ARTÍCULO

Some issues on complex networks for author characterization

Lucas Antiqueira[1,3], Thiago Alexandre Salgueiro Pardo[1,3], Maria das Graças Volpe Nunes[1,3] and Osvaldo N. Oliveira Jr.[2,3]

[1]Instituto de Ciências Matemáticas e de Computação (ICMC), Universidade de São Paulo (USP) – São Carlos, SP, Brasil

[2]Instituto de Física de São Carlos (IFSC), Universidade de São Paulo (USP) – São Carlos, SP, Brasil

[3]Núcleo Interinstitucional de Lingüística Computacional (NILC), www.nilc.icmc.usp.br

lantiq@gmail.com, taspardo@icmc.usp.br, gracan@icmc.usp.br, chu@ifsc.usp.br

Abstract

This paper presents a modeling technique of texts as complex networks and the investigation of the correlation between the properties of such networks and author characteristics. In an experiment with several books from eight authors, we show that the networks produced for each author tend to have specific features, which indicates that complex networks can capture author characteristics and, therefore, could be used for the traditional task of authorship identification.

Keywords: Authorship Attribution, Complex Networks.

1 Introduction

Recent trends in Natural Language Processing (NLP) have indicated that graphs are a powerful modeling and problem-solving technique. Graph theory, a branch of mathematics, has been studied for a long time, providing elegant and simple solutions to several problems. Recently, the field of complex networks, an intersection between graph theory and statistical mechanics, has motivated renewed interest in modeling systems as graphs. As evidence, graphs have been used in various applications, including text summarization [10, 14], information retrieval and related tasks [19, 13, 18], and sentiment analysis in texts [20].

We focus in this work on graph representation of texts, using measures usually employed in the studies of complex networks. Differently from regular networks, the dynamics and topology of a complex network follow non-trivial organization principles [17, 5]. Complex networks concepts have been applied to a number of world phenomena, e.g., internet, social networks and flight routes, which includes NLP and linguistics problems - as we briefly introduce in the next section.

We tackled the problem of authorship characterization using a network of words and a set of network measures. As known, authors present different styles of writing, resulting in diverse information flows and text structures. We therefore

look for correlations between the texts of a same author and the measures extracted from the corresponding networks. In this paper we show an experiment made with several books of 8 authors from diverse text genres, including fiction, science and poetry. We show that different measures vary considerably between some authors, which encourages the use of complex networks in the task of authorship identification.

In the next section we introduce complex networks, its main concepts and the measures used in this work. The methodology we follow and the obtained results are presented in Section 3 and 4, respectively. Some conclusions are presented in Section 5.

2 Complex Networks and Language

Complex networks have received enormous attention recently, but their origins can be traced back to the first study on graph theory, i.e., the Leonhard Euler solution to the Königsberg bridges problem in 1736 [3]. The concept of nodes linked by edges (used by Euler) provides a general mathematical way to describe discrete elements and their inter-relationships. Erdös and Rényi, two centuries later, created a way to explain the formation of real networks, which is called random graph theory. They published a series of eight papers, started in 1959 [9], that established for decades a random view of our world, an egalitarian world in which almost all nodes have a similar number of connections.

Alternatives to the random graph theory were later developed, which motivated a whole new interest in the study of networks, showing that many real networks are not random. Small-world networks were introduced more recently (e.g., Milgram [15] and Watts and Strogatz [24]), which differs from random networks in the sense that in such networks it is possible to reach every node through a relatively small number of other nodes. In other words, a small-world network has a small mean shortest path length (it also increases slowly - logarithmically - as the network grows). Moreover, these networks have a high clustering coefficient (defined later in this section), which is a tendency to form local groups of interconnected nodes. Barabási and Albert [4], introduced another special class of networks, named scale-free networks. These networks are also markedly different from the random networks, because their distributions of connections are not uniform; instead, a scale-free network has few nodes with a high number of connections (called hubs), which coexist with a much higher number of nodes that have a small number of connections. Scale-free networks are explained by two related concepts, (i) growth, which determines that a network is continually incorporating new nodes, and (ii) preferential attachment, which defines that new nodes prefer to establish connections with already well connected nodes.

Erdös and Rényi's random graph theory is a good example of an early study on complex networks, but such networks only started to receive an enormous attention from the scientific community in the recent years, after the publication of the Watts-Strogatz and Barabási-Albert papers. The modern reserch on complex networks typically incorporates concepts from mechanical statistics, and aims to characterize networks in terms of their structure and dynamics. A series of network models and measurements have been applied or created in recent network research [8], forming a profusion of tools available to almost every network study. In our present paper, we use a set of complex network measures (introduced in the next paragraphs) in order to face the problem of authorship characterization.

The tendency of the network nodes to form local interconnected groups is quantified by a measure referred to as *clustering coefficient*. For computing this coefficient for a node i in a directed network, consider S as the set of nodes that have an input edge from i, $|S|$ as the number of nodes in S, and B the number of edges among the nodes in S. Equation (1) computes the clustering coefficient of a node i:

$$\text{Clustering Coefficient } (i) = \frac{B}{|S|(|S|-1)}. \quad (1)$$

If $|S| = 0$ or $|S| = 1$, the coefficient is set to 0. The clustering coefficient of a network is the average coefficient of its complete set of nodes.

We have created another measure, based on the dynamical connectivity of the network growth. This type of measure is typical in complex network studies, since it considers the evolution of a network. It is given by the number of connected components as edges are progressively incorporated or have their associated weights increased

in the network[1,2]. The obtained dynamics is then compared to a hypothetical uniform variation of the number of components (Figure 1), and the extent to which the real dynamics departs from the hypothetical one is quantified by a measure called *components dynamics deviation*. This deviation is obtained for a network as follows:

$$\text{Deviation} = \frac{\sum_{k=1}^{E} |f(k) - g(k)|}{NE}, \qquad (2)$$

where $f(k)$ is the function that determines the number of components for k considered edge modifications (insertions or weight changes) and $g(k)$ is the function that determines the linear variation of components (see Figure 1). N is the number of different words in the text and E is the total number of edges modifications.

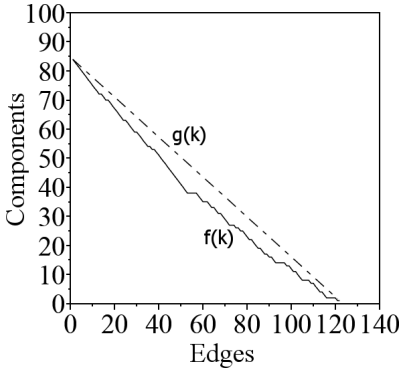

Figure 1. Component variation (vertical axis) in a network with weighted edges. The hypothetical uniform variation of the number of components is indicated by a dotted line ($g(k)$), which starts and ends, respectively, at the initial and final points of the real curve ($f(k)$). The horizontal axis indicates the number of edges inserted or modified (in respect of its weight).

As for traditional graphs, the input degree of a node is the number of edges that it receives from other nodes; similarly, the output degree is the number of edges that leave from it to other nodes. The *(in/out) degree* of a network is the average (in/out) degree of its set of nodes. Finally, the *degree correlation*, another typical complex network measure, is the Pearson correlation coefficient[3] [6] of a bidimensional data set that associates the degrees present at both ends of each edge. In other words, since each edge associates two nodes n_1 and n_2, it is possible to create a set of bidimensional points that represents the edges. Each edge is then denoted by the coordinates (degree of n_1, degree of n_2)[4]. The degree correlation is the Pearson correlation coefficient applied to this set of points.

Complex networks have been applied to several NLP and linguistics tasks. Sigman and Cecchi [23] modeled WordNet semantic relations between words into a network of word meanings, where the presence of highly polysemic words led to a small-world network. Another semantic network, which was derived from a thesaurus, was analysed by Motter et al. [16]. This network was found to be scale-free. In fact, this model is very common in linguistic networks, since it also conforms to (i) the word co-occurrence network, which models the sequence of words in a text [11], (ii) the word association network, where related concepts are linked [7], and (iii) the syntactic dependency network, which models syntactic relationships between words [12]. Antiqueira et al. [1, 2] studied the correlation among text quality and network properties, considering the words as network nodes, with edges being established between adjacent words in the text (word co-occurence model). Further details of how this modeling is processed appear in the next section. In an experiment with essays written by high school students, the authors demonstrated that text quality - as assessed by human evaluators - correlated with the clustering coefficient, the network degree (i.e., the mean degree) and the dynamics of network growth. Finally, following the same methodology, Pardo et al. [21, 22] showed the possibility of evaluating the quality of human and automatic summaries by means of network properties.

3 Representing Texts as Complex Networks

For representing texts as complex networks, we follow the proposal of Antiqueira et al. [1, 2]. Before building a network for a text, the text must

[1]A connected component is a subgraph that has no isolated nodes. Moreover, it is not possible to reach one node from a node that belongs to a different component.

[2]If an already existing edge is inserted in a network, it is possible to increase the number of times the edge was inserted and associate this number (the weight) to the edge.

[3]The Pearson correlation coefficient r quantifies the strength of the linear relationship between two variables ($0 \leq |r| \leq 1$). The sign of the coefficient gives the slope of the relation.

[4]For directed networks: (indegree of n_1, indegree of n_2), (indegree of n_1, outdegree of n_2) and so on.

be pre-processed. As traditionally done in NLP systems, the stopwords are removed, since they are very common and typically irrelevant, and the remaining words are lemmatized. Then, each word in the pre-processed text is represented as a node in the network and directed edges are established between nodes that represent adjacent words, with the node for the first word pointing to the node associated to the following word. We do not take into account sentence and paragraph boundaries for determining the adjacent words, i.e., the last word of a sentence/paragraph is considered adjacent to the first word of the next sentence/paragraph. Additionally, the edges are weighted by the number of times that the corresponding adjacent words appear in the text.

It is worth noting that the degree correlation (defined in Section 2) should take into account the edges directions, since the word co-occurrence network has outdegrees and indegrees. Since for every node the outdegree is equal to the indegree, the edges direction becomes irrelevant for the computation of the degree correlation. In fact, the only exceptions are the first and last words, which do not have a predecessor word and a successor word, respectively. We then consider, only for the computation of the degree correlation, a link between the last and the first word of a given text. Moreover, in the case of the components dynamics deviation, the dynamics is obtained as new word associations are inserted in the network in the order they appear in the text. Antiqueira et al. claim that the components dynamics deviation possibly captures the flow of new concepts introduced in texts. This is why it is also used here.

Our study is related to complex networks in two main ways. First, similar networks were proven to be scale-free and small-world by Ferrer i Cancho and Solé [11]. Although their networks are not the same as ours (e.g., the authors used the British National Corpus), we have a strong indication that the networks here employed are complex networks. Second, and more important, we apply a set of measurements frequently adopted by recent researches in complex networks (see Section 2).

4 Experiment

In order to test the possibility of using complex network properties to characterize authors, we selected books from 8 authors from varied genres, namely, fiction, science and poetry. The choice of authors depended solely on the number of books available, which were collected from the Project Gutenberg website[5] that freely distribute e-books. Concerning the reproducibility of our experiment, Table 1 shows the selected authors, the books collected for each one and the genre they belong to. The different number of books per author reflects their current availability. For the measures we report here, this is not an important factor, since each book is processed separately.

We measured (i) the network outdegree (i.e., the average outdegree, which is equal to the average indegree), (ii) the clustering coefficient, (iii) the degree correlation, and (iv) the components dynamics deviation for each text considered. We show the average results and the standard deviation for each author in Table 2. By these numbers, one can see that, in most cases, the authors present very different average measures. For instance, while Charles Darwin has an average outdegree value of 13.091, Lewis Carroll has 4.871. Although we did not evaluate the statistical significance of these differences between average measures, we have a clue that some of the network measures presented here can be used to separate texts regarding their authorship.

While any of the measures could be used to characterize authors, it is unlikely that a single measure would suffice in distinguishing all of them. Therefore, it is also interesting to investigate correlations between each pair of measures, and this is performed in the scatter plots shown in Figures 2 and 3 (it is possible to build 6 scatter plots, but we show only the 2 most significant ones).

To analyse these plots, we recall that authors whose data appear in a more limited region of the scatter plot would tend to have a more consistent writing style - obviously assuming that the complex networks features are representative of such style. Taken all the plots together (not shown), Wordsworth, Darwin and Hardy appear to be the authors with a larger overall consistency. In the comparison with all other authors, Wordsworth is particularly consistent in style, probably because it was the only one for poetry. In contrast, Dickens, Carroll and Woolf appear as the least consistent.

[5] http://www.gutenberg.org

Table 1. Data for the experiment

Author	Genre	Books
Charles Darwin	Science	Coral Reefs The Effects of Cross and Self-Fertilisation in the Vegetable Kingdom On the Origin of Species The Descent of Man The Different Forms of Flowers on Plants of the Same Species The Voyage of the Beagle
Charles Dickens	Fiction	A Tale of Two Cities American Notes David Copperfield Great Expectations Hard Times Master Humphrey's Clock Oliver Twist The Old Curiosity Shop The Seven Poor Travelers
Ernest Hemingway	Fiction	The Garden of Eden Green Hills of Africa
Lewis Carroll	Fiction	Alice's Adventures in Wonderland Phantasmagoria and Other Poems Sylvie and Bruno The Hunting of the Snark Through the Looking-Glass
Pelham G. Wodehouse	Fiction	My Man Jeeves Tales of St. Austin's The Adventures of Sally The Clicking of Cuthbert The Gem Collector The Man with Two Left Feet The Pothunters The Swoop The White Feather
Thomas Hardy	Fiction	A Changed Man and Other Tales A Group of Noble Dames Desperate Remedies Far from the Madding Crowd The Dynasts The Hand of Ethelberta
Virginia Woolf	Fiction	Jacob's Room Night and Day The Voyage Out
William Wordsworth	Poetry	Lyrical Ballads, with a Few Other Poems Lyrical Ballads, with Other Poems - Vol. 1&2 Poems, in Two Volumes - Vol. 1&2 The Poetical Works of William Wordsworth - Vol. 1,2&3

The correlation between style and the network measures is also variable. For instance, Darwin style is more consistent with regard to clustering coefficient vs. degree correlation (Figure 3) than in component dynamics deviation vs. degree correlation (not shown), especially because of the dynamics. In addition, some of the measures are more correlated to each other, as in the case of outdegree vs. clustering coefficient (Figure 2). In this case, it is possible to note that these two measures are linearly correlated, as the Pearson coefficient shows (it is equal to 0.88, in a range from 0 to 1). One notes that the clustering coefficient vs. degree correlation (Figure 3) and dynamics vs. degree correlation (not shown) are the best and worst, respectively, in terms of distinguishing the authors. It should be stressed that superimposition of points in the scatter plot reflects similarity in writing style. If we now concentrate on the clustering coefficient vs. degree correlation scatter plot (Figure 3, the one with best distinguishing ability), we note that the largest difference in style appears between Darwin and Wordsworth. This is not surprising since the texts refer to scientific writing and poetry, respectively. On the other hand, the authors with the most similar styles are Wodehouse, Hardy and Woolf.

Such results suggest that complex networks are not only useful for capturing author characteristics, but that they also could be applied in the task of authorship identification. This is one of our future plans, as stated in the next section.

Table 2. Average results and standard deviation for the network measures

Author	Outdegree	Clust. Coeff.	Deg. Correl.	Comp. Dynamics
Charles Darwin	13.091 ± 2.978	0.068 ± 0.018	0.070 ± 0.008	0.165 ± 0.023
Charles Dickens	7.946 ± 3.447	0.044 ± 0.025	0.052 ± 0.027	0.151 ± 0.028
Ernest Hemingway	8.342 ± 1.145	0.067 ± 0.016	0.062 ± 0.015	0.143 ± 0.011
Lewis Carroll	4.871 ± 2.348	0.041 ± 0.031	0.058 ± 0.032	0.107 ± 0.039
Pelham G. Wodehouse	5.193 ± 1.022	0.033 ± 0.011	0.050 ± 0.021	0.123 ± 0.016
Thomas Hardy	6.719 ± 0.988	0.035 ± 0.010	0.048 ± 0.010	0.144 ± 0.016
Virginia Woolf	7.923 ± 2.641	0.046 ± 0.017	0.056 ± 0.018	0.164 ± 0.042
William Wordsworth	4.824 ± 1.748	0.020 ± 0.008	0.053 ± 0.009	0.119 ± 0.017

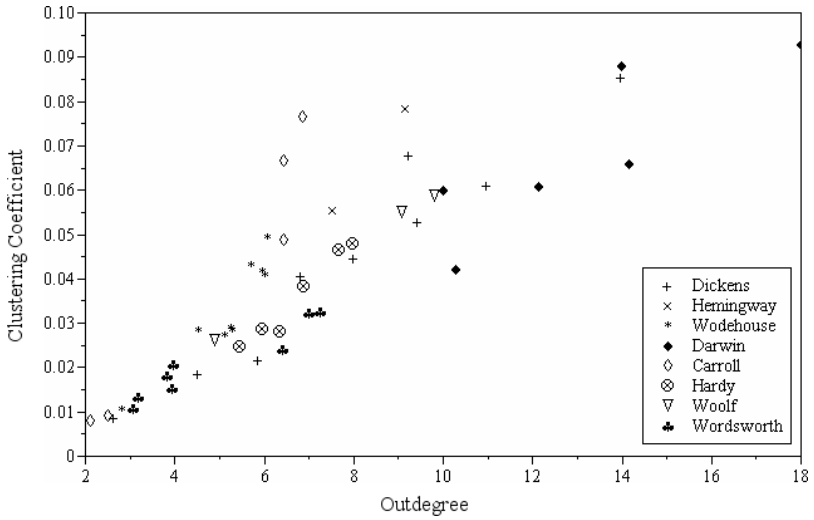

Figure 2. Scatter plot for outdegree (horizontal axis) and clustering coefficient (vertical axis)

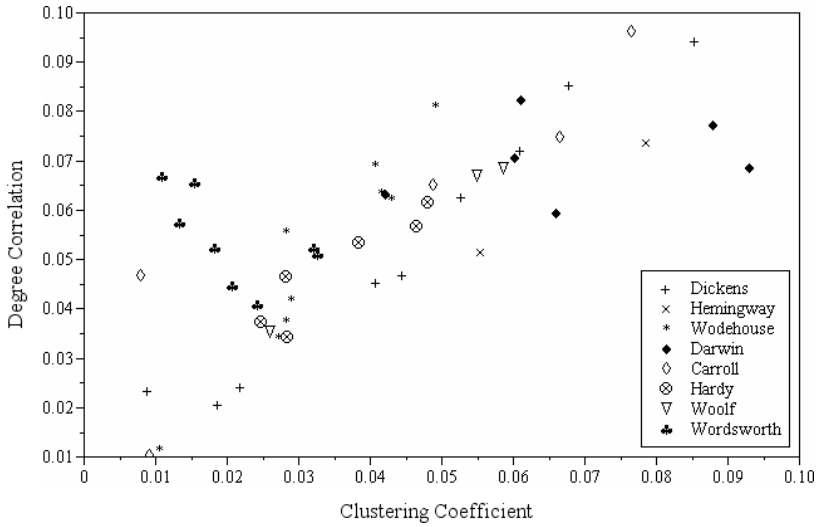

Figure 3. Scatter plot for clustering coefficient (horizontal axis) and degree correlation (vertical axis)

5 Conclusions

We showed in this paper how to model texts as networks and also introduced some network measures to be used in authorship characterization. The evaluation shows that it is possible to cluster some authors (Wordsworth, Darwin and Hardy) using combinations of three measures (ie. ploting clustering coefficient vs. outdegree or clustering coefficient vs. degree correlation). However, it is also necessary to carry out comprehensive and statistically significant tests in order to conclude with higher accuracy the usefulness of the measures. We intend to apply multivariate techniques such as Principal Components Analysis (PCA),

compare our results to other similar studies, and also apply more complex networks measures.

As in other NLP and linguistics tasks, concepts created in the field of complex networks show to be a valuable tool. Given the results presented in this paper, we believe it is possible to use complex networks for automatic authorship identification, as we plan to do in the future. Other possibilities include text categorization in general, but more investigation must be carried out first.

Acknowledgments

The authors are grateful to FAPESP and CNPq (Brazil) for financial support.

References

[1] L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr., and L. F. Costa. Modelling texts as complex networks (in Portuguese). In *Anais do XXV Congresso da Sociedade Brasileira de Computação (III Workshop em Tecnologia da Informação e da Linguagem Humana - TIL)*, 2089–2098, São Leopoldo-RS, Brasil, 2005.

[2] L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr., and L. F. Costa. Strong correlations between text quality and complex networks features. *Physica A*, 373:811–820, 2007. physics/0504033.v2.

[3] A. L. Barabási. *Linked: How Everything is Connected to Everything Else and What it Means for Business, Science and Everyday Life*. Plume, 2003.

[4] A. L. Barabási and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D.-U. Hwang. Complex networks: Structure and dynamics. *Physics Reports*, 424(4-5):175–308, 2006.

[6] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, 1990.

[7] L. F. Costa. What's in a name? *Int. J. Mod. Phys. C*, 15:371–379, 2004.

[8] L. F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[9] P. Erdös and A. Rényi. On random graphs I. *Publ. Math. Debrecen*, 6:290–297, 1959.

[10] G. Erkan and D. R. Radev. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479, 2004.

[11] R. Ferrer i Cancho and R. V. Solé. The small world of human language. *P. Roy. Soc. Lond. B Bio.*, 268:2261, 2001.

[12] R. Ferrer i Cancho, R. V. Solé, and R. Köhler. Patterns in syntactic dependency networks. *Phys. Rev. E*, 69:051915, 2004.

[13] Oren Kurland and Lillian Lee. PageRank without hyperlinks: Structural re-ranking using links induced by language models. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 306–313, 2005.

[14] R. Mihalcea. Language independent extractive summarization. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 49–52, Ann Arbor-MI, United States, June 2005. Association for Computational Linguistics.

[15] Stanley Milgram. The small world problem. *Psychology Today*, 2:60–67, 1967.

[16] A. E. Motter, A. P. S. Moura, Y. C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Phys. Rev. E*, 65:065102, 2002.

[17] M. E. J. Newman. The structure and function of complex networks. *SIAM Rev.*, 45:167–256, 2003.

[18] J. Otterbacher, G. Erkan, and D. R. Radev. Using random walks for question-focused sentence retrieval. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (HLT/EMNLP)*, 915–922, Vancouver, British Columbia, Canada, October 2005. Association for Computational Linguistics.

[19] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998. 17 p.

[20] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the ACL*, 271–278, 2004.

[21] T. A. S. Pardo, L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr., and L. F. Costa. Modeling and evaluating summaries using complex networks. In *Proceedings of the 7th Workshop on Computational Processing of Written and Spoken Portuguese (PROPOR)*, volume 3960 of *LNAI*, 1–10. Springer-Verlag, May 2006.

[22] T. A. S. Pardo, L. Antiqueira, M. G. V. Nunes, O. N. Oliveira Jr., and L. F. Costa. Using complex networks for language processing: The case of summary evaluation. In *Proceedings of the International Conference on Communications, Circuits and Systems (ICCCAS'06) - Special Session on Complex Networks*, 2678–2682, Gui Lin, China, June 2006. UESTC Press.

[23] M. Sigman and G. A. Cecchi. Global organization of the WordNet lexicon. *PNAS*, 99(3):1742–1747, 2002.

[24] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393:440–442, 1998.