Distant Reading and Visualization

A.K.A. Tools for text analysis and visualization

Rachele Sprugnoli
sprugnoli@fbk.eu

FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES

Close Reading

- Traditional method in literary criticism
- Text interpretation based on interactions between a human reader and a text

Unveil words, verbal images, elements of style, sentences, argument patterns
(Jasinski, 2001)

Distant Reading

- New method introduced in literary criticism
- Text interpretation based on general features and abstract models

A condition of knowledge: it allows you to focus on units that are much smaller or much larger than the text.
(Moretti, 2000)

Other Definitions of Distant Reading

The construction of abstract models

Jasinski, “Sourcebook on Rhetoric”, 2001

A macroanalytic approach

Matthew Jockers, “On Distant Reading and Macroanalysis“, 2011

The idea of processing content in or information about a large number of textual items without engaging in the reading of the actual text.

Johanna Drucker, “Distant Reading and Cultural Analytics”, 2013

From Texts to Models to Visualizations

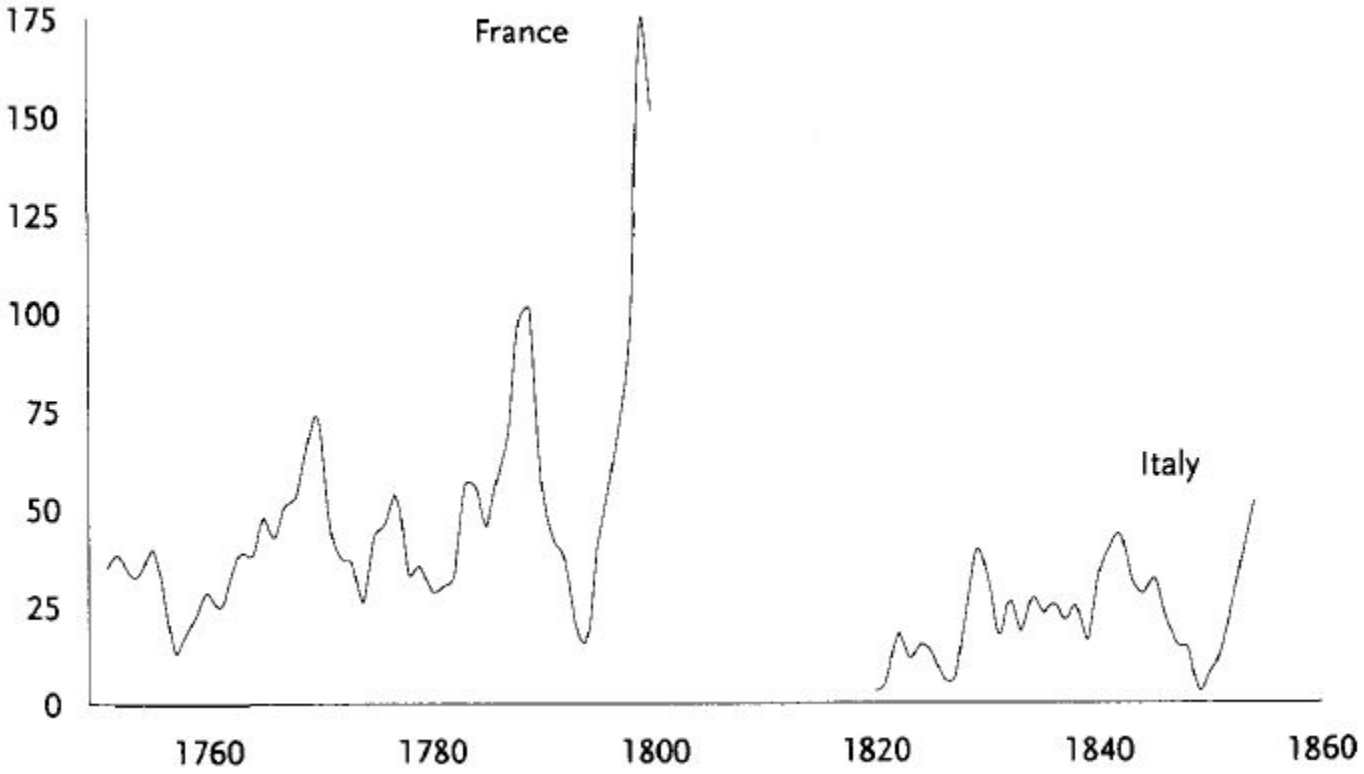
Graphs, Maps, Trees: Abstract Models for a Literary History, Franco Moretti (2007)

- Text undergoes a process of deliberate reduction and abstraction borrowing models from 3 disciplines:
    1. Graphs → quantitative history
    2. Maps → geography
    3. Trees → evolutionary theory

Graphs, maps, and trees place humanities disciplines literally in front of our eyes-and show us how little we still know about it.

Graphs

The rise and fall of the novel

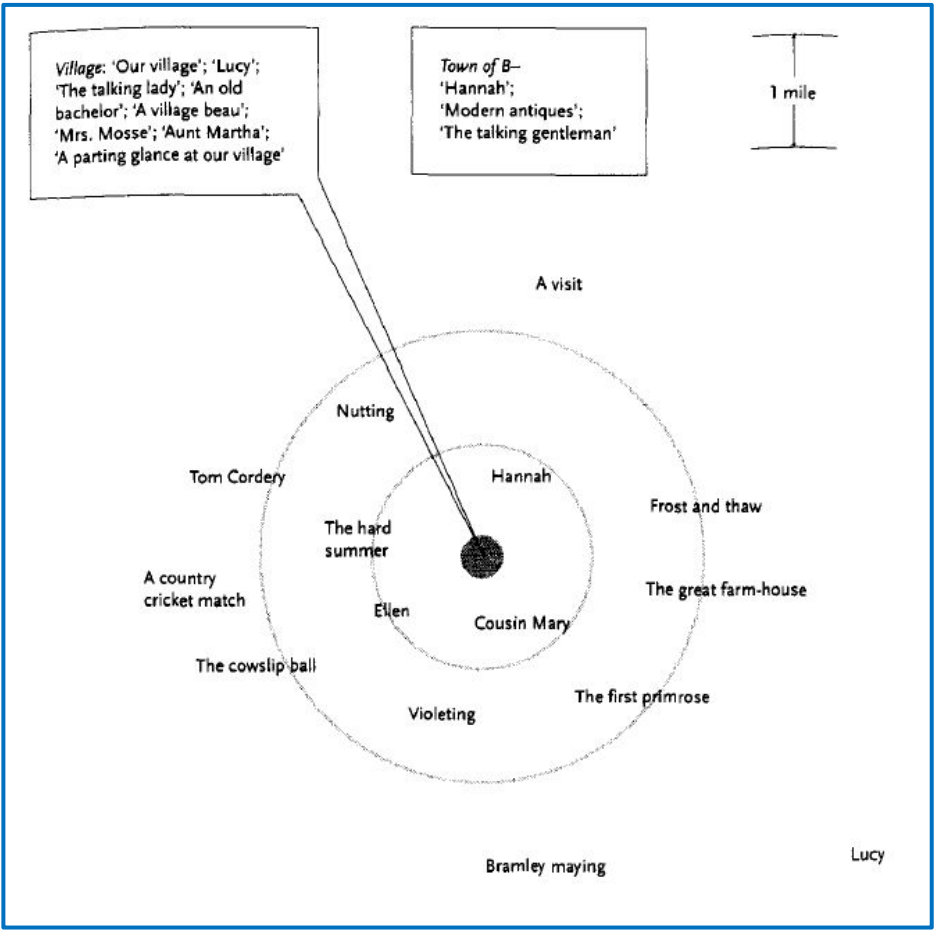


New novels per year (Moretti, 2007)

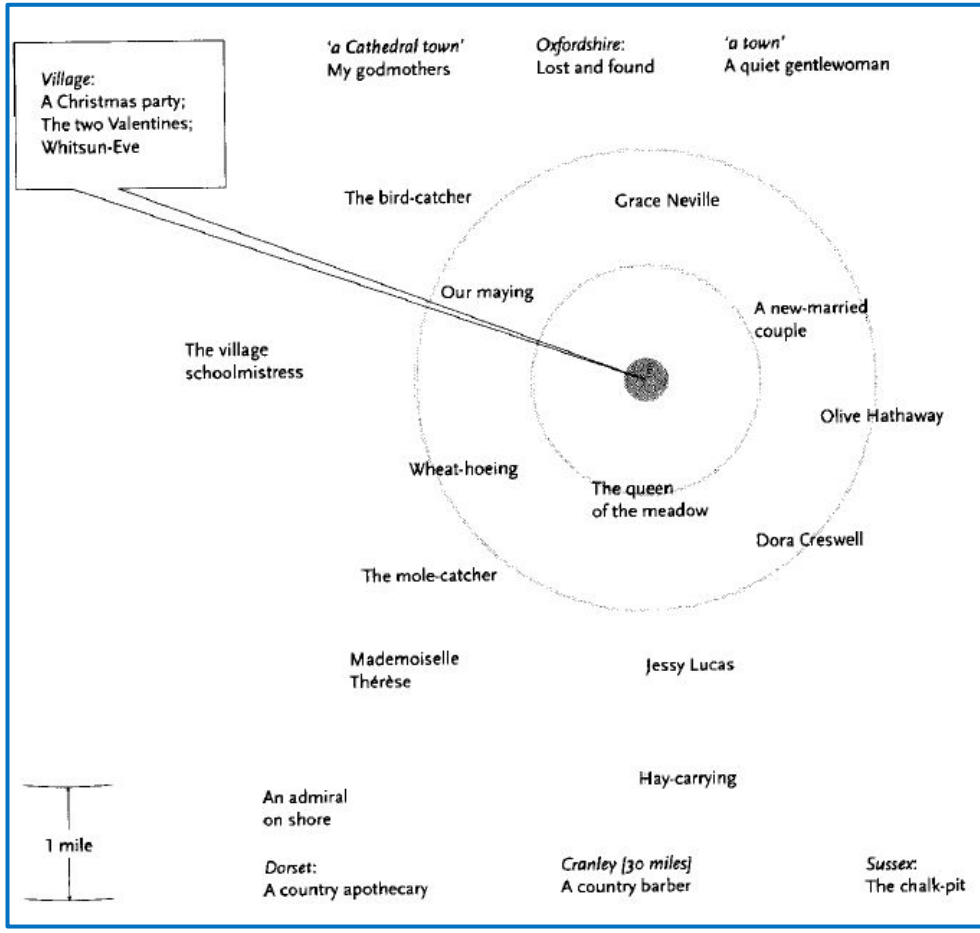
Maps

Mary Mitford, "Our Village"

1824



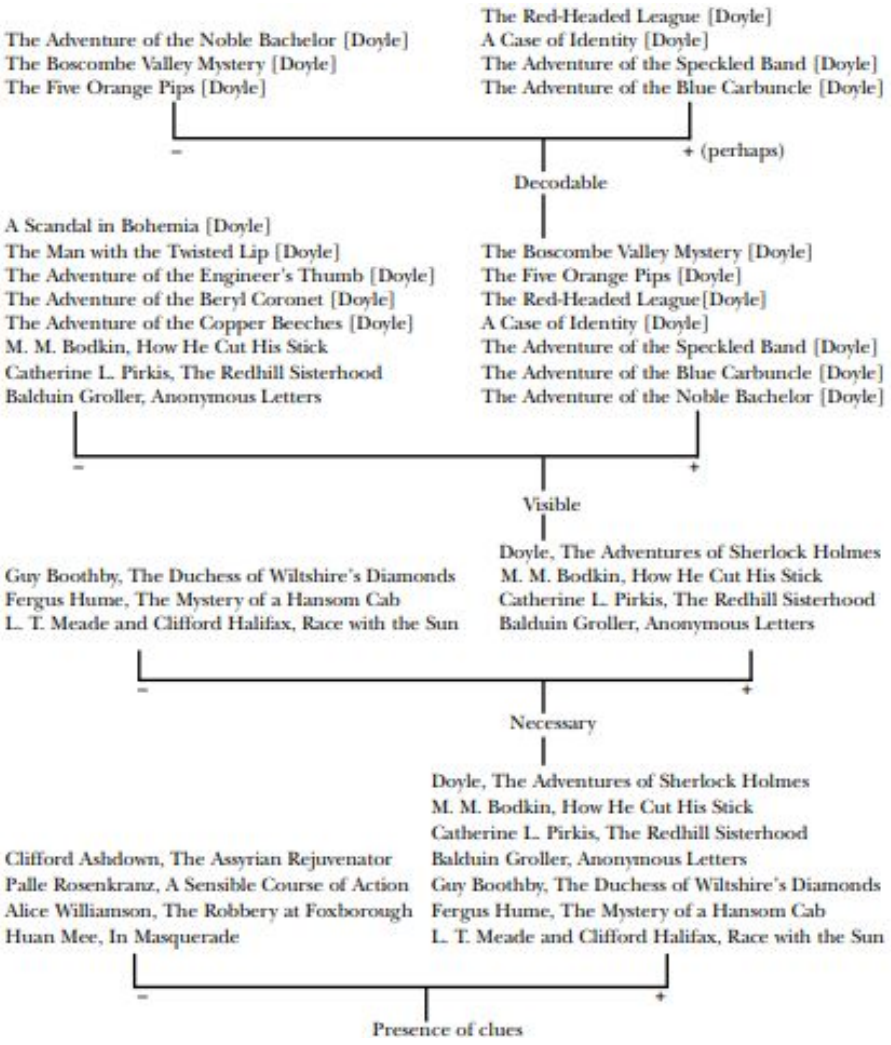
1828



(Moretti, 2005)

Trees

The success of Conan Doyle

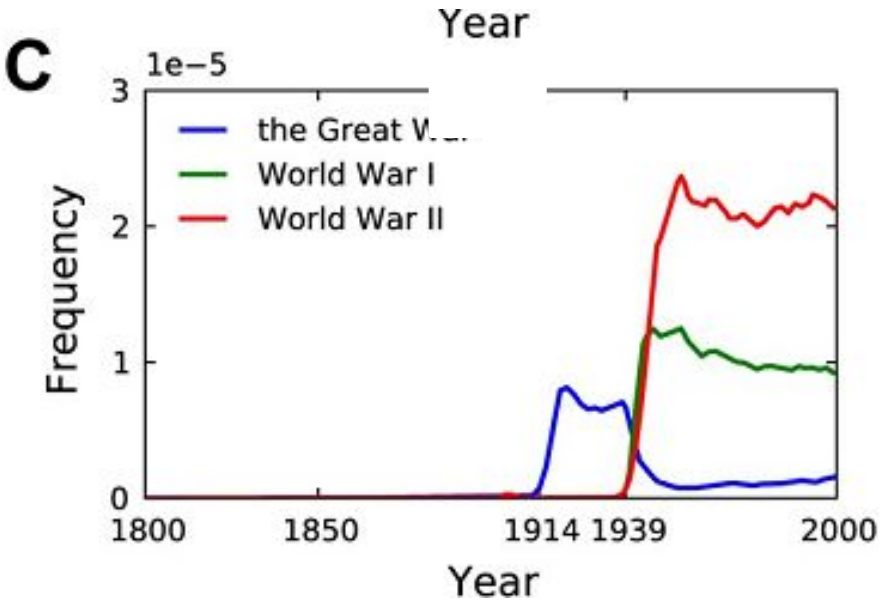


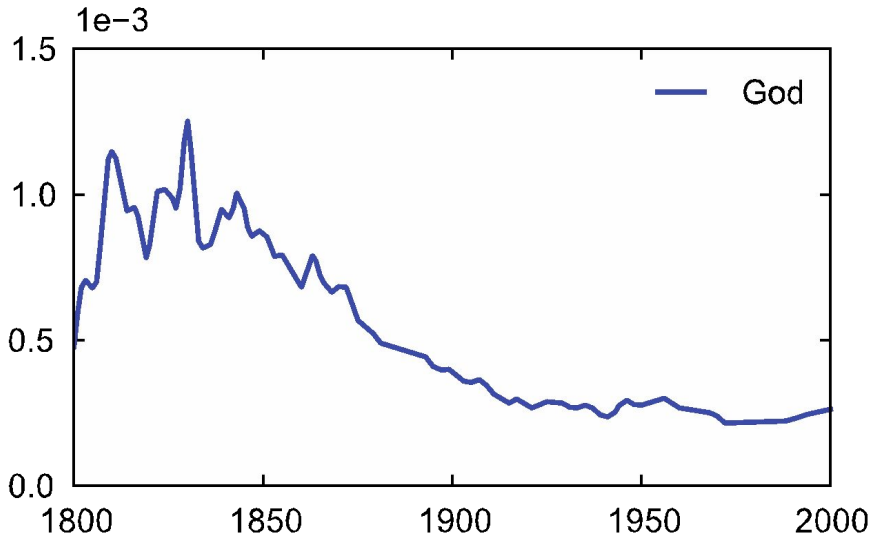
(Moretti, 2005)

Culturomics

Quantitative Analysis of Culture Using Millions of Digitized Books, Jean-Baptiste Michel et al., Science 331 (2010)

- *Culturomics*: application of data collection and analysis techniques to the study of human culture





"God is not dead but needs a new publicist"

https://books.google.com/ngrams

Well Before Moretti…

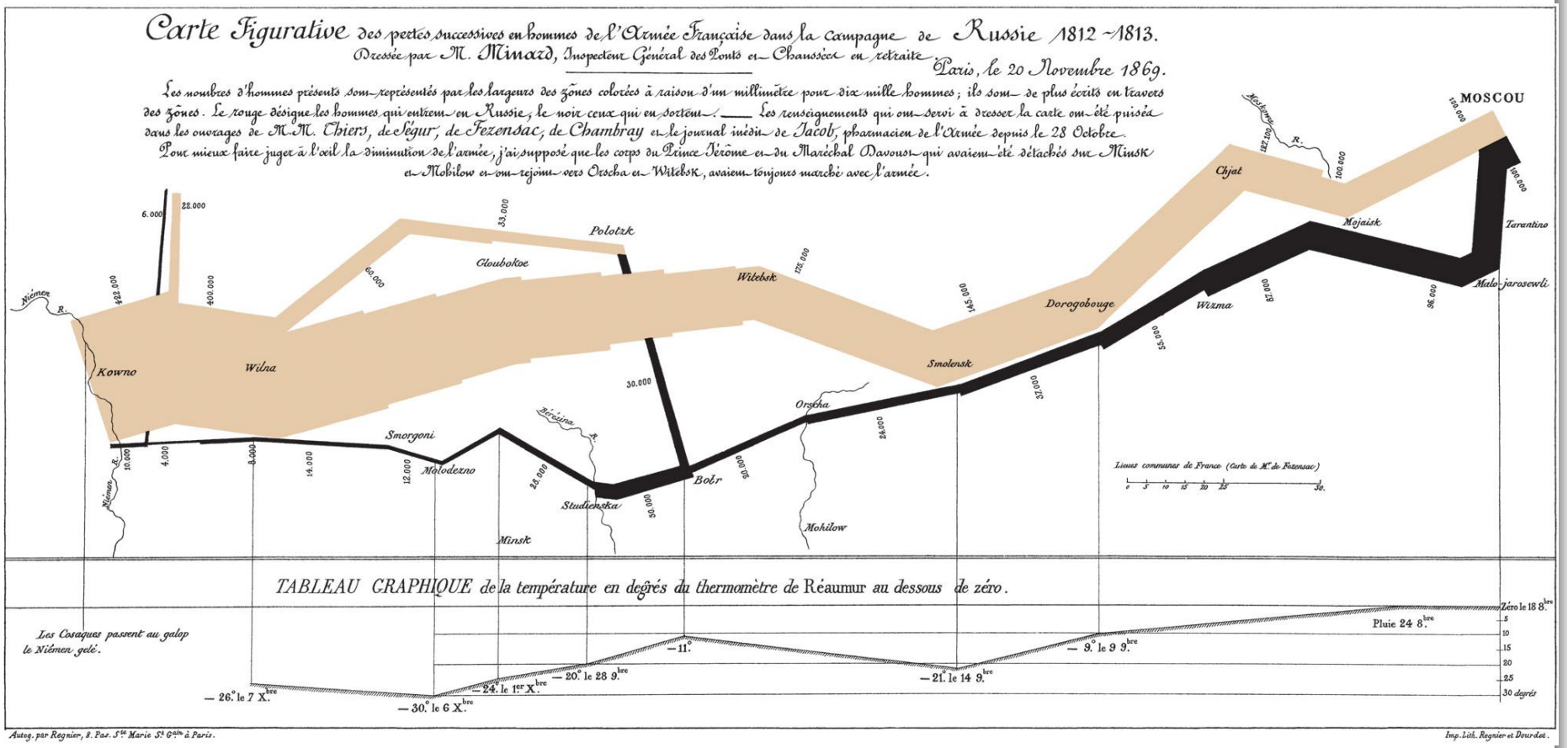
*"The great growth of **statistical research** in our times has made felt the need to record the results in forms **less dry**, **more useful**, and able to be explored more **rapidly** than numbers alone; thus, diverse representations have been imagined, among others my graphic tables and my figurative maps."*

Minard, "Graphic Tables and Figurative Maps" (1862)

translated by Edward Tufte

Visual History

Napoleon's Russian Campaign (1861)
"May well be the best statistical graphic ever drawn", Tufte



Distant Reading Procedure

- Collection of relevant data
  - format?
  - accessibility?
  - copyright?
- Data formatting and cleaning
- Data processing
- Result formatting
- Producing visualizations
- Interpretation
- Dissemination

Some Examples

- Linguistic analysis
- Topic modeling
- Sentiment analysis
- Stylometry
- Geographical referencing
- Networks of words, concepts, characters
- ...

Why is Distant Reading Important?

- It is reproducible, repeatable
- It surfaces hidden patterns not seeable otherwise
- It helps in dealing with the ever-increasing amount of digital sources

BUT DO NOT FORGET THAT…

Close reading (micro-analysis) remains fundamentally important!

SCALABLE READING

Hands-on Session

- TODAY
  - RAW
  - Voyant
  - Palladio
  - Gephi

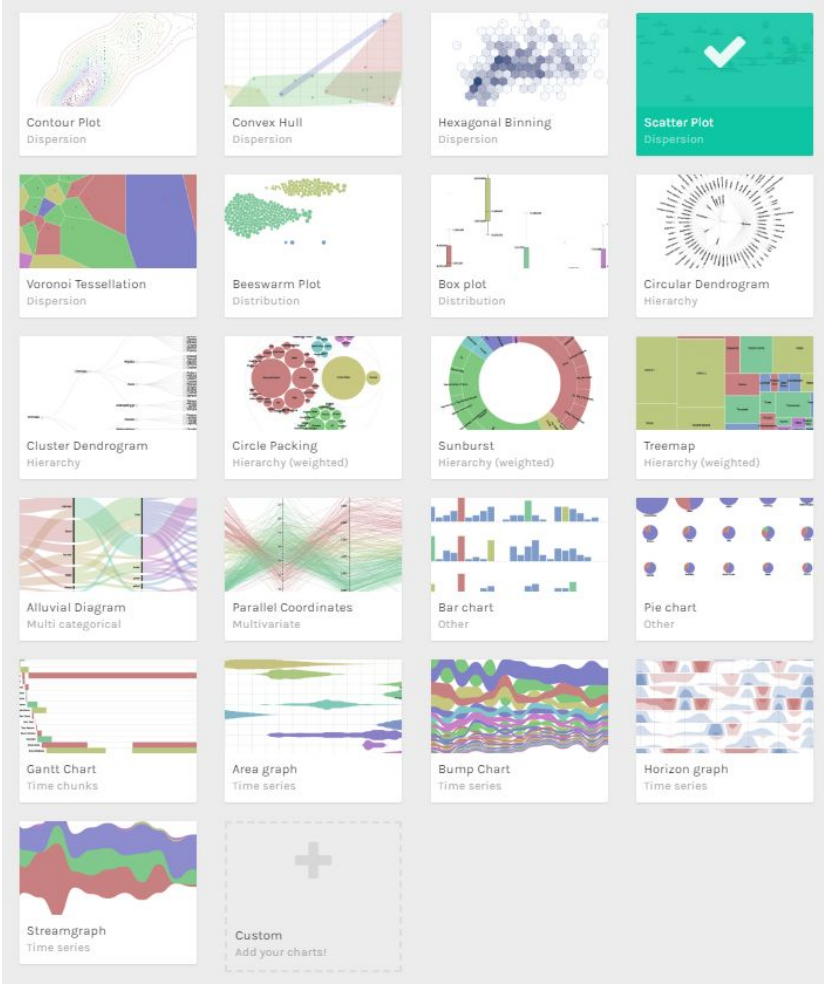
DOWNLOAD THE SHARED FOLDER:

https://drive.google.com/drive/folders/19IqrrHYIPDg6A8wA3US3Z0mJZU3hLmrY?usp=sharing

RAW

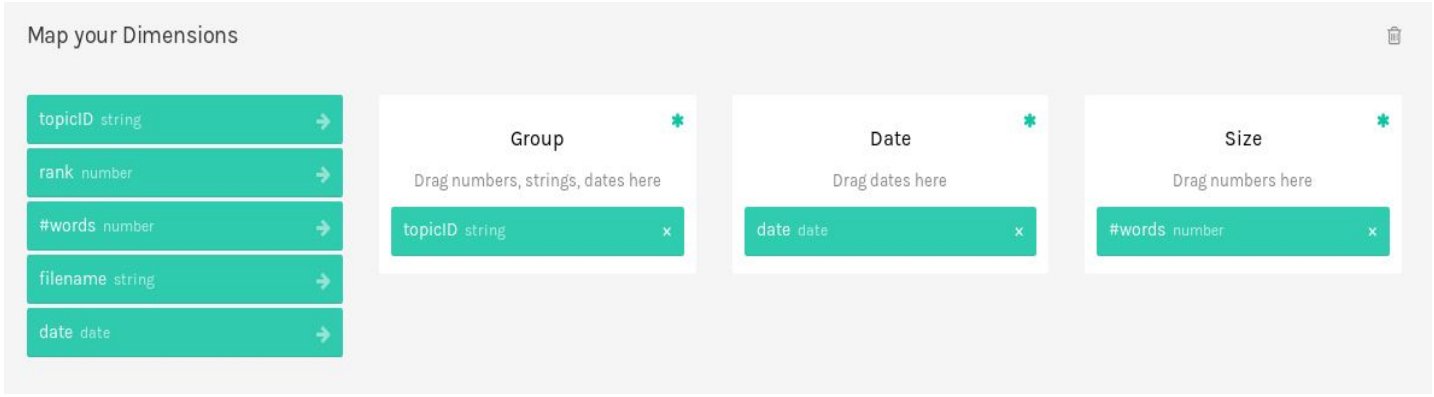
- RAW is “an open source data visualization framework built with the goal of making the visual representation of complex data easy for everyone”

  - from spreadsheets to (static) data viz
  - web-based
  - http://app.rawgraphs.io/



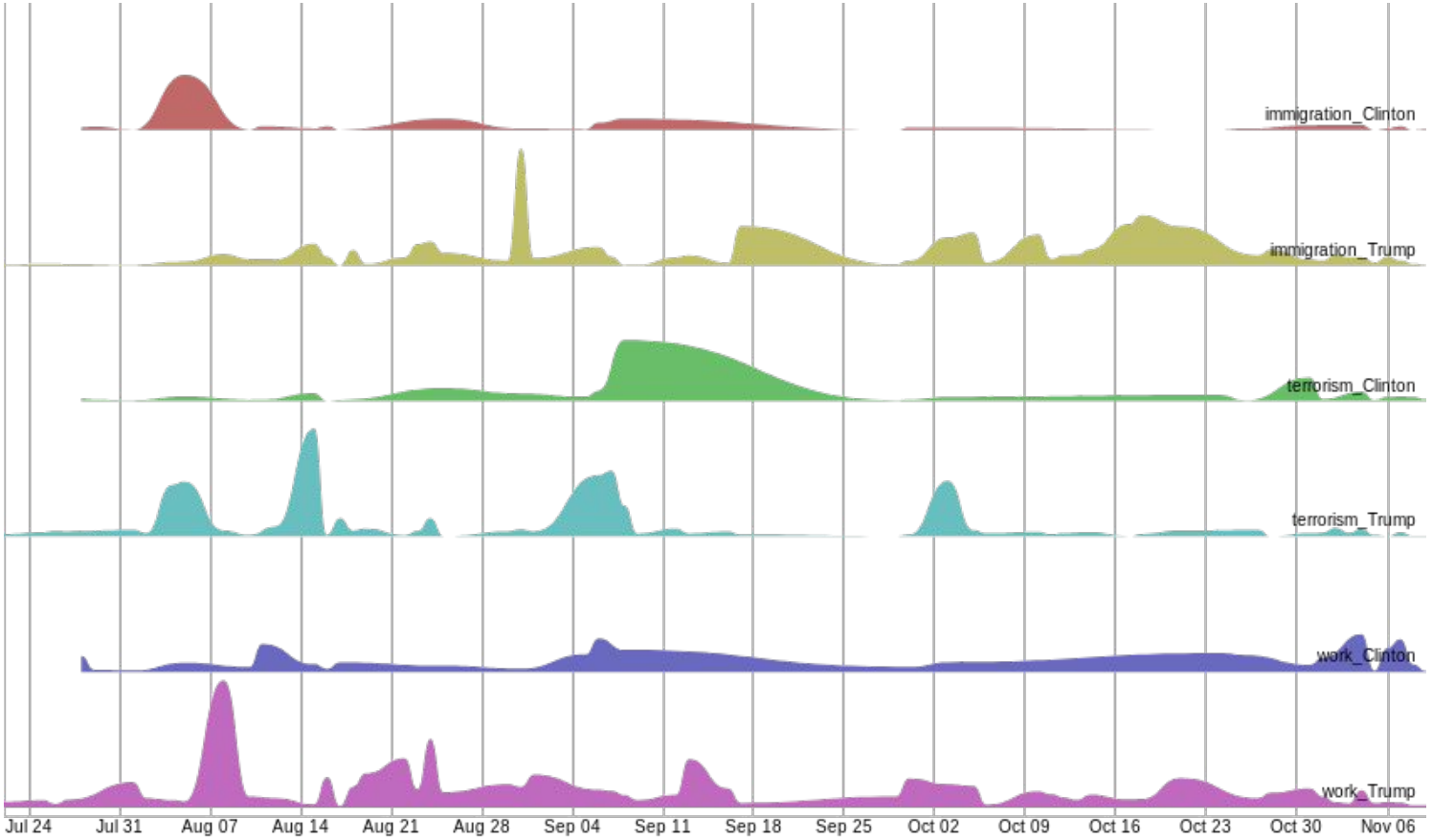
RAW and Topic Modeling

1. Go to RAW: http://app.rawgraphs.io/
2. Open TopicsInDocs-Trump_Clinton_new.ods
3. Copy the content of the spreadsheet
4. Select the area chart and map the dimensions as follows:



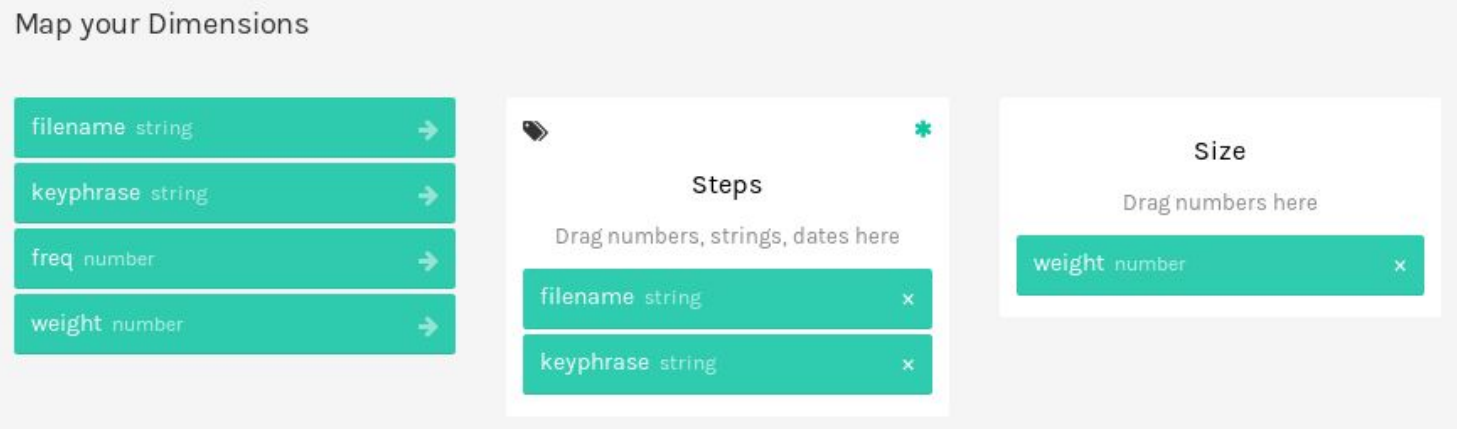
RAW and Topic Modeling

5. Output:



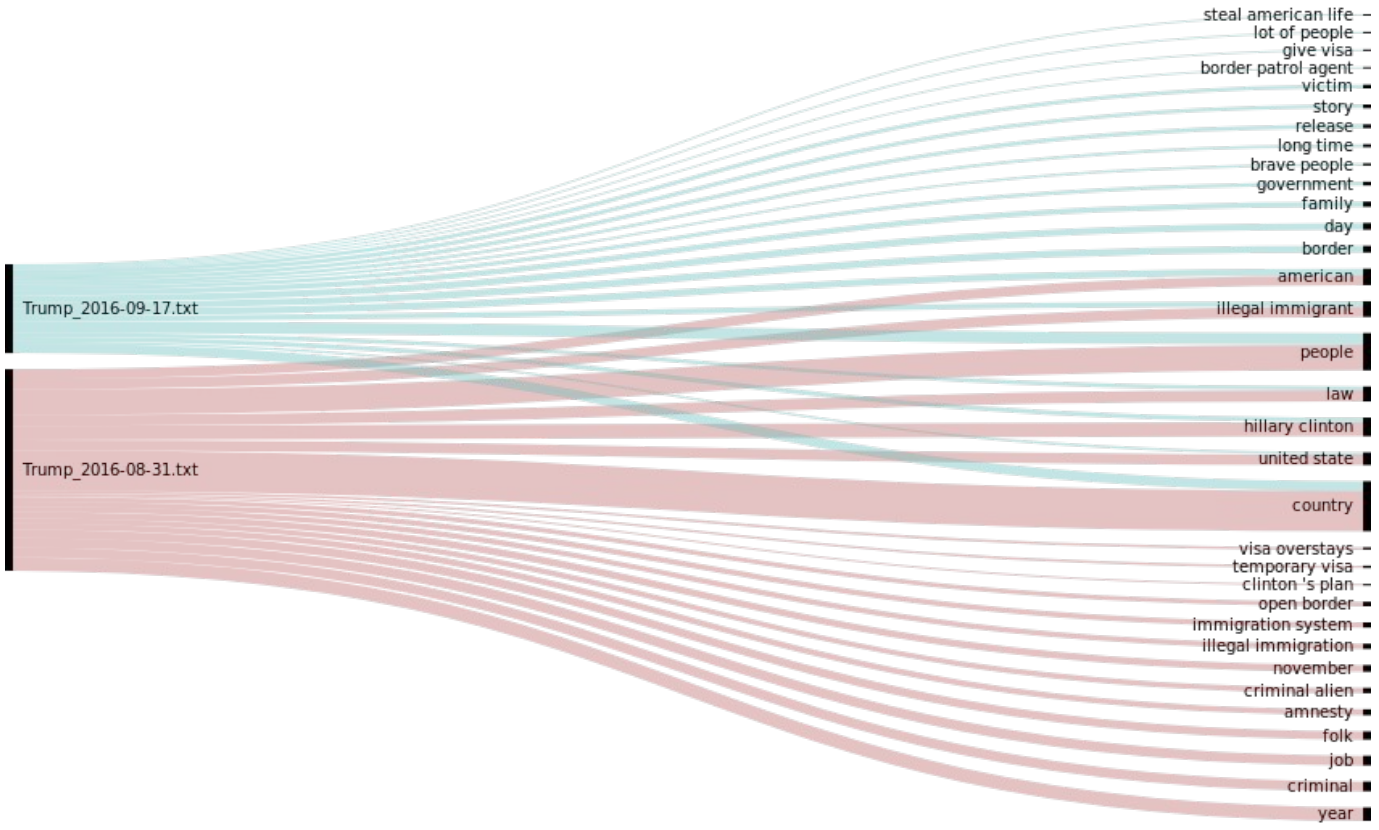
RAW and KD

1. Open KD_Trump.ods
2. Copy the content of the spreadsheet and paste it in RAW.
3. Select the Alluvial Diagram chart with the following dimensions:



RAW and KD

4. Output:



Voyant

- Voyant Tools is “a web-based text reading and analysis environment”
  - several text format as input
  - can be added to websites and online collections
  - interactive
  - scalable reading
  - web-based
  - https://voyant-tools.org/

Voyant

1. Go to http://voyant-tools.org/
2. Click on Carica and upload lincoln_sermons.zip
3. Try different tools: (examples below)
   - check the position of the word “God” in different sermons using Bubblelines
   - select a sermon in Sommario and check the trend of words “peace” in Andamenti
   - analyse the contexts of all the words starting with “slave*” in Contesti
   - modify the stopword list

Palladio

- Palladio is “a graphical interfaces based on humanistic inquiry”
  - examining data across time and space
  - from spreadsheet to maps and graphs
  - interactive
  - web-based
  - **http://hdlab.stanford.edu/palladio/**

Palladio Tutorial

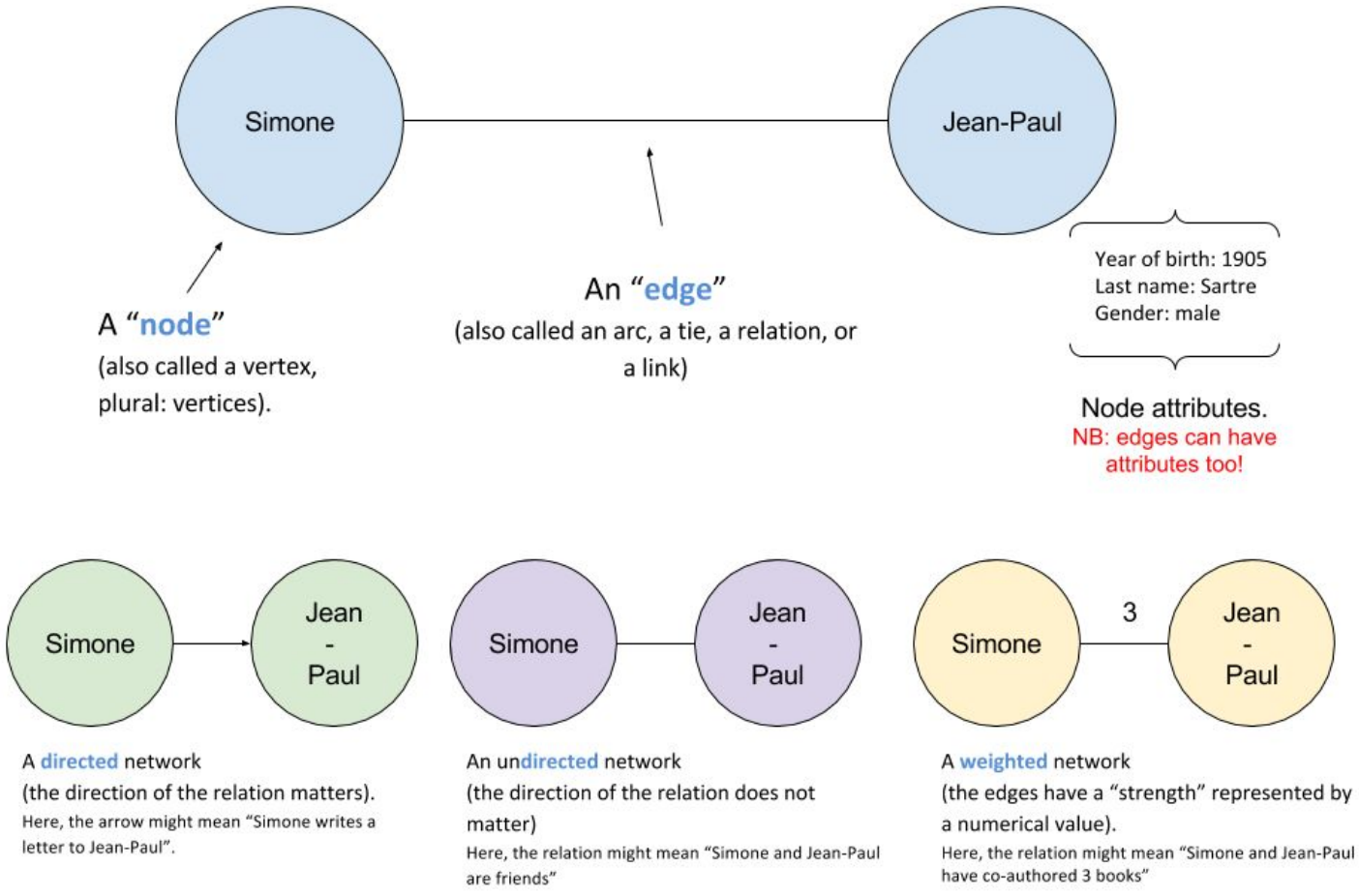
1. Go to http://hdlab.stanford.edu/palladio/ and click on Start
2. Copy the content of artist_data and paste in "Create a new project"
3. Click on Load
4. Check the dimensions highlighted in red by clicking on it
5. Correct the Data Type or Verify special characters if needed
6. Click on the name dimension and Add a new table
7. Copy the content of artwork_data and click Load
8. Check the dimensions highlighted in red in the added table
9. Click on Map at the top of the page
10. Click on New layer and choose Point to Point as Map type
11. Select coordinates_birth as source and coordinates_death as target
12. Click on Apply
13. Try the filters: Facet, Timeline, Timespan
14. Try Gallery

Gephi

- Gephi is “a visualization and exploration software for all kinds of graphs and networks”

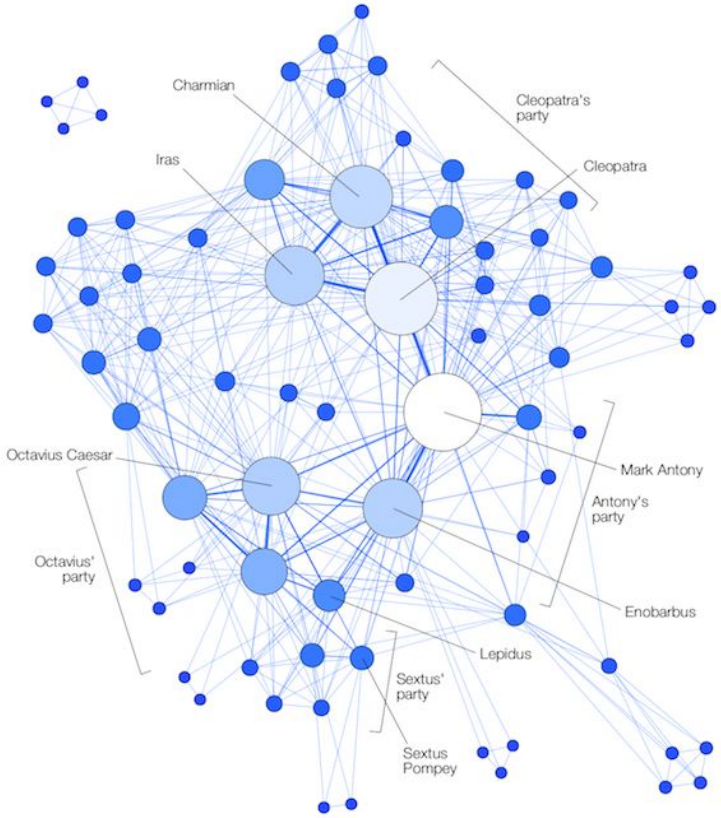
  - exploratory data analysis
  - social network analysis
  - integrated complex metrics
  - stand-alone
  - https://gephi.org/

Networks: Terminology

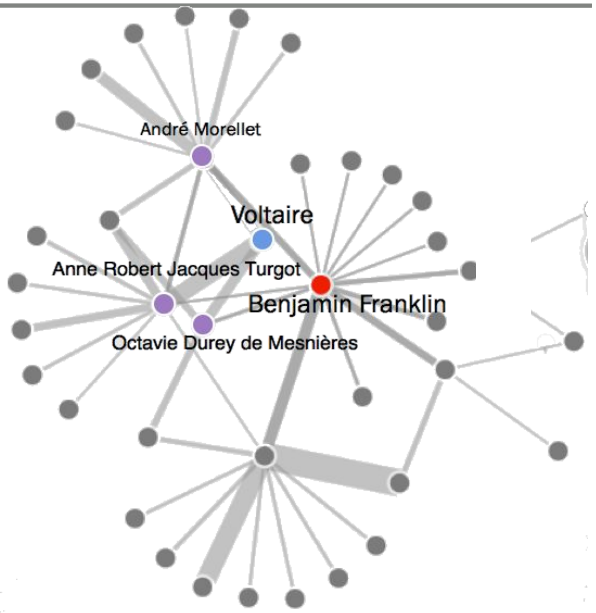


(Image taken from Clément Levallois, 2017)

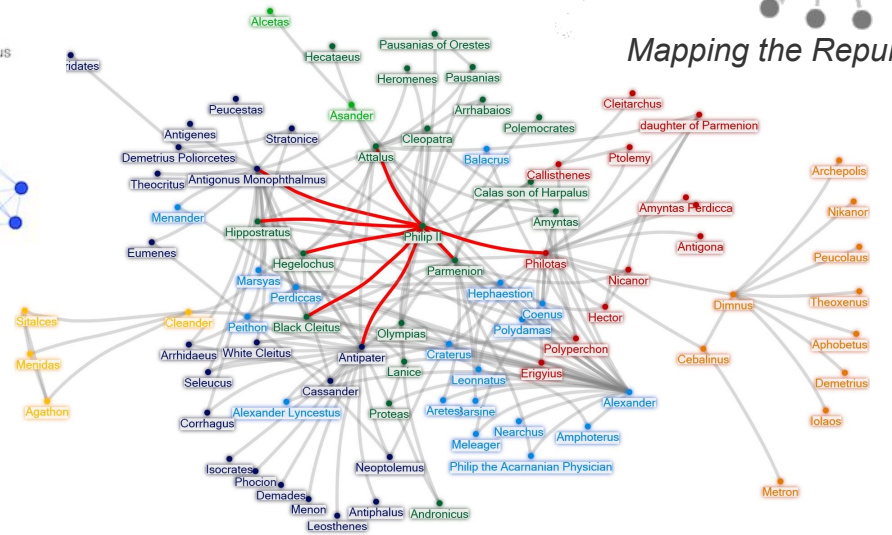
Networks in Digital Humanities



Martin Grandjean
"Mapping Shakespeare's tragedies" - 2015



Mapping the Republic of Letters



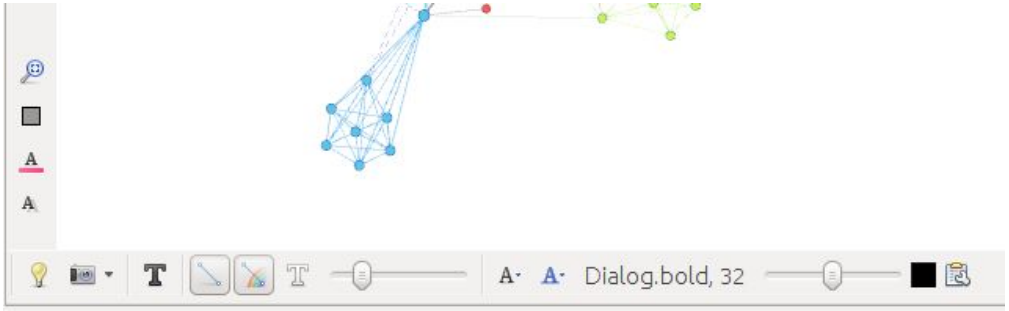
Diane Harris Cline
"Six Degrees of Alexander" - 2012

Gephi - 1

1. Open Gephi
2. Click on New Project
3. Click on Data Laboratory
4. Click on Import Spreadsheet
5. Click on the button with "..." and select the file miserables_result [Nodes]
6. Click on Successivo and then on Fine
7. Click again on Import Spreadsheet and select miserables_result [Edges]
8. Choose the option Edges Table under As table
9. Click on Successivo and then on Fine. Gephi may ask if you want to append the table to a new project or append it at the existing data table: choose this second option.
10. Click on Overview

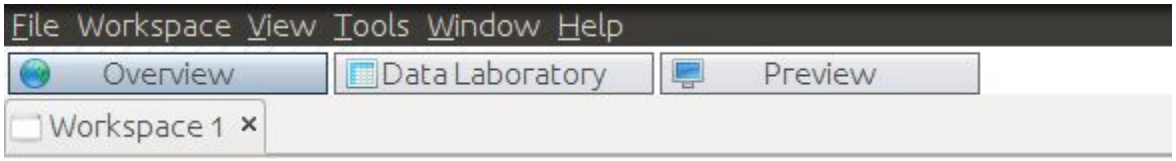
Gephi - 2

11. Among the layouts in the Layout panel: run *Fruchterman-Reingold* and *ForceAtlas2*
12. Using *ForceAtlas2*, modify the Scaling option (200.0) and select the Prevent Overlap option
13. In the Statistics panel, click Run for the Modularity metric
14. In the Appearance panel, select Partition and the attribute Modularity Class. Click on Apply
15. In the Statistics panel, click Run for the Average Degree metric
16. In the Appearance panel, select Edges, then Ranking and the attribute Degree. Click on Apply
17. Click on the T button under the graph to display node labels



Gephi - 3

18. In the Layout panel, select LabelAdjust, click on Run and then on Stop
19. Click Preview on the top of the Gephi window



20. Select the option Show Labels under Node Labels
21. Click on Refresh to update the preview
22. Export the graph in SVG/PDF/PNG

THANK YOU!

Email: sprugnoli@fbk.eu
Web Site: http://dh.fbk.eu
Twitter: https://twitter.com/DH_FBK

FONDAZIONE BRUNO KESSLER

DH DIGITAL HUMANITIES