The Topology of Semantic Knowledge

Jimmy Dubuisson **Jean-Pierre Eckmann**
Département de Physique Théorique and Section de Mathématiques
Université de Genève
Jimmy.Dubuisson@unige.ch

Christian Scheible
Institut für Maschinelle Sprachverarbeitung
University of Stuttgart
scheibcn@ims.uni-stuttgart.de

Hinrich Schütze
Center for Information
and Language Processing
University of Munich

Abstract

Studies of the graph of dictionary definitions (DD) (Picard et al., 2009; Levary et al., 2012) have revealed strong semantic coherence of local topological structures. The techniques used in these papers are simple and the main results are found by understanding the structure of cycles in the directed graph (where words point to definitions). Based on our earlier work (Levary et al., 2012), we study a different class of word definitions, namely those of the Free Association (FA) dataset (Nelson et al., 2004). These are responses by subjects to a cue word, which are then summarized by a directed, free association graph.

We find that the structure of this network is quite different from both the Wordnet and the dictionary networks. This difference can be explained by the very nature of free association as compared to the more "logical" construction of dictionaries. It thus sheds some (quantitative) light on the psychology of free association.

In NLP, semantic groups or clusters are interesting for various applications such as word sense disambiguation. The FA graph is tighter than the DD graph, because of the large number of triangles. This also makes drift of meaning quite measurable so that FA graphs provide a quantitative measure of the semantic coherence of small groups of words.

1 Introduction

The computer study of semantic networks has been around since the advent of computers (Brunet, 1974) and has been used to study semantic relations between concepts and for analyzing semantic data. Traditionally, a popular lexical database of English is Wordnet (Miller, 1995; Miller and Fellbaum, 1998), which organizes the semantic network in terms of graph theory. In contrast to manual approaches, the automatic analysis of semantically interesting graph structures of language has received increasing attention. For example, it has become clear more recently that cycles and triangles play an important role in semantic networks, see *e.g.*, (Dorow et al., 2005). These results suggest that the underlying semantic structure of language may be discovered through graph-theoretical methods. This is in line with similar findings in much wider realms than NLP (Eckmann and Moses, 2002).

In this paper, we compare two different types of association networks. The first network is constructed from an English dictionary (DD), the second from a free association (FA) database (Nelson et al., 2004). We represent both datasets through directed graphs. For DD, the nodes are words and the directed edges point from a word to its definition(s). For FA, the nodes are again words, and each cue word has a directed edge to each association it elicits.

Although the links in these graphs were not constructed by following a rational centralized process, their graph exhibits very specific features and we concentrate on the study of its topological properties. We will show that these graphs are quite different in global and local structure, and we interpret this as a reflection of the different nature of DD vs. FA. The first is an *objective* set of rela-

Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 669–680,
Seattle, Washington, USA, 18-21 October 2013.

tions between words and their meaning, as explained by other words, while the second reveals the nature of *subjective* reactions to cue words by individuals. This matter of fact is reflected by several quantitative differences in the structure of the corresponding graphs.

The main contribution of this paper is an empirical analysis of the way semantic knowledge is structured, comparing two different types of association networks (DD and FA). We conduct a mathematical analysis of the structure of the graphs to show that the way humans express their thoughts exhibits structural properties in which one can find semantic patterns. We show that a simple graph-based approach can leverage the information encoded in free association to narrow down the ambiguity of meaning, resulting in precise semantic groups. In particular, we find that the main strongly connected component of the FA graph (the so-called **core**) is very cyclic in nature and contains a large predominance of short cycles (*i.e.*, co-links and triangles). In contrast to the DD graph, bunches of triangles form well-delimited lexical fields of collective semantic knowledge. This property may be promising for downstream tasks. Further, the methods developed in this paper may be applicable to graph representations that occur in other problems such as word sense disambiguation (*e.g.*, (Heylighen, 2001; Agirre and Soroa, 2009)) or sentiment polarity induction (Hassan and Radev, 2010; Scheible, 2010).

To show the semantic coherence of these lexical fields of the FA graph, we perform an experiment with human raters and find that cycles are strongly semantically connected even when compared to close neighbors in the graph.

The reader might wonder why sets of pairwise associations can lead to any interesting structure. One of the deep results in graph theory, (Bollobás, 2001), is that in sparse graphs, *i.e.*, in graphs with few links per node, the number of triangles is extremely rare. Therefore, if one does find many triangles in a graph, they must be not only a signal of non-randomness, but carry relevant information about the domain of research as shown earlier (Eckmann and Moses, 2002).

2 The USF FA dataset

This dataset is one of the largest existing databases of free associations (FA) and has been collected at the University of South Florida since 1973 by researchers in psychology (Nelson et al., 2004). Over the years, more than 6'000 participants produced about 750'000 responses to 5'019 stimulus words.

The procedure for collecting the data is called discrete association task and consists in asking participants to give the first word that comes to mind (**target**) when presented a stimulus word (**cue**).

For creating the initial set of stimulus words, the Jenkins and Palermo word association norms (Palermo and Jenkins, 1964) proved useful but too limited as they consist of only 200 words. For this reason, additional words have been regularly added to the pool of normed words, unfortunately without well established rules being followed. For instance, some were selected as potentially interesting cues, some were added as responses to the first sets of cues and, some others were collected for supporting new studies on verbs. We still work with this database, because of its breadth.

The final pool of stimuli comprises 5'019 words of which 76% are nouns, 13% adjectives, and 7% verbs. A word association is said to be **normed** when the target is also part of the set of norms, *i.e.*, a cue. The USF dataset of free associations contains 72'176 cue-target pairs, 63'619 of which are normed. As an example, the association *puberty-sex* is normed whereas the association *puberty-thirteen* is not, because *thirteen* is not a cue.

3 Mathematical definitions

We collect here those notions we need for the analysis of the data.

A **directed graph** is a pair $G = (V, E)$ of a set V of **vertices** and, a set E of ordered pairs of vertices also called **directed edges**. For a directed edge $(u, v) \in E$, u is called the **tail** and v the **head** of the edge. The number of edges incident to a vertex $v \in V$ is called the **degree** of v. The **in-degree** (resp. **out-degree**) of a vertex v is the number of edge heads (resp. edge tails) adjacent to it. A vertex with null in-degree is called a **source** and a vertex with null out-degree is called a **sink**.

A **directed path** is a sequence of vertices such

that a directed edge exists between each consecutive pair of vertices of the graph. A directed graph is said to be **strongly connected**, (resp. **weakly connected**) if for every pair of vertices in the graph, there exists a directed path (resp. undirected path) between them. A **strongly connected component**, SCC, (resp. **weakly connected component**, WCC) of a directed graph G is a maximal strongly connected (resp. weakly connected) subgraph of G.

A **directed cycle** is a directed path such that its **start vertex** is the same as its **end vertex**. A **colink** is a directed cycle of length 2 and a **triangle** a directed cycle of length 3.

The **distance** between two vertices in a graph is the number of edges in the shortest path connecting them. The **diameter** of a graph G is the greatest distance between any pair of vertices. The **characteristic path length** is the average distance between any two vertices of G.

The **density** of a directed graph $G(V, E)$ is the proportion of existing edges over the total number of possible edges and is defined as:

$$d = |E|/(|V|(|V| - 1))$$

The **neighborhood** N_i of a vertex v_i is $N_i = \{v_j : e_{ij} \in E \text{ or } e_{ji} \in E\}$.

The **local clustering coefficient** C_i for a vertex v_i corresponds to the density of its neighborhood subgraph. For a directed graph, it is thus given by:

$$C_i = \frac{|\{e_{jk} : v_j, v_k \in N_i, e_{jk} \in E\}|}{|N_i|(|N_i| - 1)}$$

The **clustering coefficient** of a graph G is the average of the local clustering coefficients of all its vertices.

The **efficiency** Eff of a directed graph G is an indicator of the traffic capacity of a network. It is the harmonic mean of the distance between any two vertices of G. It is defined as:

$$\text{Eff} = \frac{1}{|V|(|V| - 1)} \sum_{i \neq j \in V} \frac{1}{d_{ij}}$$

The linear correlation coefficient between two random variables X and Y is defined as:

$$\rho(X, Y) = (E[XY] - \mu_X \mu_Y)/(\sigma_X \sigma_Y)$$

where μ_X and σ_X are respectively the mean and standard deviation of the random variable X.

The linear degree correlation coefficient of a graph is called **assortativity** and is expressed as:

$$\rho_D = \sum_{xy} xy(e_{xy} - a_x b_y)/(\sigma_a \sigma_b)$$

where e_{xy} is the fraction of all links that connect nodes of degree x and y and where a_x and b_y are respectively the fraction of links whose tail is adjacent to nodes with degree x and whose head is adjacent to nodes with degree y, satisfying the following three conditions:

$$\sum_{xy} e_{xy} = 1, a_x = \sum_{y} e_{xy}, b_y = \sum_{x} e_{xy}$$

When ρ_D is positive, the graph possesses **assortative mixing** and high-degree nodes tend to connect to other high-degree nodes. On the other hand, when ρ_D is negative, the graph features **disassortative mixing** and high-degree nodes tend to connect to low degree nodes.

The **intersection graph** of sets A_i, $i = 1, \ldots, m$, is constructed by representing each set A_i as a vertex $v_i \in V$ and adding an edge for each pair of sets with a non-empty intersection:

$$E = \{(v_i, v_j) : A_i \cap A_j \neq \varnothing\}$$

4 Graph topology analysis

4.1 Graph generation

Our goal being to study the FA network topology, we first concentrate on the generation of an unweighted directed graph. We generate the corresponding graph by adding a directed edge for each cue-target pair of the dataset. We only consider pairs whose target was normed in order to avoid overloading the graph with noisy data (*e.g.*, a response meaningful only to a specific participant). The graph has 5'019 vertices and 63'619 edges. It is composed of a single WCC and 166 SCCs.

For comparison with dictionary definitions (DD), we construct a graph from the Wordnet2 dictionary (nouns only), following (Levary et al., 2012). This graph contains 54'453 vertices and 179'848 edges.

4.2 Core extraction

The so-called **core** was defined previously in (Picard et al., 2009; Levary et al., 2012) as that subset of nodes in which a random walker gets trapped after only a few steps.

The shave algorithm was used in (Levary et al., 2012) to isolate this subset. It consists in recursively removing the source and sink nodes from a weakly connected directed graph and permits to get the subgraph induced by the union of its strongly connected components. Note that the dictionary graph (DD) has no sinks (*i.e.*, words that never get defined) and that it contains a giant SCC whose size is comparable to the one of the initial graph.

It turns out that the FA graph also contains a giant SCC, therefore getting the core consists more simply in extracting the main SCC of the initial graph. We use Tarjan's algorithm (Tarjan, 1972) for isolating the FA core.

4.3 Vertex degree analysis

The FA core has a maximum in-degree of 313, a maximum out-degree of 33 and an average degree of 25.42. The in-degree distribution follows a power law ($\gamma = 1.93$) and the out-degrees are Poisson-like distributed with a peak at 14 (Steyvers and Tenenbaum, 2005; Gravino et al., 2012).

Words having a high in-degree are **targets** that tend to be cited more frequently. On the other hand, words having a high out-degree are **cues** that evoke many different targets.

The most evocative cues are, in decreasing order of out-degree: *field* (33), *body* (31), *condemn* (29), *farmer* (29), *crisis* (28), *plan* (28), *attention* (27), *animal* (27), and *hang* (27). Interestingly, the most cited targets (*i.e.*, targets with highest in-degree) are in decreasing order: *food* (313), *money* (295), *water* (271), *car* (251), *good* (246), *bad* (221), *work* (187), *house* (183), *school* (182), *love* (179).

4.4 Cycle decomposition of the core

We define the **vertex k-cycle multiplicity** (resp. **edge k-cycle multiplicity**) as the number of k-cycles a given vertex (resp. edge) belongs to. We call **core-ER** the set of Erdös-Rényi (ER) random graphs $G(n, M)$ having the same number of nodes and the same number of edges as the FA core. We start by extracting the 2- and 3-cycles by using a customized version of Johnson's algorithm (Johnson, 1975). The first thing we observe is that the core has a very high density of short cycles: the subset of nodes belonging to 2-cycles (*i.e.*, nodes with 2-cycle multiplicities > 0) cover 95% of the core vertices and the 3-cycles cover 88% of the core vertices. The corresponding core-ER graphs have on average about 100 times fewer 2-cycles and almost 20 times fewer 3-cycles.

This shows that the core is very cyclic in nature and that it remains very well connected for short-length cycles: most vertices of the core indeed belong to at least one co-link or triangle.

In order to limit computation times, we only considered shortest cycles for lengths ≥ 3 and analyzed the distribution of the number of shortest cycles in the core compared to equivalent random graphs. Whereas there are many more short cycles in the core, we observe a predominance of 4, 5 and 6-cycles in core-ER graphs. However, we find again a slight predominance of long cycles (length between 7 and 15) in the core (see Fig. 1). See (Levary et al., 2012), Fig. 3, where the cycle distribution is very different, with a minimum at length 5.

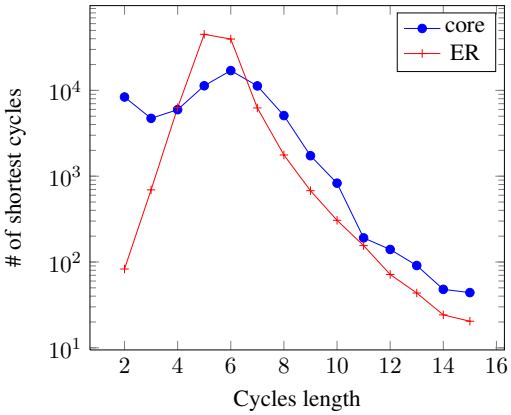

Figure 1: Distribution of shortest cycles lengths in the core compared to equivalent ER models

One should bear in mind that we only consider the set of *shortest* cycles. Thus, a k-cycle is not counted if each of its nodes belongs to a cycle whose length is $< k$. Although the number of 4-shortest cycles is comparable in the core and core-ER graphs for example, there are in reality far more 4-cycles in the core (*i.e.*, 42'738 versus 6'517). We see that when considering shortest cycles, short cycles tend to hide long ones, and, as a large proportion of nodes in the core belong to 2- and 3-cycles, many longer cycles do not get counted at all.

4.5 Interpretation of cycles

2-cycles are composed of concretely related words (*e.g.*, *drug-coke*, *destiny-fate*, *einstein-genius*, ...). The vertex with highest 2-cycle multiplicity is *music* (22).

Words in 3- and 4-cycles often belong to the same lexical field. Examples of 3-cycles: *protect-guard-defend* or *space-universe-star*. The vertex (resp. edge) with highest 3-cycle multiplicity is *car* (86) (resp. *bad-crime* (11)). Examples of 4-cycles: *monster-dracula-vampire-ghost* or *flu-virus-infection-sick*.

Longer cycles are more difficult to describe: Relations linking words of a given cycle exhibit semantic drift with increasing length (cf. (Levary et al., 2012)). Examples of 5-cycles: *yellow-coward-chicken-soup-noodles* and *sleep-relax-music-art-beauty*.

The cumulated set of free associations reflects the way in which a group of people retrieved its semantic knowledge. As the associated graph is highly circular, this suggests that this knowledge is not stored in a hierarchical way (Steyvers and Tenenbaum, 2005). The large predominance of short cycles in the core may indeed be a natural consequence of the semantic information being acquired by means of associative learning (Ashcraft and Radvansky, 2009; Shanks, 1995).

4.6 FA core clustering

4.6.1 The walktrap community algorithm

Complex networks are globally sparse but contain locally dense subgraphs. These groups of highly interconnected vertices are called **communities** and convey important properties of the network.

Although the notion of community is difficult to define formally, the current consensus establishes that a partition $P = \{C_1, C_2, \ldots, C_k\}$ of the vertex set of a graph G represents a good community structure if the proportion of edges inside the C_i is higher than the proportion of edges between them (Fortunato, 2010).

Computing such communities in a large graph is generally computationally expensive (Lancichinetti and Fortunato, 2009). We use the so-called 'Walktrap' community detection algorithm (Pons and Latapy, 2006) for extracting communities from the FA networks. The idea lying behind this algorithm is that random walks on a graph will tend to get trapped in the densely connected subgraphs.

Let P_{ij}^t be the probability of going from vertex i to vertex j through a random walk of length t. The distance between two vertices i and j of the graph is defined as:

$$r_{ij}(t) = \sqrt{\sum_{k=1}^{n} \frac{(P_{ik}^t - P_{jk}^t)^2}{d(k)}}$$

where $d(k)$ is the degree of vertex k.

One defines the probability $P_{C,j}^t$ to go from community C to vertex j in t steps: $P_{C,j}^t = \sum_{i \in C} P_{ij}^t / |C|$, and then the distance is easily generalized for two communities C_1, C_2.

The algorithm starts with a partition $P_1 = \{\{v\} \in V\}$ of the initial graph into n communities each of which is a single vertex. At each step, two communities are chosen and merged according to the criterion described below and the distances between communities are updated. The process goes on until we obtain the partition $P_n = \{V\}$.

In order to reduce complexity, only adjacent communities are considered for merging. The decision is then made according to Ward's method (Everitt et al., 2001): at each step k, the two communities that minimize the mean σ_k of the squared distances between each vertex and its community are merged:

$$\sigma_k = \frac{1}{n} \sum_{C \in P_k} \sum_{i \in C} r_{iC}^2$$

4.6.2 Clustering of the core

We first identify the communities of the FA core using the Walktrap algorithm. We immediately observe that when the path length parameter increases, the number of identified communities decreases (*i.e.*, for a length of 2, we find 35 communities whereas for a length of 9, we only find 8 communities).

For a path length of 2, the algorithm extracts 35 communities, 7 of which contain more than 100 vertices, 3 of which contain between 100 and 50 vertices and 25 of which contain less than 50 vertices.

We observe that for most small communities (*i.e.*, the ones containing less than 50 vertices), there exists a clear relation between the labels of their ver-

tices. Typically, the labels are part of the same lexical field (*e.g.*, all the planets (except *earth*) or related by a common grammatical function (such as *why*, *where*, *what*, ...).

4.6.3 Clustering of the core co-links

We define the **k-cycle induced subgraph** of a graph G as the subgraph of G induced by the set of its vertices with k-cycle multiplicity > 0.

The **co-link graph** of a graph $G(V, E)$ is the undirected graph obtained by replacing each co-link (*i.e.*, 2-cycle) of the 2-cycle induced subgraph of G by a single undirected edge and removing all other edges.

The co-link graph of the FA core has 4'508 vertices and 8'309 edges for a density of 8×10^{-4}. It is composed of a single weakly connected component that can be seen as a projection of the strongest semantic links from the original graph. Extracting the co-link graph is thus an efficient way of selecting the set of most important semantic links (*i.e.*, the set of 2-cycles that appear in large predominance in the core compared to what is found in an equivalent random graph) while filtering out the noisy or negligible ones.

The sets of communities extracted by the Walktrap algorithm exhibit different degrees of granularity depending on the length parameter. For short paths, a large number of very small communities are returned (*e.g.*, 923 communities when length equals 2) whereas for longer paths the average size of the communities increases more and more.

The community detection exhibits thus a far finer degree of granularity for the core co-links graph than for the core itself. The size of the communities being much smaller in average, it is striking to notice to which extent the words of a given community are semantically related.

Examples of communities found in the core co-links graph include (*standards*, *values*, *morals*, *ethics*), (*hopeless*, *romantic*, *worthless*, *useless*), (*thesaurus*, *dictionary*, *vocabulary*, *encyclopedia*) or (*molecule*, *atom*, *electron*, *nucleus*, *proton*, *neutron*).

4.6.4 DD core clustering vs FA core clustering

The clustering of both cores has very different characteristics: We illustrate the neighborhoods of *conflict* for both cases in Fig. 2 and 3.

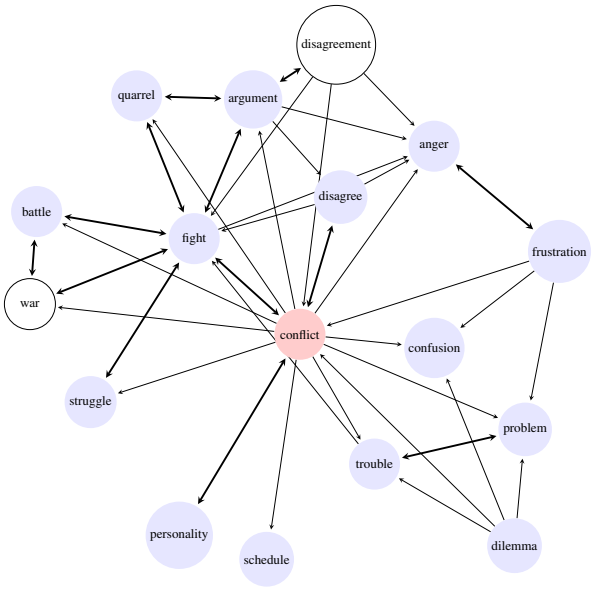

Figure 2: Neighborhood of *conflict* in the FA core
The set of words belonging to the neighborhood of *conflict* are clearly part of the same lexical field. The high density of co-links leads to cyclicity and we see that many directed triangles are present in the local subgraph (*e.g.*, *conflict-trouble-fight, conflict-argument-disagree*). We can even find triangles of co-links that link together words semantically strongly related (*e.g.*, *fight-war-battle, fight-quarrel-argument*). Nodes that are part of the neighborhood of *conflict* in both FA and DD are in empty circles.

On one hand, the words in communities of the DD core are in most cases either synonyms, *e.g.*, *(declaration, assertion, claim)* or an instance-of kind of relation, *e.g.*, *(signal, gesture, motion)* or *(zero, integer)*.

On the other hand, communities of the FA core are generally composed of words belonging to the same lexical field and sharing the same level of abstraction.

Moreover, we notice that it is often difficult to establish the semantic relation existing between words of many small communities (*i.e.*, containing less than 10 words) of the DD core. Two such examples are: (*choice*, *probate*, *executor*, *chosen*, *certificate*, *testator*, *will*) and (*numeral*, *monarchy*, *monarch*, *crown*, *significance*, *autocracy*, *symbol*, *interpretation*).

The comparison of DD and FA reveals, in a quantitative way, fundamental differences between the two realms. The interesting data are shown in table 1.

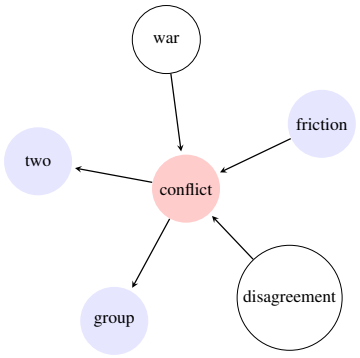

Figure 3: Neighborhood *conflict* in the DD core
First, we note that the neighborhood has a lower density than in the FA core. We also see that there is no cycle and there seems to be a flow going from source nodes to sink nodes. As it generally happens in the neighborhood subgraphs of the DD core, source nodes are rather specific words whereas sink nodes are generic words.

	FA core	DD core
# vertices	4'843	1'496
# edges	61'544	4'766
density	2.5×10^{-3}	2.1×10^{-3}
avg degree	25.4	6.37
max in-degree	313	59
directed diameter	10	29
characteristic path length	4.26	10.42
efficiency	2.5×10^{-1}	1.2×10^{-1}
clustering coefficient	8.5×10^{-2}	5.1×10^{-2}
assortativity	5.5×10^{-2}	6.1×10^{-2}

Table 1: Comparison FA vs DD

Note that while the FA core is in fact larger than the DD core, its diameter is smaller. This illustrates in a beautiful way the nature of free association as compared to the more neutral dictionary. In particular, the characteristic path length is smaller in the FA graph, because humans use generalized event knowledge (McRae and Matsuki, 2009) in free association, producing semantic shortcuts. For example, FA contains a direct link *mirage→water*, whereas in DD, the shortest path between the two words is *mirage→refraction→wave→water*.

5 The Bricks of Meaning

5.1 Extraction of the seed

We already saw that most vertices of the core belong to directed 2- and 3-cycles. Whereas 2-cycles establish strong semantic links (*i.e.*, synonymy or antonymy relations) and provide cyclicity to the underlying directed graph, we claim that 3-cycles (*i.e.*, triangles) form the set of elementary concepts of the core.

These structues are common to DD and to FA, but we will see that the links in FA are somehow more direct than in DD.

We call **seed** the subgraph of the core induced by the set V_3 of vertices belonging to directed triangles and **shell** the subgraph of the core induced by the set $V \backslash V_3$ (*i.e.*, the set of vertices with a null 3-cycle multiplicity), see Fig. 4.

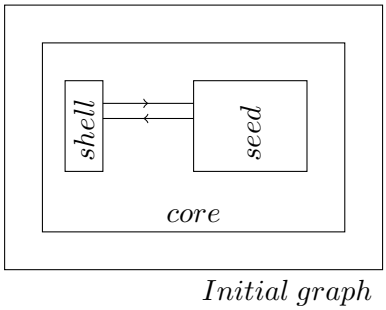

Figure 4: Composition of the FA graph
The graph of FA contains a giant SCC (the core). The subgraph of the core induced by the set of nodes belonging to at least one triangle also forms a giant component we call the 'seed'. The subgraph of the core induced by the set of nodes not belonging to any triangle is called the 'shell' and is composed of many small SCCs, including single vertices. Although the shell has a low density, its nodes are very well connected to the seed.

The shell contains 530 nodes and 309 edges. There are 7'035 edges connecting the shell to the seed. The shell consists of of many small SCCs and although its average degree is low (1.17), its vertices have on average many (13.27) connections to the seed.

The seed contains 4'313 vertices (89% of the core) and 54'197 edges. The first thing to notice is that it has 100 times more co-links (7'895) and 20 times more triangles (13'119) than an equivalent random graph. We call **shortcuts** the 32'773 edges of the seed that do not belong to 3-cycles, see Fig. 5.

The seed obviously also contains cycles whose length is greater than 3. One can check that there exist only 5 basic motifs involving 2 attached triangles and 1 shortcut for creating 4- and 5-cycles, and that linking 2 isolated triangles with 2 shortcuts also per-

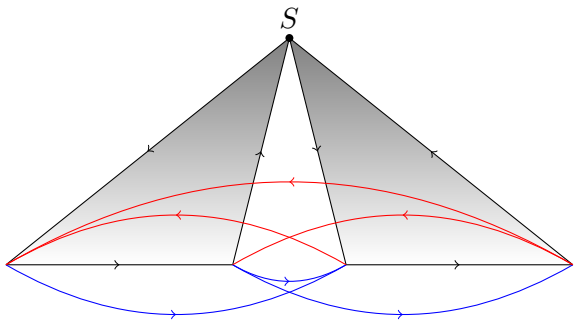

Figure 5: Shortcut edges between two triangles sharing a single vertex S
Two triangles can share 0, 1 or 2 vertices. For each of these three basic motifs, we count the maximum number of shortcut edges (*i.e.*, edges not belonging to 3-cycles) that can be added. By linking two triangles, these shortcuts permit to move two basic semantic units closer together and create longer cycles (*i.e.*, 4, 5, and 6-cycles). Long cycles can be thus considered as groupings of basic semantic units. In the case of two triangles sharing one vertex for example, it is possible to add at most 6 shortcuts, whereas, for two triangles sharing two vertices, at most 2 shortcuts can be added.

mit to form 4-, 5- and 6-cycles. All longer cycles are simply made of a juxtaposition of these basic motifs.

Furthermore, there is a limit on the number of shortcuts that can possibly be added in the seed before it gets saturated, as all its vertices belong to at least one triangle. We show that at most 16 shortcuts can be added between two isolated triangles, at most 6 between 2 triangles sharing 2 vertex and at most 2 between 2 triangles sharing 2 vertices (see Fig. 5).

5.2 The elementary lexical fields

Once the seed is isolated, we go on digging into its structure. We focus on the arrangements of triangles as they constitute the set of elementary concepts.

We start by removing all shortcuts from the seed and convert it then to an undirected graph, in order to get a homogeneous simplicial 2-complex.

Let **t** be the graph operator which transforms a graph G into the intersection graph $\mathbf{t}G$ of its 2-simplices (*i.e.*, triangles sharing an edge). We apply **t** to the homogeneous simplicial 2-complex found previously. The result represents the links between the basic semantic units of the seed. We call **seed-crux** the giant WCC in the intersection graph.

We enumerate the 8'380 maximal cliques of FA seed-crux and get the list of words composing each

Distance	Acc	κ	KS
original	–	0.404	30
1	74	0.522	42
2	97	0.899	89
∞	99	0.899	89

Table 2: Accuracy, κ, and *count*$(p < 0.05)$ for KS

clique. By removing the ones that are subsets of bigger lists, we finally obtain 3'577 lists of words .

These lists of words have a rather small and homogeneous size (between 4 and 17) and 95% have a size comprised between 4 and 10. More interestingly, they clearly define well-delimited lexical fields. We will show this through two experiments in the following sections. A few examples include *(honest, trustworthy, reliable, responsible)*, *(stress, problem, worry, frustration)* and *(data, process, computer, information)*.

From a topological perspective, we deduce that bunches of triangles (*i.e.*, cliques of elementary concepts) span the seed in a homogeneous way. These bunches form a set of cohesive lexical fields and constitute essential bricks of semantic knowledge.

5.3 Semantic similarity of the lexical fields

In order to quantify the relative meaning of words in the lexical fields of the seed-crux, we define the following semantic similarity metric based on the Wordnet *WUP* metric (Wu and Palmer, 1994) for a given set of words L:

$$S_\ell(L) = 2 \sum_{w_i,w_j \in L, w_i \neq w_j} S_w(w_i, w_j)/(|L|(|L|-1))$$

where $S_w(w_i, w_j) = \max_{S_k \ni w_i \text{and} S_\ell \ni w_j} \{wup(S_k, S_\ell)\}$ and wup is the WUP semantic metric and S_k and S_ℓ are Wordnet synsets.

The average value of S_ℓ for the set of cliques of seed-crux is 0.6 whereas it is only 0.43 for randomly sampled set of words. This suggests the corresponding lists of words are indeed semantically related. We will show the strength of this relation in the following experiment with human raters.

5.4 Human evaluation of the lexical fields

To validate our findings, we conducted an empirical evaluation through human annotators. Starting from

the 1'204 4-groups, we designed the following experiment: We corrupt the groups by exchanging one of the 4 elements with a randomly chosen word at a distance from the group of 1, 2, and "infinity" (*i.e.*, any word of the whole core). We presented 100 random samples for each of the 3 distances as well as 100 unperturbed groups (*original*) to annotators at Amazon Mechanical Turk[1], asking which word fits the group the least. Intuitively, the closer the randomly chosen words get to the group, the closer the distribution of the votes for each sample should be to the uniform distribution. We collected 10 votes for each of the 4 problems of 100 random samples. We calculated accuracy (*i.e.*, the relative frequency of correctly identified random words) for the 3 random confounder experiments and Fleiss' κ. Further, we used the Kolmogorov-Smirnov (KS) test for how uniform the label distribution is, reporting the relative frequency of samples that are significantly ($p < 0.05$) different from the uniform distribution. The results of this experiment are summarized in Table 2 and show clearly that the certainty about the "odd man out" increases together with the distance.

5.5 Error analysis

If we view our results as a resource for a downstream task, it is important to know about possible downsides. First, we note that there are words which are not in a triangle and will thus be missing in the intersection graph. This is an indication that the corresponding word is less well embedded contextually, so conversely, any prediction made about it from the data may be less reliable. Additionally, semantic leaps caused by generalized event knowledge may lead to lesser-connected groups such as (*steel*, *pipe*, *lead*, *copper*). Jumps like these may or may not be desired in a subsequent application.

6 The Case of the EAT FA dataset

The Edinburgh Associative Thesaurus (EAT) (Kiss et al., 1973) is a large dataset of free associations. We extract the EAT FA seed-crux with the previously described methods.

We start by generating the initial graph (23'219 vertices and 325'589 edges), then extract its core (7'754 vertices and 247'172 edges) and its seed (7'500 vertices and 238'677 edges). It is interesting to notice at this stage that the EAT seed contains 74% of the words belonging to the USF seed. After generating the seed-crux which contains 63'363 vertices, 6'825'731 edges, and 342'490 maximal cliques, we finally obtain 40'998 lists of words.

These lists comprise between 4 and 233 words but 80% of them have a relatively small size between 4 and 20. Although we find exceptions for this graph, most of the extracted lists again form well-delimited lexical fields (*e.g.*, *(health, resort, spa, bath, salts)* or *(god, devil, angel, satan)*.

Comparing the two association experiments, we see that the local topologies are quite similar. Both FA cores have a high density of connected triangles, whereas cycles in the DD graph tend to be longer and most triangles are isolated. This can be attributed to the different ways in which DD and FA are obtained, the former being built rationally by following a humanly-driven process and the latter reflecting an implicit collective semantic knowledge.

7 Related Work

A number of metrics like *Latent Semantic Analysis* (Deerwester et al., 1990) and *Word Association Spaces* (Steyvers et al., 2004) have been recently developed for quantifying the relative meaning of words. As the topological properties of free association graphs reflect key aspects of semantic knowledge, we believe some graph theory metrics could be used efficiently to derive new ways of measuring semantic similarity between words.

Topological analysis of the Florida Word Associations (FA) was started by (Steyvers and Tenenbaum, 2005; Gravino et al., 2012), who extracted global statistics. We follow the basic methodology of these studies, but extend their approach. First, we conduct deeper analyses by examining the neighborhood of nodes and extracting the statistics of cycles. Second, we compare the properties of FA and DD graphs.

Word clustering based on graphs has been the subject of various earlier studies. Close to our work is (Widdows and Dorow, 2002). These authors recognize that nearest-neighbor-based clustering of co-occurrence give rise to semantic groups. This type of approach has since been applied in various modified forms, *e.g.*, by (Biemann, 2006) who performs label-

[1] `http://www.mturk.com`

propagation based on randomized nearest neighbors, or Matsuo et al. (2006) who perform greedy clustering. Hierarchical clustering algorithms (*e.g.*, (Jonyer et al., 2002; Manning et al., 2008)) are related as well, however, the key difference is that in hierarchical clustering, the granularity of a cluster is difficult to determine.

Dorow et al. (2005) recognize that triangles form semantically strongly cohesive groups and apply clustering coefficients for word sense disambiguation. Their work focuses on undirected graphs of corpus co-occurrences whereas our work builds on directed associations. Building on this work, we take finer topological graph structures into account, which is one of the main contributions in this paper.

8 Conclusion

The cognitive process of discrete free association being an epiphenomenon of our semantic memory at work, the cumulative set of free associations of the USF dataset can be viewed as the projection of a collective semantic memory.

To analyze the semantic memory, we use the tools of graph theory, and compare it also to dictionary graphs. In both cases, triangles play a crucial role in the local topology and they form the set of elementary concepts of the underlying graph. We also show that cohesive lexical fields (taking the form of cliques of concepts) constitute essential bricks of meaning, and span the core homogeneously at the global level; 89% of all words in the core belong to at least one triangle, and 77% belong to cliques of triangles containing 4 words (i.e., pairs of triangles sharing an edge or forming tetrahedras). As the words of a graph of free associations acquire their meaning from the set of associations they are involved in (Deese, 1962), we go a step further by examining the neighborhood of nodes and extracting the statistics of cycles. We further check through human evaluation that the clustering is strongly related to meaning, and furthermore, the meaning becomes measurably more confused as one walks away from a cluster.

-¿ -¿I call the pairs of triangles sharing an edge the 2-clovers ;-)

Comparing dictionaries to free association, we find the free association graph being more concept driven, with words in small clusters being on the same level of abstraction. Moreover, we think that graphs of free associations could find interesting applications for *Word Sense Disambiguation* (*e.g.*, (Heylighen, 2001; Agirre and Soroa, 2009)), and could be used for detecting psychological disorders (*e.g.*, depression, psychopathy) or whether someone is lying (Hancock et al., 2013; Kent and Rosanoff, 1910).

Finally, we believe that studying the dynamics of graphs of free associations may be of particular interest for observing the change in meaning of certain words (Deese, 1967), or more generally to follow the cultural evolution arising among a social group.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '09, pages 33–41, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mark H Ashcraft and Gabriel A Radvansky. 2009. *Cognition*. Pearson Prentice Hall.

Chris Biemann. 2006. Chinese whispers: an efficient graph clustering algorithm and its application to natural language processing problems. In *Proceedings of the First Workshop on Graph Based Methods for Natural Language Processing*, TextGraphs-1, pages 73–80, Stroudsburg, PA, USA. Association for Computational Linguistics.

Béla Bollobás. 2001. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition.

Etienne Brunet. 1974. Le traitement des faits linguistiques et stylistiques sur ordinateur. *Texte d'application: Giraudoux, Statistique et Linguistique. David, J. y Martin, R.(eds.). Paris: Klincksieck*, pages 105–137.

Scott Deerwester, Susan T. Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407.

James Deese. 1962. On the structure of associative meaning. *Psychological review*, 69:161.

James Deese. 1967. Meaning and change of meaning. *The American psychologist*, 22(8):641.

Beate Dorow, Dominic Widdows, Katerina Ling, Jean-Pierre Eckmann, Danilo Sergi, and Elisha Moses.

2005. Using curvature and Markov clustering in graphs for lexical acquisition and word sense discrimination. In *MEANING-2005, 2nd Workshop organized by the MEANING Project, February 3rd-4th 2005, Trento, Italy.*

Jean-Pierre Eckmann and Elisha Moses. 2002. Curvature of co-links uncovers hidden thematic layers in the World Wide Web. *Proc. Natl. Acad. Sci. USA*, 99(9):5825–5829 (electronic).

Brian Everitt, Sabine Landau, and Morven Leese. 2001. Cluster analysis. 4th Edition. *Arnold, London*.

Santo Fortunato. 2010. Community Detection in Graphs. *Physics Reports*, 486(3):75–174.

Pietro Gravino, Vito DP Servedio, Alain Barrat, and Vittorio Loreto. 2012. Complex structures and semantics in free word association. *Advances in Complex Systems*, 15(03n04).

Jeffrey T Hancock, Michael T Woodworth, and Stephen Porter. 2013. Hungry like the wolf: A word-pattern analysis of the language of psychopaths. *Legal and Criminological Psychology*, 18(1):102–114.

Ahmed Hassan and Dragomir Radev. 2010. Identifying text polarity using random walks. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 395–403. Association for Computational Linguistics.

Francis Heylighen. 2001. Mining associative meanings from the web: from word disambiguation to the global brain. In *Proceedings of Trends in Special Language & Language Technology*, pages 15–44.

Donald B Johnson. 1975. Finding all the elementary circuits of a directed graph. *SIAM Journal on Computing*, 4(1):77–84.

Istvan Jonyer, Diane J Cook, and Lawrence B Holder. 2002. Graph-based hierarchical conceptual clustering. *The Journal of Machine Learning Research*, 2:19–43.

Grace H Kent and Aaron J Rosanoff. 1910. *A study of association in insanity*. American Journal of Insanity.

George R Kiss, Christine Armstrong, Robert Milroy, and James Piper. 1973. An associative thesaurus of english and its computer analysis. *The computer and literary studies*, pages 153–165.

Andrea Lancichinetti and Santo Fortunato. 2009. Community detection algorithms: A comparative analysis. *Physical review E*, 80(5):056117.

David Levary, Jean-Pierre Eckmann, Elisha Moses, and Tsvi Tlusty. 2012. Loops and self-reference in the construction of dictionaries. *Phys. Rev. X*, 2:031018.

Christopher D Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*, volume 1. Cambridge University Press Cambridge.

Yutaka Matsuo, Takeshi Sakaki, Kôki Uchiyama, and Mitsuru Ishizuka. 2006. Graph-based word clustering using a web search engine. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, EMNLP '06, pages 542–550, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ken McRae and Kazunaga Matsuki. 2009. People use their knowledge of common events to understand language, and do so as quickly as possible. *Language and linguistics compass*, 3(6):1417–1429.

George Miller and Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical database*. MIT Press, Cambridge, MA.

George A Miller. 1995. WordNet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Douglas L Nelson, Cathy L McEvoy, and Thomas A Schreiber. 2004. The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, 36(3):402–407.

David S Palermo and James J Jenkins. 1964. Word association norms: Grade school through college. *University of Minnesota Press*.

Olivier Picard, Alexandre Blondin-Massé, Stevan Harnad, Odile Marcotte, Guillaume Chicoisne, and Yassine Gargouri. 2009. Hierarchies in dictionary definition space. In *Annual Conference on Neural Information Processing Systems*.

Pascal Pons and Matthieu Latapy. 2006. Computing communities in large networks using random walks. In *Journal of Graph Algorithms and Applications*, pages 284–293. Springer.

Christian Scheible. 2010. Sentiment translation through lexicon induction. In *Proceedings of the ACL 2010 Student Research Workshop*, pages 25–30, Uppsala, Sweden, July. Association for Computational Linguistics.

David R Shanks. 1995. *The psychology of associative learning*, volume 13. Cambridge University Press.

Mark Steyvers and Joshua B Tenenbaum. 2005. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, 29(1):41–78.

Mark Steyvers, Richard M Shiffrin, and Douglas L Nelson. 2004. Word association spaces for predicting semantic similarity effects in episodic memory. *Experimental cognitive psychology and its applications: Festschrift in honor of Lyle Bourne, Walter Kintsch, and Thomas Landauer*, pages 237–249.

Robert Tarjan. 1972. Depth-first search and linear graph algorithms. *SIAM journal on computing*, 1(2):146–160.

Dominic Widdows and Beate Dorow. 2002. A graph model for unsupervised lexical acquisition. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Zhibiao Wu and Martha Palmer. 1994. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, pages 133–138. Association for Computational Linguistics.