EPJ manuscript No.
(will be inserted by the editor)

Identifying the Ultrametricity of Networks

Fionn Murtagh [a]

Department of Computer Science, Royal Holloway University of London, Egham, Surrey TW20 0EX, England

May 7, 2005

Abstract. High dimensional, sparsely populated data spaces have been characterized in terms of ultrametric topology. This implies that there are natural, not necessarily unique, tree or hierarchy structures defined by the ultrametric topology. The extent to which time series can be embedded in an ultrametric topology was also studied recently, with the aim of finding unique "fingerprints" for a time series, discriminating between time series from different domains, and opening up the possibility of exploiting hierarchical structures in the data. In the present work, we do the same for networks, defined as edge-weighted complete graphs. We first study random networks, i.e. networks with randomly valued edge weights. Then we investigate coherent and meaningful collections of nearly 1000 texts, comprising over one million words. The latter allow us to infer semantic similarity between the texts. We show that we can distinguish between these semantic networks in terms of extent of ultrametricity.

PACS. 89.75.Hc Networks and genealogical trees – 02.50.Sk Multivariate analysis – 89.75.Kd Patterns – 89.75.Fb Structures and organization in complex systems

1 Introduction

Through finding hierarchical structure in networks it is possible to pinpoint critical and influential nodes. Trusina et al. [21] use connectivity analysis, based on vertex degree. The special issue [6] contains other articles on this theme. Barabási [4] has extensive online visualization resources and has studied biological and other networks in terms of (scale free) power law versus exponential connectivity law, leading to robustness versus fragility as new network nodes are added or deleted. All such work shares the perspective of network connectivity and the scaling properties or dynamics of connectivity.

Our work takes a point of departure in complete graphs. Our motivation is to to have a limit model for dense biological neural networks, or semantic networks, or even

[a] E-mail fmurtagh@acm.org

communications networks. Our focus is on complete edge weighted graphs and connectivity is implicitly handled through zero-valued or near zero-valued edge weights (both of which can be considered as indicating lack of connection between associated vertices).

We report new results on scaling properties of complete graphs. What has led us to the study of complete edge-weighted graphs is past work on the remarkable scaling (relative to spatial dimensionality) properties of points in high dimensional space [13,18]. These results are based on relative spatial sparensess. In fact, the embedding spatial dimensionality does not need to be very large. Thus for spatial dimensionalities of a thousand and upwards, containing a few thousand points, we can find quite remarkable properties. What we see is that the points tend to be equidistant. Now, equidistant separation defines one case of the ultrametric inequality. So increasing relative sparseness implies a natural hierarchy for the set of points considered.

Note though that such a hierarchy is not unique. If we fix the hierarchy to be of a particular combinatorial type (unlabeled, ranked) then the number of hierarchies on n points is counted combinatorially by the André numbers [9,7,12].

The data studied in [13,18] is point pattern data: observational features with their measurements on many coordinate dimensions. Data may be instead presented as time-varying signals and in a similar way, related to the findings of [13,18], we have investigated ultrametric-related properties of time series or 1D signals in [14]. In the latter time series work, we encoded the data in a particular way.

In this work we study data which is given as a network. As for our time series work [14] our goals include the following. Firstly, we can use the extent of ultrametricity to distinguish between data sets; i.e., to characterize quantitatively different networks. Secondly, if we can find some ultrametricity – some degree or extent of ultrametricity – in a network, two questions are immediately raised: (i) what is the physics or biological or other scientific explanation for this?, and (ii) how do we exploit it? Since we are dealing with complete graphs, one answer to the latter exploitation question lies in proximity searching [13].

We recall: The triangular inequality holds for a metric space: $d(x,z) \leq d(x,y) + d(y,z)$ for any triplet of points x, y, z. In addition the properties of symmetry and positive definiteness are respected. The "strong triangular inequality" or ultrametric inequality is: $d(x,z) \leq \max \{d(x,y), d(y,z)\}$ for any triplet x, y, z. An ultrametric space implies respect for a range of stringent properties. For example, the triangle formed by any triplet is necessarily isosceles, with the two large sides equal; or is equilateral.

We begin in the next section by considering complete graphs with uniformly- or Gaussian-distributed edge weights and show that a Euclidean embedding of such graphs increases in dimensionality with graph size (viz., numbers of vertices and hence of edges). Therefore from [13,18], ultrametricity also increases.

We then turn attention to real case studies. For this, we take a large number of coherent collections of meaningful

texts. Through shared words, we can define a similarity network between all texts in each of the collections we chose. Aspects of the semantics of the given collection are captured in this way. We investigate how ultrametric each of these semantic networks is.

We select texts each containing roughly 500 to 1000 words (but as will be seen below, some texts had up to around 44,000 words). All are in English and do not contain accented characters (and this can be very easily generalized). These were: fairy tales by the Brothers Grimm; novels by the English writer, Jane Austen; in order to have very technical language, aircraft accident reports from the US National Transport Safety Board; and in order to seek linkages with biological and cognitive processes, a range of dream reports from the online DreamBank repository.

We find clear distinctions between the semantic networks (or text collections) studied, in terms of their relative (albeit small) extent of ultrametricity.

2 Least Squares Network Euclidean Embedding

Consider a network defined by the complete graph consisting of n vertices and $m = n(n-1)/2$ edges. For non-degenerate cases of distribution of edge weights, the dimensionality of the best Euclidean embedding of these edge weights increases with n and with m. We will show that the Euclidean embedding dimensionality increases linearly with n, and hence with the square root of m, for uniform and Gaussian distributions of edge weights.

If $n = 2$, then the two vertices can be located in a space of dimensionality at most $n - 1$. If $n = 3$, the 3 points (vertices) define an $(n-1)$-dimensional surface. In general, the set of pairwise distances on a set of n points implies that these points can be located in a Euclidean space of dimension at most $n - 1$.

This principle has been used in classical multidimensional scaling [11, 20].

Classical multidimensional scaling (also known as principal coordinates analysis or metric scaling) assumes a hypothetical real data matrix, X, of dimensions $n \times p$, and we seek the values of p and of X, where the latter are coordinate projections in some p-dimensional Euclidean space. A natural choice for X is such that the coordinates are principal axes ones, i.e. defined by the eigenvectors of outer product matrix $A = XX^t$ (X^t is transpose of X). Now if each axis is centered, and if d_{ik}^2 is the edge weight between graph vertices i and k, then it can be shown that matrix A has typical element $a_{ik} = -\frac{1}{2}(d_{ik}^2 - d_i^2 - d_k^2 - d^2)$ where $d_i^2 = \frac{1}{n}\sum_k d_{ik}^2$, $\quad d_k^2 = \frac{1}{n}\sum_i d_{ik}^2$, $\quad d^2 = \frac{1}{n^2}\sum_i \sum_k d_{ik}^2$. When d is Euclidean distance then it can be shown that matrix A is positive, symmetric and semidefinite, with rank $p \leq n$. Having thus been given distances, we have constructed matrix $A = XX'$; and then determined an orthogonal basis in which the coordinate projections give us X.

In practice, dissimilarities, d, may be at issue rather than distances. (We recall: dissimilarities are symmetric, $d_{ik} = d_{ki}$; positive definite, $d_{ik} > 0$ if $i \neq k$, and $d_{ik} = 0$ if $i = k$, but do *not* satisfy the triangular inequality.) Then,

matrix A will be symmetric and will have zero values on the diagonal but will not be positive semidefinite. In this case negative eigenvalues are obtained. These are inconvenient but may be ignored if the approximate Euclidean representation (given by the eigenvectors corresponding to positive eigenvalues) is satisfactory. Here, we work with these non-negative eigenvalues which provide an approximate Euclidean space representation of the given dissimilarity data.

In Tables 1 and 2 uniform and Gaussian realizations, respectively, were used. From each Gaussian random value, the overall minimum was subtracted, in order to have positive (or zero) valued dissimilarities. The best-fitting Euclidean embedding is given by the principal coordinates (defined by the eigenvectors) corresponding to non-negative eigenvalues. In Tables 1 and 2 it is clear that the dimensionality of the best-fitting Euclidean embedding increases linearly with number of nodes n for these data distributions.

Following on from these results, presented in Tables 1 and 2, and the findings of [13,18] – putting both properties together – it can be seen that increasingly large networks are increasingly ultrametric. Our conclusion holds for uniformly- and Gaussian-distributed network edge weights.

In the next section, we target real data in the form of pairwise linkages expressing certain aspects of semantic linkage between texts.

Table 1. Edge weights were defined for sets of 100, 200, 300, 400, and 500 nodes – hence 100*99/2 edge weights, 200*199/2 weights, etc. These edge weights were uniformly distributed on [0,1]. The best strictly Euclidean embedding was found using classical multidimensional scaling, by retaining non-negative eigenvalues. The dimensionality of this best Euclidean embedding is given, for 6 different random realizations. We can see a linear increase in average Euclidean best fit dimensionality (rightmost column) with n, the number of network nodes (leftmost column).

No. of nodes	Six random realizations	Average
100	53, 53, 53, 55, 55, 54	53.83
200	105, 105, 106, 106, 106, 104	105.33
300	157, 157, 156, 155, 155, 155	155.83
400	207, 206, 206, 207, 207, 206	206.50
500	258, 258, 259, 259, 257, 258	258.17

Table 2. As for Table 1 but here the edge weights were Gaussian distributed with mean 0 and standard deviation 1, followed by subtraction of the overall minimum value, to yield non-negative values. Again we see an approximate linear relationship between the number of nodes, n, and the dimensionality of the best-fitting Euclidean embedding (rightmost column).

No. of nodes	Six random realizations	Average
100	57, 57, 56, 57, 56, 57	56.67
200	109, 109, 109, 108, 109, 110	109.00
300	161, 164, 162, 163, 162, 162	162.33
400	215, 215, 214, 216, 213, 215	214.67
500	268, 267, 270, 265, 265, 267	267.00

3 Real Case Studies: Methodology

We employ correspondence analysis for metric embedding, followed by determination of the extent of ultrametricity, in factor space, based on the alpha coefficient of ultrametricity. Our motivation for using precisely this Euclidean embedding is as follows. Our input data is in the form of frequencies of occurrence. Now, a Euclidean distance defined on vectors with such values is not appropriate.

The χ^2 distance is an appropriate weighted Euclidean distance for use with such data [5,15]. Consider texts i and i' crossed by words j. Let k_{ij} be the number of occurrences of word j in text i. Then, omitting a constant, the χ^2 distance between texts i and i' is given by $\sum_j 1/k_j(k_{ij}/k_i - k_{i'j}/k_{i'})^2$. The weighting term is $1/k_j$. The weighted Euclidean distance is between the *profile* of text i, viz. k_{ij}/k_i for all j, and the analogous *profile* of text i'.

3.1 Alpha Coefficient of Ultrametricity

The definition of ultrametricity introduced in [13] was, in summary, as follows. For all triplets of points, we consider the three internal angles. We require that the smallest angle be less than or equal to 60 degrees. Then we require that the two remaining angles be approximately equal. Approximate equality is defined as less than 2 degrees, in order to cater for imprecise coordinate measurement (e.g., due to floating point values) in an acceptable way. Satisfying these angular constraints implies that the triplet of points defines an approximate isosceles (with small base) or equilateral triangle. We define a coefficient of ultrametricity of the point set as the proportion of all triangles satisfying these requirements. The coefficient of ultrametricity is 1 for perfectly ultrametric data; and if 0 no triangle satisfies the isosceles or equilateral requirements. This coefficient is referred to as alpha below in this article.

3.2 Correspondence Analysis: Mapping χ^2 into Euclidean Distances

As a dimensionality reduction technique, not unlike principal components analysis, correspondence analysis is particularly appropriate for handling frequency data. As an example of the latter, frequencies of word occurrence in text will be studied below.

The given contingency table (or numbers of occurrence) data is denoted $k_{IJ} = \{k_{IJ}(i,j) = k(i,j); i \in I, j \in J\}$. I is the set of text indexes, and J is the set of word indexes. We have $k(i) = \sum_{j \in J} k(i,j)$. Analogously $k(j)$ is defined, and $k = \sum_{i \in I, j \in J} k(i,j)$. Next, $f_{IJ} = \{f_{ij} = k(i,j)/k; i \in I, j \in J\} \subset \mathbb{R}_{I \times J}$, similarly f_I is defined as $\{f_i = k(i)/k; i \in I, j \in J\} \subset \mathbb{R}_I$, and f_J analogously. What we have described here is taking numbers of occurrences into relative frequencies.

The conditional distribution of f_J knowing $i \in I$, also termed the jth profile with coordinates indexed by the elements of I, is:

$$f_J^i = \{f_j^i = f_{ij}/f_i = (k_{ij}/k)/(k_i/k); f_i \neq 0; j \in J\}$$

and likewise for f_I^j.

Note that the input data values here are always non-negative reals. The output factor projections (and contributions to the principal directions of inertia) will be reals.

3.3 Input: Cloud of Points Endowed with the Chi Squared Metric

The cloud of points consists of the couple: profile coordinate and mass. We have $N_J(I) = \{(f_J^i, f_i); i \in I\} \subset \mathbb{R}_J$, and again similarly for $N_I(J)$.

The moment of inertia is as follows:

$$M^2(N_J(I)) = M^2(N_I(J)) = \|f_{IJ} - f_I f_J\|^2_{f_I f_J}$$

$$= \sum_{i \in I, j \in J} (f_{ij} - f_i f_j)^2 / f_i f_j \quad (1)$$

The term $\|f_{IJ} - f_I f_J\|^2_{f_I f_J}$ is the χ^2 metric between the probability distribution f_{IJ} and the product of marginal distributions $f_I f_J$, with as center of the metric the product $f_I f_J$. Decomposing the moment of inertia of the cloud $N_J(I)$ – or of $N_I(J)$ since both analyses are inherently related – furnishes the principal axes of inertia, defined from a singular value decomposition.

3.4 Output: Cloud of Points Endowed with the Euclidean Metric in Factor Space

Let us consider a tensor (Einstein) notation of transitions between probability spaces. A transition from I to J is an element of the tensor product $\mathbb{R}_J \otimes \mathbb{R}^I$. It is a function on I, but with values in the J measures; or the conditional probability of j given i. Such a transition takes masses (or probability measures or densities) from I to J; and associates every function on J with a function on I.

We write: $f_J^I \in \mathbb{R}_J \otimes \mathbb{R}^I$, signifying that f_J^I is a measure relative to I, and a function relative to J. Taking a step back in the data we analyze, the normalized frequencies viewed as probabilities are given by $f_{IJ} \in \mathbb{R}_I \otimes \mathbb{R}_J$.

The marginals or masses, which are measures on I and J, can be written:

$$f_I = \delta_I^{IJ} \circ f_{IJ} \text{ and } f_J = \delta_J^{IJ} \circ f_{IJ}$$

where δ is the Kronecker delta. Composition of measure and function transitions is denoted $\circ$.

In this notation we can also write: $f_{IJ} = (\delta_I^I \times f_J^I) \circ f_I$, and $f_{IJ} = (f_I^J \times \delta_J^J) \circ f_J$.

It can be shown ([5], p. 153) that the principal eigenvalue λ corresponding to eigenvector ϕ satisfies: $\phi^I \circ f_I^J \circ f_J^I - (\phi^I \circ f_I)\delta^I = \lambda \phi^I$. Furthermore it holds that $\delta^I \circ f_I^J \circ f_J^I = (\delta^I \circ f_I)\delta^I = \delta^I$; that is to say, δ^I is the first trivial eigenvector, i.e., the constant function equal to 1. The factor ϕ^I is zero mean for the measure f_I, i.e., $\phi^I \circ f_I = 0$.

We can right-multiply the eigen-equation above by f_I^J to get $(\phi^I \circ f_I^J) \circ (f_J^I \circ f_I^J) = \lambda(\phi^I \circ f_I^J)$. Consequently $\phi^I \circ f_I^J$ is a factor of the dual space.

Through consideration of the norms, it turns out that we can define factors on J, ϕ^J, in the following way: $\phi^J = (1/\sqrt{\lambda(\phi)})\phi^I \circ f_I^J$.

Because of the inherent link between the factors on I, ϕ^I, and the factors on J, ϕ^J, that is to say, the particular Euclidean projection used for the clouds $N_J(I)$ and $N_I(J)$, we have that the number of non-zero eigenvalues will be $\leq \min(n-1, m-1)$ (where $n = |N_J(I)|$ is set cardinality of the texts/observations, and $m = |N_I(J)|$ is set cardinality of the words/attributes; the minus 1 term

is explained by the trivial eigenvector mentioned above, and is simultaneously explained by the centering of the cloud).

3.5 Conclusions on Correspondence Analysis and Introduction to the Numerical Experiments to Follow

Some important points for the analyses to follow are – firstly in relation to correspondence analysis:

1. From numbers of occurrence data we always get (by design) a Euclidean embedding using correspondence analysis. The factors are embedded in a Euclidean metric.
2. As seen in the previous subsection, the numbers of factors, i.e. number of non-zero eigenvalues, are given by one less than the minimum of the number of observations studied (indexed by set I) and the number of variables or attributes used (indexed by set J).
3. The number of dimensions in factor space may be less than full rank if there are linear dependencies present.
4. In the experiments to follow in the next section, we take $n < m$ always, implying that inherent (full rank) dimensionality of the projected Euclidean factor space is $n-1$. We do this to make comparability of the results clear and straightforward.
5. We also take $m = 1000, 2000$ and the full attribute set (say, m_{tot}) in each case, where the attributes are ordered in terms of decreasing marginal frequency. In other words, we take the 1000 most frequent words to characterize our texts; then the 2000 most frequent words; and finally all words. Since $n < m$ it is not surprising that very similar results are found irrespective of the value of m. After all, the inherent, projected, Euclidean, factor space dimensionality is the same in each case, viz., $n - 1$.
6. From the previous remark, viz. that the results obtained for the $m = 1000, 2000$, and all most frequent words, are of the same inherent dimensionality we motivate our use of these different characterizations of the text set by the need to study the stability of our results. We will show quite convincingly that our results are characteristic of the texts used, in each case, and are in no way "one off" or arbitrary.

Some important points related to our numerical assessments below, in relation to data used, determining of ultrametricity coefficient, and software used, are as follows.

1. In line with one tradition of textual analysis associated with Benzécri's correspondence analysis (see [15]) we take the unique full words and rank them in order of importance. Thus for the Brothers Grimm work, below, we find: "the", 19,696 occurrences; "and", 14,582 occurrences; "to", 7380 occurrences; "he", 5951 occurrences; "was", 4122 occurrences; and so on. Last three, with one occurrence each: "yolk", "zeal", "zest".
2. The alpha ultrametricity coefficient is based on triangles. Now, with n graph nodes we have $O(n^3)$ possible triangles which is computationally prohibitive, so we instead sample. The means and standard deviations below are based on 2000 random triangle vertex re-

alizations, repeated 20 times; hence, in each case, in total 40,000 random selections of triangles.

3. All text collections reported on below (section 4) are publicly accessible (and web addresses are cited). All texts were obtained by us in straight (ascii) text format.

   The preparation of the input data was carried out with programs of ours, written in C, and available at www.correspondances.info (accompanying [15]). The correspondence analysis software was written in the public R statistical software environment (www.r-project.org, again see [15]) and is available at this same web address. Some simple statistical calculations were carried out by us also in the R environment.

   The classical multidimensional scaling (or principal coordinates analysis) work reported on above in subsection 2 used R command `cmdscale`).

4 Real Case Studies: Text Interrelationships Through Shared Words

We use in all over 900 short texts, given by short stories, or chapters, or short reports. All are in English. Unique words are determined through delimitation by white space and by punctuation characters with no distinction of upper and lower case. In all, over one million words are used in our studies of these texts. The study of word/text occurrences in a straightforward way, with no truncation nor stemming nor other preprocessing, typifies a great deal of the work of Benzécri, and his journal *Les Cahiers de l'Analyse des Données*, published by the French publisher Dunod over three decades up to 1996. This work of Benzécri is discussed in detail in [15].

4.1 Brothers Grimm

As a homogeneous collection of texts we take 209 fairy tales of the Brothers Grimm [17], containing 7443 unique (in total 280,629) space- or punctuation-delimited words. Story lengths were between 650 and 44,400 words.

To define a semantic context of increasing resolution we took the most frequent 1000 words, followed by the most frequent 2000 words, and finally all 7443 words. We constructed a cross-tabulation of numbers of occurrences of each word in each one of the 209 fairy tales. This led therefore to a set of frequency tables of dimensions: $209 \times 1000, 209 \times 2000$ and 209×7443. Through use of the χ^2 distance between fairy tale texts, a correspondence analysis was carried out. From the three frequency tables, the contingency table crossing all pairs of fairy tales could be examined; but it was far more convenient for us to proceed straight to the factor space, of dimension $209 - 1 = 208$. The factor space is Euclidean, so the correspondence analysis can be said to be a mapping from the χ^2 metric into a Euclidean metric space.

Table 3 (columns 4, 5) shows remarkable stability of the alpha ultrametricity coefficient results, and such stability will be seen in all further results to be presented below. The ultrametricity is not high for the Grimm Brothers' data: we recall that an alpha value of 0 means no triangle is isosceles/equilateral. We see that there is very little

Table 3. Coefficient of ultrametricity, alpha. Input data: frequencies of occurrence matrices defined on the 209 texts crossed by: 1000, 2000, and all = 7443, words. Alpha (ultrametricity coefficient) based on factors: i.e., factor projections resulting from correspondence analysis, with Euclidean distance used between each pair of texts in factor space, of dimensionality 208.

209 Brothers Grimm fairy tales				
Texts	Orig.Dim.	FactorDim.	Alpha, mean	Alpha, sdev.
209	1000	208	0.1236	0.0054
209	2000	208	0.1123	0.0065
209	7443	208	0.1147	0.0066

ultrametric (hence hierarchical) structure in the Brothers Grimm data (based on our particular definition of ultrametricity/hierarchy).

4.2 Jane Austen

To further study stories of a general sort, we use some works of the English novelist, Jane Austen.

1. *Sense and Sensibility*, published 1811 [1], 50 chapters = files, chapter lengths from 1028 to 5632 words.
2. *Pride and Prejudice*, published 1813 [2], 61 chapters each containing between 683 and 5227 words.
3. *Persuasion*, published 1817 [3], 24 chapters, chapter lengths 1579 to 7007 words.
4. *Sense and Sensibility* split into 131 separate texts, each containing around 1000 words (i.e., each chapter was split into files containing 5000 or fewer characters). We did this to check on any influence by the size (total number of words) of the text unit used (and we found no such influence).

Table 4. Coefficient of ultrametricity, alpha. Input data: frequencies of occurrence matrices defined on the 266 texts crossed by: 1000, 2000, and all = 9723, words. Alpha (ultrametricity coefficient) based on factors: i.e., factor projections resulting from correspondence analysis, with Euclidean distance used between each pair of texts in factor space. Dimensionality of latter is necessarily $\leq 266 - 1$, adjusted for 0 eigenvalues = linear dependence.

266 Austen chapters or partial chapters				
Texts	Orig.Dim.	FactorDim.	Alpha, mean	Alpha, sdev.
266	1000	261	0.1455	0.0084
266	2000	262	0.1489	0.0083
266	9723	263	0.1404	0.0075

In all there were 266 texts containing a total of 9723 unique words. We looked at the 1000, 2000 and all = 9723 most frequent words to characterize the texts by frequency of occurrence.

Table 4, again displaying very stable alpha values, indicates that the Austen corpus is a small amount more ultrametric than the Grimms' corpus, Table 3.

4.3 Air Accident Reports

We used air accident reports to explore documents with very particular, technical, vocabulary. The NTSB aviation accident database [16] contains information about civil aviation accidents in the United States and elsewhere. We selected 50 reports. Examples of two such reports used by

us: occurred Sunday, January 02, 2000 in Corning, AR, aircraft Piper PA-46-310P, injuries – 5 uninjured; occurred Sunday, January 02, 2000 in Telluride, TN, aircraft: Bellanca BL-17-30A, injuries – 1 fatal. In the 50 reports, there were 55,165 words. Report lengths ranged between approximately 2300 and 28,000 words. The number of unique words was 4261.

Sample of start of report 30: *On January 16, 2000, about 1630 eastern standard time (all times are eastern standard time, based on the 24 hour clock), a Beech P-35, N9740Y, registered to a private owner, and operated as a Title 14 CFR Part 91 personal flight, crashed into Clinch Mountain, about 6 miles north of Rogersville, Tennessee. Instrument meteorological conditions prevailed in the area, and no flight plan was filed. The aircraft incurred substantial damage, and the private-rated pilot, the sole occupant, received fatal injuries. The flight originated from Louisville, Kentucky, the same day about 1532.*

In Table 5 we find ultrametricity values that are marginally greater than those found for the Brothers Grimm (Table 3). It could be argued that the latter, too, uses its own technical vocabulary. We would need to use more data to see if we can clearly distinguish between the (small) ultrametricity levels of these two corpora.

Table 5. Coefficient of ultrametricity, alpha. Input data: frequencies of occurrence matrices defined on the 50 texts crossed by: 1000, 2000, and all = 4261, words. Alpha (ultrametricity coefficient) based on factors: i.e., factor projections resulting from correspondence analysis, with Euclidean distance used between each pair of texts in factor space. Dimensionality of latter is necessarily less than $50-1$, with an additional adjustment made for one 0-valued eigenvalue, implying linear dependence.

50 aviation accident reports				
Texts	Orig.Dim.	FactorDim.	Alpha, mean	Alpha, sdev.
50	1000	48	0.1338	0.0077
50	2000	48	0.1186	0.0058
50	4261	48	0.1154	0.0050

4.4 DreamBank

With dream reports (i.e., reports by individuals on their remembered dreams) we depart from a technical vocabulary, and instead raise the question as to whether dream reports can perhaps be considered as types of fairy tale or story, or even akin to accident reports.

From the Dreambank repository [8,10,19] we selected the following collections:

1. "Alta: a detailed dreamer," in period 1985–1997, 422 dream reports.
2. "Chuck: a physical scientist," in period 1991–1993, 75 dream reports.
3. "College women," in period 1946–1950, 681 dream reports.
4. "Miami Home/Lab," in period 1963–1965, 445 dream reports.
5. "The Natural Scientist," 1939, 234 dream reports.
6. "UCSC women," 1996, 81 dream reports.

To have adequate length reports, we requested report sizes of between 500 and 1500 words. With this criterion, from (1) we obtained 118 reports, from (2) and (6) we obtained no reports, from (3) we obtained 15 reports, from (4) we obtained 73 reports, and finally from (5) we obtained 8 reports. In all, we used 214 dream reports, comprising 13696 words.

Sample of start of report 100: *I'm delivering a car to a man – something he's just bought, a Lincoln Town Car, very nice. I park it and go down the street to find him – he turns out to be an old guy, he's buying the car for nostalgia – it turns out to be an old one, too, but very nicely restored, in excellent condition. I think he's black, tall, friendly, maybe wearing overalls. I show him the car and he drives off. I'm with another girl who drove another car and we start back for it but I look into a shop first – it's got outdoor gear in it - we're on a sort of mall, outdoors but the shops face on a courtyard of bricks. I've got something from the shop just outside the doors, a quilt or something, like I'm trying it on, when it's time to go on for sure so I leave it on the bench. We go further, there's a group now, and we're looking at this office facade for the Honda headquarters.*

With the above we took another set of dream reports, from one individual, Barbara Sanders. A more reliable (according to [10]) set of reports comprised 139 reports, and a second comprised 32 reports. In all 171 reports were used from this person. Typical lengths were about 2500 up to 5322. The total number of words in the Barbara Sanders set of dream reports was 107,791.

Table 6. Coefficient of ultrametricity, alpha. Input data: frequencies of occurrence matrices defined on the 384 texts crossed by: 1000, 2000, and all = 11441, words. Alpha (ultrametricity coefficient) based on factors: i.e., factor projections resulting from correspondence analysis, with Euclidean distance used between each pair of texts in factor space, of dimensionality $385 - 1 = 384$.

385 dream reports				
Texts	Orig.Dim.	FactorDim.	Alpha, mean	Alpha, sdev.
385	1000	384	0.1998	0.0088
385	2000	384	0.1876	0.0095
385	11441	384	0.1933	0.0087

First we analyzed all dream reports, furnishing Table 6.

In order to look at a more homogeneous subset of dream reports, we then analyzed separately the Barbara Sanders set of 171 reports, leading to Table 7. (Note that this analysis is on a subset of the previously analyzed dream reports, Table 6). The Barbara Sanders subset of 171 reports contained 7044 unique words in all.

Compared to Table 6 based on the entire dream report collection, Table 7 which is based on one person shows, on average, higher ultrametricity levels. It is interesting to note that the dream reports, collectively, are higher in ultrametricity level than our previous values for alpha; and that the ultrametricity level is raised again when the data used relates to one person.

We carried out a preliminary study of James Joyce's *Ulysses*, comprising 304,414 words in total. We broke this

Table 7. Coefficient of ultrametricity, alpha. Input data: frequencies of occurrence matrices defined on the 171 texts crossed by: 1000, 2000, and all = 7044, words. Alpha (ultrametricity coefficient) based on factors: i.e., factor projections resulting from correspondence analysis, with Euclidean distance used between each pair of texts in factor space, of dimensionality $171 - 1 = 170$.

171 Barbara Sanders dream reports				
Texts	Orig.Dim.	FactorDim.	Alpha, mean	Alpha, sdev.
171	1000	170	0.2250	0.0089
171	2000	170	0.2256	0.0112
171	7044	170	0.2603	0.0108

text into 183 separate files, comprising approximately between 1400 and 2000 words each. The number of unique words in these 183 files was found to be 28,631 words. The ultrametricity alpha values for this collection of 183 Joycean texts were found to be less than the Barbara Sanders values, but higher than the global set of all dream reports.

5 Conclusion

First, we took network weights to be either uniformly distributed or Gaussian-distributed. These distributions are highly "unclustered". Notwithstanding the fact that such data ought not to have inherent hierarchical structure, we in fact show (relying in part on [13]) that a Euclidean embedding leads to scaling properties such that the network nodes become increasingly ultrametric. As seen in this way, as networks grow they become more naturally hierarchical (but these hierarchies are not unique).

Then we looked at a range of text corpora, comprising about 1000 texts containing over one million words. We found very stable ultrametricity quantifications of the text collections, across numbers of most frequent words used to characterize the texts, and sampling of triplets of texts. We also found that in all cases (save, perhaps, the Brothers Grimm versus air accident reports) there was a clear distinction between the ultrametricity values of the text collections.

Some very intriguing ultrametricity characterizations were found in our work. For example, we found that the technical vocabulary of air accidents did not differ greatly in terms of inherent ultrametricity compared to the Brothers Grimm fairy tales. Secondly we found that novelist Austen's works were distinguishable from the Grimm fairy tales. Thirdly we found dream reports to be have higher ultrametricity level than the other text collections. Further exploration of these issues will require availability of very high quality textual data. Open questions include whether meaningful texts with far higher levels of ultrametricity can be found; whether data encodings could facilitate finding such higher values of alpha; and how do we best exploit the hierarchical structure that we find in textual data.

References

1. J. Austen, Sense and Sensibility (1811). Available at: http://www.pemberley.com/etext/SandS

2. J. Austen, Pride and Prejudice (1813). Available at: http://www.pemberley.com/etext/PandP
3. J. Austen, Persuasion (1817). Available at: http://www.pemberley.com/etext/Persuasion
4. A.-L. Barabási, "Self-organized networks: resources", at www.nd.edu/∼networks/database (2004).
5. J.P. Benzécri, L'Analyse des Données Tome 2, Correspondances, 2nd ed. (Dunod, Paris, 1979).
6. G. Caldarelli, A. Erzan and A. Vespignani, Eds., Special issue on Networks, European Physical Journal B **38**, no. 2 (2004).
7. L. Comtet, Advanced Combinatorics (Reidel, Dordrecht, 1974).
8. G.W. Domhoff, The Scientific Study of Dreams: Neural Networks, Cognitive Development and Content Analysis, American Psychological Association (2003).
9. R. Donaghey, "Alternating permutations and binary increasing trees", J. Combin. Theory (A) **18**, 141 (1975).
10. DreamBank, Repository of dream reports, www.dreambank.net (2004).
11. J.C. Gower, "Some distance properties of latent root and vector methods used in multivariate analysis", Biometrika **53**, 325 (1966).
12. F. Murtagh, "Counting dendrograms: a survey", Discrete Applied Math. **7**, 191 (1984).
13. F. Murtagh, "On ultrametricity, data coding, and computation", Journal of Classification **21**, 167 (2004).
14. F. Murtagh, "Identifying the ultrametricity of time series", European Physical Journal B **43**, 573 (2005).
15. F. Murtagh, Correspondence Analysis and Data Coding with Java and R (Chapman and Hall/CRC Press, New York, 2005).
16. Aviation Accident Database and Synopses, National Transport Safety Board, accessible from http://www.landings.com (2003).
17. J.M. Ockerbloom, Grimms' Fairy Tales, http://www-2.cs.cmu.edu/∼spok/grimmtmp (2003).
18. R. Rammal, G. Toulouse and M.A. Virasoro, "Ultrametricity for physicists", Reviews of Modern Physics **58**, 765 (1986).
19. A. Schneider and G.W. Domhoff, The quantitative study of dreams, http://dreamresearch.net (2004).
20. W.S. Torgerson, Theory and Methods of Scaling (Wiley, New York, 1958).
21. A. Trusina, S. Maslov, P. Minnhagen and K. Sneppen, "Hierarchy measures in complex networks", Physical Review Letters **92**, 178702(4) (2004).