Effect of Associative Rules on the Dynamics of Conceptual Combination in a Neurodynamical Model

Sarjoun Doumit and Ali A. Minai, Senior Member, IEEE

Abstract—Most models of thought are based on the idea of associative memory: The combination of distinct concepts through their association with each other based on experience. Thus, the pattern of associations in the mind is thought to be a critical factor in determining the ideas – both familiar and unfamiliar – that thought can generate. The Itinerant Dynamics with Emergent Attractors (IDEA) model has been proposed as a computational representation of the process by which ideas might emerge in an associative memory system. However, the effect of how “association” is defined in this context remains an important unexplored issue. In this paper, we consider three different ways of defining associations between concepts based on their joint usage in texts, and show that these lead to very different dynamics in the process of emergent conceptual combinations. We use data from two real-world text corpora to instantiate the model: The collected poems of Dylan Thomas, and the Memoirs of the economist Ludwig von Mises. The two sources differ greatly in the nature of their content as well as the statistics of their word usage, which makes them interesting subjects for comparison.

I. INTRODUCTION

It is widely recognized that the association plays a fundamental role in all cognitive processes. Associations between cues and memories, stimuli and responses, features and objects have all been studied extensively through experiments and modeling. Associations are also fundamental to the process of thinking – both goal-oriented and unconstrained – and associative combination of concepts is regarded as a primary mechanism for the recall of familiar ideas and the generation of new ones [1], [2], [3], [4], [5], [6]. In an influential paper, Mednick [2] proposed that the pattern of associations in the mind plays a central role in creative thinking, and that creative individuals tended to have a larger number of associations with each concept than non-creative individuals, including some uncommon ones, termed remote associates.

Most of the work in the explicit study of conceptual associations has focused on determining word association norms [7], i.e., standard associates for individual words over a large number of speakers of a language. An alternative approach to determining the associative structure of a language is to extract associative networks from dictionaries and thesauri [8]. However, neither of these methods provides information about the pattern of associations in individuals. Cued recall and priming experiments have been used to explore both the structure of associative memory and the process of associative recall in individual subjects [9], [10], [11], [7]. More recently, the associative patterns of individuals have been obtained explicitly through experiments [12], [13], and Mednick’s hypothesis has been tested experimentally with mixed results [14]. However, thinking is a fluent process, with one thought generating another, rather than the “stop-and-go” implicit in cued recall. To model this, we recently proposed the itinerant dynamics with emergent attractors (IDEA) model, where associations among concepts are embodied in a recurrent semantic neural network whose dynamics naturally generates an itinerant sequence [15] of metastable attractors comprising combinations of coherently associated concepts. To find real-world data to instantiate this model, we have proposed using text and speech produced by individuals in natural settings. It is reasonable to assume that a sufficiently large body of natural expressions by an individual would provide information about the associations in their mind. The results of using this approach with the works of poets and writers have been interesting [16], but an important issue remains unaddressed: What is the proper way to translate the statistics of a source text into associations between concepts in the neural model? In this paper, we consider this issue by looking at three intuitively reasonable associative models and analyzing the dynamics that emerges from them in the IDEA framework.

II. BACKGROUND AND MOTIVATION

The research in this paper is motivated ultimately by the desire to understand the neural basis of thinking – especially creative thinking – which remains poorly understood. This has led to the proposal of various computational models at the abstract and neural levels [17], [3], [18], [19], [20], [21], [22], [23]. Indeed the first such model came from the work of Hebb [24], who suggested that combinatorial activity of cell assemblies in the brain may underlie the emergence of new ideas. This has been instantiated in a recent model by Thagard and Stewart [23] and also in a comprehensive model developed in our research group [25], [26], [27], [22], [28], [29]. Consistent with experimental data, our model looks at creativity as an interaction of two processes: 1) Associatively-driven conceptual combination; and 2) Top-down modulation of the associative dynamics by contextual information and evaluative feedback [22], [29]. Recently, we have focused on studying the first component separately, giving rise to the IDEA model [30], [31], [16], [32]. This model is based on the following postulated framework:
1) Semantic knowledge is represented in the brain as an associative network between conceptual elements, which we will term an epistemic network.

2) During thinking, ideas—both familiar and novel—arise emergently as metastable activity patterns (emergent attractors) in the epistemic network as a result of the natural itinerant dynamics [15] of neural activity in the network.

3) The dynamics of the epistemic network depends fundamentally on its connectivity pattern, i.e., the pattern of association between concepts [2], [12], [33].

The main issue in instantiating the IDEA model (or any model like it) is the difficulty of constructing a realistic epistemic network. As discussed in the introduction, we have proposed to solve this problem by using large bodies of text authored by individuals to—partially and imperfectly—reconstruct the pattern of conceptual associations in the minds of the authors [16]. In this earlier work, we defined an association between two words as the probability of their co-occurrence in the same sentence. However, there are other equally plausible mathematical ways to define the magnitude of association between words [34], [35], [36].

In this paper, we study three such variants using two epistemic networks—one a “creative” network with a flatter association hierarchy, obtained from the works of a famous poet, and the other a “non-creative” network based on the memoirs of an economist.

We extracted epistemic networks from the following two corpora using the above methods:

1) Ninety-nine poems by Dylan Thomas (20th century surrealist English poetry); and
2) The Memoirs by Ludwig von Mises (20th century Austrian economist). In each case, stop words were removed using a standard list, and the remaining words were stemmed to group together variants of the same root, e.g., “cry”, “cried”, “crying”, etc. This resulted in 17,257 total word tokens, and a vocabulary of 3,602 unique words for Dylan Thomas, and M = 17,257 total word tokens/reading frames and N = 3,951 unique words for von Mises. The corpora were chosen deliberately to be of very different types—one (Thomas) considered highly creative and the other (von Mises) mainly descriptive. As shown below, the two exhibited systematic differences in their statistics.

For each given corpus $T_q$, three networks were simulated using three different definitions of associative weights. The networks had $N = 500$ neural units, corresponding to the 500 most frequent words in each corpus. The three types of associative weights, $a_{ij}^q$, between word units $v_i$ and $v_j$ were:

- **Joint (co-occurrence) probability:** $a_{ij}^q = p_{ij}^q$.
- **Covariance:** $a_{ij}^q = \frac{p_{ij}^q - p_i^q p_j^q}{\sqrt{p_i^q (1 - p_i^q) p_j^q (1 - p_j^q)}}$, with only positive values retained. This is actually the correlation coefficient, which is normalized covariance.
- **Pointwise mutual information (PMI):** $a_{ij}^q = \log \frac{p_{ij}^q}{p_i^q p_j^q}$, with only positive values retained.

The matrix of all association weights is designated as $A = [a_{ij}]$. In all three cases, the resulting networks were very sparse, with most words having no association. Figure 1 shows the log-log degree distribution for words/nodes across the networks. Both distributions show a wide range of power-law degree distribution [37] followed by a sharp cutoff. The distribution for Thomas is notably more fat-tailed than that for von Mises, suggesting that the poetic epistemic network is more “web-like” while the prose network relies more heavily on a smaller number of words—possibly domain- or genre-specific terms.

Figure 2 shows the distribution of non-zero association weights for each type of network for both corpora. It is clear that the three prescriptions lead to very different association patterns. In particular, the joint probability network is dominated by small weights with very few large weights, the covariance network has a Poisson-like asymmetric distribution of weights with a peak around 0.1 for both corpora, and the PMI network has a nearly symmetric unimodal weight.
distribution with a peak near 0.5 for Thomas and near 0.65 for von Mises. Thus, the networks are very different in their connectivity and are likely to produce distinct dynamics. However, their weights come from the same underlying corpus (in the case of each author), and are based on the same underlying statistics of word usage.

Another quantity of interest is how strongly, on average, a word in the network is associated with other words. To get the distribution of these associations, we proceed as follows. For each node $i$, the outgoing associative weights, $a_{ji}$, are sorted in descending order, giving the vector $h_i$. The mean association weight rank (MAWR) vector, $H$, for the whole network is calculated as:

$$H = \frac{1}{N} \sum_{i=1}^{N} h_i$$  \hspace{1cm} (1)

This is then normalized by dividing each component of $H$ by the value of the largest (first) component.

Figure 3 shows the plots of $H$ for all three networks and both corpora. The poetry corpus has a flatter curve in all cases compared to the prose corpus, indicating that, on average, the poet makes more strong associations with each word. We have also shown previously that this is a generic difference between poetic and prose corpora [16].

IV. NEURAL NETWORK MODEL

A. Model Description

Having characterized the extracted associative networks, we instantiated recurrent neural networks using each of them and compared the resulting dynamics of conceptual combination in these networks. The network model used is similar to the one we have used previously [30], [31], [16], [32].

The associative network has $N$ units, one for word. Each unit can be seen neurobiologically as a cell assembly that acts together in an all-or-none fashion [38]. The connection weight matrix for the network is given by $W = A/\max(a_{ij} \in A)$, with $w_{ij}$ denoting the weight from unit $j$ to unit $i$. Due to the normalization, the strongest association has a weight of 1, and the rest are scaled in proportion. As noted earlier, all negative values are set to 0.

The stimulus for a unit $i$ and time $t$ is given by:
where $x_j$ is the output of unit $j$, $\xi(t)$ is uniform white noise, and $\gamma_{\text{noise}}$ is a fixed noise gain. The division by total activity represents a shunting inhibition that smooths out the effect of fluctuation in overall activity level, and $\epsilon$ is a very small constant that precludes division by zero. The excitation of unit $i$ at time $t$ is given by:

$$y_i(t) = \alpha y_i(t - 1) + (1 - \alpha)x_i(t)$$

where $\alpha$ is an inertial parameter set to $\alpha = 0.95$ in all simulations, thus approximating continuous-time dynamics.

At any given time $t$, the $K$ most highly activated units allowed to fire, provided they are not refractory and have $y_i(t) > y_{\text{min}}$, where $y_{\text{min}} > 0$ is a fixed parameter. The output of unit $i$ is calculated as:

$$x_i(t) = f(y_i(t)) = \begin{cases} 
1, & \text{if } y_i(t) \in \{K \text{ most excited units} \} \\
0, & \text{otherwise}
\end{cases}$$

The competitive activity rule is applied with some flexibility so that a unit $i$, with $y_i(t)$ within 1% of the $K$th most excited unit is also allowed to fire, which means that though we set $K = 8$, the number of active units can vary around this value.

A crucial feature of the model that generates the itinerancy of the dynamics is the use of two types of modulation in the system:

**Refractoriness:** Each unit $i$ has a variable resource, $r_i(t)$ that determines whether it can fire. This depletes during the unit’s active period and is replenished when it is inactive as follows:

$$r_i(t) = \begin{cases} 
(1 - \lambda^-)r_i(t - 1), & \text{if active} \\
r_i(t - 1) + \lambda^+(1 - r_i(t - 1)), & \text{if inactive}
\end{cases}$$

where $\lambda^-$ is the resource depletion rate, and $\lambda^+$ is the resource recovery rate. If $r(i)$ falls below a threshold $\theta^-$, unit $i$ becomes refractory and cannot fire again until the resource recovers to a level $\theta^+$. The depletion and recovery rates for individual units are similar but not identical due to a small random variation.

**Synaptic Modulation:** It is known from experiments that synaptic modulation is a critical aspect of short-term memory[39], [40]. When activated repeatedly over short periods, synapses can temporarily become weaker [41], and then recover gradually to their nominal levels after activity ceases. This is a short-term change independent of any long-term synaptic change induced by the activity. This modulation is represented in the model as follows:

$$w_{ij}(t) = \begin{cases} 
(1 - \psi^-)w_{ij}(t - 1), & \text{if active} \\
w_{ij}(t - 1) + \psi^+ [\bar{w}_{ij} - w_{ij}], & \text{if inactive}
\end{cases}$$

where $\psi^-$ and $\psi^+$ are the synaptic decay and recovery rates, respectively, and $\bar{w}_{ij}$ is the nominal (pre-modulation) weight of the synapse.

The depletion and recovery rate parameters for both types of modulation are kept fixed for all simulations, and are set such that one “duty cycle” of activity nexts several cycles of synaptic modulation, allowing the dynamics to potentially linger in a semantic neighborhood before moving on.

### B. Model Dynamics

As a result of the competitive activity rule and the two types of modulation, the dynamics of activity in the network is an itinerant “wandering” between metastable attractors [15]. When the group of active units have strong mutual connections within the group, the pattern can keep winning the competition and persist as an attractor. However, this also depletes the activity resources of the participating units and weakens the synapses involved, eventually leading to the pattern’s destabilization. The activity then goes through a transient phase before finding another metastable attractor. This type of dynamics (though not using the same mechanisms) has been proposed by others [42], [38] as a fundamental feature of cortical computation. We hypothesize that the metastable attractors that persist for a period longer than an awareness threshold correspond to conscious conceptual combinations, or ideas, while the transient dynamics remains subconcious. Thus, the dynamics of the epistemic network represents a “train of thought” [22], [30], [31]. This is similar to the notion of “mental saccades” proposed in [43]. We set the awareness threshold to 20 time-steps based on the various modulation rates.

An important aspect of the dynamics described above is that it is emergent. The attractors – or ideas – are not embedded explicitly in the network – as, for example, in Hopfield networks [44], [45] – but become become embedded implicitly as the system is configured from the underlying texts. In particular, some of the attractors configured into the system implicitly may not correspond to any explicit combination of words in the text, but represent “recombinations” – which is exactly how novel ideas have been hypothesized to arise [1], [2], [3], [5], [4], [46], [6]. Of course, this can also be seen as “false memories”, which suggests that creativity can be seen as productive confabulation, as has been noted by others [19], [47].

### V. Simulations and Results

Networks with $N = 500$ neural units corresponding to the 500 most frequent words in each corpus were simulated for both corpora and all three weight types using the same parameter settings. Each case was run multiple times, but we use two runs for each case in this report. Each run consisted of 20,000 time-steps divided into four episodes each of 5,000 time-steps, i.e., the state of the network (but not the parameter values) were randomly re-initialized after every 5,000 time-steps. The purpose was to see if the system transitioned to very different regions of semantic space after each reset or not. The activity level $K$ was set to 8.
We focus on the following issues:

1) **Productivity**: How many ideas were produced by the system in 20,000 steps with 5,000-step blocks?

2) **Within-Trial Repetition**: What fraction of the generated ideas were unique?

3) **Cross-Trial Repetition**: What fraction of ideas produced in one simulation are repeated in the other. For this, we only compare simulations using the same association weights, though a cross-type comparison would also be interesting.

4) The frequency of word usage across ideas and trials.

The second and third observables attempt to quantify the “richness” of the semantic space defined by the associations in each case. Since the dynamics within each trial is reset after 5,000 steps, producing non-unique ideas across all 20,000 steps indicates a qualitative “ergodicity” in the dynamics. This is even clearer if ideas are repeated across trials, indicating that ideas are organized into a single semantic domain rather than several. Looking at word usage frequency shows whether ideas are being built around a few “core” words.

Figures 4-6 are *recurrence plots* showing the pairwise normalized Hamming distance between all ideas generated sequentially during two runs of each type of network from

Fig. 4: Recurrence plots for two *Thomas* networks with joint probability weights. Each $x$-$y$ pixel in a plot shows the pairwise normalized hamming distance between the ideas generated in the same run. Black indicates perfect match between the ideas on the $x$ and $y$ axes, and lighter colors progressively indicate lower match. The ideas on each axis are ordered identically in the sequence generated.

Fig. 5: Recurrence plots for two *Thomas* networks with covariance weights.

Fig. 6: Recurrence plots for two *Thomas* networks with PMI weights.

Fig. 7: Recurrence plots for two *von Mises* networks with joint probability weights.

Fig. 8: Recurrence plots for two *von Mises* networks with covariance weights.

Fig. 9: Recurrence plots for two *von Mises* networks with PMI weights.
Fig. 10: Comparison of two Thomas networks with joint probability weights. Top row: Maximum similarity of each idea in one run with any idea of the other (in normalized hamming distance). Middle row: Mean similarity of each idea in a run with all ideas of the other. Bottom row: Mean normalized word frequency of each idea in each run.

Fig. 11: Comparison of two Thomas networks with covariance weights.

Fig. 12: Comparison of two Thomas networks with PMI weights.

the Thomas corpus. Figures 7-9 show the same thing for the von Mises corpus. The plots provide a concise snapshot of how each network’s dynamics moves from idea to idea in semantic space. Several interesting observations can be made from these. First, the pattern of how the dynamics moves in semantic space is qualitatively very similar in the two runs for each case, but quite different for each type of network. In the case of Thomas, the dynamics of the joint probability network (Figure 4) repeatedly generates similar ideas in a reverberating fashion, i.e., it tends to hover in a particular region of semantic space, and often returns to the same regions. The net result is that, while many ideas are generated, the proportion of unique ideas is only 70.5%. The covariance network, in contrast, generates a large number of ideas with less repetition: a unique idea proportion of about 82% (Figure 5). Finally, the PMI network (Figure 6) generates fewer ideas but almost 94% of them are unique. Looking at the same set of figures for the von Mises corpus (Figure 7), the most noticeable thing is the qualitative similarity with the pattern seen in the Thomas dataset. As with Thomas, the joint probability network for von Mises tends to linger in and return more often to the same semantic neighborhoods, possibly reflecting the highly asymmetric weight distribution in these networks, where a few large (Figure 8) weights hold the network together. It again generates only 70.5% unique ideas. The covariance network in this case actually generates fewer ideas with somewhat greater repetition, ending with only 70.1% unique ideas. This may be because, in contrast to the Thomas network, the von Mises covariance network is more sparsely connected. Finally, as with Thomas, the von Mises PMI network generates fewer ideas but almost all of them unique (98.2%).

Next, we look at how similar two runs of the same network from different initial conditions and different random resets are in terms of the similarity between generated ideas, and the frequency with which the ideas generated by each type of network involve highly frequency words. Figures 10-15 answer these questions. The first three (Figures 10-12) consider two runs of each type of Thomas network and Figures 13-15 two runs each of the von Mises networks. The corresponding plots in all figures use the same range scales to allow direct visual comparison. Several interesting observations emerge from these plots:

Similarity of Ideas across Runs: The top row of Figure 10 shows that the joint probability network tends to repeat many of the same ideas across multiple runs, which is consistent
with its behavior of repeating ideas even within the same run. This also accounts for the relatively high levels of mean similarity with all ideas across runs (middle row, Figure 10). However, the covariance network is more interesting: While it repeats only two ideas in each direction, there is a lot of near-repetition of ideas (top row, Figure 11). However, the (near) repetitions are different from each other, so that the mean similarity between the ideas of the two runs is low (middle row, Figure 11). The PMI network repeats only three ideas across runs in one direction and only two in the other, though there are also a few near-repeats (top row, Figure 12), with low mean similarity (middle row, Figure 12). On average, for Thomas, the joint probability network repeats 31% of ideas across runs, the covariance network 2.5% and the PMI network only 5%. The results for the corresponding von Mises networks are similar, with across-run repeat percentages of 19.6%, 31.8% and 8.9%, respectively.

Involvement of High-Frequency Words: The bottom rows of Figures 10-15 show the mean word frequency (in fraction of reading frames) for each idea produced in two runs of each network for both authors. Consistently, the plots show that: 1) In each case, two runs of the same network produce qualitatively similar results; and 2) The joint probability network makes preferential use of high frequency words compared to the other two networks. The latter observation is in concordance with the fact that the joint probability network tends to repeat ideas, indicating that these repeated ideas often involve common words with high weights in the tail of the weight distribution.

Overall, the results from the comparison between different runs and network types suggests that the covariance and PMI networks tend to produce more atypical word combinations, while joint probability networks repeatedly produce fairly typical combinations. It is worth remembering that all combinations produced as ideas are coherent in each case. Other work in our research group has indicated that using covariance as the definition of association is likely to over-emphasize words that may occur at low frequencies but are more semantically relevant because they occur more reliably in combination, whereas the more frequent words emphasized by joint probability often have mainly syntactic or modal function.

VI. CONCLUSIONS

The results reported in this paper represent a preliminary but systematic study of an important issue: How should word associations be inferred from texts? We have used the neural network-based IDEA model to study this issue, but it is of broad interest beyond just the model. The inference of asso-
citations from texts is likely to be increasingly important in applications such as stylometry, document classification and idea mining. Using the best definition of associations, and understanding the relative pros and cons of each definition, will be critical for this. The present study represents a small step in this direction.

REFERENCES


