A Social Network of The *Commentarii De Bello Gallico*

Elia Rizzetto, Digital Humanities and Digital Knowledge

Context

The *Commentarii De Bello Gallico* (BG) by Julius Caesar is a central work **in Latin Literature and Roman History and Historiography:** for centuries it has been used as a reference book for **teaching Latin** in schools and universities, due to its **simple and straightforward prose**; it represents a great **testimony** for key events in Caesar's era; it constitutes an important example of the **genre of military diaries** and an early witness of **political propaganda** in **pre-imperial Rome**. The work dates back to 51 BC, when the first seven books were released by Caesar as an evidence of his military capabilities, in order to obtain an extension of his proconsular command. The BG was defined as an "impressive and comprehensive edition of the campaign reports": it is the most trustable and detailed survived account of the series of Roman expeditions to Gaul started in 58 BC, but, at the same time, it is a firsthand work written by the general of the Roman armies, who took part to the battles he gives witness to, referring to himself in third person. For these reasons, the narrative constraints determined by the biassed nature of the report should be taken into account, assuming that the BG books depict the relations and the interactions between literary characters more than the social network of the real historical group whom the storytelling is based on. Coherently, the network analysis was based on the aforementioned premise, with a particular attention to the implications of studying a text whose protagonist is the literary representation of the author himself.

Problem and Motivation

Data for building social networks from literary texts can either be compiled manually by a human, who identifies the entities (nodes) and the relations amongst them (edges) from the text's semantics, or extracted automatically. The latter approach is typically based on the co-occurrence of two automatically recognized entities (e.g. the characters of a novel) in a defined text window: for example, two entities (nodes) are said to have a relation of some kind (edge) based on the fact that they occur together within the same window of 500 words, or in the same paragraph or chapter.

Despite its practicality, this approach falls short of precision, since two entities could often occur together in the same context without actually being linked by a semantic relation. To solve this issue, a possible option is that of reducing the span of the text window at the expense of performance in terms of recall, given that some semantic relations might be identifiable only in a wider context than the one chosen.

Since inferring relations between nodes relying on the sole co-occurrence is prone to false assumptions, this paper proposes a workflow based on an alternative approach to data extraction and management. More in detail, we aim at verifying the possibility to infer (more) precise semantic information from a text that has been manually annotated for morphology and syntax, i.e. if the morphological categories and the syntactic structure assigned to each sentence by expert annotators bear any processable information for:

- **Correctly identifying named entities** (people);
- **Correctly identifying existing relations among the people mentioned in the text,** assuming that the great majority of "real" (i.e. semantic) links between two named entities are represented at the syntactic level as well. For example, by identifying the syntactic relations in a sentence like "Caesar [subject] talks to [verb phrase] Vercingetorix [object]", we should be able to infer, by abstraction, a semantic relation to be represented in a social network graph.

The main goals of this project are the following:

- **Identifying and visually rendering the people mentioned in the Gallic War report by Caesar**, together with the relations that link them to one another, where each relation is represented as a mutual interaction. This is intended to provide a tool for an additional, though general and high-level, understanding of a key work in Latin literature and historiography.
- **Provide a possible workflow for the processing of structured data**, specifically morpho-syntactically annotated data of Latin texts, encoded in CoNLL-U format[1].

In addition to that, the application of the aforementioned methodology to this specific case of study is also aimed at drawing conclusions about the adequacy of the proposed methodology for analysing a text where the author is at the same time the narrator and one of the characters. In particular, one of the aspects which will be taken into consideration is the degree of polarisation introduced by the autobiographical aspect and whether or not the third-person narration neutralises, or at least mitigates, the impact of this condition on the network structure.

Datasets

The source data is a ConLL-U formatted .txt file containing morphological and syntactic annotations marked by expert linguists and Latin scholars. The file was downloaded from the repository of Latin data from the PROIEL treebank[2], which collects the texts curated by the Syntacticus project[3] and makes it publicly available on github. The file stored in this repository contain texts which were originally annotated in PROIEL-XML[4], and then were converted into CoNLL-U by the organisation itself in order to increase interoperability. Hence, the data used for the network analysis was provided in digital format and the contained information was already structured.

The distinctive characteristic of the ConLL-U format is that the data layout results from the morphological and syntactical manual annotation of each sentence, based on the dependencies model. Formally, the processed document exposes a series of tables, each corresponding to a sentence, where each row represents a specific token inside the sentence, and columns contain morphologic and syntactic annotations, all expressed in a standard and reasonably language-independent format.

[1] https://universaldependencies.org/format.html

[2] https://github.com/UniversalDependencies/UD_Latin-PROIEL

[3] http://dev.syntacticus.org/

[4] https://github.com/proiel/proiel-treebank

The syntax annotation is based on Dependency Grammar[5], therefore it is possible to represent each sentence as a tree structure[6], where the token (word, i.e. tree node) is marked as dependent from another token. The token on which is based the dependency of another token is named "dominant head", and it can be either another word node or the sentence root itself.

The document is processed using a minor Python library named conllu[7], which provides the possibility to read the CoNLL-U document as a list of Python dictionary-like objects. This allows the user to manipulate the data and set the content for the graphing phase, by building two lists of dictionaries, one for the nodes and the other one for the edges.

The further analysis of the pre-processed data is taken care of by means of the Python NetworkX[8] library, used also for the computation of measures. As a final step, visualisations are realised with Matplotlib[9].

Validity and Reliability

For the network analysis, we decided to use an approach which represents an alternative to the most commonly used methods for studying the relations in literary productions. In particular, the adopted approach is based on the assumption that grammatical and syntactical dependencies may reveal semantic information about the relations between the characters mentioned in the text, i.e. the named entities of the grammatical structures. Despite the positive side of granting that all relations present in the network are "actual", i.e. semantic, relations between the characters of text, this approach leaves out a great number of relations, and consequently nodes, that are not represented on the syntactic level. It should be bear in mind that the reality this model aims at representing is the one of the Bellum Gallicum as a text, and the entities and the relations among the entities are the ones expressed within it; it is not to be intended as a representation of the social network of historical people.

Measures

The graph

The network is represented as a monomodal, undirected, simple weighted graph. It consists of 62 nodes and 65 edges. From a quick visual analysis, we can see that it is a disconnect graph, with a central node corresponding to Caesar and a periphery of isolated dyads, most of which probably correspond to Roman consuls, i.e. mentioned as a way of determining time and not as actual agents in the text.

[5] https://taweb.aichi-u.ac.jp/tmgross/DG.html
[6] https://en.wikipedia.org/wiki/Treebank
[7] https://pypi.org/project/conllu/
[8] https://pypi.org/project/networkx/
[9] https://pypi.org/project/matplotlib/

Nodes: Centrality Measures

For understanding the role of the characters in the interaction network, we can observe the features at the node level. First, we get each node's **degree**, i.e. the number of edges connected to each of them, and **degree centrality**, i.e. its degree divided by the maximum possible number of connections it could have (which equals the total number of nodes in the network minus one). Despite its simplicity, **degree distribution** and **degree centrality** are very representative measures of a node's relevance in the network, and they are meaningful to our analysis in that they can show - together with **closeness centrality**, **betweenness centrality**, and **eigenvector centrality -** if the most important characters in the work are represented as such in the interaction network. **Betweenness centrality** measures the importance of a node based on the concept of shortest paths (or geodesic paths), which are the paths between any two nodes in a graph that have the minimum number of edges (for unweighted graphs) or the minimum sum of edge weights, interpreted as distances (for weighted graphs): for every pair of nodes in a *connected* graph, there exists at least one shortest path between the nodes. The betweenness centrality for a given node is defined as the number of these shortest paths that pass through that node. Factually, this means that the more shortest paths pass through a specific vertex, the higher is the betweenness centrality of that vertex. The algorithm used by the aforementioned NetworkX library assumes that the input graph is connected, therefore it only considers the shortest paths within the connected component containing the source and target nodes.

Closeness centrality, similarly, could be taken into account for gaining a deeper understanding of the most relevant characters in the narration. It measures a node's centrality within the network considering the node's average farness (inverse distance) to all other nodes: nodes with a high closeness score have the shortest distances to all other nodes. This approach, though, only works with connected graphs. To deal with a disconnected graph, we will have to resort to a variant of closeness centrality, **harmonic centrality**. Harmonic centrality is an alternative measure to closeness centrality, which provides the possibility to handle disconnected parts of the graph. Instead of summing the distances of a given node to all other nodes (as closeness centrality does), the harmonic centrality algorithm calculates the harmonic mean of the reciprocals of those distances, giving more weight to smaller distances and making it less affected by large distances.

In studying the key actors of the network, we can consider the nodes with the highest value for **eigenvector centrality**. This measure, besides taking into account the number of edges connected to a node, also considers the centrality of its neighbours, whereby nodes connected to (more) nodes with high centrality values get higher centrality values themselves. Therefore, the value of eigenvector centrality of a node is directly proportional to the sum of the eigenvector centrality of its neighbours.

Network: K-cores

Interesting observations on the structure and shape of the network can be made starting from the quantitative and qualitative analysis of sets of nodes. Here, k-cores and cliques are considered. K-cores are groups of nodes in which each node is connected to *at least k* other members of the group, while cliques are groups of nodes in which each node is connected to *all* the others. Both these kinds of groups can be extracted and analysed to investigate the presence of groups of actors in the network that interact mainly between each other, how many of these groups are there, and what nodes they include.

Applied Measures

With respect to the present study, it is relevant to understand whether or not Caesar's node is present in the clearly identifiable groups of nodes. On the one hand, coherently, we expect most of the actions (syntactic relations between people NE, represented as nodes linked by an edge in the graph) to concern the protagonist (therefore to include Caesar's node), since the input text is a propagandistic, self-celebrative work with a strongly oriented perspective. On the other hand, there is also a more "objective" dimension represented in the text, i.e. the one produced by Caesar as an author, or rather Caesar as a historian. The factual concretization of this dimension is investigated in the graph by studying the presence of relations that do not involve the Caesar-character.

In conclusion, further considerations can be made by calculating the network **clustering coefficient**.

The measures applied in the study are listed below:

- **Degree centrality**, based on the degree of a node, i.e. the number of edges that are connected to it (the higher the degree, the more central the node);
- **Closeness centrality**, which measures a node's average farness (inverse distance) to all other nodes (nodes with a high closeness score have the shortest distances to all other nodes);
- **Eigenvector centrality**, which computes the centrality of a node based on the centrality of its neighbours;
- **Hub centrality**, i.e. the property of a node of having a relatively (to the average) high number of edges.

By applying these measures, we want to see if the resulting graph let us see new information and/or reflects some important aspects of the social relations described in the *Bellum Gallicum*, therefore answering the following questions:

- *Are the most "important" actors represented as such in the graph?* Degree centrality and closeness centrality are used as a proxy for an individual's relevance. It is assumed that prominent individuals are involved in a higher number of interactions with other individuals.
- *Is it possible to infer any hypothesis on the role of the characters from the graph?* The value of the Eigenvector centrality of each node can be interpreted as a measure of how "important" is each individual in the military scene, since we would expect that prominent individuals (such as Caesar and other Roman generals) interact more frequently with other important individuals rather than with "minor" actors.

- *Is it possible to identify any "unexpectedly" relevant actor or community from the graph?* We analyse secondary patterns in the network, besides the expected major central node corresponding to Caesar, in order to discover whether or not unforeseen sets of nodes can be identified.

Results

As we would expect, the degree distribution reflects the importance of Caesar's node: the great majority of the nodes have a really low degree, being connected to one or two other nodes;10 nodes have a relatively small degree (2-5 edges) and finally only one node shows a peak in the number of connection: Caesar's, of course, for which the graph represents 20 connections. The analysis of the raw degree is already a first suggestion for the centralised nature of our network, reflected also in the histogram representing degree centrality.

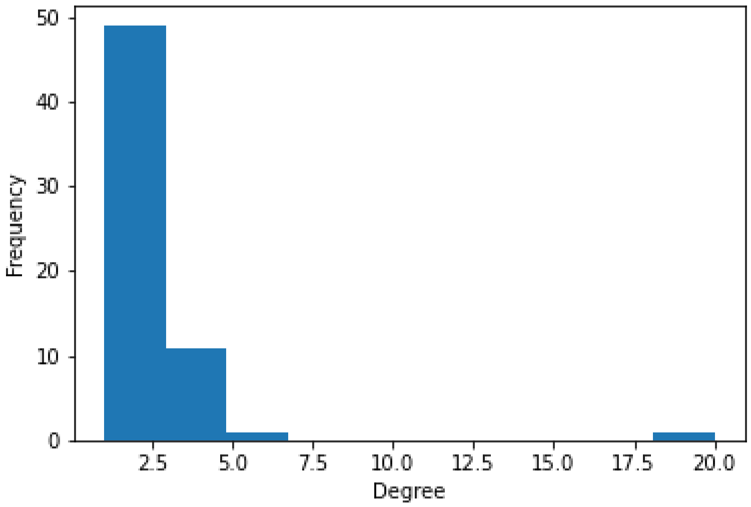


Figure 1.

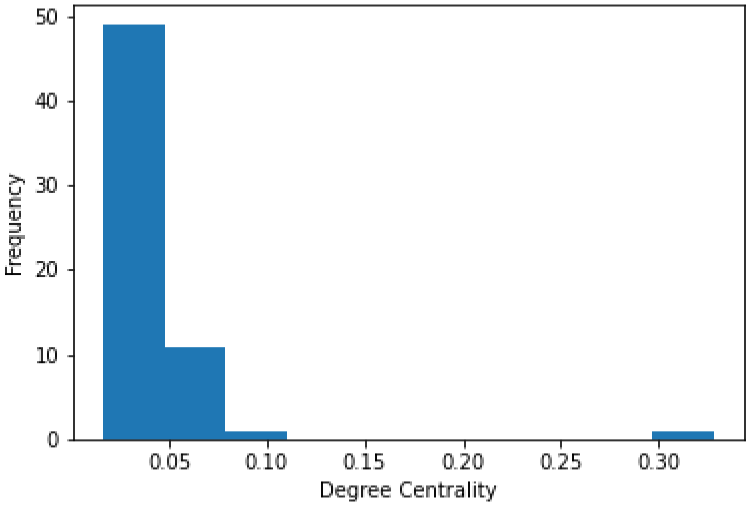
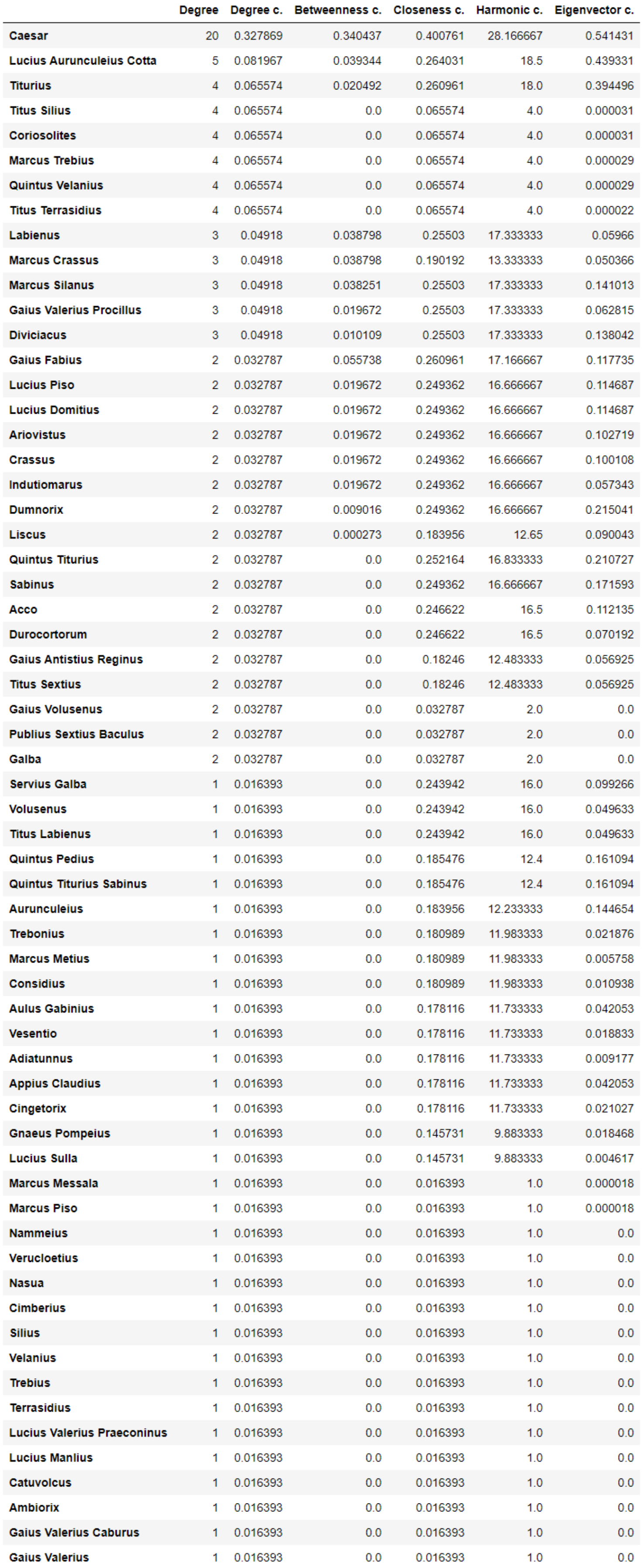


Figure 2.

In order to identify which nodes build up the short tail of "intermediate" degree, we can start looking at the table below, listing the value of each centrality measure for every node in the network. From here, we will try to identify *who* are the most important nodes in the graph, and see if they comply with the general "importance" of their corresponding character in the work.



	Degree	Degree c.	Betweenness c.	Closeness c.	Harmonic c.	Eigenvector c.
Caesar	20	0.327869	0.340437	0.400761	28.166667	0.541431
Lucius Aurunculeius Cotta	5	0.081967	0.039344	0.264031	18.5	0.439331
Titurius	4	0.065574	0.020492	0.260961	18.0	0.394496
Titus Silius	4	0.065574	0.0	0.065574	4.0	0.000031
Coriosolites	4	0.065574	0.0	0.065574	4.0	0.000031
Marcus Trebius	4	0.065574	0.0	0.065574	4.0	0.000029
Quintus Velanius	4	0.065574	0.0	0.065574	4.0	0.000029
Titus Terrasidius	4	0.065574	0.0	0.065574	4.0	0.000022
Labienus	3	0.04918	0.038798	0.25503	17.333333	0.05966
Marcus Crassus	3	0.04918	0.038798	0.190192	13.333333	0.050366
Marcus Silanus	3	0.04918	0.038251	0.25503	17.333333	0.141013
Gaius Valerius Procillus	3	0.04918	0.019672	0.25503	17.333333	0.062815
Diviciacus	3	0.04918	0.010109	0.25503	17.333333	0.138042
Gaius Fabius	2	0.032787	0.055738	0.260961	17.166667	0.117735
Lucius Piso	2	0.032787	0.019672	0.249362	16.666667	0.114687
Lucius Domitius	2	0.032787	0.019672	0.249362	16.666667	0.114687
Ariovistus	2	0.032787	0.019672	0.249362	16.666667	0.102719
Crassus	2	0.032787	0.019672	0.249362	16.666667	0.100108
Indutiomarus	2	0.032787	0.019672	0.249362	16.666667	0.057343
Dumnorix	2	0.032787	0.009016	0.249362	16.666667	0.215041
Liscus	2	0.032787	0.000273	0.183956	12.65	0.090043
Quintus Titurius	2	0.032787	0.0	0.252164	16.833333	0.210727
Sabinus	2	0.032787	0.0	0.249362	16.666667	0.171593
Acco	2	0.032787	0.0	0.246622	16.5	0.112135
Durocortorum	2	0.032787	0.0	0.246622	16.5	0.070192
Gaius Antistius Reginus	2	0.032787	0.0	0.18246	12.483333	0.056925
Titus Sextius	2	0.032787	0.0	0.18246	12.483333	0.056925
Gaius Volusenus	2	0.032787	0.0	0.032787	2.0	0.0
Publius Sextius Baculus	2	0.032787	0.0	0.032787	2.0	0.0
Galba	2	0.032787	0.0	0.032787	2.0	0.0
Servius Galba	1	0.016393	0.0	0.243942	16.0	0.099266
Volusenus	1	0.016393	0.0	0.243942	16.0	0.049633
Titus Labienus	1	0.016393	0.0	0.243942	16.0	0.049633
Quintus Pedius	1	0.016393	0.0	0.185476	12.4	0.161094
Quintus Titurius Sabinus	1	0.016393	0.0	0.185476	12.4	0.161094
Aurunculeius	1	0.016393	0.0	0.183956	12.233333	0.144654
Trebonius	1	0.016393	0.0	0.180989	11.983333	0.021876
Marcus Metius	1	0.016393	0.0	0.180989	11.983333	0.005758
Considius	1	0.016393	0.0	0.180989	11.983333	0.010938
Aulus Gabinius	1	0.016393	0.0	0.178116	11.733333	0.042053
Vesentio	1	0.016393	0.0	0.178116	11.733333	0.018833
Adiatunnus	1	0.016393	0.0	0.178116	11.733333	0.009177
Appius Claudius	1	0.016393	0.0	0.178116	11.733333	0.042053
Cingetorix	1	0.016393	0.0	0.178116	11.733333	0.021027
Gnaeus Pompeius	1	0.016393	0.0	0.145731	9.883333	0.018468
Lucius Sulla	1	0.016393	0.0	0.145731	9.883333	0.004617
Marcus Messala	1	0.016393	0.0	0.016393	1.0	0.000018
Marcus Piso	1	0.016393	0.0	0.016393	1.0	0.000018
Nammeius	1	0.016393	0.0	0.016393	1.0	0.0
Verucloetius	1	0.016393	0.0	0.016393	1.0	0.0
Nasua	1	0.016393	0.0	0.016393	1.0	0.0
Cimberius	1	0.016393	0.0	0.016393	1.0	0.0
Silius	1	0.016393	0.0	0.016393	1.0	0.0
Velanius	1	0.016393	0.0	0.016393	1.0	0.0
Trebius	1	0.016393	0.0	0.016393	1.0	0.0
Terrasidius	1	0.016393	0.0	0.016393	1.0	0.0
Lucius Valerius Praeconinus	1	0.016393	0.0	0.016393	1.0	0.0
Lucius Manlius	1	0.016393	0.0	0.016393	1.0	0.0
Catuvolcus	1	0.016393	0.0	0.016393	1.0	0.0
Ambiorix	1	0.016393	0.0	0.016393	1.0	0.0
Gaius Valerius Caburus	1	0.016393	0.0	0.016393	1.0	0.0
Gaius Valerius	1	0.016393	0.0	0.016393	1.0	0.0

Table 1.

For all the measures, the top ranking node is naturally Caesar. The ones with immediately lower values, interestingly enough, are all Romans, besides a group of nodes (Titus Silius, Coriosolites, Marcus Trebius, Quintus Velanius, Titus Terrasidius) which will be analyzed later. We can suppose that this is also a symptom of Caesar's perspective on the narration, since interactions involving Gauls seem to be underrepresented with respect to the ones involving Romans: a natural phenomenon, if we consider that the number of nodes for Gauls is lower than the number of "Roman" nodes, and at the same time assume that the interactions described in the work are mostly of the kind Roman-Roman and Gaul-Gaul.

From the observation of the results for betweenness centrality, we can further see the centralization of the network. Caesar has the highest value of 0.34; the Romans Lucius Auruculeius Cotta, Labienus, Marcus Crassus, Marcus Silanus and Titurius are the most central characters besides Caesar, but have much lower values (<0.03). The great majority of the entities have betweenness centrality equal to 0, which suggest that they are either peripheral nodes, with respect to the central node of Caesar, or members of groups disconnected from the main one.

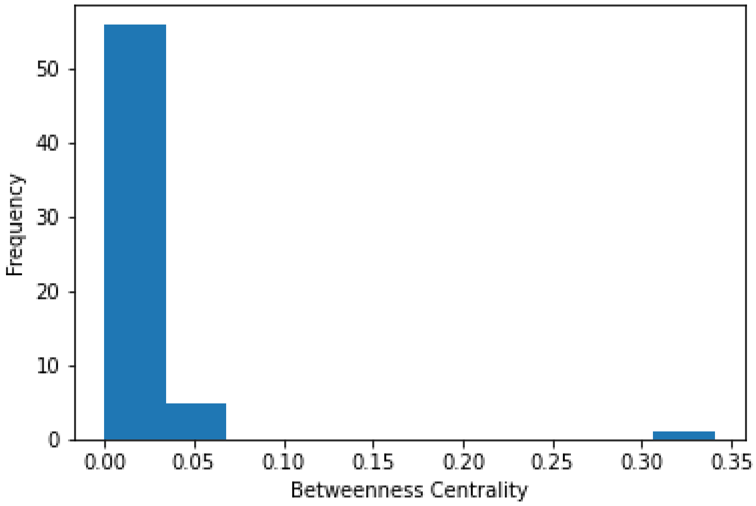


Figure 3.

A comparison of the results from the calculation of closeness centrality and harmonic centrality shows the inaptness of the former for a disconnected network, as sensibly different results can be seen in the central part of the two histograms. No significant results are deducible from the analysis of the harmonic centrality, despite the fact that, again, the great majority of the nodes seem to be peripheral with respect to Caesars' node.

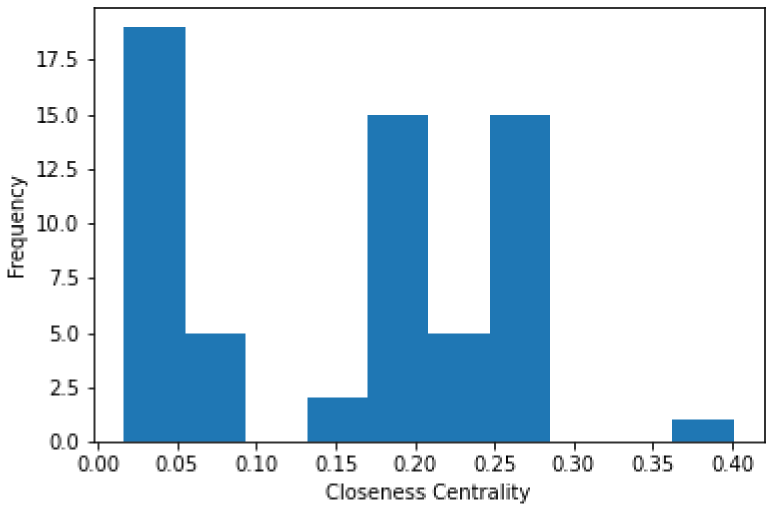


Figure 4.

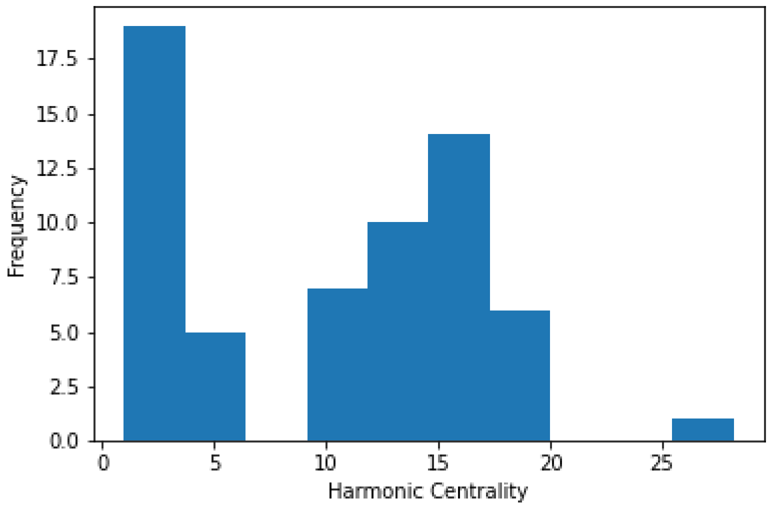


Figure 5.

Eigenvector centrality, on the other hand, shows clearly both the widespread presence of peripheral nodes, and the role of “intermediate” ones in our network: the centrality of Lucius Aurunculeius Cotta and Titurius probably “benefit” from a connection with other Roman characters.

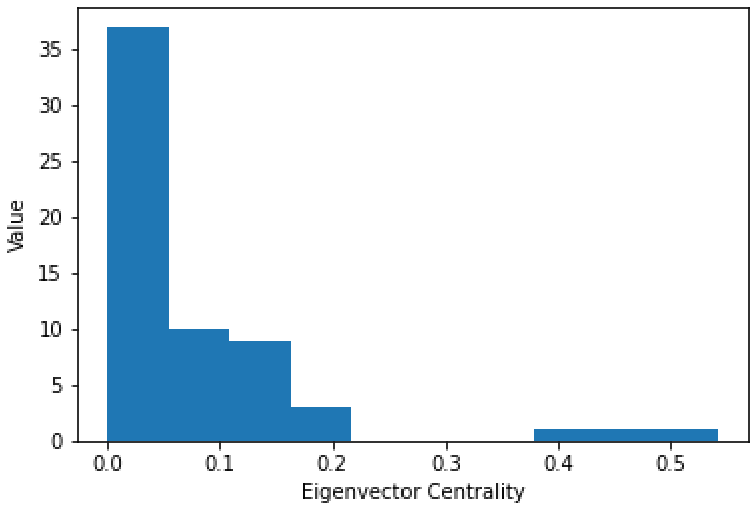


Figure 6.

In order to identify relevant sets of nodes, k_cores are looked for in the graph. We saw that the great majority of nodes have core number 1 (the core number of a node is the largest value k of a k-core containing that node), therefore we expect the subgraph for k=2 to be substantially different from the original one. For k=2, Caesar's node is still in the (sub)graph, while for k>=3, only one core remains. This decomposition of the graph in cores could be used to identify the core-periphery structure of the network, where the nodes "remaining" in the graph at the highest value of k would be the central "core" of the network, and the ones removed the periphery. The disconnectedness of the graph, though, makes this interpretation deceiving: the remaining core is definitely less relevant for the whole network than the bigger cores we find at k=2.

Conclusion

In conclusion, it is possible to summarise the qualitative and quantitative outcomes as follows.

- As expected, **Caesar’s node results to be prominent** compared to all the other nodes. This aspect demonstrates that the third-person narration aimed at presenting the autobiographical storytelling as an objective report is not enough to balance the polarisation of Caesar as the central node. Nevertheless, the network shows the presence of some minor set of nodes where the central node is not involved.
- **The approach based on grammatical and syntactical dependencies showed some limitations** with respect to the analysis of the human network in exam. In particular, the results highlighted that not all the relations between the entities are represented as syntactic and morphological information derivable from the

sentence structure. Further, another part of the semantic information expressed in the grammatical structures was not represented with structures properly involving the named entities.

Critique

A critical step of building this network was in the data collection phase, for the following reasons:

- Most entities are mentioned across the text both with their “full name” and with their “first” name only. Since some characters share their first name, in some cases it is impossible to disambiguate these nodes, therefore we end up producing two different nodes (full name and first name) for what is actually a single entity.
- In the phase of building the graph, an edge was added to the graph for each syntactic relation between two tokens corresponding to a person, and weighted according to the “distance” from an entity to the other within the syntactic tree, in terms of tree nodes to traverse to get from one entity to the other. This method, despite allowing us to be sure that all the interactions represented in the resulting graph are semantic relations in the text, forces us to leave out a high number of semantic relations that are not expressed on the syntactic level (for example, in case a named entity is referred to with a pronoun).

Bibliography

Brandes, Ulrik. «A Faster Algorithm for Betweenness Centrality*». *The Journal of Mathematical Sociology* 25, fasc. 2 (giugno 2001): 163–77. https://doi.org/10.1080/0022250X.2001.9990249.

Cesare, Gaio Giulio, e Sossio Giametta. *De bello gallico. Testo latino a fronte*. 8° edizione. Bompiani, 1984.

Disney, Andrew. «Social Network Analysis: Understanding Centrality Measures». Cambridge Intelligence, 2 gennaio 2020. https://cambridge-intelligence.com/keylines-faqs-social-network-analysis/.

Everett, Martin, e Stephen P. Borgatti. «Ego Network Betweenness». *Social Networks* 27, fasc. 1 (gennaio 2005): 31–38. https://doi.org/10.1016/j.socnet.2004.11.007.

Han, Aaron Li-Feng, Derek F. Wong, Lidia S. Chao, Yi Lu, Liangye He, e Liang Tian. «A Universal Phrase Tagset for Multilingual Treebanks». In *Chinese*

Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, a cura di Maosong Sun, Yang Liu, e Jun Zhao, 8801:247–58. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-12277-9_22.

Osborne, Timothy. *A Dependency Grammar of English: An Introduction and Beyond*. Amsterdam: John Benjamins Publishing Company, 2019. https://doi.org/10.1075/z.224.

Welch, Kathryn. *Julius Caesar as Artful Reporter: The War Commentaries as Political Instruments*. ISD LLC, 2009.