A (Dis)similarity Index for Comparing Two Character Networks Based on the Same Story

François Bavaud[1,2], Coline Métrailler[1]

[1]*Faculty of Arts, Department of Language and Information Sciences, University of Lausanne, bâtiment Anthropole, 1015 Lausanne, Switzerland*

[2]*Faculty of Geosciences and Environment, Institute of Geography and Sustainability, University of Lausanne, bâtiment Géopolis, 1015 Lausanne, Switzerland*

Abstract

Comparing networks is always a complicated matter, whose effective implementation strongly depends on the amount of shared information between them, in particular whether nodes, edges, weights etc. are identical, or not. In the case of character networks and adaptations (from book to movie, from movie to theater, and so on), the formal challenge proves stimulating: some characters will be mapped from one work to the other, some will have no correspondence, and their weights, measuring their relative occurence, are bound to differ.

This formal contribution, rooted in Multivariate Data Analysis, proposes a presumably novel similarity index, the generalized weighted RV coefficient, taking into account both the difference in character weights (nodes) and in character interactions (edges). This approach first requires to transform the character networks into weighted squared Euclidean configurations. We then compare a novel of C.S. Lewis, part of the series *The Chronicles of Narnia*, and the script of its film adaptation to illustrate the proposal and the results.

1. Introduction

Networks of fictional characters often exist in two or more versions. For instance (section 4), network A is built from a novel, and network B from a movie adaptation. Besides the main characters common to both versions, there are characters proper to a single version only. Also, the importance of common characters (as measured, e.g., by their relative occurrence), is bound to vary between the two versions, as is the strength of their mutual relations (as e.g. measured by their relative co-occurrence). Hence, the networks A and B differ both along character weights (nodes) and interaction weights (edges).

Our contribution proposes the definition of a single index measuring the overall similarity between A and B. This index, noted RV, constitutes an innovative generalization, involving *two distinct sets of object weights*, of the *weighted RV-coefficient* [1], which is itself a generalization of the original, unweighted RV-coefficient [2] (where R did refer to "correlation" and V to "vector"). In particular, $RV \in [0, 1]$, with $RV = 1$ iff A and B are identical (i.e., same character weights and dissimilarities between characters), and $RV = 0$ iff A and B have no character in common. This

---

similarity coefficient can in turn be transformed into a dissimilarity coefficient, which can be additively decomposed into five components.

The formalism, exposed in section 2, is rooted into Weighted Data Analysis. It involves three major steps:

1. Transforming a weighted network into a weighted Euclidean configuration (section 2.1). This step is chiefly dictated by the formalism, which requires squared Euclidean dissimilarities, but permits as a byproduct a visualization of the character network (by weighted multidimensional scaling) of interest in itself.
2. Transforming the weighted Euclidean configuration into a kernel, whose eigendecomposition permits to visualize the network nodes (section 2.2)
3. Computing the generalized RV coefficient (beginning of section 3), assessing the similarity between two networks whose node weights may differ, and its exact decomposition into five terms (sections 3.1 and 3.2).

2. Visualization of character networks: a few "reminders"

We formalize character networks as *weighted networks* $(\mathbf{f}, \mathbf{C})$, where $\mathbf{f}$ is the vector of the n character weights, obeying $f_i \geq 0$ and $\sum_{i=1}^{n} f_i = 1$. The $n \times n$ matrix of edge weights $\mathbf{C} = (c_{ij})$ is non-negative, and quantifies the importance of edge ij, reflecting some kind of affinity between the characters i and j, such as the their co-occurrence in the present study (section 4). It can be symmetric (as for $\mathbf{C} = \mathbf{A}$, where $\mathbf{A}$ is a binary adjacency matrix in a non-directed network), or not (as for asymmetric social relationships), in which case the network is directed.

2.1. Extracting Euclidean dissimilarities from a weighted network

Network visualization is a boundless research topic, even restricted as done here to the *Euclidean embedding* of networks. Specifically, one seeks to extract from the network data $(\mathbf{f}, \mathbf{C})$ a matrix $\mathbf{D} = (D_{ij})$ of squared Euclidean dissimilarities between nodes, that is of the form $D_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|^2$, where $\mathbf{x}_i$ is the representative vector of node i. Two proposals only, among many others possibilities (extracting squared Euclidean dissimilarities from a weighted network is a topic in itself), are considered in this contribution: *commute-time distances* (section 2.1.1) and *diffusive distances* (section 2.1.2).

2.1.1. Commute-time distances

By construction, the square matrix $\mathbf{P} = (p_{ij})$ with components $p_{ij} = c_{ij}/c_{i\bullet}$ (where $c_{i\bullet} = \sum_{j=1}^{n} c_{ij}$) is non-negative, with $p_{i\bullet} = 1$: it therefore constitutes the *transition matrix* of a *Markov chain*, defining a *random walk* on the network. The *commute time* D_{ij}^{com} is the average time needed to go from i to j and then back to i [see e.g. 3]. The matrix $\mathbf{D}^{\mathrm{com}} = (D_{ij}^{\mathrm{com}})$ of commute times is well-known to be squared Euclidean [see e.g. 4], irrespectively of the properties of $\mathbf{C}$, as far as $\mathbf{C}$ is reducible – that is as far as any two states can be directly or indirectly connected by the random walk.

2.1.2. Diffusive distances

One considers the edge weights $\mathbf{V}$ as generating a so-called *instantaneous jump process*, permitting to navigate from node i to node j with a rate given by (minus) the components of the *weighted Laplacian* $\mathbf{L} = \boldsymbol{\Pi}^{-\frac{1}{2}}[\mathrm{diag}(\mathbf{C}\mathbf{1}_n) - \mathbf{C}]\boldsymbol{\Pi}^{-\frac{1}{2}}$, where $\boldsymbol{\Pi} = \mathrm{diag}(\mathbf{f})$, and $\mathbf{C}$ must be taken as symmetric, that is replaced if necessary by $\frac{1}{2}(\mathbf{C} + \mathbf{C}^\top)$. Then choose a *diffusion time* $t > 0$, and compute the joint probability $e_{ij}(t)$ to be initially in i and in j at time t (or the other way round) as

$$\mathbf{E}(t) = (e_{ij}(t)) = \boldsymbol{\Pi}^{\frac{1}{2}} \exp(-t\,\mathbf{L})\,\boldsymbol{\Pi}^{\frac{1}{2}}$$

The squared Euclidean *diffusive distances* $\mathbf{D}^{\mathrm{diff}}(t) = (D_{ij}^{\mathrm{diff}}(t))$ finally obtain as [see e.g. 5]

$$D_{ij}^{\mathrm{diff}}(t) = \frac{e_{ii}(t)}{f_i^2} + \frac{e_{jj}(t)}{f_j^2} - 2\frac{e_{ij}(t)}{f_i f_j} \ .$$

2.2. Visualizing a weighted Euclidean configuration by weighted MDS

Weighted multidimensional scaling constitutes the canonical procedure for the low-dimensional visualisation of a *weighted configuration* $(\mathbf{f}, \mathbf{D})$ (see figure 1):

1. define $\boldsymbol{\Pi} = \mathrm{diag}(\mathbf{f})$, as well as the weighted centering matrix $\mathbf{H} = \mathbf{I}_n - \mathbf{1}_n\mathbf{f}^\top$
2. obtain by double centering the matrix of *scalar products* $\mathbf{B} = -\frac{1}{2}\mathbf{H}\,\mathbf{D}\,\mathbf{H}^\top$
3. define the matrix of *weighted scalar products* or *kernel* as $\mathbf{K}$ as:

$$\mathbf{K} = \sqrt{\boldsymbol{\Pi}}\,\mathbf{B}\sqrt{\boldsymbol{\Pi}} \qquad\qquad K_{ij} = \sqrt{f_i f_j}\,B_{ij} \tag{1}$$

4. perform the spectral decomposition $\mathbf{K} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ where $\mathbf{U} = (u_{i\alpha})$ is orthogonal and $\boldsymbol{\Lambda} = \mathrm{diag}(\boldsymbol{\lambda})$ diagonal
5. finally, define $x_{i\alpha} = u_{i\alpha}\sqrt{\lambda_\alpha}/\sqrt{f_i}$, which is the MDS coordinate of node i in dimension α. By construction, $\sum_\alpha (x_{i\alpha} - x_{j\alpha})^2 = D_{ij}$. Also, the total dispersion or inertia of the weighted configuration $(\mathbf{f}, \mathbf{D})$ reads $\Delta = \frac{1}{2}\sum_{i,j=1}^n f_i f_j D_{ij} = \mathrm{tr}(\mathbf{K}) = \sum_{\alpha=1}^{n-1} \lambda_\alpha$.

3. The generalized weighted RV coefficient

At this stage, the book and the movie networks of characters have been expressed into commute time or diffusive kernels (1), namely $\boldsymbol{K}_A$ and $\boldsymbol{K}_B$. Each kernel defines a weighted configuration $(\mathbf{f}_A, \mathbf{D}_A)$, respectively $(\mathbf{f}_B, \mathbf{D}_B)$, and conversely. The *weighted RV coefficient* between both configurations is defined as [1]

$$\mathrm{RV} = \mathrm{RV}_{AB} = \frac{\mathrm{CV}_{AB}}{\sqrt{\mathrm{CV}_{AA}\,\mathrm{CV}_{BB}}} \qquad\qquad \text{where} \quad \mathrm{CV}_{AB} = \mathrm{Trace}(\boldsymbol{K}_A \boldsymbol{K}_B) \tag{2}$$

and constitues a straightforward generalization of the original, unweighted RV coefficient [2], providing the cosine similarity between the vectorized matrices $\mathbf{K}_A$ and $\mathbf{K}_B$ (figure 1).

In particular, $\mathrm{RV}_{AB} \geq 0$ (since $\mathbf{K}_A$ and $\mathbf{K}_B$ are positive semi-definite: this condition is equivalent to the squared Euclidean nature of $\mathbf{D}_A$ and $\mathbf{D}_B$), $\mathrm{RV}_{AB} \leq 1$ (by the Cauchy-Schwarz inequality) and $\mathrm{RV}_{AA} = 1$.
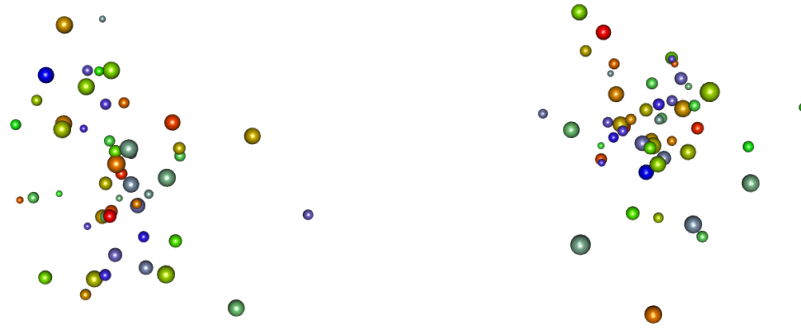
Figure 1: The weighted RV coefficient measures the similarity between two weighted configurations $(\mathbf{f}, \mathbf{D}_A)$ (left) and $(\mathbf{f}, \mathbf{D}_B)$ (right) embedded in $\mathbb{R}^{n-1}$. Here the n objects (characters) are endowed with the *same* weights $\mathbf{f}$ in both configurations, and the object coordinates are obtained from weighted MDS applied on *squared Euclidean dissimilarities* $\mathbf{D}_A$, respectively $\mathbf{D}_B$.

The null distribution of the weighted RV coefficient, i.e. assuming no relationships between the two configurations, and in particular its statistical significance, have been extensively investigated during the last decades [see e.g. 6, 1, and references therein].

The crucial issue here is that the character weights $\mathbf{f}_A$ and $\mathbf{f}_B$ **differ** in the two character networks, whence the naming *generalized weighted RV coefficient* for the same quantity (2), where

$$\mathbf{K}_A = \sqrt{\mathbf{\Pi}_A}\,\mathbf{B}_A\,\sqrt{\mathbf{\Pi}_A} \qquad\qquad \mathbf{\Pi}_A = \mathrm{diag}(\mathbf{f}_A)$$

$$\mathbf{B}_A = -\frac{1}{2}\mathbf{H}_A\,\mathbf{D}_A\,\mathbf{H}_A^\top \qquad\qquad \mathbf{H}_A = \mathbf{I}_n - \mathbf{1}_n\mathbf{f}_A^\top$$

$$\mathbf{K}_B = \sqrt{\mathbf{\Pi}_B}\,\mathbf{B}_B\,\sqrt{\mathbf{\Pi}_B} \qquad\qquad \mathbf{\Pi}_B = \mathrm{diag}(\mathbf{f}_B)$$

$$\mathbf{B}_B = -\frac{1}{2}\mathbf{H}_B\,\mathbf{D}_B\,\mathbf{H}_B^\top \qquad\qquad \mathbf{H}_B = \mathbf{I}_n - \mathbf{1}_n\mathbf{f}_B^\top$$

This circumstance generates new challenging issues, whose investigation was one of the motivations for embarking on the present piece of research.

Note that well-established statistical procedures permitting to test the statistical significance of the weighted coefficient RV_h are available [see e.g. 6, 1, and references therein]. However, testing the generalized weighted RV coefficient is, at the present time, a completely open issue.

3.1. Decomposition of the generalized weighted RV coefficient

As mentioned, the relative weights $\mathbf{f}_A$ and $\mathbf{f}_B$ (set to the uniform weights $1/n$ in most applications of Multivariate Analysis) may *differ* to a spectacular extent: their supports $\mathrm{supp}(\mathbf{f}_A)$ and $\mathrm{supp}(\mathbf{f}_B)$ do not even coincide in general, since version A may contain characters absent in version B, and vice-versa.

Define the *compromise weight* $\mathbf{h}$ as

$$h_i = \frac{\sqrt{f_i^A f_i^B}}{Z} \qquad \text{where} \quad Z = \sum_j \sqrt{f_j^A f_j^B} \in [0, 1] \ .\tag{3}$$

This choice, initially dictated by formal considerations, permitting to further transform the square roots in (1) into a tractable expression, turns out to be conceptually convenient and interpretable as well: Z is a measure of weights dissimilarity, appearing in identity (6) below. Also, $h_i = 0$ unless character i appears in both versions (figure 3). A little algebra demonstrates the numerator of the similarity index (2) to express as

$$\frac{\mathrm{CV}_{AB}}{Z^2} = \mathrm{trace}(\mathbf{K}_{hA}\mathbf{K}_{hB}) + \kappa_{AB}\tag{4}$$

where $\boldsymbol{K}_{hA}$ is the kernel associated to configuration $(\boldsymbol{h}, \boldsymbol{D}_A)$ and $\boldsymbol{K}_{hB}$ is the kernel associated to configuration $(\boldsymbol{h}, \boldsymbol{D}_B)$. Also,

$$\kappa_{AB} = D_{h f_A}^A D_{h f_B}^B + 2\sum_i h_i \ell_i^A \ell_i^B\tag{5}$$

where $D_{h f_A}^A$ is the squared Euclidean distance between the gravity centers of $(\mathbf{h}, \mathbf{D}_A)$ and $(\mathbf{f}_A, \mathbf{D}_A)$. The quantity $D_{h f_B}^B$ is defined analogously. Naturally, the gravity centers of $(\mathbf{f}_A, \mathbf{D}_A)$ and $(\mathbf{h}, \mathbf{D}_A)$ generally differ, as are the gravity centers of $(\mathbf{f}_A, \mathbf{D}_B)$ and $(\mathbf{h}, \mathbf{D}_B)$, but the differences are extremely small in the case study (see Figure 6). The second component in (5) involves a weighted covariance between the $\mathbf{h}$-centered vectors $\boldsymbol{\ell}_A = \mathbf{B}_{hA}\mathbf{f}_A$ and $\boldsymbol{\ell}_B = \mathbf{B}_{hB}\mathbf{f}_B$, where

$$\mathbf{B}_{hA} = -\frac{1}{2}\mathbf{H}_h \, \mathbf{D}_A \, \mathbf{H}_h^\top \qquad \mathbf{B}_{hB} = -\frac{1}{2}\mathbf{H}_h \, \mathbf{D}_B \, \mathbf{H}_h^\top$$

This second component is again zero if the compromise centroid coincides with the original centroid in configuration A, or B, or both. In short, the term κ_{AB} in (5), which can be negative (as here in the two distance variants), represents a correction due the non-coincidence of the $\boldsymbol{f}_A$- and $\boldsymbol{h}$-centroids in configuration A (respectively the $\boldsymbol{f}_B$- and $\boldsymbol{h}$-centroids in configuration B).

3.2. An exact additive decomposition formula

A similarity coefficient such as the generalized weighted coefficient $\mathrm{RV} \in [0, 1]$ can be simply converted into a *dissimilarity* coefficient $d \in [0, \infty)$ by $d = -\ln \mathrm{RV}$. Applying the transformation to (2), taking into account the previous definitions and performing direct, down-to-earth algebraic operations finally yields the following exact decomposition for the dissimilarity between character networks A and B:

$$d_{AB} = \underbrace{-\ln \mathrm{RV}}_{\text{composite dissimilarity}} = \underbrace{-\ln \mathrm{RV}_h}_{\text{adjusted dissimilarity } d_{AB}^h} \underbrace{-2\ln Z}_{\text{dissimilarity between character weights}} \underbrace{-\frac{1}{2}\ln \Gamma_A}_{\text{relative dispersion, book}}$$

$$\underbrace{-\frac{1}{2}\ln \Gamma_B}_{\text{relative dispersion, movie}} \underbrace{-\ln(1+\epsilon)}_{\text{centroid correction}}\tag{6}$$

where

- the "compromise" RV coefficient, RV_h, defined as

$$\text{RV}_h = \frac{\text{trace}(\boldsymbol{K}_{hA}\boldsymbol{K}_{hB})}{\sqrt{\text{trace}(\boldsymbol{K}_{hA}^2)\text{trace}(\boldsymbol{K}_{hB}^2)}} \in [0,1] \tag{7}$$

  which measures the similarity between dissimilarities $\boldsymbol{D}_A$ and $\boldsymbol{D}_B$ in the common compromise weighting $\boldsymbol{h}$.

- $Z \in [0,1]$ in (3) is a measure of similarity between weights $\boldsymbol{f}_A$ and $\boldsymbol{f}_B$, taking on its maximum value $Z = 1$ iff $\boldsymbol{f}_A = \boldsymbol{f}_B$, and its minimum value $Z = 0$ iff the two versions have no character in common.

- $\Gamma_A = \frac{\text{trace}(\boldsymbol{K}_{hA}^2)}{\text{trace}(\boldsymbol{K}_A^2)}$ is a measure of the ratio of the (quartic) dispersion of configuration $\boldsymbol{D}_A$ in the compromise weighting $\boldsymbol{h}$ to the dispersion of $\boldsymbol{D}_A$ in the original weighting $\boldsymbol{f}_A$ (Γ_B is defined analogously).
  $-\frac{1}{2}\ln\Gamma_A > 0$ essentially means that the average contrast between characters (as expressed by $\boldsymbol{D}_A$) is stronger in the original version $\boldsymbol{f}_A$ than in the compromise version $\boldsymbol{h}$, which is in particular likely to occur when "eccentric" characters in version A occur less often in version B.

- the quantity

$$\epsilon = \frac{\kappa_{AB}}{\text{RV}_h\sqrt{\Gamma_A\Gamma_B\text{trace}(\boldsymbol{K}_A^2)\text{trace}(\boldsymbol{K}_B^2)}}$$

  is a normalized measure of the centroid correction occurring in (4). It reflects a "polarization effect" due to centroid change $\bar{\mathbf{x}}_{\mathbf{f}_A} \to \bar{\mathbf{x}}_{\mathbf{h}}$ and $\bar{\mathbf{x}}_{\mathbf{f}_B} \to \bar{\mathbf{x}}_{\mathbf{h}}$, since the overall dispersions $\boldsymbol{D}_A$ and $\boldsymbol{D}_B$ are bound to vary when the reference point is moved to from the centroid configuration. Its magnitude is expected to be small since *main common characters* (i.e. those with large compromise weights $\mathbf{h}$) are precisely the most frequent in both versions A and B.

4. The case study

The Lion, the Witch and the Wardrobe was the second of the seven novels of the *The Chronicles of Narnia*, written by C. S. Lewis in 1950, and adapted into a film directed by A. Adamson released in 2005 (figure 4).

After semi-manual annotation of all named entities throughout the book and the movie script with the module *charnetto* [7], then gathered into groups of aliases, a list of 37 distinct characters were identified:

- 16 characters are common to the book and the movie
- 8 characters occur in the book only
- 13 characters occur in the movie only.

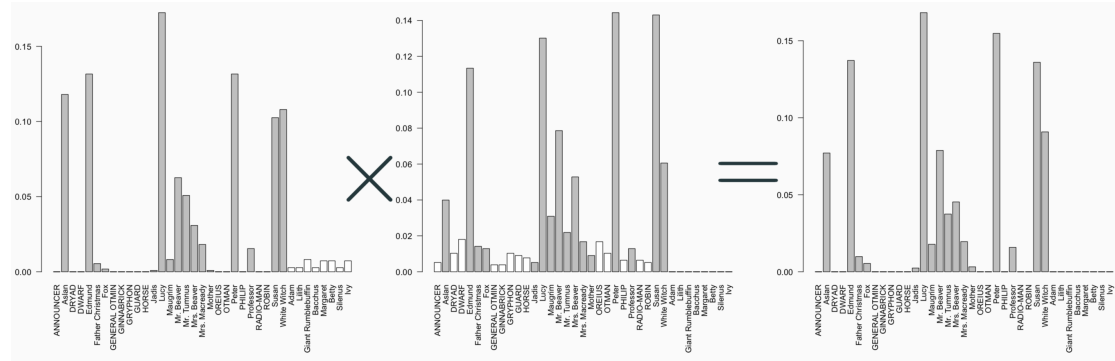**Figure 2:** The two works under study: the book (A) and the movie (B)

Figure 3: Book character weights $\mathbf{f}_A$ (left), movie character weights $\mathbf{f}_B$ (middle) and compromise character weights $\mathbf{h}$ (right), with $h_i = \sqrt{f_i^A f_i^B}/Z$. Characters appearing in both versions are represented by grey bars, and otherwise by white bars.

For each work, we defined the edge weights as the cross-count matrix c_{ij} = "number of co-occurrences of characters i and j within a window of 5 paragraphs" (each paragraph being delimited by a line break), with $c_{ii} = 0$ (see figure 4). Similarly, the character weights were, for a given work, simply defined as $f_i = c_{i\bullet}/c_{\bullet\bullet}$. Figure 4 depicts the corresponding networks.

The cross-count matrices $\mathbf{C}$ permit to compute commute-time distances (section 2.1.1) and diffusive distances (section 2.1.2). Weighted MDS (section 2.2) allows to extract in turn character coordinates, as depicted in figure 6.

In the present study, the centroids of configurations $(\mathbf{f}_A, \mathbf{D}_A)$ and $(\mathbf{f}_B, \mathbf{D}_B)$ are located at the origin by construction, while the first coordinates of the centroids of $(\mathbf{h}, \mathbf{D}_A)$ and $(\mathbf{h}, \mathbf{D}_B)$ are $(\bar{x}_h^A, \bar{y}_h^A) = (0.002, 0.004)$, respectively $(\bar{x}_h^B, \bar{y}_h^B) = (-0.0007, -0.005)$, and fairly close to the origin: $D_{hf_A}^A = 5.6 \cdot 10^{-5}$, respectively $D_{hf_B}^B = 4.0 \cdot 10^{-5}$. As a consequence, the terms κ_{AB} in (5) and $\bar{\varepsilon}$ in (6) are small.
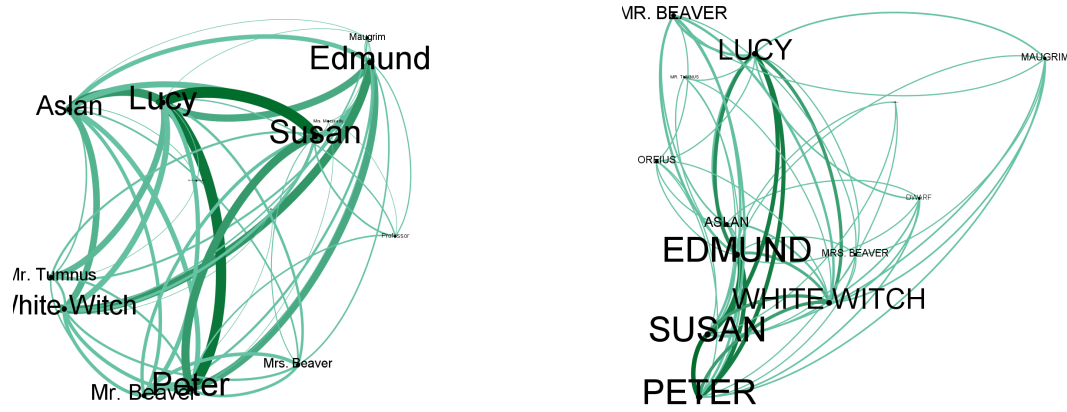
Figure 4: Character networks of the book (left) and movie (right). Edge widths reflect the co-occurrences c_{ij} between nodes, and name sizes the corresponding degree $c_{i\bullet}$.

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]	[,10]	[,11]	[,12]	[,13]	[,14]	[,15]	[,16]	[,17]	[,18]	[,19]	[,20]	[,21]	[,22]	[,23]	[,24]
Adam	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1
Aslan	0	0	0	0	20	0	0	1	0	0	0	28	0	0	0	11	2	4	0	24	0	0	18	22
Bacchus	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0
Betty	0	0	0	0	1	0	0	0	1	0	0	1	1	0	0	0	0	0	1	1	1	0	1	0
Edmund	0	20	0	1	0	0	1	1	1	0	0	28	1	3	1	4	2	4	3	22	2	0	13	38
Father Christmas	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	1	1
Fox	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Giant Rumblebuffin	0	1	0	0	1	0	0	0	0	0	0	3	0	0	0	0	1	0	0	1	0	0	1	1
Ivy	0	0	0	1	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1	1	0	1	0
Jadis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0
Lilith	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
Lucy	0	28	1	1	28	0	0	3	1	0	0	0	1	0	0	12	29	4	2	32	2	1	31	14
Margaret	0	0	0	1	1	0	0	0	1	0	0	1	0	0	0	0	0	0	1	1	1	0	1	0
Maugrim	0	0	0	0	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0	4
Mother	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mr. Beaver	1	11	0	0	4	1	0	0	0	0	1	12	0	0	0	0	3	10	0	11	0	0	5	10
Mr. Tumnus	0	2	1	0	2	0	0	1	0	1	0	29	0	0	0	3	0	2	0	3	0	1	3	8
Mrs. Beaver	0	4	0	0	4	1	0	0	0	0	0	4	0	0	0	10	2	0	0	3	0	0	4	2
Mrs. Macready	0	0	0	1	3	0	0	0	1	0	0	2	1	0	0	0	0	0	0	5	2	0	5	0
Peter	0	24	0	1	22	0	1	1	1	0	0	32	1	2	0	11	3	3	5	0	4	0	21	13
Professor	0	0	0	1	2	0	0	0	1	0	0	2	1	0	0	0	0	0	2	4	0	0	4	0
Silenus	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0
Susan	0	18	0	1	13	1	0	1	1	0	0	31	1	0	0	5	3	4	5	21	4	0	0	4
White Witch	1	22	0	0	38	1	0	1	0	0	1	14	0	4	0	10	8	2	0	13	0	0	4	0

Figure 5: Symmetric cross-count matrix $\mathbf{C} = (c_{ij})$ for the book (column categories are identical to row categories)

The generalized coefficient RV defined in (2) (with differing weights) and the compromise coefficient RV_h defined in (7) turn out to be

$$\mathrm{RV} = 0.113 \qquad \mathrm{RV}_h = 0.391 \qquad \text{(diffusive distance)}$$

$$\mathrm{RV} = 0.531 \qquad \mathrm{RV}_h = 0.611 \qquad \text{(commute-time distance)}$$

In both cases, the magnitude of the term κ_{AB} in (4), is negligible in comparison to trace($\mathbf{K}_{hA}\mathbf{K}_{hB}$). Also, (6) reads here (in order)

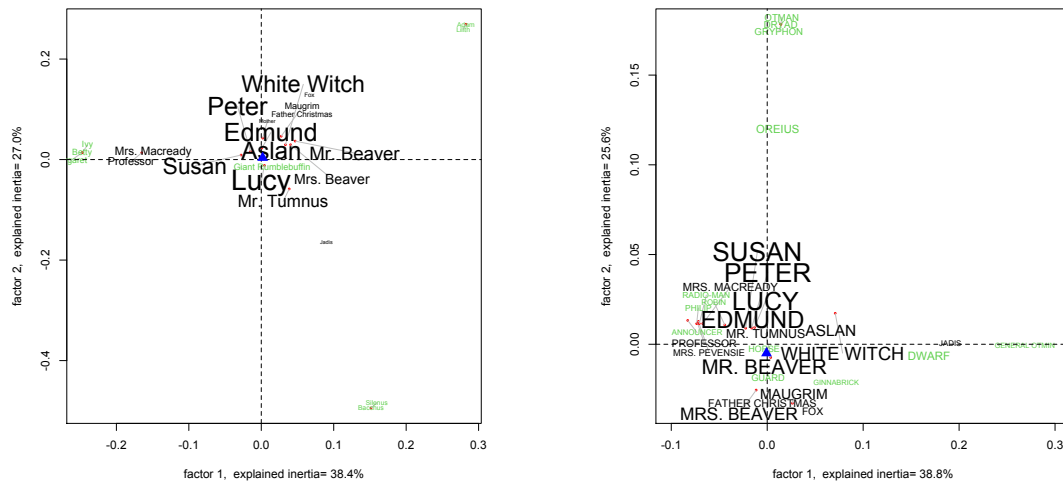$$2.1829 = 0.9385 + 0.2323 + 0.4349 + 0.5762 + 0.0011$$

Figure 6: First MDS coordinates of the characters of the book (left) and movie (right). They have been extracted from the inter-character diffusive distances $\mathbf{D}_A^{\text{diff}}(t)$ for the book (left), respectively $\mathbf{D}_B^{\text{diff}}(t)$ for the movie (right), with a diffusion time arbitrarily set to $t = 10$. Characters in black appear in both works, characters in green in one work only. The blue point depicts the corresponding centroids obtained with the compromise distribution $\mathbf{h}$ (section 3.1).

for the diffusive distances, and

$$0.6333 = 0.4917 + 0.2323 - 0.3431 + 0.2437 + 0.0087$$

for the commute-time distances.

5. Conclusion

Representing the relations between characters of a work as a weighted Euclidean configuration $(\mathbf{f}, \mathbf{D})$ arguably constitutes an instance of *very distant reading*, but not more distant than the usual representation by a weighted network. In both cases, the underlying dyadic formalism (i.e. based upon character pairs) could, and maybe should, be extended to p-adic formalism, taking into account the simultaneous co-occurrences of $p = 0, 1, 2, 3, \ldots$ characters (cliques). Also, the simple co-occurrence relation is in itself particularly rudimentary, yet surprisingly efficient as attested in many applications of Data Analysis, Natural Language Processing and Machine Learning.

On the one hand, we recognize that the mathematical requirements needed to appreciate (or not) the present proposal may distress some amateurs of character networks. Also, a fully convincing literary interpretation of the various terms in decomposition (6) is yet to establish. Furthermore, obtaining a single index (such as RV $= 0.113$) is neither terribly enlightening nor helpful. Comparing *more than two* character networks is more satisfactory, but multiple versions of character networks are alas rare.

On the other hand, quantifying the dissimilarity between two networks cannot ignore mathematical issues, and the proposed formalism permits to propose a procedure which can be made

fully automatic, and yields dissimilarities which can be shown to be metric, namely such that $d_{AB} \leq d_{AC} + d_{CB}$ (triangle inequality) for three versions A, B and C. Also, the exact decomposition permits a detailed, systematically comparable, analysis of sources of (dis)similarities between two character networks.

More generally, the present formalism may contribute to better anchor the study of character networks into mainstream Data Analysis, and draw attention to otherwise overlooked phenomena: for instance, the weights similarity index Z can be related to the Chernoff information occurring in the Neyman-Pearson statistical testing framework [see e.g. 8]; dealing with distinct distributions on the same objects endowed with pair distances evokes Optimal Transportation theory, together with as the possible involving of the *earth mover's distance* [see e.g. 9, 10] in comparing two character networks; finally, the quantities Γ_A and Γ_B, which can exceed or be inferior to one, should be interpreted as indicators of the diversity loss entailed by the disappearance of A-specific characters in the compromise weighting, or, to the contrary, as a diversity gain reflecting distinct emphasis and specificity between two variants A and B. But, as demonstrated in the case study, such a behaviour turns out to depend on the choice of the character dissimilarities $\mathbf{D}$, whose suitability from a more literary perspective should certainly be further investigated in future developments.

References

[1] F. Bavaud, Exact first moments of the RV coefficient by invariant orthogonal integration, Journal of Multivariate Analysis (2023) 105227.

[2] P. Robert, Y. Escoufier, A unifying tool for linear multivariate statistical methods: the RV-coefficient, Journal of the Royal Statistical Society: Series C (Applied Statistics) 25 (1976) 257–265.

[3] J. G. Kemeny, J. L. Snell, Finite Markov chains: with a new appendix "Generalization of a fundamental matrix", Springer, 1983.

[4] M. Saerens, F. Fouss, L. Yen, P. Dupont, The principal components analysis of a graph, and its relationships to spectral clustering, in: European conference on machine learning, Springer, 2004, pp. 371–383.

[5] F. Bavaud, Spatial weights: Constructing weight-compatible exchange matrices from proximity matrices, in: M. Duckham, E. Pebesma, K. Stewart, A. U. Frank (Eds.), Geographic Information Science, Springer International Publishing, Cham, 2014, pp. 81–96.

[6] J. Josse, J. Pagès, F. Husson, Testing the significance of the RV coefficient, Computational Statistics & Data Analysis 53 (2008) 82–91.

[7] C. Métrailler, *charnetto* : a module designed to create an automated character network based on a book or a movie script, 2021. https://pypi.org/project/charnetto/.

[8] T. M. Cover, Elements of information theory, John Wiley & Sons, 1999.

[9] C. Villani, Optimal transport: old and new, volume 338, Springer, 2009.

[10] M. Cuturi, D. Avis, Ground metric learning, The Journal of Machine Learning Research 15 (2014) 533–564.