Development of hierarchy theory for digraphs using concentration theory based on a new type of Lorenz curve

Development of hierarchy theory for digraphs using concentration theory based on a new type of Lorenz curve

by

L. Egghe, LUC, Universitaire Campus, B-3590 Diepenbeek, Belgium[1]
and
UIA, Universiteitsplein 1, B-2610 Wilrijk, Belgium
e-mail : leo.egghe@luc.ac.be

ABSTRACT

In digraphs one has a hierarchy based on the unidirectional order between the vertices of the graph. We present a method of measuring degrees of hierarchy as expressed by the inequality that exists between the vertices' hierarchical numbers. In order to do so, we need to extend the classical Lorenz theory of concentration (curves and measures) for a set of numbers $x_1,...,x_N$ to the case that $\sum_{i=1}^{N} x_i=0$. This is then applied to the set of hierarchical numbers of the vertices of the graph. A graph has a more concentrated hierarchy than another one if the Lorenz curve of the first one is above the Lorenz curve of the second one, hereby expressing that the inequality in domination in the first case is larger than in the second case and that the inequality in subordination in the first case is larger than in the second case. We also determine maximal and minimal Lorenz curves in this setting and characterise the graphs that yield these curves. Based

---

[1]Permanent address

Keywords and phrases : digraph, hierarchy, Lorenz, concentration theory.

on this theory, we also determine good measures of hierarchical concentration in graphs. Applications can be given in the study of organigrams in companies and administrations and in citation analysis.

I. Introduction

Consider a general digraph (directed graph) G in which there are no loops. let the number of vertices be $N \in \mathbb{N}$. We also suppose that the graph is weakly connected, i.e. that the underlying undirected graph of G consists of one component (see Wilson (1972)). In fact if this is not the case, we can apply the results from this paper to the different components of the graph. We will number the vertices by denoting them as i, i=1,...,N. For each vertex $i \in \{1,...,N\}$ we can consider all chains that have i as the starting point. Their lengths are an indication of the role of vertex i in the graph from the point of view of "domination". In terms of a graph representing an organigram in a company or an administration, they indicate in what way vertex (person) i is a direct or indirect boss of other vertices (persons) j. Conversely we can consider all chains that have i as an endpoint. Now their lengths are an indication of the role of vertex i in the graph from the point of view of "subordination". In the same example as above, they indicate in what way this person has direct or indirect bosses. The two together indicate the general hierarchical position of vertex i in the graph G. Let us denote, for every $i \in \{1,...,N\}$

$$\sigma_i^+ = \text{the sum of the lengths of the chains that start in i} \tag{1}$$

$$\sigma_i^- = \text{the sum of the lengths of the chains that end in i} \tag{2}$$

We will denote

$$\sigma_i = \sigma_i^+ - \sigma_i^- \tag{3}$$

and call this the hierarchical number (or degree) of i in graph G. Note that it is obvious that

$$\sum_{i=1}^{N} \sigma_i = 0 \tag{4}$$

Examples (N=4)

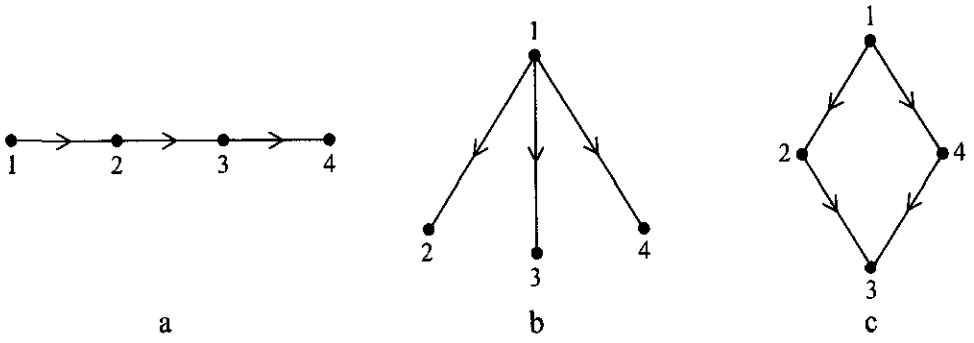

Fig. 1 Three examples, N=4

In case of the chain (Fig. 1a) we have $\sigma_1^+ = 6$, $\sigma_1^- = 0$, $\sigma_2^+ = 3$, $\sigma_2^- = 1$, $\sigma_3^+ = 1$, $\sigma_3^- = 3$, $\sigma_4^+ = 0$, $\sigma_4^- = 6$. Hence $\sigma_1 = 6$, $\sigma_2 = 2$, $\sigma_3 = -2$, $\sigma_4 = -6$.

In case of the graph in Fig. 1b we have $\sigma_1^+ = 3$, $\sigma_1^- = 0$, $\sigma_i^+ = 0$, $\sigma_i^- = 1$ (i=1,2,3,4). Hence $\sigma_1 = 3$, $\sigma_i = -1$ (i=1,2,3).

In case of the graph in Fig. 1c we have $\sigma_1^+ = 6$, $\sigma_1^- = 0$, $\sigma_2^+ = \sigma_4^+ = 1$, $\sigma_2^- = \sigma_4^- = 1$, $\sigma_3^+ = 0$, $\sigma_3^- = 6$. Hence $\sigma_1 = 6$, $\sigma_2 = \sigma_4 = 0$, $\sigma_3 = -6$.

So, for each graph G as described above, we have a family of numbers of hierarchy $\sigma_1,\ldots,\sigma_N$, representing degrees of domination and of subordination. The examples in Figs. 1a and 1c show that the domination degrees have (apart from the - sign) the same pattern as the subordination degrees, but the example in Fig. 1b shows that this is not always the case : here vertex 1 dominates but the subordination degrees of 2, 3 and 4 are the same. In the sequel we want to study the inequality of the numbers $\sigma_1,\ldots,\sigma_N$ in its totality : inequality in domination and subordination.

The study of inequality (also called concentration) goes back to the beginning of the twentieth century when it was used to measure social inequality, as e.g. expressed by the income inequality in a social group. We mention Muirhead (1903), Lorenz (1905), Gini (1909), Dalton (1920), Shannon (1948), Theil (1967), Atkinson (1970), Allison (1978) as some historical papers amongst the many other ones. Of course, in Shannon (1948) one emphasizes more on similarity (being opposite to concentration) as one also does in biometry where one uses the

term diversity, see e.g. Rousseau and Van Hecke (1999). The basics of concentration theory can be summarized as follows. Let $X=(x_1,...,x_N)$ be a vector of positive numbers, incl. zero (but not all of them zero). We will always arrange the x_i decreasingly, although an equivalent theory can be given for the increasing order (see Marshall and Olkin (1979)).

Define, for each i=1,...,N

$$a_i = \frac{x_i}{\sum_{k=1}^{N} x_k} \tag{5}$$

Note that

$$\sum_{k=1}^{N} x_k > 0 \tag{6}$$

The Lorenz curve L_X of X is the polygonal line connecting (0,0) with $(\frac{1}{N}, a_1)$, then connecting this point with $(\frac{2}{N}, a_1+a_2)$, and so on, hence connecting

$$\left(\frac{i}{N} , \sum_{j=1}^{i} a_j \right) \tag{7}$$

i=1,...,N. Note that for i=N this point is (1,1). See Fig. 2 for an example.

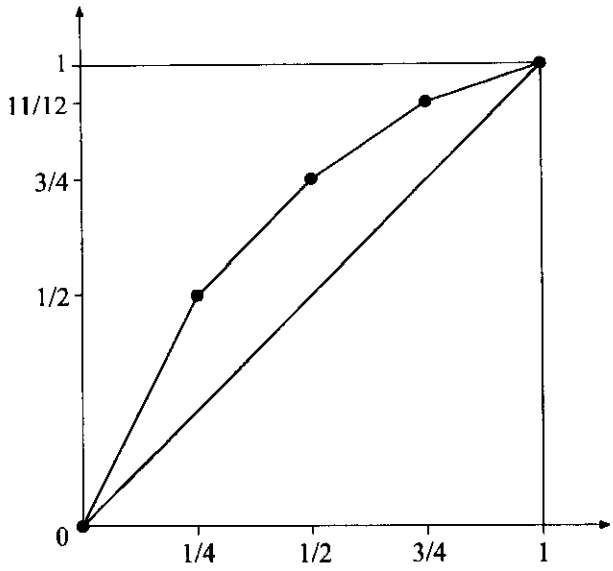

Fig. 2 Lorenz curve L_X for X=(6,3,2,1) connecting (0,0), $(\frac{1}{4},\frac{1}{2})$, $(\frac{1}{2},\frac{3}{4})$, $(\frac{3}{4},\frac{11}{12})$ and (1,1).

The diagonal of the unit square, connecting (0,0) and (1,1) represents the Lorenz curve for the vector X=(x,x,...,x) (x>0), the least concentrated situation since all x_i are equal. This is the lowest Lorenz curve that is possible. The highest possible Lorenz curve (for fixed N∈ℕ) is the one of X=(1,0,...,0) connecting (0,0) with $(\frac{1}{N},1)$ and then with (1,1).

Let $X=(x_1,...,x_N)$ and $X'=(x'_1,...,x'_N)$ be two vectors. We say that X′ is larger than X in the Lorenz sense, denoted X--<X′ if $L_X \leq L_{X'}$. Unless X=X′, X′ is then more concentrated than X (as can be seen by applying elementary transfers - we do not go into this and we refer the reader to Marshall and Olkin (1979) or to Egghe and Rousseau (1990)). The "degree" of concentration can be measured by using good measures of concentration C, i.e. measures that respect the Lorenz order --<. In other words, this are measures satisfying X--<X′ and X≠X′ ⇒ C(X)<C(X′). Examples of good measures of concentration abound (see the references given above). We only give three examples. The coefficient of variation

$$V = \frac{\sigma}{\mu} \tag{8}$$

(also V^2 can be used), where σ and μ are the standard deviation and the average of the vector X. Note that V^2 is the normalized version of $\sum_{i=1}^{N} a_i^2$, where a_i is given by (5). Another measure is the one of Theil :

$$Th = \frac{1}{N} \sum_{i=1}^{N} \left(\frac{x_i}{\mu} \right) \ln \left(\frac{x_i}{\mu} \right) \tag{9}$$

which is nothing else than the "concentration" version of the diversity measure entropy (Shannon (1948)). Finally the area between the Lorenz curve and the diagonal connecting (0,0) and (1,1) is also obviously a good concentration measure. The normalized version of it is nothing else than the famous Gini index, used in econometrics (Gini (1909)).

It is this type of concentration theory that we will introduce in graph theory in order to measure the concentration (inequality) in domination and subordination. However, here, the numbers σ_i can also be negative and they even add up to zero (4), contradicting (6) and making the construction of the Lorenz curve of $(\sigma_1,...,\sigma_N)$ impossible because of (5). In the next section we will extend the theory of concentration to vectors $X=(x_1,...,x_N)$ where some x_i can be negative, including the case $\sum_{k=1}^{N} x_k=0$. Maximal and minimal Lorenz curves will be determinated and good measures of concentration will be given. The third section will apply this concentration theory to the vector $(\sigma_1,...,\sigma_N)$ of hierarchical degrees of the vertices of a graph. This represents the way to measure the inequality in domination and subordination at the same time. The maximal and minimal Lorenz curves will be characterized by the graphs (for general $N \in \mathbb{N}$) that yield these extreme Lorenz curves. We will also compare this theory with some existing "measures of hierarchy" which are - in the author's opinion - too weak to describe hierarchy in a graph. Examples of application are given. In section four we illustrate the results on a general chain of N vertices.

II. Lorenz concentration theory for general vectors $X=(x_1,\ldots,x_N)$

We will begin with the case that

$$\sum_{k=1}^{N} x_k > 0 \tag{10}$$

(but some x_i can be negative). If $\sum_{k=1}^{N} x_k<0$, this model can still be used by applying it to the vector $-X=(-x_1,\ldots,-x_N)$. The case $\sum_{k=1}^{N} x_k=0$ will be handled in subsection II.2.

II.1 Concentration theory in case $\sum_{k=1}^{N} x_k > 0$.

This case is very simple. In fact we act exactly as in the case that all x_i are positive, by applying (5). We again obtain a Lorenz curve connecting (0,0) with (1,1) but now the curve can leave the unit square. Fig. 3 gives an example

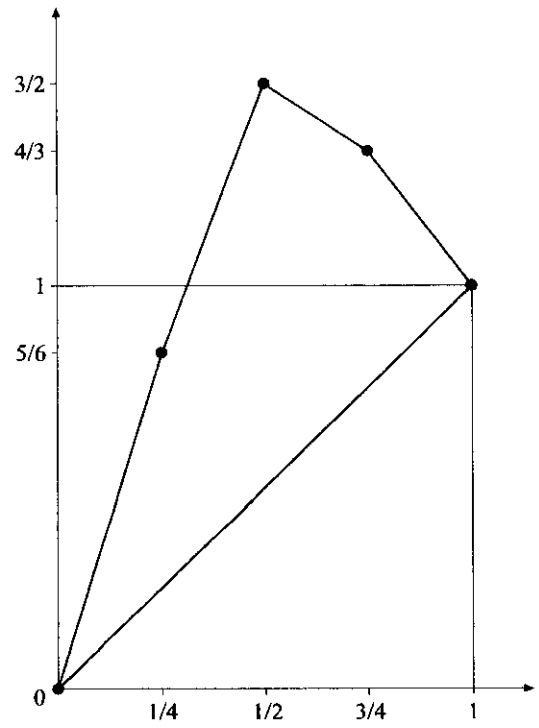

Fig. 3 Lorenz curve L_X for X=(5,4,-1,-2) connecting
(0,0), $(\frac{1}{4},\frac{5}{6})$, $(\frac{1}{2},\frac{3}{2})$, $(\frac{3}{4},\frac{4}{3})$, (1,1)

Let $X=(x_1,...,x_N)$, $X'=(x'_1,...,x'_N)$ be two vectors such that $\sum_{k=1}^{N} x_k \neq 0$, $\sum_{k=1}^{N} x'_k \neq 0$. Upon multiplication by -1 (possibly) we can assume $\sum_{k=1}^{N} x_k > 0$, $\sum_{k=1}^{N} x'_k > 0$ and we also suppose X and X′ to be decreasing. We say that X′ is larger than X in the Lorenz sense, denoted $X \prec X'$ if $L_X \leq L_{X'}$ and we say that X′ is more concentrated than X (unless X=X′). Because of this construction the following theorem of Hardy, Littlewood and Pólya (1929,1952) - see also Marshall and Olkin (1979) applies.

Theorem II.1 (Hardy, Littlewood and Pólya)

If $X=(x_1,...,x_N) \prec X'=(x'_1,...,x'_N)$ and if they are decreasing, then

$$\sum_{k=1}^{N} \varphi(a_k) \leq \sum_{k=1}^{N} \varphi(a'_k) \tag{11}$$

(and < if $X \neq X'$) for all continuous convex functions φ. Here

$$a_i = \frac{x_i}{\sum_{j=1}^{N} x_j} \tag{12}$$

$$a'_i = \frac{x'_i}{\sum_{j=1}^{N} x'_j} \tag{13}$$

Since e.g. $\varphi(x)=x^2$ is continuous and convex we are able to provide a good measure of concentration for our model :

$$\sum_{k=1}^{N} a_k^2 \tag{14}$$

This measure can then be normalized, if necessary. Of course, as in the classical case, the measure = area between L_X and the diagonal connecting (0,0) and (1,1), is also a good measure of concentration, hereby generalizing Gini's index.

Applications of this theory can be the measurement of inequality (fluctuations) of the temperature in a certain area over a certain time period (e.g. a year), which then can be compared with the same in another area. We will not go into this since our main goal is the study of the hierarchy of digraphs.

II.2 Concentration theory in case $\sum_{k=1}^{N} x_k = 0$.

Of course, we assume that not all x_i are zero and that the x_i are decreasing. Denote

$$I_+ = \{i \in \{1,...,N\} \| x_i > 0\} \tag{15}$$

$$I_- = \{i \in \{1,...,N\} \| x_i < 0\} \tag{16}$$

Hence

$$\sum_{k=1}^{N} x_k = \sum_{i \in I_+} x_i + \sum_{i \in I_-} x_i = 0 \tag{17}$$

so

$$\sum\nolimits_+ = : \sum_{i \in I_+} x_i = -\sum_{i \in I_-} x_i \tag{18}$$

Equation (18) enables us to develop a concentration theory (or inequality theory) for vectors X for which the coordinates add up to zero, hereby studying the concentration in $(x_i)_{i \in I_+}$ as well as the concentration in $(x_i)_{i \in I_-}$. We proceed as follows. Instead of (5) we calculate

$$\alpha_i = \frac{x_i}{\sum\nolimits_+} \tag{19}$$

for all i=1,...,N. Because of (18) we have

$$\sum_{i\in I_+} \alpha_i = 1 \tag{20}$$

$$\sum_{k=1}^{N} \alpha_k = 0 \tag{21}$$

We now form the polygonal curve connecting (0,0) with $\left(\frac{1}{N},\alpha_1\right)$, connecting $\left(\frac{1}{N},\alpha_1\right)$ with $\left(\frac{2}{N},\alpha_1+\alpha_2\right)$ and so on. Because of the above this curve goes from (0,0) to (x,1), where

$$x = \frac{|I_+|}{N}, \tag{22}$$

then from (x,1) to (y,1), where

$$y = \frac{N-|I_-|}{N} \tag{23}$$

and from (y,1) to (1,0) via the points $\left(\frac{i}{N}, \sum_{k=1}^{i} \alpha_k\right)$ where $i\in I_-$. $x\le y$ since $|I_-|+|I_+|\le N$. If no x_i is zero, then x=y since $|I_-|+|I_+|=N$ then. An example is given in Fig. 4.

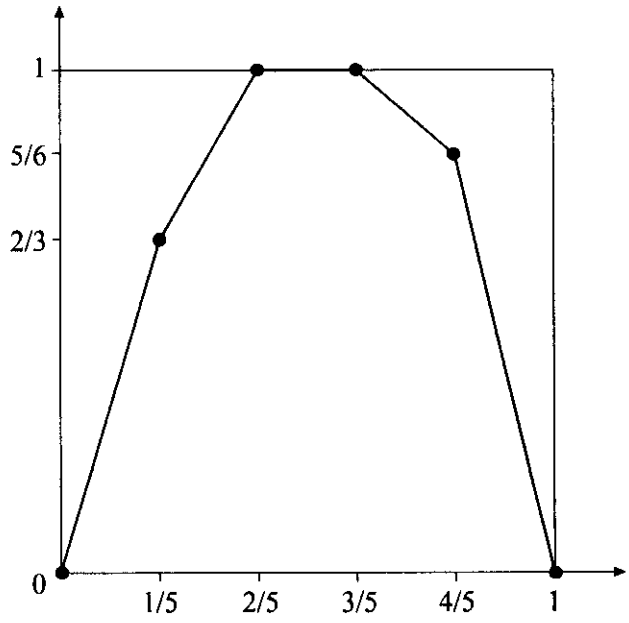

Fig. 4 Lorenz curve L_X for X=(4,2,0,-1,-5) connecting (0,0), $(\frac{1}{5},\frac{2}{3})$, $(\frac{2}{5},1)$, $(\frac{3}{5},1)$, $(\frac{4}{5},\frac{5}{6})$, (1,0).

Intuitively speaking, L_X consists of a "Lorenz curve" for the $(x_i)_{i\in I_+}$ (from (0,0) to (x,1)) and of a "Lorenz curve" for the $(x_i)_{i\in I_-}$ (from (y,1) to (1,0) and mirrored over the vertical line with abscissa y). This is why we have here a method of measuring the concentration in the $(x_i)_{i\in I_+}$ as well as the one in the $(x_i)_{i\in I_-}$. The global degree of inequality can then be compared with the same for another vector.

Let, indeed, be $X=(x_1,...,x_N)$ and $X'=(x'_1,...,x'_N)$ be two decreasing vectors such that

$$\sum_{k=1}^{N} x_k = \sum_{k=1}^{N} x'_k = 0 \tag{24}$$

We say that X' is larger than X in the Lorenz sense, denoted $X--<X'$ if $L_X \leq L_{X'}$. If $X \neq X'$, then X' represents a more concentrated situation in both the positive and negative values. This will enable us, in applications (see next section) to measure the hierarchical degree (both in domination and subordination as one system) in a digraph, with obvious practical applications. Using again, the theorem of Hardy, Littlewood and Pólya (which is also valid for the order --< here), we arrive at the following good measures of concentration in this case

$$\sum_{k=1}^{N} \left(\frac{x_k}{\sum_+} \right)^2 \tag{25}$$

and

$$\sum_{k=1}^{N} \left| \frac{x_k}{\sum_+} \right| \ln \left| \frac{x_k}{\sum_+} \right|. \tag{26}$$

Again, the area between L_X and the x-axis is also a good concentration measure.

Note

We repeat that the present concentration theory measures (at the same time) the concentration of $(x_i)_{i\in I_+}$ and of $(x_i)_{i\in I_-}$. It does not measure the concentration of $X=(x_1,...,x_N)$ itself. This is illustrated by the following example. Let $X=(2,1,-1,-2)$ and $X'=(2,2,-2,-2)$. Although X' looks more "concentrated" than X (in an intuitive way of speaking), the inequality in the positive as well as in the negative coordinates of X' is smaller than in the comparable coordinates of X.

This is also verified by using e.g. (25) :

$$\sum_{k=1}^{4}\left(\frac{x_k}{\sum_{+}}\right)^2 = 2\left(\left(\frac{2}{3}\right)^2+\left(\frac{1}{3}\right)^2\right) = \frac{10}{9}$$

$$\sum_{k=1}^{4}\left(\frac{x'_k}{\sum'_{+}}\right)^2 = 2\left(\left(\frac{2}{4}\right)^2+\left(\frac{2}{4}\right)^2\right) = 1$$

This clearly illustrates the value of concentration theory for vectors $X=(x_1,\ldots,x_N)$ for which $\sum_{k=1}^{N} x_k=0$.

The highest possible Lorenz curve in this setting (and $N\in\mathbb{N}$ fixed) is (obviously) the one connecting (0,0) with $\left(\frac{1}{N},1\right)$, $\left(\frac{1}{N},1\right)$ with $\left(\frac{N-1}{N},1\right)$ and the latter point with (1,0). It is obtained for

$$X = (\underbrace{x,0,\ldots,0}_{N-2},-x)$$

where $x>0$, and only for this type of vector. There is no lowest possible Lorenz curve but we can characterise all minimal Lorenz curves that are possible, i.e. curves L for which there does not exist another Lorenz curve L′ such that $L'<L$. The following theorem can be proved.

Theorem II.2.

Let $N\in\mathbb{N}$ be fixed. Let $X=(x_1,\ldots,x_N)$ be decreasing with $\sum_{k=1}^{N} x_k=0$. Then L_X is minimal iff L_X consists of 2 straight lines : one connecting (0,0) with $\left(\frac{i}{N},1\right)$ and one connecting $\left(\frac{i}{N},1\right)$ with (1,0), (i=1,...,N-1). Such a curve is obtained for

$$X = (\underbrace{x,\ldots,x}_{i},\underbrace{-y,\ldots,-y}_{N-i})$$

(and only for this type of vector) where $x,y \neq 0$ and where $ix=(N-i)y$. Hence there are exactly N-1 minimal curves. No minimal curve is lowest, except if N=2.

Proof :

Every Lorenz curve L_X has at least one point $\frac{i}{N}$(i=1,...,N-1) where the ordinate is one (by construction). Since also (0,0) and (1,0) belong to any L_X, the lowest Lorenz curve, given $\left\{\left(\frac{i}{N},1\right), \left(\frac{i}{N},1\right)\right\}$, that is possible is the one consisting of the straight lines connecting (0,0) with $\left(\frac{i}{N},1\right)$ and the one connecting $\left(\frac{i}{N},1\right)$ with (1,0). So every minimal Lorenz curve is of this type and, obviously, every curve of this type is minimal.They all exist, given N, since the above curve obtains for

$$X = (\underbrace{x,...,x}_{i},\underbrace{-y,...,-y}_{N-i})$$

where $ix=(N-i)y$ which is clear from the construction (L_X connects (0,0) linearly with $\left(\frac{i}{N},\frac{ix}{\sum_+}\right) = \left(\frac{i}{N},1\right)$ and connects the latter point with (1,0) because $ix=(N-i)y$. Hence since all these Lorenz curves exist, they intersect each other and, hence, no minimal curve can be the lowest except if N=2, since then there is only one minimal curve, the one connecting (0,0) linearly with $\left(\frac{1}{2},1\right)$ and connecting $\left(\frac{1}{2},1\right)$ with (1,0). □

Corollary II.3.

Let A be the good concentration measure giving the area under the Lorenz curve. Then $A=\frac{1}{2}$ for every minimal Lorenz curve and $A=1-\frac{1}{N}$ for the highest Lorenz curve. Hence $\lim_{N\to\infty} A=1$ for the highest Lorenz curves.

So a normalization of this measure is given by 2A-1.

The proof is trivial.

We will now apply this concentration theory to the study of hierarchy in digraphs, characterise graphs that yield the highest and the minimal Lorenz curves and interpret the results in terms of hierarchy in companies or in the administration and in terms of hierarchy in citation analysis.

III. Hierarchy theory for digraphs

Combining the results of the two previous sections we can describe the hierarchy theory for digraphs (without loops and weakly connected) as follows. Let $N\in\mathbb{N}$ and consider any graph (as described above) with N vertices. For each vertex $i\in\{1,...,N\}$ we consider all chains that start in i and all chains that end in i. They are defined as follows. [i,j] is a chain that starts in i if (i,j) is an edge of the graph or if there exist $k_1,....,k_m \in \{1,...,N\}$ such that $(i, k_1),(k_1, k_2),....,(k_{m-1}, k_m),(k_m, j)$ are edges of the graph (in that order of course). In the same way we can define a chain [j,i] that ends in i. Note that, in the above notation, all $[i,k_\ell]$, $\ell\in\{1,...,m\}$ are also chains that start in i (the same for chains ending in i). The length of a chain is the number of consecutive edges it contains (e.g. the length of [i,j] above is 1 if (i,j) is an edge of the graph or is m+1 in the other case, using the same notation). Note that, given $i,j\in\{1,...,N\}$ the chain [i,j] (if it exists) is not always unique. Example : consider a (part) of a graph as shown in Fig. 5.

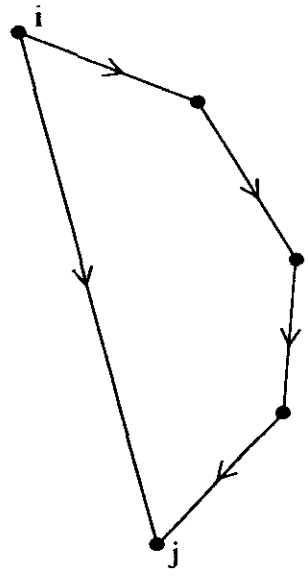

Fig. 5 Existence of two chains between i and j.

Here we have a chain of length 4 between i and j and one of length 1 between i and j. In our hierarchy theory, we will use all these chains since, e.g. only considering the one of length one in Fig. 5 (corresponding to d(i,j)=1, the distance between i and j) does not reveal that j is subordinated in 2 different ways w.r.t. i (or otherwise said, i is dominating j in 2 different ways), which should be taken into account in hierarchy theory : Fig. 5 is completely different (e.g. as an organigram in an administration) than the simple direct relation from i to j.

We repeat now the construction of the hierarchy vector $X=(\sigma_1,...,\sigma_N)$. For all $i\in\{1,...,N\}$:

$$\sigma_i^+ = \text{the sum of the lengths of all the chains that start in i} \tag{27}$$

$$\sigma_i^- = \text{the sum of the lengths of all the chains that end in i} \tag{28}$$

The hierarchical number (or degree) of i is then

$$\sigma_i = \sigma_i^+ - \sigma_i^- \tag{29}$$

Note (as explained above) that this definition is different from the one in which we only use the distances d(i,j) (from i) or d(j,i) (to i), for the reasons given. With the vector $X=(\sigma_1,...,\sigma_N)$ one can then apply the concentration models as explained in section II (since $\sum_{i=1}^{N} \sigma_i=0$). Let us consider, as an example, some cases with N=4. Considering all cases is virtually impossible since it is easy to derive from R.C. Reid (1997), section 2.9 that there are 209 weakly connected digraphs without loops for N=4. To the three examples in Fig. 1a,b,c, we add the ones in Fig. 6.

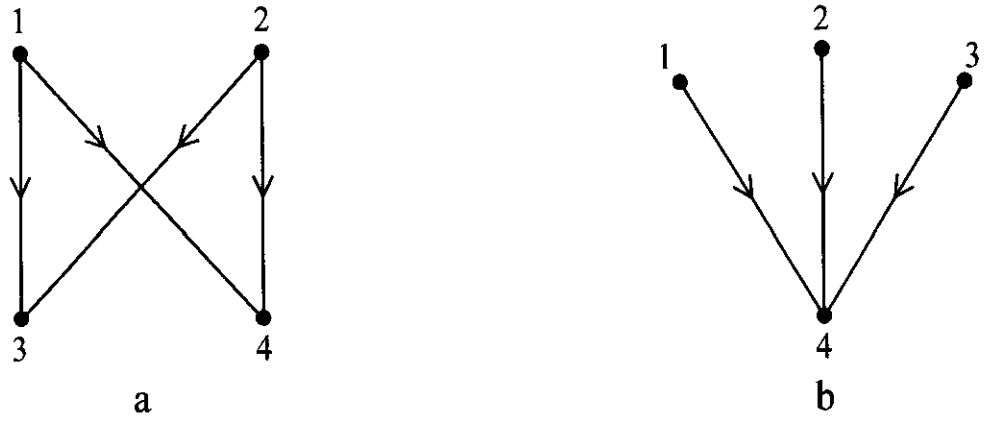

Fig. 6 Two more examples, N=4

The hierarchical situation of each of these graphs is given by the Lorenz curves in Fig. 7.

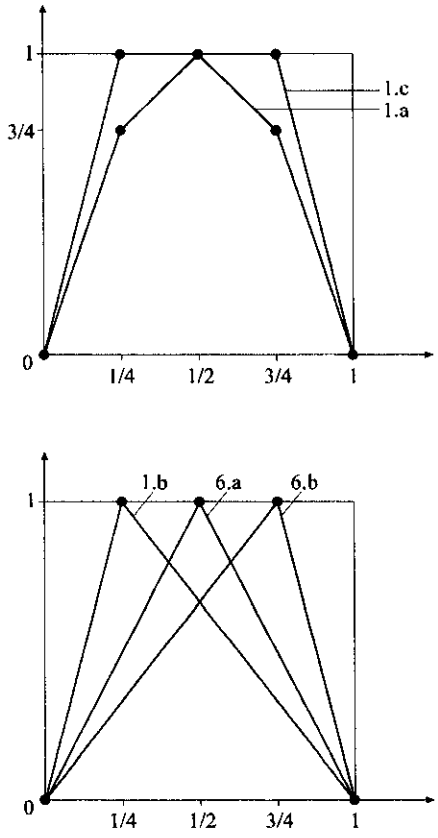

Fig. 7 L_X for the hierarchy of the graphs in Figs. 1a, 1b, 1c, 6a, 6b.

Fig. 1.c represents the highest Lorenz curve. Here the highest possible inequality exists in the sense of domination and subordination : X=(6,0,0,-6). The Lorenz curves of Figs. 1.b, 6.a, 6.b represent the 3 minimal Lorenz curves that are possible in this case (cf. Theorem II.2), resp. for the vectors X=(3,-1,-1,-1), X′=(2,2,-2,-2), X″=(1,1,1,-3). Note that all these cases represent equality in domination as well as equality in subordination. Finally Fig. 1.a is the chain, where domination and subordination are not extreme, which is intuitively obvious. The above example shows that the largest Lorenz curve as well as the minimal ones are all realized by existing graphs.

We will now give a characterization, for general $N\in\mathbb{N}$, of graphs that give the largest Lorenz curve as well as the minimal Lorenz curves. We will furthermore show that they are all realized by existing graphs, for all $N\in\mathbb{N}$. We first state and prove a simple but crucial lemma.

Lemma III.1. Suppose that 1 and 2 are two vertices in a general weakly connected digraph without loops and with any number N of vertices in total. Suppose that (1,2) belongs to the edges of this graph. Then

$$\sigma_1 > \sigma_2 \tag{30}$$

Proof : Using definitions (27) and (28), it is clear that $\sigma_1^+ > \sigma_2^+$ since every chain that starts in 2 is part of a chain that starts in 1. Also $\sigma_1^- < \sigma_2^-$ since every chain that ends in 1 is part of a chain that ends in 2. Hence

$$\sigma_1 = \sigma_1^+ - \sigma_1^- > \sigma_2^+ - \sigma_2^- = \sigma_2 \qquad \square$$

Theorem III.1 **Characterization of weakly connected digraphs without loops yielding the largest Lorenz curve.**

For every fixed $N\in\mathbb{N}$, the largest Lorenz curve is only obtained for the graph (upon a permutation of the vertices) in which only vertex 1 has a positive hierarchical degree (namely N-1), only vertex N has a negative hierarchical degree (namely 1-N) and where the vertices 2,...,N-1 have a zero hierarchical degree. Moreover (1,2),...,(1,N-1),(2,N),...,(N-1,N) and possibly (1,N) are the only edges of the graph.

Summarizing, we have the graph of Fig. 8.

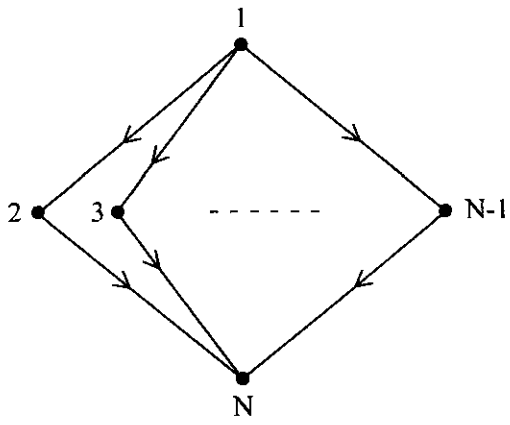

Fig. 8 Graph yielding the largest hierarchical Lorenz curve

Proof : In the previous section it was found that the maximal hierarchical Lorenz curve is obtained only for the vector

$$X = (x, \underbrace{0,\ldots,0}_{N-2}, -x)$$

, where $x>0$. Since $\sigma_2 = \ldots = \sigma_{N-1} = 0$ we have that no edge between the vertices 2,...,N-1 exists, by the lemma. Since the graph is weakly connected, the vertices 2,...,N-1 are linked to 1 or to N or both. Only the last possibility is valid since $\sigma_2 = \ldots = \sigma_{N-1} = 0$. Suppose there exist $i,j \in \{2,\ldots,N-1\}$, $i \neq j$ such that (1,i) and (N,j) are edges. Then since $\sigma_i = \sigma_j = 0$ we have that also (i,N) and (j,1) are edges. But this implies the existence of the loop $1 \to i \to N \to j \to 1$, which is excluded. So, only the edges (1,2),...,(1,N-1),(2,N),...,(N-1,N) and possibly (1,N) exist apart from an interchange of the vertices 1 and N, concluding the proof (yielding also that $x=3(N-2)$ or $x=3(N-2)+1$). □

Theorem III.2 **Characterization of weakly connected digraphs without loops yielding the minimal Lorenz curves**.

For the typical minimal Lorenz curve L_X of Theorem II.2 consisting of two straight lines : one connecting (0,0) with $(\frac{i}{N},1)$ and one connecting $(\frac{i}{N},1)$ with (1,0), we have the following characterization of graphs of N vertices that yield L_X as their hierarchical Lorenz curve : upon a permutation of {1,...,N}, we have that only the vertices 1,...,i have a positive equal hierarchical degree x and that the vertices i+1,...,N have a negative equal hierarchical degree y and the relation between x and y is : ix = (N-i)y. No edges between the vertices 1,...,i exist and no edges between the vertices i+1,...,N exist. The only edges that exist are from a vertex in {1,...,i} to one in {i+1,...,N}, in that order. The case that every vertex in {1,...i} is connected to every vertex in {i+1,...,N} is always a solution. This solution is unique iff ℓ.c.d(i,N-i)=1.

Proof :

The given minimal Lorenz curve necessarily has i vertices (say 1,...,i) of equal positive hierarchical degree x and N-i vertices (say i+1,...,N) of equal negative hierarchical degree y, since L_X is constructed from

$$X=(\underbrace{x,...,x}_{i},\underbrace{-y,...,-y}_{N-i})$$

(theorem II.2). Since these coordinates add up to zero, we have that ix=(N-i)y. No edges between the vertices 1,...,i exist, otherwise their hierarchical degree would be different, by the lemma. The same goes for the vertices i+1,...,N. So the only edges that can exist is between 1,...,i and i+1,...,N. Suppose there would exist an edge (k,ℓ) and (ℓ,k') with $k,k' \in \{1,...,i\}$ and $\ell \in \{i+1,...,N\}$. Applying the lemma twice yields $\sigma_k > \sigma_{k'}$, contradicting what we have. Hence only edges between 1,...,i and i+1,...,N, in that order, exist. The number of solutions is determined by the equation ix=(N-i)y. It is immediately clear that x=N-i, y=i always is a solution. It represents the case where every vertex in {1,...,i} is connected with every vertex in {i+1,...,N}. We now show that this solution is the unique one iff ℓ.c.d.(i,N-i)=1. Indeed, since $y \in \{1,...,i\}$ and $x \in \{1,...,N-i\}$ necessarily, any other solution than the one above satisfies y<i. Since ix=(N-i)y, we have that there exists at least one prime divisor (>1) of i that is also a divisor of N-i.

Hence ℓ.c.d(i,N-i)>1. Conversely, let the ℓ.c.d(i,N-i)>1. Hence there exists a prime divisor q>1 such that N-i=qk_1 and i=qk_2 ($k_1,k_2\in\mathbb{N}$). Hence, because ix=(N-i)y, also x=(q-ℓ)k_1, y=(q-ℓ)k_2 (ℓ=1,...,q-1) are solutions, if x,y≠1 and they are all realisable : connect vertex 1 with the vertices i+1,...,(i+(q-ℓ)k_1) (mod(N-i+1)), vertex 2 with vertices (i+(q-ℓ)k_1+1) (mod(N-i+1)),..., (i+2(q-ℓ)k_1) (mod(N-i+1)) and so on until : connect vertex i with vertices (i+(i-1)(q-ℓ)k_1+1) (mod(N-i+1)),...,(i+i(q-ℓ)k_1) (mod(N-i+1)). Here a (mod (N-i+1)) denotes the rest of the division of a by N-i+1. In total, the vertices i+1,...,N receive ix=i(q-ℓ)k_1=(N-i)(q-ℓ)k_2=(N-i)y links and each of these vectors have the same (negative) hierarchical degree (being -(q-ℓ)k_2) by construction and since (N-i)(q-ℓ)k_2 is an (N-i) multiple. Of course, each of the 1,...,i vertices also have an equal positive hierarchical degree (being (q-ℓ)k_1), also by construction. This ends the proof of this theorem. □

The construction of nonunique minimal curves, as discussed in the above proof, is illustrated by the next example.

Example : The two graphs of Fig. 9 yield the same minimal Lorenz curve (N=6, i=3). The freedom is expressed by ix=(N-i)y, as given in Theorem II.2.

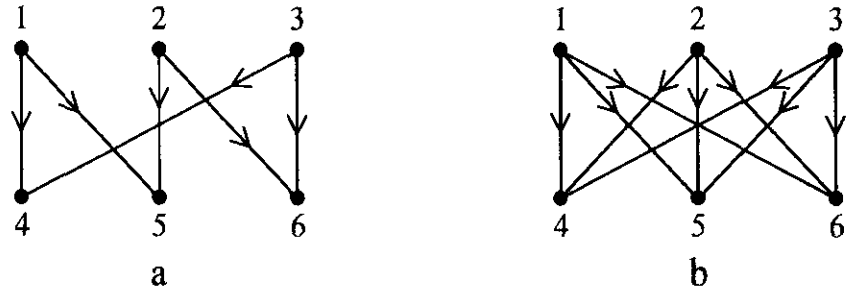

Fig. 9 Two different (non-isomorphic) graphs yielding the same minimal Lorenz curve

In case of Fig. 9a we have X=(2,2,2,-2,-2,-2) and for Fig. 9b we have X′=(3,3,3,-3,-3,-3) yielding the minimal Lorenz curve connecting linearly (0,0) with $(\frac{1}{2},1)$ and connecting linearly $(\frac{1}{2},1)$ with (1,0).

The above theorem also shows that, via the weakly connected digraphs without loops, we can obtain all minimal Lorenz curves, showing again that, also in this framework, no minimal curve is the lowest (since they all intersect).

Note :

It is clear from the above characterization of minimal Lorenz curves that they are obtained in the cases that all dominators have equal hierarchical degree (i.e. equal power) and that all subordinated vertices have equal hierarchical degree (i.e. are dominated in an equal way).

The other extreme is reached in the case of the graph yielding the largest Lorenz curve : if we consider the vertices 2,...,N-1 (all with zero hierarchical degree) as dominators as well as being dominated, we have indeed that, in this situation, the inequality (in hierarchical degree) between the dominators is maximal ; the same goes for the inequality between the ones that are dominated.

This shows that our model yields a good way to measure hierarchy in graphs. This is an important tool for measuring hierarchy in companies and administrations and in citation graphs (see Egghe and Rousseau (1990)).

Note :

In Botafogo, Rivlin and Shneiderman (1992), an attempt has been given to describe hierarchy in graphs. Instead of using the lengths of all chains (as we do), they use, per vertex $i \in \{1,...,N\}$, all d(i,j) and all d(j,i). They are analogous (but different) from our σ_i^+ and σ_i^-. The difference is discussed in the beginning of this section, cf. Fig. 5. Also, our proofs of the results of the hierarchy theory for graphs, especially the one of lemma III.1 is false if we use distances : see Fig. 10.

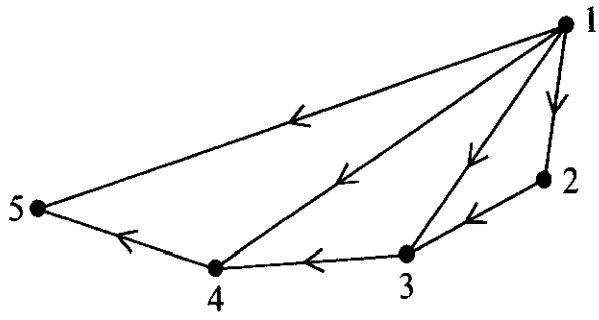

Fig. 10 Counterexample to lemma III.1 if we use distances
instead of lengths of chains in the definition of σ_1, σ_2

Indeed, denote σ'_1, σ'_2 the analogues of σ_1, σ_2 but now using "all distances" instead of "lengths of all chains". Then we have

$$\sigma'_1 = 4 < \sigma'_2 = 6\text{-}1 = 5,$$

showing that lemna III.1 is not true for the σ'_1, σ'_2.

But, as explained in the beginning of this section, it is more logical to use lengths of chains, instead of distances (as is done in Botafogo, Rivlin and Shneiderman (1992) and De Bra (2000)) to explain hierarchy (in short : what "makes" hierarchy <u>are</u> the chains!). The two notions coincide for chains, obviously. In Botafogo, Rivlin and Shneiderman (1992) and De Bra (2000), $\sigma'^{+}_i - \sigma'^{-}_i$ (for a vertex i) is called the prestige of i. They only use

$$\sum'_+ = \sum_{i=1}^{N} |\sigma'^{+}_i - \sigma'^{-}_i|. \qquad (31)$$

Twice this number is, what they call, "stratum" (as is easily seen from their formulae). For chains we have that $\sum_+ = \sum'_+$. It is clear from our theory that only using $\sum_+$ (or $\sum'_+$) is a too weak measure to describe hierarchy.

Open problem.

Characterise the graphs for which

(i) $\sigma_i = \sigma'_i$ for all $i \in \{1,...,N\}$

(ii) $\sum_+ = \sum'_+$

(of course, (i) $\Rightarrow$ (ii)). It is clear that chains satisfy this but it is easy to see that also other graphs can satisfy this, see the graph in Fig. 11

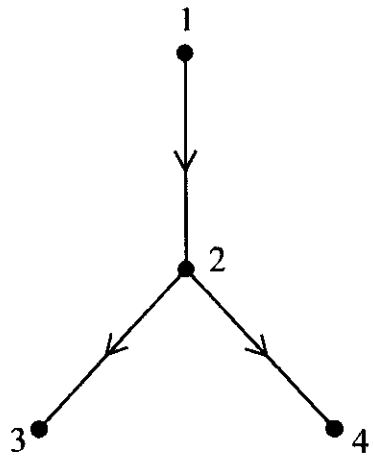

Fig. 11 Example of a graph where prestige equals hierarchical degree

Note that, by definition, we always have

$$\sigma_i^+ \geq \sigma_i'^+$$
$$\sigma_i^- \geq \sigma_i'^-$$

for all $i \in \{1,...,N\}$ since distance is determined by the length of the shortest chain between two points : this is used in $\sigma_i'^+$ and $\sigma_i'^-$. The lengths of all chains between two points is used in σ_i^+ and σ_i^-.

IV. Hierarchy theory, applied to general chains.

We will give general formulae for the hierarchical Lorenz curves for chains (general $N \in \mathbb{N}$) and for the good concentration measure 2A-1 of section II (corollary II.3)

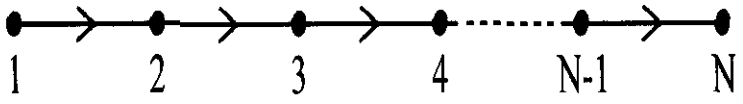

Fig. 12 A general chain of N vertices

Theorem IV.1

Let G be a unidirectional chain with N vertices. If N is even, then the vector $X=(\sigma_1,\ldots,\sigma_N)$ of hierarchical degrees is given by

$$X = \left(\frac{N(N-1)}{2}, \frac{N(N-3)}{2}, \frac{N(N-5)}{2}, \ldots, \frac{N}{2}, -\frac{N}{2}, \ldots, -\frac{N(N-5)}{2}, -\frac{N(N-3)}{2}, -\frac{N(N-1)}{2} \right) \quad (32)$$

$$\sum_{+} = \frac{N^3}{8} \quad (33)$$

and L_X connects (linearly) the points (in that order) :

$$(0,0), \left(\frac{1}{N}, \frac{4(N-1)}{N^2}\right), \left(\frac{2}{N}, \frac{8(N-2)}{N^2}\right), \left(\frac{3}{N}, \frac{12(N-3)}{N^2}\right), \ldots, \left(\frac{1}{2}, 1\right)$$

and then symmetrically to (1,0).

If N is odd, then the vector $X=(\sigma_1,\ldots,\sigma_N)$ of hierarchical degrees is given by

$$X = \left(\frac{N(N-1)}{2}, \frac{N(N-3)}{2}, \frac{N(N-5)}{2}, \ldots, N, 0, -N, \ldots, -\frac{N(N-5)}{2}, -\frac{N(N-3)}{2}, -\frac{N(N-1)}{2} \right) \quad (34)$$

$$\sum_{+} = \frac{N^3-N}{8} \tag{35}$$

and L_X connects (linearly) the points (in that order) :

$$(0,0), \left(\frac{1}{N},\frac{4}{N+1}\right), \left(\frac{2}{N},\frac{8(N-2)}{N^2-1}\right), \left(\frac{3}{N},\frac{12(N-3)}{N^2-1}\right),\ldots, \left(\frac{2}{(N-1)N},1\right),\left(\frac{2}{(N+1)N},1\right)$$

and then symmetrical to (1,0)

Proof : N even

It is easy to see that

$$\sigma_1 = \sum_{i=1}^{N-1} i = \frac{N(N-1)}{2}$$

$$\sigma_2 = \sum_{i=1}^{N-2} i-1 = \frac{N(N-3)}{2}$$

$$\sigma_3 = \sum_{i=1}^{N-3} i-1-2 = \frac{N(N-5)}{2}$$

....

$$\sigma_{\frac{N}{2}} = \sum_{i=1}^{\frac{N}{2}} i-1-2-\ldots-\left(\frac{N}{2}-1\right) = \frac{N}{2}$$

$$\sigma_{\frac{N}{2}+1} = -\frac{N}{2}$$

....

$$\sigma_N = -\frac{N(N-1)}{2}$$

$\sum_{+} = \sum_{k=1}^{\frac{N}{2}} \sigma_k = \frac{N^3}{8}$ as is easily seen and hence L_X is as given.

N odd

It is easy to see that

$$\sigma_1 = \sum_{i=1}^{N-1} i = \frac{N(N-1)}{2}$$

$$\sigma_2 = \sum_{i=1}^{N-2} i-1 = \frac{N(N-3)}{2}$$

....

$$\sigma_{\frac{N-1}{2}} = \sum_{i=1}^{\frac{N+1}{2}} i-1-2-\ldots-\left(\frac{N-1}{2}-1\right) = N$$

$$\sigma_{\frac{N+1}{2}} = \sum_{i=1}^{\frac{N-1}{2}} i-1-2-\ldots-\frac{N-1}{2} = 0$$

$$\sigma_{\frac{N+3}{2}} = -N$$

.....

$$\sigma_N = -\frac{N(N-1)}{2}$$

$\sum_+ = \sum_{k=1}^{\frac{N-1}{2}} \sigma_k = \frac{N^3-N}{8}$ as is easily seen and hence L_X is as given. □

Note :

The last note of the previous section explain that in chains (as studied here), the number $2\sum_+$ is what is called in Botafogo, Rivlin and Shneiderman (1992) "stratum" - see also De Bra (2000). The formulae for $\sum_+$ above also appear in Botafogo, Rivlin and Shneiderman (1992).

From theorem IV.1 we can derive the values of 2A-1 (the normalization of A, being the area under the Lorenz curve) for general chains.

Theorem IV.2.

Let G be a unidirectional chain with N vertices. Let A be the area under its hierarchical Lorenz curve. Then the normalized measure 2A-1 equals

$$\frac{1}{3}\left(1-\frac{4}{N^2}\right) \tag{36}$$

if N is even and equals $\frac{1}{3}$ for all N odd. So, if N is large,

$$2A-1 \approx \frac{1}{3} \tag{37}$$

for all chains and $\approx$ becomes = for all odd N.

Proof :

Rather than calculating the area A from left to right, separating the area in pieces of abscissa length $\frac{1}{N}$, we will work from bottom to top, separating the area in pieces of ordinate length $\frac{\sigma_i}{\sum_+}$ (i=1,..., $\frac{N}{2}$ (N even) or $\frac{N-1}{2}$ (N odd)) which is much easier, based on theorem IV.1.

Let, first, N be even. We have

$$A = \frac{8}{N^3}\left[\frac{N(N-1)}{2}\left(1-\frac{1}{N}\right)+\frac{N(N-3)}{2}\left(1-\frac{3}{N}\right)+\ldots+\frac{N(N-(N-3))}{2}\left(1-\frac{N-3}{N}\right)+\frac{N}{2}\left(1-\frac{N-1}{N}\right)\right]$$

$$= \frac{4}{N^3}\left[(N-1)^2+(N-3)^2+\ldots+3^2+1^2\right]$$

$$= \frac{4}{N^3}\sum_{j=1}^{\frac{N}{2}}(N-(2j-1))^2$$

$$= \frac{4}{N^3}\sum_{j=1}^{\frac{N}{2}}((N+1)^2-4j(N+1)+4j^2)$$

$$= \frac{4}{N^3}\left[(N+1)^2\frac{N}{2}-4(N+1)\frac{\frac{N}{2}\left(\frac{N}{2}+1\right)}{2}+4\frac{\frac{N}{2}\left(\frac{N}{2}+1\right)(N+1)}{6}\right]$$

$$A = \frac{2}{3}\left(1-\frac{1}{N^2}\right). \text{ Hence } 2A-1 = \frac{1}{3}\left(1-\frac{4}{N^2}\right).$$

Let now N be odd. Now

$$A = \frac{8}{N^3-N}\left[\frac{N(N-1)}{2}\left(1-\frac{1}{N}\right)+\frac{N(N-3)}{2}\left(1-\frac{3}{N}\right)+\ldots+\frac{N.2}{2}\left(1-\frac{N-2}{N}\right)\right]$$

$$= \frac{4}{N^3-N}\left[(N-1)^2+(N-3)^2+\ldots+2^2\right]$$

$$= \frac{4}{N^3-N}\sum_{j=1}^{\frac{N-1}{2}}(N-(2j-1))^2.$$

Proceeding as in the case N even, we finally reach $A = \frac{2}{3}$, for all odd N. Hence

$$2A - 1 = \frac{1}{3}$$

for all chains with an odd number of vertices. □

Note : It is easy to see that, if N is even,

$$L_{N+1} > L_N,$$

where L_N, L_{N+1} denote the hierarchical Lorenz curve of the chain with N resp. N+1 vertices. But as indicated above (theorems IV.1, IV.2), $L_N \approx L_{N+1}$ for all N large.

References

P.D. Allison (1978). Measures of inequality. American Sociological Review 43, 865-880.

A.B. Atkinson (1970). On the measurement of inequality. Journal of Economic Theory 2, 244-263.

R.A. Botafogo, E. Rivlin and B. Shneiderman (1992). Structural analysis of hypertexts : identifying hierarchies and useful metrics. ACM Transactions on Information Systems 10(2), 142-180.

H. Dalton (1920). The measurement of the inequality of incomes. The Economic Journal 30, 348-361.

P. De Bra (2000). Using hypertext metrics to measure research output levels. Scientometrics 47(2), 227-236.

L. Egghe and R. Rousseau (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier Science Publishers, Amsterdam.

C. Gini (1909). Il diverso accrescimento delle classi sociali e la concentrazione della richezza. Giornali degli Economisti serie 11, 37.

G.H. Hardy, J.E. Littlewood and G. Pólya (1929). Some simple inequalities satisfied by convex functions. Messenger Math. 58, 145-152.

G.H. Hardy, J.E. Littlewood and G. Pólya (1952). Inequalities. Cambridge University Press. Cambridge, UK. Reprinted in 1988.

M.O. Lorenz (1905). Methods of measuring concentration of wealth. Journal of the American Statistical Association 9, 209-219.

A.W. Marshall and I. Olkin (1979). Inequalities : Theory of Majorization and its Applications. Mathematics in Science and Engineering, Volume 143, Academic Press, New York.

R.F. Muirhead (1903). Some methods applicable to identities and inequalities of symmetric algebraic functions of n letters. Proceedings of the Edinburgh Mathematical Society 21, 144-157.

R.C. Reid (1997). Enumeration. In : Graph Connections (L.W. Beineke and R.J. Wilson, eds). Oxford Science Publications, Clarendon Press, Oxford.

R. Rousseau and P. Van Hecke (1999). Measuring biodiversity. Acta Biotheoretica 47, 1-5.

C. Shannon (1948). A mathematical theory of communication. Bell System Technical Journal 28, 379-423.

H. Theil (1967). Economics and Information Theory. North-Holland, Amsterdam.

R.J. Wilson (1972). Introduction to Graph Theory. Longman, London.