Text mining a Portuguese book on Freemasonry: Disclosing network communities' features

Conceição Rocha[1,3]; Alípio Mário Jorge[2,3]; Márcia Oliveira[1,3];
Paula Brito[1,3]; João Gama[1,3]; Carlos Pimenta[1,4]

[1]Faculdade de Economia da Universidade do Porto

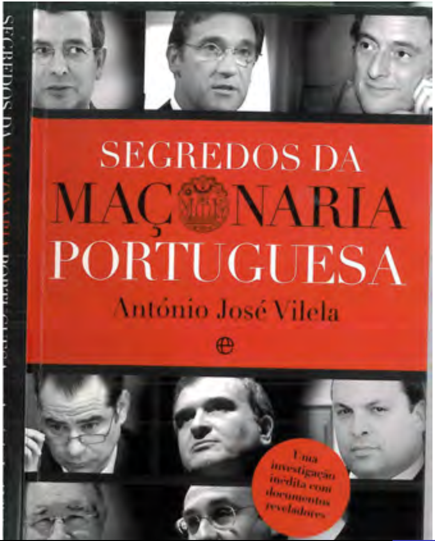
[2]Faculdade de Ciências da Universidade do Porto

[3]LIAAD/INESC TEC

[4]OBEGEF

INForum 2014

Objective



Extract named entities from a Portuguese book on Freemasonry and explore network communities based on their co-occurrences in the same sentences

Difficulties

Some difficulties:

1. Pre-processing the text
   - the structure of the scanned book text - page breaks
   - 'junk' like page numbers
   - mistakes/limitations of Optical Character Recognition (OCR)
2. Named entities extraction
   - limitations of free software on Portuguese language
   - different designations used for the same entity

Software

Program developed in R

packages: **tm**, **gdata**, **stringr**, **cwhmisc**, **openNLP** and **Hmisc**

Social network analysis - Gephi software

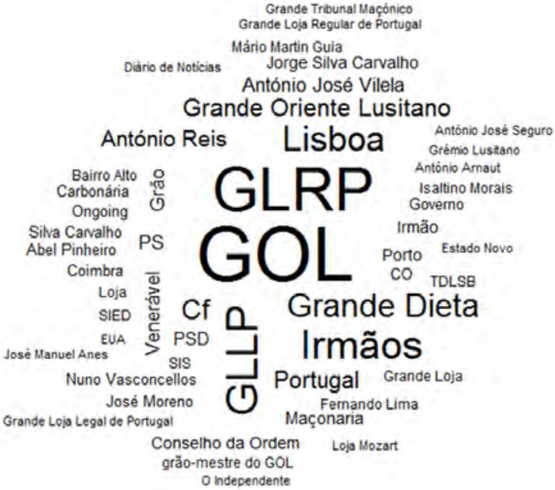
Methodology

Process main steps:

- Phase 1
  1. remove page numbers and empty lines
  2. remove 'junk' based on their patterns
  3. extract the named entities using regular expressions (capital letters and lower — *e.g. presidente*, *câmara*, *deputado*)
- Phase 2
  1. tag terms list as part-of-speech
  2. remove all the terms that do not have at least one tag 'prop'
  3. remove the first word from terms starting by 'pron-det'
  4. remove some stop words
  5. identify the named entities

Word cloud

Entities appearing 20 or more times in the text



Validation

12650 events corresponding to 5502 unique terms in the book

To evaluate the term extraction:

125 book' pages with 3866 named entities have been manually labeled (1/3 of the text book)

	Phase 1	Phase 1 + Phase2
extracted terms	5089	3075
named entities	3205	2982
recall	0.84	0.78
precision	0.63	0.97
$F - measure$	0.72	0.865

Network characteristics

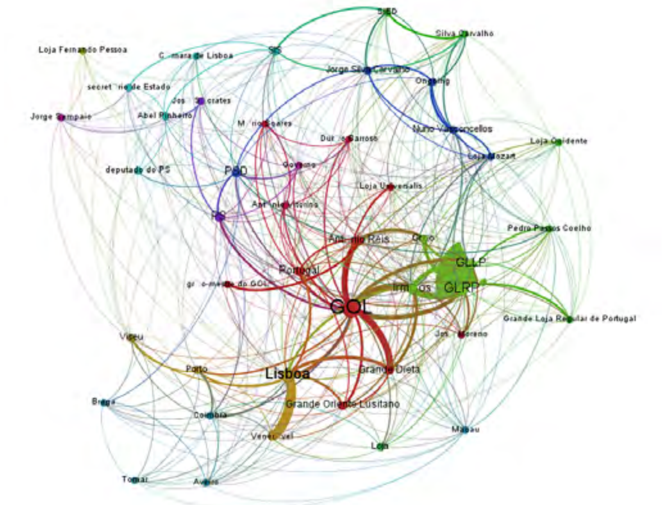
4730 nodes and 24997 edges
undirected and weighted graph

Table : Statistics attributes

Network	average clustering coefficient	0.851
	average path length	3.445
	average degree	10.57
	average weighted degree	11.75
	network diameter	12
	network radius	1
	graph density	0.002
	modularity	0.682
	$N^{\underline{o}}$ of communities	268
	$N^{\underline{o}}$ weakly connected components	238

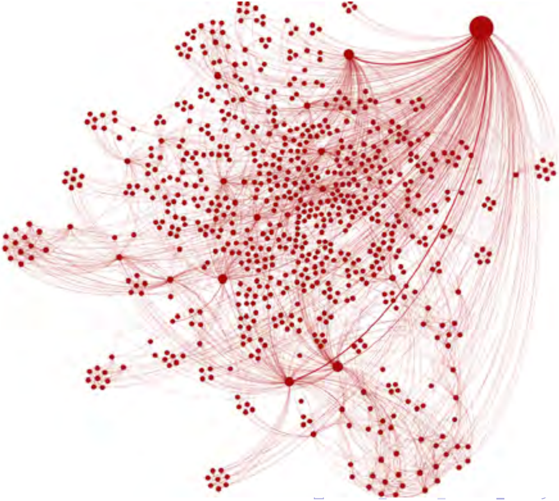
Social network graph

Social network graph filtered by degree - minimum 81

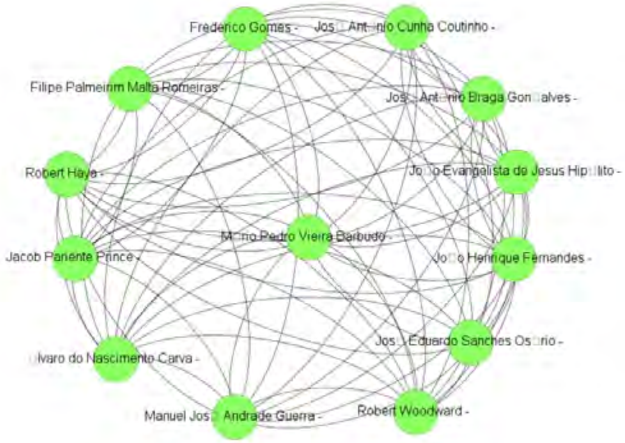


Main community

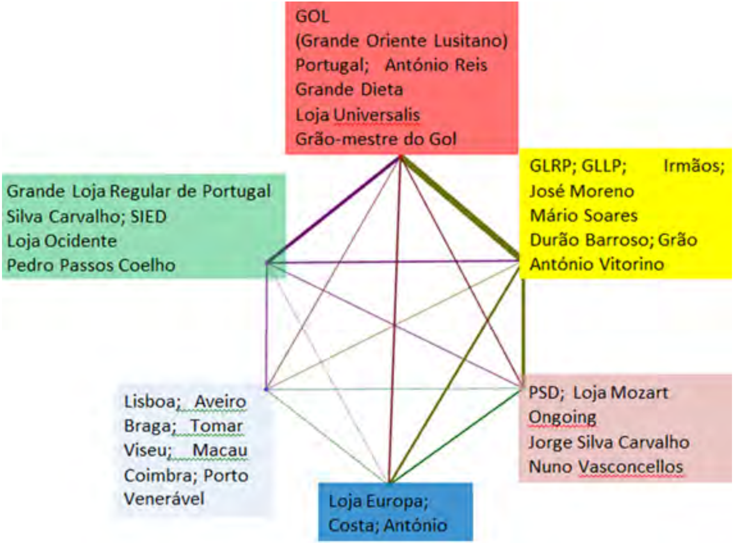
- comprises the majority of nodes
- heterogeneous
- exhibits a rich internal structure



A component and a clique of the network


Frederico Gomes -
Jos Ant nio Cunha Coutinho -
Filipe Palmeirim Malta Romeiras -
Jos Ant nio Braga Gon alves -
Robert Haye -
Jo o Evangelista de Jesus Hip lito -
M rio Pedro Vieira Barbudo -
Jacob Parente Prince -
Jo o Henrique Fernandes -
Jos Eduardo Sanches Os rio -
lvaro do Nascimento Carva -
Manuel Jos Andrade Guerra -
Robert Woodward -

Higher-level representation of the Social network



GOL
(Grande Oriente Lusitano)
Portugal; António Reis
Grande Dieta
Loja Universalis
Grão-mestre do Gol

GLRP; GLLP; Irmãos;
José Moreno
Mário Soares
Durão Barroso; Grão
António Vitorino

Grande Loja Regular de Portugal
Silva Carvalho; SIED
Loja Ocidente
Pedro Passos Coelho

Lisboa; Aveiro
Braga; Tomar
Viseu; Macau
Coimbra; Porto
Venerável

PSD; Loja Mozart
Ongoing
Jorge Silva Carvalho
Nuno Vasconcellos

Loja Europa;
Costa; António

Remarks and Conclusions

⇒ Inclusion of the second phase on the process improves the quality

⇒ $F - measure$ increases from 0.72 to 0.865

Considering that:

- the text mining procedure to extract entity names is not finished
- the relation between entities is given by their co-occurrence in the same sentence

⇒ The results are quite meaningful and we can see relevant connections in terms of some political organizations, politicians and other public figures

Further Work

- including an entity synonymy step and a disambiguation step
- adjusting the network model so that links between entities are based on the verbs

The results obtained so far may also be considered a step towards the creation of a text intelligence system to be used in the study of the social context of possible economic and financial offenses.

Acknowledgments

This work is partially funded by FCT/MEC through PIDDAC and ERDF/ON2 within project NORTE-07-0124-FEDER-000059 and through the COMPETE Programme (operational programme for competitiveness) and by National Funds through the FCT - Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) within project FCOMP-01-0124-FEDER-037281. Márcia Oliveira acknowledges funding from FCT, through PhD grant SFRH/BD/81339/2011.