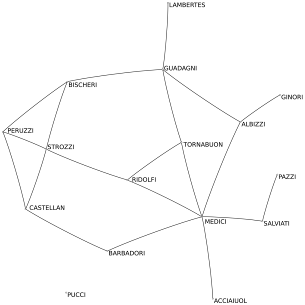
# Mining parliamentary data and news articles to find patterns of collaboration between politicians and third party actors.

Francisco Rodríguez Drumond

DAMA & LARCA - UPC

July 7,2014

# Social Networks: a natural tool for political analysis.



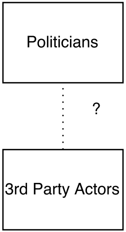

- *Nodes*: Families of the political landscape of XV century Florence.
- *Links*: marriages between families (alliances).

# Analizing parliaments through SNs.

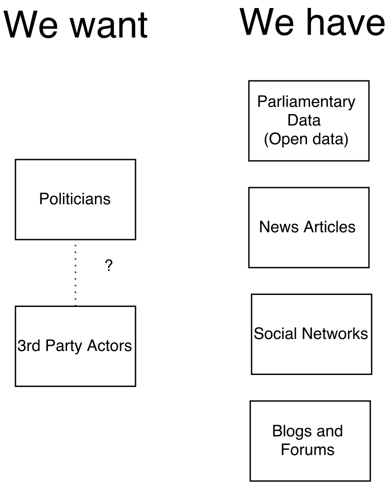
- Why?
- Main challenge: source of information (nodes and relationships)
    - Co-sponsorship. [Fow06]
    - Speeches. [TPL06]
    - Strong and weak ties. [Kir11]
- Can we discover relationships involving third-party actors?
    - Third party discovery
    - Defining meaningful relationships.

# An overview of our task

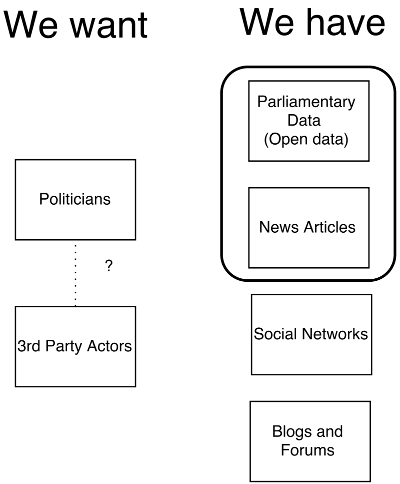
We want



# An overview of our task



# An overview of our task



# An overview of our task

**We want**

Politicians

?

3rd Party Actors

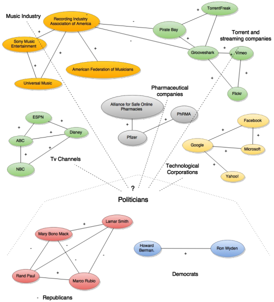
**We have**

Parliamentary Data (Open data)

News Articles

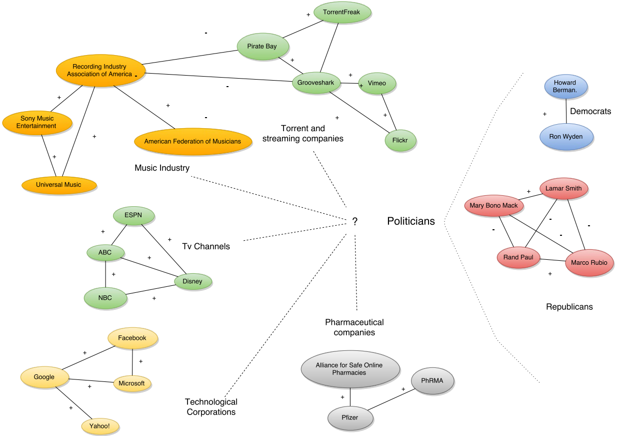
Social Networks

Blogs and Forums



# SOPA: A motivating example.

Policy Networks (PN): Social networks for political analysis.



# An overview of the literature.

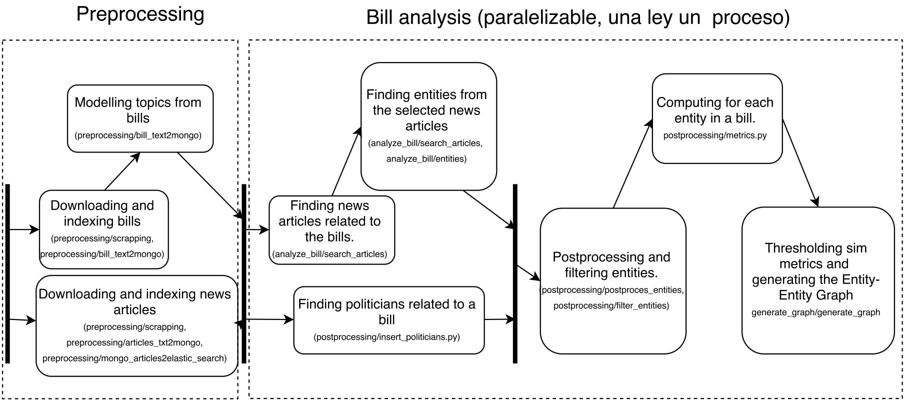
- Co-occurrence. [EESGGHAC14], [PSIO06].
- Enriching links with the strength and semantics of relations. [Tan07],[PSB07],[ZAR03].
- Beyond document co-occurrence. [NCSS06],[Bra06].
- A (very) related paper. [MID$^{+}$13]

# A (very) related paper.

Moschopoulos (2013) *Toward the automatic extraction of policy networks using web links and documents*

- Two pre-computed PNs: Ireland and Greece.
- Ground truth used for measuring correlations with similarity measures.
- Web based.
- Three types of similarity metrics:
  - Co-occurrence metrics (Set comparisons).
  - Text-based metrics.
  - Link-based metrics.

# Generating bill based Policy Networks: the architecture.



# Finding news articles that talk about a bill.

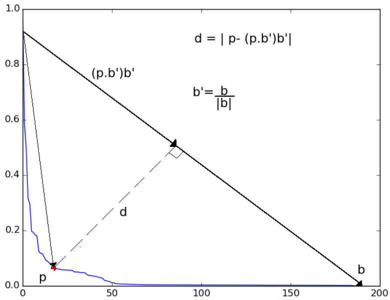
## Topic modeling:

- TF-IDF for keyword extraction.
  - One bill - one document.
  - Whole set of bills as the corpus.
- 1,2,3-ngrams.
- Top 1000 keywords for each bill.

## Querying news articles:

- Bills and news articles modeled as vectors
  - Cosine similarity for comparison.
- Rocchio's rule for improving queries.

# Selecting relevant news articles.




Threshold: point that maximizes:

$$threshold = \underset{p}{\mathrm{argmax}} |p - (p.b')b'|$$

*Intuition:* point at which there is no significant gain in score.

# Entity extraction and preprocessing.

MITIE for entity extraction +

1. **Entity Normalization**
   - 'The Univ. lumiere Lyon 2' → 'Univ Lumiere Lyon 2'
2. **Mapping organization initials to the whole name**
   - 'The **World Life Fund** (**WLF**) has...'
     → 'World Life Fund' = 'WLF'
3. **Mapping partial names with full names**
   - '**George Harrison** preferred .... **Harrison** also...'
     → 'George Harrison' = 'Harrison'
4. **Expanding names based on the news corpus**
   - 'Politècnica de Catalunya'
     → 'Universitat Politècnica de Catalunya'

# Filtering relevant entities.

**Problem:** +3000 entities per bill

- Noise.
- Expensive comparisons.

**Solution:**

- Document co-occurrence + Latent Semantic Indexing (LSI) for fast similarity computation.
- Hierarchical Agglomerative Clustering (HAC) for grouping entities based on their similarity.
  - Politicians $\rightarrow$ seed entities.
- Silhoutte for detecting the best cluster containing seed entities.

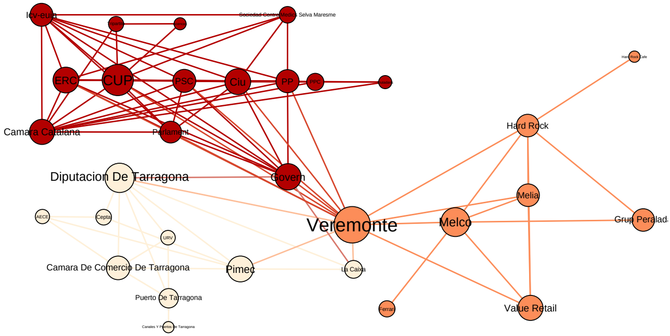
# Computing and thresholding entity similarities.

- Entities represented as vectors of 1...3-grams occurring in paragraphs they are mentioned in.
  - TF-IDF with sublinear TF scaling ($tf = 1 + log(frequency)$)
- Cosine similarity for comparing the vectors.
- Elbows for detecting relevant entities for each entity.
  - Two entities $e1$ and $e2$ are related iff they are in each others relevant entities list.

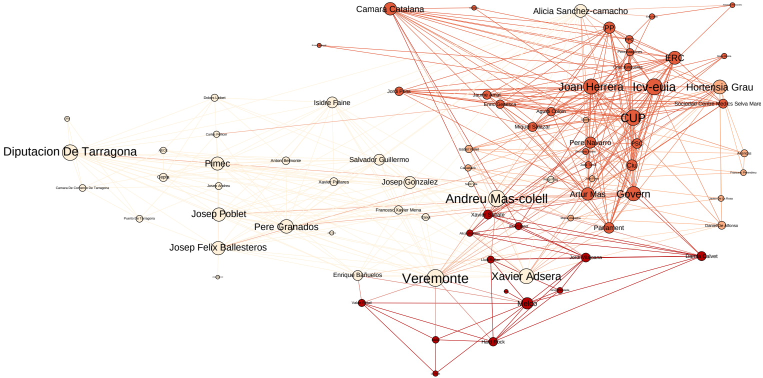
# Results.

- Two bills:
    - BCN-World.
    - Law of Popular Non-referendary Consults.
- Look at:
    - Communities → colors.
    - Influencers → node size.

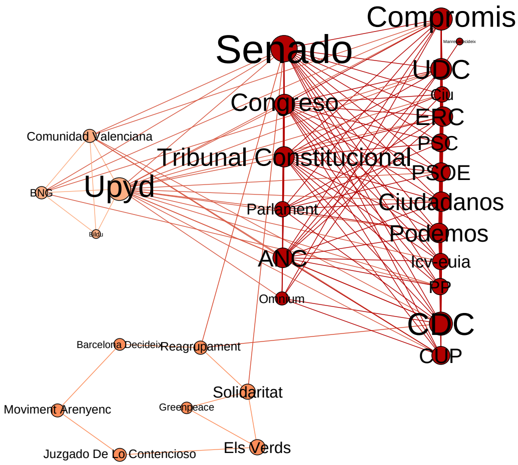
# BCN-World - Organizations.



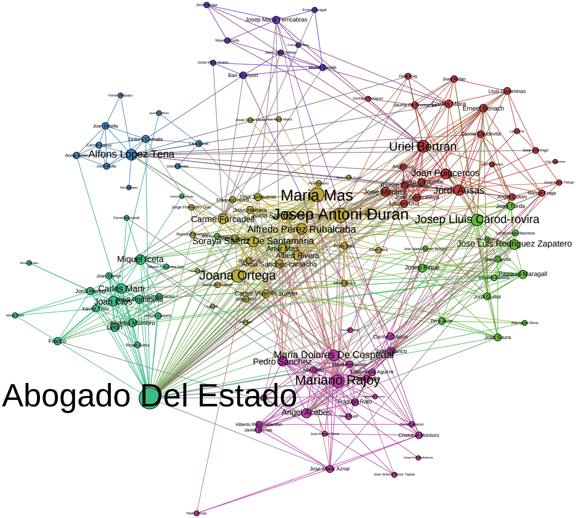
# BCN-World - Persons-Organizations.



# Law of Popular Non-referendary Consults. - Organizations.



# Law of Popular Non-referendary Consults. - Persons.



# Conclusions.

1. An unbiased, low-cost, automated tool to aid the process of Policy Network generation and analysis.
2. The system automatically:
   1. Detect entities related to a bill.
   2. Computes and thresholds similarity measures for SN generation.
3. The method works better for finding relationships between organizations than for persons, particularly politicians.

# Contributions.

1. The use of bills as a cornerstone relating political actors, allowing to:
    - Understand better the discovered relations.
    - Find fine-grained relationships which would otherwise be missed.
2. A method for combining parliamentary open data and news papers for PN generation.
3. An unsupervised method for automatically detecting relevant entities of a given topic from a corpus of documents given a set of seed entities.

# Future work.

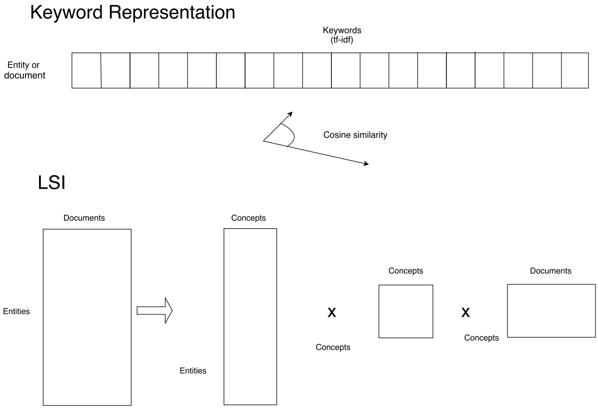
1. A more rigorous evaluation and problem definition.
2. Improving the PN generation phase.
3. Generative models.
4. Use-case driven PN generation.
5. Time component.
6. Signed Social Network Analysis

# The end.

Merci beacoup!
Gràcies!
Grazie!
Mulțumesc!

Questions?

# Understanding the representation of entities and documents.



# References I


- Roger B Bradford, *Application of latent semantic indexing in generating graphs of terrorist networks*, Intelligence and Security Informatics, Springer, 2006, pp. 674–675.
- Jesús Espinal-Enríquez, J Mario Siqueiros-García, Rodrigo García-Herrera, and Sergio Antonio Alcalá-Corona, *A literature-based approach to a narco-network*, Social Informatics, Springer, 2014, pp. 97–101.
- James H Fowler, *Connecting the congress: A study of cosponsorship networks*, Political Analysis **14** (2006), no. 4, 456–487.
- Justin H Kirkland, *The relational determinants of legislative outcomes: Strong and weak ties between legislators*, The Journal of Politics **73** (2011), no. 03, 887–898.

# References II


- Theodosis Moschopoulos, Elias Iosif, Leeda Demetropoulou, Alexandros Potamianos, and Shrikanth Shri Narayanan, *Toward the automatic extraction of policy networks using web links and documents*, Knowledge and Data Engineering, IEEE Transactions on **25** (2013), no. 10, 2404–2417.
- David Newman, Chaitanya Chemudugunta, Padhraic Smyth, and Mark Steyvers, *Analyzing entities and topics in news articles using statistical topic models*, Intelligence and Security Informatics, Springer, 2006, pp. 93–104.
- Bruno Pouliquen, Ralf Steinberger, and Clive Best, *Automatic detection of quotations in multilingual news*, Proceedings of Recent Advances in Natural Language Processing, 2007, pp. 487–492.

# References III


- Bruno Pouliquen, Ralf Steinberger, Camelia Ignat, and Tamara Oellinger, arXiv preprint cs/0609066 (2006).
- Hristo Tanev, *Unsupervised learning of social networks from a multiple-source news corpus*, MuLTISOuRcE, MuLTILINguAL INfORMATION ExTRAc-TION ANd SuMMARIzATION (2007), 33.
- Matt Thomas, Bo Pang, and Lillian Lee, *Get out the vote: Determining support or opposition from congressional floor-debate transcripts*, Proceedings of the 2006 conference on empirical methods in natural language processing, Association for Computational Linguistics, 2006, pp. 327–335.

# References IV

Dmitry Zelenko, Chinatsu Aone, and Anthony Richardella, *Kernel methods for relation extraction*, The Journal of Machine Learning Research **3** (2003), 1083–1106.