Small-world semantic networks

Postma, E.O.; Wiesman, F.; van den Herik, H.J.

Published in:
BNAIC'00

Publication date:
2000

Small-world semantic networks

Eric Postma, Floris Wiesman and Jaap van den Herik

Institute for Knowledge and Agent Technology (IKAT)
International Institute of Infonomics (III)
Universiteit Maastricht, P.O.Box 616, 6200 MD Maastricht

Abstract

The efficiency of associating concepts in humans and machines depends on the structure of the underlying semantic network. This paper focuses on the associative properties of semantic networks by measuring two network characteristics: (i) the *clustering coefficient* (i.e., the tendency of concepts to cluster), and (ii) the *average association length* (i.e., the number of association steps between two concepts). We study these characteristics in two networks: a human semantic network derived from behavioural association data and a man-made semantic network derived from the ARIA database of the *Rijksmuseum*. Our results reveal that both semantic networks exhibit the *small-world* phenomenon: a large clustering coefficient is combined with a small average association length. It is argued that the small-world character of the human semantic network enables an efficient mental association of disparate concepts. We found that in the man-made semantic network the small-world effect is smaller than in the human semantic network. We suggest that the smaller effect is due to an imperfect representation of semantic associations in the artificial semantic network as compared to the human semantic network. Implications for theories and models of knowledge representation are discussed. It is concluded that the small-world analysis of semantic networks enhances our understanding of the statics and dynamics of associative structures in man and machine.

1. Semantic networks

Semantic networks play a central role in artificial intelligence and cognitive science. In their simplest form [2], semantic networks consist of a collection of nodes (i.e., concepts) with links representing the associations between concepts. Mental processes, such as thinking or reasoning, are modelled by the activation of concepts (nodes) and by the spreading of activation to associated nodes [2, 3]. According to the semantic network model, the dynamics of mental processes are strongly dependent on the underlying semantic-network structure. After all, the presence of a link between two concepts facilitates the spread of activation that results in the successive activation of both concepts as part of a mental process [10]. Cognitive scientists have stressed that the *organisation* of knowledge is crucial for the mental capabilities of humans [1,5]. Insofar as the interconnected concepts of a semantic-network model reflect the organisation of knowledge, it is clear that the nature of the connectivity pattern of semantic networks deserves closer inspection.

The present study aims at analysing a generic human semantic network by deliberately ignoring the underlying causes of the associations present in the network. The structure of the semantic network is represented by a connected graph of nodes (i.e., concepts) and edges (i.e., associations). As models of knowledge representation semantic networks have their limitations (see, e.g., [11]). However, in our analysis we employ semantic networks for *describing* associations, rather than modelling knowledge.

Through a computational analysis of two semantic networks: a human semantic network and a man-made semantic network, we show that both networks are so-called *small-world networks* [13]. In both cases, the association of concepts from disparate semantic domains is facilitated by their reduced average association lengths.

The outline of the paper is as follows. In section 2 the data and the resulting semantic-network models are discussed. Section 3 presents our method of analysis by discussing the seminal work of Watts and Strogatz [13] on small-world networks. Then, in section 4 the results of our analysis are presented. Section 5 discusses the results and proposes several applications in cognitive science and artificial intelligence. Finally, section 6 concludes by stating that the formal analysis of (inferred) semantic networks is a fruitful approach for studying the structure and development of knowledge representations in man and machine.

2. Data collection and models

Our starting point is a generic semantic network model in which concepts are either associated or not associated, i.e., links are present or absent. We defined two semantic networks. The first network reflects an "average" human semantic network; the second one is based on a man-made database of objects and their (shared) properties.

2.1 The human semantic network

Our approximation of the human semantic network in terms of nodes and connections requires a measurement of the associations between concepts in humans. To assess the presence of associative links in the human semantic network, we considered various publicly-available dictionaries and thesauri. The *Edinburgh Associative Thesaurus* database of word-association data published by Michael Wilson[1] on the Internet as part of the *Psych* psycholinguistic database turned out to be most suitable for our purposes. The word-association data were obtained in an experimental study [7] where subjects were asked to associate freely on target words. As the aim of the study was to "to obtain a reasonably large complete mapping of the associative network for a large set of words" [7], the appropriateness for our study is evident.

The experimental procedure ran as follows. Subjects were given sheets of target words and requested to write down the first word that came to mind for each word in the list. Most of the data were collected in a classroom setting under supervision. In case subjects failed to write down associated words for more than 25% of the target words, the response sheets were discarded. The procedure was repeated several times. After

[1] http://www.itd.clrc.ac.uk/Activity/Psych

each repetition, new response words were added to the list of target words. Finally, hundred subjects were tested with the stable set of words so obtained. For each word the responses and their frequency of occurrence was collected. As an example, table 1 lists the six most-frequent responses to the target words "apple" and "Newton".

target	most-frequent responses (percent mentioned)
APPLE	PIE (20) PEAR (17) ORANGE (13) TREE (8) CORE (7) FRUIT (4)
NEWTON	APPLE (22) ISAAC (15) LAW (8) ABBOT (6) PHYSICS (4) LAW (5)

Table 1. Two examples of target words and their six most-frequent responses.

Although the experiment is rather uncontrolled for, e.g., context effects, we believe that the results obtained in the experiment represent a genuine manifestation of the average associative structure of the human semantic network.

We translated the target words and responses into a graph representation where target words are represented by nodes. For each target word t, a response r generated by at least a proportion $\theta = 0.2$ of subjects was translated into an undirected edge (or associative link) between the concepts t and r, which introduced a few associations not present in the data (e.g., the association apple $\rightarrow$ Newton; see Table 1). The largest connected sub-graph was identified and used for further analysis.

2.2 The man-made semantic network

For creating a man-made semantic network we used the *ARIA* (*Amsterdam Rijksmuseum InterAction*) database of the Rijksmuseum in Amsterdam. The objects in the ARIA database are the most important art objects in the collection of the Rijksmuseum. To facilitate the search through the database (e.g., using the ANS metabrowser [14]), each object is assigned to a content-based class. For paintings, two examples of such classes are "interiors" and "family portraits". With these classes, users can navigate through the database by searching paintings of interiors or by searching for paintings depicting family portraits. We derived a semantic-network graph from a part of the ARIA database by equating objects to nodes and translating features shared by two objects into undirected edges. As with the human semantic network, the largest connected sub-graph was identified and used for analysis.

3. The large-world to small-world transition

Our method of analysis is based on Watts and Strogatz's [13] seminal work on large-world to small-world transitions in networks. We discuss their findings in terms of semantic networks. Let us consider a regular semantic network (i.e., regular graph) and a random semantic network (i.e., random graph) and characterise their structures by (i) a *clustering coefficient* C and (ii) an *average association length* L.

The clustering coefficient expresses the degree to which the concepts associated with a central concept are associated with each other. A high clustering coefficient signifies a large clique of concepts. For instance, the central concept *apple* is associated to the concepts *pear* and *orange* (cf. table 1), which, in turn, are also associated and therefore contribute to the clustering coefficient.

The average association length is the average number of associative links separating a pair of concepts along the shortest path. A high average association length indicates large distances among concepts from different semantic domains, such as, for instance, *apple* (as part of the semantic domain *fruit*) and *car* (as part of the semantic domain *vehicles*).

A regular semantic network is characterised by a large clustering coefficient and a large average path length. Alternatively, a random semantic network has a small clustering coefficient and a small average association length. Figure 1 illustrates both extremes. In the regular semantic network shown on the left of figure 1, concepts (the disks) are part of partially overlapping semantic domains as defined by their two nearest neighbouring concepts (which are also associated). The absence of long-range associative connections yields a large average association length in regular networks. In the random semantic network there are no semantic domains. However, the presence of long-range associative connections results in a small average association length.

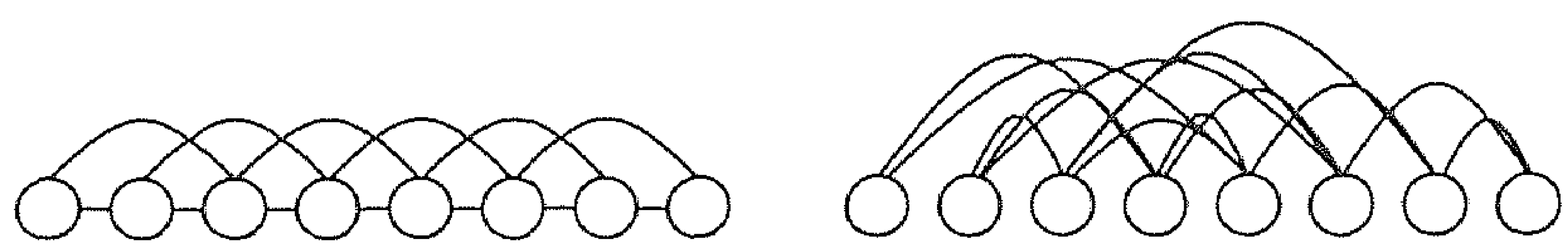

Figure 1. Schematic illustration of a (part of a) regular semantic network (left), and a random semantic network (right). The disks represent concepts and the lines represent associations between concepts. The average association length, i.e., the shortest path connecting two arbitrary concepts, is larger for the regular than for the random network.

Apparently, minimising the average path length results in the elimination of order (i.e., low clustering coefficient) in the network. Since ordered representations are the hallmark of semantic networks, this poses a serious problem when attempting to minimise association length in these networks. However, as Watts and Strogatz showed for general regular and random graphs [13], introducing only a small amount of randomness in a regular graph preserves the order (high clustering coefficient) while reducing the average path length drastically (see also [6]). Their results from computational analyses of artificial graphs are illustrated in figure 2. The figure shows the normalized values of C and L as a function of p, the probability of randomly reconnecting edges in an otherwise regular network. (A logarithmic scale is used for the horizontal axis.)

With only a small amount of randomness (small p), the average path length drops rapidly towards zero, whereas the clustering coefficient remains virtually unchanged. This implies that the transition from a large world to a small world is hardly detected at the microscopic or local scale.

Watts and Strogatz [13] applied their analysis to the collaboration graph of film actors, the neural network of the worm *Caenorhabditis elegans*, and the power grid of the western United States, and showed them to be small-world networks.

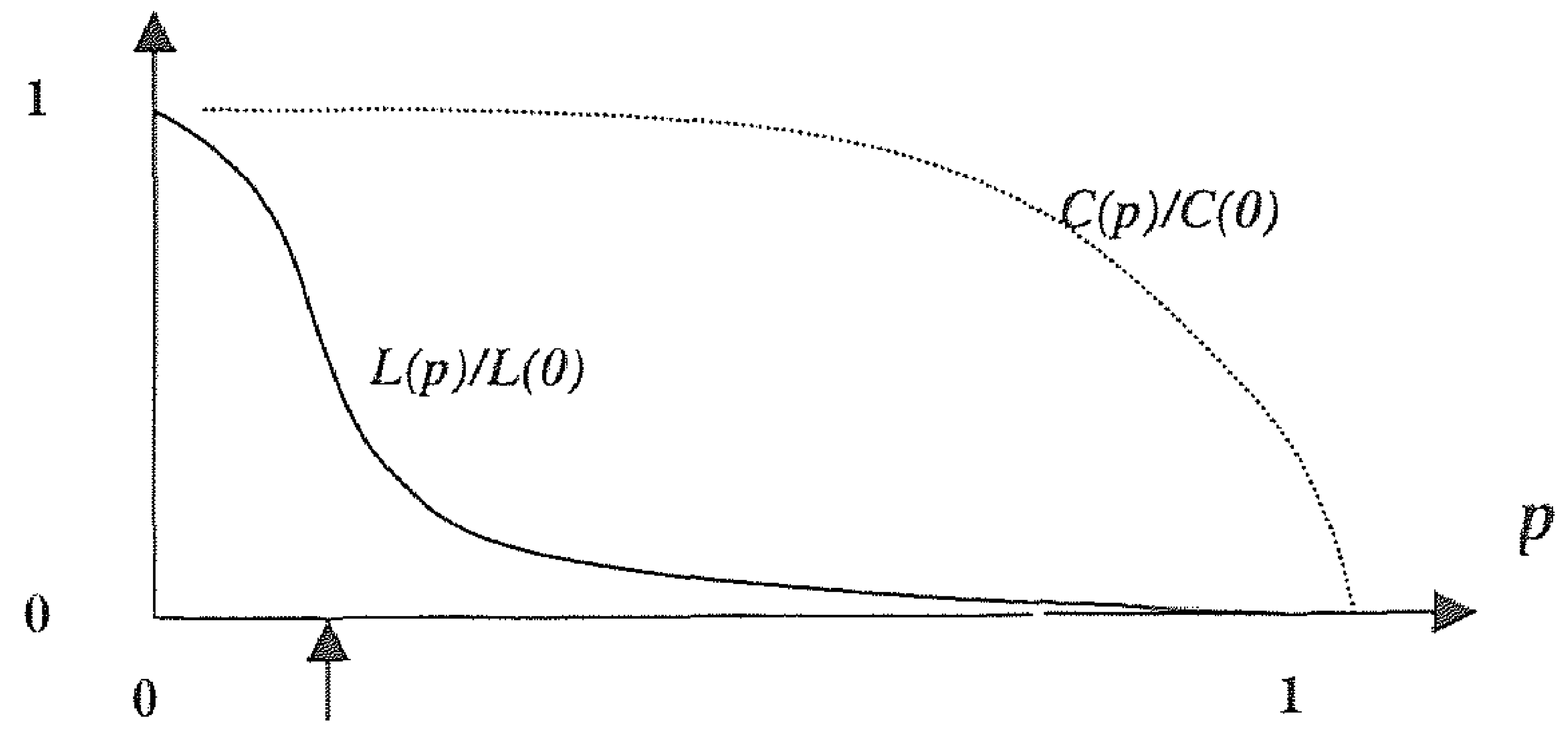

Figure 2. The large-world to small-world transition as evident from the normalized clustering coefficient C(p)/C(0) (dotted curve) and the normalised average path length L(p)/L(0) (solid curve), as a function of randomness p, drawn on a logarithmic scale (redrawn from [13]). At a small value of p (arrow), the path length decreases drastically whereas the clustering coefficient remains almost unchanged.

Framed in terms of the large-to-small world transition, our analysis aims at characterising semantic networks in terms of either a large or a small world. We measure the clustering coefficient and association length of the semantic network (p unknown) and compare their values with the case $p=1$ obtained by randomly rewiring the network while preserving the average number of associations per concept k. In case $L \approx L(1)$ and $C >> C(1)$ the human semantic network exhibits the small-world phenomenon.

4. Semantic network analysis

We submitted our semantic network to a computational analysis involving the calculation of the clustering coefficient and the average association length. The clustering coefficient was computed for each concept by counting the number of nearest-neighbour concepts (i.e., concepts connected through an association link with the concept under consideration). Subsequently, the number of associations among the nearest-neighbour concepts was assessed. Together, these numbers determine the fraction of associated nearest neighbours. The average association length was computed by averaging over the shortest paths between all pairs of concepts computed using Floyd-Warshall's algorithm [4].

4.1 Results

Table 2 displays the results for both semantic networks. The quantity $\mu = (C/L)/(C_{random}/L_{random})$ is a quantitative measure of "small-worldliness" defined by Walsh [12]. Small-world networks are characterised by a value $\mu >> 1$ whereas random graphs have a value $\mu = 1$. Evidently, both semantic networks exhibit the small-world phenomenon: a large clustering coefficient C is combined with a small average association path L. In addition, the human semantic network (HSN) has a larger small-worldliness than the man-made artificial semantic network (ASN).

	n	k	C	C_{random}	L	L_{random}	μ
HSN	2169	2.1	0.004	0.00013	15.01	18.54	34.5
ASN	708	54.2	0.537	0.07653	2.24	1.94	6.1

Table 2. Results of the analysis of the human semantic network (HSN, $\theta = 0.2$) and the artificial semantic network (ASN). Legend: n, the total number of concepts retained in the graph; k, the (average) number of associations per concept; C, the clustering coefficient; L, the average association length; and μ, the "small-worldliness" measure.

4.2 Mean-field approximation

Our empirical results can be considered in the context of analytical work on small-world networks to reveal the position on the L-curve (cf. the solid line in figure 1). Newman *et al.* [9] derived a mean-field solution of small-world networks yielding an approximation of the average association path L as a function of p, n, and k:

$$L(p,k,n) = \frac{1}{2p\left(\frac{k}{2}\right)^2\sqrt{1+\frac{2}{p(k/2)n}}}\tanh^{-1}\frac{1}{\sqrt{1+\frac{2}{p(k/2)n}}}. \qquad (1)$$

Using this equation we plotted L as a function of p for the two semantic networks using their computed values of n and k (cf. table 2). The results are displayed in figure 3. Our empirical values for L are represented by the black dots. Evidently, in terms of the average association length, the human semantic network is closer to a random network (i.e., p = 1) than the artificial semantic network. Taken together with our empirical results this implies that the human semantic network exhibits a more efficient structure for fast spreading of activation than the man-made semantic network.

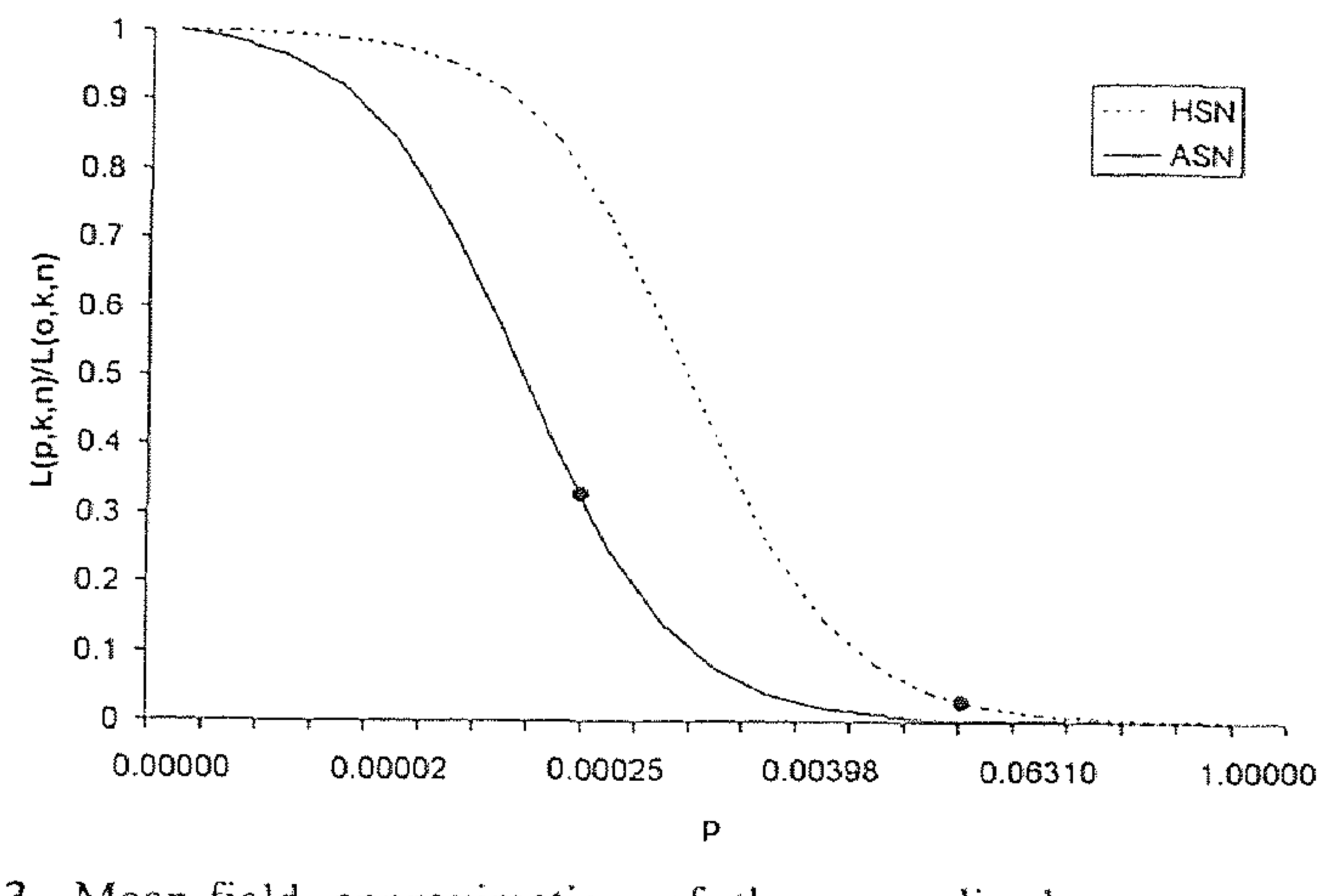

Figure 3. Mean-field approximation of the normalised average association path $L(p,k,n)/L(0,p,n)$ as a function of the random-rewiring probability p, for the human semantic network (HSN) and the artificial semantic network (ASN). The black dots correspond to the empirical values of L for both networks (cf. table 1).

5. General discussion

The small-world character of the (inferred) human semantic network and the man-made ARIA semantic network suggests that mental representations obey the laws of physical network structures such as societal networks and energy-distribution networks [13]. Our results imply that the human semantic network as derived from our data, combines a high degree of structure (a high value of C/C_{rand}) with small average association lengths (a low value of L/L_{rand}). Structure is required to represent closely related concepts such as *apple*, *pear*, and *orange* into a tight cluster. A moderate number of random associations (i.e., associations between concepts from different semantic domains) is required to be able to associate concepts from different clusters into new meaningful wholes. In terms of our example of section 3, the (alleged) discovery of the laws of gravitation associates the concepts *apple* and *Newton* that belong to different semantic domains. An important consequence is that the associational distance between concepts related to the newly associated concepts, e.g., *banana* and *Einstein*, is reduced too. This effect lies at the heart of the small-world transition as it is responsible for the sudden decrease in average association length. In addition, there is some psychological evidence for the existence of the effect [8].

6. Conclusions

The omnipresence of semantic-networks structures in artificial intelligence and cognitive science is an indication of their importance for modelling and understanding knowledge processes. Our analysis of a human and a man-made semantic network is only a first step towards a formal approach of the associative structure of semantic networks. Nevertheless, the main result indicates the feasibility of applying graph analysis to network-based knowledge representations. The small-worldliness μ expresses the association efficiency in a network-based knowledge structure. Therefore, in terms of spreading activation, semantic networks with a large μ value are good networks for making associating arbitrary pairs of concepts as required for mental processes and searching databases. We conclude by stating that the small-world analysis of natural and man-made semantic networks is a fruitful approach for studying the statics and dynamics of associative structures in man and machine.

Acknowledgement

This study is performed as part of the NWO (*The Netherlands Organisation for Scientific Research*) ToKeN2000 project which studies the access and disclosure of knowledge in The Netherlands. We gratefully acknowledge the Amsterdam Rijksmuseum for providing us their ARIA database in the context of the ToKeN2000 project. Also we thank the anonymous referees, Leo Plugge and Alard Roebroeck for their helpful comments on an earlier version of this manuscript.

References

[1] W.G. Chase and H.A. Simon. The mind's eye in chess. In W.G. Chase, editor, *Visual Information Processing*, pp. 215-281, Academic Press, New York, 1973.

[2] A.M. Collins and E.F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82: 407-429, 1975.

[3] A.M. Collins and M.R. Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8: 240-248, 1969.

[4] R.W. Floyd. Algorithm 97: Shortest path. *Communications of the ACM*, 5: 345, 1962

[5] A.D. de Groot. *Thought and choice in chess*. Mouton, The Hague, Netherlands, 1965.

[6] H. Herzel. How to quantify 'small-world networks'? *Fractals*, 6(4): 301-303, 1998.

[7] G.R. Kiss, C. Armstrong, R. Milroy, and J. Piper. An associative thesaurus of English and its computer analysis. In A.J. Aitken, R.W. Bailey, and N. Hamilton-Smith, editors, *The Computer and Literary Studies*, University Press, Edinburgh, 1973.

[8] J.H. Neely. Semantic priming effects in visual word recognition: A selective review of current findings and theories. In D. Besner and G.W. Humphreys, editors, *Basic processes in reading: Visual word recognition*. Lawrence Erlbaum Associates, Hillsdale, NJ, 1991.

[9] M.E.J. Newman, C. Moore, and D.J. Watts. Mean-field solution of the small-world network model. Submitted to *Physical Review Letters*, 1999. http://acoma.santafe.edu/sfi/publications/Abstracts/99-09-066abs.html

[10] D.A. Norman and D.E. Rumelhart. *Explorations in cognition*. Freeman, San Francisco, 1975.

[11] S.J. Russel and P. Norvig. *Artificial intelligence: a modern approach*. Prentice Hall, New Jersey, 1995.

[12] T. Walsh. Search in a small world. In T. Dean, editor, *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI'99), Volume* 2, pp. 1172-1177, Morgan Kaufmann Publishers, San Francisco, 1999.

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

[14] F. Wiesman. *Information retrieval by graphically browsing meta-information*. Dissertation Universiteit Maastricht. Phidippides, Cadier en Keer, 1998.