Advances in Complex Systems
Vol. 15, Nos. 3 & 4 (2012) 1250054 (22 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0219525912500543

COMPLEX STRUCTURES AND SEMANTICS IN FREE WORD ASSOCIATION

PIETRO GRAVINO

Dipartimento di Fisica, Sapienza Università di Roma,
Piazzale Aldo Moro 5, 00185 Roma, Italy

Dipartimento di Fisica, "Alma Mater Studiorum"
Università di Bologna
Viale Berti Pichat 6/2 40127, Bologna, Italy
pietro.gravino@gmail.com

VITO D. P. SERVEDIO

Dipartimento di Fisica, Sapienza Università di Roma,
Piazzale Aldo Moro 5, 00185 Roma, Italy
vito.servedio@roma1.infn.it

ALAIN BARRAT

Centre de Physique Théorique (CNRS UMR 6207),
Luminy, 13288 Marseille Cedex 9, France

Institute for Scientific Interchange (ISI), Torino, Italy
Alain.Barrat@cpt.univ-mrs.fr

VITTORIO LORETO

Dipartimento di Fisica, Sapienza Università di Roma,
Piazzale Aldo Moro 5, 00185 Roma, Italy

Institute for Scientific Interchange (ISI), Torino, Italy
vittorio.loreto@roma1.infn.it

Received 19 July 2011
Revised 28 March 2012
Published 24 May 2012

We investigate the directed and weighted complex network of free word associations in which players write a word in response to another word given as input. We analyze in details two large datasets resulting from two very different experiments: On the one hand the massive multiplayer web-based Word Association Game known as *Human Brain Cloud*, and on the other hand the *South Florida Free Association Norms* experiment. In both cases, the networks of associations exhibit quite robust properties like the small world property, a slight assortativity and a strong asymmetry between in-degree and out-degree distributions. A particularly interesting result concerns the existence of a characteristic scale for the word association process, arguably related to specific conceptual contexts for each word. After mapping, the Human Brain Cloud network onto the WordNet semantics network, we point out the basic cognitive mechanisms underlying word associations when they are represented as paths in an underlying semantic network. We derive in particular an expression describing the growth of the HBC graph and we highlight the existence of a characteristic scale for the word association process.

Keywords: Complex networks; language dynamics; word association; semantic network; WordNet.

1. Introduction

In the last years, the physicists' toolbox has become increasingly used for the study of complex systems in areas traditionally far from the pure realm of physics. Many works have concerned the interdisciplinary field of complex networks [11, 27, 5, 1], as well as the statistical physics of social dynamics [6] where the collective properties of population of human individuals are investigated. From this perspectives a lot of emphasis has been put on opinion, cultural and language dynamics, crowd behavior, hierarchy formation, human dynamics, and social spreading phenomena, just to quote the most important examples.

In social phenomena, the basic constituents are not particles but humans [4]. Though in many cases one can neglect the intrinsic complexity of human beings, as for instance for collective phenomena like traffic or crowd behavior [16, 23], the detailed behavior of each of them is already the complex outcome of many cognitive, psychological and physiological processes, still largely unknown. A very interesting example is represented by social annotations processes [18, 14] through which users annotate resources (web pages, bibliographic references, digital photographs, etc.) with free-form text keywords, known as tags. The emergent data structures are complex networks that represent an externalization of semantic structures (networks of concepts [28]) grounded in cognition and typically hard to access. In addition these networks are collectively built by the uncoordinated activity of thousands to millions of users, entangling semantics and user behavior into the so-called techno-social systems [2, 7].

In Ref. 8 it has been shown that the process of social annotation can be seen as a collective exploration of a semantic space, modeled as a graph, through a series of random walks that should represent sequences of word associations in an hypothetical conceptual space. This simple approach reproduces several aspects of social annotation, among which there is the peculiar growth of the size of the vocabulary used by the community [15] and its complex network structure [9].

In Ref. 12 the human language, through co-occurence of words in sentences, was used to build a network which showed the typical features of the complex networks, being small-world and presenting a scale-free distribution for the degree of the nodes. In Ref. 8 the semantic space was modeled as a graph. Though very reasonable, this hypothesis has never been tested in a quantitative way. Since the elementary cognitive processes underlying social annotation are word associations, it is quite natural to investigate the structures of our conceptual spaces by looking at the experiments where word associations have been measured in a quantitative way. This is precisely the point of view we take in this paper and we perform a thorough analysis of the two most important word associations databases obtained by collecting responses (target words) given by humans to specific input words (cue words). We consider in particular the *Human Brain Cloud* (HBC) database and the *University of South Florida Free Association Norms* database [24].

HBC represents the largest available word association database. It was obtained through the implementation of a massively multiplayer online game. As HBC was not conceived with a specific scientific purpose in mind and may thus suffer from uncontrolled biases, we also consider the smaller, though scientifically more controlled, database known as the *University of South Florida Free Association Norms* [24]. We analyze the graphs built from both databases though we focus more specifically on HBC. We derive in particular an expression describing the growth of the HBC graph and we highlight the existence of a characteristic scale for the word association process. Next we present an analysis aimed at grounding the observations made in the word association graphs by a direct comparison with WordNet (WN) [20, 22], the largest lexical database of words and semantic relations. The aim here is to associate a semantic value to each node of the graph as well as labeling the word associations in terms of semantically charged cognitive paths on the WN graph. The ensemble of these results is very valuable since they help to shed light on the cognitive processes underlying free word associations.

2. Free Word Associations: Data and Networks

Word associations have been experimentally studied in details, especially by linguists and cognitive scientists [10, 24]. An interesting point of view, not yet fully explored, is to look at the ensemble of words and associations as a complex network, where nodes are words and links are associations, and to analyze it as such [29, 19]. But previous works performed the typical analysis of the complex networks' framework on words association networks obtained from scientific experiment performed by researchers in extremely controlled environments. This makes data reliable but at a quite high cost. Typical word association experiments involve a relatively limited number of subjects (100–1000), in controlled conditions. A word (called cue word) is presented to the recruited subjects, who are asked to write the first related word that come to their minds (called target word). Due to high costs in terms of time and money, the number of cue-target associations gathered in these classical experiments has been relatively limited, at maximum of the order of 100,000.

The experimental data we analyze were obtained from the "massively multiplayer word association game" HBC.[a] HBC was designed as a web-based game in English language that simply proposed a cue word to the player, asking for a target word (e.g., volcano-lava, house-roof, dog-cat, etc.). The cue was each time

[a] Dataset courtesy of Kyle Gabler [13].

taken independently and randomly with uniform probability from an internal self-consistent set of words, constructed by gathering the answered target words at the end of each game session. With respect to usual experiments, no control is performed on the number of players, the number of cue-target associations given by each player, etc. On the other hand, the obtained dataset is considerably larger than that of previous experiments, and consists of approximately 600,000 words and 7,000,000 associations, gathered within a period of one year (while pre-existing experiments involved teams of specialists for much longer periods of time). As each player could enter whichever word or set of words, the data contain a certain volume of inconsistent words, which have to be discarded. After a suitable filtering procedure, which we describe in Appendix A, we obtain a strongly connected directed weighted graph with almost 90,000 words and 6,000,000 associations, i.e., a dataset still considerably larger than those of previous experiments.

In order to detect possible systematic biases in HBC, we also compare some properties of the HBC dataset with the dataset of the most important word association experiment, the Nelson *et al.* "University of South Florida Free Association Norms" database (SF) [24]. SF is the outcome of a great effort started back in 1973 and lasted almost thirty years. It consists of 5000 words and $700,000$ associations. Each dataset yields a graph whose nodes are the words and whose edges correspond to the associations between words made by the players/subjects. Moreover, each edge is weighted by the number of times that the corresponding association has been made. A snapshot of a part of the HBC graph is shown in Fig. 1. We compare in Fig. 2 the in and out-degree distributions of the two networks. It turns out that almost all (99%) the words used as cues in SF were present in HBC and 72% of the SF associations were present in the HBC dataset. To deepen our analysis, we consider for each word w_i the set of associations in which w_i was proposed as a cue word to players, and define the out-strength $s_{\text{out}}(w_i)$ of the corresponding node as the total number of associations in which w_i was the cue, and the out-degree $k_{\text{out}}(w_i)$ as the number of *distinct* target words answered in response to w_i. Analogously, we define the in-strength $s_{\text{in}}(w_i)$ and the in-degree $k_{\text{in}}(w_i)$ by referring to the associations in which w_i was answered as target: $s_{\text{in}}(w_i)$ is the total number of times that w appears as a target, and $k_{\text{in}}(w_i)$ is the number of distinct other words which yielded w as an answer.

Figure 2 shows the distribution of in- and out-degrees in both HBC and SF, as a function of $k/\langle k\rangle$ (where $\langle k\rangle$ is the average of the distribution). The distribution of out-degrees is narrow, which means that the number of distinct answers given by distinct persons to a given cue is rather limited, as could be expected. On the other hand, the distribution of in-degrees is broad: The number of distinct cues from which a given target can be obtained ranges from 1 to 10 $\langle k\rangle$ with $\langle k\rangle_{\text{in}}^{HBC} = \langle k\rangle_{\text{out}}^{HBC} = 34$, $\langle k\rangle_{\text{in}}^{SF} = 36.5$ and $\langle k\rangle_{\text{out}}^{SF} = 6.15$.

The asymmetry between the average k_{in} and k_{out} in SF is due to the asymmetry in the choice of the cue words by the experimentalists. In SF only selected words were used as cues, while many of the answered target words were never used as

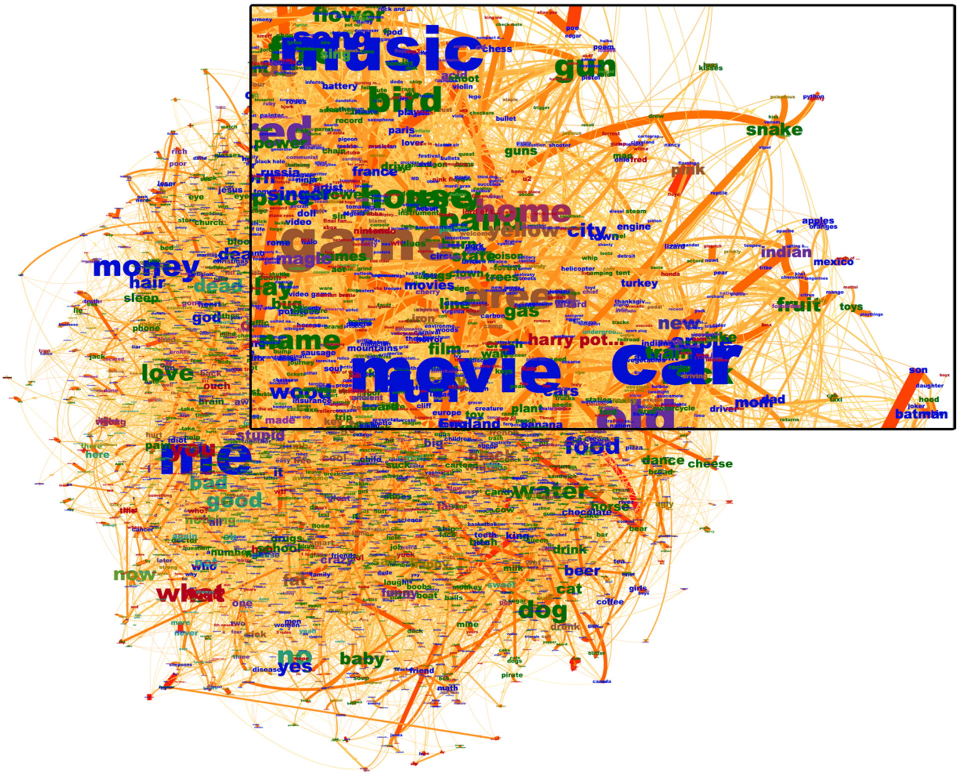

Fig. 1. Graphical representation of the word association network obtained from the HBC dataset. Labeled nodes represent the first 5000 words entered in the system while arcs represent word associations. Colors associated to labels distinguish different parts of speech (adjectives, nouns, verbs, etc.) or any of their combinations (noun/verbs as "saw", noun/adjectives as "red"). The width of arcs codes the weight of the corresponding association, i.e., how many people ever made that association.

cues. For these reasons, in order to compare SF and HBC we took the asymmetry into account by calculating the averages only on those words of SF which were used as cues.

The degree distributions of both HBC and SF exhibit very similar shapes, except for a difference caused by the filtering procedure (and described in Appendix A) for small in-degrees in HBC. It is worth noticing that even though the systems have different sizes, the out-degrees distributions show a peak at close values of $k/\langle k\rangle$, a clear indication of an intrinsic similarity of this specific kind of networks, despite their very different origin.

Figure 3 reports the strength distributions both for HBC and SF, as a function of $s/\langle s\rangle$, with $\langle s\rangle_{\text{in}}^{HBC} = \langle s\rangle_{\text{out}}^{HBC} = 68.7$, $\langle s\rangle_{\text{in}}^{SF} = 144$ and $\langle s\rangle_{\text{out}}^{SF} = 24.3$. As for the degree, the asymmetry between the strength averages of SF is caused by the asymmetry of the cue choice. The out-strength distributions are completely determined by the way in which the cue words were chosen. In fact, in HBC cue words were

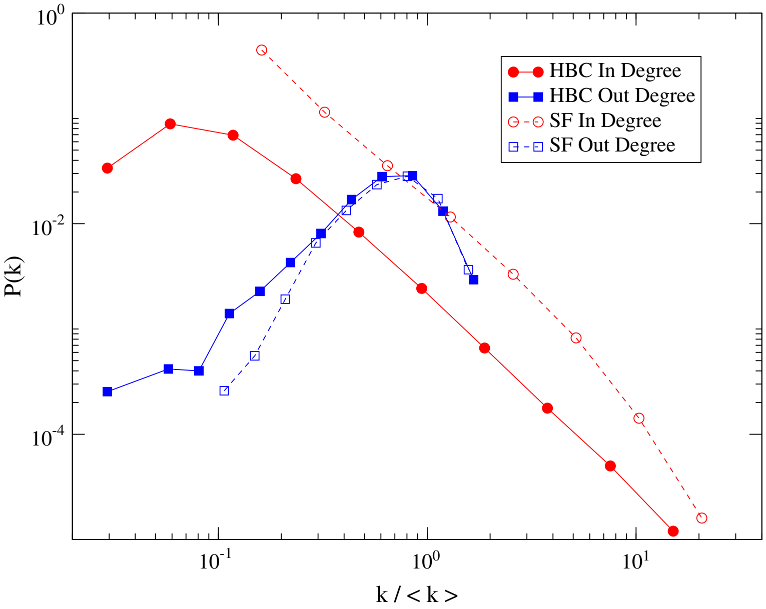

Fig. 2. The log-binned degree distributions for the word association networks of HBC and SF, in log-log scale. Since we are matching datasets of different sizes we divided the degree by the average degree value. In red, the in-degree distributions for HBC (filled circles) and for SF (empty circles). In blue, the out-degree for HBC (filled squares) and for SF (empty squares). Both the in-degree distributions show a power-law form (a straight line in log-log) for sufficiently large $k/\langle k\rangle$, i.e., for values larger than 0.1. Both out-degree distributions have a characteristic scale, showing a peak slightly before 1. The difference in shape at lower degrees is a consequence of the filtering procedure described in Appendix A.

chosen randomly from a growing set of words, which passed a quality requirement (see Appendix A). So, after filtering, we see a peak for older words which, being in the set for longer time, have been statistically used more than the others, and a growing shape for lower value of s_{out} corresponding to more recent words. In SF, instead, the cue words set has been chosen from the beginning and words were used as cue with similar frequencies. Both in-strength distributions show a power law form (a straight line in log-log). This power law is consequence of the Zipf's Law [31] observed in the frequency of words of the english language. The different shape at lower degrees is a consequence of the filtering procedure described in Appendix A.

Note that the difference of distributions between in- and out-degrees is not a surprise: This kind of asymmetry has been studied also for other different kind of complex network, such as the Web [3], the mail network [26] and the citation networks [30]. In our specific case, first of all we have to consider the relation between k and s. We will later see that the degree can be thought as a monotonic function of the strength. So the asymmetry we see in Fig. 2 is a consequence of what we see in Fig. 3.

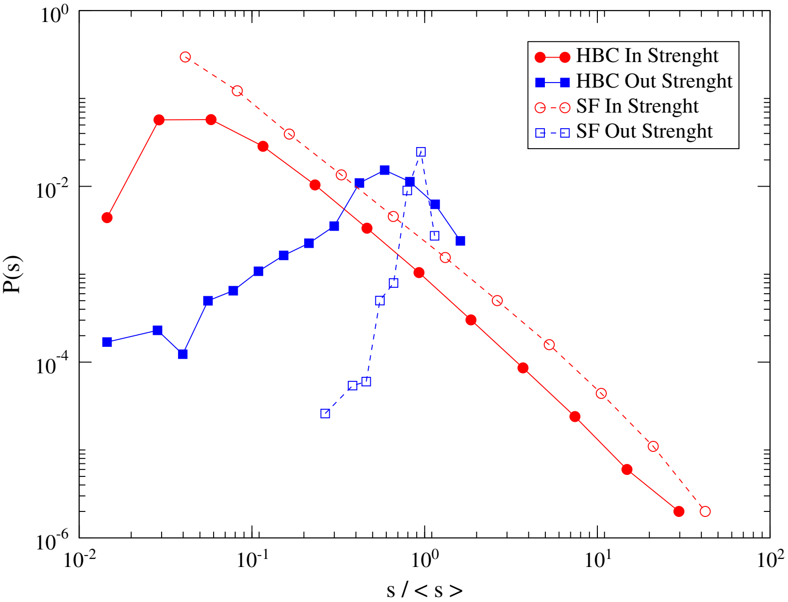

Fig. 3. The log-binned strength distributions for the word association networks of HBC and SF, in log-log scale. Since we are matching datasets of different sizes we divided the strength with the average strength value. In red, the in-strength distributions for HBC (filled circles) and for SF (empty circles). In blue, the out strength for HBC (filled squares) and for SF (empty squares). Both the in-strength distributions show a power-law form (a straight line in log-log) for sufficiently large $s/\langle s\rangle$, i.e., for values larger than 0.1. The different shape at lower degrees is a consequence of the filtering procedure described in Appendix A. The out-strength distributions have a peaked but different shape, because of the different ways in which cue words were chosen.

In order to quantify more precisely the similarity between SF and HBC, we computed the cosine similarity between the neighborhoods of each word. Given a word w_i which belongs to both SF and HBC, we are interested in how much the associations in our two datasets having w_i as a cue are similar. We define l_{ij}^{SF} (resp. l_{ij}^{HBC}) as the number of times that the association between w_i and w_j has been made in the SF (resp. HBC) dataset. The cosine similarity between the associations made with w_i in the two datasets is then defined as:

$$CS_i = \frac{\sum_j l_{ij}^{HBC} \cdot l_{ij}^{SF}}{\sqrt{\sum_j (l_{ij}^{HBC})^2 \cdot \sum_j (l_{ij}^{SF})^2}}, \tag{1}$$

where the sums run over all common words between the two datasets. The cosine similarity ranges between zero, no common word is associated to w_i in SF and HBC, and 1, if all the l_{ij} are equal. The average cosine similarity turns out to be 0.215. This number seems to be relatively low but this was somehow to be expected since HBC is much larger than SF and it contains many associations not contained in SF. These associations will contribute to the denominator of Eq. (1) but not to the numerator so they will lower the similarity. In order to avoid this kind of bias, we repeated the measure only considering those associations belonging to both datasets and we obtained in this way an average cosine similarity of 0.851. In order to evaluate the significance of these figures of the similarity, we performed 200 independent reshufflings of the SF system's nodes obtaining the following values: $\langle CS \rangle = 0.111 \pm 0.001$ by considering all the associations and $\langle CS \rangle = 0.483 \pm 0.003$ by considering only those associations belonging to both databases. This result demonstrates a strong correlation between the data obtained in the SF controlled experiment and in the web-based HBC game. This is an important information pointing to the reliability of the HBC dataset.

3. HBC in Depth

Once acquired enough confidence on the reliability of the HBC dataset, let us now deepen our analysis only focusing on HBC, as the largest available Word Association dataset. The directed graph of HBC is composed by $88,747$ nodes and $3,013,125$ edges in total, corresponding to $6,097,806$ registered associations. As we have seen in Fig. 2, while the in-degree distribution follows a power law, the out-degree distribution is peaked. We measure a power-law in-degree distribution of HBC with an exponent $\beta \simeq 1.76$ smaller than 2, which implies that the average value of the in-degree $\langle k_{\text{in}} \rangle$ is not well defined, diverging with the size of the system: The number of cues that can yield a given target word as outcome has no characteristic value. On the contrary, the out-degree shows a Gamma-like distribution peaked around a characteristic value $k_{\text{out}} \simeq 28$, corresponding therefore to a well-defined characteristic number of distinct words obtained as targets for each word inserted in the game as a cue. The distribution drops very rapidly at large k_{out} with a sharp cut-off around 100, indicating that the number of distinct words that humans spontaneously associate to a given cue is quite restricted. For a better understanding of the origin of this k_{out} scale, we can study the average growth of k_{out} as a function of s_{out}, i.e., the average number of different target words obtained from a given cue word as a function of the number of times the given cue was extracted for a game (see Fig. 4).

The inset of Fig. 4 shows that $\langle k_{\text{out}} \rangle$ can be approximately described by a sublinear power-law with exponent of ~0.87, but a deviation from this law is observed at large s_{out}. The sublinear power-law behavior is well known to appear in the dictionary growth of texts and is known as Heaps' law in this framework [15]. The Heaps' law is symptomatic of an underlying generalized Zipf's law [17], which in our case should appear in the marginal distribution of the target word frequencies, given a fixed cue word. More formally, we can say that the target word associated to a given cue is our random variable and we expect its frequency to be distributed as a Zipf's Law. In other words, let us consider a given word X and all the words that have been associated when X has been used as cue. Some of those target words have been used just once, while some others have been answered more frequently. If we order all those target words from the most frequent to the rarest and plot their

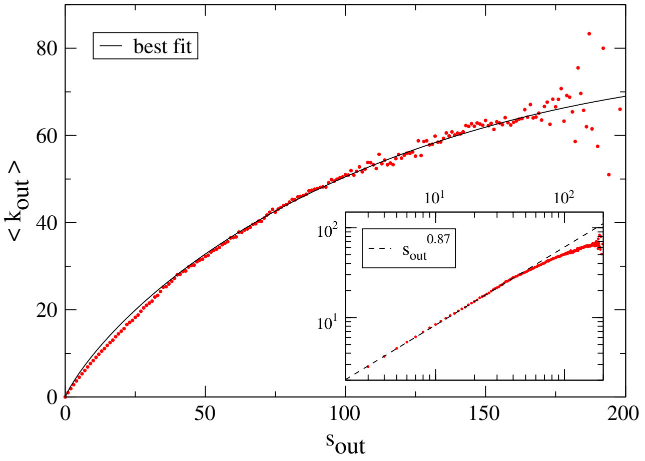

Fig. 4. Average $\langle k_{\text{out}} \rangle$ as a function of s_{out} (red), obtained by averaging the k_{out} values of all those cue words with the same s_{out} value. The fitted curve obtained from Eq. (2) is shown in black (parameters: $V \simeq 74$ and $B \simeq 0.92$). Inset: same data in log-log scale together with a fit to a pure sublinear power law, to highlight the deviation from such a law at large s_{out}.

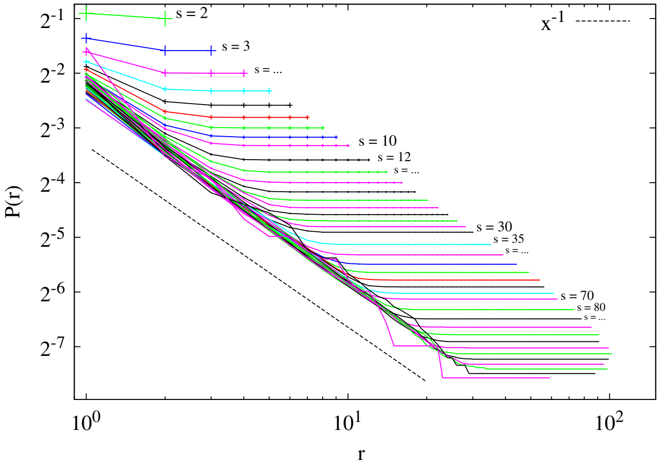

Fig. 5. For each cue word we calculated the rank distribution (RD) of its target words. Each curve in the picture corresponds to an averaged RD plot obtained by averaging the RD curves corresponding to cue words with the same out-strength value. When s_{out} is sufficiently large (i.e., we have sufficiently large statistics) we obtain a power law with an exponent close to (minus) unity, followed by a flat tail.

frequencies, because of the Heaps' Law observed in the inset of Fig. 4, we expect to observe a decreasing Zipf's Law-like power law. Figure 5 shows the average ranks distribution for different values of s_{out}, confirming the presence of a Zipf's law, which becomes better defined as s_{out} grows.

To estimate the functional shape of $k_{\mathrm{out}}(s_{\mathrm{out}})$, we investigate in Appendix B the impact of finite size effects combined with the above mentioned Zipf's law. We define B as the exponent of the rank distribution of the target to a given cue and assume that there exists a maximum number, V, of different target words that can be associated to a cue word. In this way we are implicitly assuming that k_{out} will converge asymptotically to V (possibly with a residual logarithmic growth that we neglect here). With these hypotheses in mind we derive an expression (see Appendix B) for the behavior of k_{out} as a function of s_{out} that reads:

$$k_{\mathrm{out}}(s_{\mathrm{out}}) = V\frac{B{s'}^{1/B}_{\mathrm{out}} - s'_{\mathrm{out}}}{B-1}, \qquad (2)$$

where

$$s'_{\mathrm{out}} = \frac{s_{\mathrm{out}}}{s^{\mathrm{lim}}_{\mathrm{out}}} \qquad s^{\mathrm{lim}}_{\mathrm{out}} = \frac{(V^B - V)}{(B-1)}.$$

We also defined the auxiliary parameter $s^{\mathrm{lim}}_{\mathrm{out}}$: it is the value of s_{out} for which $k_{\mathrm{out}} = V$.

We can use Eq. (2) to fit the experimental data and to check whether our finite-size analysis holds in this case. Then, we can also deduce an estimate for B and V. The best fit is shown in Fig. 4, leading to the values $B \sim 0.92$ and $V \simeq 74$. It is worth to note that, while Eq. (2) is significantly different from a typical Heaps' law, one recovers the pure Heaps' behavior for $V \gg s_{\mathrm{out}}$. In this limit, in fact one has that for small s_{out} and for $B > 1$ the behavior of k_{out} can be approximated by a power law with the correct exponent $1/B$. If instead $B \leq 1$ the dominant term of the Eq. (2) will be the linear one. This finding is consistent with the fact that the growth may not be more than linear since it must always hold that $k_{\mathrm{out}} \leq s_{\mathrm{out}}$.

It is important to remark that we cannot exclude that the actual formula fitting the data of Fig. 4 contains logarithmic corrections implying that k_{out} would keep growing asymptotically with s_{out} though with a logarithmic or sub-logarithmic law. Here, we are neglecting this possibility.

The value of V obtained is somewhat surprising since it implies that the asymptotic number of different targets obtained for a given cue is indeed several orders of magnitude smaller than the total number of words in the system ($\sim$90,000). Hence, we find that on average, there is a limited number of target words for a given cue, reflecting the existence of a limited semantic context associated by humans to each word. This is in contrast with the fact that the number of words (cues) that yield a given target does not show any particular scale: The semantic context is thus not a "symmetric" concept in terms of free word associations.

Other measures can help in characterizing in a more detailed way the topology of the graph. A measure of the distances between the nodes of the graph, i.e., the shortest path between nodes, indicates that the filtered strongly connected component of the HBC network satisfies the "small world" property with an average node distance of 4.0 ± 0.7. This means that with path of around five edges we can reach almost every node of the graph, so that we can easily and quickly explore it. This is not a surprise as most complex networks have been shown to share this property [27, 5, 1].

We also analyzed the mixing patterns [25] of the HBC graph, in order to measure the tendency of nodes with similar degree to preferably connect to each other. The assortativity coefficient r is defined as the Pearson's correlation coefficient calculated between the out-degrees of the nodes linked by an edge. If, for each edge, k_c is the out-degree of the cue and k_t is the out-degree of the target, we have:

$$r = \frac{\langle k_c \cdot k_t \rangle - \langle k_c \rangle \cdot \langle k_t \rangle}{\sqrt{(\langle k_c^2 \rangle - \langle k_c \rangle^2) \cdot (\langle k_t^2 \rangle - \langle k_t \rangle^2)}}, \tag{3}$$

where the averages are calculated on the ensemble of edges. For the unfiltered data we measured $r = 0.0517$ while for the filtered graph we found $r = 0.0644$. The value obtained for the filtered data is slightly larger but still of the same order of magnitude of the result obtained for the unfiltered graph. Hence, it is reasonable to neglect the correlation effects that the filtering procedure may have introduced. We tested the significance of the r values found by performing the same measure on an ensemble of 100 occurrences obtained by reshuffling the edges in the network. We came up with $(1.0 \pm 5.8) \cdot 10^{-4}$, i.e., three order of magnitudes lower than the corresponding value of the actual network, confirming that the values of r obtained for the real system are small but still significant. Furthermore, if we consider the number of times a given cue has been associated with a given target as the weight w of that edge, we can define the weighted average of a given function f of the generic edge e as:

$$\langle f \rangle_w = \frac{\sum_e w(e) \cdot f(e)}{\sum_e w(e)}.$$

Using this weighted average in Eq. (3) we can measure, for the filtered values, the weighted assortativity coefficient $r_w = 0.108$, which, again, points to a slight assortativity. We performed a set of 50 reshufflings to evaluate the significance of the result, obtaining $\langle r_w \rangle = (1.3 \pm 11.9) \cdot 10^{-4}$ which, again, points out a small but still meaningful assortativity.

These results give us an overview of the properties of the HBC network. Using the HBC word association network as a proxy of the way in which our mind stores and organizes all words and related meanings, we observe that it has indeed the properties we should expect from such a network: Every word defines a limited context; we can explore the network in a fast and efficient way, for example to recover meanings; and the results about the graph seems to suggest that the network is robust so that is still connected even in case we forget some word or if we do not know it. While this first analysis has concerned the network in itself, considering nodes and edges as abstract entities, in the next section we shall deal with their semantic content.

4. Introducing Semantics

The nodes of the HBC network are words, and as such, they have a semantic value. Let us therefore aim at measuring quantitatively what kind of semantic relations are used while associating two words. In order to analyze these semantic connections, a database of semantic relations between words is needed. We choose to use WN [20, 22], a large lexical database developed at Princeton University. In WN, words are related to each other according to their mutual semantic relations, a list of which is given in Appendix C. WN allows us to built another directed graph in which nodes (143963) are again words but edges (1345801) represent now the semantic relations between them. In order to find what kind of relation corresponds to a given free association between words (i.e., a link in the HBC graph), we can examine the shortest paths in the WN graph between every pair of words linked by an association in the HBC graph, as shown in Fig. 6.

In this mapping procedure, we have to take into account that words of the WN database are actually lemmas, i.e., are reported in their dictionary form, while HBC words are subject to any kind of morphological derivation (third person, plurals, etc). For this reason, in performing the mapping we assign to pairs words of HBC the path existing between their relative lemmas. For example since the WN connection between "buy" and "purchase" is of the synonymy type, we consider also the couple "bought" and "purchased" as synonyms.

Moreover, it may happen that two words associated in HBC correspond to multiple shortest paths with the same length in WN. In such a case, we consider all paths as equiprobable by assigning to each possibility a normalized weight. For example "oak" and "pine" are linked by an association in HBC. In the WN graph, to go from "oak" to "pine" we can go either through "tree" or "forest". The two

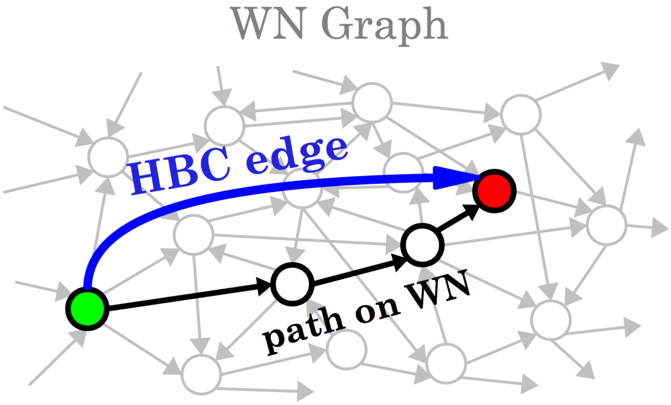

Fig. 6. The mapping of HBC onto WN: Given a couple of words linked by an association in the HBC graph (the blue arrow) we consider for the shortest path (black arrows and nodes) between the same words in the WN graph (gray graph in the background). In this example a HBC association maps into a three-edges WN path.

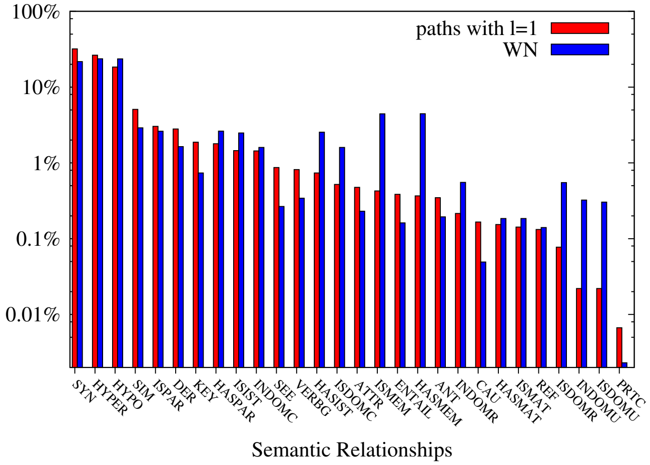

Fig. 7. Red bars represent (in logarithmic scale) the normalized distribution of the semantic relations characterizing those HBC associations that map into a one edge path on the WN network. In other words, red bars are the normalized semantic relations distribution of those edges common to both graphs. Blue bars represent the normalized semantic relations distribution in the whole WN graph. Data are in decreasing order for the distribution of those edges common to both graphs (red bars).

alternative semantic paths would be: oak $\models$ hyperonymy[b] $\Rightarrow$ tree $\models$ hyponymy[c] $\Rightarrow$ pine or oak $\models$ holonymy[d] $\Rightarrow$ forest $\models$ meronymy[e] $\Rightarrow$ pine; we consider both by assigning to each of them a weight 0.5.

We first analyze the 74903 HBC associations that map into paths of length 1 on WN. They correspond to the directed edges which are present in both WN and HBC graphs. Figure 7 shows the normalized distribution of their semantic relations.

The distribution shows that synonymy[f] (32% of the common links between HBC and WN), hyperonymy (26%) and hyponymy (18%) are the most commonly used semantic relations for an association. In Fig. 7, we also show the distribution of the semantic relations in the whole WN graph for comparison. We notice that, for some relations there is a substantial difference between their occurrence in HBC and WN. For example, the causal (CAU) nexus normalized occurrence is much larger in the mapped HBC (1.6%) than in the WN graph (0.05%). Even if the HBC value is still low in absolute terms, it is worth to note that is more than thirty times greater than the WN value. This means that if a given cue word has a causal link in WN then players will be very oriented to follow it while making

[b] Hyperonymy is the link between a word with a particular meaning and another word with a more general meaning which includes the first one; e.g., “red” is the hyperonym of “scarlet”.
[c] Hyponymy is the inverse relation of hyperonymy.
[d] Holonymy is the link between a word denoting a part of a whole and the word denoting the whole itself; e.g., “hand” is holonym of “finger”.
[e] Meronymy is the inverse relation of holonymy.
[f] Synonymy is the link between two words with the same meaning; e.g., “student” and “pupil”.

the association, more than one could expect looking at the number of causal links in WN, even if it will not be the preferred relation in absolute terms. It is then clear that we need to quantify the occurrences of semantic relations in HBC with respect to our reference system of WN by analyzing the effective possibility of players of doing an association in HBC following a certain semantic relation of WN. If we look at all associations starting from a cue word w we could measure the ratio of associations following the synonymy relation or the hyperonimy relation. In general, we could measure all the ratios for any kind of semantic relation. If we have enough statistics, we may assume these ratios are the probabilities of doing these associations starting from w. Let us consider the example case in which w does have $k_{\text{out}} = 5$ and $s_{\text{out}} = 10$. Consider also that four links are synonymy links and the other one is a causal link. Finally, consider also that, of the 10 associations, four follow a causal link and six one of the synonymy links. Even though there are overall more synonymy associations it is clear that the causal link is more used than we could expect by just looking at the links of w. This means that human beings construct associations with probabilities that could strongly deviate from what would be expected from the pure statistical structure of WN. This is a very interesting feature to further quantify. Sticking on the example of the causal link, it is evident that in order to quantify its actual relevance, we should not estimate the probability of making a causal link association, but rather the probability of making a causal link association conditioned to the existence of a certain number of available causal links for w. In this case, it would be one causal link out of the five total links, i.e., $P(cau|(1\,cau;\ 5\,syn))$.

In general, the fraction of associations of a given type represents the probability of choosing a certain semantic link given the underlying WN structure of the possible available links. In order to quantify the relative importance of a given semantic association we normalize its frequency of occurrence in WN. To this end, we define ρ_i as the percentage of relations of kind i in one edge paths and p_i as the percentage of relations of kind i in WN and we define a normalized effective probability as:

$$\pi_i = \frac{\frac{\rho_i}{p_i}}{\sum \frac{\rho_j}{p_j}}, \tag{4}$$

π_i represents thus the probability of performing the association i as if at each edge all the semantic associations were equally possible. The results of this computation is presented in Fig. 8, where we also report the information about the size of set of each type of semantic association, to give a qualitative and quantitative idea of the reliability of these results.

The analysis reported in Fig. 8 demonstrates how the word association process in HBC features important deviations with respect to WN. If the semantic relation occurrences in HBC were exactly the same as in WN, all symbols would lie on the dashed line. The ones which lie below occur less frequently and viceversa. Deviations are in both directions. For example, hyperonymy is slightly more frequent than expected while hyponymy is less frequent. This means that the target word tends

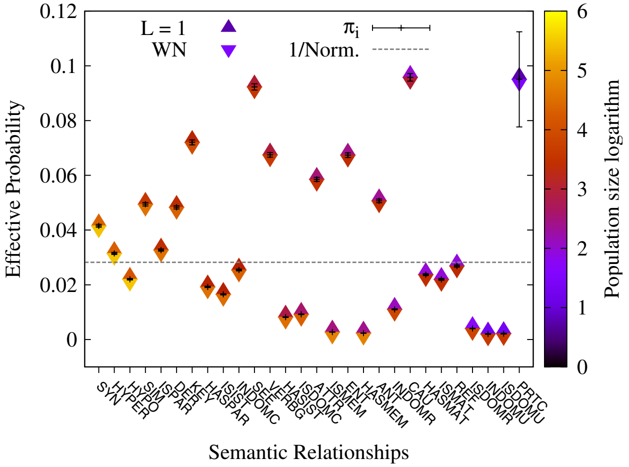

Fig. 8. Normalized "effective" probabilities of the different types of WN semantic relations with error bars. The dashed line represents the probabilities of the associations in WN. So, if the semantic relation occurrences in the mapped HBC were exactly the same as in WN, all symbols would lie on the dashed line. To give an idea of the significance of the values, for each value we draw two triangles whose colors indicate the logarithm of the number of relations of the given type and therefore give an idea of the measure accuracy (blue and violet values correspond to smaller sample sizes than yellow and red ones). Upward triangles refer to the mapped HBC with unity length paths, while downward triangles refer to the whole WN.

to be a more general term instead of a more specific one. The already mentioned causal link is largely overrepresented in HBC. This means that, if the cue word have a causal link, while making an associations we will tend to prefer it. On the other hand there are associations poorly represented in HBC like "is member" (ISMEM) or "has member" (HASMEM). We also see how many not purely semantic relations (as "see also" or "keyword") are chosen frequently.

We also considered the HBC associations that map into a path of two or three edges in WN, as reported in Fig. 9. The first positions in the distribution correspond again to combinations of hyperonymy, hyponymy and synonymy, which were already the most frequent relations in the previously discussed case of one edge paths.

By normalizing as we did before for the one edge path case, and retaining the most significant results (i.e., those corresponding to the largest statistics, such as the yellow and the red ones in Fig. 8) we obtain the situation shown in Fig. 9.

These probabilities seem to show the emergence of a particular kind of exploration paths of the WN graph. Many of the WN relations are hierarchical. For example, hyperonymy binds a specific term to a more general one (e.g., tree is the hyperonym of oak). Other examples are holonymy (the relation between a whole and a part: Tree is the holonym of bark) or geographical collocation. In Figs. 9 and 10 we can see how in the sequences exploring these hierarchical structures there is a recurrent pattern, made of one edge toward a more general term followed by one edge toward a more specific term. We call this pattern the "brother" pattern,

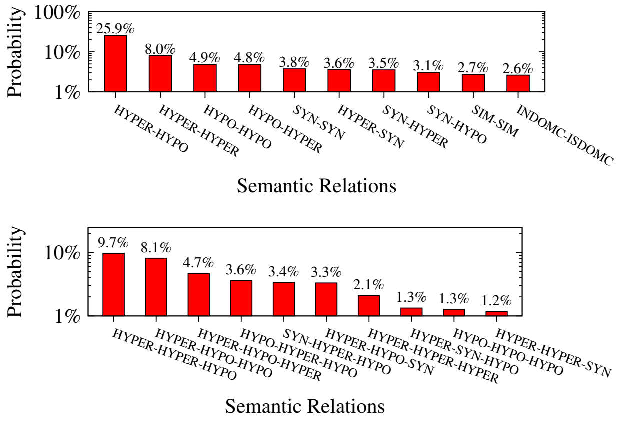

Fig. 9. The top ten occurrence ranking of the semantic relations sequences for those HBC associations that map onto a path of two (top chart) and three (bottom chart) edges in WN.

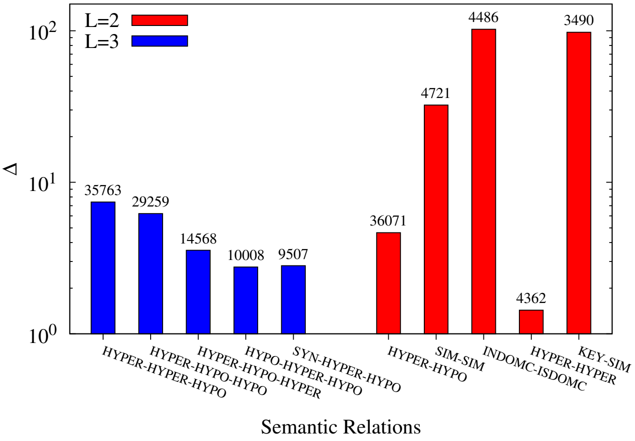

Fig. 10. The occurence of the five most significant normalized semantic relations sequences for those HBC associations that map onto a path of two and three edges in WN. Above the bars, the number of paths of the relative sequence. Δ is defined as the ratio $\frac{\rho_i}{p_i}$ with p_i and ρ_i defined in the text.

because, starting from a given node it takes us to another child (specific term) of the same parent (general term). In order to study the importance of this pattern we measure its occurrence together with two other patterns: grandparent (two edges towards more general terms) and grandchild (two edges towards more specific terms). For paths of two and three edges, we find:

Pattern	Two edges (%)	Three edges (%)
Brother	31	51
Grandparent	8	18
Grandchild	5	14

For paths composed of three edges, we measured how many of them included a brother pattern, e.g., the sequence hyperonymy-hyponymy-hyperonymy includes the brother pattern (the first two steps) while hyponymy-hyperonymy-hyperonymy does not (in the last two steps we find the grandparent pattern). These results confirm the existence of preferential patterns of exploration of hierarchical structures in the process of word association, in a way that most often maintains the same level of specificity between cue and target. Note that we limit the analysis to paths of up to three edges because the distribution for longer paths is practically equal to the distribution occurring in an uncorrelated artificial system built by associating randomly extracted words.

5. Conclusion

In this paper, we have presented the results of the analysis performed on the Word Association Graph constructed in two different experiments: HBC and the *South Florida Free Association Norms* (SF). Word association graphs are quite interesting because they represent a proxy of the way in which our mind stores and organizes all words and related meanings. The HBC dataset is substantially larger than any previously available datasets and it has been constructed through a web-based game while the SF is the outcome of a controlled linguistic experiment. The comparison between the two databases has been an important preliminary step in our analysis to ensure that no major biases were present in the HBC database. After a filtering procedure we ended up with a clean HBC database whose statistical properties exhibit strong correlations with that of SF, giving us confidence on the generality and reliability of the approach. For the first time, the huge database of the HBC experiment has been brought to the attention of the scientific community as a valuable tool to shed light on the possible mechanisms of word and meaning retrieval processes underlying human language skills.

Our results have shown that the associations to a given cue word tend to result in a limited set of words, whose size is considerably smaller than that of the system, while the number of words that yield a given target can fluctuate enormously. Hence, the input of a word seems to define a sort of “semantic context” as an output, which can be subject of further analysis. It is worth to point out how the existence of these contexts should influence the realization of future word association experiment. Equation (2) may be used to understand how much statistics is needed in order to fully explore these semantic contexts, or at least their most significant part. On the other hand, the size of the context which leads to a given word target has unbounded fluctuations.

We found that the HBC network can be quickly explored due to its small world property. In the framework of the hypothesis of the existence of a cognitive networked structure revealed by the free word association games or experiments, the navigational efficiency of such a network is thus compatible to what we should expect from our mental “meaning management system”.

We further extended our analysis by grounding the observations made on the HBC graph through a direct comparison with the largest lexical database of words and semantic relations, WN [20, 22]. Through WN, we classified the semantic character of word associations collected in HBC. This classification leads to a preliminary understanding of the cognitive processes underlying word association. The most used semantic relations result to be synonymy, hyperonymy and hyponymy. By comparison with the overall number of semantic relations present in WN, we have shown that other types of less common relations are in fact important, such as for instance the causal nexus. Moreover, when the association corresponds to a sequence of semantic steps, preferential combinations of semantic relations emerge, with an overall tendency to keep the same level of specificity between the cue word and the target.

Acknowledgments

The authors wish to warmly thank Kyle Gabler for providing us the HBC database and information about data gathering, Filippo Menczer and Francesca Tria for very interesting discussions. This research has been partly supported by the EveryAware project funded by the Future and Emerging Technologies program of the European Commission under Grant Agreement Number 265432.

Appendix A. Filtering HBC Data

HBC was conceived as a game, and not designed for scientific purposes. Nevertheless, the system included a “quality control” for the word dataset based on two factors: word popularity, i.e., a word used often was assumed to be a “valid word”, and reports of users about misspelled or offensive words. These two factors contributed to a sort of quality score. A word was used as cue by the system only if its quality score was above a given threshold. In the first part of our study we used the same strategy to filter words. In a set of ∼600,000 words only ∼90,000 satisfied this quality requirement. Then we did further filtering considering only the strongly connected component (it is worth to note that it has been obtained by removing just few hundreds nodes, less than the 1% of the quality-filtered words). Anyway, for the measure of the assortativity coefficient r, as we said previously, the filtering procedure implies just a low alteration of the value which allows us to neglect eventual correlations introduced by the filtering. In the second part, the matching procedure with WN itself provided a filtering mechanism, as invalid words do not have a match in WN. Even in this case the filtered words were ∼90,000.

Appendix B. A Limited Heaps’ Law

To obtain Eq. (2), we start from the assumption that target words to a given cue have a power-law rank distribution. This assumption is justified by the measured average frequency rank displayed in Fig. 5.

Let us now describe the problem in an abstract way. We consider a set of V events, each with a given probability. These probabilities, arranged in decreasing order, form a power law of the form:

$$f_{\mathrm{th}}(r) = A(V, B) \cdot r^{-B}, \tag{B.1}$$

where r is the rank, $A(V, B)$ is the normalization coefficient and B is a parameter of the distribution. We perform s extractions of such events, and we want to compute the number k of *distinct* events obtained. After s extractions, the minimum finite value for the frequency of an event is $1/s$. The empirically obtained rank distribution will therefore give a curve such as the ones of Fig. 5, i.e., a power-law decay followed by a plateau at $1/s$. Only the probabilities larger than $1/s$ can thus be correctly estimated. The number of distinct events with probability larger than $1/s$, $n_{>1/s}$, is given by the rank corresponding to the probability $1/s$, i.e., from Eq. (B.1):

$$f(r) = A(V, B) \cdot r^{-B} \simeq 1/s \Rightarrow n_{>1/s} \simeq r(1/s) = (s \cdot A)^{1/B}. \tag{B.2}$$

We also have to estimate the number of distinct events which have been extracted although their probability is lower than $1/s$, as the cumulative probability

$$P_{<1/s} = \sum_{r_{1/s}}^{V} f(r) \tag{B.3}$$

can be larger than $1/s$.

We can reasonably assume that these events are extracted at most once, and estimate their number $n_{<1/s}$ as:

$$n_{<1/s} \simeq P_{<1/s} \cdot s \tag{B.4}$$

The total number of distinct events observed k after s extractions is then obtained by summing (B.2) and (B.4):

$$k(s) \simeq n_{<1/s} + n_{>1/s} = P_{<1/s} \cdot s + (s \cdot A)^{1/B}. \tag{B.5}$$

After straightforward computations we obtain

$$k(s) = V \cdot \frac{B(s/s_{\mathrm{lim}})^{1/B} - s/s_{\mathrm{lim}}}{B - 1}, \tag{B.6}$$

where $s_{\mathrm{lim}} = V^B/A = (V^B - V)/(B - 1)$, which is equivalent to Eq. (2).

The treatment presented so far does not include the case $B = 1$. In this case it is easy to find:

$$k(s) = V \frac{s}{s_{\mathrm{lim}}} \left(1 - \log\left(\frac{s}{s_{\mathrm{lim}}}\right)\right) \tag{B.7}$$

It is worth to note that, if $V \gg s$ one recovers Heaps' law, which is a particular case of Eq. (2). In this limit, in fact, $s/s_{\mathrm{lim}} \ll 1$ and the dominant term in (B.6) is the one with exponent $1/B$ for $B > 1$ and the linear one for $B < 1$. For $B > 1$ one recovers the right behavior with a power law with an exponent $1/B$:

$$k(s) \sim V \frac{B}{B-1} \frac{s}{s_{\mathrm{lim}}}^{1/B} \tag{B.8}$$

while for $B < 1$ one recover a linear behavior:

$$k(s) \sim \frac{V}{1-B} \frac{s}{s_{\mathrm{lim}}}. \tag{B.9}$$

Of course in the limit $V \to \infty$ one is able to observe these behaviors in a very large range of values for s.

Appendix C. Wordnet Semantic Relations Abbreviations

Here we report a list of the semantic relations recognized by WordNet with their abbreviation. For further information we refer to the WordNet documentation [20, 22, 21].

Appendix Table 1.

Semantic relation	Abbreviation	Example
keyword	KEY	"cold" is the keyword for "icy"
synonym	SYN	"auto" is a synonym of "car"
antonym	ANT	"up" is an antonym of "down"
hypernym	HYPER	"tree" is an hypernym of "oak"
hyponym	HYPO	"trout" is an hyponym of "fish"
entails	ENT	"dream" entails "sleep"
similar	SIM	"mistaken" is similar to "wrong"
is member	ISMEM	"juror" is member of "jury"
is material	ISMAT	"iron" is material of "steel"
is part	ISPAR	"month" is part of an "year"
has member	HASMEM	"fleet" has "ship" as member
has material	HASMAT	"air" has "oxygen" as material
has part	HASPAR	"hour" has "minute" as part
cause to	CAU	"teach" cause to "learn"
is participle	PRTC	"forced" is participle of "force"
see also	SEE	"race" see also "speed"
refers to	REF	"italian" refers to "Italy"
is attribute	ATTR	"height" is the attribute of "tall"
verb group	VERBG	"incinerate" belongs to the verb group of "burn"
derivation	DER	"sailor" derives from "sail"
belongs to domain (category)	INDOMC	"lithograph" belongs to the "art" domain (category)
belongs to domain (use)	INDOMU	"google" belongs to the "trademark" domain (use)

Appendix Table 1. (*Continued*)

Semantic relation	Abbreviation	Example
belongs to domain (region)	INDOMR	"chili" belongs to the "Mexico" domain (region)
is domain (category)	ISDOMC	"biology" is the domain (category) for "cell"
is domain (use)	ISDOMU	"jargon" is the domain (use) for "trash"
is domain (region)	ISDOMR	"Japan" is the domain (region) for "origami"

References

[1] Barrat, A., Barthélemy, M. and Vespignani, A., *Dynamical Processes on Complex Networks* (Cambridge University Press, 2008).

[2] Benz, D., Grobelnik, M., Hotho, A., Jäschke, R., Mladenic, D., Servedio, V. D. P., Sizov, S. and Szomszor, M., Analyzing tag semantics across tagging systems, in *Social Web Communities*, number 08391 in Dagstuhl Seminar Proceedings (Schloss Dagstuhl — Leibniz-Zentrum für Informatik, Germany, 2008), http://drops.dagstuhl.de/opus/volltexte/2008/1785.

[3] Broder, M *et al.*, Graph structure in the Web, *Comput. Netw.* **33**(1–6) (2000) 309–320.

[4] Buchanan, M., *The Social Atom* (Bloomsbury, New York, NY, USA, 2007).

[5] Caldarelli, G., *Scale-Free Networks* (Oxford University Press, 2007).

[6] Castellano, C., Fortunato, S. and Loreto, V., Statistical physics of social dynamics, *Rev. Mod. Phys.* **81** (2009) 591–646.

[7] Cattuto, C., Baldassarri, A., Servedio, V. D. P. and Loreto, V., Emergent community structure in social tagging systems, *Adv. Compl. Syst.* **11** (2008) 597–608.

[8] Cattuto, C., Barrat, A., Baldassarri, A., Schehr, G. and Loreto, V., Collective dynamics of social annotation, *PNAS* **106** (2009) 10511–10515.

[9] Catutto, C., Schmitz, C., Baldassarri, A., Servedio, V. D. P., Loreto, V., Hotho, A., Grahl, M. and Stumme, G., Network properties of folksonomies, *AI Communications Journal, Special Issue on 'Network Analysis in Natural Sciences and Engineering'* (2007).

[10] Church, K. and Hanks, P., Word association norms, mutual information, and lexicography, *Comput. Linguist.* **16** (1990) 22–29.

[11] Dorogovtsev, S. N. and Mendes, J. F. F., *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford University Press, 2003).

[12] Ferrer i Cancho, R. and Solé, R., V., *The Small World of Human Language*, *Proc. R. Soc. Lond. B, November 7* **268** (2001) 2261–2265; doi:10.1098/rspb.2001.1800.

[13] Gabler, K., Kyle gabler's web page, http://kylegabler.com/.

[14] Golder, S. and Huberman, B. A., The structure of collaborative tagging systems, *J. Inf. Sci.* **32** (2006) 198–208.

[15] Heaps, H. S., *Information Retrieval: Computational and Theoretical Aspects* (Academic Press, Inc., Orlando, FL, USA, 1978).

[16] Helbing, D., Traffic and related self-driven many-particle systems, *Rev. Mod. Phys.* **73** (2001) 1067–1141.

[17] van Leijenhorst, D. C. and van der Weide, T. P., *A Formal Derivation of Heaps Law*, Vol. 170 (Information Sciences, 2005), p. 263.

[18] Mathes, A., Folksonomies — cooperative classification and communication through shared metadata (2004), http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html.

[19] Mehler, A., Large text networks as an object of corpus linguistic studies (2007).

[20] Miller, G. A., Wordnet — about us, http://wordnet.princeton.edu.

[21] Miller, G. A., Wordnet: A lexical database for english., *Commun. ACM* **38** (1995) 39–41.

[22] Miller, G. A. and Fellbaum, C., *WordNet: An Electronic Lexical Database* (MIT Press, Cambridge, MA, 1998).

[23] Nagatani, T., The physics of traffic jams, *Rep. Prog. Phys.* **65** (2002) 1331–1386.

[24] Nelson, D., McEvoy, C. and Schreiber, T., The university of south florida free association, rhyme, and word fragment norms, *Behav. Res. Methods* **36** (2004) 402–407.

[25] Newman, M. E. J., Assortative mixing in networks, *Phys. Rev. Lett.* **89** (2002) 208701.

[26] Newman, M. E. J., Forrest, S. and Balthrop, J., Email networks and the spread of computer viruses, *Phys. Rev. E* **66** (2002) 035101.

[27] Pastor-Satorras, R. and Vespignani, A., *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge University Press, New York, NY, USA, 2004).

[28] Sowa, J. F., *Conceptual Structures: Information Processing in Mind and Machine* (Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984).

[29] Steyvers, M. and Tenenbaum, J. B., The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth, *Cognitive Sci.* **29** (2005) 41–78.

[30] Vazquez, A., *Statistics of Citation Networks*, 05/2001, arXiv:cond-mat/0105031.

[31] Zipf, G. K., *Human Behavior and the Principle of Least Effort* (Hafner, New York, 1965).