COMPLEX NETWORKS: MODELLING KNOWLEDGE IN ASTROPHYSICS. THE CASE OF KH15D..

I.Joncour, *LAOG,414 rue de la piscine, Grenoble 38041 Cedex, France, (isabelle.joncour@obs.ujf-grenoble.fr).*

Abstract

Since a decade and the work of Watts and Strogatz (1998) and Barabasi Albert (1999), complex networks are used in various research fields to represent and to modelize complex systems as a whole for studying both their structure and their dynamics. Here, we propose a language-based approach for enhancing the dynamics of research through the statistical analysis of scientific papers. We applied that to the controversial nature of pre-main sequence star KH15D in order to experiment the methodology and to show first preliminary results.

Introduction

At the end of the XXe century, a new approach for dealing with social, natural and physical interacting dynamical systems has emerged with the so-called complex networks field combining graph theory and statistical physics.

Frequently cited examples of such real systems include the biological cell, a network composed of chemical species (nodes) interlinked by chemical reactions, the internet network composed by computers linked by physical connections, social networks (such as scientific citation networks)... These networks shared in common global and local properties which put them in a state which is neither totally disordered, nor totally organized. (see for a complete review, Albert et Barabasi, 2001; Dorogovtsev and Mendes, 2003). These non-equilibrium networks are typical of evolving self-organized structures.

Within this context of complex networks, semantic networks have been studied ; the nodes are words/concepts and links between the nodes semantic relationships such as, Synonymy/Antonymy, Meronymy/Holonymy ("part of" relation) or Hyponymy/Hypernymy ("kind of" relation)... Semantic networks are often used as a form of knowledge representation, the meaning of a concept being, at least, partly constituted by its connections to other concepts. I will focus on two types of works, the first one is performed on english thesaurus (a thesaurus is defined as a list of entries, the root words, followed by a list of terms belonging to semantic or hierarchical categories, Motter et al. 2002, Steyvers and Tenenbaum, 2005) the second one is performed on corpus, where the immediate co-occurence word networks are studied (Cancho and Solé, 2001; Milo et al., 2004). These works show that these semantic networks belong to the class of the other real networks ; complex networks are then especially well-suited to the study of semantic knowledge, where the structural general semantics seem to have significant implications for semantic growth (discovery).

Results

In this poster, statistical text analysis is carried out on a scientific paper in astrophysics field to build a complex linguistic network of specific "interacting" words proposed as a representation of the underlying background of scientific content. This approach offers an unique opportunity to cartography the conceptual current content of the scientific paper.

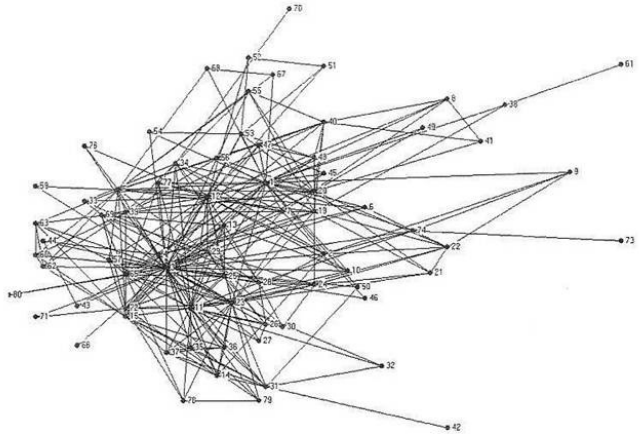

Figure 1: Scientific knowledge graph of the hamilton et al. work, 2001

As a preliminary work, we choose the pre-main sequence star KH15D which displays large variation in magnitude and for which the origin of the phenomena is not totally clear. This case is simple enough since it is based on observations and interpretations are in the context of star and planet formation. We start to analysis the the first study dedicated to this young object (hamilton et al., 2001), reporting observations and interpretation within the general paradigm of star and planet formation. In the context of complex networks, we define the nodes of the network as the words of the astronomical thesaurus (Shobbrook and Shobbrok, 1993) when included in the studied scientific paper. The compilation of this astronomy thesaurus was at first requested during a meeting of the International Astronomical Union at the New Delhi General Assembly of the IAU in 1984, to standardise the terminology in the field of astronomy. This thesaurus is then increased, when needed, by terms adapted to the specific thematic developped in the paper. The "interaction" between words is basically defined as the co-occurrence of two words in a same sentence and counted through the selected scientific paper, whatever the syntactic dependency between them. The number of co-occurrences between words defined the matrix element of the adjacency matrix used to built the network; such then, the bigger the number of co-occurrence between two words, the more they "interact", the smaller the distance between the two will be in the network.

In figure 1, preliminary visual results on the built graph are presented, where labels of central nodes may be seen in Table 1. Properties of this "scientific knowledge network" share the five usual features with other real complex networks:

1. Sparsity : the average number of links $\bar{k}$ per node is typically much smaller that the total number of nodes

Table 1: Table

N	*Node*	*N*	*Node*	*N*	*Node*	*N*	*Node*
1	Eclips	2	Dust	3	Star	4	Photomet
5	Spectr	12	Disk	13	Orbit	24	Circumstellar

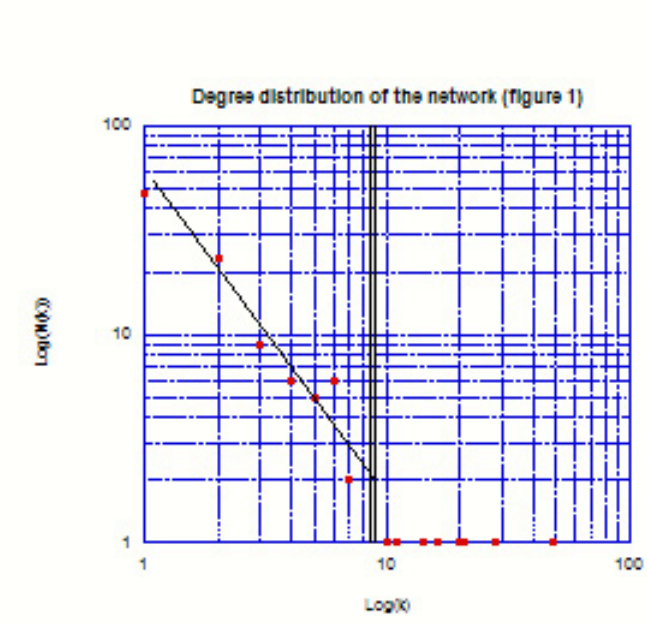

Figure 2: Degree spectrum Log(N(k) = fct(Log(k))

N (*ie* $\bar{k} \ll N$).

2. Connectedness : the network is composed by a single large connected components or is totally connected

3. Short Path length : the average minimal number of links $\bar{l} = 1/N^2 \sum_{i=1}^{N} \sum_{j=1}^{N} l_{min}(i,j)$ connecting two nodes randomly i and j chosen in the network is rather small and comparable to the path length in random network $l_{rand} = lnN/ln\bar{k}$

4. High local clustering : the clustering coefficient in real network, that is the probability C_i that two neighbours of a randomly chosen node i will themselves be neighbours, is high compared to the clustering coefficient in random network $C_i \ll C_{rand}$ with $C_i = T_i/N_i$, T_i the number of connections between the neighbours of node i in the network, $N_i = k_i(k_i - 1)/2$, the number of connections between the neighbours of node i in a fully connected network and $C_{rand} = \bar{k}/(N-1)$, the clustering coefficient in random network; a property shared with the totally ordered networks.

5. A scale-free distribution for the degree spectrum : the degree distribution, the probability that a randomly chosen node will a degree k (*ie* k neighbours), estimated with the frequencies N(k) of node degrees found through the network, shows no characteristic scale of node degree : all scales of connectivity showed simultaneously ($N(k) \propto k^{-\gamma}$), this long-tailed distribution is unlike random networks which obeys to an exponential's law distribution for the degree spectrum. Here, we find two regims for the degree spectrum, but different from the two regims seen by Cancho and Solé (2001) in their semantic network; work remains to be done to analyse the origin of the difference.

Conclusions

This work intends to propose a new methodology and visual tools to study scientific knowledge through the complex networks approach. Our first results show that the built scientific knowledge network shared common properties with other real networks.Future work remains to be done to fully interpret and analyse the networks built within a paper and to study the changes/modifications of the network following the chronological survey of publications on the subject.

We expect as a final result that these studies may help (1) to understand the discovery and innovation process (dynamics studies) and (2) to capture and enhance the underlying hierarchy of scientific knowledge levels (structure studies).

Acknowledgements *This work was done with the collaboration of A. Sandor (Xerox, Grenoble) using the linguistic tool "Xerox Incremental Parser" (XIP) tool (Roux, 1999),and of S. Robert (LANCI, uqam, Montreal)who granted the use of Ucinet, Pajek and Netdraw work package (Borgatti et al,1999) for network analysis). I thank also J. Bouvier (LAOG, Grenoble* for determining the choice of the KH15D for this study.

References

Albert, R., Barabási, A.-L., 2002, Reviews of modern Physics, 74, 47-97

Barabasi, A.-L., Albert R., 1999, Science, 286, 509-512,

Borgatti, S.P., Everett, M.G., Freeman, L.C., 1999, UCINET 6.0 Version1.00. Analytic Technologies, Natick, Massachusetts

Cancho, R.F. , Solé, Ricard V., 2001; Proceedings of the Royal Society B-Biological Science, 268, 2261-2265

Dorogovtsev S.N., Mendes J.F.F., 2003, Evolution of networks, from Biological Nets to the internet and WWW, Oxford University Press, Oxford

Hamilton, C. M., Herbst, W., Shih, C., Ferro, A. J. 2001, ApJ 554L, 201

Milo, R., Itzkovitz, S., Kashtan, N., Levitt, R., Shen-Orr, S., Ayzenshtat, I., Sheffer, M. and Alon, U., 2004, Science 303, 1538-1542.

Motter, A.E., de Moura, A.P.S., Lai, Y.-C., Dasgupta, P., 2002, Physical Review E 65, 065102

Roux, C., 1999, Proceedings of VEXTAL '99, Nov. 22-24,Venezia, San Servolo.

Shobbrook, R.M., Shobbrook, R.R., 1993, The Astronomy Thesaurus, Version 1.1, Anglo-Australian Observatory, Epping, Australia, 115 pp.

Steyvers, M., Tenenbaum, J.B, 2005, Cognitive Science 29, 41-78

Watts, J.D., Strogatz,S.H., 1998, Nature, 393, 440-442, Collective dynamics of Small-world networks.