé | -1.5 | 1.4 | 2.8 | 1.5 | 2.1 | 2.0 | 2.2 | 1.9 | 2.6 | -0.6 | 18.1 |

# Data examples

## Adjacency matrix from Haberman's residuals matrix

```
. coin Unamuno-Jugo, adjacencies
329 scenarios. 51 probable coincidences amongst 11 events. Density: 0.93. Components: 1.
11 events(n>=5): Unamuno Lizarraga Fernando Pablo Salome Felisa Jose Maria Rafael Ramon Jugo
```

| Adjacency matrix | Una~o | Liz~a | Fer~o | Pablo | Sal~e | Fel~a | Jose | Maria | Raf~l | Ramon | Jugo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unamuno y Jugo, Migu~e | 0 | | | | | | | | | | |
| Lizárraga, Concepción | 1 | 0 | | | | | | | | | |
| Unamuno, Fernando de | 1 | 1 | 0 | | | | | | | | |
| Unamuno, Pablo de | 0 | 1 | 1 | 0 | | | | | | | |
| Unamuno, Salomé de | 1 | 1 | 1 | 1 | 0 | | | | | | |
| Unamuno, Felisa de | 1 | 1 | 1 | 1 | 1 | 0 | | | | | |
| Unamuno, José de | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | | |
| Unamuno, María de | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | |
| Unamuno, Rafael de | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | |
| Unamuno, Ramón de | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| Jugo, Salomé | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

# Data examples

## Adjacency matrix from significant Haberman's residuals matrix

```
. coin Unamuno-Jugo, adjacencies pvalue(.05)
329 scenarios. 44 statistically probable(p<=.05) coincidences. Density: 0.80. Components: 1.
11 events(n>=5): Unamuno Lizarraga Fernando Pablo Salome Felisa Jose Maria Rafael Ramon Jugo
```

| Adjacency matrix | Una~o | Liz~a | Fer~o | Pablo | Sal~e | Fel~a | Jose | Maria | Raf~l | Ramon | Jugo |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Unamuno y Jugo, Migu~e | 0 | | | | | | | | | | |
| Lizárraga, Concepción | 0 | 0 | | | | | | | | | |
| Unamuno, Fernando de | 0 | 1 | 0 | | | | | | | | |
| Unamuno, Pablo de | 0 | 1 | 1 | 0 | | | | | | | |
| Unamuno, Salomé de | 1 | 1 | 1 | 1 | 0 | | | | | | |
| Unamuno, Felisa de | 1 | 1 | 1 | 1 | 1 | 0 | | | | | |
| Unamuno, José de | 0 | 1 | 1 | 1 | 1 | 1 | 0 | | | | |
| Unamuno, María de | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | | |
| Unamuno, Rafael de | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | | |
| Unamuno, Ramón de | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | |
| Jugo, Salomé | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 |

# Data examples

## Links list

```
. coin Unamuno-Jugo, list key(normalized) lminimum(10)

329 scenarios. 51 probable coincidences amongst 11 events. Density: 0.93. Components: 1.
11 events: Unamuno Lizarraga Fernando Pablo Salome Felisa Jose Maria Rafael Ramon Jugo

        N Edge
 -------- ---------------------------------------
    14.38 Unamuno, Felisa de <-> Unamuno, María de
    14.12 Unamuno, María de <-> Unamuno, Rafael de
    13.48 Unamuno, Salomé de <-> Unamuno, María de
    13.40 Unamuno, Salomé de <-> Unamuno, Rafael de
    12.81 Unamuno, Felisa de <-> Unamuno, Rafael de
    12.54 Unamuno, José de <-> Unamuno, María de
    12.43 Unamuno, Salomé de <-> Unamuno, Felisa de
    12.00 Unamuno, José de <-> Unamuno, Rafael de
    11.93 Unamuno, Pablo de <-> Unamuno, María de
    11.91 Unamuno, Salomé de <-> Unamuno, José de
    11.37 Unamuno, Felisa de <-> Unamuno, José de
    11.22 Lizárraga, Concepción <-> Unamuno, María de
    10.86 Unamuno, Pablo de <-> Unamuno, José de
    10.65 Unamuno, Pablo de <-> Unamuno, Rafael de
    10.47 Lizárraga, Concepción <-> Unamuno, Felisa de
    10.22 Lizárraga, Concepción <-> Unamuno, José de
```

# netcoin

## What is it?

- netcoin is a new ado command in its development phase, which is capable of create interactive graphs in html format.
- Its input is a dataset with scenarios as rows and events as columns.
- It can also use another dataset with the characteristics of the events
- Its output is an interactive graph in html format.
- Its syntax is very simple as it uses coin to calculate its statistics.

# Command

## netcoin

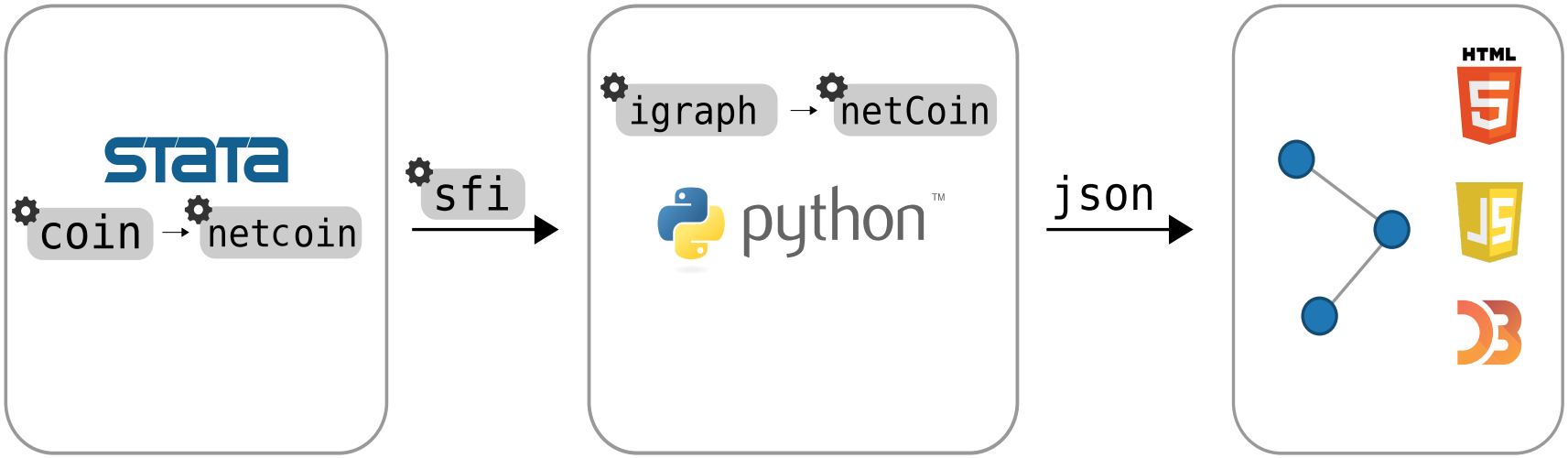
`netcoin` *varlist* [ *if* ] [ *in* ] [ *weight* ] [ `using` *filename* ] [ , *options* ]

Options can be classified into the following groups:

- **Controls**: `minimum(#)` `directory(`*dirname*`)` `language(en|es|ca)`
- **Outputs** (only if using): `name(`*varname*`)` `label(`*varname*`)` `size(`*varname*`)` `color(`*varname*`)` `shape(`*varname*`)` `image(`*varname*`)`

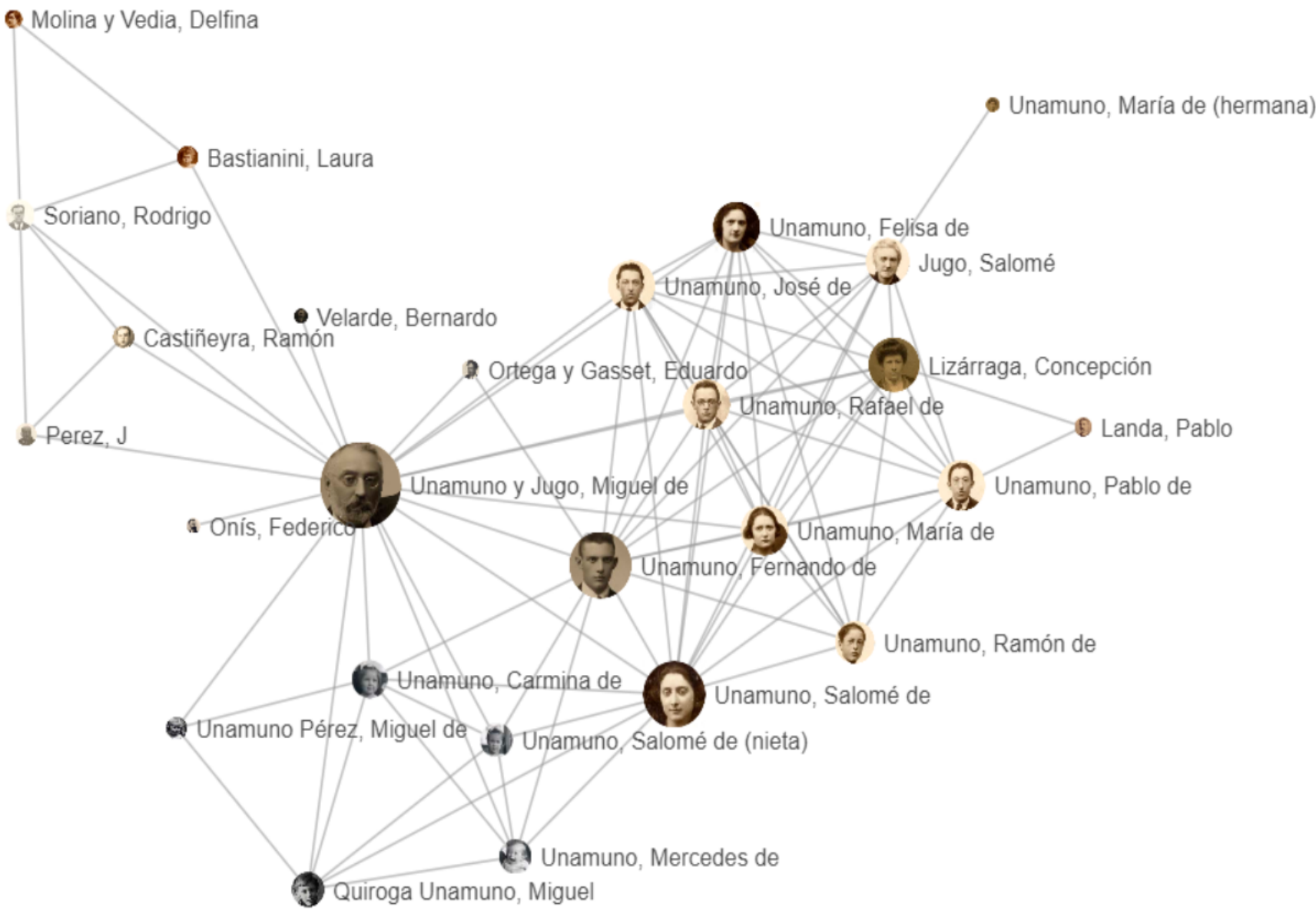
# Process

## From Stata to D3-JavaScript-html



# Output

## Network representation of Unamuno's family album





# Remarks

## About coincidence analysis

- I've proposed a manner of analyzing coincidences mixing different statistical tools.
- I think that the novelty of coincidence analysis is combining several techniques in order to represent data with interactive html graphs.
- This may be useful in analyzing dichotomous variables, but also to represent regressions, structural equation models and other networked graphs.
- I think that this approach could be extensively used with the aid of the `coin`, `precoin`, `netcoin` and other forthcoming programs.

# Availability of coin and netcoin

## Frame Subtitle

- If you are users of a version superior to the 11.2 of Stata, you can have a free copy of coin by typing:
  - net install coin, from(https://sociocav.usal.es/me/stata/)
- It is still a beta version, but it works reasonably well and it is being improved. It could be updated as follows:
  - adoupdate, update
- `netcoin` is more difficult to install as it requires Stata 16.0, Python and the `igraph` module.
- Comments and suggestions will be welcome!!

# Last slide

## Thanks



¡Gracias por la atención prestada!

modesto@usal.es