

Categorizing words through semantic memory navigation

J. Borge-Holthoefer¹ and A. Arenas¹

Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, 43007 Tarragona, Catalonia, Spain

Received: date / Revised version: date

Abstract. Semantic memory is the cognitive system devoted to storage and retrieval of conceptual knowledge. Empirical data indicate that semantic memory is organized in a network structure. Everyday experience shows that word search and retrieval processes provide fluent and coherent speech, i.e. are efficient. This implies either that semantic memory encodes, besides thousands of words, different kind of links for different relationships (introducing greater complexity and storage costs), or that the structure evolves facilitating the differentiation between long-lasting semantic relations from incidental, phenomenological ones. Assuming the latter possibility, we explore a mechanism to disentangle the underlying semantic backbone which comprises conceptual structure (extraction of categorical relations between pairs of words), from the rest of information present in the structure. To this end, we first present and characterize an empirical data set modeled as a network, then we simulate a stochastic cognitive navigation on this topology. We schematize this latter process as uncorrelated random walks from node to node, which converge to a feature vectors network. By doing so we both introduce a novel mechanism for information retrieval, and point at the problem of category formation in close connection to linguistic and non-linguistic experience.

PACS. 89.75.Hc Networks and genealogical trees – 89.75.Fb Structures and organization in complex systems

1 Introduction

Semantic memory is the cognitive system where conceptual knowledge is stored. Empirical evidence from experiments with subjects and other lexical resources (*thesauri* [1], *corpus* [2], etc.) suggest that this system can be suit-

ably represented as a semantic network, where each node corresponds to a word, and edges stand as pairwise associations. The network reconstructed from semantic information is in contrast with hierarchies created by individuals for computer storage and retrieval -which are

trees- [3], the network has an intricate topology of cyclic relationships. Estimations that on average a healthy adult knows from 20000 to 40000 words [4] raise challenging questions about storage capacity, organization of the information and verbal performance. Regarding organization, some words are linked by virtue of their semantic similarity (intra-categorical relations, e.g. *car* and *automobile*). Other types of associations fall under the more general semantic relatedness, which includes the former and any kind of functional or frequent association [5], e.g. *car* and *road*. This implies that many types of association exist undistinguished in the network structure. In particular, categorical (similarity) relations are embedded in a much richer structure of superposed relationships.

In this article we propose a computational model to extract semantic similarity information from the track of a dynamical process upon word association data. The main idea is that categorical relations emerge from navigation on the topology of semantic memory. Although we focus on cognitive phenomena and data, our efforts can be more generally interpreted in terms of the extraction of the backbone of a network, which entails that there exist “master relations” between elements (long-lasting similarity relations) and “incidental” (experience-dependent) ones that are entangled with the previous.

We use two empirical data sets to test the model: a general association semantic network as substrate of a dynamic process, and a feature similarity network for comparison purposes. Both are characterized in the next section. After that, the model itself is detailed. We name

it the Random Inheritance Model (RIM) because it is based on uncorrelated *random walks* from node to node that propagate an inheritance mechanism among words. The results obtained yield significant success both at the macro- and the microscopic level when compared to actual data. Finally, we discuss that the key to such success is the modular structure of the substrate network, which retains significant meta-similitude relationships.

2 Topology of semantic networks

Before focusing on the model it is necessary to characterize the data under consideration. The algorithm that implements our model runs on general word association data, which are typically called Free Association. It is widely accepted that such data offer the most general and realistic insight of the structure of semantic memory, because they are not restricted to a particular kind of association. On the contrary, feature similarity data reports only the amount of features two words have in common, thus displaying strictly pairwise similarity information.

2.1 Free-Association Norms

Nelson *et al.* collected these norms (FA from now on) by asking over 6000 participants to produce (write down) the first word (*target*) that came to their mind when confronted with a *cue* (word presented to the subject) [6]. The experiment was performed using more than 5000 distinct cues. Among other information, a frequency of coincidence between subjects for each pair of words is obtained.

As an example, words *mice* and *cheese* are neighbors in this database, because a large fraction of the subjects produced the target *mice* in response to the cue *cheese*. Note, however, that the association of these two words is due to their similarity but other relationships (in this case mice eat cheese). The network empirically obtained is directed (asymmetric) and weighted, weights represent the frequency of association in the sample. We maintain the asymmetry property in our approach to preserve the meaning of the empirical data.

2.2 Feature Production Norms

Feature Production Norms (FP from now on) were collected by McRae *et al.* [7] by asking subjects to produce features when confronted with a certain word. This feature collection is used to build up a vector of characteristics for each word, where each dimension represents a feature. The value of each component of the final vector represents the production frequency of the corresponding feature across participants. These norms include 541 concepts. Semantic similarity is computed as the cosine (overlap) between pairs of vectors of characteristics, obtained as the dot product between two concept vectors, divided by the product of their lengths. For example, words like *banjo* and *accordion* are very similar (i.e. they have a projection close to 1) because they share many features as musical instruments, their vector representations show a high overlap. On the contrary, vectors for *banjo* and *spider* are very different, showing an overlap close to 0 (orthogonal vectors). In terms of network representation an edge

is laid between a pair of nodes whenever their vectors projection is different from 0, and its weight is the features similarity between the two words. The network is thus undirected (symmetric relationships).

The differences in the nature of edges has drastic effects on the topology of these semantic networks, this can be analyzed in terms of statistical descriptors. In table 1 we highlight some of such descriptors. $\langle s \rangle$ is the average strength per node; L is the average path length, defined as the average of the geodesic paths (minimal distance) between any pair of nodes; D is the diameter of the network, i.e. the longest geodesic path in the network; C_i is the clustering coefficient of a single node, its average across N (network size) is indicative of the cohesion in data. Strength distribution $P(s)$ is a cumulative distribution function, which gives the probability that the strength of a node is greater than or equal to s . It is helpful to gain a global vision of a network's connectivity profile, in fig. 3 we see FA's and FP's distributions. A complete review of these descriptors can be found in [8–10].

It is readily understood from table 1 that the structures differ largely. The high connectivity in FP gives raise to a dense network, which in turn allows that any node is reachable in less than 2 steps on average. It also has the effect of a highly cohesive structure, i.e. clustering is prominent. In order to avoid size effects (the difference between FA and FP sizes), the same statistics are computed for the common subset of words, the differences between both topologies still hold. Strength distribution, which is

plotted for FA's and FP's common subgraphs, also evidences deep structural disagreement, fig. 3.

Table 1. Main statistical descriptors of the networks FA and FP, and their respective common words' subnetworks. N is the number of nodes; $\langle s \rangle$ is the average strength; L is the average shortest path length; D is the diameter of the network and C is clustering coefficient.

	FA (all)	FP (all)	FA (subset)	FP (subset)
N	5018	541	376	376
$\langle s \rangle$	0.77	20.20	0.26	13.43
L	3.04	1.68	4.41	1.68
D	5	5	9	3
C	0.1862	0.6344	0.1926	0.6253

We have analyzed quantities that describe macro and micro levels of networks. Also at the level of groups or communities (mesoscale) differences arise between FA and FP. This is expected, both because reviewed topological features differ largely, and the semantics of links is different from construction. Modularity optimization methods [11–13] yield partitions in which groups of words are gathered differently. The statistical significance of modularity is performed in a sample obtained by randomizing the original network and applying the same method of optimization [14]. FA shows a highly modular structure $Q = 0.6162$, compared to its random counterpart $Q = 0.091 \pm 0.001$. FP reaches a modularity value $Q = 0.4288$ also very significant compared to its random counterpart $Q = 0.323 \pm 0.002$. Lower modularity implies that clear boundaries are harder to define, this fits well with evi-

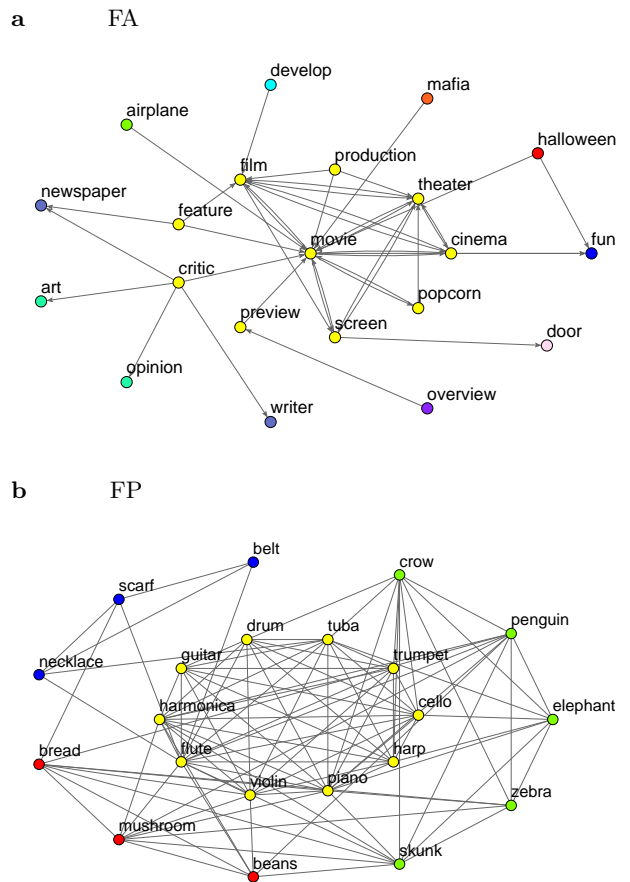


Fig. 1. A sample of words that conform communities, from partitions obtained through modularity optimization in (a) FA and (b) FP. For the sake of simplicity edges leaving the depicted subgraph have been removed (color online).

dence of humans' fuzzy categorical system [15] and with computational models of verbal fluency [16]. Despite this, a close look to the words that conform communities, either in FA or FP, correctly reflect the distinct underlying associations, see fig. 1.

3 The Random Inheritance Model (RIM)

Up to now we have some clues about the type of topology our algorithm will be run on (FA), and what the output of

the model should resemble (FP). From this knowledge we move on to develop the logic steps behind our proposal and describe the mathematical framework behind it. Recent works have pointed out the ability of a random navigation to explore the complexity of networks [17–19]. Here we propose a random navigation process and an inheritance mechanism to disentangle categorical relationships from a semantic network. Our intuition about the expected success of our approach relies on two facts: the modular structure of the FA network retains significant meta-similitude relationships, and random walks are the simplest dynamical processes capable of revealing the local neighborhoods of nodes when they persistently get trapped into modules. The inheritance mechanism is a simple reinforcement of similarities within these groups. We call this algorithm the Random Inheritance Model (RIM).

The RIM proceeds in three steps, (i) initialization, (ii) navigation and inheritance, and (iii) output construction. Step (i) tags every word in the FA network with an initial features vector. The vectors are orthogonal in the canonical basis to avoid initial bias. That means that every word has associated a vector of N -dimensions, being N the size of the network, with a component at 1 and the rest at zero. The second step consists of launching random walks of length S from every word i in the network. The inheritance mechanism changes the vector of i , v_i depending on the navigation behavior. Let $s = \{s_1, s_2, \dots, s_n\}$ the set of visited nodes. Then the new vector for node i is computed as:

$$v_i = \sum_{s_i \in s} v_{s_i} \quad (1)$$

Note that (a) update of the feature vectors is synchronized, final values are computed after completion of the inheritance for every word; and (b) a random walk is a time-reversible finite Markov chain, which implies that node i can be itself in the set of visited nodes, see [20] for a survey on the topic. A new (synthetic) network FS is built in step (iii). Nodes in the new structure are those from the substrate network, weights between them are the result of projecting all pairs of updated vectors.

Steps (i)-(iii) are iterated (by simulating several runs) up to convergence of the average of the synthetic feature similarity networks generated at each run. The final average is the synthetic feature similarity network to be compared to FP.

This algorithm can be algebraically described in terms of Markov chains. Before we must define the transition probability of the FA network. The elements of FA (a_{ij}) correspond to frequency of first association reported in [6]. However, note that the 5018 words that appear on the data set are not all the words that appeared in the experiment, but only those that were at the same time cues in the experiment. Therefore data need to be normalized before having a transition probability matrix. We define the transition probability matrix P as:

$$P_{ij} = \frac{a_{ij}}{\sum_j a_{ij}} \quad (2)$$

As the original matrix, this one is also asymmetric. Once the matrix P is constructed, the random walkers of different lengths are simply represented by powers of P . In practice, this means that if we perform random

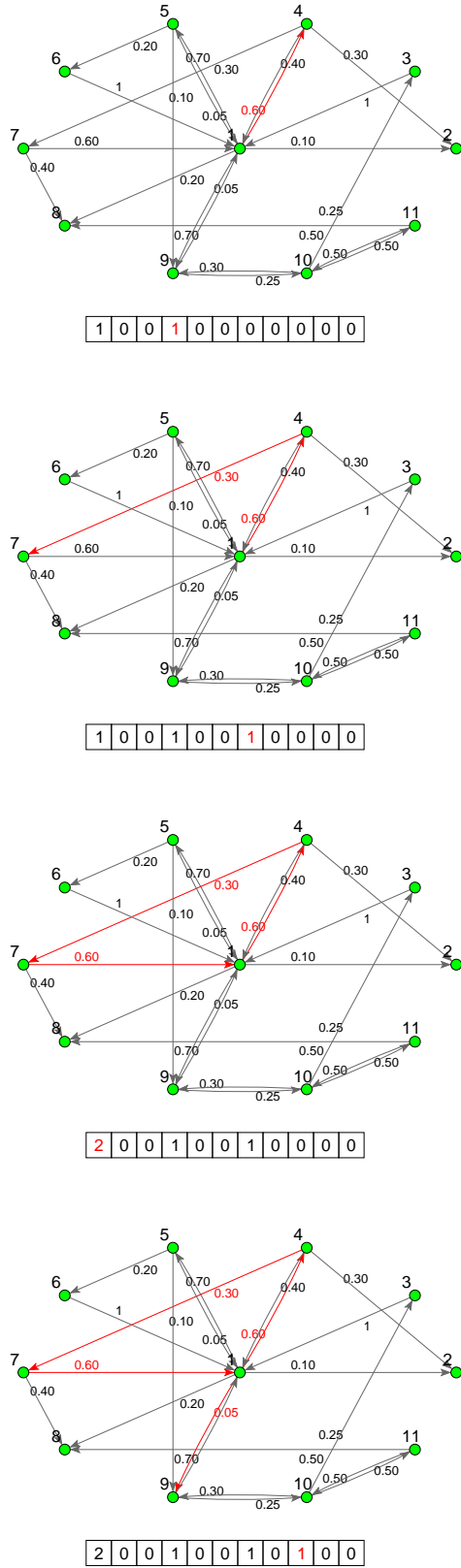


Fig. 2. In RIM, the visits of a random walker starting at node i trigger the inheritance mechanism, which modifies the features vector of a node i . In the figure, a random walk of 4 steps changes the vector of node 1 (color online).

walks of length S , after averaging over many realizations we will converge to the transition matrix P^S , every element $(P^S)_{ij}$ represents the probability of reaching j , from i , in S steps. The inheritance process corresponds, in this scenario, to a change of basis, from the orthogonal basis of the N -dimensional space, to the new basis in the space of transitions T :

$$T = \lim_{S \rightarrow \infty} \sum_{i=1}^S P^i = (I - P)^{-1} \quad (3)$$

The convergence of eq. (3) is guaranteed by the Perron-Frobenius theorem. In practice, the summation in eq. (3) converges, in terms of the matrix 1-norm, very fast, limiting the dependence on indirect associative strengths [21]. Although computations were done up to $S = 10$, $S = 4$ is enough to reach quasi-stationary states in T . Results for RIM in this work are expressed for $S = 4$ from now on. Finally, FS is the matrix that will represent in our model the feature similarity network (synthetic features network), where similarity is calculated as the cosine of the vectors in the new space, given by the scalar product of the matrix and its transpose, $FS = TT^\dagger$.

RIM fits naturally in the family of path-based similarity measures [22–28]. Jaccard index [22], cosine similarity [24] and the like have an inherent constraint, they can only account for short range similarities. This limitation is overcome in measures that take into consideration also long-range relationships [26–28]. However, a subtle distinctive feature of RIM is that similarity between nodes i and j is not a function of the number of paths from i

to j , but depends on their navigational characteristics to the whole network, i.e. two nodes are similar if random walkers departing from them behave similarly. Cosine of vectors at the end of the navigation process accounts for random walkers' global performance. We think this particular feature is adequate in a cognitive-inspired dynamical mechanism, where navigation matters.

4 Model performance

The algorithm sketched above yields a new synthetic network, FS. The capacity of RIM to extract similarity information must be tested against the empirical FP. We first check statistical macroscopical resemblance between FS and FP, by direct comparison of network descriptors and $P(s)$. We also point out results from Latent Semantic Analysis, LSA [29,30]. LSA uses truncated Singular Value Decomposition to infer semantic similarity between pairs of words. We report results for LSA trained on the TASA corpus and truncation at $d = 300$, for the subset of common words in FA and FP. We will refer to this network as LSA-N. This LSA TASA-based representation is an appropriate benchmark because it largely succeeds at predicting human synonym test judgments [31].

In fig. 3 we plot the cumulative strength distribution $P(s)$ of the empirical networks FA, FP, and the synthetic ones LSA-N and FS. The statistical agreement between FP and FS is remarkable. Note that all distributions present an exponential decay instead of a power-law decay, being the cutoff of the distribution in FA more pronounced due to its original sparseness. Random homogeneous networks

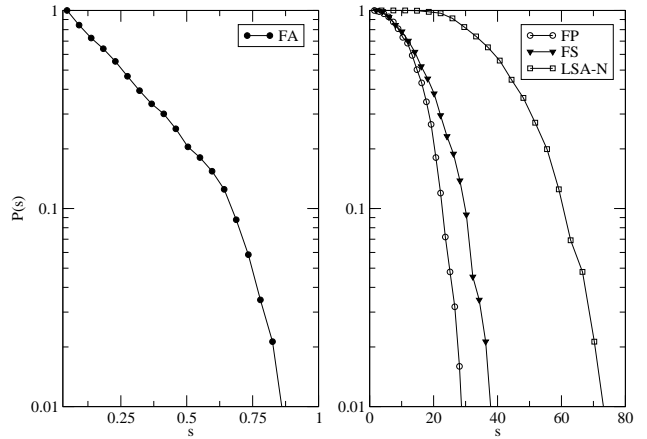


Fig. 3. Log-linear plots of the cumulative strength distribution of the networks. Left: Free Association norms FA (substrate of the dynamic process). Right: Feature Production norms FP (empirical target) , and the synthetic networks obtained using Latent Semantic Analysis (LSA-N) and Random Inheritance Model (FS).

typically show this specific form of the distributions. Main descriptors of the four networks are presented in table 2. Again, the agreement between FP and FS is remarkable, the model reproduces with significant accuracy average strength, average path length, diameter, and clustering of the FP target network. The descriptors indicate that LSA-N is even denser than FP, close to complete connectivity.

Though informative and important, agreement on average or global descriptors does not determine to state the validity of RIM to extract actual categorical information from the original substrate. The reason for this is that nodes are tagged, conformity must be sought down to the local level. In practice, we intend to test whether the specific neighborhood of a word in FP is replicated for the same word in FS (and LSA-N). We proceed as follows: given a specific word i , we start sorting its neighbors ac-

Table 2. Statistical parameters for Free Association norms FA (substrate of the dynamic process), Feature Production norms FP (empirical target), and the synthetic networks obtained using Latent Semantic Analysis LSA and Random Inheritance Model RIM.

Descriptor	FA	FP	LSA-N	FS
N	376	376	376	376
$\langle s \rangle$	0.26	13.43	39.60	15.62
L	4.41	1.68	0.02	1.77
D	9	3	2	3
C	0.1926	0.6253	0.9611	0.5848

ording to their linking weight. We apply this for each word in our data sets forming lists. The list of each word in FP is the empirical reference, and the lists we want to compare with, are those obtained for each word in the synthetic data sets, FS and LSA-N. We restrict our analysis up to the first 15 ordered neighbors, assuming that these are the most significant ones.

We now need a convenient measure to compare pairs of lists. To this end, we design a restrictive expression that assigns an error score between a list and its reference. Error depends on the number of mismatches between both lists, and also on the number of misplacements in them. A mismatch (M) corresponds to a word that exist in the reference list and not in the synthetic list and vice versa. A misplacement (O) is an error in the order of appearance of both words in each list. The error score E is then defined as:

$$E = E_M + \frac{E_O}{l - E_M} \quad (4)$$

where E_M stands for the number of mismatches, E_O the number of displacements and l the length of the list. This quantity is inspired in Levenshtein edit distance [32] and its generalization, Damerau-Levenshtein distance [33]. In them, similarity between two strings depends on the amount of insertions/deletions and transpositions that one has to perform on a string in order to completely match another one. Notice that E is strongly increased when a mismatch appears, movements are less punished. Note also that $E = 0$ when lists match perfectly, we prescribe $E = l + 1$ for two completely different lists.

Besides a proper measure, we also define a suitable micro null case. To this end, we check whether categorical information is available just by listing a word’s closest neighbors in the original FA. This implies the calculation of all-to-all shortest paths, weighting links as $d_{ij} = \frac{1}{p_{ij}}$, stronger relatedness is equivalent to shorter distance. Note that a direct neighbor of word i , i.e. a word with an edge from i , might lie at a longer distance than a second-level word. Success with this strategy would imply that RIM’s retrieval capacity is merely due to topological closeness.

Success, i.e. $100(1 - \frac{E}{l+1})$, with E as defined in eq. (4), is plotted in fig. 4 for FS and LSA-N. Error in the null model is close to 100%, it has been left out in this plot. On average the success of FS is about 10% higher than that of LSA-N, the null model evidences that categorical information demands a stronger model to be disentangled.

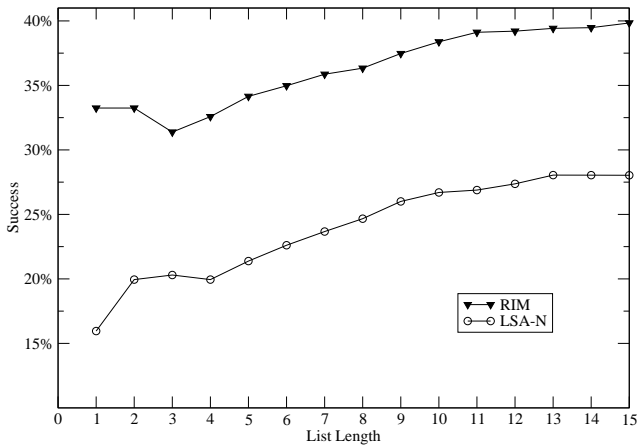


Fig. 4. For each synthetic network (LSA and FS) we have measured the mean error (for $l = 1$ to $l = 15$) against FP, according to eq. (4). We plot $100(1 - E)$ to obtain a percentage measure.

5 Summary and Conclusions

We have designed a simple information retrieval algorithm (RIM). This algorithm yields a measure of similarity between all pairs of vertices in a network. RIM is naturally related to a class of path-based similarity measures, but its aim is not the discovery of *structural similarity*. Inspired by cognitive mechanisms of memory search and retrieval, RIM highlights similar words, i.e. words that belong to the same category. From this point of view, the focus is not to spot words with structural similarities, but words with similar meaning.

Along the article we propose that RIM is related to open problems in natural language processing and cognitive science, the understanding of conceptual knowledge organization. For this reason empirical data is related to cognitive science, and output interpretation is in terms of semantic knowledge, the capacity of RIM to predict se-

ments through semantic memory navigation

semantic similarity. RIM’s results are compared to those of LSA, which has a long history of success in many machine learning linguistic-related tasks.

However we suspect that RIM has a more general interpretation. The meaning of a word (its defining features) is reduced to a dynamic process of probabilistic walks and inheritance, blind to semantic content. Then, semantic similarity is just similarity of the behavior of random walkers: two vertices are highly similar when random walkers departing from them visit, on average, the same nodes. The close connection of RIM to random walkers allows its reduction to an algebraic description in terms of Markov chains. All these facts yield an algebraic and topological interpretation of conceptual knowledge.

Indeed, topology is a key factor to understand RIM’s success. In a highly modular scenario, such as FA, random walkers tend to get trapped [34,35] reinforcing inheritance among vertices in the same community. Topological communities then enable meta-similitude relationships. While immediate neighborhood does not suffice to infer categorical relationships, see fig. 4, mesoscale relationships matter.

Acknowledgments

We thank T. L. Griffiths, M. Steyvers, G. Zamora and S. Gómez, for helpful comments. This work has been supported by the Spanish DGICYT Project FIS2009-13730-C02-02.

References

1. P. Roget, *Roget's Thesaurus of English Words and Phrases* (TY Crowell co, 1911)
2. C. Fellbaum et al., *WordNet: An electronic lexical database* (MIT press Cambridge, MA, 1998)
3. K. Klemm, V. Eguíluz, M. San Miguel, *Physical Review Letters* **95**(12), 128701 (2005)
4. A. Baddeley, *Human memory: Theory and practice* (Allyn & Bacon, Boston, MA, 1990)
5. A. Budanitsky, G. Hirst, *Computational Linguistics* **32**(1), 13 (2006)
6. D.L. Nelson, C.L. McEvoy, T.A. Schreiber, <http://www.usf.edu/FreeAssociation/> (1998)
7. K. McRae, G. Cree, M. Seidenberg, C. McNorgan, *Behavior Research Methods* **37**(4), 547 (2005)
8. R. Albert, A. Barabasi, *Reviews of modern physics* **74**(1), 47 (2002)
9. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D. Hwang, *Physics Reports* **424**(4-5), 175 (2006)
10. M.E.J. Newman, *SIAM Review* **45**, 167 (2003)
11. M.E.J. Newman, *Physical Review E* **69** (2004)
12. J. Duch, A. Arenas, *Physical Review E* **72** (2005)
13. L. Danon, J. Duch, A. Arenas, A. Díaz-Guilera, *Journal of Statistical Mechanics: Theory and Experiment* p. P09008 (2005)
14. M. Sales-Pardo, R. Guimerà, A. Moreira, L. Amaral, *PNAS USA* **104**(39), 15224 (2007)
15. E. Rosch, B. Lloyd, *Cognition and categorization* (Erlbaum Hillsdale, NJ, 1978)
16. J. Goñi, I. Martincorena, B. Corominas-Murtra, G. Arondo, S. Ardanza-Trevijano, P. Villoslada, *International Journal of Bifurcation and Chaos* (2009)
17. L. Costa, G. Travieso, *Physical Review E* **75**(1), 16102 (2007)
18. J. Noh, H. Rieger, *Physical review letters* **92**(11), 118701 (2004)
19. S. Yang, *Physical Review E* **71**(1), 16107 (2005)
20. L. Lovász, *Bolyai Soc. Math. Stud.* **2**, 353 (1996)
21. D.L. Nelson, N. Zhang, *Psychonomic Bulletin and Review* **7**, 604 (2000)
22. P. Jaccard, *Bulletin de la Société Vaudoise des Sciences Naturelles* **37**, 241 (1901)
23. E. Ravasz, A. Somera, D. Mongru, Z. Oltvai, A. Barabási, *Hierarchical organization of modularity in metabolic networks* (2002)
24. G. Salton, M. McGill, *Introduction to modern information retrieval* (McGraw-Hill, Auckland, 1983)
25. G. Salton, *Automatic text processing: the Transformation, Analysis and Retrieval of Information by Computer* (Addison-Wesley, Reading, MA, 1989)
26. E. Leicht, P. Holme, M. Newman, *Physical Review E* **73**(2) (2006)
27. V. Blondel, A. Gajardo, M. Heymans, P. Senellart, P. Van Dooren, *SIAM review* **46**(4), 647 (2004)
28. G. Jeh, J. Widom, *SimRank: a measure of structural-context similarity*, in *Proceedings of the 8th ACM SIGKDD international conference on Knowledge discovery and data mining* (2002)
29. S. Deerwester, S. Dumais, G. Furnas, T. Landauer, R. Harshman, *Journal of the American Society for Information Science* **41**(6), 391 (1990)
30. T. Landauer, S. Dumais, *Psychological Review* **104**, 211 (1997)

31. T. Landauer, P.W. Foltz, D. Laham, *Discourse Processes* **25**, 259 (1998)
32. V.I. Levenshtein, *Soviet Physics Doklad.* **10**(8), 707 (1966)
33. F. Damerau, *Communications of the ACM* (1964)
34. P. Pons, M. Latapy, *Lecture notes in computer science* **3733**, 284 (2005)
35. M. Rosvall, C. Bergstrom, *Proceedings of the National Academy of Sciences* **105**(4), 1118 (2008)

Table 3. Some illustrative examples of LSA and RIM’s predictive capacity, when compared to our FP (list size $l = 10$).

TUBA		
FP	LSA	RIM
trombone	clarinet	trombone
trumpet	violin	saxophone
drum	flute	trumpet
cello	guitar	flute
clarinet	trombone	clarinet
saxophone	fork	cello
flute	trumpet	violin
harp	cake	harp
banjo	drum	banjo
piano	piano	stereo
ERROR	4.83	2.5
ROOSTER		
FP	LSA	RIM
chicken	cat	chicken
goose	gate	turkey
pigeon	donkey	crow
sparrow	barn	robin
penguin	turnip	sparrow
pelican	owl	bluejay
bluejay	pig	pigeon
dove	fence	pelican
hawk	lion	goose
turkey	strawberry	hawk
ERROR	11	2.87