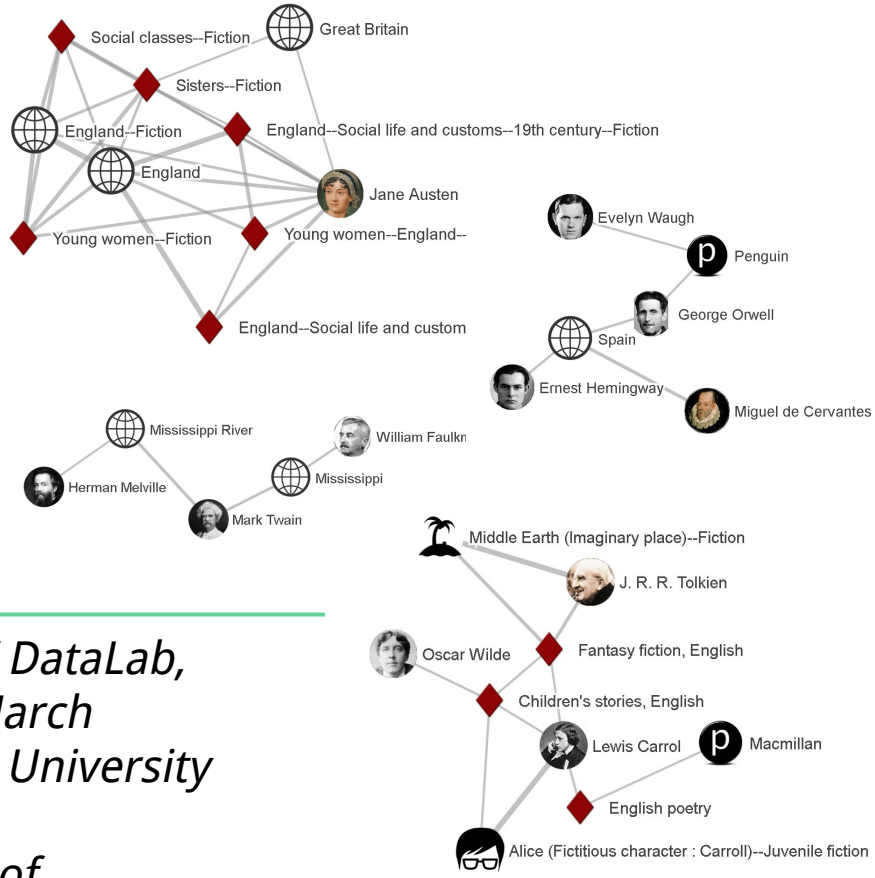
Analyzing big and small collections of books with Network Coincidence Analysis



Luis Martinez-Uribe *University of Salamanca / DataLab, Library, Fundación Juan March*

Modesto Escobar *Department of Sociology , University of Salamanca*

Carlos Prieto *Bioinformatics, University of Salamanca*

David Barrios *Bioinformatics, University of Salamanca*

BigSurv18

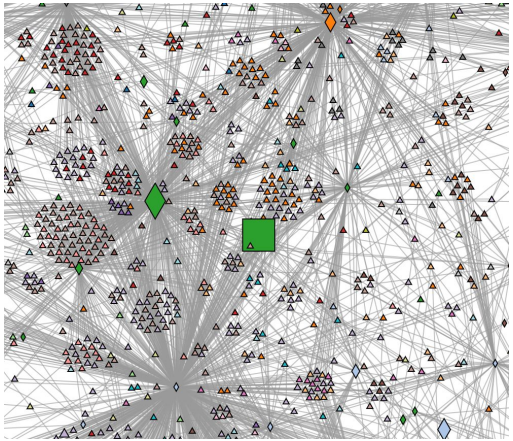
OCTOBER 25–27, 2018 ■ BARCELONA, SPAIN

Background

- ***Computational social science*** emerging as unprecedented capacity to collect and analyze data in depth and scale offering society a better understanding of individuals and collectives (*Lazer et al. 2009*)
- ***Exploratory data analysis*** as part of the research process of understanding, pattern-seeking, comparison-making and knowledge-gaining (*Tukey 1977*)
- ***Visual analytics*** as interactive means in assisting with synthesizing information, detecting patterns and communicating results (*Keim et al. 2008; Healy & Moody 2014*)
- ***Cultural/library curated data as alternative data sources*** available for research (*Moretti 2005; Cohen 2006; Tuppen 2016*)

Motivation

When exploring visually large graphs, there is limited number of pixels, computer processing power and user brainpower → ***network hairball***



Research question

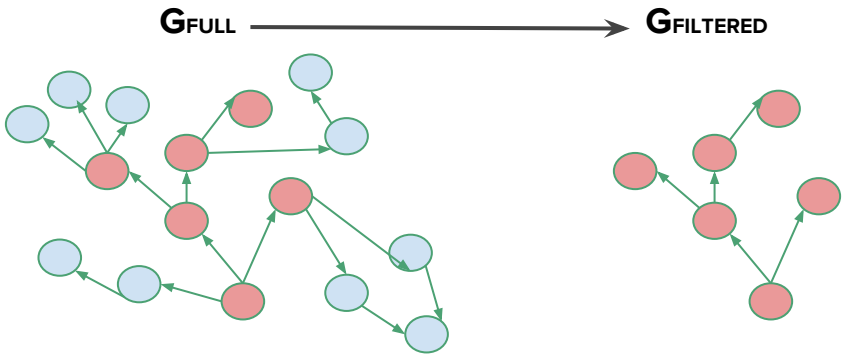
- How can we explore, visualize and interact with large social networks?

Contribution

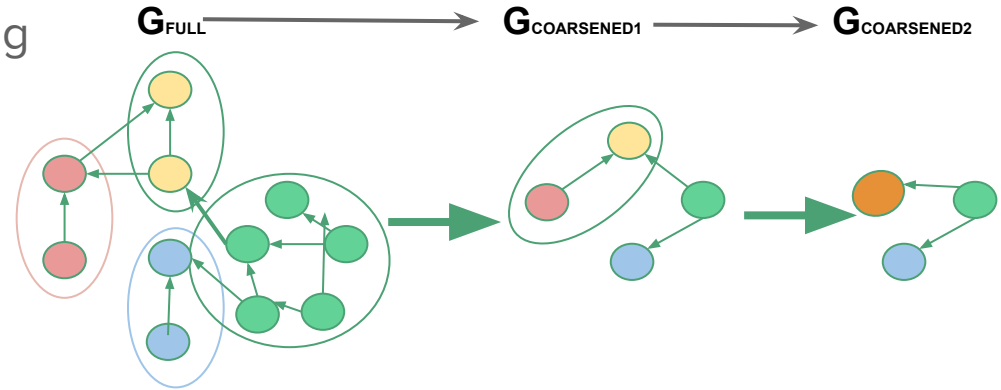
- We present a combination of methods that help representing smaller versions of large graphs through several case studies using the netCoin R package to create interactive network visualizations

Strategies to visualize large networks

1. Filtering of nodes and edges to show subsections of the graph
    a. Stochastic (random selection of network elements)
    b. Deterministic (selection based on properties of nodes and edges)



2. Coarsening, i.e. grouping and merging to reduce network density
    a. According to node attributes
    b. Based on node hierarchies
        i. Thesaurus
        ii. Clustering classification



Method: Network Coincidence Analysis

Aims to detect events, characters, objects or attributes that tend to occur together within limited spaces (Escobar 2015).

- The spaces are call ***scenarios*** (S) and are the units of analysis.
- In each scenario, a series of J ***events*** X_j, may or may not occur. Scenarios and events constitute an ***incidence matrix (I)***.
- From ***I***, a ***coincidence symmetric matrix (C)*** can be obtained with frequencies of X_j in the diagonal, and number of coincidences between two events in the rest of the matrix.
- From ***C***, similarity matrices (Matching, Jaccard, Hamman, etc) and other measures (frequencies, Haberman, etc) can be obtained.

Tool: netCoin R package

R package: interactive analytic networks (Escobar, Barrios, Prieto, Martinez-Uribe 2018)

- Apply Network Coincidence Analysis.
- Create interactive networks using D3.js Javascript library via web browser with the capacity to:
  - Modify the label, size, color and shape of nodes
  - Adjust width, weight, color and text of the edges
  - Filter manually or dynamically nodes and edges

https://cran.r-project.org/web/packages/netCoin/index.html

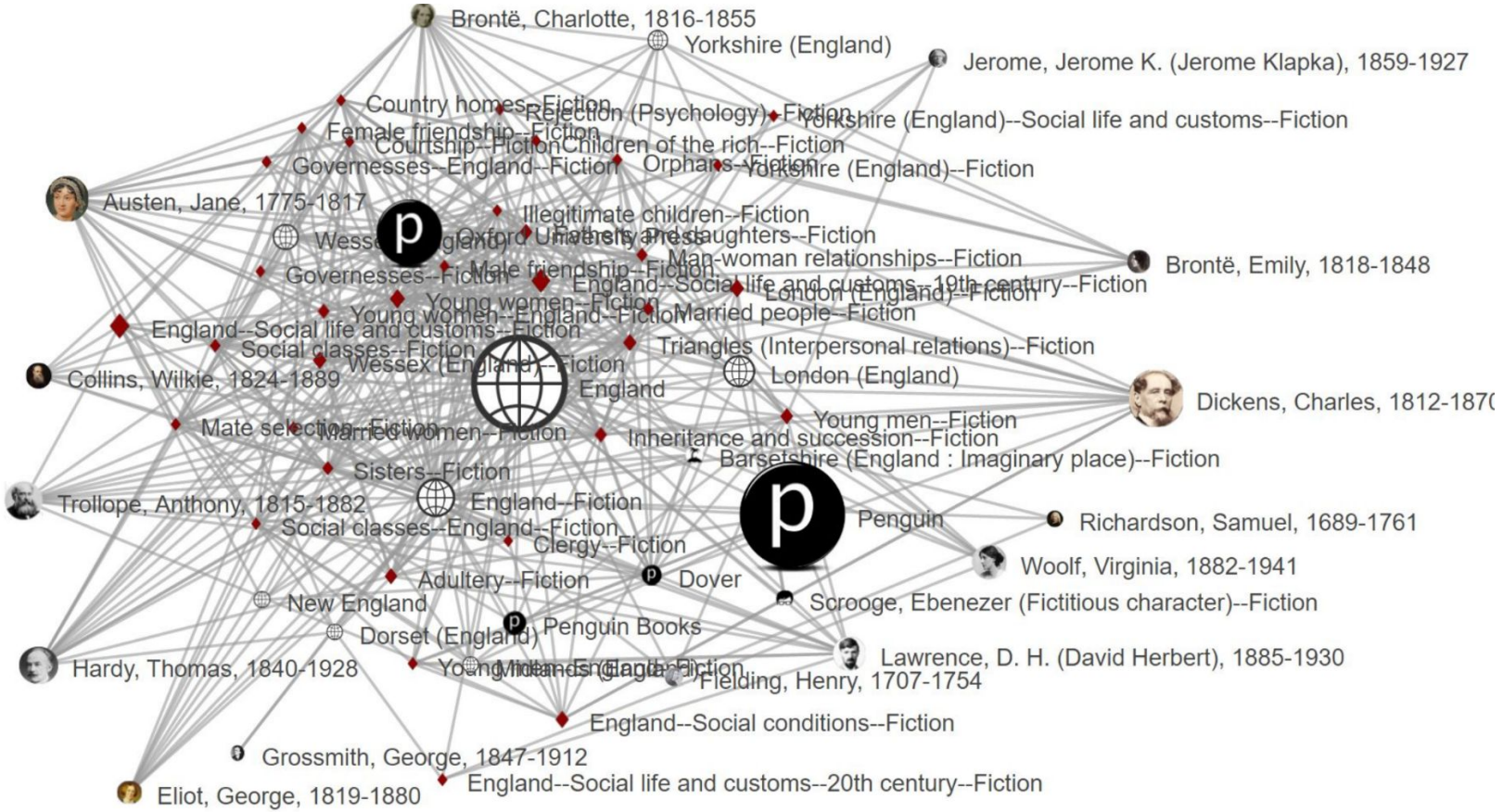
Data: British National Bibliography

- The British Library's **British National Bibliography** data
  - 3M records with 1.4M subjects
- The Guardian's **100 best novels of all time**
- Selection of 13K books written by the 100 greatest authors, and published between 1950-2015.
- Wikipedia (data and images)

Data dimensions:

- Scenarios: 13.216 books.
- Events: 100 authors, 120 subjects, and 20 publishers.

First interactive network



Data: Microsoft Academic Graph Dataset

Publications	Authors	Fields of Study	Journals	Conferences	Institutions
206,252,196	249,999,501	229,304	47,989	4,029	25,374

Big Data cloud infrastructure

Arnab Sinha, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June (Paul) Hsu, and Kuansan Wang. 2015. An Overview of Microsoft Academic Service (MAS) and Applications. In *Proceedings of the 24th International Conference on World Wide Web* (WWW '15 Companion). ACM, New York, NY, USA, 243-246. DOI=http://dx.doi.org/10.1145/2740908.2742839

Second interactive network

1. Use Fields of Study (FoS) to impose a smaller structure (230k FoS) on top of 200 million publications
2. Filter the top 400 terms at levels 1,2 and 3 (nodes) plus their 700 links (edges)
   a. ***Node size***= # publications with that FoS
   b. ***Links*** = FoS thesaurus relationships

Sample: a selection of sociology publications

Top 20 Sociology Journals*

* According to the 2015 ArticleInfluence (TM) Score in

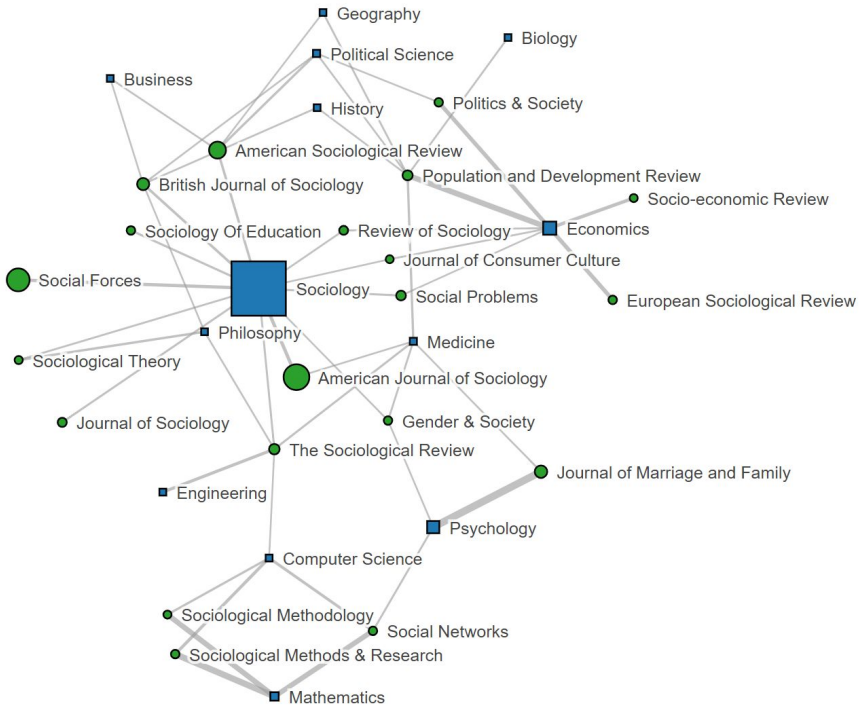
ISI Journal Citation

Publications	Fields of Study
82.903	**12.960**

Journal Title	MAG articles in Azure
Americal Journal of Sociology	21459
Social Forces	17943
American Sociological Review	11012
Journal of Marriage and Family	5789
British Journal of Sociology	5458
The Sociological Review	3480
Population and Development Review	2972
Social Problems	2518
Journal of Sociology	1772
Review of Sociology	1500
Sociology of Education	1176
European Sociological Review	1163
Politics & Society	1141
Social Networks	1113
Sociological Methods & Research	1104
Gender & Society	1089
Sociological Theory	675
Sociological Methodology	569
Socio-Economic Review	501
Journal of Consumer Culture	469

Third interactive network

1. Use sample data 98K articles from top 20 sociology journals and 13K FoS
2. Apply Coincidence Analysis to top 2.300 most frequent FoS and 20 journals (nodes) and their 30K links (edges)
    a. ***Node size***= frequencies
    b. ***Links*** = coincidence analysis relations
    c. ***Tree*** = FoS tree structure to initially show only level 1 FoS and expand when needed



Conclusion and next steps

- The combination of tools for social network exploration, Coincidence Analysis, filtering and coarsening strategies supports the process of conducting visual analytics on large graphs
- Cultural/Library curated data as yet another data source available to researchers
- Further improve coarsening (group and merge) and un-coarsening functionalities in the netCoin R package
- Take advantage of R shiny interfaces to bring the computation capacity for data processing and cluster generation on the fly for large datasets