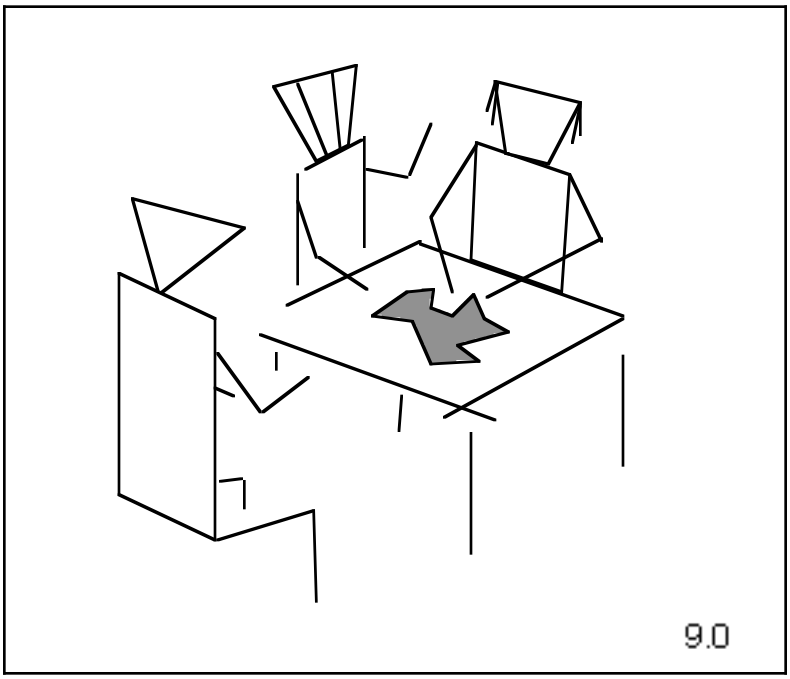
Anigraf9: Mind Games



8 Jun 04

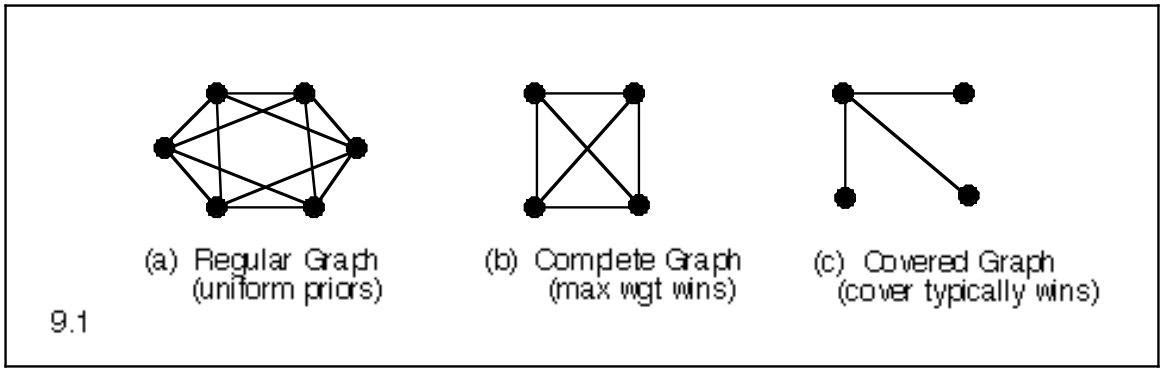
9.0 Theory of Mind

Powerful alliances between anigrafs consolidate the diverse abilities of its members into one cohesive unit. Such consolidations increase the dominance of the species. In the cognitive world of anigrafs, the greater the diversity of models in an alliance, the greater the likelihood of predicting intentions of competitive entities. But, as yet, there are no anigrafs that explicitly examine models of others to determine a strategy for optimizing social encounters. Specifically, we need anigrafs that have the ability to "read minds." If one anigraf could deduce another's internal model and hence that anigraf's intentions, then there is not only the possibility for greater clarity in communication, but also the chance for strategic manipulation – i.e. mind games. A new level of intelligence and social awareness emerges.

9.1 The Analyst

Let us focus on one anigraf "P" who is endowed with special analytical powers. Think of him as a psychiatrist. His task is to observe the behavior of other anigrafs, and to infer their mental models. P's current target is Q. We begin with the assumption that Q's choices are honest, with no attempt to mislead others. Also, let P and Q have the same numbers and types of constituent mental organisms, which carry out the same set of possible actions. So P and Q differ only in how these actions are seen as related.

After many observations of Q's behavior when different weights are placed on alternatives, P will obtain a very good estimate of the apriori probability distribution for Q's winners. If these priors are flat, then P knows that Q's model must be a regular graph, such as a ring, or a ring with chords such that all vertices have the same degree. If the alternative with the maximum weight always wins, then P knows that Q believes everything is similar to everything else, having as an internal model the complete graph Kn. If one vertex almost always wins, then that alternative very likely covers the other alternatives in the anigraf model. The priors on the actions Q takes thus provides important clues to Q's internal model.



A second piece of information gleaned from priors will be the ranking of the degree of vertices in Q's anigraf. An example will be given shortly.

But with sufficient memory capacity, P can do much better. He can keep track of the relation between his own choices and Q's choices for actions, in effect constructing an n x n table of conditional probabilities. Such a table shows the odds for Q's choices given P's choice. If P and Q have identical graphical models, the entries will be all 1's along the diagonal, with 0's everywhere else. When P and Q differ, however, the entries will create a more complex pattern. Decoding this pattern allows P to infer Q's internal graphical form -- his mental model for the domain. This is a first step in creating anigrafs who are able to view themselves and others from an overarching perspective –a superego if you will.

9.2 An Example

Let the psychiatrist's model P be a tree of four nodes {A, B, C, D} with C as the a covering vertex (Fig. 9.2, left.) Anigraf Q is known to have the same four goals, namely {A…D}, but Q's relations between these goals are unknown. P knows that Q's answers will be honest, and that Q also uses a Condorcet tally to calculate winners. Rather than asking Q questions as a typical psychiatrist, P makes note of Q's actions for sets of input weights on the four mental organisms. This is easy in this context because the same weights also apply to P's choices. After observing Q's actions to a host of such "questions", P is able to construct the following table of conditional probabilities:

		Q's winner			
		A	B	C	D
P's winner	A	1	0	0	0
	B	0	1	0	0
	C	0.1	0.1	0.8	0
	D	0	0	0	1

The entries give the probability p(Xq | Yp) of Q picking the action Xq □{A, B, C, D}, given that P picks Yp from the same set of actions, for the same set of weights.

P also knows the following:

(i) The priors on his own winners, which for our example are calculated to be (0.04, 0.04, 0.88, 0.04) for alternatives A - D respectively.

(ii) Estimates of the priors for Q, again calculated to be (0.125, 0.125, 0.71, 0.04).

The psychiatrist P can now recover Q's mental model (and vice versa if Q were to have access to P's winners.) The analysis is as follows:

1. Table 1: *Whenever P chooses A, B, or D as the winner, then Q also chooses the same winner.*

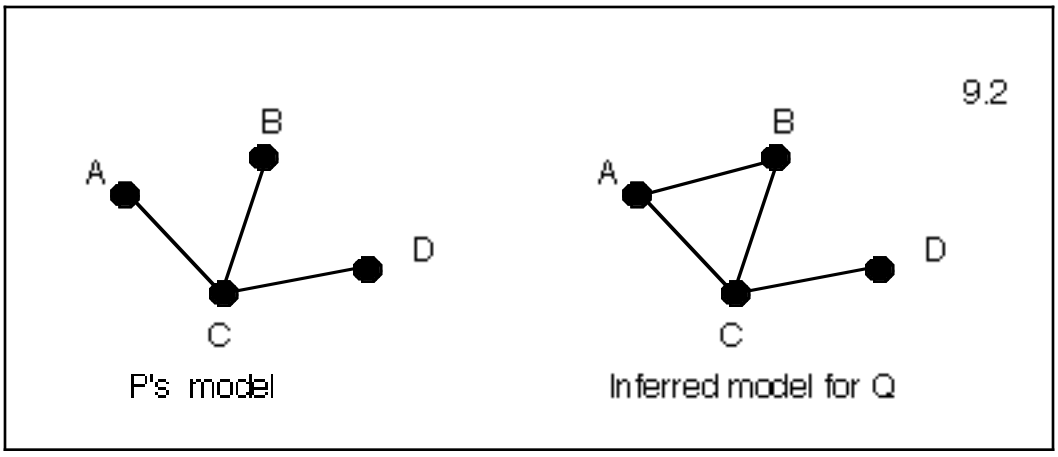
At first blush, one might infer that Q's preference orders (i.e. neighbors in Q's anigraf) for A, B and D must be identical to P's neighbors for the same alternatives. But this is not correct. All we know for sure is that SOME of the neighbors for these actions are very likely the same. In other words, if each node or agent in P's model has a smaller number of second choices, then if that alternative wins for P, the same alternative can only have an equal or

greater weight for Q. Hence P deduces that Q 's model has at least the following edges: AC, BC, DC, and remains uncertain about the remaining possible edges, namely AB, AD, BD.

2. Priors: *The odds for Q picking D as a winner are very low (0.04) and are the same as for P picking D. The odds for C in both cases are very high, namely 0.7 and 0.8.*

Guiding Hypothesis: The rank ordering of vertex degree in any graphical model is the same as the rank ordering of priors.

Because Q's priors have three levels, namely ~3/4, 1/8, 1/25, and because the maximum degree is 3, P concludes that the vertex for alternative C has degree 3, the vertices for A and B have degree 2, and the vertex for D has degree 1, the latter being the same as P's model for D. Hence Q's model has edge AB, but not edges AD or BD, as shown in Fig. 9.2.



We note in passing that prob(Aq/Cp) = prob(Bq/Cp) = ~1/12. This is consistent. Because both Aq and Bq gain strength in a vote from each other as well as from Cq, they will win more often than Ap and Bp. The difference will be absorbed by Cp.

9.3 Poker Intelligence

Most psychiatrists do not reveal their own choice for any given set of "questions". Clearly, without knowledge of P's winners, Q is at a considerable disadvantage, if indeed his answers honestly reveal choices. P knows Q's internal mental model that guides decision-making, but not vice versa. If P and Q are playing for some shared resource, P will dominate, for he can predict all of Q's choices in any situation. Adding new anigrafs R and S to the group will present no new hurdles for P, provided his memory capacity for past winners is sufficient. Thus, P has a strong advantage over others in any competitive game. Many would regard P as the most intelligent of the four because he most quickly grasps the mind sets of the others.

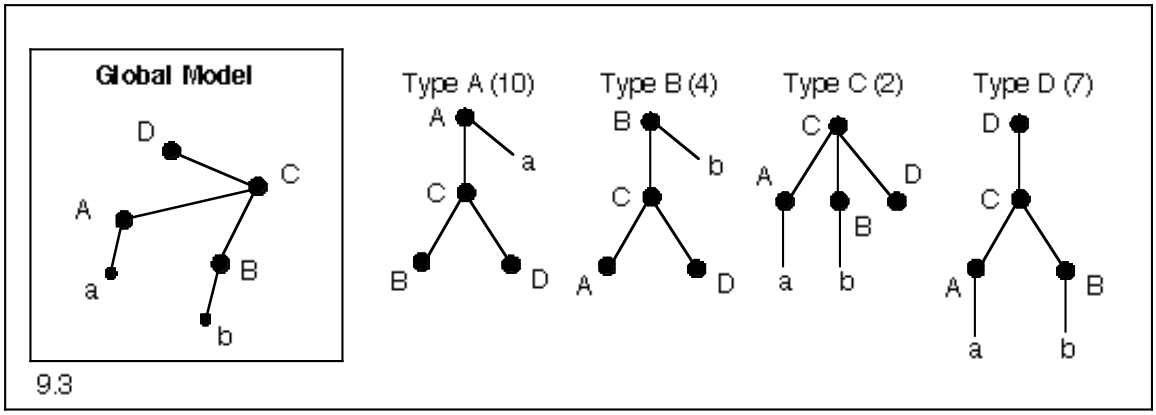
Recognizing P's advantage, what strategies should Q, R, and S take? Should answers (and consequent actions) be random? Or perhaps misleading answers are better, implying a mental model different from their true preferences. Are these subterfuges worth taking actions that are not really rewarding in the short term? What if circumstances change, and now cooperation is mutually beneficial? Will each player lack the other's trust, or at a minimum be confused by new choices inconsistent with earlier evidence? Why should P come to believe that Q is now giving honest answers? Our social setting is quite different from the case when one of the parties is an automaton, where any mismatching of models can only come from a defect or fault in the machine, not through intentional strategizing. Elsewhere, there has been much effort studying best strategies to restore cooperativity. In the face of non-cooperative behaviors, "tit-for-tat" is one popular choice (Axelrod, 1987.) However, for many situations, honesty is the best policy. But such a policy does not exclude the dominant, more intelligent anigraf from bargaining, or attempting to modifying another's mental model to advantage.

9.4 Making Deals

When anigrafs reflect on cooperative vs. competitive choices, rewards obviously play an important role. Voting behaviors are influenced, or, in anigraf world, the weights placed on the individual actions of the mental organisms will be adjusted. Classical game theory casts these rewards as gains or losses. Here, in contrast with the poker game where each player strives to win as much as possible from others, the gains and losses will apply to the

entire population of anigrafs. In other words, the rewards are dispersed throughout the society, according to votes of the anigrafs that comprise this social structure. The optimization problem is to maximize the assignment of rewards to the group members.

In this society, just like in all previous anigrafs, we assume that there is one shared global model for the similarity relations between the items to be bartered. Initially, all individual anigraf's models are consistent with this global model. Thus, one anigraf may prefer certain fruits over others, and these over various vegetables. Whereas other anigrafs may prefer vegetables to fruit or meat. But in spite of these different preferences, all have agreed at the outset that fruits are more similar to vegetables than they are to meat. The global similarity relationship among the items of interest establishes the exchange value for any individual anigraf.



9.3

Figure 9.3 illustrates using a very simple global model. There are four different types of items, or rewards {A, B, C, D}, two of which come in half quantities, designated a,b. For simplicity, all items have positive attributes, and we assume that larger amounts of items are more preferred. Each of these four types of rewards is associated with one type of anigraf. Their graphical forms are shown at the right. Note the consistency with the global model. The number of members of each type are given in the parentheses. (Members of each type are assumed to vote as a block.) Normally, our anigrafs would truncate their preference orders at level two in the graph, with all remaining alternatives being treated indifferently, regardless of the actual depth in the

ordering. Here, because of the simplicity of the global model, such truncation only affects the type D anigrafs, with lesser amounts of items A and B, namely a, b, lying at the lowest, fourth level.

With the populations of each type as shown, all anigrafs want the others to help gather their own first choice. This gathering process proceeds by having members of all types work together for the society's first choice, then the second choice of the society is gathered, etc. There may not be enough time to complete all the harvesting, so the agreed order becomes important. (Note that this process could also be cast as a dispersion game, where participants are ranked by ability, rewards are limited, and an optimal social assignment of roles is determined [Grenager et al 2002.]) There are no defections in this game.

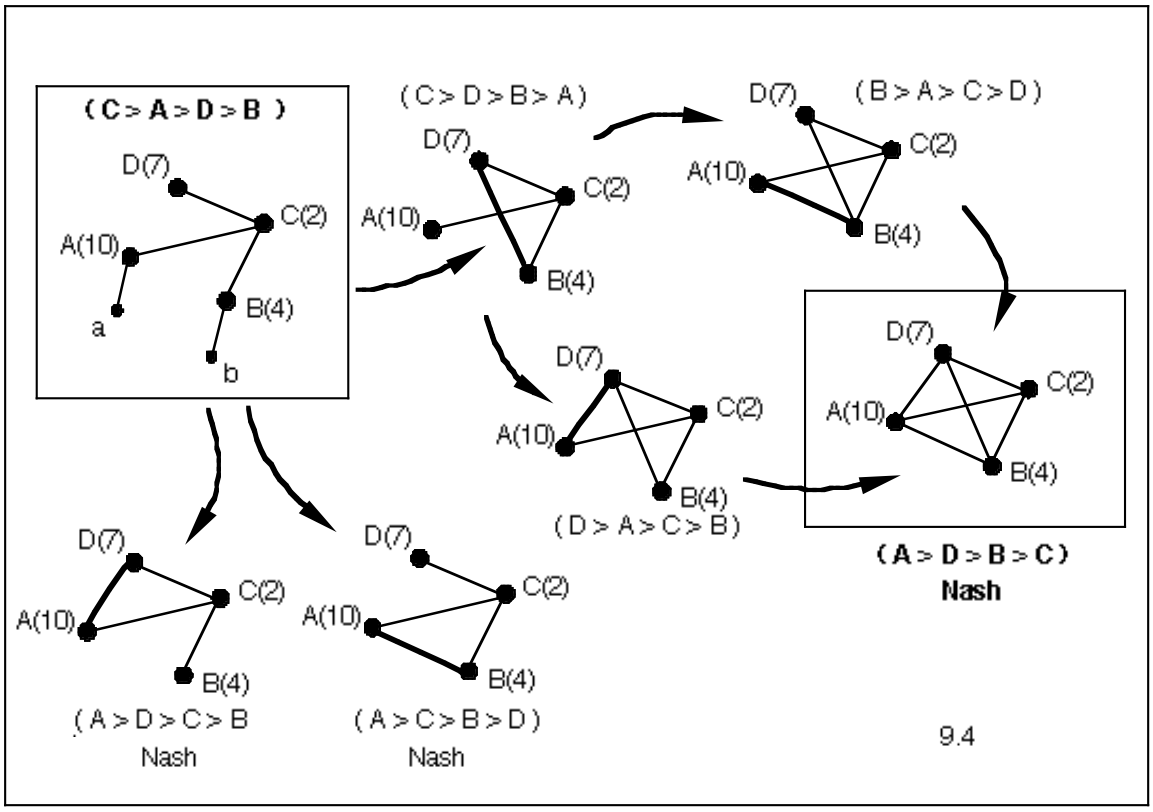
The result of our first Condorcet tally among this population is the following harvesting order: $S^*1 = (C > A > D > B)$. In other words, given the current anigraf forms and their group sizes, the greatest social benefit is when C is harvested first. Thus, although C is the choice of the smallest sub-population, the tally favors first gathering "fruit C", and then moving on to A's favored reward, then D, with B being the last to be harvested.

9.5 Manipulating Global Models

Typically, we would terminate our aggregation process after this vote. But our anigrafs have now reached a level of intelligence that allows them to read the mind-set of others, opening the door to model manipulation. What will happen to the social outcome for harvesting {A…D} if one type of anigraf could convince another type to change its mental model – i.e. to accept a small revision in the graphical relations of Fig. 9.3? Obviously this can succeed only if both types benefit in the social outcome, and if the consequent change in the global model is not inconsistent with the preference orderings of the remaining types of anigrafs.

To illustrate, let each type of anigraf pick as their leader that anigraf with the greatest ability to deduce the graphical relations of others. For example, let B* be the leader of type B anigrafs. Note that in the first vote, B anigrafs lie at the bottom of the social order. But B* observes that their

position as well as that of type D anigrafs can be improved if their leader D* can be convinced that there is now a similarity relation between items B and D. This requires adding the edge BD to the global model, but such an addition will not affect the preference orderings for A and C. Hence they will have no immediate grounds for objections. B* succeeds in her negotiation with D* and



types B and D anigrafs then modify their models and preference orderings, leaving types A and C unaffected. They now request another the tally. The new social outcome becomes S*2 = (C > D > B > A). So both B and D have improved their positions in the ranking at the expense of A! Of course, now A* realizes with the new global model, there is a benefit in establishing a relationship with either B* or D*. If A* and B* reach an agreement using the revised global model, then again the preference orderings of the remaining two anigraf types are unaffected. But now a third tally yields yet another social ordering: S*3 = (B > A > C > D). In most cases, if the deal-making continues, then everything will be deemed similar to everything else, and the final global model will be the complete graph Kn. The resultant social order is S*final =

(A > D > B > C), where the ordering is determined by the sizes of the sub-populations.

When edges can only be added but not deleted from a global model, Kn is an obvious Nash equilibrium: no two players can further negotiate to improve their positions. But other fixed points in the social order are also possible, even with the constraint of edge additions only. Consider again the original global model, but with A* and D* negotiating first. Then S*2 = (A > D > C > B). This also is a Nash equilibrium. Yet another is shown in Fig. 9.4.

9.6 Evolution and Global Orders

The above are simple examples of the many possible paths in the evolution of a global model for a set of anigraf types. Each step in the evolutionary process improves the position of two groups in the population, without (initially) altering the cognitive structures of the other members. For four alternatives, there are twenty possible anigraf types, and 30 global models, making possible many evolutionary paths even for so few alternatives. Just which sequences will be picked at any instant will rest on the equivalent of a coin flip. Each of these sequences will end in a fixed point that need not be the complete graph Kn (D. Richards, 2003). Hence there will be many sub-optimal fixed points, which are stable until a new element in the social structure is introduced (such as a link to a new population.)

In the biological world, these fixed points are Natural Modes: tight clusters of highly correlated properties. These modes are robust to change because they are associated with very successful designs. (Thompson, McMahon, Richards.) Analogously, in the cognitive world of anigrafs, the fixed points are those cognitive structures that are robust predictors and highly successful and stable models for social encounters. Clearly, the Natural Modes of biology and the Cognitive Modes of the mind are at least loosely coupled, because designs of one domain influence designs in the other. Although predicting the dynamics of the evolution of modes seems ne'er impossible, the structure of the fixed points may be accessible. In the cognitive arena of anigrafs, this structure will be the ubiquity of particular graphical forms and their relationships. Metagrafs begins this journey.