Hidden Patterns in Combined and Adaptive Knowledge Networks

Bart Kosko
University of Southern California

ABSTRACT

Uncertain causal knowledge is stored in fuzzy cognitive maps (FCMs). FCMs are fuzzy signed digraphs with feedback. The sign (+ or −) of FCM edges indicates causal increase or causal decrease. The fuzzy degree of causality is indicated by a number in [−1, 1]. FCMs learn by modifying their causal connections in sign and magnitude, structurally analogous to the way in which neural networks learn. An appropriate causal learning law for inductively inferring FCMs from time-series data is the differential Hebbian law, *which modifies causal connections by correlating time derivatives of FCM node outputs. The differential Hebbian law contrasts with Hebbian output-correlation learning laws of adaptive neural networks.*

FCM nodes represent variable phenomena or fuzzy sets. An FCM node nonlinearly transforms weighted summed inputs into numerical output, again in analogy to a model neuron. Unlike expert systems, which are feedforward search trees, FCMs are nonlinear dynamical systems. FCM resonant states are limit cycles, or time-varying patterns. An FCM limit cycle or hidden pattern *is an FCM inference. Experts construct FCMs by drawing causal pictures or digraphs. The corresponding connection matrices are used for inferencing. By additively combining augmented connection matrices, any number of FCMs can be naturally combined into a single knowledge network. The credibility w_i in [0, 1] of the ith expert is included in this learning process by multiplying the ith expert's augmented FCM connection matrix by w_i.*

Combining connection matrices is a simple type of adaptive inference. *In general, connection matrices are modified by an unsupervised learning law, such as the*

This research was supported by the Air Force Office of Scientific Research (AFOSR F49620-86-C-0070) and the Advanced Research Projects Agency of the Department of Defense under ARPA Order No. 5794.

Address Correspondence to Bart Kosko, Department of Electrical Engineering—Systems, Signal and Information Processing Institute, University of Southern California, Los Angeles, CA 90089-0272.

International Journal of Approximate Reasoning 1988; 2:377-393

52 Vanderbilt Ave., New York, NY 10017 0888-613X/88/$3.50

differential Hebbian learning law. Under special conditions, differential Hebbian dynamical systems are proved globally stable: they resonate on fixed-point attractors.

KEYWORDS: ***neural networks, fuzzy cognitive maps, dynamical systems, unsupervised learning, causal inference.***

KNOWLEDGE NETS VS. EXPERT SYSTEMS: DIGRAPHS VS. TREES

What is an expert system? A decision tree with graph search. A search-tree structure underlies all expert systems: logic trees, game trees, Markov trees, Bayesian causal trees, frame-based inheritance systems, etc. An expert system's tree structure permits graph search; otherwise inferencing algorithms would get stuck in infinite logic loops.

There are three fundamental problems with tree representations. First, *trees are feedforward structures.* They have no dynamical behavior. Feedback cannot be represented by a tree, at least not naturally, and feedback can be expected to abound in a universe everywhere connected by physical laws. Expressive power is lost.

Second, *search time increases with tree size.* The more rules or branches in an expert system tree, the longer it takes to make an inference, to enumerate a path. Path enumeration is exponentially complex. Real-time behavior is in principle impossible for large search trees.

Third, *trees do not naturally combine to yield a tree.* In general, two or more trees can only be forced-fit into a single tree. (To "open" a closed loop of length n in a graph, one of the n edges, or a subset of the edges, must be removed. But which edge or edges? Such edge removal is inherently ad hoc.) Knowledge representation accuracy is compromised. Search trees are in this sense *noncombinable.* That is why expert systems are built with very few experts.

Ideally the knowledge-acquisition process should allow each expert to build his or her own expert system. These individual knowledge bases could then be combined into a representative knowledge base. Larger expert sample sizes should produce more reliable knowledge bases. But with search trees, which are inherently noncombinable, larger sample sizes produce less reliable knowledge bases.

A directed graph is the minimal knowledge representation structure that overcomes the difficulties of search trees. In general a *fuzzy cognitive map* (FCM) (see [1–5] and [6]) is a fuzzy signed digraph with feedback. An FCM is the feedback generalization of a search tree. An FCM graphically represents uncertain causal reasoning. Its matrix representation allows causal inferences to be made as feedback associative memory recollections. FCM cycles naturally

allow feedback to be represented. Abandoning graph search, the FCM (temporal associative memory; see [7]) dynamic system immediately reverberates [7] on an inference or prediction no matter how large the FCM. An arbitrary number of weighted FCMs of arbitrary structure can naturally be combined by summing the underlying augmented connection matrices. Moreover, the strong law of large numbers ensures that as expert sample size increases, knowledge base reliability increases. We now explicate these properties.

A *simple* FCM has causal edge weights in $\{-1, 0, 1\}$. All causality is nonfuzzy. It occurs to maximal degree. In general, FCM causal edge weights are numbers in $[-1, 1]$, allowing degrees of causality to be represented. An example of a simple FCM is abstracted as Figure 1 from an article by political economist Walter Williams [8].

Searching this FCM knowledge base is not easy! This nonfuzzy signed digraph with feedback is equivalent to the connection matrix

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9
C_1	0	1	1	0	0	0	0	1	1
C_2	0	0	1	0	0	0	0	1	0
C_3	0	0	0	1	0	−1	0	1	1
C_4	0	0	0	0	0	1	1	0	−1
C_5	0	−1	−1	0	0	1	1	0	0
C_6	0	0	0	1	0	0	−1	−1	0
C_7	0	0	0	0	1	0	0	−1	0
C_8	0	0	0	0	0	0	−1	0	0
C_9	0	0	0	0	−1	0	0	1	0

where the ith row lists the connection strengths of the edges e_{ik} directed out from C_i, and the ith column lists the edges e_{ki} directed in to C_i. C_i causally increases C_j if $e_{ij} > 0$, decreases C_j if $e_{ij} < 0$, and does not affect C_j if $e_{ij} = 0$. This matrix isomorphism with an FCM allows experts to graphically represent their knowledge by drawing causal pictures and allows that knowledge to be processed in feedback associative memory fashion by operating on the underlying connection matrices. Simple signed FCMs, rather than real-valued FCMs, are easier to get from experts. Simple FCMs are also usually more reliable, because experts are more likely to agree on causal signs than on magnitudes. The FCM matrix combination scheme [4] described below allows simple signed FCMs to be combined into a nonsimple FCM that naturally represents causal magnitudes as the expert sample size increases.

FCM inference proceeds by nonlinear spreading activation. This implies [7] that an FCM inference or prediction is a reverberating limit cycle or temporal sequence of events. Each causal node C_i is a nonlinear function that transforms the path-weighted activation flowing into it into a value in [0, 1]. This nonlinear function is in general a bounded monotone increasing transformation, such as

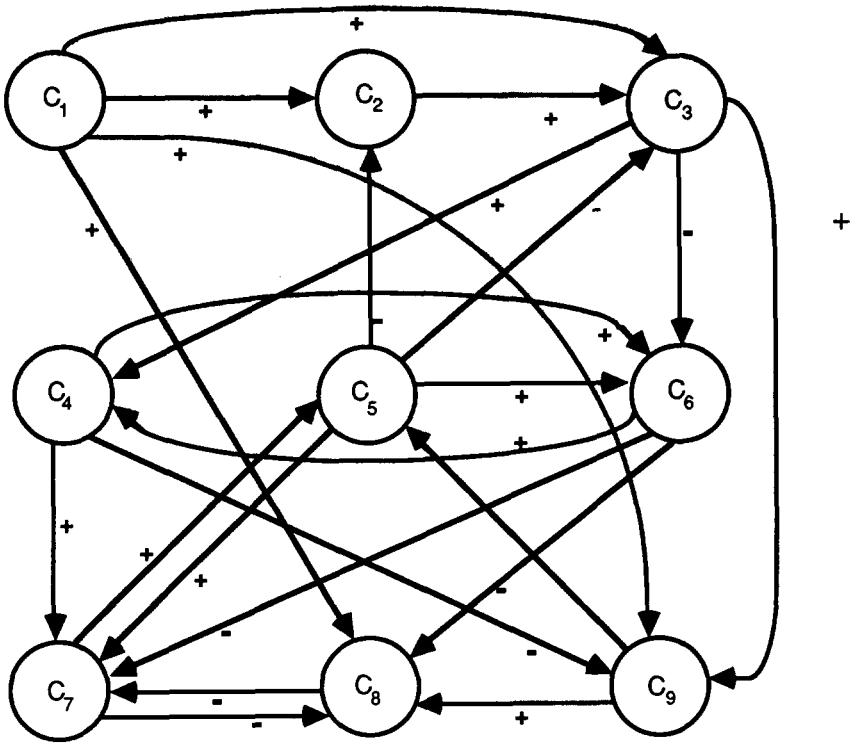

Figure 1. A nine node political FCM (after 14).

the sigmoid or S-shaped functions—for example, the logistic function $C(x) = (1 + e^{-x})^{-1}$. If each C_i is binary, as we shall assume for simplicity, the simplest nonlinear operation that turns real inputs into binary values is thresholding: $C(x) = 1$ if $x > T$, $C(x) = 0$ otherwise, for some threshold T that we shall take as zero. The threshold operation is also the limiting case of a steep sigmoid. We can write this synchronous state-transition law as

$$C_i(t+1) = \begin{cases} 1 & \text{if } C(t)E^i > 0 \\ 0 & \text{if } C(t)E^i \leq 0 \end{cases} \tag{1}$$

where $C(t) = [C_1(t), \cdots, C_n(t)]$ is the state vector of causal activation at discrete time t, and E^i is the ith column of the FCM connection matrix E. For example, a six-node FCM with input activations [6 −4 2 9 −1 −5] thresholds to [1 0 1 1 0 0]. Only C_1, C_3, and C_4 are active.

Causal flow on an FCM is easily maintained with vector-matrix operations and thresholding: $C(0) \rightarrow E \rightarrow C(1) \rightarrow E \rightarrow \cdots$. In state-transition notation, $C(t + 1) = T[C(t) E]$, where T is the vector threshold operation. This is equivalent to tracing causal flow around an FCM by inspection, as one might

inspect the feedforward flow on a search tree. Fortunately, for simple FCMs and many nonlinear FCMs, the causal flow immediately stabilizes on a limit cycle. This can be established [7] with attractor-basin arguments using the sum of "energy functions" of the form $-C_{i-1}EC_i^T$ as a limit-cycle Lyapunov function. (Arbitrary differentiable FCM models with time-varying edges can in principle resonate on chaotic attractors.) For the synchronous operation of simple FCMs, convergence is obvious, since the threshold operation is deterministic and there are 2^n possible binary states or "what-if" questions. So the FCM must converge in at most 2^n iterations. In practice it will converge after very few iterations. The first binary state of the limit cycle is the first state that is causally recalled twice.

Consider how sustained C_1 affects C_7 activity [8]. We can model this policy question by simply keeping C_1 on during the inference cycle. The initial FCM input is the state

$$S = [1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0 \quad 0]$$

Then

$$\begin{aligned} SE &= [0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1] \rightarrow S^1 \\ &= [1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1], \end{aligned}$$

which reminds us that C_1 has four outward causal arrows with positive weights and C_1 is itself exogenously kept active. At the next iteration,

$$\begin{aligned} S^1E &= [0 \quad 1 \quad 2 \quad 1 \quad -1 \quad -1 \quad -1 \quad 4 \quad 1] \rightarrow S^2 \\ &= [1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1]. \end{aligned}$$

Then

$$\begin{aligned} S^2E &= [0 \quad 1 \quad 2 \quad 1 \quad -1 \quad 0 \quad 0 \quad 4 \quad 1] \rightarrow S^2 \\ &= [1 \quad 1 \quad 1 \quad 1 \quad 0 \quad 0 \quad 0 \quad 1 \quad 1]. \end{aligned}$$

So $S^2 = \{C_1, C_2, C_3, C_4, C_8, C_9\}$ is a fixed point of the dynamical system, a degenerate limit cycle.

The resonant limit cycle S^2 is a *hidden pattern* in the causal edges E. The hidden patterns in an expert's FCM presumably correspond to the expert's characteristic set of responses to what-if questions. As with an expert's answer, the resonant hidden pattern can be tested against the available evidence, and the responsible FCM can be modified accordingly as needed.

Continuing the example, Williams claims that the absence of C_1 activity leads to, among other things, the absence of C_8 activity. The simplest way to model this is to perturb the fixed-point equilibrium by turning off C_1 in the stable state S^2. Thus we present [0 1 1 1 0 0 0 1 1] to the FCM. Quickly the FCM resonates on the limit cycle $\{C_4, C_5\} \rightarrow \{C_6, C_7\} \rightarrow \{C_4, C_5\} \rightarrow \cdots$. C_8 activity has disappeared as predicted. This two-step oscillation admits an interesting political interpretation [8].

Binary limit cycles can be directly encoded into an FCM matrix E. This follows from the correlation decoding properties of temporal associative memories [7]. Let $[A_1, A_2, \cdots, A_m, A_1]$ be a binary limit cycle. Each A_i is a binary n-vector, a point in $\{0, 1\}^m$. Convert each binary A_i into a bipolar X_i by replacing zeros with -1s. Then each X_i is a point in $\{-1, 1\}^n$. Then, using row vectors, the limit cycle can be directly encoded in an $n \times n$ matrix F by summing contiguous bipolar correlation matrices:

$$F = X_1^T X_2 + X_2^T X_3 + \cdots + X_{m-1}^T X_m + X_m^T X_1$$

(This encoding scheme allows nonzero diagonal entries f_{ii} to occur). The encoding procedure breaks down as the length m of the limit cycle (or the sum of the lengths of all encoded limit cycles) approaches the network dimension n. It also tends to fail to the extent that similar patterns do not contiguously abut similar patterns. For example, if we encode alphabet letters with large binary matrices, then NETWORK is encodable but BABY is not. The two B's in BABY are similar, but the A and Y are not.

Consider, for example, the binary limit cycle $[A_1, A_2, A_3, A_1]$ given by

$$A_1 = [1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1]$$

$$A_2 = [1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1]$$

$$A_3 = [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0]$$

Then the appropriate FCM (temporal associative memory) is found by

$$F = X_1^T X_2 + X_2^T X_3 + X_3^T X_1$$

$$\begin{bmatrix}
3 & -1 & -1 & -1 & 1 & 1 & -1 & -3 & 1 & 1 \\
-1 & -1 & 3 & -1 & 1 & 1 & 1 & 1 & 1 & -3 \\
-1 & -1 & -1 & 3 & -3 & -3 & 1 & 1 & -3 & 1 \\
-1 & 3 & -1 & -1 & 1 & 1 & -3 & 1 & 1 & 1 \\
1 & -3 & 1 & 1 & -1 & -1 & 3 & -1 & -1 & -1 \\
1 & -3 & 1 & 1 & -1 & -1 & 3 & -1 & -1 & -1 \\
1 & 1 & -3 & 1 & -1 & -1 & -1 & -1 & -1 & 3 \\
-3 & 1 & 1 & 1 & -1 & -1 & -1 & 3 & -1 & -1 \\
1 & -3 & 1 & 1 & -1 & -1 & 3 & -1 & -1 & -1 \\
1 & 1 & 1 & -3 & 3 & 3 & -1 & -1 & 3 & -1
\end{bmatrix}$$

For example

$$A_1 F = [4 \quad 4 \quad -4 \quad -4 \quad 4 \quad 4 \quad -6 \quad -4 \quad 4 \quad 4]$$

$$\rightarrow [1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1 \quad 0 \quad 0 \quad 1 \quad 1] = A_2$$

$$A_2 F = [6 \quad -10 \quad 6 \quad -2 \quad 2 \quad 2 \quad 8 \quad -6 \quad 2 \quad -6]$$

$$\rightarrow [1 \quad 0 \quad 1 \quad 0 \quad 1 \quad 1 \quad 1 \quad 0 \quad 1 \quad 0] = A_3$$

and

$$A_3F = [6 \quad -10 \quad -2 \quad 6 \quad -6 \quad -6 \quad 8 \quad -6 \quad -6 \quad 2]$$
$$\rightarrow [1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1 \quad 0 \quad 0 \quad 1] = A_1$$

completing the limit cycle. A key insight is that the limit cycle can be played backwards by passing information through the FCM matrix transpose F^T. In an FCM, passing information through F^T is a crude attempt to reverse the causal arrow of time. It produces a rough "backward chaining" inference.

COMBINING FUZZY KNOWLEDGE NETWORKS

Any set of FCMs can be naturally combined [4, 5]. Each expert can draw a different size FCM with different FCM causal concepts. There is no restriction on the number of experts or on the number of concepts. Indeed, the more experts the better. We are not restricted to the prejudices of a small number of experts. Larger sample sizes yield more reliable combined FCMs. Moreover, each expert can have a credibility weight w_i in [0, 1]. Combined weighted FCMs naturally reflect the different levels of expertise of the acquired knowledge. The hidden patterns of each combined FCM, modulated by w_i, blend into the hidden patterns of the combined FCM.

FCMs are combined by adding augmented connection matrices [4]. This is a mathematical transform trick: Transform digraphs to augmented connection matrices, combine additively, then inverse transform back to a single representative FCM fuzzy signed digraph. Suppose k-many experts each draw an FCM. The ith expert's FCM is equivalent to an $n_i \times n_i$ connection matrix E_i. These different connection matrices are not likely to be conformable for addition. In general they involve different concepts. Or do they? Suppose the second expert uses a concept C in his analysis that the first expert does not use. The first expert presumably does not believe that C is causally relevant. This means that every concept the first expert uses has zero causal connectivity to C, as if C were a phantom concept. E_i can be augmented to include C by adding a row and column of all zeros.

This procedure can be extended to augment every connection matrix to account for every concept discussed by all the experts. If the total number of distinct concepts is n, then each connection matrix E_i is augmented to an $n \times n$ matrix, perhaps quite sparse. The rows and columns of these new matrices are then permuted as needed to bring them into mutual coincidence, relabeling the row/column concepts from C_1 to C_n. This produces k-many conformable augmented connection matrices $F_1, \cdots, F_k$. The augmented connection matrices are combined by adding pointwise , $F = F_1 + \cdots + F_k$. If each expert has a credibility weight w_i in [0, 1], with $w_i = 1$ as the default weight, then the weighted combined FCM matrix F^w is found by summing multiplicatively weighted augmented connection matrices $F^w = w_1F_1 + \cdots + w_kF_k$. This is shown pictorially in Figure 2.

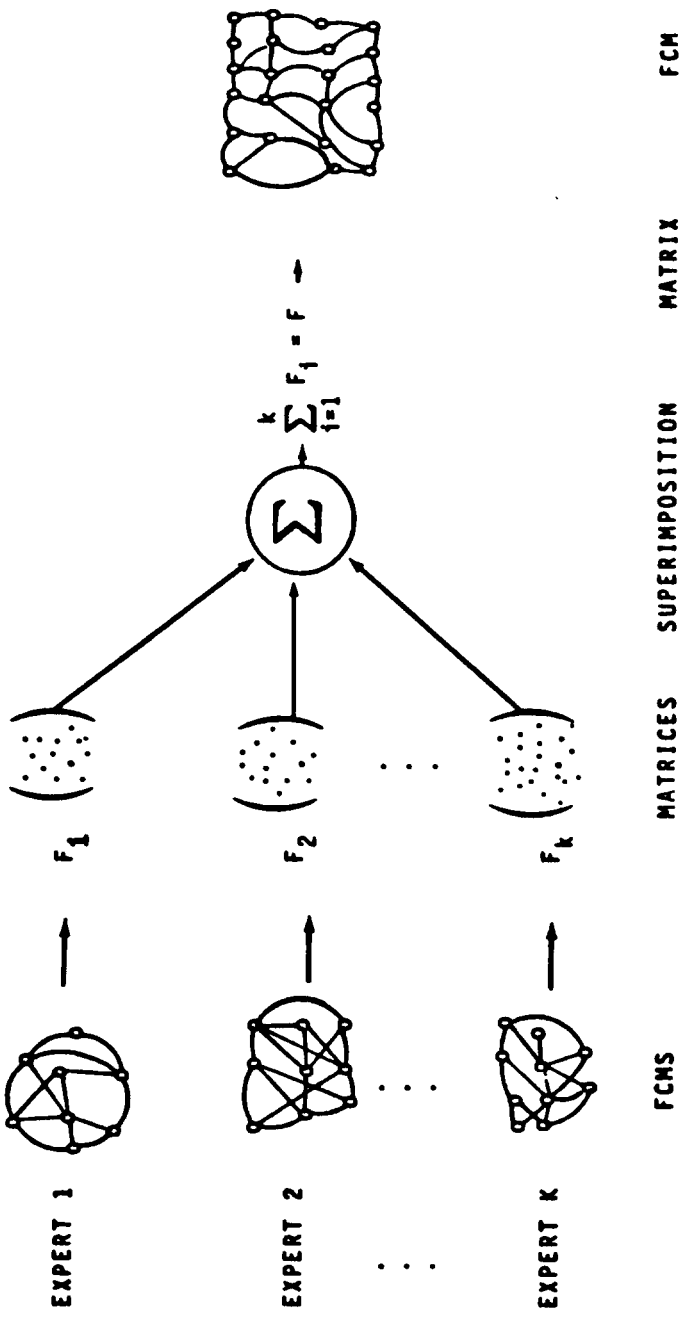
EXPERT 1
EXPERT 2
EXPERT K
F_1
F_2
F_k
$\sum_{i=1}^{k} F_i = F$
FCMS
MATRICES
SUPERIMPOSITION
MATRIX
FCM

Figure 2.

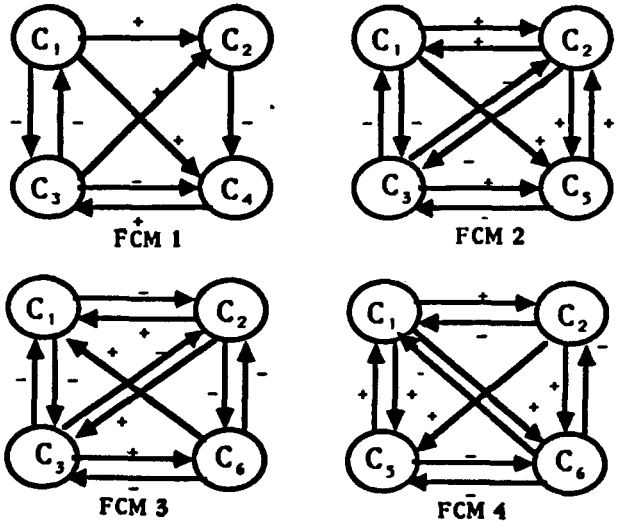

Figure 3.

The combined FCM connection matrix F naturally weights the knowledge of the experts. Suppose each expert assigns causal edge weights in $[-1, 1]$. Then if only one expert out of k asserts a particular causal connection, that connection can have maximum magnitude $1/k$. If the experts are equally weighted and half assert that $f_{ij} = 1$, say, and the other half assert that $f_{ij} = -1$, the combined weight $f_{ij} = 0$ reflects the perfect disagreement of the experts. Large sample sizes tend to produce stable connection strengths.

In general, f_{ij} is not a number in $[-1, 1]$. This causes no problem for nonlinear associative recall. In particular, it is clear that, for zero thresholds in (1), F can be replaced with $(1/k)F$ to normalize edges. For weighted experts, one can also use the normalization factor $1/w$, where w is the sum of credibility weights, $w = w_1 + \cdots + w_k$.

Consider a simple example. Suppose four unweighted experts provide the four simple FCMs shown in Figure 3. There are six distinct concept nodes. Each expert uses only four concepts. We can represent these FCMs by four 6×6 augmented connection matrices:

$$F_1 = \begin{pmatrix} 0 & 1 & -1 & 1 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 \\ -1 & 1 & 0 & -1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$F_2 = \begin{pmatrix} 0 & 1 & -1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 & 1 & 0 \\ -1 & -1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$F_3 = \begin{pmatrix} 0 & 1 & -1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & -1 \\ -1 & -1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & -1 & -1 & 0 & 0 & 0 \end{pmatrix}$$

$$F_4 = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & -1 \\ -1 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & -1 \\ 1 & -1 & 0 & 0 & -1 & 0 \end{pmatrix}$$

which combine to yield FCM connection matrix F (see also Figure 4).

$$F = \begin{pmatrix} 0 & 4 & -3 & 1 & 2 & -1 \\ 1 & 0 & 0 & -1 & 2 & 0 \\ -3 & -1 & 0 & -1 & 1 & 1 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & -1 & 0 & 0 & -1 \\ 2 & -2 & -1 & 0 & -1 & 0 \end{pmatrix}$$

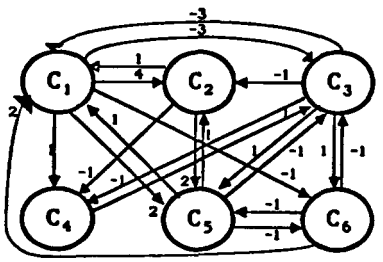

Figure 4. Combined FCM.

FCM reliability increases with expert sample size. The simplest way to prove this is to view the experts as independent identically distributed (i.i.d.) random variables [5]. Independence models individuality; identical distribution, problem domain focus; randomness, inherent error. The response of the ith expert is a number in $[-1, 1]$. Then matrix entry f_{ij} in the combined FCM matrix F is, when normalized by n, the sample mean of the ijth-distributed random variables. Thus the Kolmogorov strong law of large numbers ensures that, with probability one, as sample size increases, $(1/n)F$ approaches the underlying matrix of distribution means. In a probabilistic setting these distribution means are arguably the knowledge we seek to represent. This seems reasonable if we assume that an expert consensus is an informed average opinion.

But how reasonable are averages? We must be careful because statistical pathologies abound when density functions have thick tails and the expert random variables have infinite variances. For example, consider the one-dimensional Cauchy probability density function,

$$f_C(x) = \frac{1}{\pi(1+x^2)}$$

which is locally indistinguishable from the standard normal density function

$$f_N(x) = \frac{1}{\sqrt{2\pi}\, e^{x^2/2}}$$

If one million experts are Cauchy-distributed, then the sample mean is also Cauchy-distributed. So any one of the million expert responses predicts as well (or as badly) as the sample mean! Random variables with infinite variances in general do not obey laws of large numbers or central limit theorems. Logically speaking, they are as likely to occur in nature as are finite-variance random variables. A nonprobabilistic approach to combining knowledge is developed by Kosko in Ref. 5.

Another sample-size property of the FCM additive combination scheme is that experts need only give causal signs ($-$ or $+$), not magnitudes in $[-1, 1]$. This is extremely practical. Causal signs are more safely acquired than magnitudes, especially when the FCM is constructed from documents. It is easier to evoke dissent, indifference, or assent—$\{-1, 0, 1\}$—from a queried expert than to evoke shades of these responses. A good exercise is to construct 10 or more 4×4 matrices with values in $[-1, 1]$ and compute the matrix of sample means. Then threshold each matrix according to sign, find the new matrix of sample means of the thresholded matrices, and compare the original matrix and new matrix of sample means. The difference tends to be even less when a more equitable threshold rule is used, such as thresholding all numbers in $[-1, -1/3)$ to -1, all in $[-1/3, 1/3]$ to 0, and all in $(1/3, 1]$ to 1.

These rational sample means tend to approach the underlying real-valued

sample means provided the underlying probability density functions are sufficiently "spread out" (do not have too little variance) over $[-1, 1]$. It should, for instance, be intuitively clear that if the values are normally distributed with zero mean and unit variance, then an equal proportion of values will tend to fall in $[-1, -1/3)$ and in $(1/3, 1]$. If the normal mean is shifted to $1/2$, proportionately more values will tend to fall in $(1/3, 1]$ than in $[-1, -1/3)$, and so on.

ADAPTIVE INFERENCE THROUGH CONCOMITANT VARIATION

Inference occurs on an FCM as data-driven activation flows through FCM edges and nodes. The causal edge structure is the logical structure that represents empirical hypotheses. This is clear when causal edges are viewed as fuzzy logical implications (or conditional probabilities). *Adaptive* inference occurs when this logical structure itself is modified by data. Persistent activation of FCM concepts gradually shapes the inferential mechanisms housed by the FCM connection matrix. In the next section we prove sufficient conditions for global stability of a wide variety of adaptive inference dynamical systems.

The FCM combination technique just discussed is a simple form of adaptive inferencing. It represents a recursive "learning" procedure for gradually modifying connection strengths as new causal information becomes available in the form of weighted expert opinion. A new FCM can always be added to the current combined FCM by suitably augmenting the new FCM matrix or the combined FCM matrix, or both. Once several FCMs have been combined in a problem domain, new FCMs are less likely to contain new concepts. More important, since each new FCM matrix contains weighted elements in $[-1, 1]$ and the unnormalized strength of a combined edge is in $[-k, k]$, the addition of a new FCM matrix is not likely to significantly change the magnitude of an arbitrary combined edge and is even less likely to change its sign.

The problem of adaptive inference can be stated as a set of questions. How can the causal structure of an environment be inferred? Which things are connected to which things, and how? One answer is to ask an expert. To the extent that such information is available it should be incorporated in the inferential structure. But how did the expert obtain his knowledge? Ideally it came directly or indirectly from observation, from sensing and perhaps measuring the flux of experience. Experience enters the FCM model by additively entering the equation for a concept node's activation. If the activation of the ith node C_i is some real number x_i, then the simplest model for causal activation is equivalent to the additive short-term memory model of a neuron's activation:

$$x_i = -x_i + \sum C_j(x_j)e_{ji} + I_i \tag{2}$$

where C_j is a sigmoid function. Several state activation models other than (2) are possible. The first term in (2) is passive causal decay. Something happens if and only if something causally makes it happen. The second term is path-weighted internal feedback. The third term is external input, raw observation.

Which causal "learning" law best infers the causal structure of an environment? Ask God. A more tractable question is, Which learning laws reproduce an FCM in sign and magnitude when applied to data generated from that FCM? A good answer is: those laws that measure *changes* in the environmental parameters or, in the terminology of John Stuart Mill [9], that measure *concomitant variation*. The "causes" of a phenomenon's behavior are the variables of which that phenomenon's behavior is a function: $B = f(v_1, v_2, \cdots)$. If changes in a variable quantity Q_1 are repeatedly followed by changes in another quantity Q_2, what can an empiricist conclude but that, to some extent, Q_1 "causes" Q_2? The greater the concomitant or lagged variation in frequency and magnitude, the bolder the causal conjecture.

A *differential Hebbian* learning law [1, 3, 4, 6, 10] is the minimal unsupervised learning law for measuring change. It correlates time derivatives of node activations or of node outputs, or some mix thereof. For example,

$$\dot{e}_{ij} = -e_{ij} + \dot{C}_i \dot{C}_j = -e_{ij} + C_i' C_j' \dot{x}_i \dot{x}_j \tag{3}$$

where $C_i' = dC_i/dx_i$. Expert opinion can be added to (3). By (3), e_{ij} converges to an exponentially weighted average of correlated change. Simulations have shown that (3) and its variates tend to reproduce in sign, and often in magnitude, the FCM used to generate time-series concept node data. In contrast, the Hebbian law of neural associative learning, which simply correlates activations or outputs, if used by itself, connects all active concepts and rapidly produces a connection matrix full of spurious causal conjectures.

GLOBAL STABILITY OF DIFFERENTIAL HEBBIAN LEARNING

Most nonlinear dynamic systems are unstable. They persistently oscillate on "noise" patterns. Some resonate or equilibrate on chaotic (aperiodic) attractors. Others, like discrete binary FCMs in synchronous operation, resonate on limit cycles [7] or repeating temporal patterns. Rarest of all is equilibration to fixed points. This is global stability. As Cohen and Grossberg [11] phrase (absolute) global stability: The limits of system trajectories exist for all inputs and all choices of system parameters. All input balls rapidly roll down a local "energy" or Lyapunov minimum.

Most global stability results are for nonlearning networks, that is, for $dm_{ij}/dt = 0$, where in the tradition of neural networks we use m_{ij} instead of e_{ij} to denote the long-term memory trace or synapse between the ith and jth units. We shall also denote the bounded monotone-nondecreasing output of the ith unit by the

signal function S_i instead of the concept C_i. When only one field F_A of neurons or concept variables is involved—the *autoassociative* case—the fixed connection matrix M must be symmetric ($M = M^T$) to ensure global stability. The widest class of globally stable nonlearning, symmetric dynamical systems are those found in the Cohen-Grossberg theorem [11]. These models subsume upon change of variables many popular neural network models, including the Hopfield neural circuit, as well as Lotka–Volterra predator–prey models of population biology and Eigen–Schuster hypercycle models of macromolecular evolution. These results are extended to the $n \times p$ matrix M that interconnects the two fields F_A and F_B—the *heteroassociative* case—in the bidirectional associative memory (BAM) model [7]. Information flows from F_A to F_B in the forward direction by passing through M and from F_B to F_A in the backward direction by passing through M^T, thus symmetrizing the (in general asymmetric or rectangular) matrix M. Every matrix is globally stable in a BAM. The autoassociative case is obtained when $F_A = F_B$ and $M = M^T$.

The adaptive BAM (Kosko [7, 12–14]) extends global stability to learning, or adaptively inferring, networks. Let x_i be the activation of the ith unit in F_A and y_j the activation of the jth unit in F_B. Then, using the notation of Cohen and Grossberg [11], and thus achieving the same generality of dynamical models, every dynamical system of the form

$$\dot{x}_i = -a_i(x_i)[b_i(x_i) - \sum_j S_j(y_j)m_{ij}] \tag{4}$$

$$\dot{y}_j = -a_j(y_j)[b_j(y_j) - \sum_i S_i(x_i)m_{ij}] \tag{5}$$

where a_i is nonnegative, b_i is essentially arbitrary and S_i is bounded with $S_i' = dS_i/dx_i > 0$, and with the learning law the signal Hebb law

$$\dot{m}_{ij} = -m_{ij} + S_i(x_i)S_j(y_j), \tag{6}$$

is globally stable. Constants can be added anywhere as desired. The Cohen–Grossberg theorem [11] is the special case of the ABAM theorem when $F_A = F_B$, $M = M^T$, and the time derivative of M is identically zero. The ABAM stability theorem is proved by noting that the bounded function L is an appropriate Lyapunov function for the dynamical system (4)–(6):

$$L = -\sum_i \sum_j S_i(x_i)S_j(y_j)m_{ij} + \frac{1}{2}\sum_i \sum_j m_{ij}^2$$

$$+ \sum_i \int_0^{x_i} S_i'(z_i)b_i(z_i)\,dz_i + \sum_j \int_0^{y_j} S_j'(w_j)b_j(w_j)\,dw_j \tag{7}$$

since $\dot{L} \leq 0$ upon time differentiating L, rearranging, and eliminating the time derivatives of x_i, y_j, and m_{ij} with (4), (5), and (6), respectively. If $S_i' > 0$, $S_j' > 0$, $a_i > 0$, $a_j > 0$, then $\dot{L} = 0$ iff $\dot{x}_i = \dot{y}_j = \dot{m}_{ij} = 0$ for all i and j. The

ABAM theorem extends to any number of fields interconnected in BAM fashion. It also extends to any number of higher-order correlations [14]. Careful examination of the proof of the ABAM theorem reveals, though, that *only* the signal Hebb law (6) is compatible with the quadratic-form structure of the Lyapunov function L in (7). Other learning laws that modify m_{ij} on the basis of locally available information are not in general globally stable unless additional dynamical assumptions are made. A like remark holds for higher-order networks and the higher-order forms used in the accompanying Lyapunov functions.

Is the differential Hebbian law globally stable? In full generality we cannot expect it to be stable, since FCMs in general house nontrivial limit cycles. The analysis of these limit cycles—how the underying basins of attractions in the network state space gradually evolve as learning unfolds—is difficult. We can gain insight into the hidden-pattern dynamics of the differential Hebbian law by examining when which form of the learning law with which state model leads to fixed points, when such adapting networks globally stabilize.

A globally stable differential Hebbian model allows activations x_i and y_i to be driven by the time derivatives of the signals $S_j(y_j)$ and $S_i(x_i)$ as they flow back and forth over the pathway m_{ij}. For instance, the simple additive model (2) must be extended to the model

$$\dot{x}_i = -x_i + \sum_j C_j(x_j)e_{ji} + \sum_j \dot{C}_j(x_j)e_{ji} + I_i \tag{8}$$

The globally stable form of the learning law (3) includes a Hebbian product as well as a differential Hebbian product:

$$\dot{m}_{ij} = -m_{ij} + S_i(x_i)S_j(y_j) + \dot{S}_i(x_i)\dot{S}_j(y_j) \tag{9}$$

stated in two-field or heteroassociative notation. In general, the general dynamical systems (10) and (11) are compatible with the adaptation law (9):

$$\dot{x}_i = -a_i(x_i)[b_i(x_i) - \sum_j S_j(y_j)m_{ij} - \sum_j \dot{S}_j(y_j)m_{ij}] \tag{10}$$

$$\dot{y}_j = -a_j(y_j)[b_j(y_j) - \sum_i S_i(x_i)m_{ij} - \sum_i \dot{S}_i(x_i)m_{ij}] \tag{11}$$

To eliminate the signal derivative terms in (9)–(11), a *kinetic* energy [3] quadratic form must be added to the Lyapunov function in (7) to give the appropriate L:

$$\begin{aligned} L = &-\sum_i \sum_j S_i(x_i)S_j(y_j)m_{ij} + \frac{1}{2}\sum_i \sum_j m_{ij}^2 \\ &+ \sum_i \int_0^{x_i} S_i'(z_i)b_i(z_i)\,dz_i + \sum_j \int_0^{y_j} S_j'(w_j)b_j(w_j)\,dw_j \\ &- \sum_i \sum_j \dot{S}_i(x_i)\dot{S}_j(y_j)m_{ij} \end{aligned} \tag{12}$$

But now to prove global stability we must make a crucial assumption on signal accelerations. They must approximate signal velocities:

$$\dot{S}_i(x_i) \approx \ddot{S}_i(x_i) \qquad \text{and} \qquad \dot{S}_j(x_j) \approx \ddot{S}_j(x_j) \qquad \text{for all } i, j \tag{13}$$

(More generally, signal velocities and accelerations must agree, or tend to agree, in sign). If (13) holds, then $\dot{L} \leq 0$. And if all $a_i > 0$ and $S_i' > 0$, then $\dot{L} = 0$ iff $\dot{x}_i = \dot{y}_j = \dot{m}_{ij} = 0$ for all i and j. Hence global stability eliminates the contribution of signal velocity information. This may explain why neural network global theorists have overlooked velocity information in state and learning models: It is "observable" only as a transient phenomenon.

In general, (13) is violated if for no other reason than that sigmoid signal functions contain accelerations and decelerations. [A thresholded exponential signal function satisfies (13).] Causal concepts in FCMs tend to behave as sigmoids or inverted sigmoids, suitably scaled.

The Hebb product in (9) can, of course, be scaled so small that it makes no contribution to learning, and so can the signal derivative terms in (8), (10), and (11). Then only concomitant variation drives adaptation, and the original adaptive inference model (2) and (3) returns. As (13) is violated, limit cycles, and perhaps more complicated attractors, can occur.

References

1. Kosko, B., *Adaptive Inference,* monograph (Verac, Inc. Technical Report), in review, June 1985.
2. Kosko, B., Fuzzy cognitive maps, *Int. J. Man-Mach. Stud.* **24**, 65–75, 1986.
3. Kosko, B., Differential Hebbian learning, in *American Institute of Physics Conference Proceedings: Neural Networks for Computing* (J. S. Denker, Ed.), 277–282, 1986.
4. Kosko, B., Fuzzy associative memories, in *Fuzzy Expert Systems* (A. Kandel, Ed.), Addison-Wesley, Reading, Mass., 1986.
5. Kosko, B., Fuzzy knowledge combination, *Int. J. Intelligent Syst.* **1**, 293–320, 1986.
6. Kosko, B., and Limm, J. S., Vision as causal activation and association, *Proc. SPIE: Intelligent Robots and Computer Vision,* **579**, 104–109, 1985.
7. Kosko, B., Bidirectional associative memories, *IEEE Trans. Syst., Man, Cybern.* **18**, 49–60, Jan./Feb. 1988.
8. Williams, W. E., South Africa is changing, *San Diego Union,* Heritage Foundation Syndicate, August 1986.
9. Mill, J. S., *A System of Logic,* 1843.

10. Klopf, A. H., A drive-reinforcement model of single neuron function, in *American Institute of Physics Conference Proceedings: Neural Networks for Computing* (J. S. Denker, Ed.), 265–270, 1986.

11. Cohen, M. A., and Grossberg, S., Absolute stability of global pattern formation and parallel memory storage by competitive neural networks, *IEEE Trans. Syst. Man, Cybern.* **SMC-13**, 815–826, 1983.

12. Kosko, B., Adaptive bidirectional associative memories, *Appli. Opt.* **26**, (23), 4947–4960, 1987.

13. Kosko, B., Global stability in neural networks, in review, Sep 1987.

14. Kosko, B., Sampling adaptive bidirectional associative memories, *Proc. 21st Asilomar Conf. on Signals, Systems, and Computers,* Nov 1987.