Content Analysis with Stata

M. Escobar(modesto@usal.es) y J.L. Alonso Berrocal(berrocal@usal.es)

Universidad de Salamanca

8th Spanish Stata Users Group meeting

Madrid, 22th October-2015

Overview

- Background
  - Content analysis
  - Social network analysis
  - Coincidence analysis
  - Stata users-written commands
- The command precoin
  - Multiple variables
  - Thesaurus strings
  - Words
- The command coin
- Next steps

Content Analysis

Definitions

Content analysis is a technique used in the social sciences for the systematic study of the contents of the communication.

- "A systematic, replicable technique for compressing many words of text into fewer content categories based on explicit rules of coding" [Berelson, 1952].
- "Any technique for making inferences by objectively and systematically identifying specified characteristics of messages" [Holsti, 1969].
- "Content analysis is a research technique for making replicable and valid inferences from data to their context" [Krippendorff, 1980].

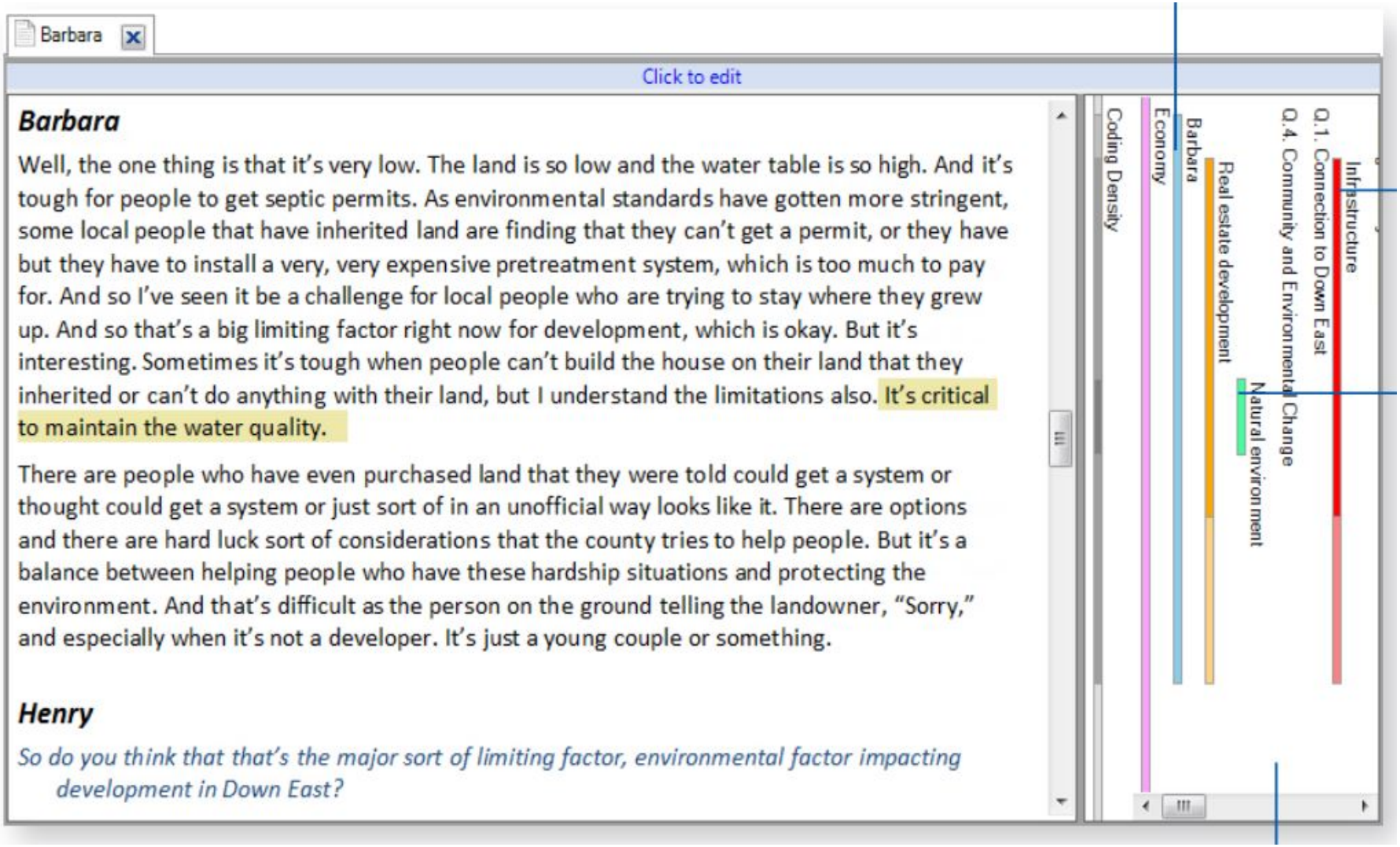
Software for content analysis

Programs

- Qualitative analyzers
  - Nvivo
  - Atlas-ti
  - QDA miner
- Statistical analyzers
  - WordStat
  - TextAnalyst
  - LIWC

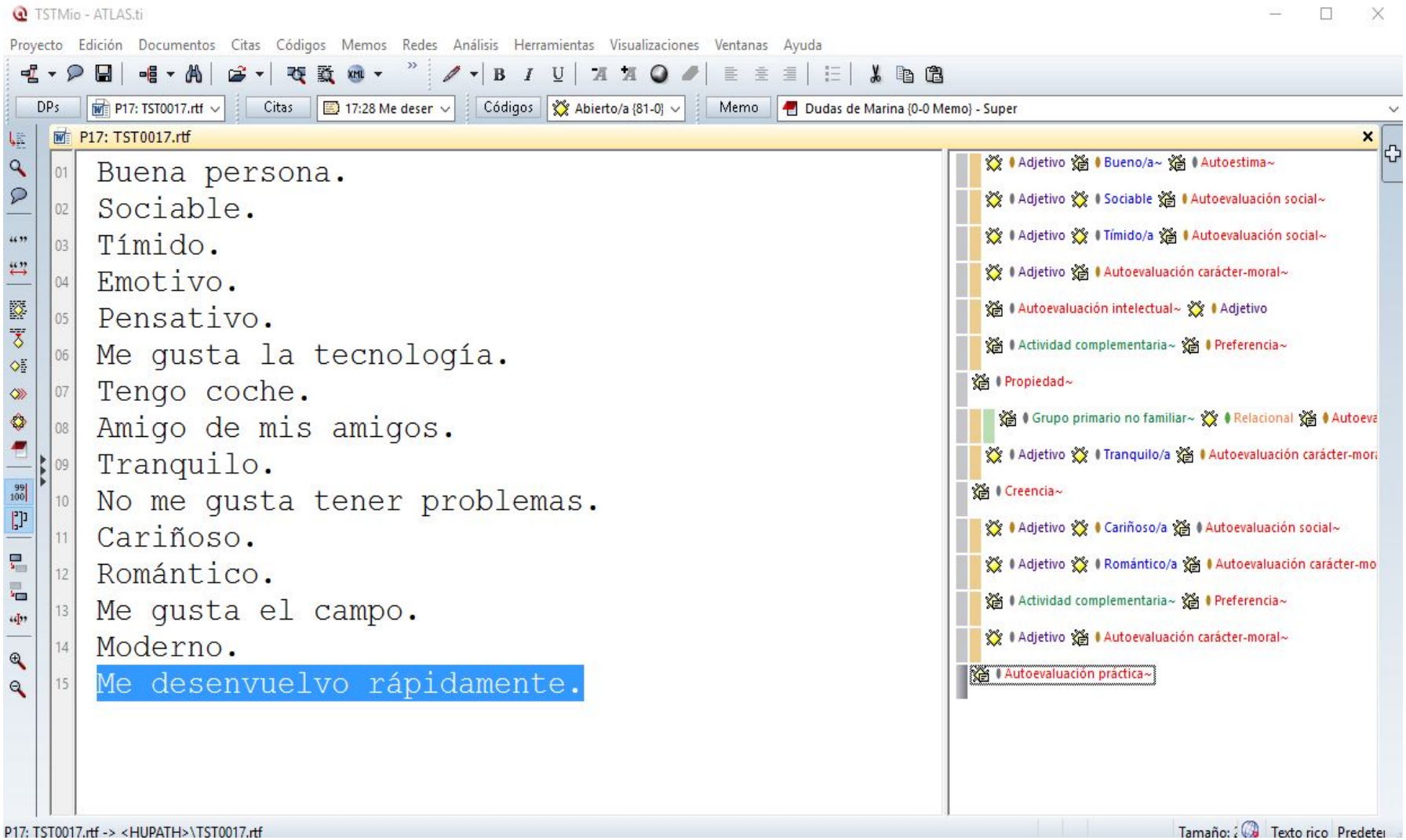
Qualitative analysis programs

Nvivo



Qualitative analysis programs

Atlas-ti



Statistical analysts

WordStat for QDA (and for Stata)

	FREQUENCY	% SHOWN	% PROCESSED	% TOTAL	NO. CASES	% CASES	TF•IDF
AUTÉNTICO	361	12,57%	4,68%	1,76%	344	25,94%	211,5
VIDA	133	4,63%	1,72%	0,65%	124	9,35%	136,9
ELPODERDELOAUTENTICO	114	3,97%	1,48%	0,56%	114	8,60%	121,5
PLAYA	90	3,13%	1,17%	0,44%	90	6,79%	105,1
FAMILIA	83	2,89%	1,08%	0,41%	83	6,26%	99,9
GARNIER	76	2,65%	0,98%	0,37%	76	5,73%	94,4
DISFRUTAR	76	2,65%	0,98%	0,37%	75	5,66%	94,8
MAR	61	2,12%	0,79%	0,30%	60	4,52%	82
AMIGOS	52	1,81%	0,67%	0,25%	52	3,92%	73,1
PELO	52	1,81%	0,67%	0,25%	52	3,92%	73,1
SONRISA	52	1,81%	0,67%	0,25%	51	3,85%	73,6
VIVIR	47	1,64%	0,61%	0,23%	41	3,09%	71
VERANO	46	1,60%	0,60%	0,22%	42	3,17%	69
AMOR	44	1,53%	0,57%	0,21%	42	3,17%	66
SOL	43	1,50%	0,56%	0,21%	41	3,09%	64,9
MOMENTOS	43	1,50%	0,56%	0,21%	40	3,02%	65,4
SENTIR	42	1,46%	0,54%	0,21%	41	3,09%	63,4

Social network analysis

Stata programs

- Although there are no tools for SNA in Stata, some advanced users have begun to write some routines. I wish to highlight the following works from which I have obtained insights:
  - Corten [2011] wrote a routine to visualize social networks [netplot]
  - Miura [2012] created routines (SGL) to calculate networks centrality measures, including two Stata commands [netsis and netsummarize]
  - White presented a suite of Stata programs for network meta-analysis which includes the network graphs of Anna Chaimani in the 2013 UK users group meeting. Cerulli and Zinilii presented a procedure [datanet] to prepare a dataset for analysis purposes in the 2014 Italian Stata Users Group meeting.
  - Grund [2014] have created a collection of programs to plot and analyze social networks in the Nordic and Baltic Stata Users Group [nwcommands].

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 scenarios (n), 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.
- The starting point is an 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.

Pictures analysis

4 pictures (scenarios) & 8 different people (events)

Example with names

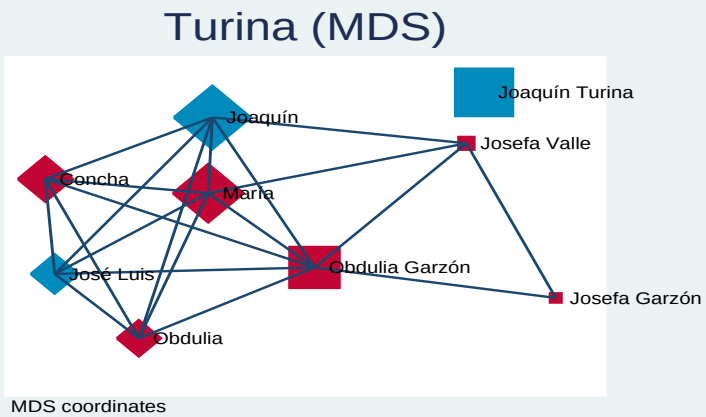
Father, mother, grandmother and 5 children

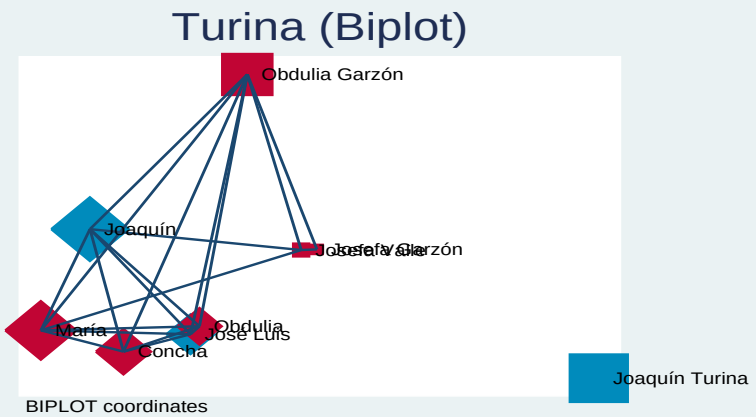
Example with codes

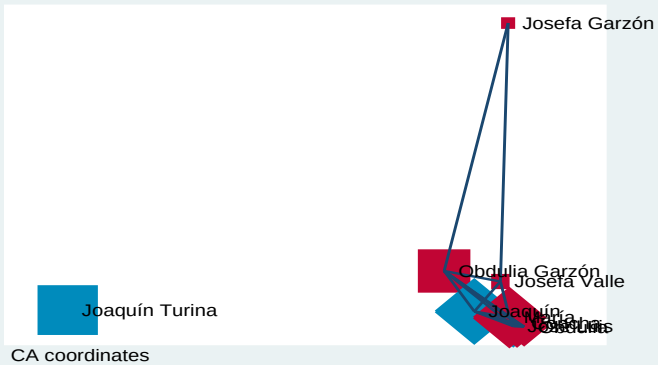
Turina, Garzón, Joaquín, María, Concha, José Luis, Obdulia, Valle

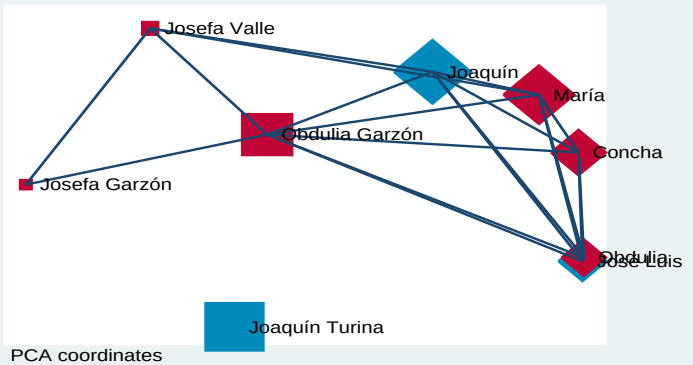
Coincidences graphs

MDS-Biplot-CA-PCA









Other uses of coincidence analysis

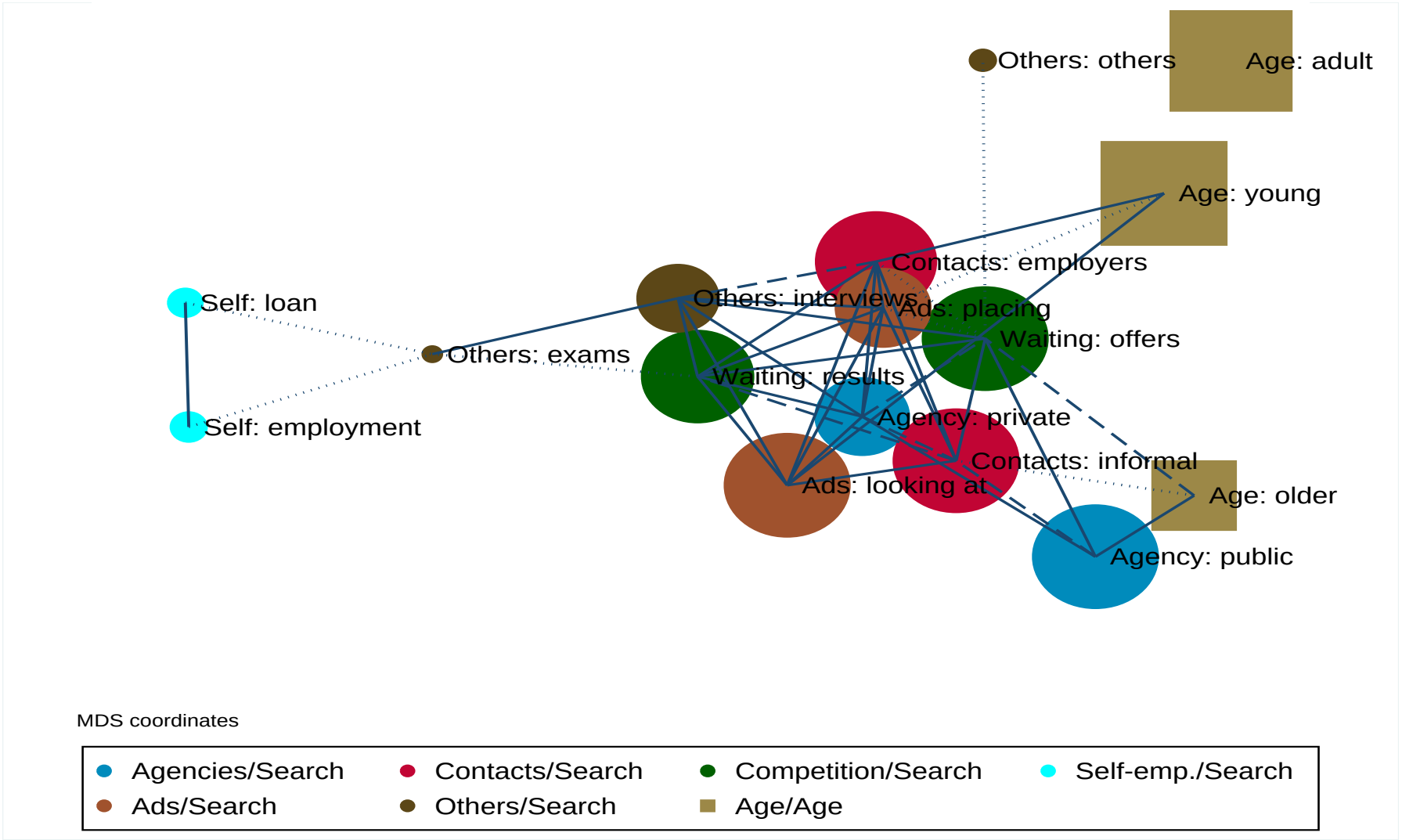
From survey analysis to cultural trends

Coincidence analysis has many applications. Among others:

- Survey analysis
  - Unemployment
  - Social problems
  - Mass media audience
- Data Mining
  - Samples (Composition of genes)
  - Corruption (Black cards)
- Cultural trends
  - Composers
  - Painters
  - Creators
- Content analysis

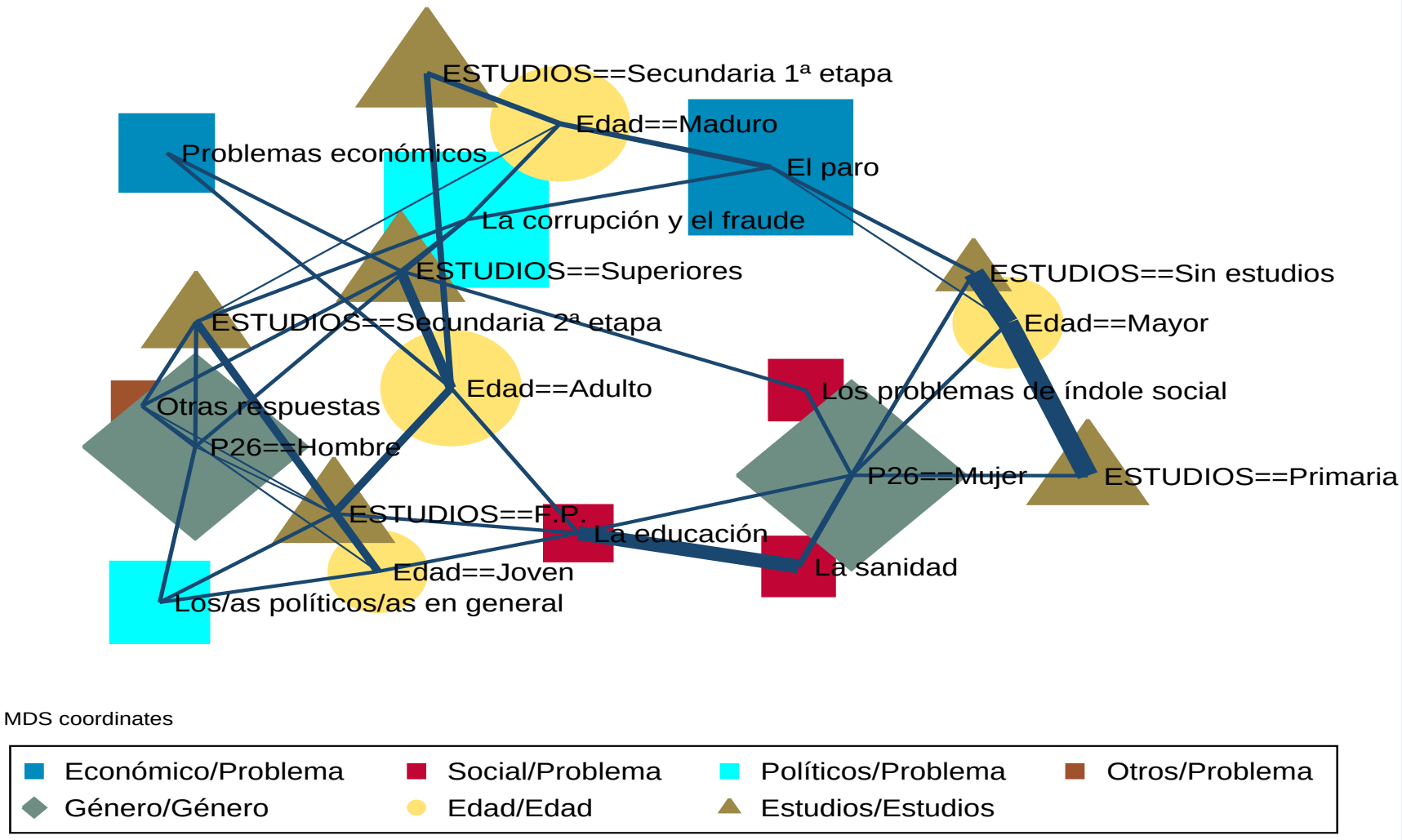
Survey analysis

Ways of looking for jobs (EPA-2014)



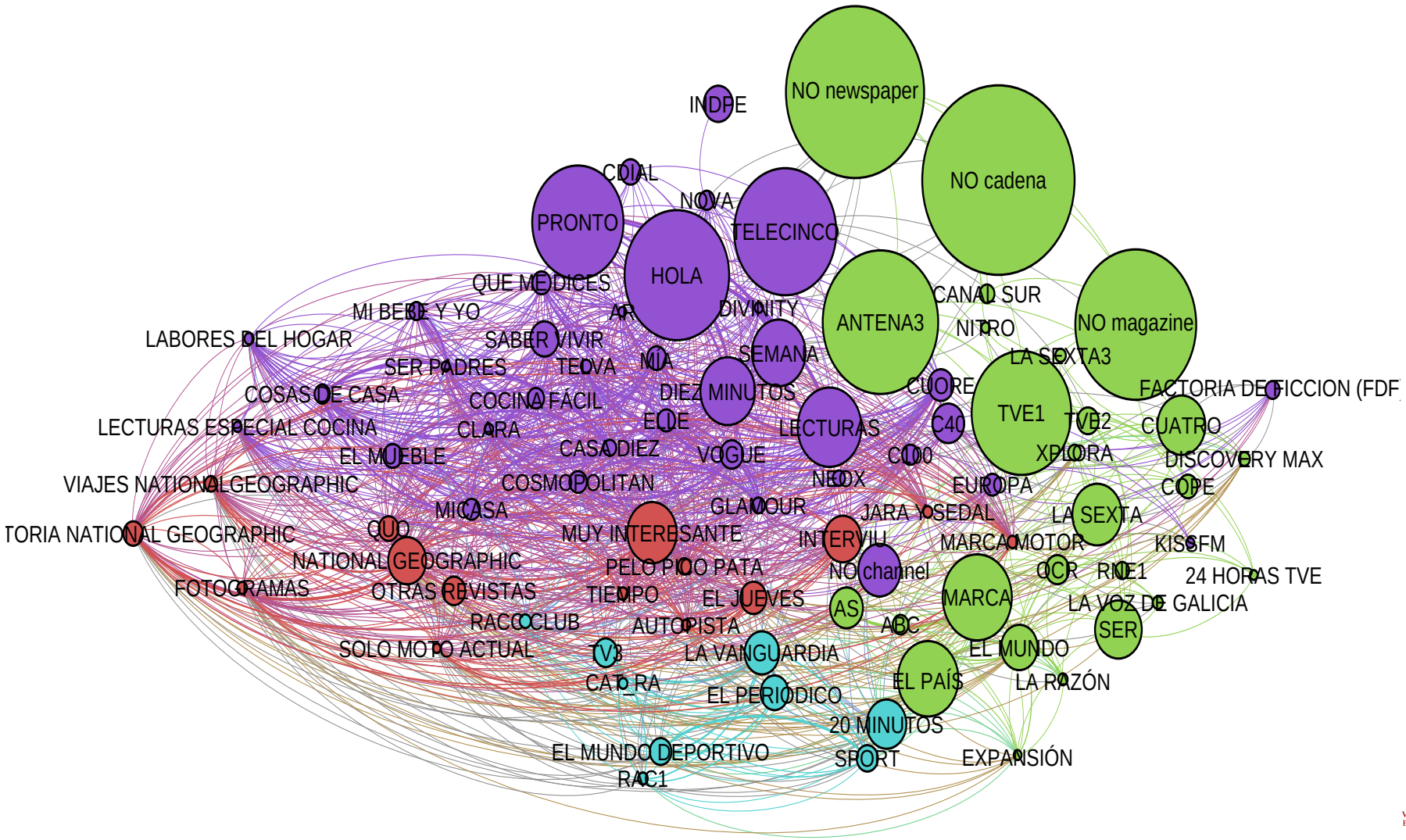
Survey analysis

Social problems in Spain (2014) CIS-3045



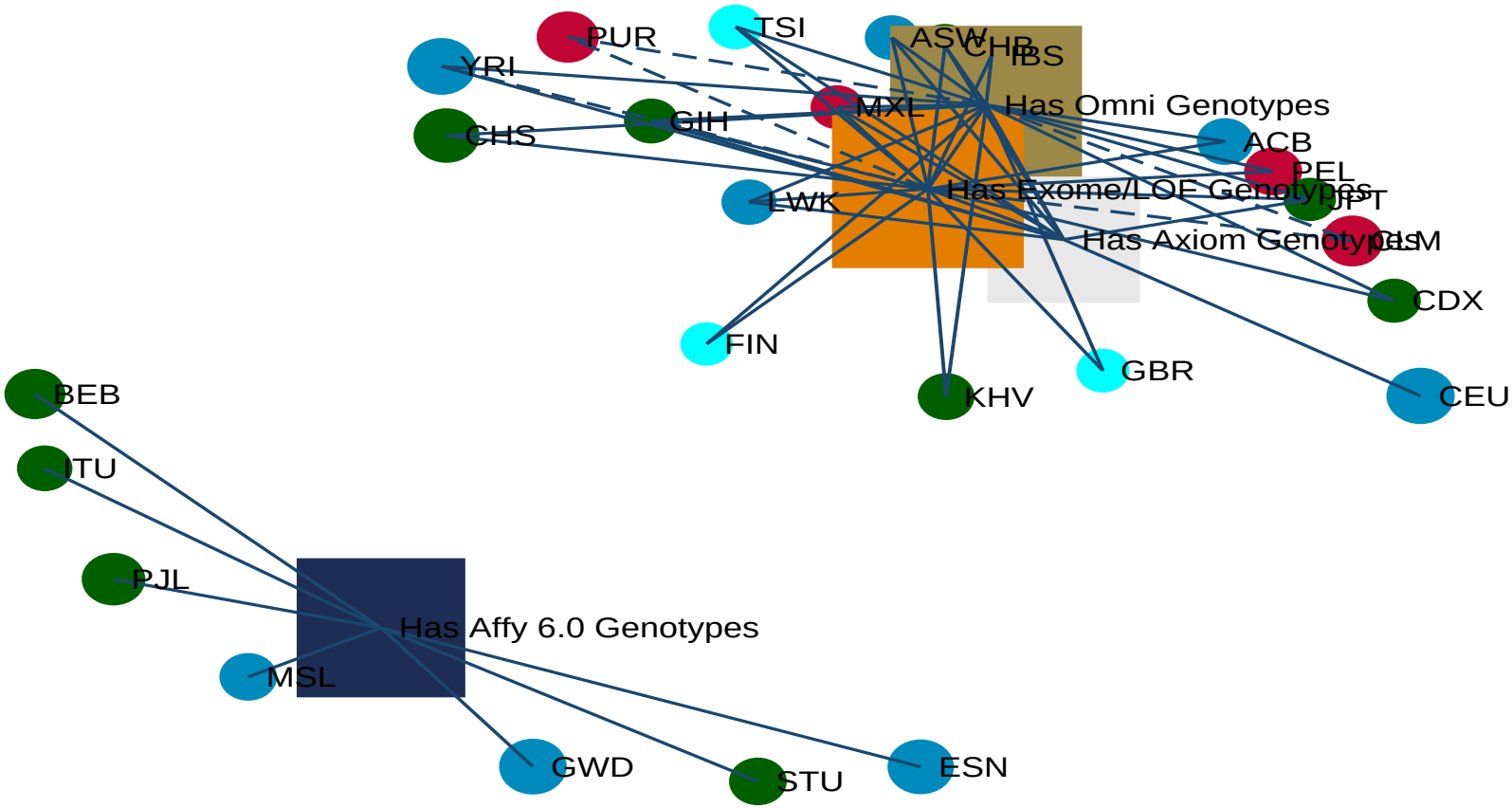
Survey analysis

Mass media audience (EGM-2013)



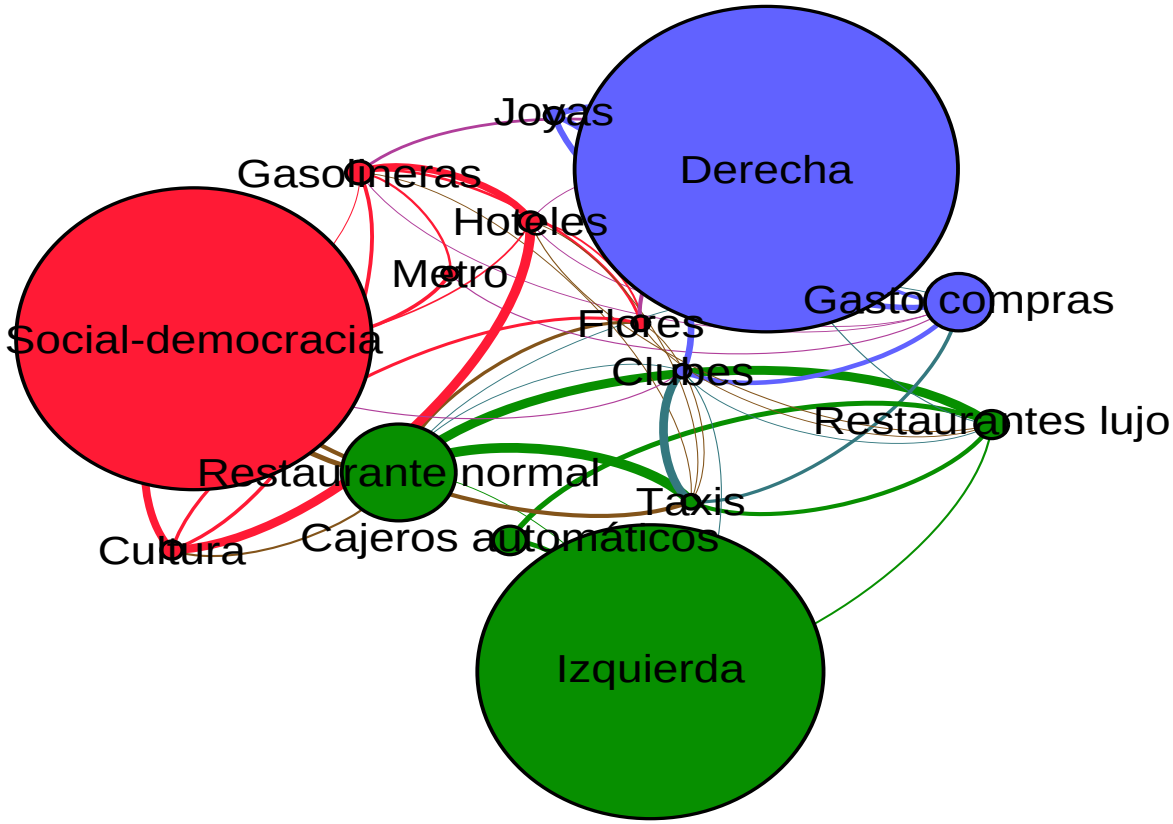
Data mining

Genes composition of samples. Fuente: http://www.1000genomes.org/data



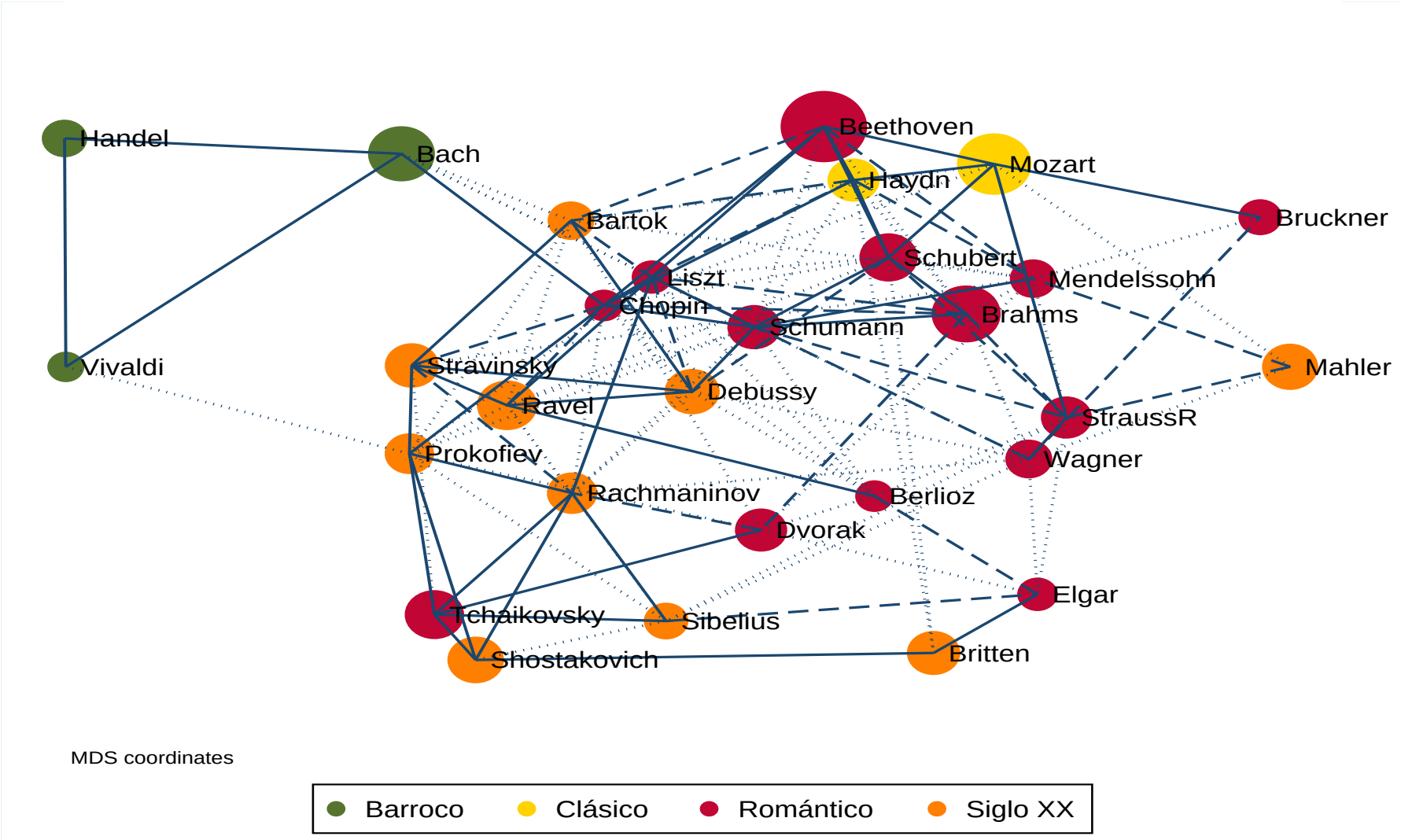
Data mining

Mean expenses per person with Bankia black cards (2003-2011)



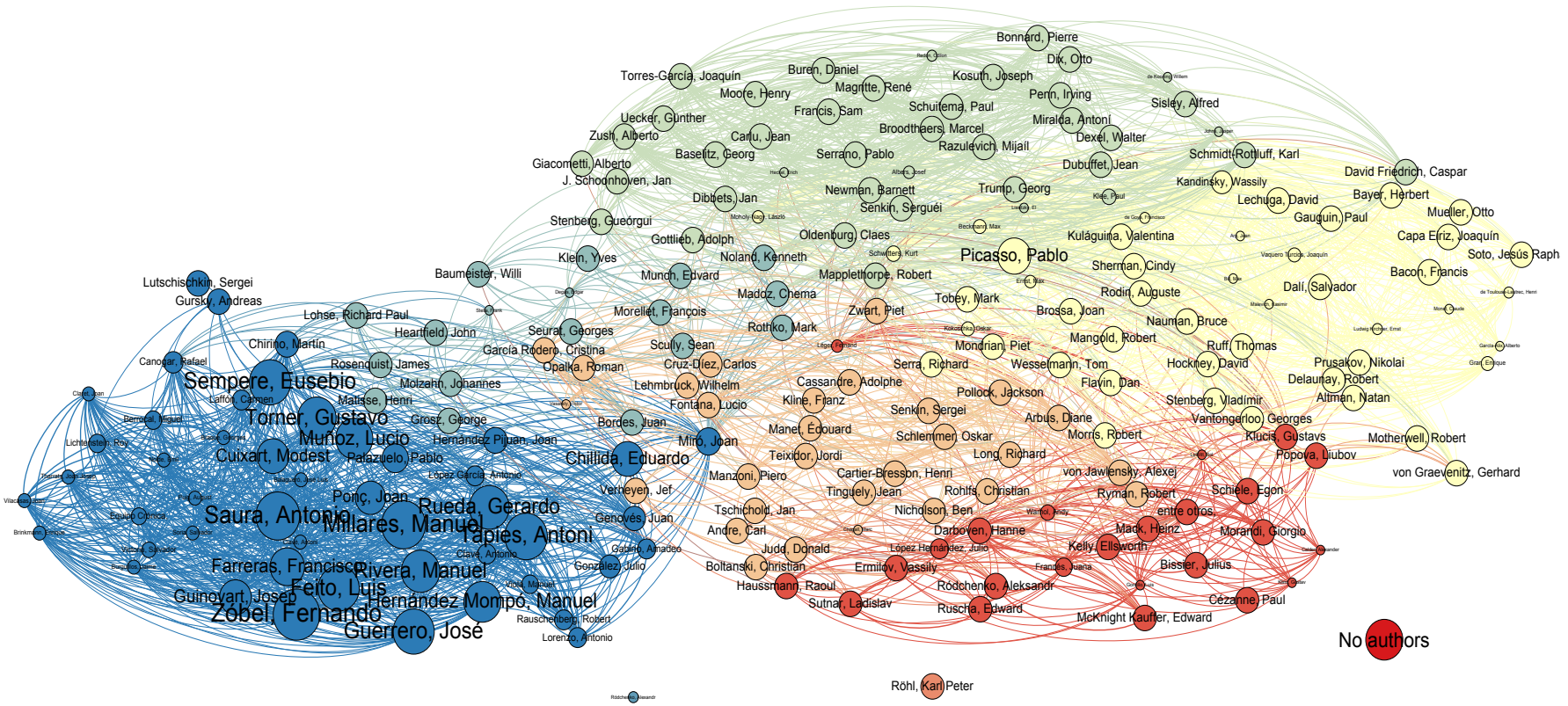
Cultural trends

Bachtack concerts reviewed (2009-2015)



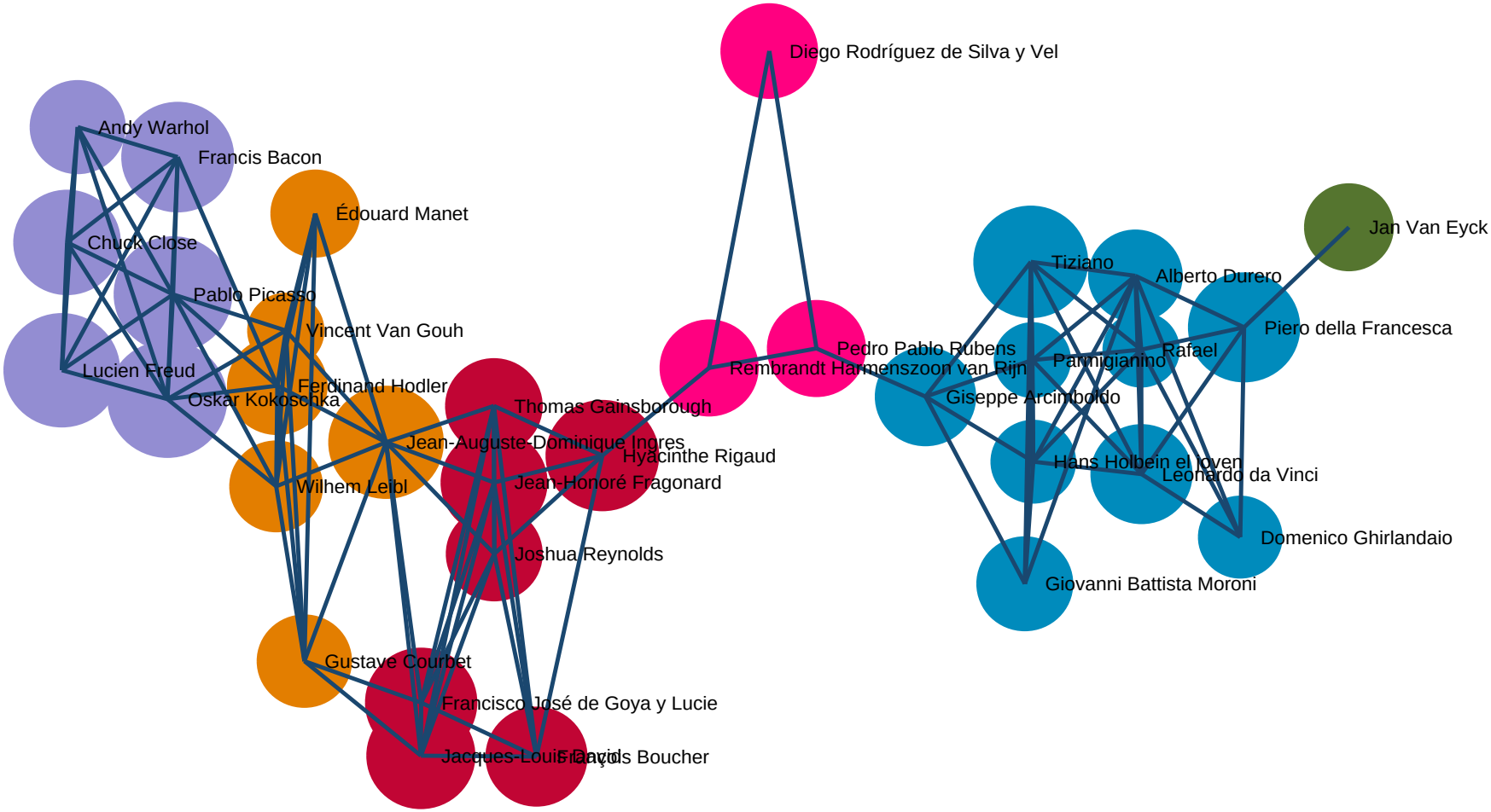
Cultural trends

Creators in Juan March exhibitions (1975-2015)



Cultural trends

Timeline of famous portrait painters



Stata user-written commands

Main

- `txttool` provides a set of tools for managing and analyzing free-form text. The command integrates several built-in Stata functions with new text capabilities, including a utility to create a bag-of-words representation of text and an implementation of Porter's word-stemming algorithm.
- `wordfreq` inputs a set of text files and produces in memory a set of frequencies of all words that occur in at least one of the input texts. The resulting dataset consists of a text variable word containing a list of the words themselves.
- `wordscores` implements the computerized content analysis techniques described in "Extracting Policy Positions From Political Texts Using Words as Data" by Laver et al. [2003]

Stata user-written commands

Others

- `strdist` module to calculate the Levenshtein distance (or edit distance) between strings.
- `matchit` is a tool to join observations from two datasets based on string variables which do not necessarily need to be exactly the same. It performs many different string-based matching techniques, allowing for a fuzzy similarity between the two different text variables.

precoin

Distinct uses

`precoin` converts politomous variables into binary variables for coincidence analysis. Original variables can be either numerical or string.

It also can divide the content of just one variable into different dichotomous variables according to a separator.

It has three kind of uses:

- Multiple variables
- Thesaurus strings
- Words

precoin uses

Multiple variables

	P701	P702	P703
1	Los/as políticos/as en general, los partidos y ...	El paro	Las drogas
2	El paro	La corrupción y el fraude	N.C.
3	El paro	La corrupción y el fraude	N.C.
4	La corrupción y el fraude	Los/as políticos/as en general, los partidos y ...	Los problemas de índole económica
5	Los/as políticos/as en general, los partidos y ...	La corrupción y el fraude	N.C.
6	La corrupción y el fraude	El paro	Los problemas de índole económica
7	El paro	Los problemas de índole económica	La corrupción y el fraude
8	El paro	Los/as políticos/as en general, los partidos y ...	La corrupción y el fraude
9	El paro	La corrupción y el fraude	N.C.
10	El paro	N.C.	N.C.
11	Los/as políticos/as en general, los partidos y ...	La corrupción y el fraude	N.C.
12	El Gobierno y partidos o políticos/as concretos	El funcionamiento de los servicios públicos	N.C.
13	La corrupción y el fraude	El paro	Los/as políticos/as en general, los partidos y ...
14	La inmigración	La corrupción y el fraude	Los problemas relacionados con la calidad del e...
15	La corrupción y el fraude	El paro	La violencia contra la mujer
16	El paro	La corrupción y el fraude	N.C.
17	La corrupción y el fraude	El paro	N.C.
18	La educación	La sanidad	N.C.
19	El paro	La corrupción y el fraude	N.C.
20	La corrupción y el fraude	El paro	N.C.
21	Los/as políticos/as en general, los partidos y ...	El paro	N.C.
22	El paro	La corrupción y el fraude	N.C.
23	La corrupción y el fraude	El paro	Los/as políticos/as en general, los partidos y ...

precoin uses

Frequencies of multiple variables

```
. precoin P701-P703, stub(problem) min(.02) sort freq replace
```

Categories	f	%/events	%/scenar
El paro	1897	30.6	77.1
La corrupción y el fraude	1573	25.4	63.9
Los problemas de índole económic	629	10.1	25.6
Los/as políticos/as en general,	574	9.3	23.3
Others	327	5.3	13.3
Los problemas de índole social	220	3.5	8.9
La sanidad	213	3.4	8.7
La educación	190	3.1	7.7
Otras respuestas	145	2.3	5.9
Los recortes	101	1.6	4.1
La Administración de Justicia	88	1.4	3.6
El Gobierno y partidos o polític	68	1.1	2.8
La inmigración	62	1.0	2.5
Los problemas relacionados con l	58	0.9	2.4
La crisis de valores	54	0.9	2.2

Events:	6199
Scenarios:	2461
Missing scenarios:	4

precoin uses

Transformation of multiple variables

```
. describe problem*
```

variable name	storage type	display format	value label	variable label
problem01	byte	%8.0g		El paro
problem02	byte	%8.0g		La corrupción y el fraude
problem03	byte	%8.0g		Los problemas de índole económic
problem04	byte	%8.0g		Los/as políticos/as en general,
problem05	byte	%8.0g		Los problemas de índole social
problem06	byte	%8.0g		La sanidad
problem07	byte	%8.0g		La educación
problem08	byte	%8.0g		Otras respuestas
problem09	byte	%8.0g		"Los recortes"
problem10	byte	%8.0g		La Administración de Justicia
problem11	byte	%8.0g		El Gobierno y partidos o polític
problem12	byte	%8.0g		La inmigración
problem13	byte	%8.0g		Los problemas relacionados con l
problem14	byte	%8.0g		La crisis de valores
problem99	byte	%8.0g		Others
problem_miss	byte	%8.0g		No events

precoin uses

Thesauri strings

	concert_title	venue	city	date	composers
1	Barenboim une a Wagner y Elgar en el Palau de l...	Palau de la Música Catalana	Barcelona	06/07/2015	Wagner;Wagner;Wagner;Elgar
2	Leonskaja, o cuando Schubert daña	Auditorio Nacional de Musica: Sa...	Madrid	30/06/2015	Schubert;Schönberg;Schubert;Schubert
3	Goerne cantando a la muerte: último recital del...	Teatro de la Zarzuela	Madrid	29/06/2015	Berg;Schubert;Schubert;Schubert;Brahms;Wolf;Sho...
4	Sinfónica Juvenil en dos conciertos de alto vol...	Gran Teatro Nacional	Lima	28/06/2015	Ravel;Prokofiev;Bruch;Ravel;Mozart;Milhaud;Vivaldi
5	Tres nuevas creaciones escénicas en el festival...	Teatros del Canal	Madrid	14/06/2015	Bernal;Galindo;Rappoport
6	Afkham y Beethoven en el Auditorio Nacional: el...	Auditorio Nacional de Musica	Madrid	12/06/2015	Brahms;Brahms;Brahms;Beethoven
7	Khatia Buniatishvili: la cristalina sensualidad...	Auditorio de Valladolid	Valladolid	12/06/2015	Rachmaninov;Stravinsky
8	WDR Köln en Ibercamera y el ruido de fondo...	L'Auditori: Sala Pau Casals	Barcelona	08/06/2015	Dvorák;Tchaikovsky
9	Un tenor que llegó para quedarse: Javier Camare...	Palacio Euskalduna	Bilbao	07/06/2015	Gounod;Gounod;Bizet;Rossini;Rossini;Donizetti;V...
10	Variaciones Goldberg por Konjar y Espasa: revis...	L'Auditori: Sala Tete Montoliu	Barcelona	05/06/2015	Bach
11	El Latinoamericano acomete la integral de Cuart...	Teatro Mayor Julio Mario Santo D...	Bogotá	21/05/2015	Mignone;Villa-Lobos
12	Un Liszt a la española, pero sin ayeo, en la Fu...	Fundación Juan March	Madrid	20/05/2015	(del Pópulo Vicente Rodríguez) García;Liszt;Lis...
13	Ashkenazy con la Philharmonia Orchestra: espejo...	Auditorio Nacional de Musica	Madrid	18/05/2015	Sibelius;Sibelius;Sibelius
14	Bach y Pärt llegan a nuevos públicos en el Rein...	Museo Nacional Centro de Arte Re...	Madrid	17/05/2015	Pärt
15	Tango a tres en la Fundación Juan March	Fundación Juan March	Madrid	16/05/2015	Piazzolla;Gardel;Guastavino;Piazzolla;Piazzolla...
16	Ciclo de Lied: Vivica y Viardot en homenaje a T...	Teatro de la Zarzuela	Madrid	11/05/2015	Haydn;Bellini;Viardot-García
17	Música de Steve Reich en el Auditorio Nacional ...	Auditorio Nacional de Musica: Sa...	Madrid	23/04/2015	Reich;Reich;Reich
18	Bruckner y Afkham en el Auditorio Nacional: sin...	Auditorio Nacional de Musica	Madrid	17/04/2015	Berg;Bruckner
19	Resurrección en el Auditorio Nacional, o la arr...	Auditorio Nacional de Musica	Madrid	11/04/2015	Mahler
20	La OFGC interpreta a Beethoven junto al solista...	Auditorio Alfredo Kraus	Las Palmas d...	10/04/2015	Beethoven;Beethoven
21	Biondi: el juego infructuoso de las medias tintas	Auditorio Nacional de Musica	Madrid	08/04/2015	Mozart;Haydn;Mozart
22	Jóvenes colombianos demuestran un fino estilo m...	Teatro Colón	Bogotá	04/04/2015	Mozart;Mozart;Mozart;Mozart;Mozart
23	J.S. Bach, de procesión en Gran Canaria	Auditorio Alfredo Kraus	Las Palmas d...	02/04/2015	Bach
24	Gatti enmienda a Tchaikovsky: ¡luz, más luz!	Auditorio Nacional de Musica	Madrid	25/03/2015	Debussy;Ravel;Tchaikovsky
25	Emotivo debut de Natalie Dessay en A Coruña	Palacio de la Ópera	La Coruña	24/03/2015	Brahms;Schumann;Duparc;Strauss R.;Fauré;Poulenc...
26	Cuando la prisa es mala consejera	Auditorio Nacional de Musica	Madrid	20/03/2015	Shostakovich;Beethoven
27	Pórtico de Zamora: el encanto de lo bien hecho	Iglesia de San Cipriano	Zamora	20/03/2015	Nebra;García Fajer
28	Pasión española en el Wigmore Hall a cargo del ...	Wigmore Hall	Londres	14/03/2015	Ravel;De Falla;De Falla;De Falla;Traditional Sp...
29	Un mundo en retirada: War Requiem en el Teatro ...	Teatro Real	Madrid	12/03/2015	Britten
30	Una tarde \Entre gigantes\" en el Auditorio Nac...	Auditorio Nacional de Musica: Sa...	Madrid	11/03/2015	Bach;Cabezon;Bach;Cabezon;Bach;Cabezon;Bach;Cab...
31	Sokolov en el Auditorio Nacional: nuevo retorno...	Auditorio Nacional de Musica	Madrid	09/03/2015	Bach;Beethoven;Schubert;Schubert
32	El neotonalismo se afirma en la Carta Blanca de...	Auditorio Nacional de Musica	Madrid	06/03/2015	Pärt;Pärt;Pärt;Pärt

precoin uses

Frequencies of thesauri strings

```
. precoin composers, stub(composer) sep(;) freq sort min(.05) missing replace
```

Categories	f	%/events	%/scenar
Others	3606	57.2	86.4
Beethoven	528	8.4	12.7
Mozart	387	6.1	9.3
Brahms	329	5.2	7.9
Bach	309	4.9	7.4
Ravel	247	3.9	5.9
Tchaikovsky	239	3.8	5.7
Schubert	231	3.7	5.5
Shostakovich	215	3.4	5.2
Mahler	214	3.4	5.1
No events	46	0.7	1.1
Events:	6305		
Scenarios:	4173		

precoin uses

Variables of thesauri strings

```
. describe composer*
```

variable name	storage type	display format	value label	variable label
composers	str100	%-50s		Composers
composer0001	byte	%8.0g		Beethoven
composer0002	byte	%8.0g		Mozart
composer0003	byte	%8.0g		Brahms
composer0004	byte	%8.0g		Bach
composer0005	byte	%8.0g		Ravel
composer0006	byte	%8.0g		Tchaikovsky
composer0007	byte	%8.0g		Schubert
composer0008	byte	%8.0g		Shostakovich
composer0009	byte	%8.0g		Mahler
composer1823	byte	%8.0g		Others
composer_miss	byte	%8.0g		No events

precoin uses

Thesauri strings

	NombreUsuario	Mensaje	Plataforma	Instagram	Twitter	Web
1	nmolsan	#elpoderdeloautentico #originalremedies @garnier_es	Instagram	1	0	0
2	noelialuque	Tarde beauty #summerbeautyday #elpoderdeloautentico #originalremedies @garnier_es	Instagram	1	0	0
3	supernala1982	@garnier_es #elpoderdeloautentico #originalremedies#	Instagram	1	0	0
4	123patataaa	Tarde beauty #summerbeautyday #elpoderdeloautentico #originalremedies @garnier_es	Instagram	1	0	0
5	Gara Charada	La inspiración es enamorarse de algo que todavía no existe	Web	0	0	1
6	Gara Charada	I love you but don´t know what to do.	Web	0	0	1
7	X	Corazón, se te va la olla	Web	0	0	1
8	Swimmer	Si de verdad te gusta cuando lleva un gorro de piscina puesto, es ella	Web	0	0	1
9	edriesgo	Esos días que sales a tomarte solo una y te acabas encontrando con un plan imparable #elpoderdeloautentico	Instagram	1	0	0
10	EdRiesgo	Nada más asturiano que nuestro segundo plan de reserva :) #elpoderdeloauténtico #elpoderdeloautentico	Twitter	0	1	0
11	Paula	Que bonito es el amor mas que nunca en primavera	Web	0	0	1
12	JDev	La sensación agridulce de ir a un baño público y que la taza este calentita…#elpoderdeloautentico	Web	0	0	1
13	Gon	Si no existieras habrá que inventarte #contigoterapia	Web	0	0	1
14	sandrahache	Comer lo que te gusta, escuchar esa canción que te emociona, un beso apasionado, aquella mirada, el sol en la cara cuando hace frío. El agua ...	Web	0	0	1
15	Maria	No te olvides de traer pan y leche	Web	0	0	1
16	Paula Álvarez Rioja	El spa que monto en mi baño cada vez que me lavo el pelo	Web	0	0	1
17	Sandra	El momento en que tu pareja te enjabona en la ducha #elpoderdeloauténtico	Web	0	0	1
18	David	Cuando las palabras sobran y hablan las miradas #amorautentico	Web	0	0	1
19	Sepharad	Mirar por la ventana y ver a mi hijo jugando en la calle, feliz y despreocupado #simplementeamor	Web	0	0	1
20	Maly	Los paseos en primavera, por la orilla del mar, sintiendo la arena y el agua en los pies	Web	0	0	1
21	Luz Stella Zuluaga	#elpoderdeloauténtico es aplicarme la nueva crema de Garnier de aguacate y macadamia y sentir mi cabello auténticamente fabuloso llevo de vida	Web	0	0	1
22	Marta Montero Toro	No hay nada mas auténtico que ser tú mismo. No imites a nadie, que el mundo no se pierda tu verdadero yo.	Web	0	0	1
23	Ariadna	Ese momento en el que te levantas sin despertador con tu pareja al lado, el sol entrando por la ventana…#elpoderdeloautentico	Web	0	0	1
24	Rakel Rosales Fortuño	Creer en ti mismo y no dejarse influenciar por los demás	Web	0	0	1
25	Mari Angeles Martín Entonado	El mejor momento del día es la siesta debería considerarse deporte Nacional!!!	Web	0	0	1
26	Lucía Cáceres	Auténtico es dormirse con el pelo algo mojado y despertarse con un peinado que ni el mejor peluquero podría hacer natural y desenfadado.	Web	0	0	1
27	Nuria Villalón	En la playa, un día de sol con mi pareja y amigos	Web	0	0	1
28	maria teresa fernandez fer...	Las fiestas de verano	Web	0	0	1
29	Yolanda Isabel Hurtado Garcia	Lo auténtico para mi es ser sinceros pese a quien pese y poder disfrutar de mi familia y amigos que me quieren pese a mis defectos	Web	0	0	1
30	karolina	Un día de risas, sol, playa, juegos, secretos con mis VERDADERAS amigas, no hay nada como la amistad auténtica	Web	0	0	1
31	Rocio Angel	Vivir sin que importe el mañana y disfrutar de las pequeñas cosas que nos da la vida	Web	0	0	1
32	Alarcos	Lo más auténtico para mi es el amor verdadero, ese amor inolvidable y único que cambia tu vida	Web	0	0	1

precoin uses

Simple conversion

```
. precoin Plataforma, stub(labels) freq
Warning: separator has been set to space
```

Categories	f	%/events	%/scenar
Instagram	79	6.0	6.0
Twitter	225	17.0	17.0
Web	1020	77.0	77.0

Events:	1324
Scenarios:	1324
Missing scenarios:	2

```
. describe Instagram-Web
```

variable name	storage type	display format	value label	variable label
Instagram	byte	%8.0g		Instagram
Twitter	byte	%8.0g		Twitter
Web	byte	%8.0g		Web

precoin uses

Words

```
. precoin Mensaje, stub(labels) sort freq replace stop(stopwords.txt) separator(" ") min(.03)
```

Categories	f	%/events	%/scenar
Others	1275	49.4	96.5
Autentico	369	14.3	27.9
Elpoderdeloautentico	146	5.7	11.1
Vida	121	4.7	9.2
Playa	87	3.4	6.6
Familia	82	3.2	6.2
Disfrutar	73	2.8	5.5
Mar	58	2.2	4.4
GarnierEs	55	2.1	4.2
Amigos	51	2.0	3.9
Pelo	51	2.0	3.9
Sonrisa	50	1.9	3.8
Amor	42	1.6	3.2
Sol	41	1.6	3.1
Verano	40	1.5	3.0
Sentir	40	1.5	3.0

Events:	2581
Scenarios:	1321
Missing scenarios:	5

precoin uses

Simple convertion

```
. describe Autentico-Mensaje_miss
              storage   display    value
variable name   type    format     label      variable label
------------------------------------------------------------------------------
Autentico       byte    %8.0g                 Autentico
Elpoderdeloau~o byte    %8.0g                 Elpoderdeloautentico
Vida            byte    %8.0g                 Vida
Playa           byte    %8.0g                 Playa
Familia         byte    %8.0g                 Familia
Disfrutar       byte    %8.0g                 Disfrutar
Mar             byte    %8.0g                 Mar
GarnierEs       byte    %8.0g                 GarnierEs
Amigos          byte    %8.0g                 Amigos
Pelo            byte    %8.0g                 Pelo
Sonrisa         byte    %8.0g                 Sonrisa
Amor            byte    %8.0g                 Amor
Sol             byte    %8.0g                 Sol
Verano          byte    %8.0g                 Verano
Sentir          byte    %8.0g                 Sentir
Mensaje_others  byte    %8.0g                 Others
Mensaje_miss    byte    %8.0g                 No events
```

coin

What is it?

`coin` is an ado program 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 Ucinet, Pajeck, nwcommands, Excel and csv files
- Its syntax is simple, but flexible. Many options (output, bonferroni, p value, minimum, special event, graph control and options, ...)

Command

coin

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

Options can be classified into the following groups:

- **Outputs**:
  - Frequencies: frequencies g-relative-frequencies vertical% horizontal%, expected-frequencies odd-ratios,
  - Residuals: residuals standard-residuals normalized-residuals
  - Significance: phaberman podd ratios pfisher-exact-test
  - Others: tetrachoric-correlations, adjacencies-matrix distances list-key centrality measures, all-previous-statistics
  - Coordinates: x (with plot) xy(circle|mds|ca|pca|biplot).
- **Plots**
  - Bar: bar, cbar(*varlist*) and ccbar(*varlist*)
  - Residuals: rgraph(*varlist*) and ograph(*varlist*)
  - Graph: graph(circle|mds|ca|pca|biplot)
  - Dendrograms: dendrogram(single|complete|average|wards)

Command

coin (continued)

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

Options can be classified into the following groups (continued):

- **Controls**: head(*varlist*), variable(*varname*), ascending, descending, minimum (#), support(#), pvalue(#), levels(# # #), bonferroni, lminimum(#), iterations(#).
- **Exports**
  - Edges: export(filename) with .csv .xls .nw .pjk and .dl extensions
  - Nodes: varsave(filename) o export(filename) with .csv or .xls extensions

coin example (I)

Matrix of coincidences in L'Oreal's messages

```
. coin Vida-Mar Amigos-Sentir, frequencies

1326 scenarios. 32 probable coincidences amongst 12 events. Density: 0.48. Components: 1.
12 events(n>=5): Vida Playa Familia Disfrutar Mar Amigos Pelo Sonrisa Amor Sol Verano Sentir
```

Frequencies	Vida	Playa	Fam~a	Dis~r	Mar	Ami~s	Pelo	Son~a	Amor	Sol	Ver~o	Sen~r
Vida	121											
Playa	2	87										
Familia	6	15	82									
Disfrutar	13	6	9	73								
Mar	4	5	1	8	58							
Amigos	1	9	17	8	3	51						
Pelo	4	2	2	3	6	1	51					
Sonrisa	7	0	0	1	2	1	1	50				
Amor	8	0	1	1	1	0	4	0	42			
Sol	3	11	0	3	5	2	4	3	2	41		
Verano	2	7	6	1	2	4	3	0	0	1	40	
Sentir	6	0	1	1	2	0	6	0	3	0	3	40

coin example (II)

Matrix of expected coincidences in L'Oreal's messages

```
. coin Vida-Mar Amigos-Sentir, expected
1326 scenarios. 32 probable coincidences amongst 12 events. Density: 0.48. Components: 1.
12 events(n>=5): Vida Playa Familia Disfrutar Mar Amigos Pelo Sonrisa Amor Sol Verano Sentir
```

Expected frequencies	Vida	Playa	Fam~a	Dis~r	Mar	Ami~s	Pelo	Son~a	Amor	Sol	Ver~o	Sen~r
Vida	11.0											
Playa	7.9	5.7										
Familia	7.5	5.4	5.1									
Disfrutar	6.7	4.8	4.5	4.0								
Mar	5.3	3.8	3.6	3.2	2.5							
Amigos	4.7	3.3	3.2	2.8	2.2	2.0						
Pelo	4.7	3.3	3.2	2.8	2.2	2.0	2.0					
Sonrisa	4.6	3.3	3.1	2.8	2.2	1.9	1.9	1.9				
Amor	3.8	2.8	2.6	2.3	1.8	1.6	1.6	1.6	1.3			
Sol	3.7	2.7	2.5	2.3	1.8	1.6	1.6	1.5	1.3	1.3		
Verano	3.7	2.6	2.5	2.2	1.7	1.5	1.5	1.5	1.3	1.2	1.2	
Sentir	3.7	2.6	2.5	2.2	1.7	1.5	1.5	1.5	1.3	1.2	1.2	1.2

coin example (III)

Matrix of normalized residuals in L'Oreal's messages

```
. coin Vida-Mar Amigos-Sentir, normalized
1326 scenarios. 32 probable coincidences amongst 12 events. Density: 0.48. Components: 1.
12 events(n>=5): Vida Playa Familia Disfrutar Mar Amigos Pelo Sonrisa Amor Sol Verano Sentir
```

Haberman residuals	Vida	Playa	Fam~a	Dis~r	Mar	Ami~s	Pelo	Son~a	Amor	Sol	Ver~o	Sen~r
Vida	36.4											
Playa	-2.3	36.4										
Familia	-0.6	4.4	36.4									
Disfrutar	2.7	0.6	2.2	36.4								
Mar	-0.6	0.6	-1.4	2.8	36.4							
Amigos	-1.8	3.3	8.2	3.3	0.5	36.4						
Pelo	-0.3	-0.8	-0.7	0.1	2.6	-0.7	36.4					
Sonrisa	1.2	-1.9	-1.9	-1.1	-0.1	-0.7	-0.7	36.4				
Amor	2.3	-1.7	-1.0	-0.9	-0.6	-1.3	1.9	-1.3	36.4			
Sol	-0.4	5.3	-1.7	0.5	2.5	0.3	2.0	1.2	0.6	36.4		
Verano	-0.9	2.8	2.4	-0.8	0.2	2.1	1.2	-1.3	-1.2	-0.2	36.4	
Sentir	1.3	-1.7	-1.0	-0.8	0.2	-1.3	3.7	-1.3	1.6	-1.1	1.7	36.4

coin example (IV)

Adjacencies matrix in L'Oreal's messages

```
. coin Vida-Mar Amigos-Sentir, adjace
1326 scenarios. 32 probable coincidences amongst 12 events. Density: 0.48. Components: 1.
12 events(n>=5): Vida Playa Familia Disfrutar Mar Amigos Pelo Sonrisa Amor Sol Verano Sentir
```

Adjacency matrix	Vida	Playa	Fam~a	Dis~r	Mar	Ami~s	Pelo	Son~a	Amor	Sol	Ver~o	Sen~r
Vida	0.0											
Playa	0.0	0.0										
Familia	0.0	1.0	0.0									
Disfrutar	1.0	1.0	1.0	0.0								
Mar	0.0	1.0	0.0	1.0	0.0							
Amigos	0.0	1.0	1.0	1.0	1.0	0.0						
Pelo	0.0	0.0	0.0	1.0	1.0	0.0	0.0					
Sonrisa	1.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0				
Amor	1.0	0.0	0.0	0.0	0.0	0.0	1.0	0.0	0.0			
Sol	0.0	1.0	0.0	1.0	1.0	1.0	1.0	1.0	1.0	0.0		
Verano	0.0	1.0	1.0	0.0	1.0	1.0	1.0	0.0	0.0	0.0	0.0	
Sentir	1.0	0.0	0.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0	1.0	0.0

coin example (V)

Centrality measures in L'Oreal's messages

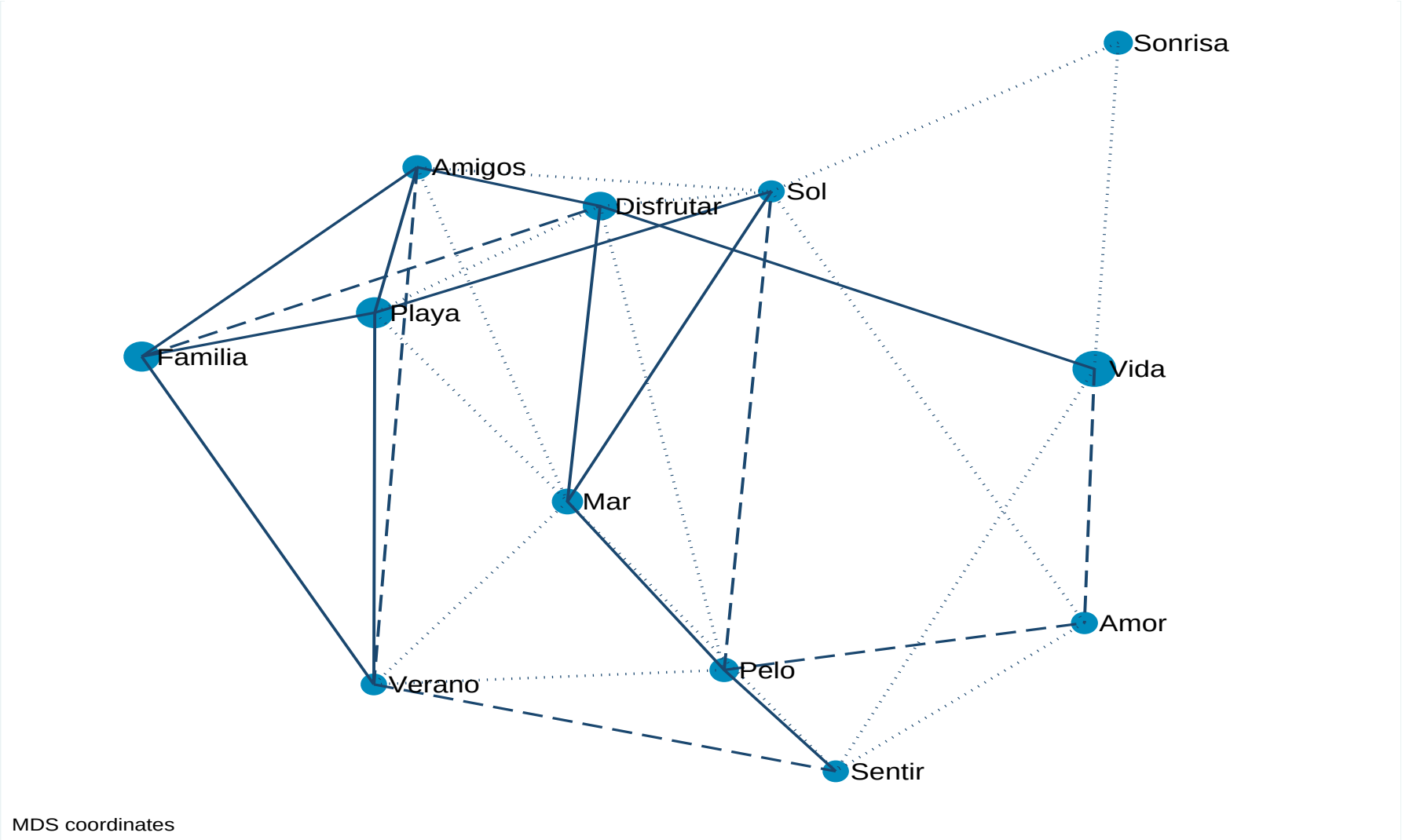
```
. coin Vida-Mar Amigos-Sentir, centrality
1326 scenarios. 32 probable coincidences amongst 12 events. Density: 0.48. Components: 1.
12 events(n>=5): Vida Playa Familia Disfrutar Mar Amigos Pelo Sonrisa Amor Sol Verano Sentir
```

Centrality measures	Degree	Close	Between	Inform
Vida	0.36	0.61	0.06	0.07
Playa	0.55	0.69	0.03	0.09
Familia	0.36	0.55	0.00	0.07
Disfrutar	0.64	0.73	0.12	0.10
Mar	0.64	0.73	0.05	0.10
Amigos	0.55	0.69	0.03	0.09
Pelo	0.55	0.69	0.05	0.09
Sonrisa	0.18	0.50	0.01	0.05
Amor	0.36	0.58	0.02	0.07
Sol	0.64	0.73	0.18	0.10
Verano	0.55	0.65	0.06	0.09
Sentir	0.45	0.65	0.05	0.08

coin example (VI)

Simple graph

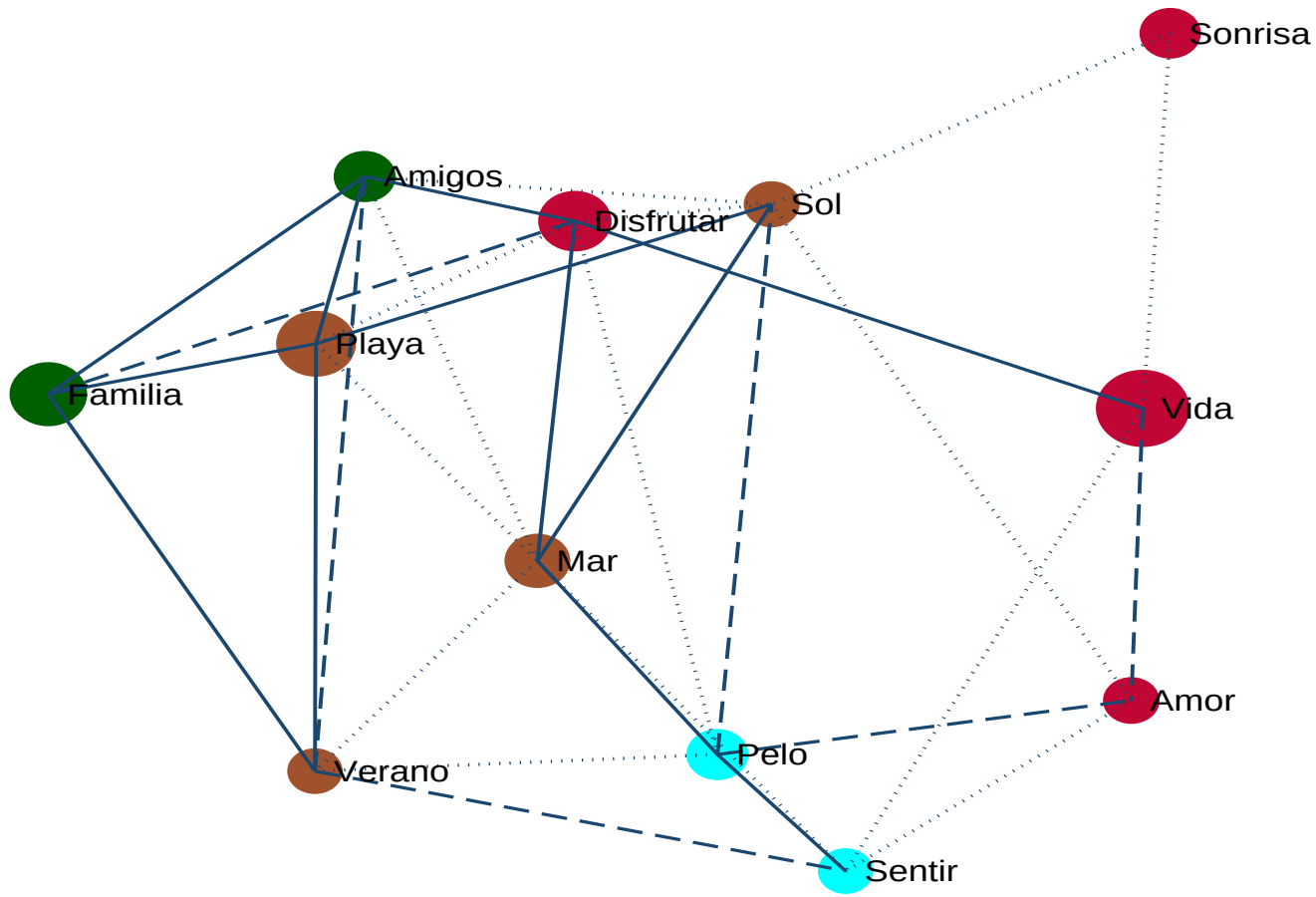
coin Vida-Mar Amigos-Sentir, graph(mds) levels(.5 .05 .01) goptions(name(Network))



coin example (VII)

Color graph

coin Vida-Mar Amigos-Sentir using Words, graph(mds) levels(.5 .05 .01) color(Tipo) legend



coin example (VIII)

Words in their context

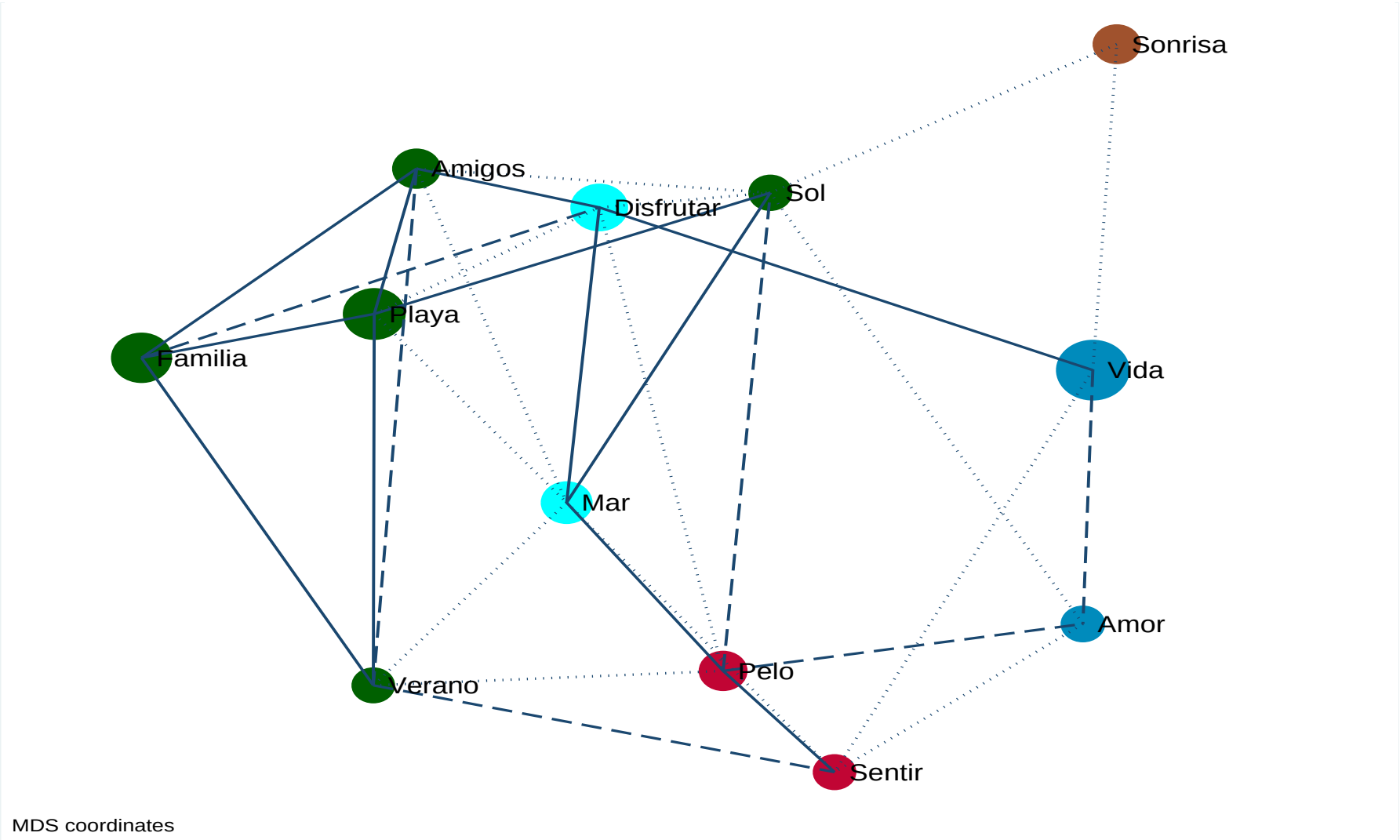
```
. list Mensaje if Pelo & Amor, clean string(120)

       Mensaje
 234.  Que hay más auténtico que tu hija de 1 año acariciándote el pelo puro amor
 237.  Que hay más auténtico que tu hija de 1 año acariciándote el pelo puro amor
 449.  Disfrutar de un atardecer con el sonido de las olas y el aroma en mi pelo de Original Remedies, en compñía del amor de m..
 636.  Lo realmente auténtico es el amor de mi familia. A mi hermana y a mi nos encanta peinarnos y tener un pelo suave, con br..
```

coin example (IX)

Automatic color graph (Communities)

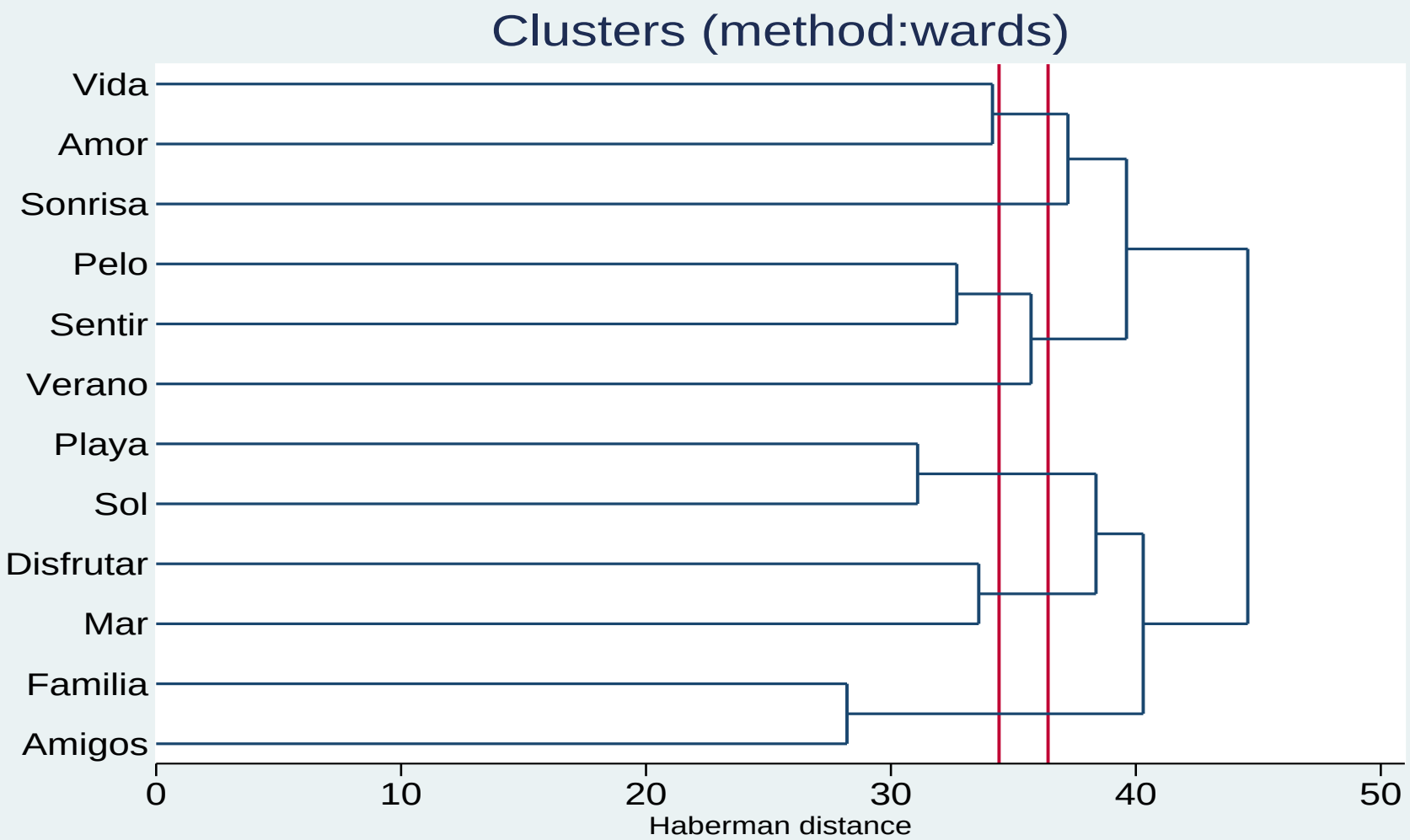
coin Vida-Mar Amigos-Sentir using Words, graph(mds) groups(5)



coin example (X)

Color graph

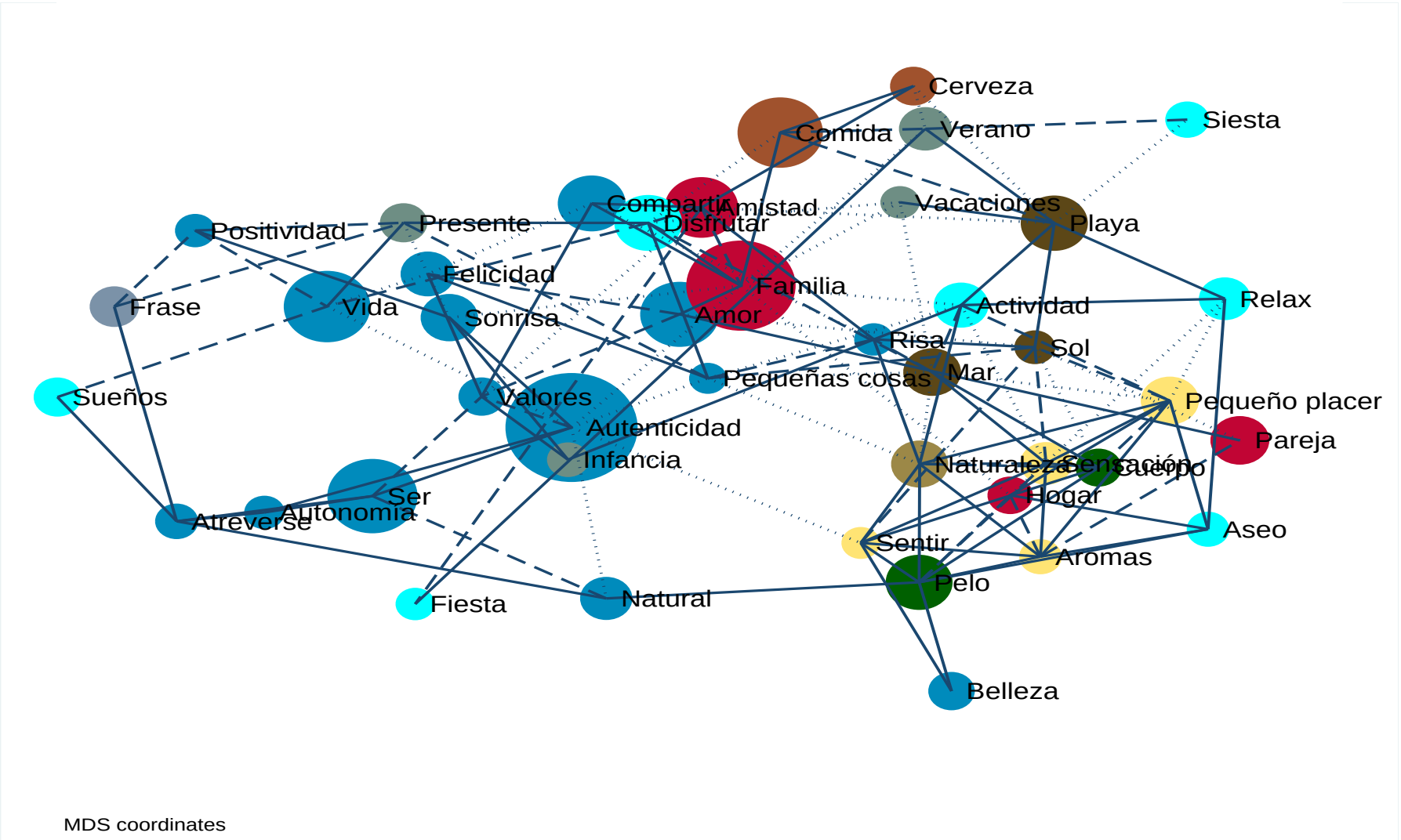
coin Vida-Mar Amigos-Sentir, dendrogram(ward)



coin example (XI)

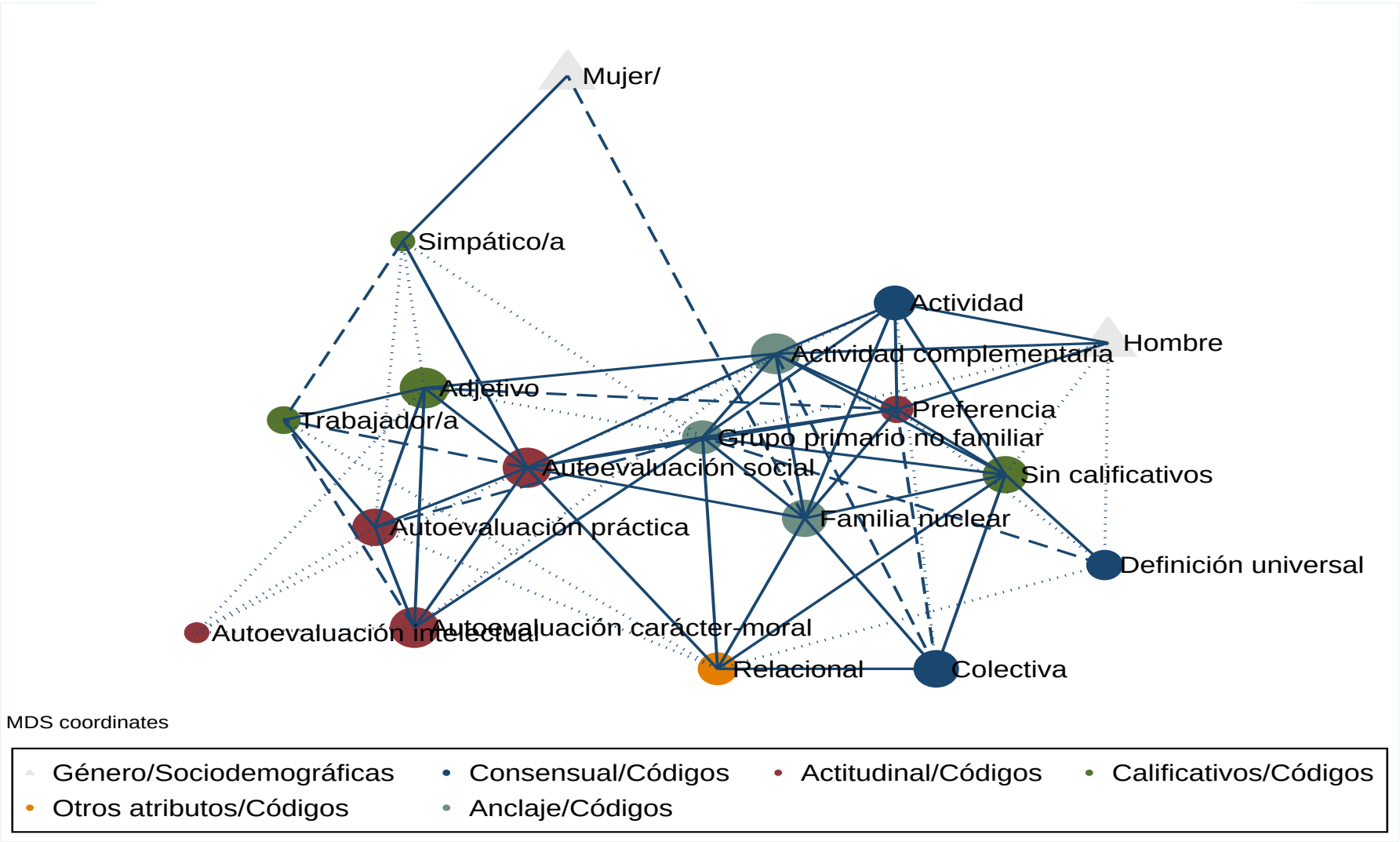
Graph with manual codes

coin K1-K69 using Nodes, graph(mds) color(tipo)



Last example

Components of self identity



Availability of precoin and coin

Frame Subtitle

- If you are an user of a version superior to the 11.2 of Stata, you can have a free copy of `coin` by typing:
  - net install coin, from(http://sociocav.usal.es/stata/)
- It is still their first version, but it works reasonably well and it is being improved. It could be updated as follows:
  - adoupdate, update
- Comments and suggestions will be welcome!!

Next steps

For coin and precoin

- Automatic codification through regular expressions.
- Similar graphs representation of correlations among quantitative variables.
- Use of log-lineal models to discover n-coincidences.
- Time based study of coincidences using dynamic networks.
- Using objects in the Mata code of the command coin.
- It would be great if Stata implemented sparse matrices in Mata!!.

References

Bernald Berelson. *Content Analysis in Communication Research*. Free Press., New York, 1952.

Ole R. Holsti. *Content Analysis for the Social Sciences and Humanities*. Addison-Wesley., Reading, 1969.

Klaus Krippendorff. *Content Anlysis. An Introduction to its Methodology*. Sage., Beverly Hills, 1980.

Rense Corten. Visualization of social networks in Stata using multidimensional scaling. *The Stata Journal*, 11(1):52–63, 2011.

Hirotaka Miura. Stata graph library for network analysis. *The Stata Journal*, 12(1):94–129, 2012.

Thomas E. Grund. nwcommands: Software tools for statistical modeling of network data in Stata, 2014. URL `http://nwcommands.org`.

Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(2):311–331, 2003.

Last slide

Thanks

Thank you very much!

modesto@usal.es & berrocal@usal.es