Interactive network regression graphs with Stata

Cristina Calvo (cristinacalvolopez@usal.es)

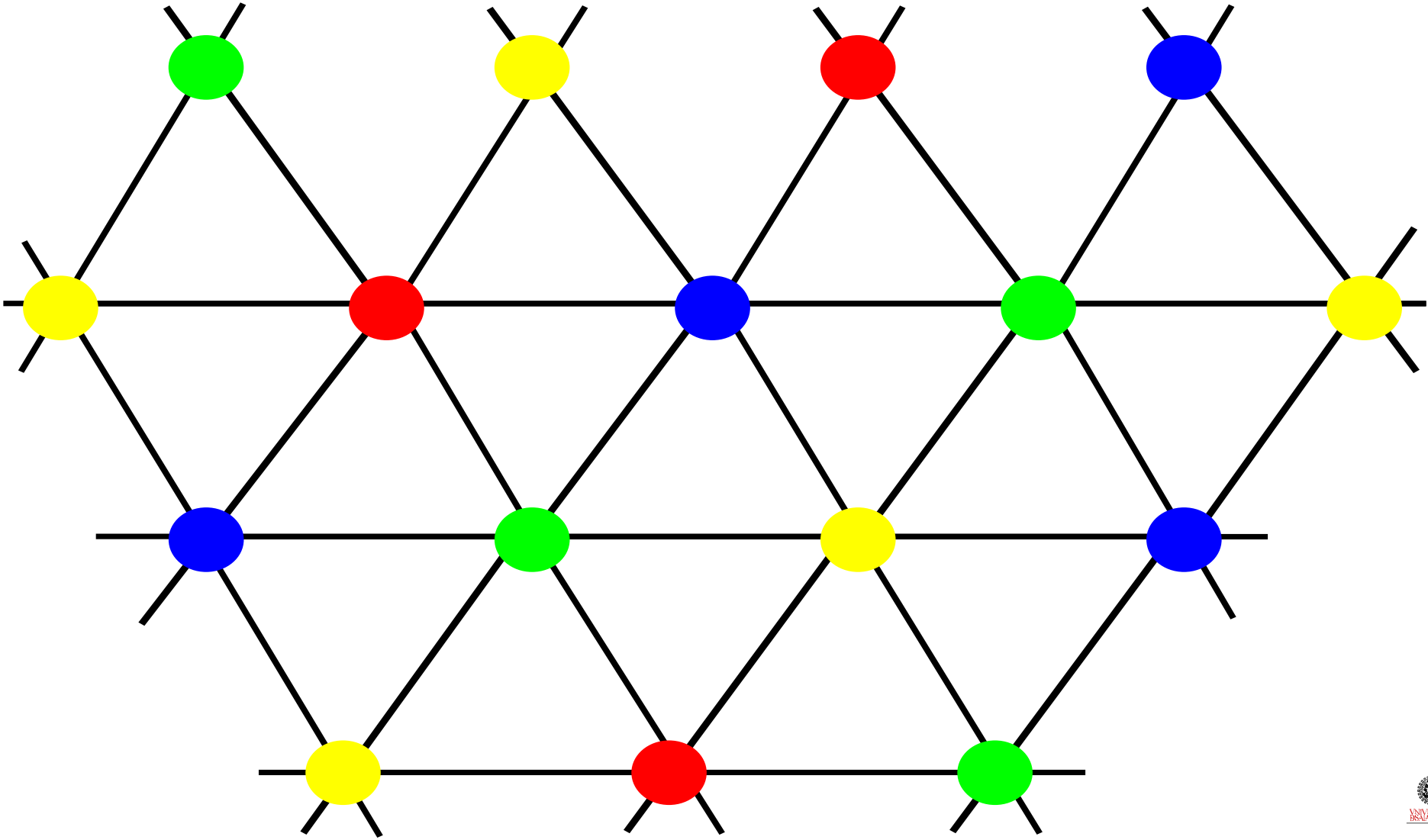
Departament of Sociology and Communication
G.A.S.

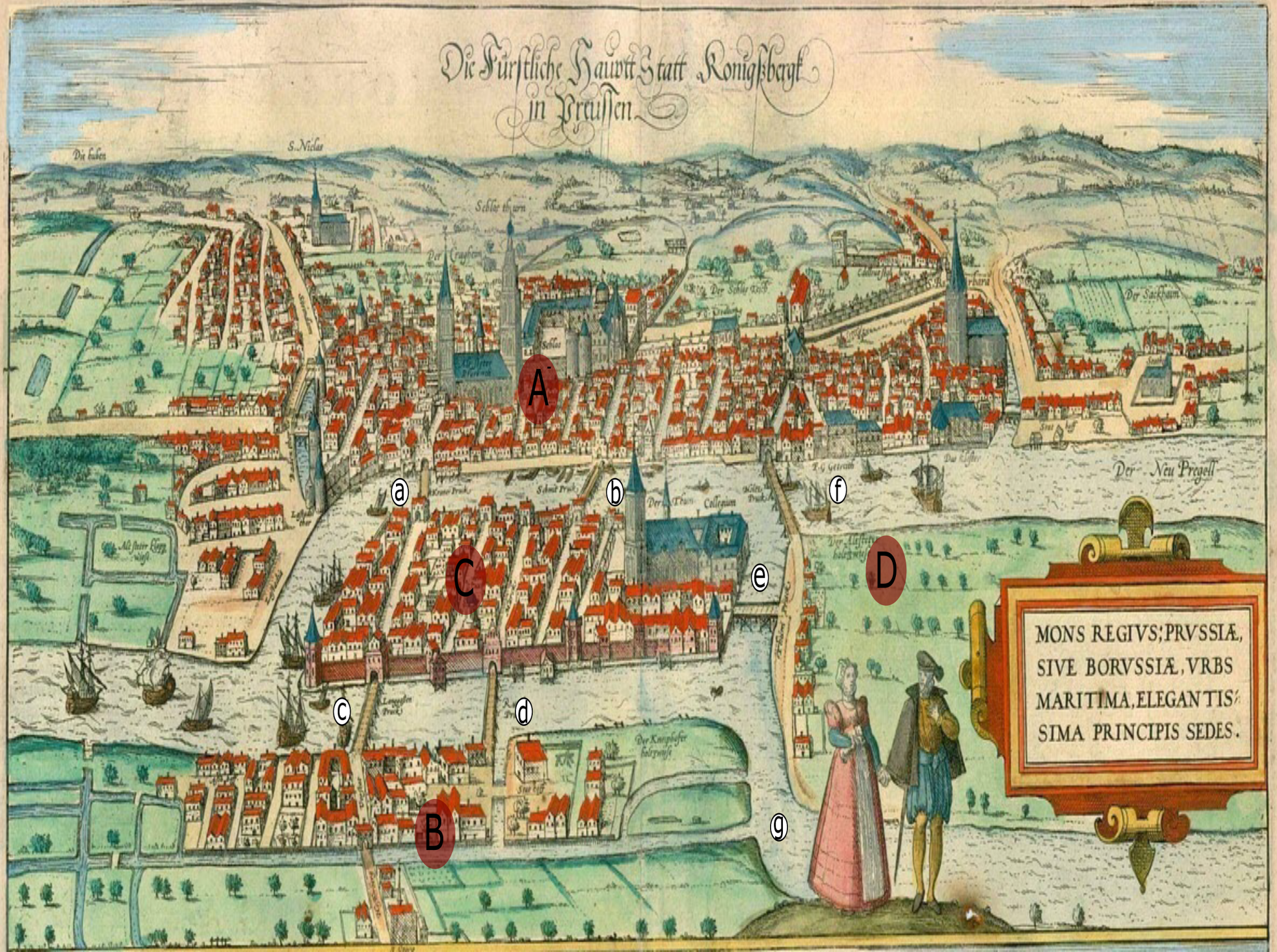
2023 Spanish Stata Conference
University of Salamanca

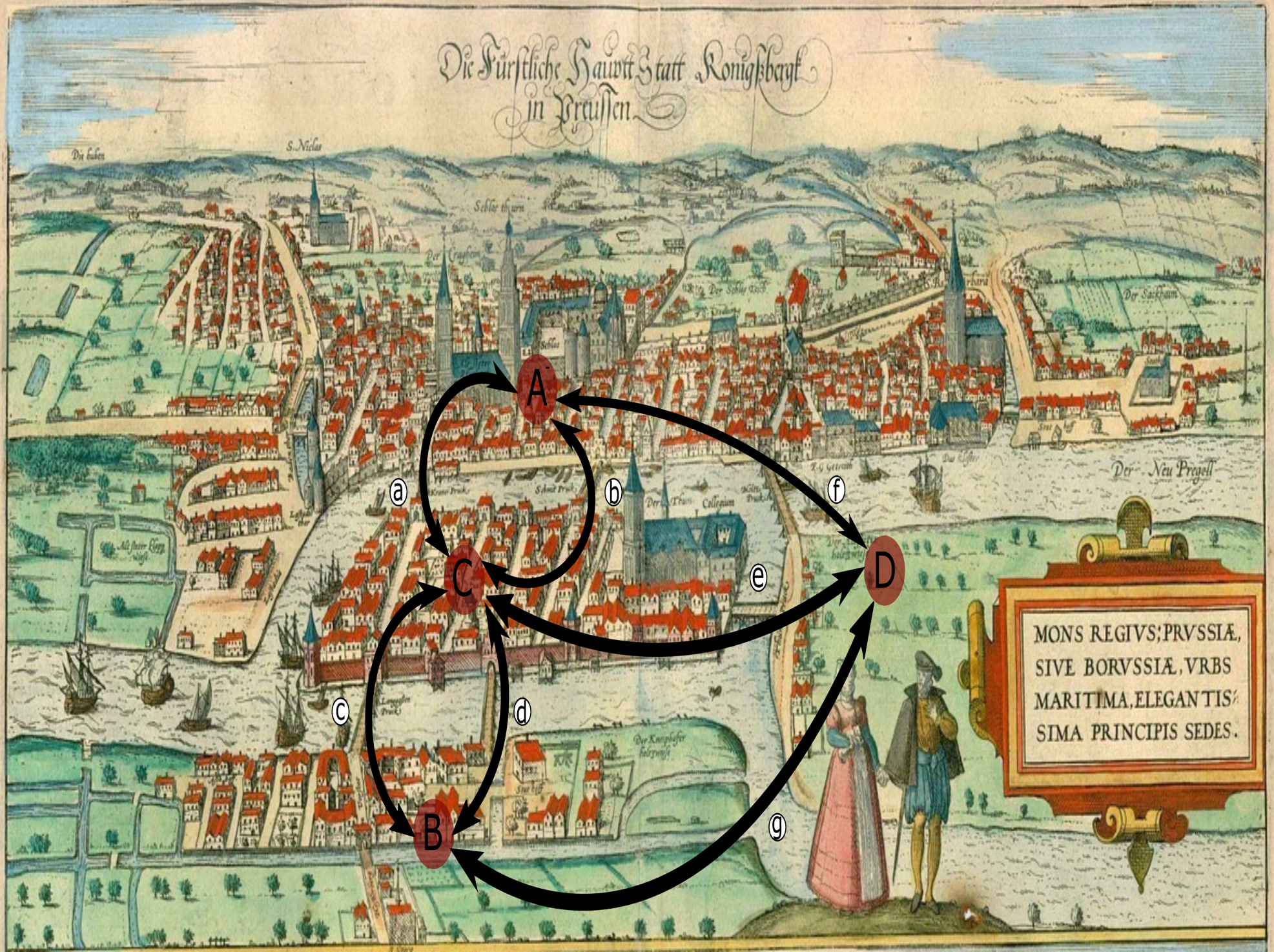
Madrid, 19 october de 2023

What is a network?

Set of points linked by lines




Die Fürstliche Hauptt Statt Konigsbergk in Preussen
S. Niclas
Schlos thurn
Der Sackheim
Der Neu Pregell
Der Thum
Collegium
A
B
C
D
a
b
c
d
e
f
g
MONS REGIVS; PRVSSIÆ, SIVE BORVSSIÆ, VRBS MARITIMA, ELEGANTISSIMA PRINCIPIS SEDES.


Die Fürstliche Hauptt Statt Konigsbergk in Preussen
S. Niclas
Schlos thurn
Der Sackheim
Der Neu Pregell
Der Thum Collegium
MONS REGIVS; PRVSSIÆ, SIVE BORVSSIÆ, VRBS MARITIMA, ELEGANTISSIMA PRINCIPIS SEDES.
A
B
C
D
a
b
c
d
e
f
g

Network graphs

Definition

- A graph $\mathcal{G}$ is a collection of points or vertices $x_1, x_2, ..., x_n$ (denoted by the set $\mathcal{N}$), and a collection of lines $l_1, l_2, ..., l_m$ (denoted by the set $\mathcal{L}$) joining all or some of these points. Graph $\mathcal{G}$ is then fully described and denoted by the doublet $(\mathcal{N},\mathcal{L})$. (Christofides, 1975)

Network graphs

Operationalization

- This doublet $(\mathcal{N},\mathcal{L})$ can be represented just by a $n \times n$ matrix $\mathbf{C}$ whose elements m_{jk} represent the (strength of) connection of point x_j to point x_k.
- There are, however, two other ways of storing graphs:
  - Adjacency list where n rows are points, and columns are only neighboring points
  - Edge list where rows are m connections with a first column indicating the source point and a second indicating the target point. (Mihura, 2011:7)

Network input structures

Disposition examples

Adjacency matrix

	x_1	x_2	x_3	x_4
	A	B	C	D
x_1 A	–			
x_2 B	0	–		
x_3 C	2	2	–	
x_4 D	1	1	1	–

Adjacency list

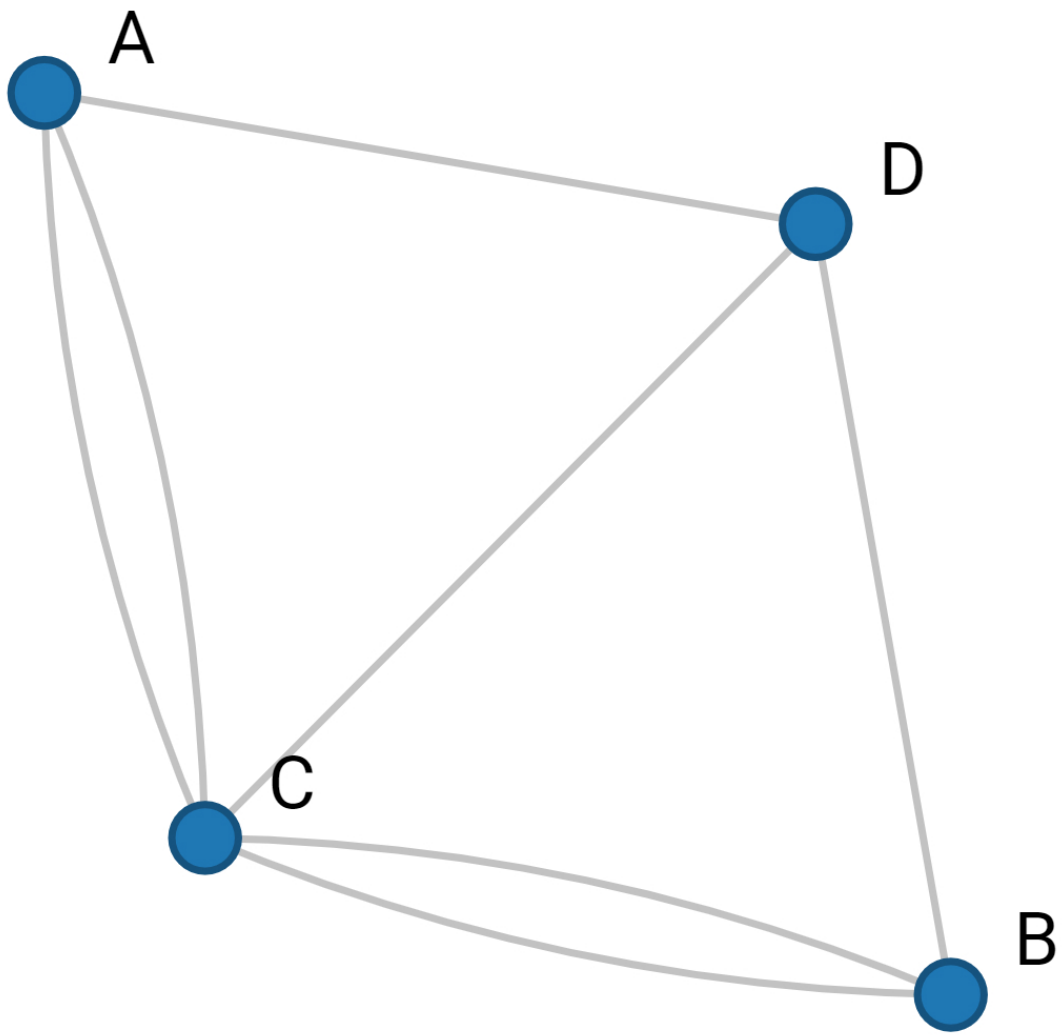
Nodes	Neighbors
A	C C D
B	C C D
C	D

Edge list

Source	Target
A	C
C	A
B	C
C	B
C	D
A	D
B	D

Graph representation

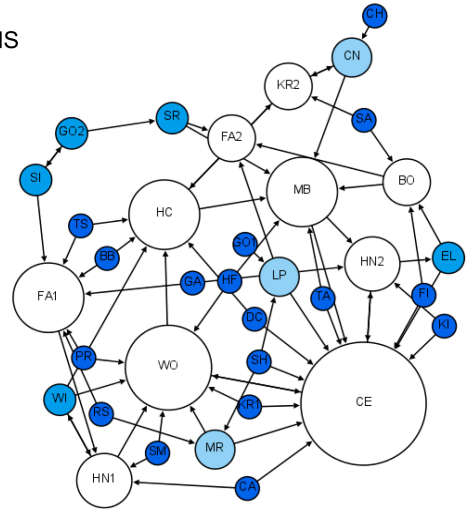
Its elements (A, B, C, D) and its links (AC AC BC BC CD AD BD) represented



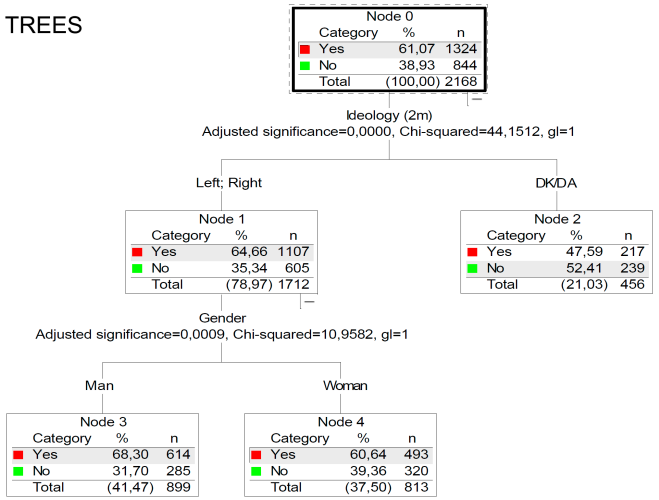
Graphs in social research

Examples of graphs

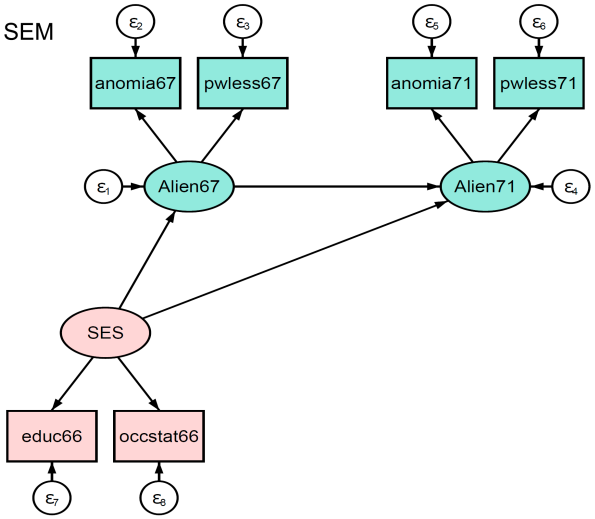
GRAPHS



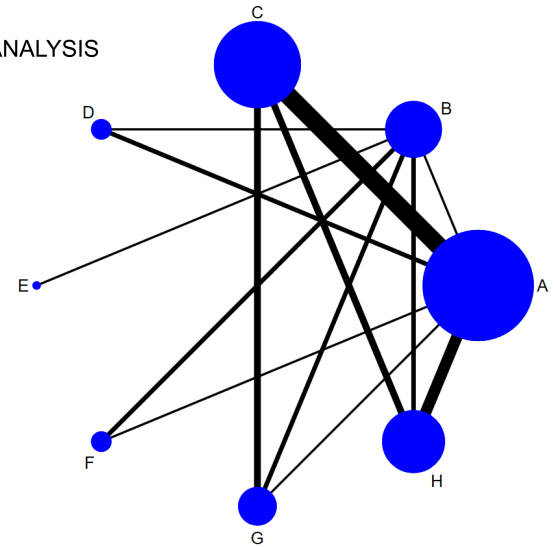
TREES



SEM



META-ANALYSIS



Coincidence analysis

Definition

- Coincidence analysis is a set of techniques whose object is to detect which J categories, 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.
- This incidence matrix is converted into a $J \times J$ coincidence matrix ($\mathbf{C}$) in whose cells appear frequencies or normalized residuals, which, if significant, represent the adjacencies of the graphs among the events or categories.

coin

What is it?

- `coin` is an ado program published in the *Stata Journal*, 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)

coin

Coincidence matrix

Table: Coincidence matrix

```
. coin i.gender i.ager i.ideology i.religion i.intencionr, f
22777 scenarios. 56 probable coincidences amongst 17 events. Density: 0.41. Components: 1.
17 events(n>=5): 1.gender 2.gender 1.ager 2.ager 3.ager 4.ager 1.ideology 2.ideology 3.ideology 1.religion 2.religion 3.religion 1.intencionr
> onr 8995.intencionr
```

Frequencies	1. gen~r	2. gen~r	1. ager	2. ager	3. ager	4. ager	1. ide~y	2. ide~y	3. ide~y	1. rel~n	2. rel~n	3. rel~n	1. in~nr	2. in~nr	3. in~nr	21. in~nr	8995 in~nr
Gender																	
Male	11966																
Female	0	10811															
ager																	
18-34	2194	1768	3962														
35-49	3372	2826	0	6198													
50-64	3647	3503	0	0	7150												
>65	2753	2714	0	0	0	5467											
Ideology																	
Left	5109	4921	2031	2663	3060	2276	10030										
Center	3571	2619	807	1852	2174	1357	0	6190									
Right	3286	3271	1124	1683	1916	1834	0	0	6557								
Religion																	
Practicing	1828	2470	423	917	1233	1725	736	1189	2373	4298							
Non-practicing	4577	4191	1143	2355	3145	2125	2832	2999	2937	0	8768						
Others	5561	4150	2396	2926	2772	1617	6462	2002	1247	0	0	9711					
Voting intention																	
PSOE	3251	3745	1041	1474	2294	2187	5240	1253	503	840	2696	3460	6996				
PP	3737	3651	980	1941	2433	2034	236	3185	3967	2412	3570	1406	0	7388			
VOX	1707	809	616	923	672	305	69	715	1732	675	1188	653	0	0	2516		
Sumar	1872	1615	872	1151	976	488	3116	230	141	125	552	2810	0	0	0	3487	
Others	1399	991	453	709	775	453	1369	807	214	246	762	1382	0	0	0	0	2390

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

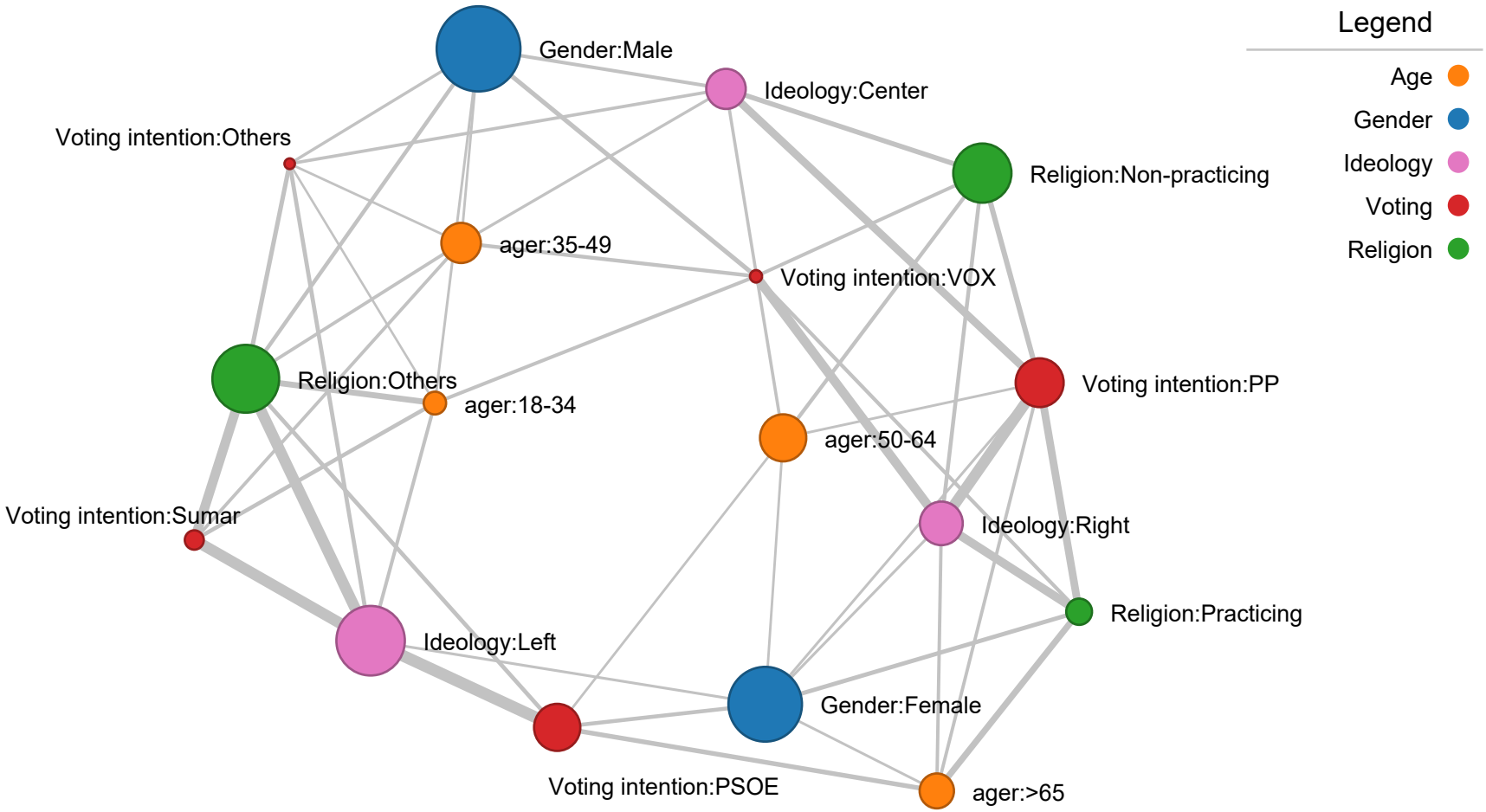
`netcoin` *varlist* [*if*] [*in*] [*weight*] [`using` *filename*] [`,`*options*]

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*`)`

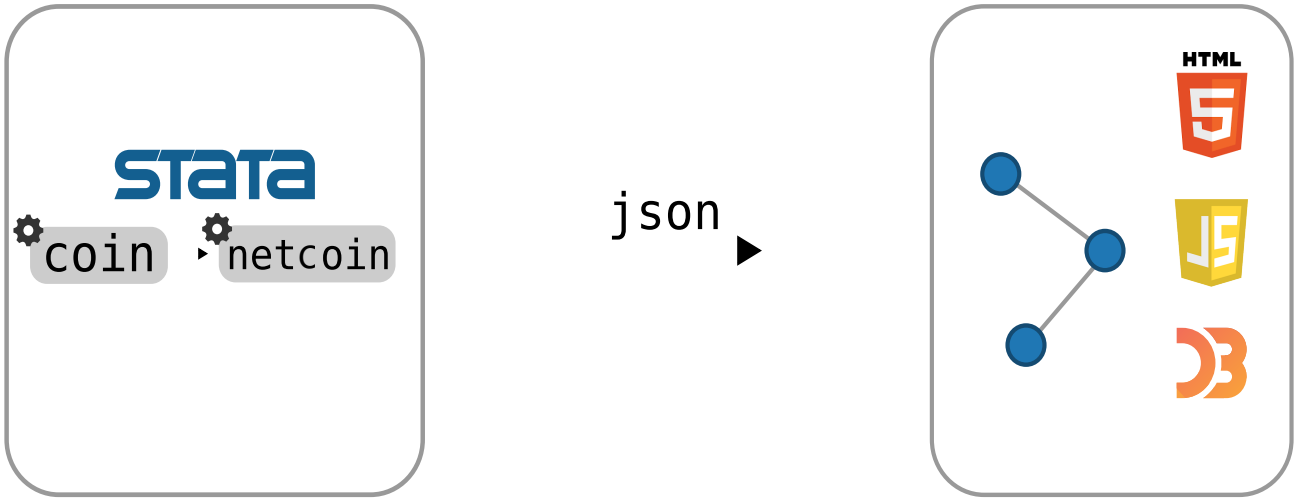
netCoin

Graph



Process

From Stata to D3-JavaScript-html



From coincidences to regressions

Differences between them

- In the analysis of coincidences we have seen how adjacencies between categories of variables were determined by the standardized residuals. Moreover, all categories have a similar status: they are categories and there is no difference between dependent and dependent.
- In regression analysis we find a dependence model in which there are dependent variables (or categories) and independent variables (or categories) due to the fact that these variables can be numerical or categorical (factor variables).
- Therefore the two most important differences between coincidence and regression are:
  1. The model of the relationship between the variables must be specified in advance in case of regression.
  2. Instead of standardized residuals, the marginal effects with their corresponding one-sided significance will be used.

Regression procedure

With marginal effects

- A way to make regression graphs would be through marginal effects.
- Marginal effects in regression refer to the change in the dependent or response variable for a small change in one of the independent variables, holding other variables constant.
- An ado has been written to obtain these marginal effects while generating two matrices: that of nodes (variables or categories) and that of links (marginal effects) positive and significant.
- These matrices are stored in the return list, and are the input for the `netreg` command.

Multiple regression

p(PP) on gender, age, religion and ideology

Table: Multiple regression of voting PP on gender, age, religion and ideology

```
. regress pPP i.gender age i.religion i.ideology
```

Source	SS	df	MS
Model	169294.931	6	28215.8219
Residual	228208.808	27,798	8.20954055
Total	397503.739	27,804	14.2966386

Number of obs	=	27,805	
F(6, 27798)	=	3436.96	
Prob > F	=	0.0000	
R-squared	=	0.4259	
Adj R-squared	=	0.4258	
Root MSE	=	2.8652	

pPP	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
gender						
Female	.0654696	.0346946	1.89	0.059	-.0025335	.1334727
age	-.0056908	.0010927	-5.21	0.000	-.0078326	-.003549
religion						
Non-practicing	-.6920244	.0502462	-13.77	0.000	-.7905095	-.5935393
Others	-1.840955	.0538855	-34.16	0.000	-1.946573	-1.735337
ideology						
Center	3.344857	.0423155	79.05	0.000	3.261916	3.427797
Right	4.890855	.0466369	104.87	0.000	4.799445	4.982266
_cons	2.696335	.085501	31.54	0.000	2.528748	2.863921

Regression general mean contrasts

p(PP) on gender, age, religion and ideology

Table: Contrasts (marginal effects) of voting PP on gender, religion and ideology

```
. contrast gw.gender gw.religion gw.ideology, nowald
Contrasts of marginal linear predictions
```

	Contrast	Std. err.	[95% conf.	interval]
gender				
(Male vs mean)	-.0315563	.0167228	-.0643338	.0012211
(Female vs mean)	.0339133	.0179718	-.0013124	.0691389
religion				
(Practicing vs mean)	1.067551	.0393402	.9904426	1.14466
(Non-practicing vs mean)	.375527	.0224017	.3316186	.4194354
(Others vs mean)	-.7734034	.0218841	-.8162973	-.7305096
ideology				
(Left vs mean)	-2.296222	.0214	-2.338167	-2.254277
(Center vs mean)	1.048635	.0261675	.9973453	1.099925
(Right vs mean)	2.594634	.0308719	2.534123	2.655144

Command

dime

`dime sureg` (*depvar1 varlist1*) (*depvar2 varlist2*) ... (*depvarN varlistN*) [*if*] [*in*] [*weight*] [, *options*]

It has three kinds of options:

- vce(vcetype) specifies the type of standard error reported.
- export(filename.suffix) to export the table to a file.
- graph, pvalue, bonferroni, and linkbipolar are for network graphs.

`dime mlogit` *depvar varlist* [*if*] [*in*] [*weight*] [, *options*]

It as the same options as `dime sureg`.

Regression marginal effects

p(PP) on gender, age, religion and ideology

Table: Margins and marginal effects of voting PP

```
. dime sureg (pPP i.gender age i.religion i.ideology)
Table of marginals and global mean differences
```

	Total		PP	
Total	(27,805)	3.7		
Male	(14,403)	3.6	-0.0	
Female	(13,402)	3.7	0.0	
Age	(29,201)	3.7	-0.0	***
Practicing	(5,067)	4.7	1.1	***
Non-practicing	(10,598)	4.0	0.4	***
Others	(12,140)	2.9	-0.8	***
Left	(12,056)	1.4	-2.3	***
Center	(8,525)	4.7	1.0	***
Right	(7,224)	6.3	2.6	***
R2		0.43		

*** p<.001, ** p<.01, * p<.05

Regression marginal effects

p(PP PSOE VOX Sumar) on gender, age, religion and ideology

Table: Margins and marginal effects of p(voting)

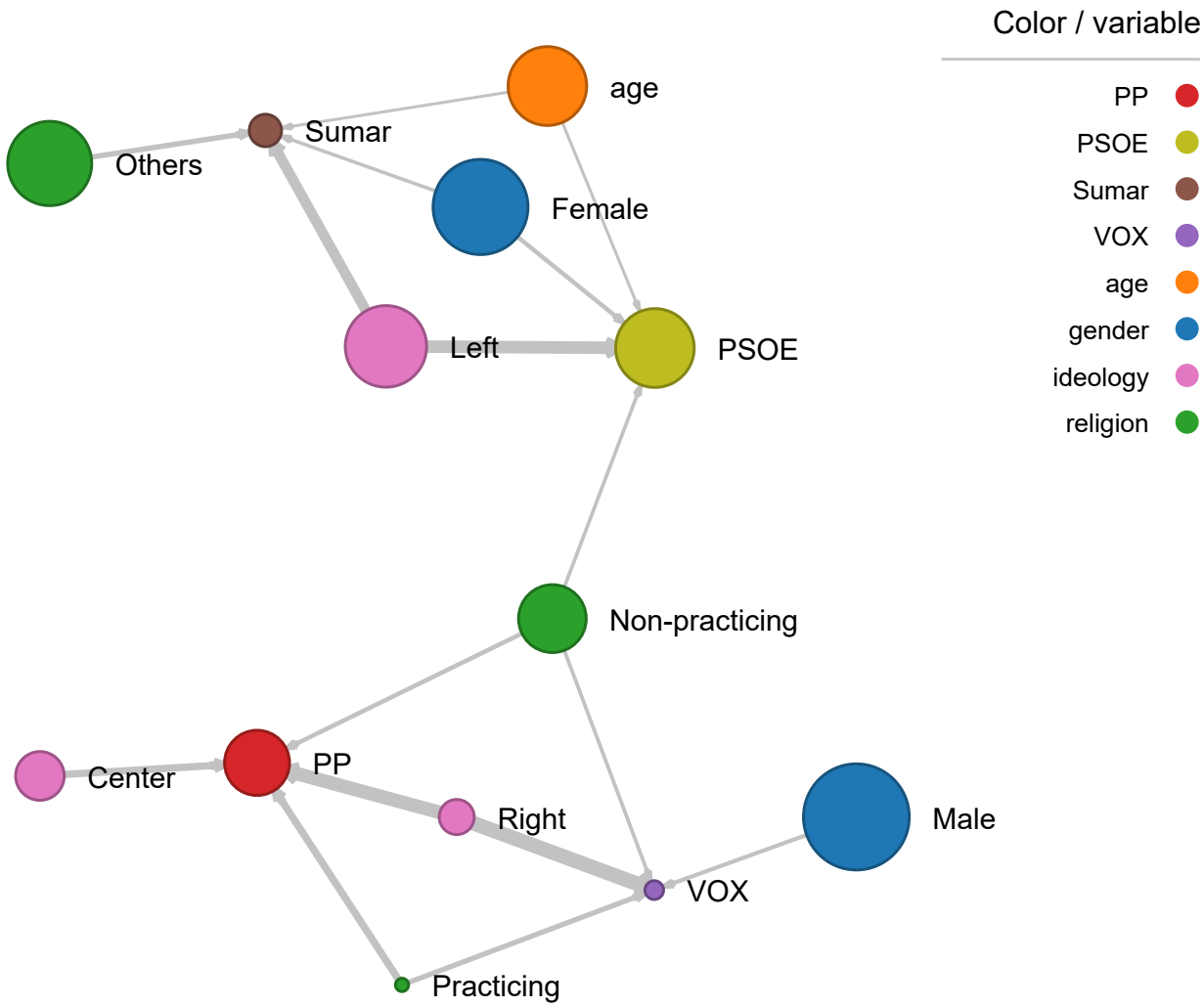
```
. dime sureg (pPP i.gender age i.religion i.ideology)(pPSOE)(pVOX)(pSumar), all graph
      export(regresion.xlsx, replace)
Table of marginals and global mean differences
```

	Total		PP		PSOE		VOX		Sumar	
Total	(27,713)	3.7		4.2		2.0		2.5		
Male	(14,364)	3.6	-0.0	3.9	-0.3 ***	2.3	0.3 ***	2.4	-0.1 ***
Female	(13,349)	3.7	0.0	4.6	0.4 ***	1.7	-0.3 ***	2.7	0.1 ***
Age	(29,201)	3.7	-0.0 ***	4.3	0.0 ***	2.0	-0.0 ***	2.5	0.0 ***
Practicing	(5,027)	4.7	1.1 ***	3.9	-0.4 ***	2.6	0.5 ***	1.9	-0.6 ***
Non-practicing	(10,569)	4.0	0.4 ***	4.4	0.2 ***	2.2	0.2 ***	2.1	-0.4 ***
Others	(12,117)	2.9	-0.8 ***	4.2	-0.0	1.6	-0.4 ***	3.1	0.6 ***
Left	(12,029)	1.4	-2.3 ***	6.6	2.3 ***	0.3	-1.7 ***	4.3	1.8 ***
Center	(8,505)	4.7	1.1 ***	3.2	-1.1 ***	2.0	0.0	1.3	-1.3 ***
Right	(7,179)	6.3	2.6 ***	1.6	-2.7 ***	4.9	2.8 ***	0.9	-1.6 ***
R2		0.43		0.32		0.37		0.41	

```
*** p<.001, ** p<.01, * p<.05
(collection netmlogit exported to file regresion.xlsx)
```

Regression graph

Regression of p(voting) on gender, age, religion and ideology



Multinomial marginal effects

Voting on gender, age, religion and ideology

Table: Margins and marginal effects of voting

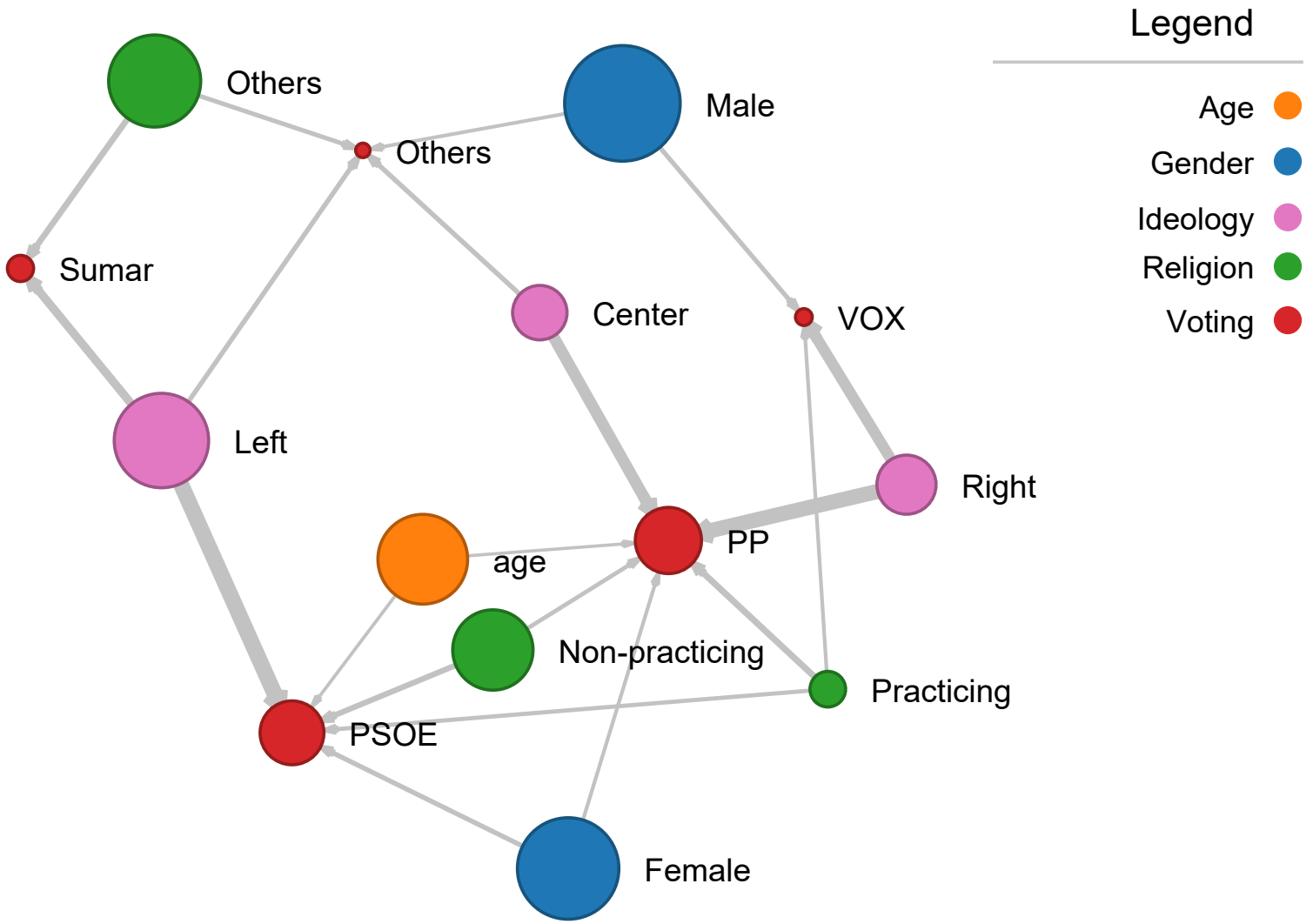
```
. dime mlogit intencionr i.gender age i.religion i.ideology, graph
Table of marginals and global mean differences
```

	Total	PSOE		PP		VOX		Sumar		Others	
Total	(22,777) 30.7		32.4		11.0		15.3		10.5		
Male	(11,966) 27.7	-3.0 ***	31.4	-1.0 **	14.0	3.0 ***	15.5	0.1	11.4	0.9	***
Female	(10,811) 34.1	3.4 ***	33.5	1.1 **	7.7	-3.4 ***	15.2	-0.1	9.5	-1.0	***
Age	(29,201) 31.1	0.4 ***	32.5	0.1 ***	10.8	-0.3 ***	15.2	-0.1 ***	10.4	-0.0	***
Practicing	(4,298) 32.8	2.1 **	38.9	6.5 ***	12.1	1.0 *	6.8	-8.5 ***	9.4	-1.1	
Non-practicing	(8,768) 35.8	5.1 ***	34.6	2.1 ***	11.4	0.4	8.5	-6.8 ***	9.7	-0.8	*
Others	(9,711) 31.1	0.3	24.3	-8.1 ***	10.3	-0.8 *	21.1	5.8 ***	13.2	2.7	***
Left	(10,030) 56.4	25.7 ***	3.2	-29.3 ***	0.9	-10.2 ***	25.8	10.5 ***	13.8	3.3	***
Center	(6,190) 21.2	-9.6 ***	49.3	16.9 ***	11.2	0.2	4.4	-10.9 ***	13.9	3.4	***
Right	(6,557) 8.8	-21.9 ***	56.4	24.0 ***	26.9	15.9 ***	3.6	-11.8 ***	4.2	-6.3	***

```
*** p<.001, ** p<.01, * p<.05
Pseudo R2: 0.279; Nagelkerke´s R2: 0.596; chi2: 19033.47; p: 0
```

Multinomial graph

Multinomial regression of voting on gender, age, religion and ideology



Some proposals

- Network coincidence and regression graphs are proposed as a visual analytic framework.
  - Coincidence graphs are employed mainly for interdependent categorical variables.
  - Regression graphs are employed to represent models of dependence between variables and categories.
- Represent the size of nodes by their importance.
  - Frequency (or percentages)for the categorical variables.
  - Distance of the mean from the minimum value for numerical variables.
- Represent the width of links by the association between categories or variables.
  - Normalized residuals in case of coincidence analysis.
  - Marginal effects in case of regression graphs.
- And express only positive associations.

References

CARING

References on Classsification & Regression Interactive Netgraphs

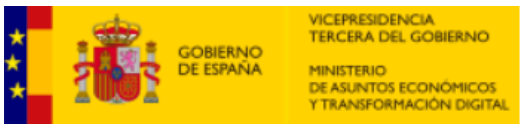
- Escobar, M. (2009). Redes semánticas en textos periodísticos: propuestas técnicas para su representación. *Empiria*, 17, 13-39.
- Escobar, M., y Gómez Isla, J. (2015). "La expresión de la identidad a través de la imagen: los archivos fotográficos de Miguel de Unamuno y Joaquín Turina". *Revista Española de Investigaciones Sociológicas*, 152, 23-46.
- Escobar, M. (2015). "Studying Coincidences with Network Analysis and Other Multivariate Tools". *The Stata Journal*, 15(4), 1118-1156.
- Escobar, M. (2016). "Ensayo sobre las coincidencias". En A. Almarcha, P. González, y L. Román (Eds.), *Donde la Sociología te lleve*. A Coruña: Universidad de A Coruña.
- Escobar, M., y C. Tejero (2018). "El análisis reticular de coincidencias". *Empiria*, 39, 129-148.
- Escobar, M. y L. Martínez (2020) "Network Coincidence Analysis: the netCoin R Package". *International Journal of Software*, 93, 11.

Last slide

Acknowledgment

Thank you very much for your attention!

cristinacalvolopez@usal.es



This work has been partially supported by the Spanish Government with the project Análisis de Coincidencias (PGC2018-093755-B-I00) and the FPI pre-doctoral scholarship number PRE2019-088733.