Alma Mater Studiorum · Università di Bologna

FACOLTÀ DI SCIENZE MATEMATICHE, FISICHE E NATURALI
Corso di Laurea Magistrale in Matematica

Relations among Shakespeare's characters: an analysis in terms of centrality measures and new techniques from graph theory

Tesi di Laurea in Teoria dei Grafi

Relatore:
Chiar.mo Prof.
Marco Lenci

Presentata da:
Roberta Perissinotti Bisoni

Correlatore:
Chiar.mo Prof.
Rocco Coronato

I Sessione
Anno Accademico 2016/2017

Contents

Introduction

The representation and the mathematical analysis of connections between individual parts or components of a system (of any kind) have been very useful in enormous amount of contexts. Whether we are talking about small communities of individuals, network of electronic devices, system of proteins, abstract theoretical structures, the study through graph theory of the nature of system of connections system has been deeply developed over the recent years ([4], [5]). For these needs, graph theory has been enriched with many abstract structures, models, algorithms to satisfies a lot of different requests aimed at understanding, analyzing and even improving the functioning of a general network. We can see a system of units that interact with each other as a graph, i.e., a mathematical model consisting of a set of nodes (the units) and a set of links between the nodes (the relations): a node i is connected with a node j if i and j are related, in some pre-fixed sense, in the system. This leads us to see a graph as a binary matrix $A = (a_{ij})$ where a_{ij} is 1 if i and j are connected, 0 otherwise ([4]). We can do more and put weights on the edges, so the matrix is no longer binary and gives more information on the interactions.

The first graph theory problem dates back to 1736, with Leonhard Euler's work on the "Seven bridges of Königsberg": the problem was to exhibit a walk through the city of Königsberg that would cross each of the seven bridges of the town once and only once ([15]). Euler proved that this walk does not exist. Nowadays we would approach this problem talking about nodes, vertices, degree, paths, distances from a node to another, connectivity and so on, and we would have all the tools to rigorously prove the non existence of this path. Many other problems were discussed after that and now we are able to study and manage even huge networks such as the World Wide Web, or the social network that "live" within.

The existence of a path between nodes that satisfies a particular property, such as that considered by Euler, is only one of the classical problems in graph theory, and it is applied in many situations: think of the need to improve transport efficiency on the internet, e.g. understanding if a shortest path

is the fastest path, or the need to organize the websites in the World Wide Web, hence the analysis of the structure of the connections is crucial when it comes to optimizing network efficiency ([4], [5], [13]). As for smaller networks, one of the most popular example is the "Zachary's karate club", that is the graph describing the relationships between 34 members of a university-based karate club monitored from 1970 to 1972 by Wayne W. Zachary ([14]). The interesting part was that at one particular moment there was a conflict between two members A and B and the connections in the entire club deeply changed, leading to fission of the social circle. Through a graph theory algorithm, Zachary predicted the two final communities (except one person).

These are two examples that give us a little overview of the importance and adaptability of graph theory, but this theory can be used for a big variety of purposes: connectivity, shortest path problem, visualization, partitioning, etc. In particular, in this thesis we will focusing on two of them: centrality measures and clustering ([8], [9], [10], [11], [12], [4]).

In many situations it is useful to find the most important nodes of a graph by assigning to each of them a numerical value which is supposed to represent or estimate a feature of that node: this is what is called a centrality measure. Clearly, we cannot define a general ranking of 'importance' in a graph because any ranking clearly depends on the particular nature of the network and especially on what we mean by 'important'. For example, a vertex i of a graph G can be important because it has local influence on the information flows on the graph (like in certain electrical circuits) or a global one (e.g., the rumors spreading in a community) ([4]). So the first step in approaching this problem is to understand what it means to be important in that network based on the task or purpose at hand. After figuring out how we want to interpret the network, we can use an appropriate centrality measure to rank the nodes. This has often proved very useful especially in complex and big networks.

The first chapters of this thesis will be dedicated to the theory behind these concepts, which involves linear algebra, probability theory and stochastic processes. We will exhibit some of the best-known centrality measures, including the world-famous Google PageRank, i.e., the algorithm that Google uses to order search engine results ([2], [13], [3]).

Another important problem in graph theory is the division of a graph in sub-graphs, depending on certain requirements. This is another challenge, in general especially in big graphs (for example in social network analysis for which an overview of the connections in the graph cannot be done visually, so a mathematical approach is mandatory). But it is a challenge even for smaller networks (see Zachary's karate club as a simple example).

This topic of research is often divided in two: graph partitioning and

community detection ([4]). The main difference is that, in the first case, one needs to know a priori the sizes of the searched communities, the second case, one does not. There are plenty of algorithms and techniques that have been developed and used to find suitable sub-graphs in a graph and we will exhibit some of them. We will also propose a new method of graph partitioning that is based on Voronoi cells associated to central nodes. The method defines Voronoi cells for a predetermined group of important nodes, and then determines whether to merge a certain cells depending on the average distance of its node to the central nodes of the other cells. In this thesis we apply all these techniques to drama. In particular, we studied five plays by William Shakespeare: *Macbeth, Romeo and Juliet, Richard III, Twelfth Night* and *The Winter's Tale*. The idea of using graphs to visualize and study plays is not new and it is developing in recent times, especially around William Shakespeare's literary work. We cite as an example the work of Martin Grandjean (an overview is available online) or the Shakespearean Network in Wolfram Demonstrations Project. Grandjean compares the graphs of the characters relations in Shakespeare's tragedies. The Shakespearean Network is an online interactive project that allows one to visualize the relations among characters of ten plays and rank the most important nodes based on PageRank centrality.

The following is a summary of the content of this dissertation. We investigate two types of graphs that represent two different interpretations of the interactions among characters: they are based on the data set of how many times i talks to j and how many times i talks about j . The natural approach is to define a graph where the vertices are the characters and the edges are the relations among them, according to what kind of analysis we want to make. Consequently we studied the roles and functions of the characters using centrality measures in the two graphs. We also propose a new point of view in "reading" relations between characters through graphs, defining a finite number of graphs associated to K characters. Each such graph is associated to a particular interesting character i: its nodes are the other characters, and a link exists between the nodes j, k if the characters j, k talked about i (counting how many times). This kind of approach brought us to define two new centrality measures, not on the nodes, as it is generally done, but on the K graphs, leading us to a ranking of importance of the nodes i. On our five plays we tested the standard method for community detection and graph partitioning, plus our new algorithm based on the Voronoi cells. Finally, with the help of prof. R. Coronato of Università di Padova, we interpret our final results and obtain several clear indications –some obvious, some not- as to the meaning of the plays and the role of the characters in them.

Chapter 1

Background

1.1 A brief introduction to Graph Theory

A *graph* G is defined as the ordered pair $G = (V, \mathcal{E})$ where V is a set of $n := |V|$ elements called *nodes* and $\mathcal{E}$ is a set of $e := |\mathcal{E}|$ elements called *edges*. In particular, $\mathcal{E}$ can be written as follows:

$$\mathcal{E} = \{(i, j) \in V \times V : \ i \text{ connected to } j\}.$$

G is called *undirected* if $(i, j) \in \mathcal{E}$ implies $(j, i) \in \mathcal{E}$; otherwise G is called *directed*.

A graph G can be also represented by a $n \times n$ matrix $A = (a_{ij})$ called *adjacency matrix* defined (with abuse of notation on the nodes' label) as follows:

$$a_{ij} = \begin{cases} 1 & \text{if } (i, j) \in \mathcal{E} \\ 0 & \text{else} \end{cases}$$

We remark that G is undirected $\iff$ A is symmetric.

A *path* in a (un)directed graph G is an ordered k-tuple of nodes $w = (i_1, \ldots, i_k) \in V^k$ such that $i_l \neq i_s$ *for* $l, s \in \{1, \ldots, k\}$, $l \neq s$, and there is a (un)directed edge between i_l and i_{l+1} for all $l = 1, \ldots, k-1$. We call k the *length* of the path. We denote the set of all directed paths of length k in a directed graph as $P_k^+ \subseteq V^k$ and (P_k) for the undirected case. Finally, we denote the set of all shortest paths of length k as P_k^{geo}.

An undirected graph is *connected* if there exists a path between every pair of nodes. A directed graph is *strongly connected* if there exists a directed path between every pair of nodes.

Finally, a graph G is called *simple* if $(i, i) \notin \mathcal{E} \ \forall \ i \in V$ (absence of *loops*). In literature *simple* means also that G must not have *multiple edges*: that is

an example of the more generic case of a *weighted* graph, which is an ordered triple $G = (V, \mathcal{E}, w_G)$ in which V and $\mathcal{E}$ have the same meaning as before and w_G is a map:

$$\begin{aligned} w_G : \mathcal{E} &\longrightarrow \mathbb{R} \\ (i,j) &\mapsto w_{ij} \end{aligned}$$

$w_G(\mathcal{E})$ can be seen as a set of "weights" that we associate to every edge in $\mathcal{E}$. In particular, if $w_G(\mathcal{E}) \subset \mathbb{N}$ then every weight w_{ij} can be seen as the number of repeated edges (i,j) in $\mathcal{E}$.

In the weighted case we define (with abuse of notation) the adjacency matrix as $W = (w_{ij})$ of G and $G' = (V, \mathcal{E})$ the *associated unweighted graph.*

For ease of exposition, when it is not specified, the graph G has every weight equals to 1 and we write as before $G = (V, \mathcal{E})$.

1.2 Perron-Frobenius Theorem

In order to understand how some centrality measures work we start by recalling an important linear algebra result for positive matrices and for irreducible non-negative matrices. We will write $A \geq 0$ if $a_{ij} \geq 0 \ \ \forall\ i,j = 1,\dots,n$ and $A > 0$ if the inequality is strict. All the materials in this section can be found in [1] and [8].

Theorem 1.2.1 (Perron-Frobenius Theorem for positive matrix)**.** *Let $A = (a_{ij}) > 0$ be a $n \times n$ matrix. We write the set of the eigenvalues of A as $\Lambda := \{\lambda_1, \dots, \lambda_n\}$ such that $\rho(A) = |\lambda_1| \geq |\lambda_2| \geq \cdots \geq |\lambda_n|$. Then:*

1. *$\lambda_1 \in \mathbb{R}^+$ (Perron root) and $\lambda_1 > |\lambda_i|$ for $i = 2, \dots, n-1$ (i.e. A is primitive).*
2. *$\exists\ \mathbf{v} > 0$ and $\mathbf{w} > 0$ such that $A\mathbf{v} = \lambda_1 \mathbf{v}$ and $\mathbf{w}^T A = \lambda_1 \mathbf{w}^T$ (Perron vectors).*
3. *The algebraic multiplicity of λ_1 is 1.*
4. *If $\mathbf{v_i}$ (or equivalently $\mathbf{w_i}$) is a right (left) eigenvector associated to λ_i then if $\mathbf{v_i} > 0$ ($\mathbf{w_i} > 0$) $\Rightarrow i = 1$.*
5. *$\lim_{k\to\infty} \frac{A^k}{{\lambda_1}^k} = \mathbf{v}\mathbf{w}^T =: \boldsymbol{\mu}$ (Perron projection) where $\mathbf{v}$ and $\mathbf{w}$ are normalized such that $\mathbf{w}^T\mathbf{v} = 1$. The rate of convergence is of the order of $\left(\frac{\lambda_2}{\lambda_1}\right)^k$.*
6. *$\lambda_1 \leq \max\limits_{i=1,\dots,n} \sum_j a_{ij} = \|A\|_\infty$.*

Proof. First, we remark that if $\rho(A) = 0$ then A is nilpotent for some $n \in \mathbb{N}$, so $\|A^n\|_\infty = 0$ and that contradicts the positivity of A. Thus $\rho(A) > 0$.

Let $PJP^{-1} = A$ be the Jordan decomposition of A. Hence, $PJ^kP^{-1} = A^k$. We denote J_{λ_i} the $m_i \times m_i$ Jordan block associated to the eigenvalue λ_i and $N_i = (n_{js})$ the $m_i \times m_i$ nilpotent matrix such that:

$$n_{js} = \begin{cases} 1 & \text{if } s = j+1 \\ 0 & \text{else} \end{cases}$$

Then $N^{m_i} = 0$ and $N,\ \lambda_i I$ commute, so we can apply the binomial theorem:

$$J_{\lambda_i}^k = (\lambda_i I + N)^k = \sum_{l=0}^{k} \binom{k}{l} \lambda_i^{k-l} N^l = \sum_{l=1}^{\min(k, m_i - 1)} \binom{k}{l} \lambda_i^{k-l} N^l. \tag{1.1}$$

If $\rho(A) < 1$ then

$$\lim_{k\to\infty} \sum_{l=1}^{\min(k,m_i-1)} \binom{k}{l} \frac{\lambda_i^{k-l}}{{\lambda_1}^k} = 0 \ \ \forall\ i = 1, \dots, n.$$

So $\exists \lim_{k\to\infty} J^k = 0 \Rightarrow \exists \lim_{k\to\infty} PJ^kP^{-1} = \lim_{k\to\infty} A^k = 0$. We will need this after. For now on, we will assume $\rho(A) = |\lambda_1| = 1$.

If $A\mathbf{v} = \lambda_1 \mathbf{v}$ and if we denote $|\mathbf{v}| = (|\mathbf{v}_i|)_{i=1,\dots,n}$ then:

$$\|\mathbf{v}\| = |\lambda_1| \|\mathbf{v}\| = \|\lambda_1 \mathbf{v}\| = \|A\mathbf{v}\| \leq \|A\|_\infty \|\mathbf{v}\| \Rightarrow |\lambda_1| \leq \|A\|_\infty$$

and

$$|\mathbf{v}| = |\lambda_1||\mathbf{v}| = |\lambda_1 \mathbf{v}| = |A\mathbf{v}| \leq |A||\mathbf{v}| \Rightarrow \mathbf{y} := |A||\mathbf{v}| - |\mathbf{v}| \geq 0.$$

If by contradiction $\mathbf{y} > 0 \Rightarrow A\mathbf{y} > 0 \Rightarrow \mathbf{z} := A|\mathbf{v}| > 0 \Rightarrow \exists\ \epsilon > 0$:

$$A\mathbf{y} = A(|A||\mathbf{v}| - |\mathbf{v}|) = |A|\mathbf{z} - \mathbf{z} > \epsilon \mathbf{z} \Rightarrow \frac{A}{1+\epsilon}\mathbf{z} > \mathbf{z}$$

Defining $B := \frac{1}{1+\epsilon} A$ and multiplying by B the inequality above we obtain the chain:

$$B^k\mathbf{z} > B^{k-1}\mathbf{z} > \cdots > B\mathbf{z} > \mathbf{z} \ \ \forall k \in \mathbb{N}.$$

We have that $\rho(B) = \rho(\frac{1}{1+\epsilon}A) = \frac{1}{1+\epsilon}\rho(A) = \frac{1}{1+\epsilon} < 1$ so, for the argument on the Jordan decomposition of a positive matrix, we have $\lim_{k\to\infty} B^k = \mathbf{0}$ thus $\mathbf{z} < \lim_{k\to\infty} B^k\mathbf{z} = \mathbf{0}$ and that is absurd $\Rightarrow \mathbf{y} = 0$, hence $|\mathbf{v}|$ is an

eigenvector of $|A| = A$ with eigenvalue $1 \Rightarrow 1 \in \Lambda$, $1 = |\lambda_1| = \rho(A) \Rightarrow 1 = \lambda_1 > 0$.

If $\rho(A) \neq 1$ then we consider the matrix $\frac{1}{|\lambda_1|}A$. We have:

$$\rho(\frac{1}{|\lambda_1|}A) = 1 \Rightarrow \exists \mathbf{v} > 0 : \frac{A}{|\lambda_1|}\mathbf{v} = \mathbf{v} \Rightarrow A\mathbf{v} = |\lambda_1|\mathbf{v}$$

so $\mathbf{v}$ is an eigenvector of A with eigenvalue $|\lambda_1| \Rightarrow |\lambda_1| \in \Lambda$ and, as before, $\lambda_1 > 0$.

Now we have to show that $\forall\ i = 2, \ldots, n : \lambda_1 \neq |\lambda_i|$ so that $\lambda_1 > |\lambda_i|$ and that the algebraic multiplicity of λ_1 is 1. Again, we can suppose $\lambda_1 = 1$. In order to do that we prove that:

- 1 is the only eigenvalue with absolute value equals to 1;
- its index (i.e. the dimension of its Jordan block in the Jordan decomposition of A) is 1 so that we can say that the algebraic multiplicity of 1 is equal to its geometric multiplicity;
- finally, its geometric multiplicity is 1.

Let assume that $\exists\ i : |\lambda_i| = 1$. As before, if $A\mathbf{x} = \lambda_i\mathbf{x}$ then $|A||\mathbf{x}| = |\mathbf{x}| > 0$. In particular:

$$\sum_{j=1,\ldots,n} a_{kj}|x_j| = \sum_{j=1,\ldots,n} |a_{kj}x_j| = |x_k| = |\lambda_i||x_k| =$$
$$= |\lambda_i x_k| = |(A\mathbf{x})_k| = |\sum_{j=1,\ldots,n} a_{kj}x_j|.$$

This means that we have equality in the triangle inequality so

$$\exists\ \alpha_j > 0 : a_{kj}x_j = \alpha_j(a_{k1}x_1)\ \forall j = 2, \ldots, n$$
$$\Rightarrow \mathbf{x} = x_1(1, \frac{\alpha_2 a_{k1}}{a_{k2}}, \ldots, \frac{\alpha_n a_{k1}}{a_{kn}})^T =: x_1\mathbf{p} > 0$$

Then $A(x_1\mathbf{p}) = \lambda_i x_1\mathbf{p} \Rightarrow \lambda_i\mathbf{p} = A\mathbf{p} = |A\mathbf{p}| = |\lambda_i\mathbf{p}| = \mathbf{p} \Rightarrow \lambda_i = 1$.

By contradiction, we assume that the index of 1 is $m > 1$. We write A in its Jordan form, i.e. $P^{-1}AP = J$ and by hypothesis there exists a $m \times m$ Jordan block J_1 associated to the eigenvalue 1 in J. Then $\lim_{k\to\infty} \|J_1^k\|_\infty = \infty \Rightarrow \lim_{k\to\infty} \|J^k\|_\infty = \infty$.

Now:

$$\|J^k\|_\infty = \|(P^{-1}AP)^k\|_\infty \leq \|P^{-1}\|_\infty \|A^k\|_\infty \|P\|_\infty$$
$$\Rightarrow \|A^k\|_\infty \geq \frac{\|J^k\|_\infty}{\|P^{-1}\|_\infty \|P\|_\infty} \Rightarrow \|A^k\|_\infty \to \infty$$

We denote i_k the index that realizes $||A^k||_\infty$. We know that $\exists\ \mathbf{p} = (p_i) > 0 :\ A\mathbf{p} = \mathbf{p}$ so:

$$\|\mathbf{p}\|_\infty \geq p_{i_k} = \sum_{j=1,\dots,n} a^{(k)}_{i_k j} p_j \geq \sum_{j=1,\dots,n} a^{(k)}_{i_k j} \min_{i=1,\dots,n} p_i = \|A^k\|_\infty \min_{i=1,\dots,n} p_i \to \infty$$

as $k \to \infty \Rightarrow \lim_{k\to\infty} \|\mathbf{p}\|_\infty = \infty$ and that is absurd because $\mathbf{p}$ does not depend on $k \Rightarrow m = 1$.

We recall that $m_{geo}(1)$ is the number of Jordan blocks associated to 1 and $m_{alg}(1)$ is the sum of the sizes of the Jordan blocks associated to 1, so we just proved that $m_{geo}(1) = m_{alg}(1)$.

If $m := m_{geo}(1) =$dim Ker$(A - \lambda_1 I) = m_{alg}(1) > 1 \Rightarrow \exists \mathbf{v_1}, \dots, \mathbf{v_m}$ positive and linearly independent vectors : $A\mathbf{v_i} = \mathbf{v_i}\ \forall i = 1, \dots, m$. In particular, $\mathbf{v_i} \neq \alpha \mathbf{v_j} \Rightarrow \exists\ l \in \{1, \dots, n\} : \mathbf{v}_{il} \neq \mathbf{v}_{jl} \Rightarrow \mathbf{z} := \mathbf{v_i} - \frac{\mathbf{v}_{il}}{\mathbf{v}_{jl}} \mathbf{v_j}$ satisfies $A\mathbf{z} = \mathbf{z} \Rightarrow A|\mathbf{z}| = |\mathbf{z}| > 0$ but this cannot be true because $\mathbf{z}_l = 0 \Rightarrow m = 1 =$dim Ker $(A - I) \Rightarrow \exists\ \mathbf{v} > 0$ eigenvector of 1, that is the searched Perron vector.

Clearly, all the proof can be used to find the unique left-hand Perron vector $\mathbf{w} > 0$ replacing A with A^T.

If the eigenvector $\mathbf{v_i}$ associated to λ_i is positive and $\mathbf{w}$ is the left-hand Perron vector, then $\mathbf{w}^T\mathbf{v_i} > 0$ and $\mathbf{w}^T = \mathbf{w}^T A \Rightarrow \mathbf{w}^T\mathbf{v_i} = \mathbf{w}^T A\mathbf{v_i} = \mathbf{w}^T\lambda_i\mathbf{v_i} \Rightarrow \lambda_i = 1$.

We study now the behaviour at the limit of $\frac{1}{{\lambda_1}^k} A^k$. As we said before we can decompose A and study the Jordan blocks $\left(\frac{1}{\lambda_1} J_{\lambda_i}\right)^k\ \forall i = 1, \dots, n$.

If $i \neq 1$ for the same argument as the one at the beginning of the proof we have:

$$\left(\frac{J_{\lambda_i}}{\lambda_1}\right)^k = \sum_{l=1}^{\min(k, m_i - 1)} \binom{k}{l} \frac{{\lambda_i}^{k-l}}{{\lambda_1}^k} N^l \tag{1.2}$$

Since $\lambda_1 > \lambda_i$ for $i \neq 1$, then $\lim_{k\to\infty} \binom{k}{l} \frac{{\lambda_i}^{k-l}}{{\lambda_1}^k} = 0 \Rightarrow \lim_{k\to\infty} \left(\frac{J_{\lambda_i}}{\lambda_1}\right)^k = 0$.

If $i = 1$ then $J^k_{\lambda_1} = {\lambda_1}^k \Rightarrow \left(\frac{J_{\lambda_i}}{\lambda_1}\right)^k = 1$. Finally, we have that:

$$\exists \lim_{k\to\infty} \left(\frac{J_{\lambda_i}}{\lambda_1}\right)^k = (1, 0, \dots, 0)^T (1, 0, \dots, 0) \Rightarrow \lim_{k\to\infty} P\left(\frac{J_{\lambda_i}}{\lambda_1}\right)^k P^{-1} = \lim_{k\to\infty} \frac{A^k}{{\lambda_1}^k}$$

with speed of convergence of the order of $\left(\frac{\lambda_2}{\lambda_1}\right)^k$ as we can see in 1.2.

Now, we normalize the Perron vectors $\mathbf{v}$ and $\mathbf{w}^T$ such that $\mathbf{w}^T\mathbf{v} = 1$. With abuse of notation we still call them $\mathbf{v}$ and $\mathbf{w}^T$. Then we can write a Jordan decomposition of A as $PAP^{-1} = J$, where the first column of P is $\mathbf{v}$

and the first row of P^{-1} is $\mathbf{w}^T$. So:

$$P^{-1}\frac{J^k}{{\lambda_1}^k}P = \frac{A^k}{{\lambda_1}^k} \Rightarrow \lim_{k\to\infty}\frac{A^k}{{\lambda_1}^k} = \lim_{k\to\infty} P^{-1}\frac{J^k}{{\lambda_1}^k}P.$$

We remark that, to the limit, in $\frac{1}{{\lambda_1}^k}J^k$ it survives only $J_{11}^{(k)} = 1$ and the remainder tends to 0, so by a direct calculation we have

$$\lim_{k\to\infty} P^{-1}\frac{J^k}{{\lambda_1}^k}P = \mathbf{v}\mathbf{w}^T$$

. □

Definition-Proposition 1.2.1. A $n \times n$ matrix A is *irreducible* if one of the following statement is true:

1. A is not similar to a block upper triangular matrix, i.e. $\nexists P$ permutation matrix such that
$$P^{-1}AP = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix}$$
where X and Z are both square matrices.
2. $(I + A)^{n-1} > 0$
3. $\forall i, j = 1, \dots, n \ \ \exists\, m \in \mathbb{N}$ such that $(A^m)_{ij} \neq 0$

Definition 1.1. Let A be a $n \times n$ matrix, $A \geq 0$. The *period* of a state i is defined as the $GCM\{m \in \mathbb{N} : (A^m)_{ii} > 0\}$. If A is irreducible then all the states have the same period (by 3. of definition-Proposition 1.2.1). If the period is 1 then A is called *aperiodic*, otherwise it is *periodic*.

Remark 1. If A is the adjacency matrix of a graph $G = (V, \mathcal{E})$ then $A \geq 0$ and $(A^k)_{ls} = \#\{w \in P_k : w_1 = i_l,\ w_k = i_s\}$. In other words, the kth-power of the adjacency matrix gives the number of paths of length k between any pair of nodes: keeping this in mind and using 3. of the 1.2.1, it can be shown that a (directed) graph G is (strongly) connected if and only if A is irreducible.

Theorem 1.2.2 (Perron-Frobenius Theorem for irreducible non-negative matrix)**.** *Let $A = (a_{ij}) \geq 0$ be an irreducible $n \times n$ matrix with period h. We write the set of the eigenvalues of A as in theorem 1.2.2. Then:*

1. $\lambda_1 \in \mathbb{R}^+$ *(Perron root)*
2. $\lambda_1,\ \lambda_2 = \lambda_1 e^{2\pi i\frac{1}{h}},\ \lambda_3 = \lambda_1 e^{2\pi i\frac{2}{h}}, \dots,\ \lambda_h = \lambda_1 e^{2\pi i\frac{h-1}{h}} \in \Lambda$*; they all have algebraic multiplicity* 1 *and, obviously,* $\lambda_1 = |\lambda_i|\ \forall i = 1, \dots, h$.

3. $\exists\ \mathbf{v} > 0$ *and* $\mathbf{w} > 0$ *such that* $A\mathbf{v} = \lambda_1\mathbf{v}$ *and* $\mathbf{w}^T A = \lambda_1\mathbf{w}^T$ *(Perron vectors)*

4. *If* $\mathbf{v_i}$ *(or equivalently* $\mathbf{w_i}$*) is a right (left) eigenvector associated to* λ_i *and* $\mathbf{v_i} > 0$ $(\mathbf{w_i} > 0) \Rightarrow i \in \{1, \ldots, h\}$.

5. *if h=1 then* $\lim_{k\to\infty} \frac{A^k}{{\lambda_1}^k} = \mathbf{v}\mathbf{w}^T =: \mu$ *(Perron projection) where* $\mathbf{v}$ *and* $\mathbf{w}$ *are normalized such that* $\mathbf{w}^T\mathbf{v} = 1$. *The rate of convergence is of the order of* $\left(\frac{\lambda_2}{\lambda_1}\right)^k$.

There are some other results of this theorem concerning irreducible matrices. For the second proof and more details see chapter 8 of [1].

1.3 Random walks and Ergodic Theorem

The Perron-Frobenius Theorem has an interpretation in terms of random walks on a graph G that gives a more practical understanding of some important centrality measures. We recall now some basic properties and results about Markov chains that can be found in [7] and then we will exhibit the stochastic explanation of the theorem.

Definition 1.2. Let $(X_n)_{n\in\mathbb{N}}$ be a stochastic process, i.e. a sequence of random variables defined on a probability space $(\Omega, \mathcal{A}, P)$, where Ω is a set, $\mathcal{A}$ is a sigma-algebra on Ω, P is a probability and S is a discrete measurable set with respect to a sigma-algebra $\mathcal{S}$:

$$\begin{aligned} X_n : (\Omega, \mathcal{A}, P) &\longrightarrow (S, \mathcal{S}) \\ w &\mapsto X_n(w) \end{aligned}$$

$(X_n)_{n\in\mathbb{N}}$ is a *Markov chain* if it satisfies the *Markov property*:

$$P(X_n = i_n | X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_n = i_n | X_{n-1} = i_{n-1})$$

A Markov chain is *time homogeneous* if $P(X_n = j | X_{n-1} = i)$ does not depend on n. For now on, we consider time homogeneous Markov chain.

We associate to a Markov chain $(X_n)_{n\in\mathbb{N}}$ a function Q called *transition matrix* defined as:

$$Q(i, j) = P(X_n = j | X_{n-1} = i).$$

We remark that, with abuse of notations on the elements of S, Q can be seen as a $|S| \times |S|$ matrix. Now:

$$P(X_0 = i_0, \ldots, X_n = i_n) =$$

$$= P(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) P(X_n = i_n | X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) =$$

$$= P(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) P(X_n = i_n | X_{n-1} = i_{n-1}) =$$

$$= P(X_0 = i_0, \ldots, X_{n-1} = i_{n-1}) Q(i_{n-1}, i_n) =$$

$$\ldots = P(X_0 = i_0) Q(i_0, i_1) \ldots Q(i_{n-1}, i_n) = P(X_0 = i_0) \prod_{j=1}^{n} Q(i_{j-1}, i_j)$$

So:

$$P(X_n = i_n) = \sum_{i_0, \ldots, i_{n-1} \in S} P(X_0 = i_0, \ldots, X_n = i_n) =$$

$$= \sum_{i_0, \ldots, i_{n-1} \in S} P(X_0 = i_0) \prod_{j=1}^{n} Q(i_{j-1}, i_j) =$$

$$= \sum_{i_0 \in S} P(X_0 = i_0) \sum_{i_1, \ldots, i_{n-1} \in S} \prod_{j=1}^{n} Q(i_{j-1}, i_j) =$$

$$\sum_{i_0 \in S} P(X_0 = i_0) Q^n(i_0, i_n)$$

We compact this result in matrix notation: if $\boldsymbol{\mu}_n$ is the probability of X_n, we have

$$\boldsymbol{\mu}_n = \boldsymbol{\mu}_0 Q^n.$$

The Markov property guarantees that Q is a stochastic matrix, i.e. $\forall\, i, j \in S$:

1. $Q(i, j) \geq 0$
2. $\sum_j Q(i, j) = 1$

hence, 1 is an eigenvalue of Q since $Q\mathbf{1} = \mathbf{1}$.

If a time homogeneous Markov chain has *spatial homogeneity*, i.e.:

$$Q(i, j) = Q(0, j - i)$$

then the Markov chain is called *random walk*.

Definition 1.3. A Markov chain is called *irreducible* if Q is irreducible. In other words, $(X_n)_n$ is irreducible if, starting from i, at a certain time n there is a positive chance to arrive on j, for all nodes i and j.

Definition 1.4. The *period* of a state $i \in S$ is defined as

$$h_i = GCD\{n : Q^n(i,i) > 0\}.$$

A state $i \in S$ is called *aperiodic* if $h_i = 1$. Otherwise, i is *periodic*. Thus, a state i is periodic if starting from i it returns cyclically on i with a positive probability.

Definition 1.5. A probability measure $\boldsymbol{\mu} = (\mu_1, \dots, \mu_n, \dots)$ on S is *stationary for* Q if:

$$\sum_i \mu_i Q(i,j) = \mu_j, \ \forall\ j \in S.$$

In matrix notation this means $\boldsymbol{\mu} = \boldsymbol{\mu} Q$, so, if such $\boldsymbol{\mu}$ exists, it is an eigenvector for Q associated to the eigenvalue 1.

Remark 2. If a Markov chain is aperiodic and irreducible then, as we said, 1 is an eigenvalue of Q. Moreover, for Perron-Frobenius Theorem (*6.*) $\rho(Q)$= 1 since $1 = ||Q||_\infty \geq \rho(Q) \geq 1$ and all the other eigenvalues are smaller in norm. Also, there is a unique stationary probability vector $\boldsymbol{\mu}$ for Q (it exists for Perron-Frobenius Theorem and it is unique because it has to satisfy $\sum_j \mu_j = 1$). In particular, $\lim_{k\to\infty} Q^k = \mathbf{1}\boldsymbol{\mu}$, so:

$$\lim_{k\to\infty} Q^k(i,j) = \mu_j$$

with rate of convergence as $|\lambda_2|^k$.

This means that the long-term probability of being in j is independent of the initial state i and the system evolves over time to a stationary state $\boldsymbol{\mu}$. This is the probability point of view of the Perron-Frobenius Theorem, known as the *Ergodic Theorem*.

In particular, if we have a random walk on a graph G that has the properties as above, then $Q^k(i,j)$ is the probability to pass from i to j through a path on G of length k. Thus, finding its stationary state means choosing an initial state i_0, joining another connected state i_1 from i_0 with the probability specified by $Q(i_0, i_1)$ and so on, repeatedly. At the limit this process converges to a probability vector that identifies the amount of mass that passed through any node.

Chapter 2

Centrality measures

The goal of this section is to find a way to detect the most important node of a graph. Clearly, this is often a difficult request especially because "being the most important node" is not a property that can be extended to all types of networks, instead it changes depending on the context and the specific network we are studying through the model of a graph.

We introduce now some of the most common centrality measures; it is not surprising that most of them can be seen as a function of the adjacency matrix A or, at least, as an information that we can find out by a manipulation of A.

2.1 Degree centrality

Let G be an undirected graph. The simplest way to define 'importancy' of the node i in G is to count the number of nodes j such that $\exists\ (i,j) \in \mathcal{E}$, the so called *neighbors* of i. We define this number as the *degree* of i.

Note that, if $\mathbf{1}$ is the vector of all ones $(1,\dots,1)^T$, $\mathbf{d} := A\mathbf{1}$ gives the column vector such that

$$\mathbf{d}_i = \sum_{j=1}^{n} a_{ij}$$

is the degree of i.

This can be naturally extended to directed graphs. The only difference is that we have to distinguish two types of degree:

- $\mathbf{d}_i^{out} = [A\mathbf{1}]_i$ *out-degree of* i
- $\mathbf{d}_i^{in} = [A^T\mathbf{1}]_i$ *in-degree of* i

Clearly, if G is undirected $\mathbf{d}_i^{out} = \mathbf{d}_i^{in} = \mathbf{d}_i$.

If $G = (V, E, w_G)$ is a weighted graph, then we define in-degree and out-degree of i as before, but clearly the weighted in-degree and out-degree are no longer the number of incoming and outgoing neighbors.

2.2 Closeness centrality

Let G be a connected undirected graph. The *distance* between two nodes i and j is defined as: $d(i,j) = \min\{k \in \mathbb{N} : w \in P_k,\ w_1 = i,\ w_k = j\}$; in other words $d(i,j)$ is the length of the shortest path that connects i and j.

Bavelas in 1950 (see [9]) defined the *closeness centrality of* i as

$$\mathbf{C}_i = \frac{1}{\sum\limits_{j \in V} d(i,j)}$$

The more a node has short distances with all the other nodes (in this sense it is 'closer'), the more its closeness is large.

If G is not connected, then there are two generalizations of closeness centrality. The first one is the *harmonic centrality* (see [10] and [11]):

$$\mathbf{H}_i = \sum_{j \in V} \frac{1}{d(i,j)}$$

When $d(i,j) = \infty$ we use the convention $\frac{1}{\infty} = 0$.

The second one can be seen as a generic formula (that we will use in our applications) that considers also the size n of the network and the number n_i of reachable nodes from i:

$$\mathbf{c}_i = \left(\frac{n_i}{n-1}\right)^2 \cdot \mathbf{C}_i$$

The generalizations to strongly connected directed graphs are the following:

- $\mathbf{c}_i^{out} = \left(\frac{n_i}{n-1}\right)^2 \cdot \frac{1}{\sum\limits_{j \in V} d^+(i,j)}$ *out-closeness of* i
- $\mathbf{c}_i^{out} = \left(\frac{m_i}{n-1}\right)^2 \cdot \frac{1}{\sum\limits_{j \in V} d^+(j,i)}$ *in-closeness of* i

where m_i is the number of nodes that can reach i, n_i is the number of reachable nodes from i and $d^+(i,j) := \min\{k \in \mathbb{N} : w \in P_k^+,\ w_1 = i,\ w_k = j\}$.

2.3 Betweenness centrality

In a lot of networks (the most popular example is a telecommunication network) it is interesting to define 'importance' as the capacity of a node of being in the middle of a lot of shortest paths between the other nodes of the graph. That means in some sort of way that a node i with a large betweenness centrality is a good mediator/intermediary in the network.

Let G be a directed graph. Freeman (see [12]) gave the first definition of betweenness centrality of a node i as follows:

$$\mathbf{B}_i = \sum_{\substack{l,s \in V \\ l \neq s}} \frac{g_{(l,i,s)}}{g_{(l,s)}}$$

where $g_{(l,i,s)}$ is the number of shortest paths $w \in \bigcup_k P_k^{geo}$ such that if $w \in P_k^{geo}$ then $w_1 = l$, $w_k = s$, $\exists j \in \{2, \dots, k-1\}$ such that $w_j = i$ and $g_{(l,s)} = \sum_{j \in V} g_{(l,j,s)}$.

Every terms of the sum $\mathbf{B}_i$ can be seen as the probability to pass through i walking in a shortest path from l to s. So $\mathbf{B}_i$ is a quantification of how much i is present in the shortest paths between every other pair of nodes.

Note that with these definitions, in the case of an undirected graph we have $g_{(l,s)} = g_{(s,l)}$ and $g_{(l,i,s)} = g_{(s,i,l)}$ $\forall l, s \in V$, $l \neq s$, so that in the formula for $\mathbf{B}_i$ we count twice two identical probabilities; for this reason it is better to redefine in the undirected case the formula as:

$$\mathbf{B}_i^U = \frac{1}{2}\mathbf{B}_i$$

2.4 Eigenvector centrality

Let us consider two nodes i and j in an undirected network such that $\mathbf{d}_i = \mathbf{d}_j$; considering degree centrality, i and j have the exact same importance in the graph, but in a lot of networks this is not how things work. Precisely, we would like to distinguish the case in which i is connected to some important nodes and j isn't, increasing the centrality of i. This is eigenvector centrality's job, so we can consider it as a much more precise degree centrality. Let's see how we can formalize this concept, using the argument mentioned in [4].

Let $G = (V, E)$ be a strongly connected (un)directed graph, A its adjacency matrix and let $\mathbf{x}$ be the unknown vector of centrality. We want to define a sequence of vector $\mathbf{x}_k$ such that $\mathbf{x}_{k+1}$ is an improved version of $\mathbf{x}_k$, in the sense that we have increased the centrality $(\mathbf{x}_k)_i$ of a node i if i is connected to important nodes. We first remark that A satisfies the hypothesis of

Perron-Frobenius Theorem for non negative irreducible matrix; in particular, denoting the set of the eigenvalues as in the statement of the Theorem, we have that $\lambda_1 \geq |\lambda_i| \ \forall \ i \neq 1$.

We start by setting:

$$\mathbf{x}_0 = \frac{1}{\lambda_1}\mathbf{1}$$

i.e. every node has the exact same importance equals to $\frac{1}{\lambda_1}$. The need of adding the factor $\frac{1}{\lambda_1}$ will be clear soon. Now, we want to give more importance to the nodes that are connected to other important nodes for $\mathbf{x}_0$, that for now are every vertex. This leads us to define:

$$\mathbf{x}_1 := \frac{1}{\lambda_1}A\mathbf{x}_0 = \frac{1}{\lambda_1}\mathbf{d}$$

i.e. the scaled vector of degree centrality. In general, if we want to improve $\mathbf{x}_k$ in the sense that we discussed before, we have to define:

$$(\mathbf{x}_{k+1})_i = \frac{1}{\lambda_1}\sum_{\{j:\ (i,j)\in\mathcal{E}\}}(\mathbf{x}_k)_j = \sum_j \frac{1}{\lambda_1}a_{ij}(\mathbf{x}_k)_j$$

hence:

$$\mathbf{x}_{k+1} = A\mathbf{x}_k = \frac{1}{\lambda_1^k}A^k\mathbf{x}_0.$$

It is natural to think that the limit of this sequence gives the best optimization vector, i.e. the searched $\mathbf{x}$. Let's calculate it.

We know that $\exists\ c_i \in \mathbb{R}$:

$$\mathbf{x}_0 = \sum_i c_i\mathbf{v_i}$$

where $\mathbf{v_i}$ are the eigenvectors of A. Thus:

$$\mathbf{x}_k = A^k\mathbf{x}_0 = A^k\sum_i c_i\mathbf{v_i} = \sum_i c_iA^k\mathbf{v_i} = \sum_i c_i\lambda_i^k\mathbf{v_i} = \lambda_1^k\sum_i c_i\frac{\lambda_i^k}{\lambda_1^k}\mathbf{v_i}$$

$$\Rightarrow \lim_{k\to\infty}\frac{1}{\lambda_1^k}\mathbf{x}_k = c_1\mathbf{v_1}$$

We find out that the limiting vector is proportional to $\mathbf{v_1}$, i.e. the searched $\mathbf{x}$ satisfies $A\mathbf{x} = \lambda_1\mathbf{x}$. In other word, it is an eigenvector for λ_1 and this is why we added $\frac{1}{\lambda_1}$ a priori. Since centrality measures have sense if they are positive, by Perron-Frobenius Theorem we know that there is a positive eigenvector for λ_1, $\mathbf{p} > 0$. Finally, we define the *eigenvector centrality of* i as $\mathbf{E}_i := \mathbf{p}_i$.

2.5 PageRank

PageRank is the algorithm that Google uses to order the list of their search engine results. We can see the World Wide Web has a graph where the nodes are the websites and the edges are the links from a site to another.

An obvious but important remark to make is that if the website i has a link to the website j there's no guarantee that j has a link to i. So the graph can be directed, and this is one of the strength of the method. Even if it has born for the website ranking purpose, PageRank constitutes an actual centrality measure that can be used in every network and can be seen as a better eigenvector centrality. Let's see how and why, following the approach suggested in [2] of the Page algorithm explained in [13] and [3]. It is clear that this method gives perfectly strong results for undirected graph.

Let $G = (V, E)$ be a directed graph, A its adjacency matrix. Our ultimate goal is to construct an irreducible stochastic matrix so that we can use Perron-Frobenius theorem to find a positive centrality measure for the nodes.

We first want to define a matrix $H = (h_{ij})$ that represents the probability for a node i to be reached (in one step) by the node j; if there is no link from j to i the chance is 0, otherwise it is $\frac{1}{\mathbf{d}_j^{out}}$.

Let $D = (d_{ij})$ be the diagonal matrix such that

$$d_{ij} = \begin{cases} \mathbf{d}_j^{out} & \text{if } i = j \text{ and } \mathbf{d}_j^{out} > 0 \\ 1 & \text{if } i = j \text{ and } \mathbf{d}_j^{out} = 0 \\ 0 & \text{else} \end{cases}$$

We remark that the apparently not necessary condition on j such that $\mathbf{d}_j^{out} = 0$ ensures that D is invertible. Let $H = A^T D^{-1} = (h_{ij}) = (a_{ji} d_{jj}^{-1})$. Thus:

$$h_{ij} = \begin{cases} 0 & \text{if } (j,i) \notin \mathcal{E} \\ \frac{1}{d_{ij}} & \text{else} \end{cases}. \tag{2.1}$$

We remark that if $\exists\ j : \ \mathbf{d}_j^{out} = 0 \Rightarrow a_{ji} = 0\ \forall i \in \{1, \dots, n\} \Rightarrow h_{ij} = 0\ \forall i \in \{1, \dots, n\}$. This nodes are called *dangling nodes* and correspond to the columns of H made of all zeros. The goal is to make possible the walk from the dangling nodes to some other node. We fix it introducing an uniform discrete distribution for the dangling nodes as follows: we denote as $\mathbf{I}$ the vector with the indices of the dangling nodes and we construct $\mathbf{a}$ such that

$$\mathbf{a}_j = \begin{cases} 1 & \text{if } j \notin \mathbf{I} \\ 0 & \text{else} \end{cases}.$$

We define:

$$S = H + \frac{1}{n}\mathbf{1}\mathbf{a}^T$$

In other words, we are hypothetically adding $(j,i) \in \mathcal{E}\ \forall i \in \{1,\dots,n\},\ \forall j \in \mathbf{I}$. This can still be a reducible matrix, so we introduce a constant $\alpha \in (0,1)$ (the so called *Google damping factor*) and we finally define the *Google matrix*:

$$G_\alpha =: \alpha S + (1-\alpha)\frac{1}{n}\mathbf{1}\mathbf{1}^T$$

We remark that S represents a random walk on G, or equivalently a weighted directed graph where every edge (i,j) in G has as weight the probability of reaching j from i, and where we connected all the nodes to all the dangling nodes j with weight $s_{ij} = \frac{1}{n}$. But we can still have nodes that are not reachable from some others. So we introduced an imaginary node that with probability $1-\alpha$ is reached by any other node with uniform probability $\frac{1}{n}$, so the graph G becomes strongly connected and this leads us to the irreducibility of G_α. Finally, for Perron-Frobenius Theorem it exists the Perron vector $\mathbf{p}(\alpha)$ associated to the eigenvalue 1, so we define the *PageRank of node* i as $\mathbf{p}(\alpha)_i$. We could have chosen any other probability vector instead of the uniform one, but this choice in our case is the most reasonable.

This is a procedure that shows how we "modify" the properties of the graph considering G_α instead of G, in order to have a better mathematical environment.

2.6 Analytic functions and centrality measures

Since a graph can be seen as a matrix, it is natural to ask what are the conditions under which we can define a centrality measure as a function of the adjacency matrix A. We are interested in finding a class of positive measures that can be manipulated through a parameter in order to give more importance to shorter or longer walks, based on our requirements. All the concepts of this section can be found in [2].

Let $B_R := \{z \in \mathbb{C} :\ |z| < R\}$. We consider the class of analytic functions with positive coefficients:

$$\mathcal{F} = \{f : \mathbb{C} \to \mathbb{C} :\ \exists\, R_f > 0,\ \ c_0 \geq 0,\ (c_k)_{k\in\mathbb{N}} \in \mathbb{R}^+\ :\ \forall z \in B_{R_f}\ f(z) = \sum_{k=0}^{\infty} c_k z^k\}$$

In the following we refer to R_f as the radius of convergence of f and we will use all the notations of theorem 1.2.2.

Definition 2.1. A function f is *defined on the spectrum of* A if $\forall\ \lambda_i \in \Lambda$ $\exists\ f^{(j)}(\lambda_i)\ \forall\ j = 0, \dots, m_i$, where m_i is the largest size of the Jordan blocks associated to λ_i in the Jordan decomposition $A = PJP^{-1}$.

If f is defined on the spectrum of A we define:

$$f(J) := \begin{bmatrix} f(J_1) & 0 & \cdots & 0 \\ 0 & f(J_2) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(J_n) \end{bmatrix}$$

where:

$$f(J_i) := \begin{bmatrix} f(\lambda_i) & f'(\lambda_i) & \cdots & \frac{f^{(m_i-1)}(\lambda_i)}{(m_i-1)!} \\ 0 & f(\lambda_i) & \cdots & \frac{f^{(m_i-2)}(\lambda_i)}{(m_i-2)!} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & f(\lambda_i) \end{bmatrix}$$

and $f(A) := Pf(J)P^{-1}$.

Proposition 2.6.1. *Let* $f \in \mathcal{F}$, $f(z) = \sum_{k=0}^{\infty} c_k z^k\ \forall\ z \in B_{R_f}$.

If f *is defined on the spectrum of* A *then* $f(A) = \sum_{k=0}^{\infty} c_k A^k > 0$.

In general, if $\tau := \frac{R_f}{\lambda_1}$, *then* $\forall\ t \in (0, \tau)$ $f(tA)$ *is well defined and* $f(tA) > 0$.

Proof. As we remarked in 1.1 we can write $J_i = \lambda_i I + N_i$ where N_i is nilpotent.

If f is defined on the spectrum of A then:

$$f(J_i) = f(\lambda_i)I + f'(\lambda_i)N_i + \cdots + \frac{f^{(m_i-1)}(\lambda_i)}{(m_i-1)!}N_i^{m_i-1} =$$

$$= \sum_{k=0}^{\infty} c_k \lambda_i^k I + \sum_{k=1}^{\infty} k c_k \lambda_i^{k-1} N_i + \cdots + \sum_{k=m_i-1}^{\infty} \binom{k}{m_i-1} c_k \lambda_i^{k-(m_i-1)} N_i^{m_i-1} =$$

$$= \sum_{k=0}^{\infty} c_k \sum_{l=0}^{\min(k, m_i-1)} \binom{k}{l} \lambda_i^{k-l} N^l = \sum_{k=0}^{\infty} c_k J_i^k$$

With a direct calculation we have $P^{-1}f(A)P = f(J) = \sum_{k=0}^{\infty} c_k J^k$, hence:

$$f(A) = \sum_{k=0}^{\infty} c_k (PJ^kP^{-1}) = \sum_{k=0}^{\infty} c_k A^k.$$

If $f \in \mathcal{F}$ then surely f is defined in $t\lambda_1$ $\forall t$ such that $|t\lambda_1| < R_f$, i.e. $\forall\, t \in (0, \tau)$. So f is defined on the spectrum of $tA \Rightarrow f(tA) = \sum_{k=0}^{\infty} c_k t^k A^k$.

In both cases the matrix is strictly positive because $c_k > 0$ for all k. $\square$

2.6.1 Examples

We show some other well-known centrality measures that are examples of the previous concepts.

1. $f(\beta z) = e^{\beta z}$. $R_f = \infty$ so for what we said we can choose $\beta > 0$.
   The **exponential subgraph centrality** of node i is defined as:

   $$\mathbf{SC}(\beta)_i = [e^{\beta A}]_{ii} = (\sum_{k=0}^{\infty} \frac{\beta^k}{k!} A^k)_{ii}$$

   $(A^k)_{ii}$ is the number of all walks of length k that begin and end with i (*closed walks*). So $\mathbf{SC}(\beta)_i$ is a measure of how easy i can return to itself walking through any path in the entire network. The coefficients $\frac{\beta^k}{k!}$ penalize the contribution of the number of path of length k when k is large, giving more weight to the shorter paths.

   The **total communicability centrality** of node i is defined as:

   $$\mathbf{TC}(\beta)_i = [e^{\beta A}\mathbf{1}]_i$$

   The difference with exponential subgraph centrality is that total communicability centrality counts all the walks that just begin with i, i.e. it is a measure of how i "communicates" with all the other nodes j, again, in the entire network.

2. $f(\alpha z) = (1 - \alpha z)^{-1} = \sum_{k=0}^{\infty} \alpha^k z^k$ $R_f = 1$ so we can choose $0 < \alpha < \frac{1}{\lambda_1}$.
   We remark that with these conditions on α the matrix $(I - \alpha A)$ is invertible and its inverse can be written as its Neumann expansion, i.e. the power series as above.

   The **resolvent subgraph centrality** of node i is defined as:

   $$\mathbf{RC}(\alpha)_i = [(I - \alpha A)^{-1}]_{ii} = (\sum_{k=0}^{\infty} \alpha^k A^k)_{ii}$$

   and the **Katz centrality** of node i is defined as:

   $$\mathbf{K}(\alpha)_i = [(I - \alpha A)^{-1}\mathbf{1}]_i = (\sum_{k=0}^{\infty} \alpha^k A^k \mathbf{1})_i.$$

Similar considerations as the ones above can be done for this two centralities.

2.6.2 Limiting behavior

The crucial point of using this class of centralities is to find the better choice for the parameter in order to avoid the ones that will give similar ranking of importance for the nodes to other centrality formula or even to the centrality in question itself.

Let $\mathcal{F}_\infty$ be the subset of functions $f \in \mathcal{F}$ such that $R_f = \infty$ and $\mathcal{F}_{lim}$ be the subset of functions $f \in \mathcal{F}$ such that

$$R_f < \infty \text{ and } \sum_{k=0}^{\infty} c_k R_f^k = \infty.$$

We remark that $\mathcal{F}_\infty \cup \mathcal{F}_{lim} \subsetneq \mathcal{F}$ because if we consider $f(z) = \sum_{k=1}^{\infty} \frac{z^k}{k^2}$ we have $R_f = 1$ and $\sum_{k=1}^{\infty} \frac{1}{k^2} < \infty$.

Definition 2.2. Let $G = (V, \mathcal{E})$, $|V| = n$, be a graph. We name the nodes as $\{1, \ldots, n\}$. Let the map

$$\begin{aligned} c : V &\longrightarrow \mathbb{R} \\ i &\mapsto c(i) \end{aligned}$$

be a centrality measure on G. We denote $\mathbf{C} \in \mathbb{R}^n$ the vector such that $\mathbf{C}_i = c(i)$. We order the nodes of V as $\{i_1, \ldots, i_n\}$ so that if $l \leq s$ then $c(i_l) \geq c(i_s)$. We define the sequence:

$$\begin{cases} r_{i_1} = 1 \\ r_{i_{k+1}} = r_{i_k} \delta_{c(i_{k+1}),c(i_k)} + k(1 - \delta_{c(i_{k+1}),c(i_k)}) \text{ for } 1 < k < n \end{cases}$$

where δ is the Kronecker delta.

Finally we define the *ranking of* $\mathbf{C}$ as the vector $\mathbf{R_C}$ such that $(\mathbf{R_C})_i = r_i$.

Theorem 2.6.2 (Undirected case)**.** *Let $G = (V, E)$ be a connected, undirected graph with primitive adjacency matrix A. Let $f \in \mathcal{F}$ be defined on the spectrum of A and let $\mathbf{SC}(t)$ and $\mathbf{TC}(t)$ be the vectors:*

- $\mathbf{SC}(t)_i := [f(tA)]_{ii}$ *f-subgraph centrality*

- $\mathbf{TC}(t)_i := [f(tA)\mathbf{1}]_i$ *f-total communicability centrality*

Then:

1. $\lim_{t\to 0^+} \mathbf{R_{SC}}(t) = \lim_{t\to 0^+} \mathbf{R_{TC}}(t) = \mathbf{R_d}$ *where* $\mathbf{d}$ *is the degree centrality.*
2. *If* $f \in \mathcal{F}_\infty \cup \mathcal{F}_{lim}$ *then:*

$$\lim_{t\to \tau^-} \mathbf{R_{SC}}(t) = \lim_{t\to \tau^-} \mathbf{R_{TC}}(t) = \mathbf{R_E} = \mathbf{v}$$

   where $\mathbf{E}$ *is the eigenvector centrality and* $\mathbf{v}$ *the right-hand Perron vector.*

3. *If we replace in the definition of* f*-total communicability* $\mathbf{1}$ *with some other positive vector the previous statements are still true.*

Remark 3. Before proving it we remark that this theorem tells us that while the parameter decays to 0 we are basically giving more and more weight to the shorter paths (because the weights of longer walks decay faster), until near 0 we are in fact dealing with paths of length 1, i.e. the degree. Same argument with opposite conclusion can be given for t that grows. That's why the parameter is fundamental if we want to use the same centrality in different situations, whether we have a network where 'importance' has to be given to shorter paths (local influence), whether we want the contrary (global influence).

Lemma 2.6.3. *Let* $a(t) := \sum_{k=0}^{\infty} a_k t^k$, $b(t) := \sum_{k=0}^{\infty} b_k t^k$ *with* $a_k > 0$, $b_k > 0$ $\forall\ k \in \mathbb{N}$ *and* $a(t) < \infty$, $b(t) < \infty\ \forall\ t \geq 0$.

If $\lim_{k\to\infty} \frac{a_k}{b_k} = 0$ *then* $\lim_{t\to\infty} \frac{a(t)}{b(t)} = 0$

Proof. Let $\epsilon > 0$. For hypothesis $\exists\ N \in \mathbb{N}$ such that $\forall k > N\ \frac{a_k}{b_k} < \epsilon$. Then:

$$\frac{a(t)}{b(t)} = \frac{\sum_{k=0}^{N} a_k t^k}{\sum_{k=0}^{N} a_k t^k + \sum_{k=N+1}^{\infty} a_k t^k} + \frac{\sum_{k=N+1}^{\infty} \frac{a_k}{b_k} b_k t^k}{\sum_{k=0}^{N} a_k t^k + \sum_{k=N+1}^{\infty} a_k t^k} <$$

$$< \frac{\sum_{k=0}^{N} a_k t^k}{\sum_{k=0}^{N} a_k t^k + \sum_{k=N+1}^{\infty} a_k t^k} + \epsilon \frac{\sum_{k=N+1}^{\infty} b_k t^k}{\sum_{k=0}^{N} a_k t^k + \sum_{k=N+1}^{\infty} a_k t^k}$$

The first term tends to 0 as $t \to \infty$ so we have the thesis. □

Lemma 2.6.4. *Let A be a primitive matrix and $f \in \mathcal{F}_\infty \cup \mathcal{F}_{lim}$ defined on the spectrum of A. Then:*

$$\lim_{t\to\tau^-} \frac{t^j f^{(j)}(t\lambda_1)}{f(t\lambda_i)} = 0 \ \forall\ j \in \mathbb{N},\ \forall\ i = 2, \dots, n$$

Proof. If $f \in \mathcal{F}_{lim}$ then $t^j f^{(j)}(t\lambda_1)$ is a power series with radius of convergence R_f. $t\lambda_1 < R_f$ is $t \to \tau$ so $t^j f^{(j)}(t\lambda_1)$ is finite. Plus, $\lim_{t\to\tau^-} f(t\lambda_i) = \infty$ for all $i \neq 1$, so we have the thesis.

Let $f \in \mathcal{F}_\infty$. As before, $f^{(j)}$ has the same radius of convergence of f and

$$|t^j f^{(j)}(t\lambda_1)| \leq \sum_{k=0}^{\infty} (k+j)\dots(k+1)c_{k+j}t^{k+j}|\lambda_i|^k < \infty$$

We denote $a_k = (k+j)\dots(k+1)c_{k+j}\lambda_i^k$, $b_k = c_k\lambda_i^k$. For lemma 2.6.3 we have the thesis. □

Proof of theorem 2.6.2. $\mathbf{SC}(t)_i = \sum_{k=0}^{\infty} c_k t^k [A^k]_{ii}$. The ranking is invariant for translation and scalar multiplication of the elements of the centrality so if we define:

$$\mathbf{Y}_i(t) := \frac{1}{c_2 t^2}[\mathbf{SC}(t) - c_0\mathbf{1}]_i = \mathbf{d}_i + \sum_{k=3}^{\infty} \frac{c_k}{c_2} t^{k-2}[A^{k+1}]_{ii}$$

then $\mathbf{R}_{\mathbf{Y}(t)} = \mathbf{R}_{\mathbf{SC}(t)}$. So:

$$\lim_{t\to 0^+} \mathbf{R}_{\mathbf{SC}(t)} = \lim_{t\to 0^+} \mathbf{R}_{\mathbf{Y}} = \mathbf{R}_{\mathbf{d}}.$$

We can do the same defining:

$$\mathbf{Y}_i(t) := \frac{1}{c_1 t}[\mathbf{TC}(t) - c_0\mathbf{1}]_i;$$

we have the same result for $\mathbf{R}_{\mathbf{TC}(t)}$.

$\mathbf{SC}(t)_i = \sum_{k=1}^{n} f(t\lambda_k)\mathbf{v_k}_i^2$, where $\mathbf{v_k}$ is the unit right eigenvector associated to λ_k and $\mathbf{v_1} > 0$ is the right Perron eigenvector. We define:

$$\mathbf{Y}(t)_i := \frac{1}{f(t\lambda_1)}\mathbf{SC}(t)_i = \mathbf{v_1}_i^2 + \sum_{k=2}^{n} \frac{f(t\lambda_k)}{f(t\lambda_1)}\mathbf{v_k}_i^2 \to \mathbf{v_1}_i^2$$

as $t \to \tau^-$, for lemma 2.6.4. Thus, we have the thesis since $\mathbf{R_{v_1}} = \mathbf{R_{v_1}}^2$. We can do the same with $\mathbf{TC}(t)_i = \sum_{k=1}^{n} f(t\lambda_k)(\mathbf{v_k}^T\mathbf{1})\mathbf{v_k}_i$. We define:

$$\mathbf{Y}(t)_i := \frac{1}{f(t\lambda_1)(\mathbf{v_k}^T\mathbf{1})}\mathbf{TC}(t)_i$$

and we obtain the same result.

All these arguments can be done identically with any other positive vector instead of $\mathbf{1}$, so the proof is complete. □

Remark 4. If A is not primitive then we can consider the primitive matrix $A_\epsilon = \epsilon A + (1-\epsilon)I$, $0 < \epsilon < 1$ and if the limit behaviour of $f(tA_\epsilon)$ still depends on ϵ we can then calculate the limit for $\epsilon \to 0^+$. It is useful to remark that $\rho(A) = \rho(A_\epsilon)$ so the radius of convergence is identical.

The theorem 2.6.2 can be extended to directed weighted graphs but first a remark has to be done: when we are working with directed networks we have to make a difference between central nodes that are important based on their incoming edges (*authorities*) and central nodes that are important based on their outgoing edges (*hubs*). In mathematical terms, this means that we can calculate $f(tA)$ if we are interested in detecting authority nodes and $f(tA^T)$ for the hubs ones.

We now give a table that resumes the general case of the directed weighted graph. The proof is similar to the previous one, although with some variations: instead of the spectral decomposition we must use the more general Jordan decomposition. More details are in [2].

Theorem 2.6.5. *Let $G = (V, E, w_G)$ be a generic strongly connected, directed, simple graph with adjacency matrix W, with $w_G \geq 0$. We denote $\mathbf{v}$ the right-hand Perron vector of W and $\mathbf{w}$ the left-hand one. With the same hypothesis on f as in theorem 2.6.2, denoting $\mathbf{TC^T}(t) = [f(tW^T)\mathbf{1}]$, we have the same results for hubs and authorities centralities, resumed in the following table:*

	$\alpha \to 0^+$	$\alpha \to \tau^-$
$\mathbf{TC}(t)$	$\mathbf{d}^{out}$	$\mathbf{v}$
$\mathbf{TC^T}(t)$	$\mathbf{d}^{in}$	$\mathbf{w}$

Remark 5. The *spectral gap* of a matrix A is the value $|\lambda_1 - \lambda_2|$. In undirected graph we can see in the proof of the theorem that the speed of convergence of the ranking as $t \to \tau^-$ depends on the spectral gap. If the spectral gap is big, when $t \to \tau^-$ $f(t\lambda_1)$ will grow much faster than $f(t\lambda_2)$ and consequently than $f(t\lambda_k)$ for $k = 3, \ldots, n$. Thus in networks with a large spectral gap it is smart to use eigenvector as a measure of centrality.

As $t \to 0^+$ we see that $\mathbf{Y}_i(t) - \mathbf{d}_i$ is dominated by $\frac{c_3}{c_2} t(A^3)_{ii}$, so the number of triangles where i is a vertex is crucial for the speed of convergence to degree centrality. If i is in a lot of triangles then its position in the graph is similar to the one of a node of a *clique*, i.e. a (sub)graph where every node shares an edge with all the others. Thus i has the possibility to communicate (i.e. exchange information) with a dense community in the network so its centrality takes time to become just the degree one when we are focusing on local influence.

2.6.3 Extension to PageRank

It is worthy to mention a result that involves the limiting behavior of the PageRank when the parameter α tends to zero.

Theorem 2.6.6. *Let G_α be the Google matrix, H as in 2.1 and $\mathbf{p}(\alpha)$ be the PageRank of G, $0 < \alpha < 1$. Then:*

$$\lim_{\alpha \to 0^+} \mathbf{R}_{\mathbf{p}(\alpha)} = H\mathbf{1}.$$

We don't prove this theorem but we give an idea. It can be shown [3] that $\mathbf{p}(\alpha) = \mathbf{x}(\mathbf{x}^\mathbf{T}\mathbf{1})^{-1}$ where $\mathbf{x}$ satisfies:

$$(I - \alpha H)\mathbf{x} = \frac{1}{n}\mathbf{1} := \mathbf{n}$$

The system is solvable because $I - \alpha H$ is invertible and equals to its Neumann expansion. So $\mathbf{x} = \sum_{k=0}^{\infty} \alpha^k H^k \mathbf{n}$ and $\frac{\mathbf{x}-\mathbf{n}}{\alpha}$ has the same ranking as $\mathbf{x}$. But then:

$$\lim_{\alpha \to 0^+} \frac{\mathbf{x} - \mathbf{n}}{\alpha} = H\mathbf{n} = \frac{1}{n} H\mathbf{1} \Rightarrow \lim_{\alpha \to 0^+} \mathbf{R}_{\mathbf{p}(\alpha)} = \lim_{\alpha \to 0^+} \mathbf{R}_{\frac{\mathbf{x}-\mathbf{n}}{\alpha}} = \mathbf{R}_{\frac{1}{n}H\mathbf{1}} = \mathbf{R}_{H\mathbf{1}}$$

and the last term is essentially the ranking of the weighted in-degree.

The behavior near $\tau = 1$ is bad for some kind of networks in terms of accuracy, especially the one that are not enough connected. Basically, α near the endpoints tends to delete all the good characteristics of the PageRank procedure. This is not surprising because choosing $\alpha \sim 1$ means join the the damping factor very rarely and choosing $\alpha \sim 0$ means join the nodes of the graph very rarely. That's why, in general, the choice of the parameter is fundamental to avoid multiple data set rankings that are not too correlated.

The most common choice, proposed by Page, is $\alpha = 0.85$. For more details see [13].

Chapter 3

Applications to Shakespeare's plays and new centrality measures

We will show in this chapter how we can use all the previous concepts and methods to analyze drama; in fact, a study of literature works through this theory can be useful for three reasons: confirm a qualitative analysis with a purely quantitative analysis, enlighten a criterion that predicts qualitative aspects and enlighten aspects that cannot be seen immediately with a qualitative check of the plays.

We have studied five William Shakespeare's works: *Macbeth, Romeo and Juliet, Richard III, Twelfth night* and *The Winter's tale*. The choice has been made based on the different type of genres: in order, we choose two tragedies, a historical play and two comedies. All the implementations and algorithms were performed using MATLAB.

3.1 Drama as graphs

Depending on what kind of analysis we are interested in, there are many ways to represent the interactions between the characters in a story.

3.1.1 Undirected case

The first natural representation is an undirected weighted graph

$$G_1 = (V, \mathcal{E}_1, w_1)$$

where the nodes in V are the characters and $(i,j) \in \mathcal{E}_1 \iff (j,i) \in \mathcal{E}_1 \iff$ i and j talk to each other at least once during the entire play. The edge

weights can be the amount of times that i talks *with* j. The associated unweighted graph $G_1^A = (V, \mathcal{E}_1)$ is simply the graph that gives us an overview about the direct verbal interactions between the characters.

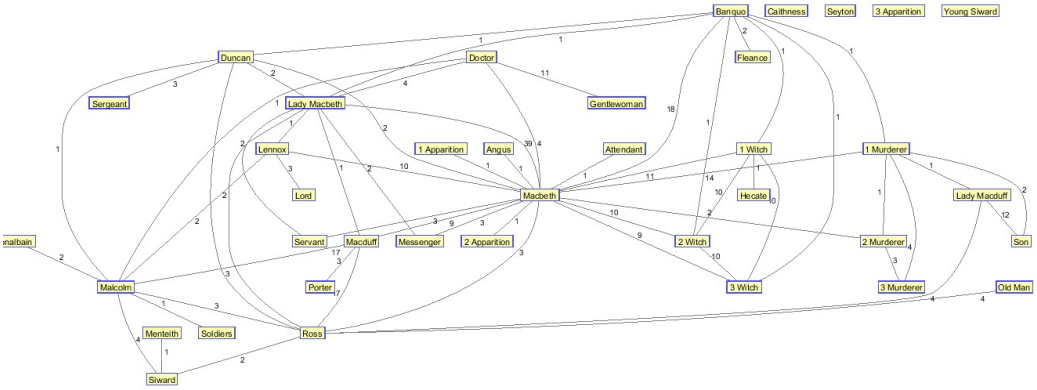


Figure 3.1: Relation among characters in *Macbeth*

3.1.2 Directed case

Even if the previous model can give perfectly reasonable results (we will see it), a complex plot cannot certainly be resumed just by how many times people talk to each other. In fact we can also consider the weighted graph

$$G_2 = (V, \mathcal{E}_2, w_2)$$

where $(i, j) \in \mathcal{E}_2 \iff i$ talks *about* j. The weight of (i, j) is, again, how many times i talks about j. Contrary of what we said about G_1, here if i talks about j, there is no guarantee that j talks about i, so this graph is directed.

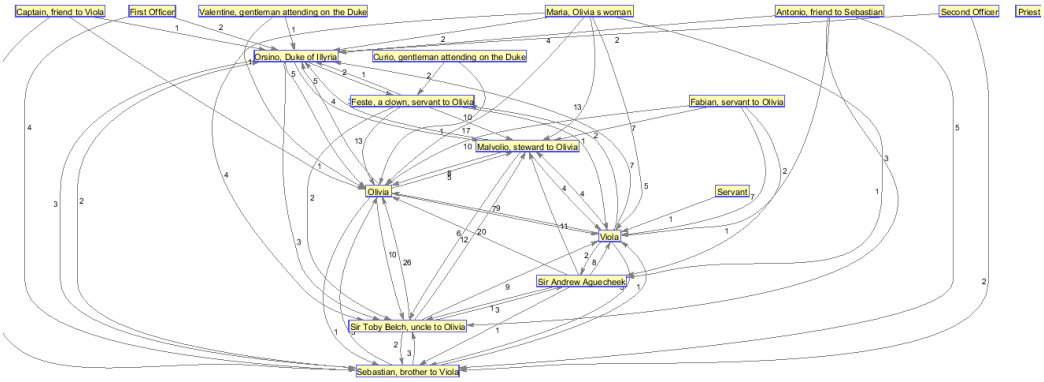


Figure 3.2: Relation among characters in *The Winter's Tale*

3.1.3 A different point of view and new centrality measures

We propose here a new idea to study the aspect of gossiping in a social circle. In addition to the directed graph G_2, we can do more. Let i be a node such that $\mathbf{d}_i^{in} > 0$. This means that there are people that talk about i. Then, for such i we can define the graph $G^i = (V^i, \mathcal{E}^i, g^i)$ where V^i are the N_i nodes that can join i through one of the $\mathbf{d}_i^{in}$ edges, and $(j,k) \in \mathcal{E}^i \iff (k,j) \in \mathcal{E}^i \iff j$ and k talked about i. The weight g^i_{jk} is the amount of times that j and k talked about i. So for every node $i \in V$ we have a graph G^i that represents if and how the other characters are talking about i and we will refer to it as "j talks about i with k".

The goal is to define a centrality measure not *of the nodes* of every G^i as we have done so far, but *of the entire graph*. This is a change of perspective that allow us to better understand the qualitative aspects inside a graph and see more.

Weighted spectral gap centrality measure

Let G^i as before. We want to use Perron-Frobenius Theorem on the adjacency matrix $A^{(i)}$ of G^i but, in general, G^i can be disconnected.

So let G^i_+ be the largest connected component of G^i and $A^{(i)}_+$ the associated submatrix of $A^{(i)}_+$. Let λ_1 be the Perron root of $A^{(i)}_+$ and λ_2 the second largest eigenvalue of $A^{(i)}_+$.

As we studied, $\frac{|\lambda_2|}{\lambda_1}$ gives a measure on the speed of convergence of the powers of the matrix. Now, we also want to raise up in the ranking the nodes i such that G^i has a large number of nodes: if a lot of people talks about

i that surely means that i is more important. Thus, we define the *weighted spectral gap centrality measure* as:

$$\boldsymbol{\lambda}^{(i)} := N_i \frac{\lambda_1 - |\lambda_2|}{\lambda_1}$$

where we recall that N_i is the number of vertices in G^i.

Weighted mean time centrality measure

Let suppose that G^i is a clique. Then the informations between the nodes are spread in the fastest way possible. In fact, if we suppose that walking through an edge takes one second, in a clique the information arrives to every node in one second. We can generalize this concept in a generic graph calculating the mean time for the nodes in G^i_+ to join each others. In other words, we want to produce an information on the topology of the graph: if the nodes are "well positioned" between them then the gossip spreads better. While the spectral gap measures a global speed (longer walks), this is a similar concept but also based on shorter walks. So we expect that the rankings with the previous centrality will be similar but not equal. So we define the *weighted mean time centrality measure* as:

$$\mathbf{MT}_i := \frac{1}{N_i} \sum_{j,k} \frac{1}{N_i(N_i - 1)} d(j,k)$$

where $d(j,k)$ is the distance from j to k and N_i as before the number of nodes that talks about i.

To better understand this formula let's suppose G^{i_1} be a clique of 3 nodes and G^{i_2} be a clique of 10 nodes. If we calculate just the mean time, it is in both cases equals to 1 and this is not exactly what we want because we planned to reward the more complex network. This is why we add the coefficient $\frac{1}{N_i}$. Another remark is that, contrary to all the other centrality measures that we use, here having a low weighted mean time centrality score means being higher in ranking.

3.2 Most important characters: results

3.2.1 Undirected case

We are interested in interpreting the results on the undirected graph, which we will refer as "i talks to j".

The first goal is to detect the characters that are the most important, based on the different centrality measures discussed in Chapter 2, i.e. based on different particular aspects about the characters that we want to enlighten.

We exhibit the table of results, where the last column are the ranking mean of all centrality rankings:

Table 3.1: *Macbeth* centrality measures for i talks to j

	PageRank **p**	$\mathbf{R_p}$	Degree **d**	$\mathbf{R_d}$	Closeness **C**	$\mathbf{R_C}$	Betweenness **B**	$\mathbf{R_B}$	Eigenvector **E**	$\mathbf{R_E}$	**Mean**
Macbeth	0,133927324	1	18	1	0,015597473	1	294,6428571	1	0,101645065	1	1
Ross	0,061317306	4	8	3	0,012626526	2	104,1738095	2	0,054863624	4	3
Lady Macbeth	0,065078636	3	9	2	0,012626526	2	37,83333333	10	0,069198759	2	3,8
Banquo	0,057928791	5	8	3	0,011872704	4	50,8452381	6	0,060750971	3	4,2
Malcolm	0,065561301	2	8	3	0,010606282	10	87,33333333	3	0,039227997	7	5
Duncan	0,045603558	7	6	6	0,011872704	4	52,34047619	4	0,053021546	5	5,2
Macduff	0,039692641	8	5	8	0,011528567	6	40,32857143	7	0,043131326	6	7
1 Murderer	0,046178336	6	6	6	0,01074961	9	51,41666667	5	0,03460068	11	7,4
Doctor	0,033489939	10	4	10	0,01104821	7	40,32857143	7	0,034199562	12	9,2
Lennox	0,033489939	10	4	10	0,01104821	7	40,32857143	7	0,034199562	12	9,2
1 Witch	0,038714654	9	5	8	0,010330794	11	32	11	0,038795026	8	9,4
2 Witch	0,029705606	12	4	10	0,010198348	12	0	15	0,037951779	9	11,6
3 Witch	0,029705606	12	4	10	0,010198348	12	0	15	0,037951779	9	11,6
2 Murderer	0,02480777	16	3	14	0,010198348	12	11,41666667	14	0,02307467	20	15,2
Servant	0,016798676	19	2	17	0,009700868	15	0	15	0,027112779	14	16
Messenger	0,016798676	19	2	17	0,009700868	15	0	15	0,027112779	15	16,2
Siward	0,028313083	14	3	14	0,009249664	21	32	11	0,015318057	26	17,2
Lady Macduff	0,025029506	15	3	14	0,009249664	21	13,01190476	13	0,015458681	25	17,6
Angus	0,010656175	30	1	21	0,00958399	17	0	15	0,016130991	21	20,8
Attendant	0,010656175	30	1	21	0,00958399	17	0	15	0,016130991	21	20,8
1 Apparition	0,010656175	30	1	21	0,00958399	17	0	15	0,016130991	21	20,8
Son	0,017969062	17	2	17	0,008117053	24	0	15	0,007944381	31	20,8
2 Apparition	0,010656175	30	1	21	0,00958399	17	0	15	0,016130991	24	21,4
3 Murderer	0,017908057	18	2	17	0,007575916	30	0	15	0,009153032	28	21,6
Old Man	0,010850608	28	1	21	0,00837338	23	0	15	0,008706814	29	23,2
Sergeant	0,010799565	29	1	21	0,008035062	25	0	15	0,008414477	30	24
Porter	0,011088047	26	1	21	0,007875952	27	0	15	0,006844907	32	24,2
Fleance	0,010490143	34	1	21	0,008035062	25	0	15	0,009641131	27	24,4
Gentlewoman	0,011458957	22	1	21	0,007648761	28	0	15	0,005427443	36	24,4
Lord	0,011458957	22	1	21	0,007648761	28	0	15	0,005427443	36	24,4
Donalbain	0,011294723	24	1	21	0,00743431	31	0	15	0,006225452	33	24,8
Soldiers	0,011294723	24	1	21	0,00743431	31	0	15	0,006225452	34	25
Menteith	0,012362962	21	1	21	0,006741281	34	0	15	0,002430964	38	25,8
Hecate	0,010917107	27	1	21	0,0072979	33	0	15	0,00615674	35	26,2
Caithness	0,00433526	35	0	35	0	35	0	15	0,026315789	16	27,2
Seyton	0,00433526	35	0	35	0	35	0	15	0,026315789	16	27,2
3 Apparition	0,00433526	35	0	35	0	35	0	15	0,026315789	16	27,2
Young Siward	0,00433526	35	0	35	0	35	0	15	0,026315789	16	27,2

Table 3.2: *Richard III* centrality measures for i talks to j

	p	R_p	d	R_d	C	R_C	B	R_B	E	R_E	Mean
Richard III	0,123731874	1	29	1	0,011963724	1	555,3270563	1	0,076839585	1	1
Queen Elizabeth	0,063078145	3	17	2	0,01004362	3	148,3770563	4	0,061694523	2	2,8
Lord Hastings	0,066692062	2	17	2	0,01004362	3	146,7573593	5	0,060379863	3	3
Duke of Buckingham	0,062916722	4	16	4	0,010429914	2	202,7307359	2	0,058179749	4	3,2
Sir William Stanley	0,037227624	6	9	6	0,009459689	5	127,3556277	6	0,040772866	7	6
Duchess of York	0,042110323	5	11	5	0,008842753	7	54,89761905	7	0,038486987	9	6,6
Lord Rivers	0,03405516	7	9	6	0,009039258	6	23,31666667	10	0,043803946	5	6,8
Marquis of Dorset	0,026084831	9	7	8	0,008747669	8	1,581818182	15	0,04022931	8	9,6
Queen Margaret	0,026068663	10	7	8	0,008747669	8	1,498484848	16	0,040914047	6	9,6
Sir William Catesby	0,022169107	12	5	11	0,008654609	10	45	8	0,02803471	11	10,4
King Edward IV	0,022816446	11	6	10	0,008654609	10	0,166666667	17	0,036765995	10	11,6
Prince Edward	0,020937027	13	5	11	0,008135333	13	2,333333333	14	0,024772049	13	12,8
Messenger	0,019963734	16	5	11	0,008054785	14	8,657575758	13	0,027064714	12	13,2
Richmond	0,031195175	8	5	11	0,007329128	23	173	3	0,011051555	28	14,6
Lady Anne	0,016530741	21	4	15	0,008217508	12	0	18	0,023473405	14	16
First Murderer	0,02004299	14	4	15	0,007463608	19	22	11	0,010847312	29	17,6
George Plantagenet	0,02004299	14	4	15	0,007463608	19	22	11	0,010847312	30	17,8
John Morton	0,013281133	25	3	18	0,007975816	15	0	18	0,021059742	15	18,2
Sir Richard Ratcliff	0,013153683	26	3	18	0,007603115	18	0	18	0,019510346	16	19,2
Lord Mayor of London	0,013508568	23	3	18	0,007822435	16	0	18	0,017222001	22	19,4
Richard Plantagenet	0,013508568	23	3	18	0,007822435	16	0	18	0,017222001	22	19,4
Sir Robert Brackenbury	0,015127773	22	3	18	0,007395757	22	0	18	0,010619824	34	22,8
Lord Lovel	0,009938416	30	2	24	0,007463608	19	0	18	0,014789243	24	23
Thomas Rotherham	0,012781574	27	3	18	0,006560752	34	0	18	0,013714348	25	24,4
First Citizen	0,019880716	17	2	24	0,000768935	48	0	18	0,019230769	17	24,8
Third Citizen	0,019880716	17	2	24	0,000768935	48	0	18	0,019230769	17	24,8
Second Citizen	0,019880716	17	2	24	0,000768935	48	0	18	0,019230769	21	25,6
Cardinal Bourchier	0,009657068	31	2	24	0,006953276	32	0	18	0,012778122	26	26,2
Lord Grey	0,009349765	35	2	24	0,006560752	34	0	18	0,011370418	27	27,6
Boy	0,009386715	32	2	24	0,006508266	36	0	18	0,010797366	31	28,2
Duke of Norfolk	0,009386715	32	2	24	0,006508266	36	0	18	0,010797366	31	28,2
Girl	0,009386715	32	2	24	0,006508266	36	0	18	0,010797366	31	28,2
Blunt	0,016936511	20	2	24	0,005282683	44	45	8	0,001205114	49	29
Third Messenger	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	35	30,2
Earl of Surrey	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	36	30,4
Second Messenger	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	36	30,4
Sir James Tyrrel	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	36	30,4
Sir Thomas Vaughan	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	36	30,4
Fourth Messenger	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	40	31,2
Gentleman	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	40	31,2
Page	0,006606172	39	1	35	0,007199409	24	0	18	0,00828162	40	31,2
Second Murderer	0,011503709	28	2	24	0,005317211	43	0	18	0,002338204	48	32,2
Sheriff of Wiltshire	0,006324824	48	1	35	0,006614092	33	0	18	0,006270499	45	35,8
Another	0,006752944	38	1	35	0,005852757	42	0	18	0,003021526	47	36
Sir Walter Herbert	0,010217775	29	1	35	0,004088107	47	0	18	0,000129885	52	36,2
Earl Oxford	0,00827371	36	1	35	0,005214957	45	0	18	0,001191115	50	36,8
Lords	0,00827371	36	1	35	0,005214957	45	0	18	0,001191115	50	36,8
Priest	0,006314352	49	1	35	0,006456613	39	0	18	0,006507623	43	36,8
Pursuivant	0,006314352	49	1	35	0,006456613	39	0	18	0,006507623	43	36,8
Christopher Urswick	0,006502068	47	1	35	0,006210178	41	0	18	0,004394419	46	37,4
Henry VI	0,002982107	51	0	51	0	51	0	18	0,019230769	17	37,6
Scrivener	0,002982107	51	0	51	0	51	0	18	0,019230769	17	37,6

Table 3.3: *Romeo and Juliet* centrality measures for i talks to j

	p	**R_p**	**d**	**R_d**	**C**	**R_C**	**B**	**R_B**	**E**	**R_E**	**Mean**
Capulet	0,090144695	1	12	1	0,0153125	1	128,0595238	1	0,092785388	1	1
Romeo	0,083297943	2	11	2	0,014723558	2	102,3928571	3	0,089584034	2	2,2
Nurse	0,061471074	3	8	3	0,013671875	4	103,9285714	2	0,071396106	4	3,2
Friar Laurence	0,060072019	4	8	3	0,014178241	3	53,6547619	6	0,076882839	3	3,8
Prince Escalus	0,050146878	6	6	5	0,01255123	6	87,30952381	4	0,041810136	8	5,8
Benvolio	0,052375217	5	6	5	0,011427239	10	56,57142857	5	0,034342196	12	7,4
Juliet	0,043552357	9	6	5	0,012976695	5	5,369047619	13	0,069809306	5	7,4
Paris	0,037771203	11	5	8	0,011962891	7	14,16666667	10	0,056616814	6	8,4
Tybalt	0,031660898	13	4	9	0,011778846	9	6,404761905	12	0,041924488	7	10
Mercutio	0,031781364	12	4	9	0,011259191	12	9,666666667	11	0,038910728	9	10,6
Montague	0,025574909	14	3	12	0,011962891	7	18,57142857	9	0,027707388	15	11,4
Peter	0,043555039	8	4	9	0,009815705	17	53	7	0,014826597	22	12,6
Balthasar	0,024589138	15	3	12	0,011427239	10	4,571428571	14	0,034159404	13	12,8
Lady Capulet	0,023752093	18	3	12	0,0109375	13	0	17	0,03837671	10	14
First Watchman	0,03800949	10	3	12	0,009114583	23	53	7	0,007247142	27	15,8
Sampson	0,045441072	7	2	16	0,001953125	30	1	16	0,037655778	11	16
First Servant	0,019218911	20	2	16	0,010074013	14	0	17	0,018203177	18	17
Second Servant	0,019218911	20	2	16	0,010074013	14	0	17	0,018203177	19	17,2
Page	0,018190952	22	2	16	0,00945216	21	1,333333333	15	0,016142953	20	18,8
First Musician	0,020452362	19	2	16	0,009570313	20	0	17	0,014141341	25	19,4
Abraham	0,023935763	16	1	21	0,001302083	31	0	17	0,026626656	16	20,2
Gregory	0,023935763	16	1	21	0,001302083	31	0	17	0,026626656	17	20,4
Second Capulet	0,011050873	31	1	21	0,009943182	16	0	17	0,015217683	21	21,2
Apothecary	0,01110229	29	1	21	0,009691456	18	0	17	0,014692631	23	21,6
Servant	0,01110229	29	1	21	0,009691456	18	0	17	0,014692631	24	21,8
First Citizen	0,01208546	27	1	21	0,008144947	24	0	17	0,005632446	28	23,4
Friar John	0,011048272	32	1	21	0,00945216	21	0	17	0,012609515	26	23,4
Lady Montague	0,01208546	27	1	21	0,008144947	24	0	17	0,005632446	28	23,4
Second Musician	0,013921057	25	1	21	0,007291667	26	0	17	0,002431702	30	23,8
Third Musician	0,013921057	25	1	21	0,007291667	26	0	17	0,002431702	31	24
Third Watchman	0,01543478	23	1	21	0,006897523	28	0	17	0,0011886	32	24,2
Second Watchman	0,01543478	23	1	21	0,006897523	28	0	17	0,0011886	33	24,4
Chorus	0,00466563	33	0	33	0	33	0	17	0,03030303	14	26

Table 3.4: *Winter's Tale* centrality measures for i talks to j

	p	**R_p**	**d**	**R_d**	**C**	**R_C**	**B**	**R_B**	**E**	**R_E**	**Mean**
Leontes	0,10122598	1	15	1	0,017379196	1	197,3239538	1	0,095295648	1	1
Paulina	0,094802635	2	13	2	0,014896453	3	129,5214286	2	0,077985722	3	2,4
Polixenes	0,055608389	3	9	3	0,015136719	2	39,12279942	8	0,082759316	2	3,6
Florizel	0,044900863	5	7	4	0,014663696	4	86,64502165	4	0,06249824	6	4,6
Camillo	0,046686224	4	7	4	0,014438101	6	40,16060606	7	0,067281415	5	5,2
Perdita	0,043795192	7	7	4	0,014663696	4	25,63946609	11	0,071613596	4	6
Old Shepherd	0,043668253	8	7	4	0,012855843	9	24,26709957	12	0,058901605	7	8
Clown	0,03909533	9	6	8	0,012348376	11	35,69105339	9	0,043745495	9	9,2
Autolycus	0,043962655	6	6	8	0,011730957	17	112,984632	3	0,027307191	15	9,8
Servant	0,032663499	11	5	10	0,01321798	7	13,75873016	13	0,046555275	8	9,8
Antigonus	0,032624501	12	4	11	0,012348376	11	30	10	0,031483165	12	11,2
Hermione	0,029135549	15	4	11	0,013034397	8	10,35	14	0,040614021	10	11,6
Mamillius	0,034302609	10	4	11	0,01187945	15	58	5	0,020982506	18	11,8
First Lord	0,029079998	16	4	11	0,012855843	9	3,424603175	16	0,03684568	11	12,6
Cleomenes	0,031277948	13	4	11	0,012188007	14	6,433333333	15	0,030437551	13	13,2
First Gentleman	0,030740102	14	3	16	0,008853552	26	58	5	0,004215839	29	18
Lord	0,015848313	26	2	18	0,012348376	11	0	18	0,023133963	17	18
Gentleman	0,016591223	25	2	18	0,01187945	15	0	18	0,025404563	16	18,4
Officer	0,017046449	24	2	18	0,011444836	18	0	18	0,018433586	19	19,4
Dion	0,017505247	23	2	18	0,010312929	20	0	18	0,015895799	20	19,8
Mopsa	0,02394047	17	3	16	0,009576291	25	2,677272727	17	0,011245603	25	20
First Lady	0,020813883	20	2	18	0,008689598	28	0	18	0,003604699	30	22,8
Dorcas	0,017676492	22	2	18	0,008609877	30	0	18	0,005652177	27	23
Second Lady	0,020813883	20	2	18	0,008689598	28	0	18	0,003604699	31	23
Emilia	0,010857825	28	1	27	0,010091146	21	0	18	0,011433388	22	23,2
First Servant	0,010399028	31	1	27	0,01117234	19	0	18	0,013971175	21	23,2
Gaoler	0,010857825	28	1	27	0,010091146	21	0	18	0,011433388	22	23,2
Second Gentleman	0,023309215	18	2	18	0,006951678	31	0	18	0,000724262	32	23,4
Second Servant	0,010857825	28	1	27	0,010091146	21	0	18	0,011433388	24	23,6
Third Gentleman	0,023309215	18	2	18	0,006951678	31	0	18	0,000724262	33	23,6
Mariner	0,011602913	27	1	27	0,008853552	26	0	18	0,004615707	28	25,2
Archidamus	0,010334837	32	1	27	0,009878701	24	0	18	0,009864043	26	25,4
Time	0,00466563	33	0	33	0	33	0	18	0,03030303	14	26,2

Table 3.5: *Twelfth Night* centrality measures for i talks to j

	p	R_p	d	R_d	C	R_C	B	R_B	E	R_E	Mean
Viola	0,106167894	1	11	1	0,043478261	1	31	2	0,098233636	2	1,4
Olivia	0,100030155	2	10	2	0,04	2	33,55555556	1	0,092249718	5	2,4
Sir Andrew Aguecheek	0,090066425	4	10	2	0,04	2	10,48888889	5	0,100181977	1	2,8
Orsino	0,092707274	3	9	4	0,04	2	27,83333333	3	0,071798725	8	4
Sir Toby Belch	0,080964474	5	9	4	0,038461538	5	8,777777778	6	0,094126105	4	4,8
Feste	0,080458202	6	9	4	0,038461538	5	3,988888889	7	0,096982729	3	5
Malvolio	0,063444066	8	7	7	0,034482759	7	0,444444444	9	0,081375831	6	7,4
Fabian	0,063444066	8	7	7	0,034482759	7	0,444444444	9	0,081375831	7	7,6
Antonio	0,065788527	7	6	9	0,033333333	9	17,83333333	4	0,046709236	11	8
Sebastian	0,05750868	10	6	9	0,033333333	9	1,233333333	8	0,065275336	10	9,2
Maria	0,05513152	11	6	9	0,03125	11	0	12	0,070191204	9	10,4
First Officer	0,034057052	12	3	12	0,029411765	12	0,4	11	0,028098718	12	11,8
Valentine	0,025296711	13	2	13	0,027777778	13	0	12	0,021846873	13	12,8
Captain	0,016539465	18	1	14	0,025641026	14	0	12	0,012621702	14	14,4
Servant	0,016828933	16	1	14	0,024390244	15	0	12	0,011852849	15	14,4
Curio	0,01709058	15	1	14	0,024390244	15	0	12	0,009225171	17	14,6
Priest	0,016828933	16	1	14	0,024390244	15	0	12	0,011852849	16	14,6
Second Officer	0,017647043	14	1	14	0,02173913	18	0	12	0,006001509	18	15,2

The first self-evident thing to remark is that we can do three divisions of the characters. In fact, after a first chaotic behavior of the mean ranking, we can see that it grows more and more slowly and tends to stabilize. There are three sections in all the plays: the apical characters, the middle roles and the rest. Now, the larger values of "importance" are concentrate to a particular position in the ranking and this happens in all the five plays, even when the number of characters visibly changes: this specific social circle is formed by the nodes that are at most at the 6-th or 7-th place in the ranking list, coherently with a qualitative analysis of the plots. Then, we have a group of nodes (relating to the moment when the ranking mean is about to stabilize) and the rest formed by the nodes that are basically read by the model as having completely irrelevant roles. This is true because they play absolutely no active role on the development of the facts in the story. So it is plausible to concentrate our analysis to the group of characters for which the ranking still has an acceptable "rapid" variation.

PageRank centrality is, as we expect, the more precise in terms of a quantitative evaluation on the principal players. To have a more qualitative sense on "why" they are important we need a less general graph as the one in the next section. The general overview detects also the most influential conflicts: Malcolm and Macbeth, Capulet and Romeo, Richard and Elizabeth, Paulina and Leontes. We can also see that, immediately after the predictable primacy of the main characters, the mediators/intermediaries of the plays are found by betweenness centrality: Duke of Buckingham and Hastings in *Richard III*,

Friar Laurence and the nurse in *Romeo and Juliet*, Viola in *Twelfth Night*.

These are characters that are high in betweenness because they share a lot of informations with the rest of the network and they contribute actively in this way to the development of the story, at least in a quantitative way.

For the other two works, the highest in betweenness are the one that are in the middle of a lot of shortest paths because they are in the entourage of the apical nodes, thus indirectly they receive a lot of important informations. This is also confirmed by their values of closeness centrality.

Another interesting way of seeing this data set is by comparison: if a node is high for PageRank and low for betweenness that says something about the function of the character on the story. This is for example the case of Juliet, lover of the main character but not so active (again, in a quantitative way) in the story and Perdita because she is important but she appears only in the second half of the play. Same for Feste and Malvolio comparing degree and betweenness, or Sir Toby and Sir Andrew: tormentor and tormented shares the aspect of talking with people but they do not spread informations in order to develop the story: for example the comic roles (*Shakespeare fools*).

3.2.2 Directed case and new point of view

We study now the case "i talks about j". First of all, not all the characters are subject of gossip in the story or talk about somebody, so the network has less nodes than the previous one. Since the graph is directed some of the centralities that we studied are no longer suitable. We focused our attention on eigenvector centrality and PageRank.

We tested the ranking behavior of PageRank centrality on the graph "i talks about j" depending on α. Here we show some of the tested values with the relative ranking. The first column is the ranking of the matrix $H\mathbf{1}$ seen in 2.1. For ease of exposition we ignored some irrelevant nodes.

Table 3.6: *Macebth*

	H	$\alpha = 0.01$	$\alpha = 0.90$	$\alpha = 0.85$	$\alpha = 0.75$	$\alpha = 0.65$	$\alpha = 0.55$
Angus	17	17	17	17	17	17	17
Attendant	17	17	17	17	17	17	17
Banquo	2	1	4	4	3	3	2
Caithness	17	17	17	17	17	17	17
Doctor	17	17	17	17	17	17	17
Donalbain	10	8	9	9	10	10	10
Duncan	1	3	1	1	1	1	1
1 Apparition	17	17	17	17	17	17	17
1 Murderer	17	17	17	17	17	17	17
1 Witch	14	14	12	12	12	13	13
Fleance	11	9	11	11	11	11	11
Gentlewoman	17	17	17	17	17	17	17
Hecate	17	17	17	17	17	17	17
Lady Macbeth	5	5	8	8	8	6	6
Lady Macduff	8	11	6	6	6	7	7
Lennox	17	17	17	17	17	17	17
Lord	17	17	17	17	17	17	17
Macbeth	3	2	2	2	2	2	3
Macduff	4	4	3	3	4	4	4
Malcolm	6	6	5	5	5	5	5
Menteith	17	17	17	17	17	17	17
Messenger	17	17	17	17	17	17	17
Old Man	17	17	17	17	17	17	17
Porter	17	17	17	17	17	17	17
Ross	17	17	17	17	17	17	17
2 Apparition	17	17	17	17	17	17	17
2 Murderer	17	17	17	17	17	17	17
2 Witch	14	14	12	12	12	13	13
Sergeant	17	17	17	17	17	17	17
Servant	17	17	17	17	17	17	17
Seyton	13	13	15	15	15	12	12
Siward	12	10	16	16	16	16	16
Soldiers	17	17	17	17	17	17	17
Son	8	11	6	6	6	7	7
3 Apparition	17	17	17	17	17	17	17
3 Murderer	17	17	17	17	17	17	17
3 Witch	14	14	12	12	12	13	13
Young Siward	7	7	10	10	9	9	9

Table 3.7: *Richard III*

	H	$\alpha = 0.01$	$\alpha = 0.90$	$\alpha = 0.85$	$\alpha = 0.75$	$\alpha = 0.65$	$\alpha = 0.55$
Another	28	28	28	28	28	28	28
Blunt	20	19	27	27	26	26	24
Boy	22	25	19	20	22	22	22
Cardinal Bourchier	28	28	28	28	28	28	28
Christopher Urswick	28	28	28	28	28	28	28
Duchess of York	17	21	14	14	14	15	17
Duke of Buckingham	9	7	9	8	8	7	7
Duke of Norfolk	14	12	7	9	9	9	9
Earl of Surrey	28	28	28	28	28	28	28
Earl Oxford	28	28	28	28	28	28	28
First Citizen	28	28	28	28	28	28	28
First Murderer	28	28	28	28	28	28	28
Fourth Messenger	28	28	28	28	28	28	28
Gentleman	28	28	28	28	28	28	28
George Plantagenet	3	4	3	4	4	4	4
Girl	26	26	22	23	23	24	26
Henry VI	10	11	12	11	12	12	12
John Morton	28	28	28	28	28	28	28
King Edward IV	2	3	2	2	3	3	3
Lady Anne	23	22	13	15	16	18	18
Lord Grey	15	16	16	16	17	16	15
Lord Hastings	6	6	6	6	6	6	6
Lord Lovel	21	20	26	25	25	23	23
Lord Mayor of London	27	27	25	26	27	27	27
Lord Rivers	12	15	11	12	13	13	13
Lords	28	28	28	28	28	28	28
Marquis of Dorset	19	18	21	19	19	19	20
Messenger	28	28	28	28	28	28	28
Page	28	28	28	28	28	28	28
Priest	28	28	28	28	28	28	28
Prince Edward	4	2	4	3	2	2	2
Pursuivant	28	28	28	28	28	28	28
Queen Elizabeth	5	5	5	5	5	5	5
Queen Margaret	8	10	10	10	10	10	10
Richard III	1	1	1	1	1	1	1
Richard Plantagenet	7	8	8	7	7	8	8
Richmond	11	13	18	18	15	14	14
Scrivener	28	28	28	28	28	28	28
Second Citizen	28	28	28	28	28	28	28
Second Messenger	28	28	28	28	28	28	28
Second Murderer	28	28	28	28	28	28	28
Sheriff of Wiltshire	28	28	28	28	28	28	28
Sir James Tyrrel	18	14	23	22	20	20	19
Sir Richard Ratcliff	28	28	28	28	28	28	28
Sir Robert Brackenbury	25	24	24	24	24	25	25
Sir Thomas Vaughan	15	16	16	16	17	16	15
Sir Walter Herbert	28	28	28	28	28	28	28
Sir William Catesby	24	23	20	21	21	21	21
Sir William Stanley	13	9	15	13	11	11	11
Third Citizen	28	28	28	28	28	28	28
Third Messenger	28	28	28	28	28	28	28
Thomas Rotherham	28	28	28	28	28	28	28

Table 3.8: *Romeo and Juliet*

	H	$\alpha = 0.01$	$\alpha = 0.90$	$\alpha = 0.85$	$\alpha = 0.75$	$\alpha = 0.65$	$\alpha = 0.55$
Abraham	7	7	7	7	7	7	7
Apothecary	7	7	7	7	7	7	7
Balthasar	7	7	7	7	7	7	7
Benvolio	7	7	7	7	7	7	7
Capulet	6	5	5	5	5	5	5
Chorus	7	7	7	7	7	7	7
First Citizen	7	7	7	7	7	7	7
First Musician	7	7	7	7	7	7	7
First Servant	7	7	7	7	7	7	7
First Watchman	7	7	7	7	7	7	7
Friar John	7	7	7	7	7	7	7
Friar Laurence	7	7	7	7	7	7	7
Gregory	7	7	7	7	7	7	7
Juliet	2	4	4	4	4	4	4
Lady Capulet	7	7	7	7	7	7	7
Lady Montague	7	7	7	7	7	7	7
Mercutio	5	6	6	6	6	6	6
Montague	7	7	7	7	7	7	7
Nurse	7	7	7	7	7	7	7
Page	7	7	7	7	7	7	7
Paris	4	3	3	3	3	3	3
Peter	7	7	7	7	7	7	7
Prince Escalus	7	7	7	7	7	7	7
Romeo	1	1	1	1	1	1	1
Sampson	7	7	7	7	7	7	7
Second Capulet	7	7	7	7	7	7	7
Second Musician	7	7	7	7	7	7	7
Second Servant	7	7	7	7	7	7	7
Second Watchman	7	7	7	7	7	7	7
Servant	7	7	7	7	7	7	7
Third Musician	7	7	7	7	7	7	7
Third Watchman	7	7	7	7	7	7	7
Tybalt	3	2	2	2	2	2	2

Table 3.9: *Twelfth Night*

	H	$\alpha = 0.01$	$\alpha = 0.90$	$\alpha = 0.85$	$\alpha = 0.75$	$\alpha = 0.65$	$\alpha = 0.55$
Antonio	9	9	9	9	9	9	9
Captain	9	9	9	9	9	9	9
Curio	9	9	9	9	9	9	9
Fabian	9	9	9	9	9	9	9
Feste	7	7	8	7	7	7	7
First Officer	9	9	9	9	9	9	9
Malvolio	2	6	5	5	6	6	6
Maria	9	9	9	9	9	9	9
Olivia	1	1	1	1	1	1	2
Orsino	4	2	4	3	3	2	1
Priest	9	9	9	9	9	9	9
Sebastian	6	4	6	6	5	4	4
Second Officer	9	9	9	9	9	9	9
Servant	9	9	9	9	9	9	9
Sir Andrew Aguecheek	8	8	7	8	8	8	8
Sir Toby Belch	5	5	3	4	4	5	5
Valentine	9	9	9	9	9	9	9
Viola	3	3	2	2	2	3	3

Table 3.10: *Winter's Tale*

	H	$\alpha = 0.01$	$\alpha = 0.90$	$\alpha = 0.85$	$\alpha = 0.75$	$\alpha = 0.65$	$\alpha = 0.55$
Antigonus	9	9	9	9	9	9	9
Archidamus	10	10	10	10	10	10	10
Autolycus	10	10	10	10	10	10	10
Camillo	2	3	3	3	3	3	3
Cleomenes	10	10	10	10	10	10	10
Clown	10	10	10	10	10	10	10
Dion	10	10	10	10	10	10	10
Dorcas	10	10	10	10	10	10	10
Emilia	10	10	10	10	10	10	10
First Gentleman	10	10	10	10	10	10	10
First Lady	10	10	10	10	10	10	10
First Lord	10	10	10	10	10	10	10
First Servant	10	10	10	10	10	10	10
Florizel	7	7	6	6	6	6	7
Gaoler	10	10	10	10	10	10	10
Gentleman	10	10	10	10	10	10	10
Hermione	4	4	5	5	5	5	4
Leontes	5	2	4	4	4	4	2
Lord	10	10	10	10	10	10	10
Mamillius	6	6	7	7	7	7	6
Mariner	10	10	10	10	10	10	10
Mopsa	10	10	10	10	10	10	10
Officer	10	10	10	10	10	10	10
Old Shepherd	10	10	10	10	10	10	10
Paulina	8	8	8	8	8	8	8
Perdita	1	1	1	1	1	1	1
Polixenes	3	5	2	2	2	2	5
Second Gentleman	10	10	10	10	10	10	10
Second Lady	10	10	10	10	10	10	10
Second Servant	10	10	10	10	10	10	10
Servant	10	10	10	10	10	10	10
Third Gentleman	10	10	10	10	10	10	10
Time	10	10	10	10	10	10	10

We immediately observe the result of theorem 2.6.6 comparing the first column with the column in which $\alpha = 0.01$.

As we said in the last chapter, the PageRank ranking behaviour near the endpoints of $(0, \tau)$ is not stable since it tends to ranking vectors that are far from doing PageRank's job. So it is clever to choose the parameter to use in the applications in a specific range far from the endpoints. Testing different values, we find that the best choice is 0.75.

We can now show the centrality measures results. Note that we have a less number of nodes because only smaller circles are subject of gossip.

Table 3.11: *Macbeth* centrality measures for i talks about j

	λ	**R_λ**	**MT**	**R_{MT}**	**E**	**R_E**	**p**	**R_p**	**Mean**
Duncan	5,733056745	2	0,181818182	2	0,705961583	1	0,155923389	1	1,5
Banquo	6,561609596	1	0,14599686	1	0,356194745	3	0,111311194	3	2
Lady Macduff	3	3	0,333333333	5	0,06007084	6	0,038153152	6	5
Son	3	3	0,333333333	5	0,06007084	7	0,038153152	6	5,25
Macbeth	1,366835021	15	0,326530612	4	0,552332337	2	0,117027467	2	5,75
Macduff	1,19880507	16	0,3	3	0,192348791	4	0,099870846	4	6,75
Young Siward	3	3	0,333333333	5	0,021585828	10	0,024682183	9	6,75
Malcolm	1,757359313	13	0,444444444	8	0,147457886	5	0,074491439	5	7,75
Lady Macbeth	2	6	0,5	11	0,030971284	9	0,037413485	8	8,5
Fleance	1,757359313	12	0,444444444	8	0,032356517	8	0,021503849	11	9,75
2 Witch	2	6	0,5	11	0,017059826	11	0,018621177	12	10
3 Witch	2	6	0,5	11	0,017059826	12	0,018621177	12	10,25
1 Witch	2	6	0,5	11	0,017059826	13	0,018621177	12	10,5
Donalbain	1,757359313	13	0,444444444	8	0,01159416	15	0,024310057	10	11,5
Seyton	2	6	0,5	11	0,015296691	14	0,01803794	15	11,5
Siward	2	6	0,5	11	0	16	0,011697486	16	12,25

Table 3.12: *Romeo and Juliet* centrality measures for i talks about j

	λ	**R_λ**	**MT**	**R_{MT}**	**E**	**R_E**	**p**	**R_p**	**Mean**
Romeo	7,565966646	1	0,119791667	1	0,647886633	1	0,230391854	1	1
Tybalt	4,938287501	3	0,139717425	2	0,346820581	3	0,147607781	2	2,5
Juliet	5,279020243	2	0,148888889	3	0,580672094	2	0,104997507	4	2,75
Paris	3,401639914	4	0,177777778	4	0,340373636	4	0,11052257	3	3,75
Capulet	2	6	0,5	6	0,083211187	5	0,085029086	5	5,5
Mercutio	2,345650413	5	0,333333333	5	0	6	0,021396766	6	5,5

Table 3.13: *Richard III* centrality measures for i talks about j

	λ	R_λ	MT	R_{MT}	E	R_E	p	R_p	Mean
King Edward IV	7,626577881	2	0,119113573	1	0,520600804	2	0,069391048	3	2
George Plantagenet	7,631855148	1	0,123958333	2	0,640164456	1	0,066927165	4	2
Richard III	4,936609956	3	0,135380623	3	0,372538097	3	0,074598092	1	2,5
Queen Elizabeth	4,570552438	5	0,182098765	6	0,202851518	4	0,052178133	5	5
Prince Edward	4,857240484	4	0,2	9	0,197920048	5	0,069826067	2	5
Lord Hastings	3,792024891	9	0,171900826	4	0,156163127	6	0,042357439	6	6,25
Duke of Buckingham	4,024027754	7	0,179292929	5	0,113131959	8	0,038734544	8	7
Richard Plantagenet	3,061287651	12	0,183471074	7	0,096001495	10	0,038792354	7	9
Queen Margaret	3,009875941	13	0,188271605	8	0,119302622	7	0,03490701	10	9,5
Lord Rivers	4,11684397	6	0,233333333	10	0,06150932	14	0,025319734	13	10,75
Duchess of York	4	8	0,25	11	0,07200923	12	0,023030245	14	11,25
Duke of Norfolk	3	14	0,333333333	16	0,102217231	9	0,037176923	9	12
Henry VI	2,99096265	15	0,254464286	13	0,067179673	13	0,026545029	12	13,25
Lord Grey	3,75	10	0,26	14	0,044338323	16	0,021684716	17	14,25
Sir Thomas Vaughan	3,75	10	0,26	14	0,044338323	17	0,021684716	17	14,5
Lady Anne	2,535898385	16	0,375	18	0,080565117	11	0,022491811	16	15,25
Richmond	1,674341674	27	0,25	11	0,056573865	15	0,022578979	15	17
Marquis of Dorset	2,535898385	16	0,375	18	0,017470769	20	0,015035055	19	18,25
Sir William Stanley	1,757359313	25	0,444444444	20	0,028332258	18	0,026605713	11	18,5
Boy	2,345650413	18	0,333333333	16	0,017952196	19	0,013785886	22	18,75
Sir William Catesby	2	19	0,5	22	0,011251796	22	0,013953694	21	21
Sir James Tyrrel	2	19	0,5	22	0,003314157	25	0,014242572	20	21,5
Sir Robert Brackenbury	2	19	0,5	22	0,013178051	21	0,012008233	24	21,5
Girl	1,757359313	25	0,444444444	20	0,009480421	23	0,012457087	23	22,75
Lord Mayor of London	2	19	0,5	22	0,005128798	24	0,010667504	27	23
Lord Lovel	2	19	0,5	22	0,003314157	26	0,011690399	25	23
Blunt	2	19	0,5	22	0,001199749	27	0,011173793	26	23,5

Table 3.14: *Twelfth Night* centrality measures for i talks about j

	λ	R_λ	MT	R_{MT}	E	R_E	p	R_p	Mean
Olivia	7,124729881	1	0,152777778	1	0,658338934	1	0,145175971	1	1
Sir Toby Belch	3,291965225	4	0,196428571	3	0,447977212	2	0,118386858	4	3,25
Viola	3,604814989	3	0,214285714	4	0,341317633	4	0,130598114	2	3,25
Malvolio	4,485083593	2	0,160714286	2	0,370502411	3	0,110269764	6	3,25
Orsino	3,004367023	5	0,25170068	5	0,28872775	5	0,12753815	3	4,5
Sebastian	2,973941195	7	0,25170068	5	0,151746309	6	0,114640307	5	5,75
Feste	3	6	0,333333333	7	0,057507118	7	0,055372399	7	6,75
Sir Andrew Aguecheek	2,345650413	8	0,333333333	7	0,049383986	8	0,053090901	8	7,75

Table 3.15: *Winter's Tale* centrality measures for i talks about j

	λ	R_λ	MT	R_{MT}	E	R_E	p	R_p	Mean
Perdita	4,400372024	1	0,204444444	1	0,467616191	2	0,125414596	1	1,25
Polixenes	3,5	3	0,224489796	3	0,482588029	1	0,092965907	2	2,25
Camillo	4,128380477	2	0,205357143	2	0,396333706	4	0,089829634	3	2,75
Hermione	3,01081305	4	0,263392857	4	0,321305949	5	0,08599107	5	4,5
Leontes	2,192235936	7	0,28	5	0,458341979	3	0,08810464	4	4,75
Mamillius	2,800303232	6	0,3	6	0,065237657	7	0,046231987	7	6,5
Florizel	1,527864045	9	0,416666667	8	0,27118929	6	0,064505202	6	7,25
Antigonus	3	5	0,333333333	7	0	9	0,024012857	9	7,5
Paulina	1,757359313	8	0,444444444	9	0,015780403	8	0,036675994	8	8,25

The first interesting thing to remark are the first places on the mean ranking. The undirected case gave us a predictable answer: the first places are the protagonists. We might have expected that this method would confirm those first places, but this is not the case in some plays. This diversity can be seen as a more qualitative information on the function of the characters: if they mostly have a role of action on the story they will be high in i talks with j but not so high in i talks about j. This is the case of Richard and Elizabeth in *Richard III*, Lady Macbeth and Macbeth, Capulet in *Romeo and Juliet*, Viola in *Twelfth Night*, Leontes and Paulina in *The Winter's Tale*. In mean ranking they were high in the undirected case, which gives us a perspective on the active function on the play, but they are not the first characters subject to the gossips. We can also see that the characters i with high degree in "j talks about i with k are often the plotters of the story. Perdita, Edward, Duncan, Banquo, Malvolio and Tybalt are the victims that turn around the events, all high in this ranking, Olivia is the bone of contention: in other words they all are the subjects of the gossiping and these graphs allow us to see these details. Another interesting remark based on our results is that the subjects of the most amount of gossip in the comedies are women and in the tragedies they are men. This is for sure an aspect that can be confirmed with a qualitative check of the plays. In order to understand how j talks about i with k we see some graphs.

Figure 3.3: Macbeth

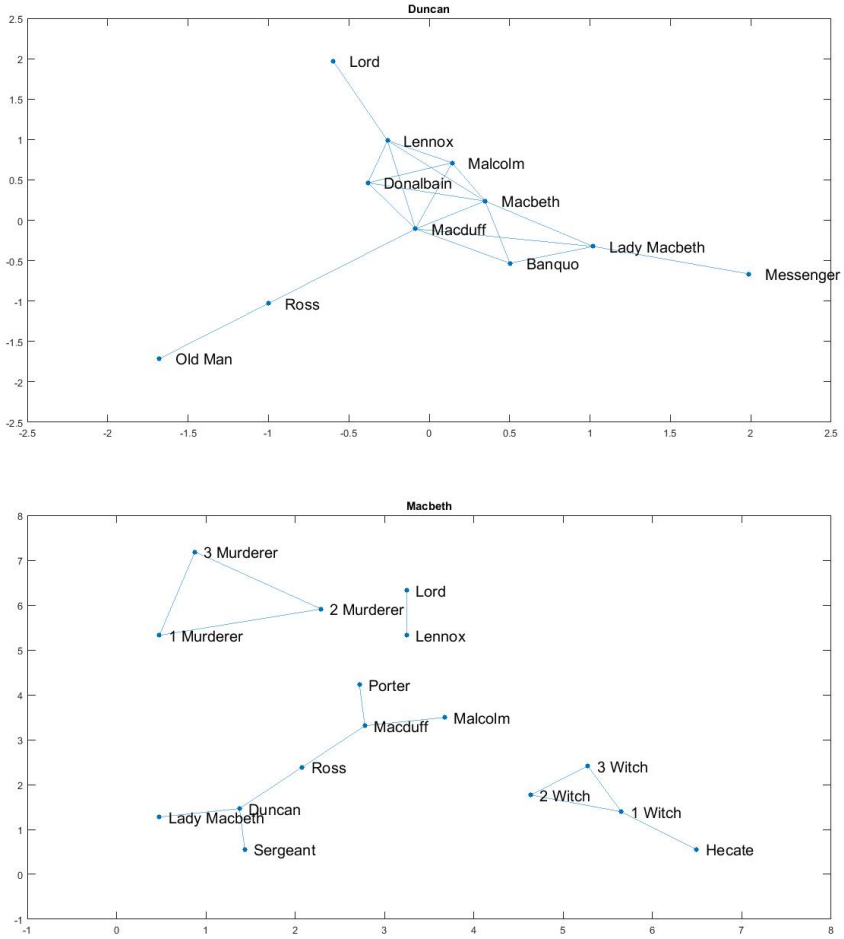


Figure 3.4: Richard III

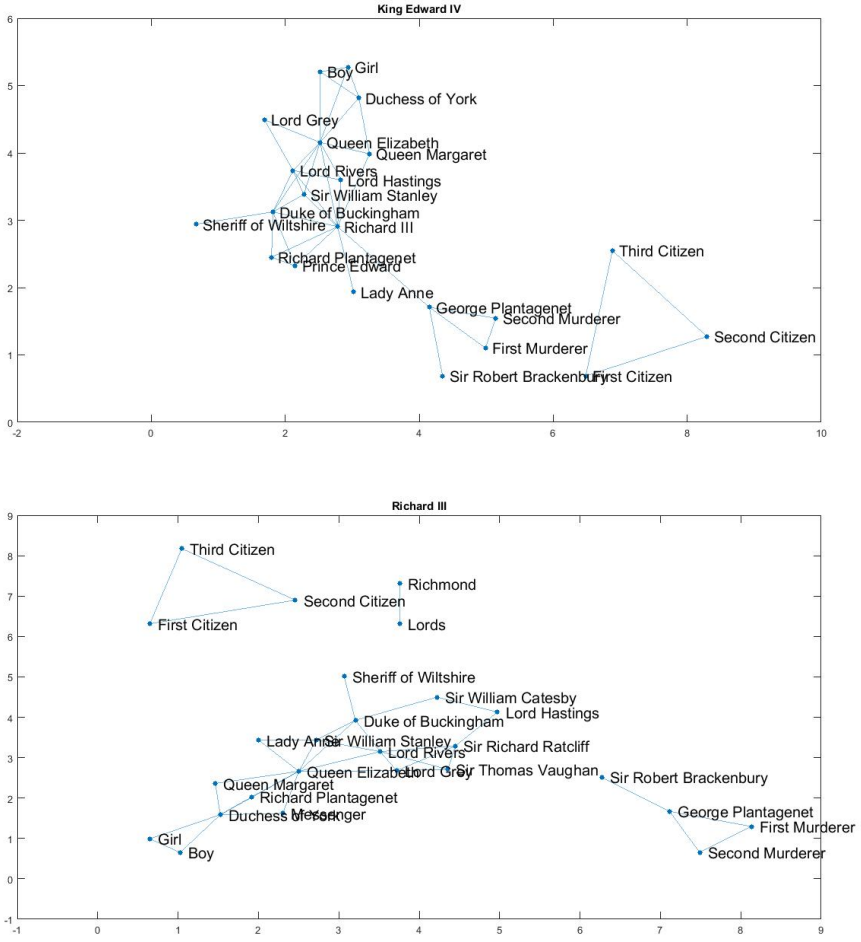


Figure 3.5: Richard III

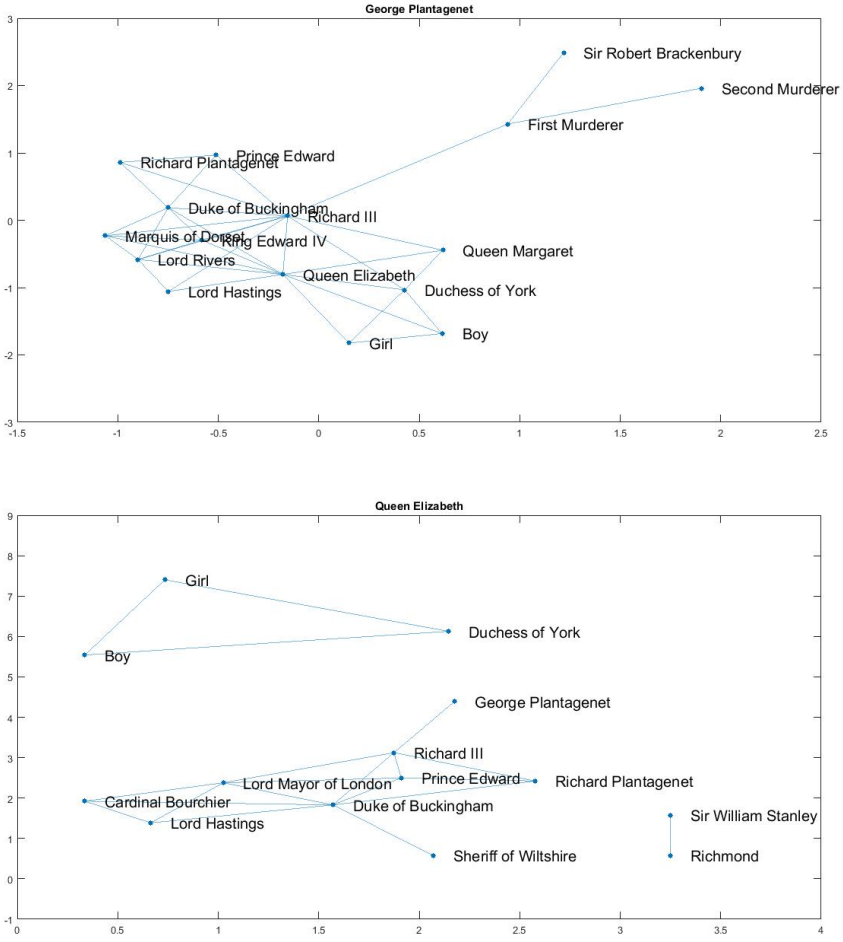
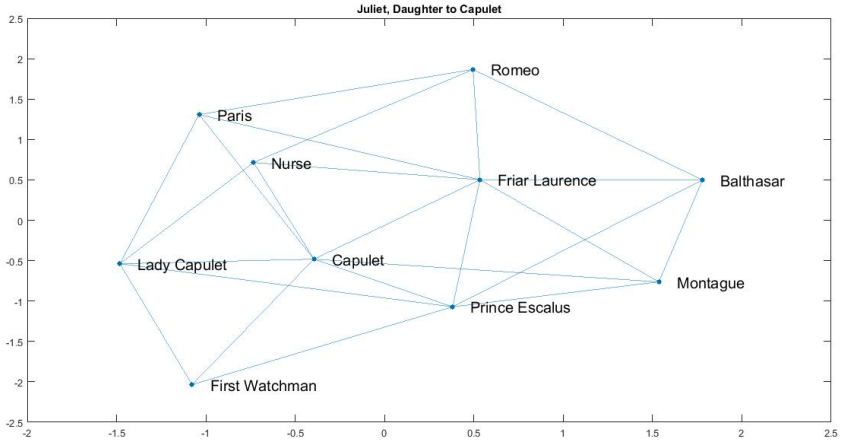
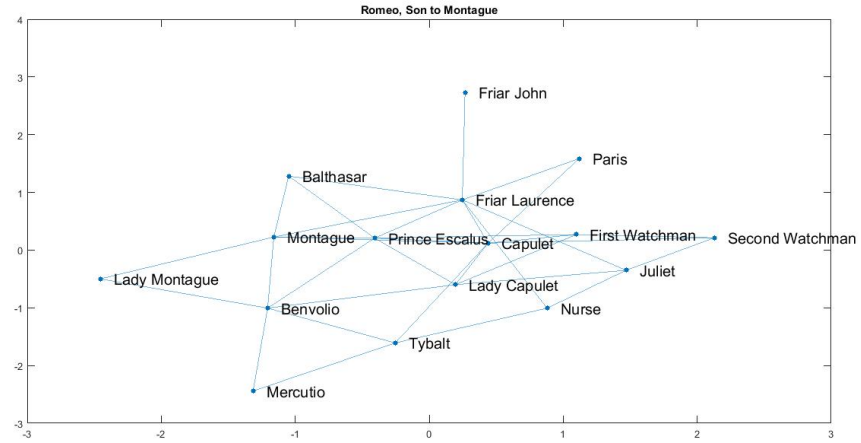


Figure 3.6: Romeo and Juliet





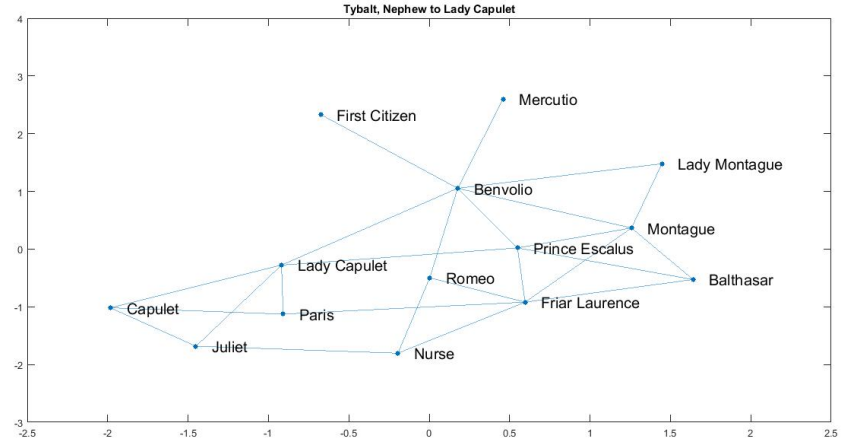


Figure 3.7: Romeo and Juliet

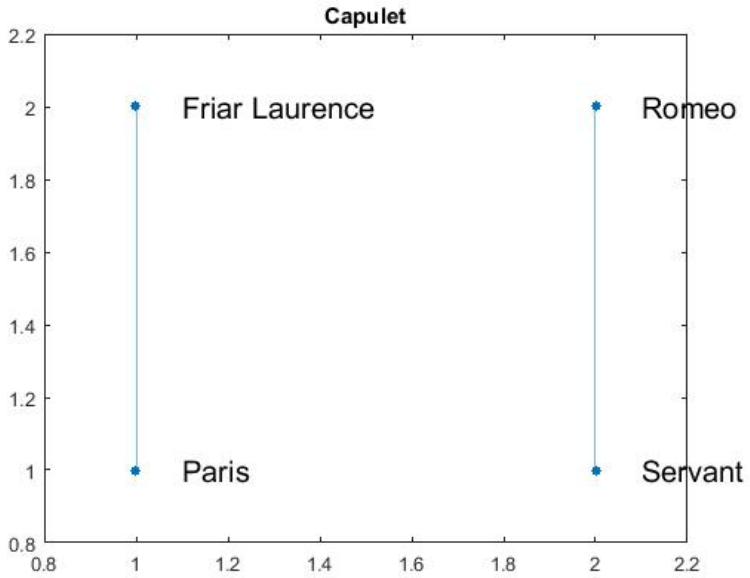


Figure 3.8: Twelfth Night

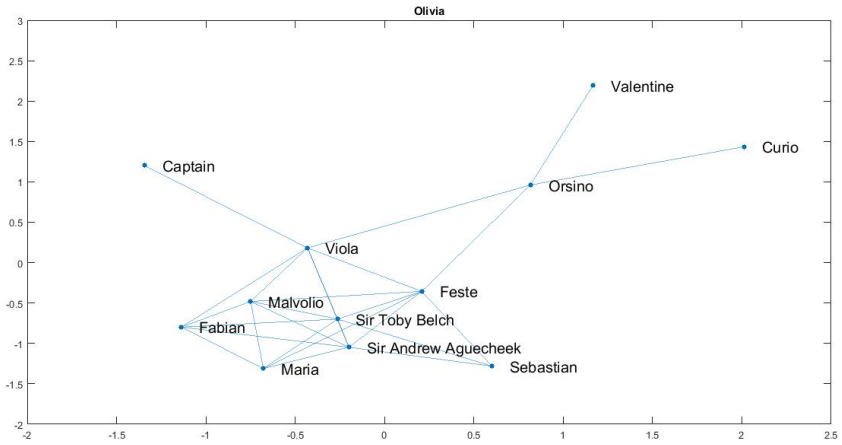


Figure 3.9: Twelfth Night

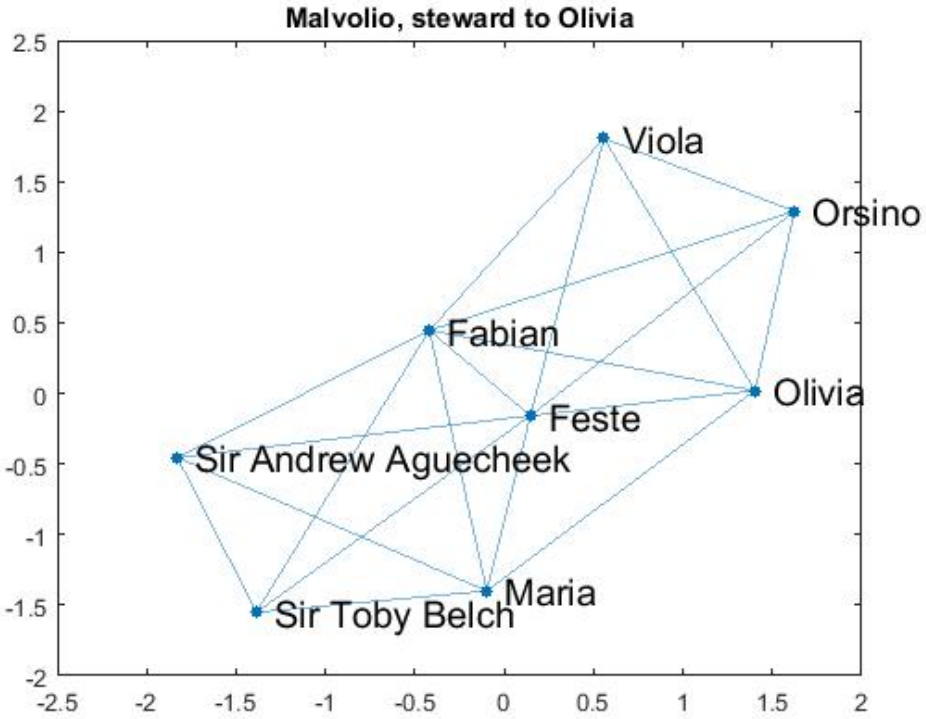


Figure 3.10: Twelfth Night

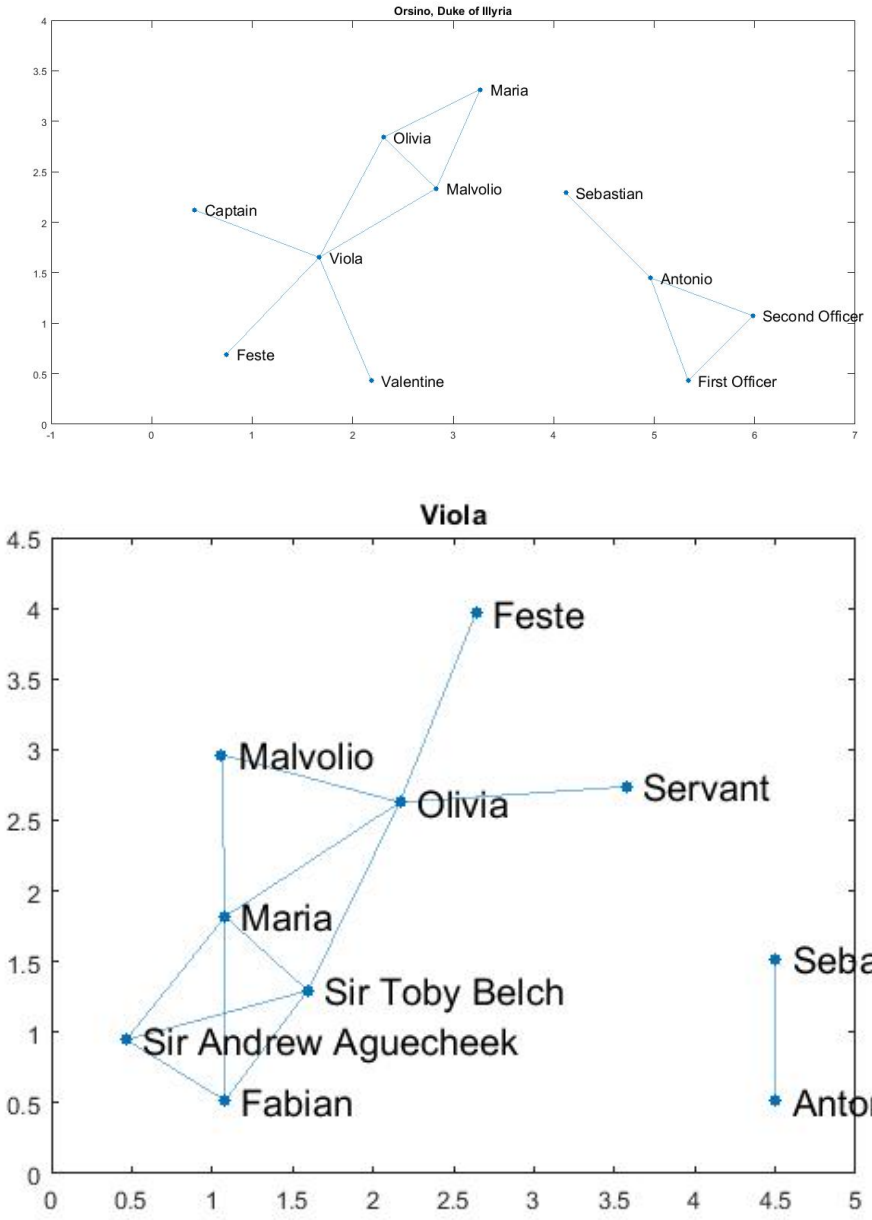
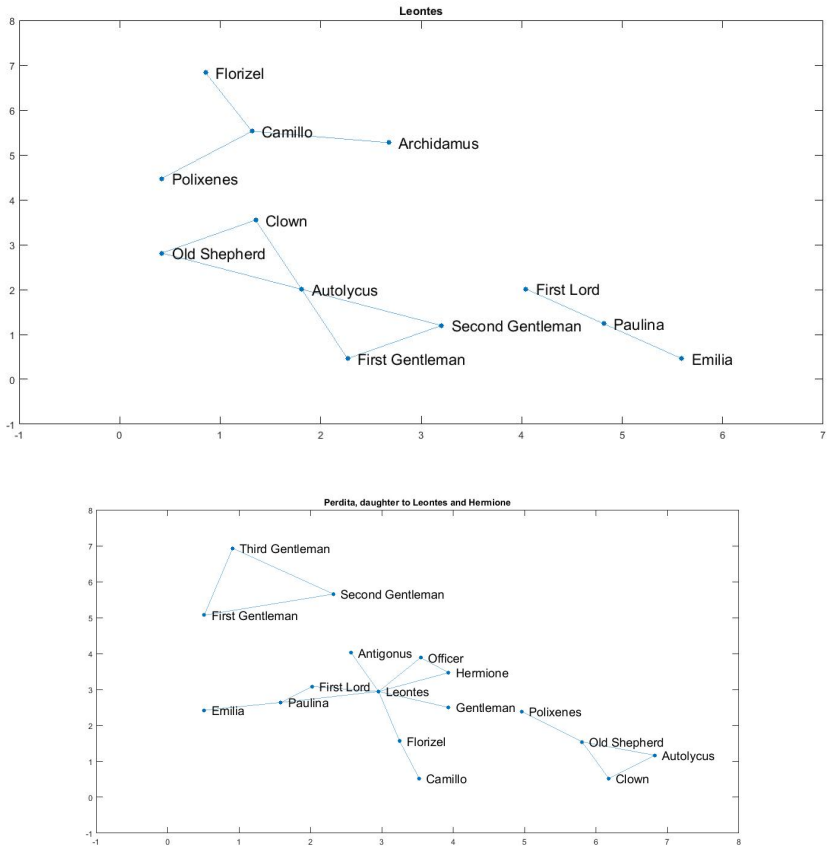


Figure 3.11: The Winter's Tale



Same for Juliet: a lot of people talks about her but in the undirected case we couldn't see her importance. Capulet is almost ignored in this case, so the difference between this centrality and the undirected one tells us that he is important as a man of action and not as a protagonist of the events. As for Olivia: she doesn't participate actively in the story because she stays at home but she is subjects of all the storyline through the gossiping of the others. Same for Malvolio, considered as a crazy person: so they talk about him but he has no power to influence the plot. Similar considerations for the other mentioned before. We also remark that for the central nodes in the undirected case there are more than one social circle that talks about them (Macbeth, Richard) and for the others the communities are dense and often unique. In general, the undirected case tells us who does the action (but not how: we will see it in the next chapters) and the directed case who is subjected to it. Clearly, if a character is high in both, this fact increases its level of general importance in the plot.

Chapter 4

Graph partitioning, community detection and a new algorithm

Another important question in network analysis is to find a mathematical method that discovers the "communities" in a network, in order to cluster nodes that are in some way more connected with each other than with the rest of the graph. One way of saying it is that we are looking for a mathematical study of the properties of a graph that can find the "dense" groups of edges and, naturally, the correspondent groups of nodes. Clearly, there can be a lot of communities in a network so for complex networks this is not a simple request, especially because the complexity order of some methods is often quite expensive.

Some algorithms assume a priori the size of the communities that they aim to detect (*graph partitioning*), some others do not need this information (*community detection*).

We introduce now some of the methods that we will use in our applications (detailed in [4]) and we will propose a new one that will also involve centrality measures. We will consider only undirected graphs.

4.1 Spectral partitioning

The most basic request is to bisect a graph in two groups. A first natural remark is that finding the two "better" communities in a graph means finding two set of disjoint nodes C_1, C_2 such that the number of edges between them (the so called *cut size*) is minimized and such that the cardinality of C_1 and C_2 are fixed, respectively n_1 and n_2.

Let $G = (V, E)$ be an undirected graph with adjacency matrix A and suppose to know C_1 and C_2.

We denote as:

$$D := (C_1 \times C_2) \cup (C_2 \times C_1)$$

the set of all pairs of nodes that belongs to different groups.

The cut size is:

$$R = \frac{1}{2} \sum_{(i,j) \in D} A_{ij}.$$

We define the function:

$$s_i = \begin{cases} +1 & \text{if } i \in C_1 \\ -1 & \text{if } i \in C_2 \end{cases}$$

Note that $s_i^2 = 1$; hence $\frac{1}{2}(1 - s_i s_j) = \chi_D(i,j)$ (i.e. the characteristic function of D) and

$$R = \frac{1}{4} \sum_{i,j} A_{ij}(1 - s_i s_j).$$

Recalling the degree $\mathbf{d}_i$ of a node i and let δ_{ij} be the Kronecker delta, we have:

$$\sum_{i,j} A_{ij} = \sum_i \mathbf{d}_i = \sum_i \mathbf{d}_i s_i^2 = \sum_{i,j} \mathbf{d}_i \delta_{ij} s_i s_j.$$

Thus:

$$R = \frac{1}{4} \sum_{i,j} (\mathbf{d}_i \delta_{ij} - A_{ij}) s_i s_j = \frac{1}{4} \sum_{i,j} L_{ij} s_i s_j.$$

The matrix $L = (l_{ij})_{i,j} = (\mathbf{d}_i \delta_{ij} - A_{ij})_{i,j}$ is called the *Laplacian matrix of* G. We remark that $L\mathbf{1} = \mathbf{0}$. The initial bisection problem can be expressed as follows: we calculate the vector $\mathbf{s} = (s_i)$ that minimizes

$$R = \frac{1}{4} \mathbf{s}^T L \mathbf{s}.$$

If $\mathbf{s}$ was allowed to take any real value then the problem would be easy to solve, but unfortunately this is not the case: we have the restrictions $s_i \in \{-1, 1\}$ and also $\mathbf{1^T s} = \sum_i s_i = n_1 - n_2$. So an approximation approach to this problem is necessary, the so called *relaxation method*.

Now, $\mathbf{s} \in \mathbb{R}^n$ such that $s_i \in \{-1, 1\}$ means that $\mathbf{s}$ points to one corner of the n-dimensional hypercube in $\mathbb{R}^n$ centered in the origin $\Rightarrow \|\mathbf{s}\|^2 = n$.

So we "relax" our constraint searching $\mathbf{s} \in \mathbb{R}^n$ such that:

$$\begin{cases} R = \frac{1}{4} \mathbf{s}^T L \mathbf{s} \\ \sum_i s_i^2 = n \\ \mathbf{1^T s} = \sum_i s_i = n_1 - n_2 \end{cases}$$

that can be expressed with the Lagrange multipliers method as the problem:

$$\frac{\partial}{\partial s_k}[\sum_{i,j} L_{ij}s_is_j + \lambda(n - \sum_i s_i^2) + 2\mu(n_1 - n_2 - \sum_i s_i)]$$

$$\Rightarrow \sum_j L_{ij}s_j = \lambda s_k + \mu \Rightarrow L\mathbf{s} = \lambda\mathbf{s} + \mu\mathbf{1}$$

$$\Rightarrow \mathbf{1^T}L\mathbf{s} = 0 = \mathbf{1^T}(\lambda\mathbf{s} + \mu\mathbf{1}) = \lambda(n_1 - n_2) + \mu$$

$$\Rightarrow \mu = -\frac{1}{n}\lambda(n_1 - n_2)$$

Defining $\mathbf{x} = \mathbf{s} + \frac{\mu}{\lambda}\mathbf{1} = \mathbf{s} + \frac{1}{n}(n_1 - n_2)$ then $L\mathbf{x} = \lambda\mathbf{x}$. So the solution of the problem depends on the choice of an eigenvector $\mathbf{x} \neq \mathbf{1}$ of L. We remark that:

$$R = \frac{1}{4}\mathbf{s^T}L\mathbf{s} = \frac{1}{4}\mathbf{x^T}L\mathbf{x} = \frac{\lambda}{n}(n_1n_2)$$

so if we want to minimize this value we have to find a unit eigenvector $\mathbf{v_s}$ associated to the smallest eigenvalue $\lambda_s \neq 0$ of L, and define $\mathbf{x} := \frac{4}{n}(n_1n_2)\mathbf{v_s}$.

We denote as I_{n_1} the set of indices of the n_1 largest components of $\mathbf{x} + \frac{1}{n}(n_1 - n_2)\mathbf{1}$, that coincides to the set of indices of the n_1 largest components of $\mathbf{x}$. The ideal not relaxed $\mathbf{s}$ is such that $\mathbf{s}^T\mathbf{s} = n$ so we do the best we can and we impose that $\mathbf{s}^T(\mathbf{x} + \frac{1}{n}(n_1 - n_2)\mathbf{1})$ is maximized. This happens for $\mathbf{s}$ such that:

$$\mathbf{s}_i = \begin{cases} 1 & \text{if } i \in I_{n_1} \\ -1 & \text{else} \end{cases}$$

So we finally find $C_1 = I_{n_1}$, $C_2 = V \setminus C_1$.

If $n_1 \neq n_2$ then the order in which we choose n_1 and n_2 is crucial, so we can invert the order of C_1 and C_2, reapply all this procedure finding a new R and decide who is C_1 and who is C_2 based on the order that gives the smaller R.

4.2 Spectral modularity detection

We can apply a similar procedure without establish the sizes of the two communities. First, we introduce the concept of modularity of a graph $G = (V, \mathcal{E})$. Let suppose to have two communities $C_1,\ C_2$ in G, i.e. two groups of vertices that are defined having "the same type", such that $C_1 \cup C_2 = V$. In order to define a measure of the presence of dense communities in the graph we can compare the expected number of edges between nodes of the same type and the actual one.

Let $G_{random} = (V, \mathcal{E}')$ be a graph such that we connect i and j through an edge randomly preserving the vertex degrees., i.e., $(i, j) \in \mathcal{E}'$ with probability $\frac{\mathbf{d}_i \mathbf{d}_j}{2e}$. In this new graph, the fraction of the expected number of edges between nodes of the same type is:

$$\sum_{ij} \frac{\mathbf{d}_i \mathbf{d}_j}{2e} \delta(C_1, C_2) ij$$

where $e = |\mathcal{E}|$ and $\delta(C_1, C_2)ij$ is 1 if i and j are in the same community, 0 otherwise.

The actual fraction of edges between nodes of the same type is:

$$\frac{1}{2e} \sum_{ij} A_{ij} \delta(C_1, C_2)_{ij}$$

So we can built the *modularity of* G as the difference between the actual number and the expected one:

$$M := \frac{1}{2e} \sum_{ij} (A_{ij} - \frac{\mathbf{d}_i \mathbf{d}_j}{2e}) \delta(C_1, C_2)_{ij}$$

To understand why this makes sense, it can be useful to think about the case where all the nodes are of one and only one type. With only one community the modularity is 0, so there is no more than one dense community in the graph.

We call

$$B_{ij} = ((A_{ij} - \frac{\mathbf{d}_i \mathbf{d}_j}{2e}) \delta(C_1, C_2))_{ij}$$

the *modularity matrix of* G and we remark that $\mathbf{1}^T B = B\mathbf{1} = 0$. With the same notation as in the previous section, we can rewrite:

$$M = \frac{1}{4e} \mathbf{s}^T B \mathbf{s}$$

with constraint $|s_i| = 1$.

This time our goal is to find $\mathbf{s}$ that maximizes M and, since we don't make any assumptions on the sizes of the communities, the only constraint is $\mathbf{s}^T \mathbf{s} = n$. Again, with Lagrange multipliers method the problem has the form:

$$\frac{\partial}{\partial s_k} [\sum_{i,j} B_{ij} \mathbf{s}_i \mathbf{s}_j + \lambda (n - \sum_i \mathbf{s}_i^2)$$

$$\Rightarrow B\mathbf{s} = \lambda \mathbf{s} \Rightarrow M = \frac{n}{4e} \lambda$$

We choose $\mathbf{v_1}$ the unit eigenvector associated to the largest eigenvalue of B and we define $\mathbf{x} = \frac{n}{4e}\mathbf{v_1}$. We denote as I the set of the indices of the positive components of $\mathbf{x}$ and we impose that $\mathbf{s}$ has to maximize the quantity $\mathbf{s}^T\mathbf{x}$. This happens for $\mathbf{s}$ such that:

$$\mathbf{s}_i = \begin{cases} 1 & \text{if } i \in I \\ -1 & \text{else} \end{cases}$$

so we can define $C_1 := I$, $C_2 := V \setminus I$.

4.3 Hierarchical clustering

We examined the case where we are satisfied with only two groups. If we would like to detect more than two communities, one way of doing it is to apply k times the previous algorithm to detect 2^k communities. Even if there are plenty of networks where this request can be enough, this is of course not the best thing we can do, especially in complex systems (naming one: social networks) where it is naive to expect that there can be only two communities, or that repeated bisections can give acceptable answers.

We present now a method that decompose in more than two parts a graph with the visual use of dendrograms.

4.3.1 Similarity measures

First, we fix a metric on the set of nodes V that has the task to compare them in the sense of dissimilarity. In order to do that we consider the rows of the adjacency matrix A as vectors in $\mathbb{R}^n$ and we denote the i-th row of A as A_i. Here some of the most common examples. For more see [4].

- *Cosine similarity:* the number of shared neighbors between i and j is $(A^2)_{ij}$; if we consider A_i and A_j, the angle θ between them is such that

$$cos(\theta) = \frac{1}{|A_i||A_j|}\langle A_i, A_j\rangle = \frac{(A^2)_{ij}}{\sqrt{\mathbf{d}_i\mathbf{d}_j}} =: \sigma_{ij}$$

  .

- *Euclidean distance:*

$$e_{ij} := \frac{1}{\mathbf{d}_i + \mathbf{d}_j}\sum_k (A_{ik} - A_{jk})^2 = 1 - \frac{2(A^2)_{ij}}{\mathbf{d}_i + \mathbf{d}_j};$$

- *Pearson correlation:*

$$r_{ij} := \frac{var(A_i)}{var(A_j)} cov(A_i, A_j) = \frac{\sum_k (A_{ik} - \mathbb{E}(A_i))(A_{jk} - \mathbb{E}(A_j))}{\sqrt{\sum_k (A_{ik} - \mathbb{E}(A_i))^2} \sqrt{\sum_k (A_{jk} - \mathbb{E}(A_j))^2}}.$$

  Note that $cov(A_i, A_j)$ is the difference between the number of shared neighbors between i and j and the expected number of shared neighbors that i and j would have if we had placed the edges randomly, so this idea is similar to the one seen for the definition of modularity matrix.

We can also define a similarity measure between subsets of nodes and this is useful when we want to compare not just pair of nodes but pair of groups of nodes, the so called *linkage clustering*. Let C_1, C_2 be two groups of nodes, and let s be a similarity measure between nodes. We define:

- *single-linkage clustering:* $S(C_1, C_2) = \max_{(i,j) \in C_1 \times C_2} s_{ij}$
- *complete-linkage clustering:* $S(C_1, C_2) = \min_{(i,j) \in C_1 \times C_2} s_{ij}$
- *UPGMA:* $S(C_1, C_2) = \frac{1}{|C_1||C_2|} \sum_{(i,j) \in C_1 \times C_2} s_{ij}$

4.3.2 The method

We call the $|V| = n$ nodes of G as $\{1, \dots, n\}$. We fix a similarity measure:

$$\begin{aligned} s : \mathcal{E} &\longrightarrow \mathbb{R} \\ (i, j) &\mapsto s_{ij} \end{aligned}$$

and let $C_1^0, \dots, C_n^0$ be the n subset of V such that $C_i = \{i\}$.

Let $m_0 := \max_{ij} s_{ij}$ and $I_0 = \{i \in V : \exists\ j \in V : \ s_{ij} = m_0\}$. Then we define:

$$\begin{cases} C_1^1 = \bigcup_{i \in I_0} C_i^0 \\ C_s^1 = C_s^0 \qquad \text{if } s \notin I_0 \end{cases}$$

so basically we created a new partition $C_1^1, \dots, C_{n_1}^1$ of V from the initial one, joining the vertices that realize the highest similarity. Note that $n_1 \leq n$.

Now, if S is the similarity measure between groups associated to s, let $m_1 := \max_{ij} S(C_i^1, C_j^1)$ and $I_1 = \{i \in V : \exists\ j \in V : \ S(C_i^1, C_j^1) = m_1\}$. Then we define:

$$\begin{cases} C_1^2 = \bigcup_{i \in I_1} \\ C_s^2 = C_s^1 \text{ if } s \notin I_1 \end{cases}$$

so we have a new partition $C_1^2, \ldots, C_{n_2}^2$ with $n_2 \leq n_1 \leq n$, and so on.

After a finite number of steps k we end with one partition C_1^k in which we have finally joined all the vertices in one class.

By construction, this method can be represented by a dendrogram. So if we want to use this algorithm to divide in m parts the network, what we can do is to read the $(k-m)$-th step of the procedure on the dendrogram, intersecting it with an horizontal line in that exact step and see under the line who are the groups. A simple example is given in the following two figures:

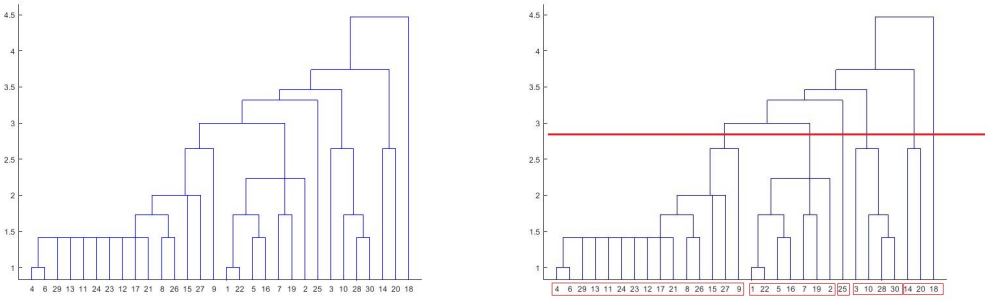


Figure 4.1: Detection of 5 communities in a graph of 30 nodes

4.4 New algorithm: Voronoi cells on centrality nodes

Let us suppose to have the task of detecting the communities in a graph based on a particular aspect of the network. Now, if we choose a good centrality measure that depends on that particular aspect, then there is a high chance that the communities strictly depend on the positions in the graph of the most important nodes. So the idea is to detect communities of the important nodes (i.e., the entourage of a particular important node). If a node is "close enough" to an important node then it is plausible to think that they belong to the same group.

We present this new technique that is the result of these remarks. We apply it for the biggest connected component of a graph because it is more suitable for our applications; by the way, it can be extended to all the connected components. In fact, if a graph has two or more disconnected components it is obvious that they are communities on their own. So in order to detect their sub-communities we can apply this method to every connected

component, not just the biggest one.

Let $G = (V, \mathcal{E})$ be an undirected graph and $G^+ = (V^+, \mathcal{E}^+)$, $|V^+| = n$, is the biggest connected component of G. With the same notations as in definition 2.2, let $\mathbf{C}$ be a centrality measure vector on G^+, $\mathbf{R_C}$ the associated ranking vector and

$$\mathcal{R} := \{i \in V : (\mathbf{R_C})_i < k\}.$$

for a pre-fixed k.

As we remarked, it is not guaranteed that $|\mathcal{R}| = k - 1$. In fact, there can be nodes with the same ranking score in $\mathbf{R_C}$. In our applications we will use a particular k chosen as a measure of how many characters form the leading groups of the most important nodes in the plot. Let r be the number of repeated ranking scores in $\mathbf{R_C}$. Then we define:

$$k' := \begin{cases} n - r + 1 & \text{if } (\mathbf{R_C})_n = (\mathbf{R_C})_{n-1}) \\ n - r & \text{else} \end{cases}$$

and

$$k_{max} = \frac{n}{k'}. \tag{4.1}$$

In our application we will use:

$$k := \begin{cases} \max\{[k_{\max}], 2\} & \text{if } k_{\max} - [k_{\max}] \leq \frac{1}{2} \\ [k_{\max}] + 1 & \text{else} \end{cases} \tag{4.2}$$

where we denoted as $[\cdot]$ the floor function.

Remark 6. The basic idea behind this choice is the following: let

$$r'_k = |\{i \in V : (\mathbf{R_C})_i = k\}|, \ 1 \leq k \leq n$$

.

Then $\sum r'_k =: k'$. We want to consider only the nodes that form the most important group in the graph, based on the behavior of the ranking $\mathbf{R_C}$. Since the ranking can have multiple equal positions, we can consider k' as the actual number of nodes, that represent the maximum value of the vector $\mathbf{R'_C} = (r'_k)_k$ such that the jumps between two consecutive rank scores are less or equal to one. Formally we define $\mathbf{R'_C} = (r'_k)_k$ such that:

$$\begin{cases} r'_1 = 1 \\ r'_{k+1} = r'_k \delta_{(\mathbf{C})_{k+1}, \mathbf{C})_k} + (k+1)(1 - \delta_{(\mathbf{C})_{k+1}, \mathbf{C})_k} \text{ for } 1 < k < n \end{cases}$$

hence,

$$k' = \max \mathbf{R'_C}.$$

Thus, $\frac{n}{\max_i(\mathbf{R}'_{\mathbf{C}})_i} = \frac{n}{k'}$ is a measure of how many characters we can consider as the group of the most important. Since our graphs in the applications are small, the proportion as before is very sensitive, so the choice 4.2 on the decimal part was necessary and this choice turned out to be good for our purposes.

We define the *Voronoi cell* of node i as the set:

$$\mathcal{V}_i := \{j \in V : d(j,i) < d(j,l),\ \forall\ l \in \mathcal{R} \setminus \{i\}\}.$$

where we recall that $d(j,i)$ is the distance from i to j.

If $\mathcal{I} := V \setminus \bigcup_i \mathcal{V}_i \neq \emptyset$ then the nodes in $\mathcal{I}$ have same distance with two ore more important nodes in $\mathcal{R}$. So we want to associate a node $l \in \mathcal{I}$ to the cell with mean distance from l. If such cells $\mathcal{V}_i$ are more than one, we associate l to the cell $\mathcal{V}_i$ with uniform probability.

In order to do it we calculate: $\forall l \in \mathcal{I}$, $\forall i \in \mathcal{R}$

$$m(l) := \min_{i \in \mathcal{R}} \{\frac{1}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} d(l,j)\}$$

i.e. the smaller mean distance between l and the elements of the Voronoi cells. Let:

$$M_l = \{i \in \mathcal{R} : m(l) = \frac{1}{|\mathcal{V}_i|} \sum_{j \in \mathcal{V}_i} d(l,j)\}$$

that is the set of s indices that realizes the smaller mean distance $m(l)$. Now, for all $l \in \mathcal{I}$ we choose a $p_l \in M_l$ with uniform probability $\frac{1}{|M_l|}$, and we define for all $i \in \mathcal{R}$

$$\mathcal{V}_i^* := \mathcal{V}_i \cup \{l \in \mathcal{I} : p_l = i\}.$$

Hence, we have $|\mathcal{R}|$ communities.

If $\mathcal{I} := V \setminus \bigcup_i \mathcal{V}_i \neq \emptyset$ we simply have $\mathcal{V}_i^* = \mathcal{V}_i$.

Now we calculate $\forall i \in \mathcal{R}$:

$$m^*(i) := \{\frac{1}{|\mathcal{V}_i^*|} \sum_{j \in \mathcal{V}_i^*} d(i,j)\},$$

i.e. the mean distance in the new cell $\mathcal{V}_i^*$.

Now, consider $i,\ j \in \mathcal{R}$, $i \neq j$, the center of two different Voronoi cells $\mathcal{V}_i^*, \mathcal{V}_j^*$. We say that i and j are *close* if

$$d(i,j) < (\frac{1}{3}) \max\{m^*(i), m^*(j)\}.$$

The partition of V that our method provides is the one obtained after all cells $\mathcal{V}_i^*$ and $\mathcal{V}_j^*$ with i close to j have been joined. In mathematical terms, this can be redescribed as follows: let us give $\mathcal{R}$ a graph structure $(\mathcal{R}, \mathcal{E}_\mathcal{R})$ where $(i,j) \in \mathcal{E}_\mathcal{R} \iff i$ close to j.

Let us consider the connected components of this graph: $\{\mathcal{C}_\beta\}$ with β in some index set $\mathcal{B}$. Then the partition we define is $\{\mathcal{V}_\beta^{**}\}_{\beta \in \mathcal{B}}$, where:

$$\mathcal{V}_\beta^{**} := \bigcup_{i \in \mathcal{C}_\beta} \mathcal{V}_i^*$$

Chapter 5

Communities in Shakespeare's plays

We now want to detect the principal communities in the stories using the undirected network $G_1 = (V_1, \mathcal{E}_1)$, i talks to j. First we begin applying the already seen methods: modularity and hierarchical clustering.

5.1 Modularity and hierarchical clustering

We start with the bisection modularity algorithm. Every pair of figures represents the two divisions of the graph. We remark that if two characters in the same community were connected through another node in G_1, here they become disconnected.

In *Romeo and Juliet* we expect that the two basic communities must have something to do with Capulets and Montagues. We will see with another method that this is actually not the only meaningful bisection, so as we said this kind of analysis allow us to enlighten some not apparently immediate considerations. Modularity algorithm acceptably succeeds in detecting the families detail. Same for *Twelfth Night*: Viola's circle and Olivia's circle are well divided:

Figure 5.1: *Romeo and Juliet*: first group

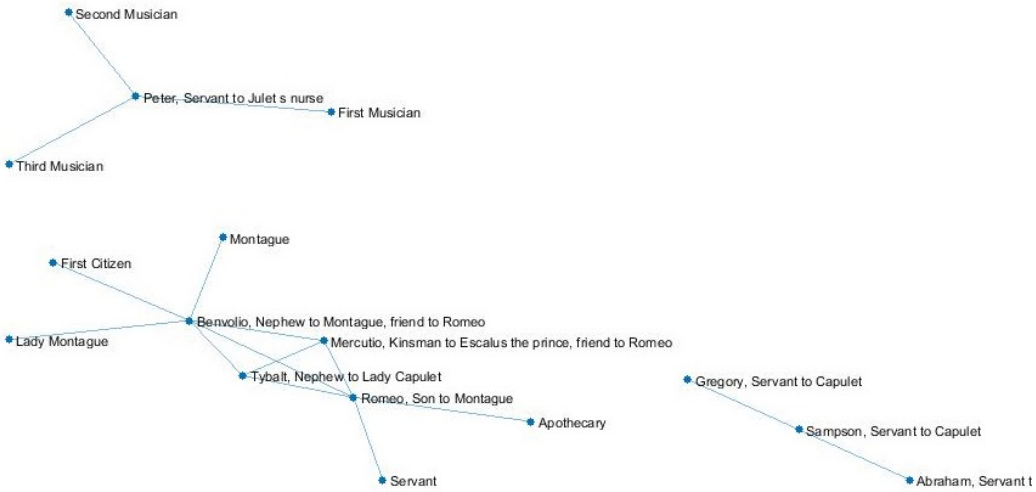


Figure 5.2: *Romeo and Juliet*: second group

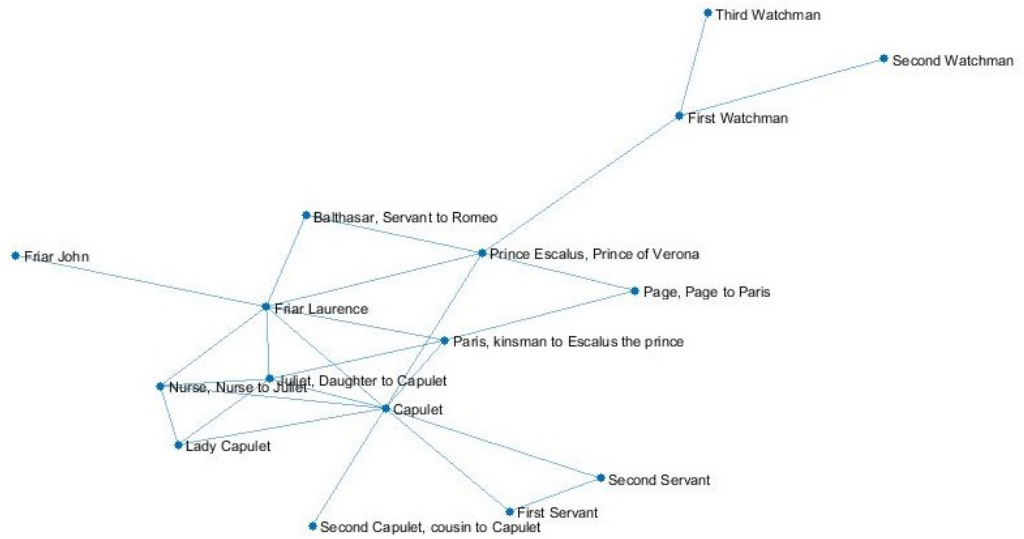


Figure 5.3: *Twelfth Night*: first group

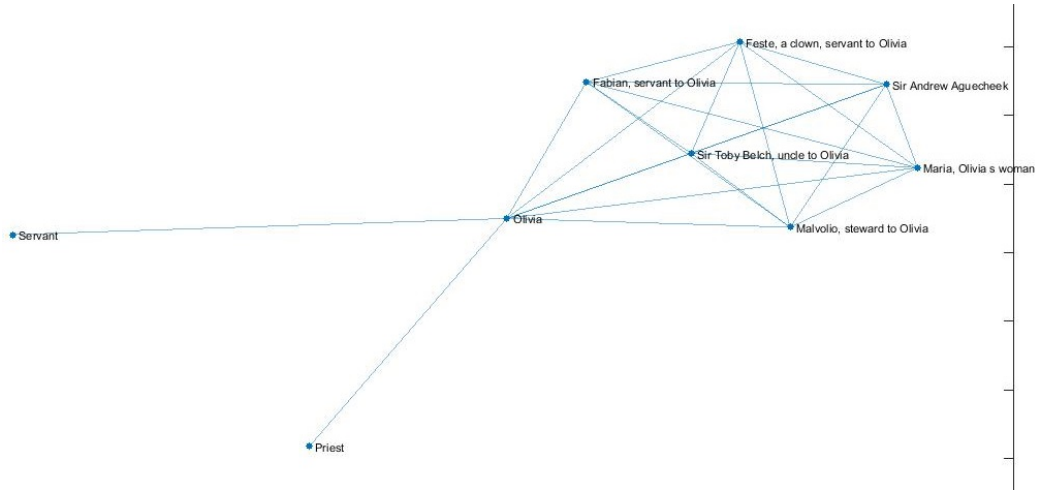
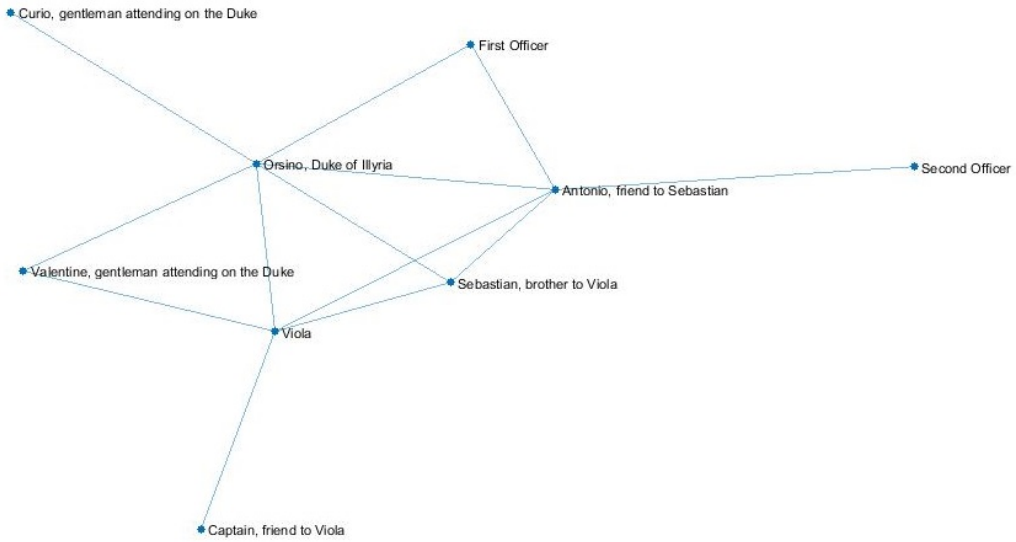


Figure 5.4: *Twelfth Night*: second group



A less but still acceptable division is done for *Richard III*. In fact, the first community is coherent with the entourage of Richard and the rest is relegated to the second community.

Figure 5.5: *Richard III*: first group

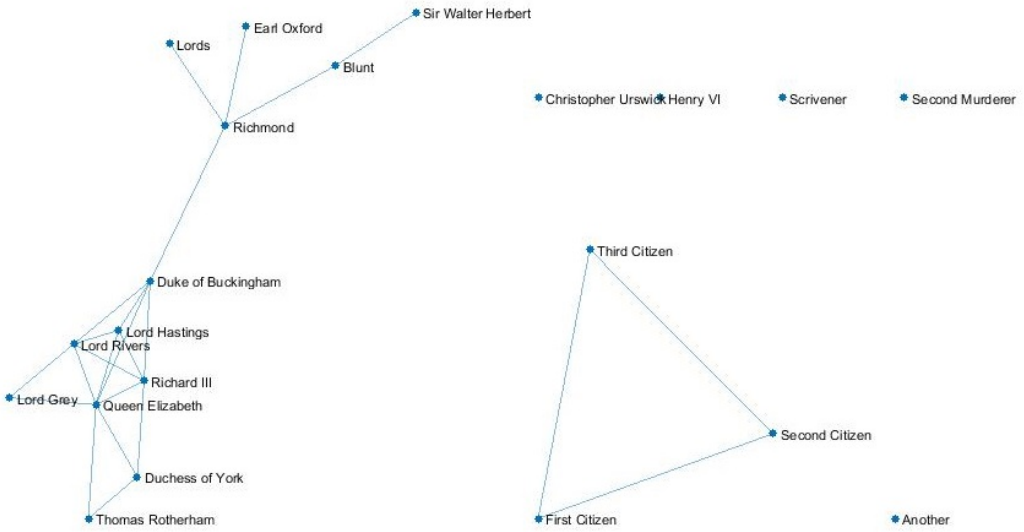
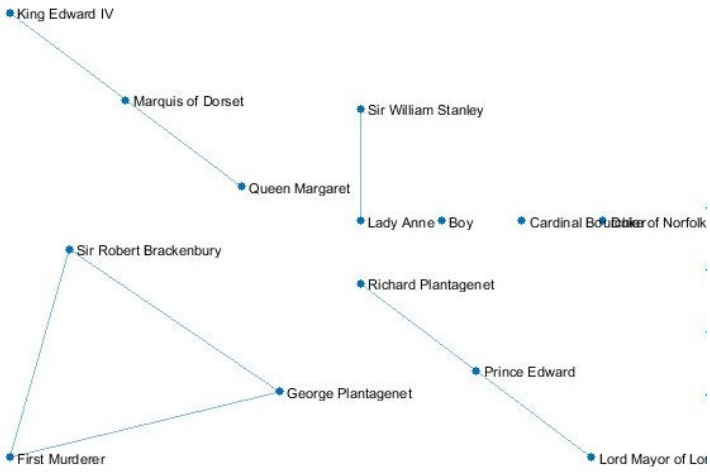


Figure 5.6: *Richard III*: second group



The method fails for the plots where the actual bisection in the plot is more qualitative than expected and it needs a more appropriate method that could fit these qualitative aspects. In fact this last two plays are not consistent for this method: it detects some aspects but not sufficiently enough if we think to read these graphs without knowing the synopsis.

Figure 5.7: *Macbeth*: first group

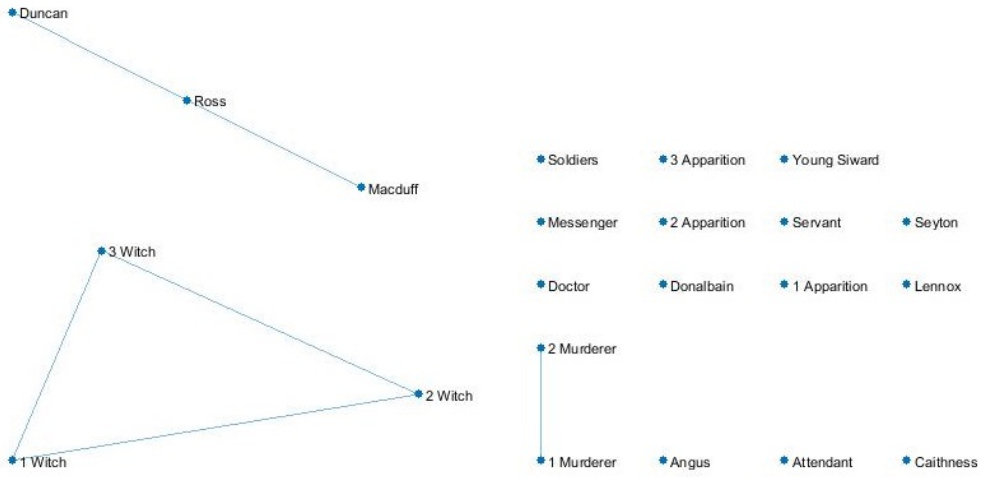


Figure 5.8: *Macbeth*: second group

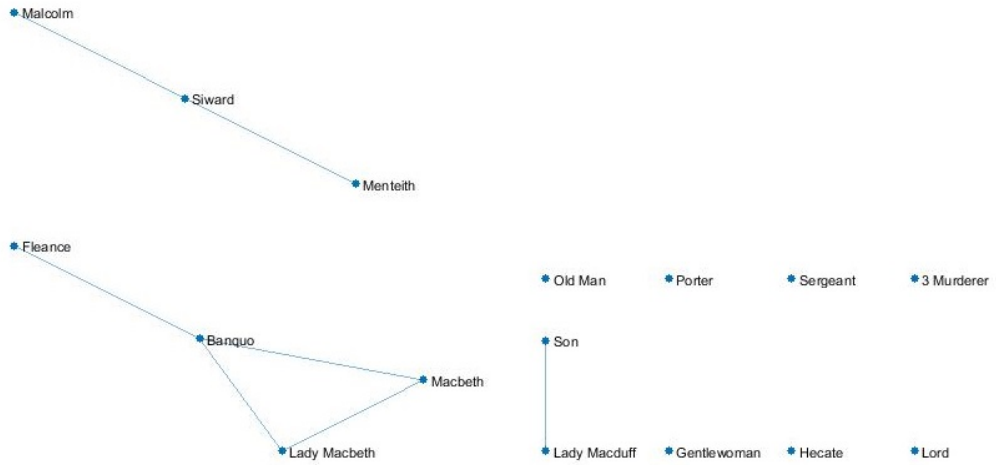


Figure 5.9: *The Winter's Tale*: first group

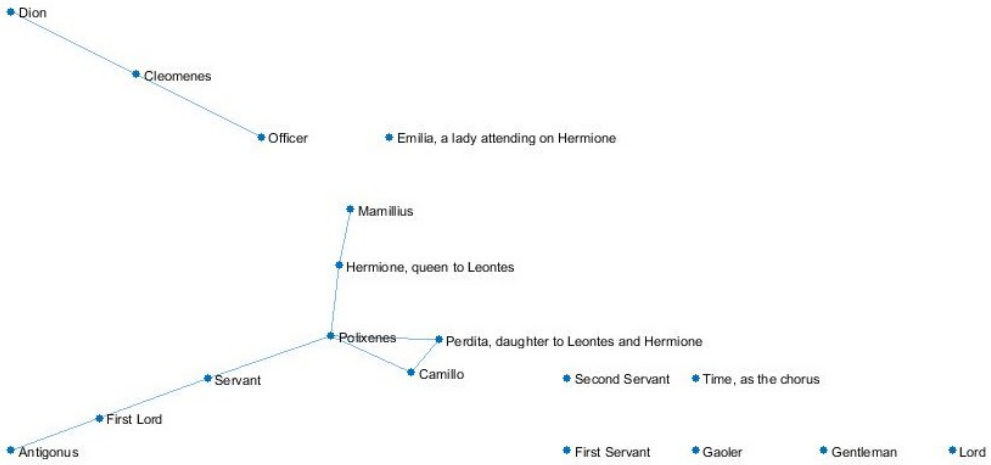
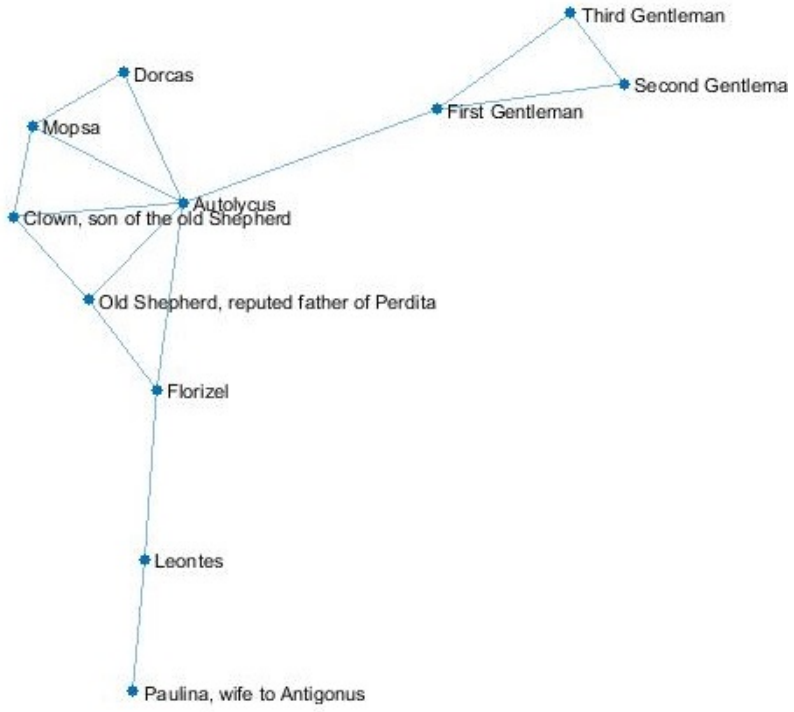


Figure 5.10: *The Winter's Tale*: second group



Dendrograms gives a different acception of the communities based, as we said, on the hierarchic of the characters. Thus it is not usefull for studying the stories that are supposed to have a more fragmented groups' structure, but they are surely usefull to understand in the two bisections of modularity who is the central character that has the power in the community: Romeo and Capulet, Viola/Orsino and Olivia.

Figure 5.11: Romeo and Juliet

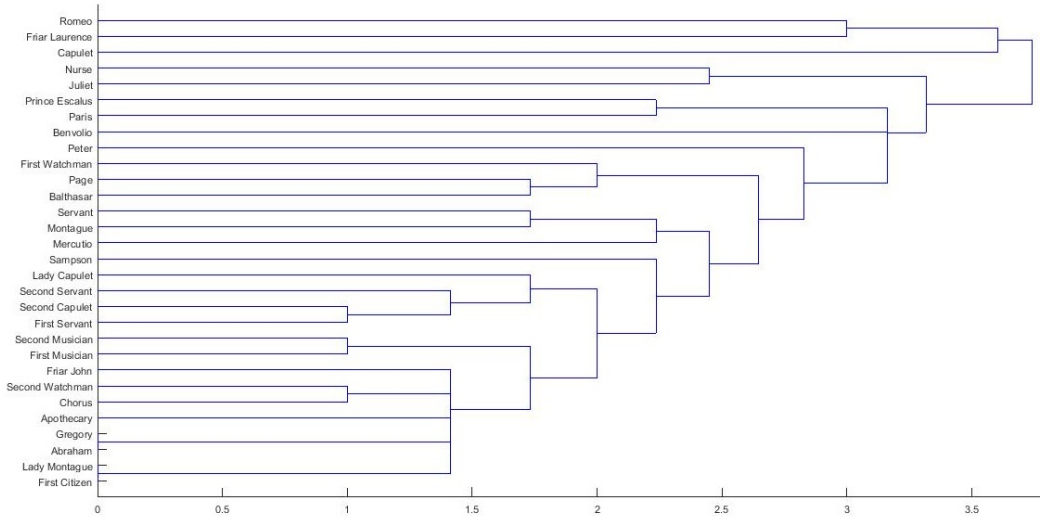


Figure 5.12: Twelfth Night

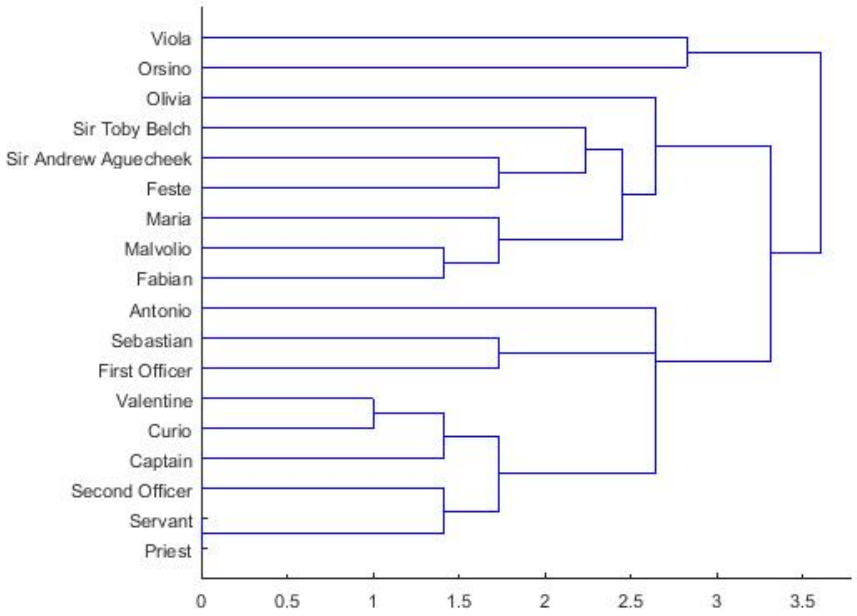
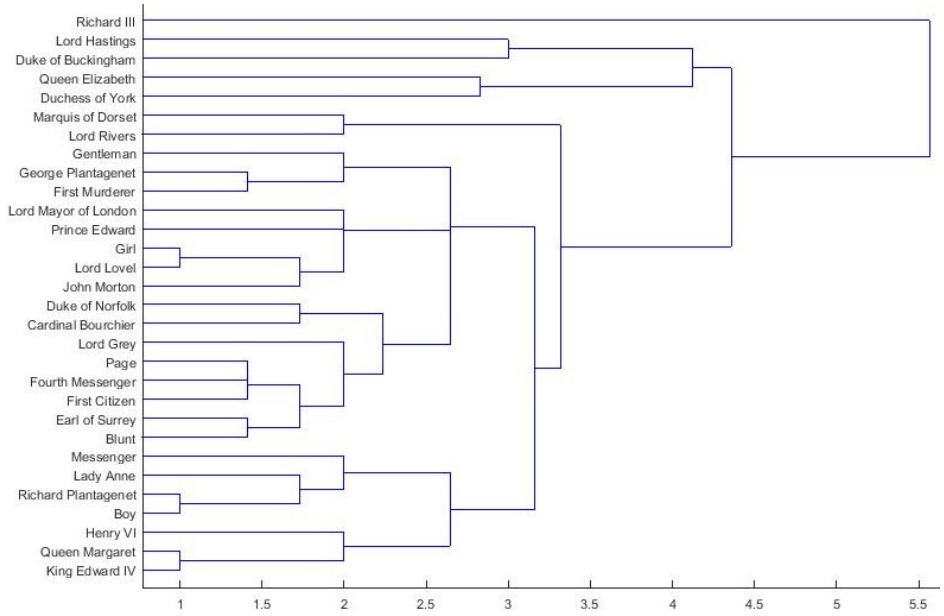


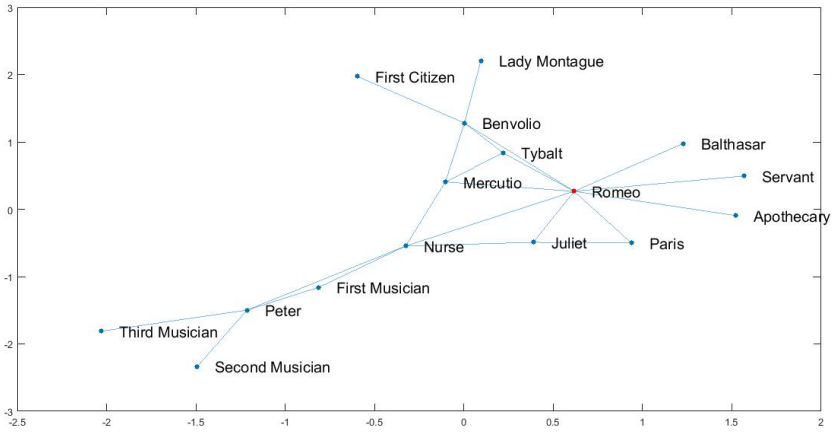
Figure 5.13: Richard III

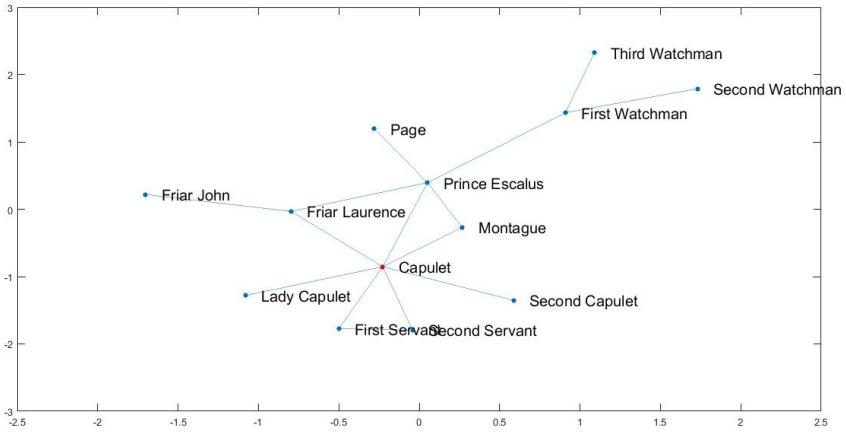


5.2 Voronoi cells

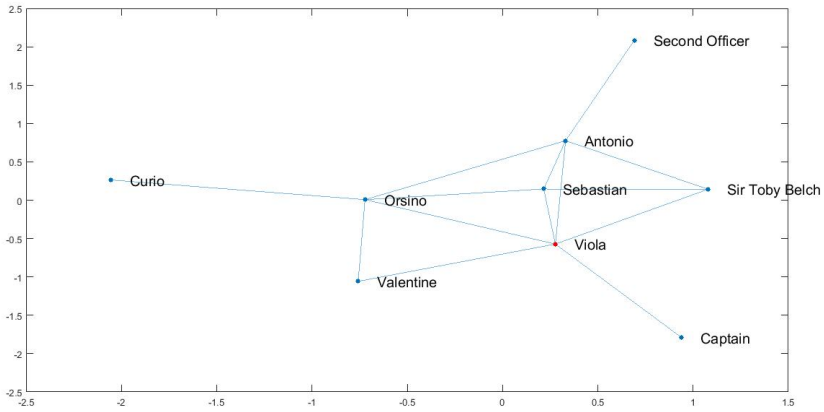
Since we are interested in detecting communities based on how they interact in terms of exchange of lines, we have chosen as centrality the degree. We choose k as in 4.1: 2 important initial nodes for *Romeo and Juliet*, *Twelfth Night* and *The Winter's Tale*, 4 for *Richard III* and 5 for *Macbeth*. Applying the method our graph is too small to merge Voronoi cells, with these choices of k. We remark once again that ranking positions can be equals: as we can see in section 3.2.1, in *Richard III* the second position for degree is shared by both Queen Elizabeth and Lord Hastings.

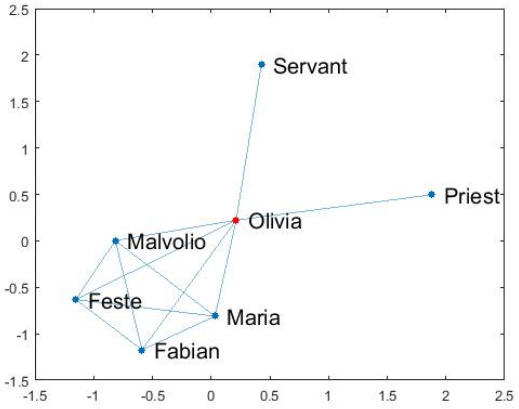
In *Romeo and Juliet* the communities are well divided. As we expected, this method gives more influence to the local interactions, that's why this partitioning leads us to the group of old people and the group of young people; this is a bisection not immediately conceivable. In the tragedy that's exactly how the events go: old people take decisions, young people react between each other.


Lady Montague
First Citizen
Benvolio
Tybalt
Balthasar
Servant
Mercutio
Romeo
Apothecary
Nurse
Juliet
Paris
First Musician
Peter
Third Musician
Second Musician

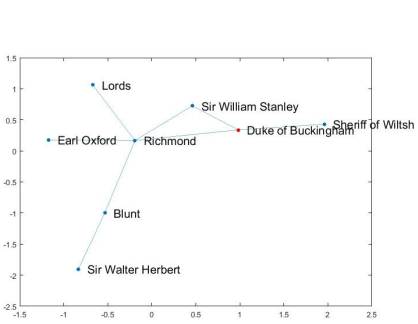

Third Watchman
Second Watchman
First Watchman
Page
Prince Escalus
Friar John
Friar Laurence
Montague
Capulet
Lady Capulet
Second Capulet
First Servant
Second Servant

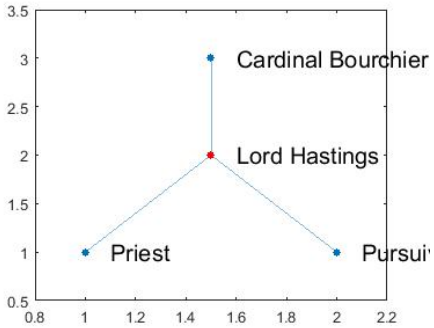
In *Twelfth Night* it is even more clear: we have the group of Viola and Olivia, in which we find their exact entourage. For now these are not so obvious results because without knowing the history (one of our silent goals) just a look on the initial graph did not say anything about the communities. For example in Twelfth Night the initial graph is very dense and the mean distance between the characters seems very small.

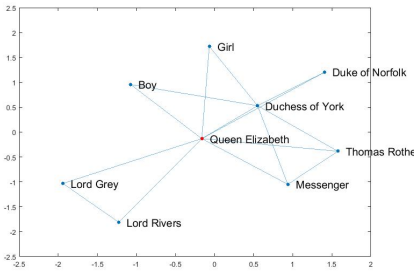


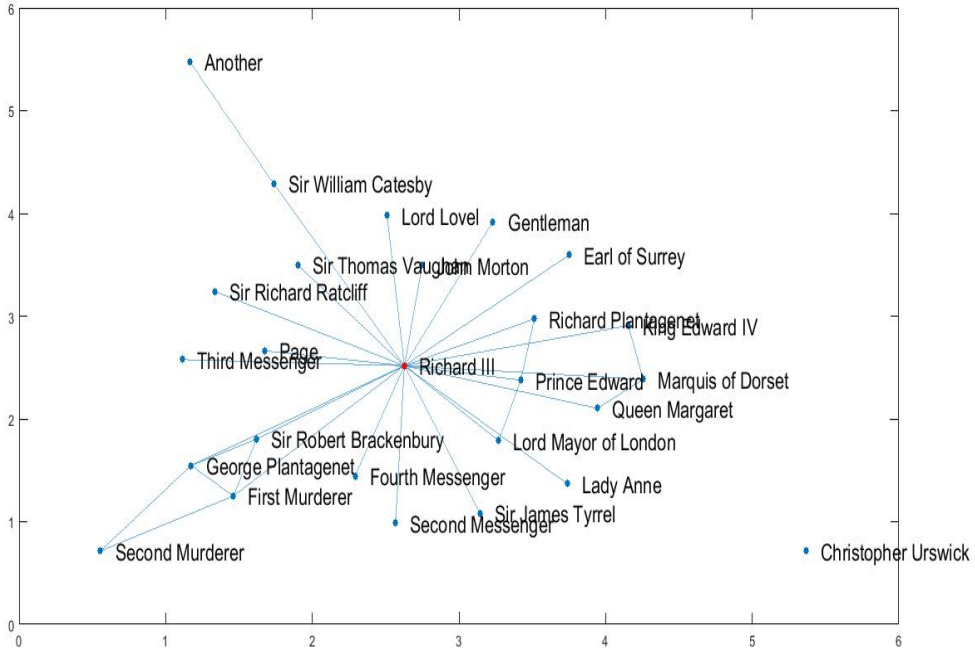


Richard III is the only plot of the five where the centrality is always concentrate in one character with a big distance from the others. This make this story the less suitable for calculations or procedure when the distances between characters are the fundamental tools. We see that with this method Richard incorporates all the characters: that's because he absolutely dominates the presence in the shortest paths, in the scenes, in the dialogues with the others. In fact, if we see the value of betweenness of Richard (555,32) it is way larger than the other characters (the second one is 202,7) in a way that we do not have in all the other plays (the highest gap between first and second place in betweenness is 190 in *Macbeth*), so it makes sense that he is the verbal center of the play: he has the biggest interactions with almost everybody.

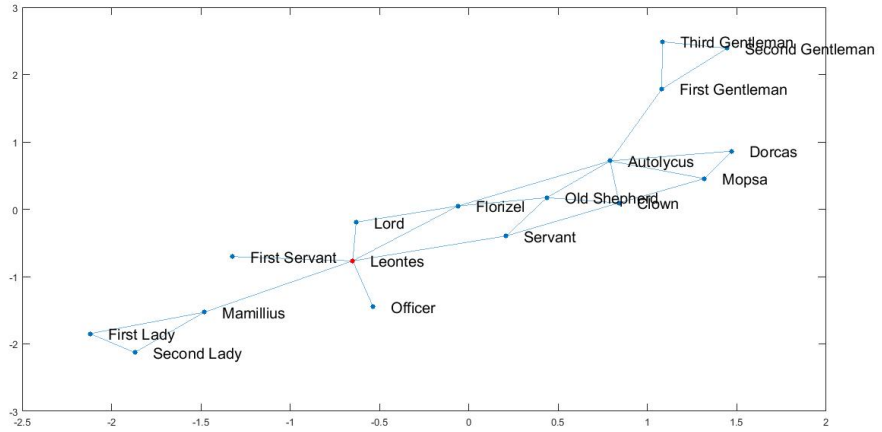


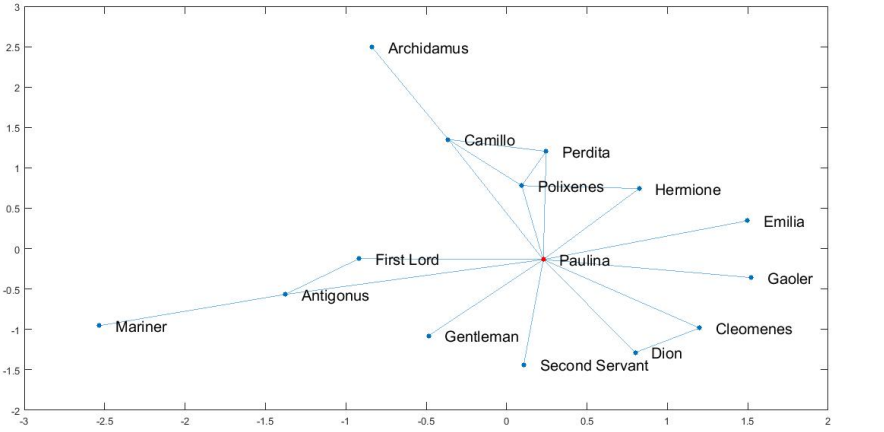




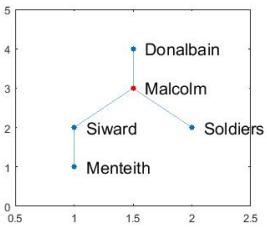

Another
Sir William Catesby
Lord Lovel
Gentleman
Earl of Surrey
Sir Thomas Vaughan
John Morton
Sir Richard Ratcliff
Richard Plantagenet
King Edward IV
Page
Third Messenger
Richard III
Prince Edward
Marquis of Dorset
Queen Margaret
Sir Robert Brackenbury
Lord Mayor of London
George Plantagenet
Fourth Messenger
Lady Anne
First Murderer
Sir James Tyrrel
Second Messenger
Second Murderer
Christopher Urswick

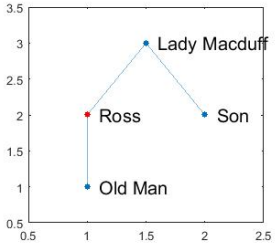
In *The Winter's Tale* the two big communities are formed as we wanted by the entourage of Paulina and the one of Leontes. We can also remark that in this case (as well as in Richard) the topology of the two graphs are not so similar. This indicates a different way to spread the informations.

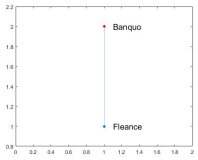


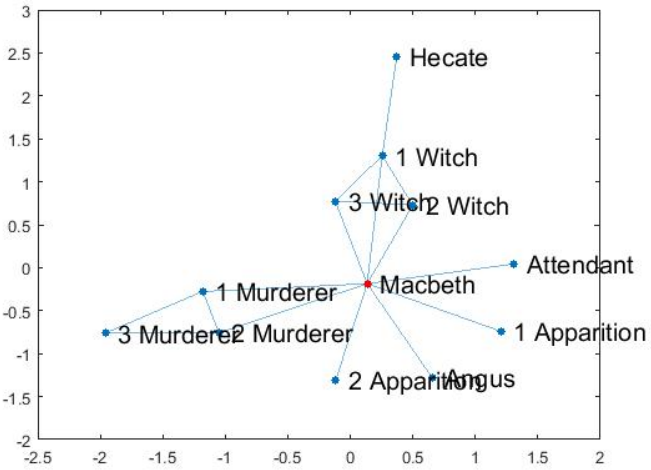


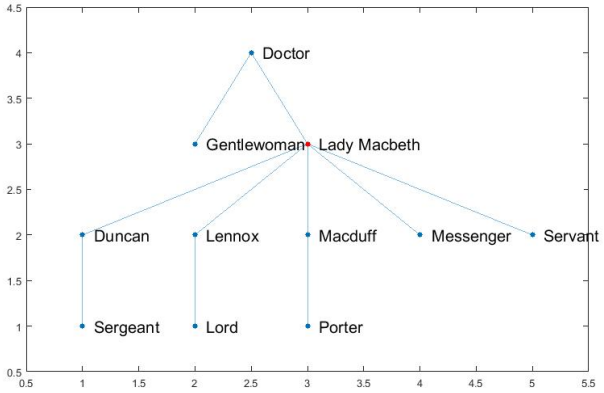
In *Macbeth* we have a qualitatively coherent partition because we start to see the central role of Lady Macbeth and see the major difference of her role with respect to the role of his husband.



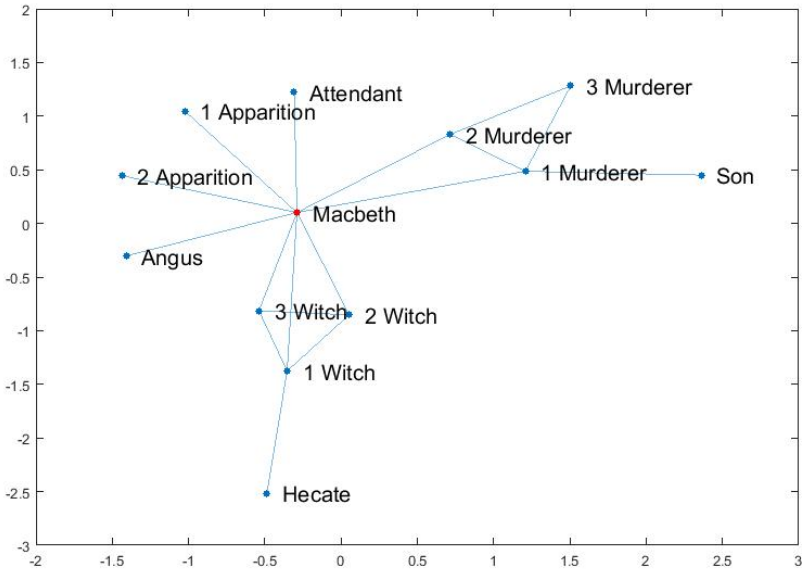


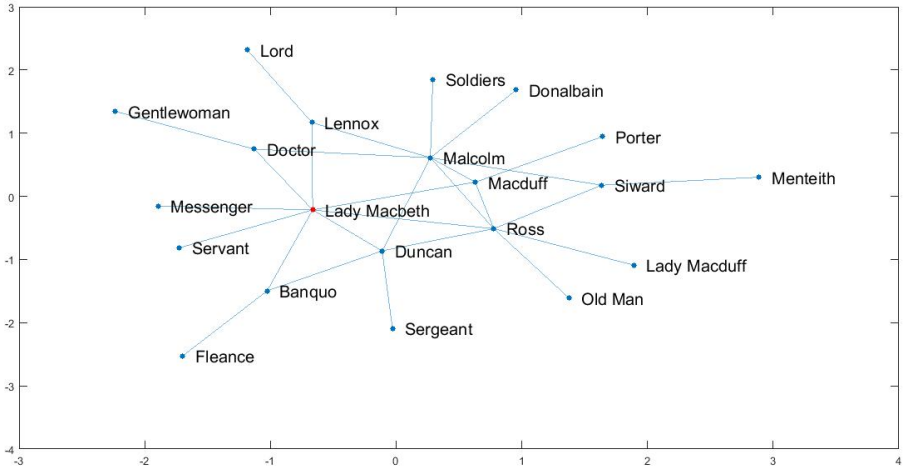






So, using different methods, we bisected in reasonable communities every graph but *Richard III* because there is a node that is too central so a bisection based on distances cannot succeed, and *Macbeth*. Now, we can force to have just two communities with this method in *Macbeth*, so we let $k = 3$. This algorithm succeed in detecting the two big communities in *Macbeth*. We are finally able to fully enlighten the central role of Lady Macbeth, that is indeed the real plotter of the story and the one that makes the decisions that create it: she persuades her husband into committing regicide, after all.





Bibliography

[1] Carl D. Meyer - Matrix analysis and applied linear algebra, SIAM: Society for Industrial and Applied Mathematics; Har/Cdr edition (February 15, 2001)

[2] M. Benzi, Christine Klymko - On the limiting behavior of parameter-dependent network centrality measures, 2015

[3] A. N. Langville and C. D. Meyer - Google's PageRank and Beyond: The Science of Search Engine Rankings, Princeton University Press, Princeton, NJ, 2006

[4] M. E. J. Newman, Networks: An Introduction. Oxford: Oxford University Press, 2010

[5] Newman, M. E. J. The structure and function of complex networks. Department of Physics, University of Michigan.

[6] P. Boldi, M. Santini, and S. Vigna, PageRank: Functional dependencies (2009)

[7] J. Norris: Markov chains. Cambridge University Press, 1998

[8] Brualdi, Richard A.; Ryser, Herbert J. (1992). Combinatorial Matrix Theory. Cambridge: Cambridge UP

[9] Alex Bavelas. Communication patterns in task-oriented groups. J. Acoust. Soc. Am, 22(6):725–730, 1950.

[10] Dekker, Anthony (2005). "Conceptual Distance in Social Network Analysis". Journal of Social Structure. 6 (3).

[11] Yannick Rochat. Closeness centrality extended to unconnected graphs: The harmonic centrality index, Applications of Social Network Analysis, ASNA 2009.

[12] Freeman, Linton (1977). "A set of measures of centrality based on betweenness". Sociometry.

[13] Page, Larry, "PageRank: Bringing Order to the Web" (1999) Engine Rankings, Princeton University Press, Princeton, NJ, 2006

[14] Wayne W. Zachary. "An Information Flow Model for Conflict and Fission in Small Groups". University of New Mexico.

[15] Alexanderson, Gerald (July 2006). "Euler and Königsberg's bridges: a historical view". Bulletin of the American Mathematical Society.