A mechanistic account of transferring structural knowledge across cognitive maps

Shirley Mark (markshir@gmail.com)
Wellcome Trust Centre for Neuroimaging,UCL. Queen Square 12
London, WC1N 3BG United Kingdom

Rani Moran (rani.moran@gmail.com)
Max Planck UCL Center for Computational Psychiatry and Aging Research. Russell Square 10-12
London, WC1B 5EH United Kingdom

Thomas Parr (thomas.parr.12@ucl.ac.uk)
Wellcome Trust Centre for Neuroimaging, UCL. Queen Square 12
London, WC1N 3BG United Kingdom

Steve Kennerley (s.kennerley@ucl.ac.uk)
Sobel department of motor neuroscience, UCL. Queen Square 33
London, WC1N 3BG United Kingdom

Tim Behrens (behrens@fmrib.ox.ac.uk)
Wellcome Centre for Integrative Neuroimaging, University of Oxford,
John Radcliffe Hospital, Oxford OX3 9DU, UK

Animals can transfer knowledge that was learnt previously and infer when this knowledge is relevant. Frequently, the relations between elements in an environment or task follow hidden underlying structure. We suggest that animals represent these underlying structures using abstract basis sets that are generalized over particularities of the current environment, such as its stimuli and size. We show that this type of representation allows inference of important task states, correct behavioural policy and the existence of unobserved routes. We further conducted two experiments in which participants learned three maps during two successive days and asked how the structural knowledge that was acquire during the first day affect participants behaviour during the second day. In line with our model, we show that participants who have a correct structural prior are able to infer the existence of unobserved routes and are able to infer appropriate behavioural policy. Therefore supporting the idea that abstract structural knowledge can be acquired and generalised across different cognitive maps.

Keywords: Inference; transfer; structural knowledge;

Introduction

Relationships between elements in different environments often follow stereotypical patterns (Kemp & Tenembaum). Social networks, for example, are organized in communities (Grivan & Newman). The cycle over the seasons and the appearance of the moon in the sky, follow a periodic pattern. Hierarchies are also abundant, for example, a management organization in a workplace. Representing such structures confers theoretical advantages in learning when encountering a new environment. In order to transfer structural knowledge from one set of sensory events to another, it should be represented in a way that is disentangled from the sensory stimuli and the particularities of the current task.

We can think of representing all tasks as graphs, each node on the graph is a particular sensory stimulus that is currently experienced, for example, observing the shape of the moon. Then, an edge between two sensory stimuli implies a transition between sensory states; a round moon will be followed by an elliptic moon. These graphs can have different structural forms (Kemp & Tenembaum) The lunar graph and the seasonal graph will all be circular; the social network graph will have a community structure; and the spatial environment will have a transition structure that respects the translational and rotational invariances of 2D space. Can humans extract such abstract information and use it to facilitate new inferences? If so, how can this knowledge be represented efficiently by the brain? Here we show that humans extract structural regularities in graph-learning tasks. When observing a new sensory environment with familiar structural form, they infer the existence of paths they have never seen, and make novel choices that are likely beneficial. In order to understand these effects, we investigate computational mechanisms in which flexible generalisation of structural knowledge is achieved using a basis set representation for each structural form. This type of representation highlights key statistical properties of the graph structure but suppress environment-specific particularities (Freguson & Mahadevan) We use the Hidden Markov

Model (HMM) framework to show how structural form can be inferred and transferred.

Results

We created a task in which agents and humans learn abstract graphs (Figure 1). The graphs belong to two different structural forms, a graph with transition matrix that obeys translational and rotational invariant symmetry (Hexagonal graph) and graphs that have underlying community structure (Figures 1). Each node on the graph corresponds to a sensory stimulus (a picture). Each edge implies that these sensory stimuli can appear directly one after the other. The agents and participants learn the graphs during repeated blocks of the task. During each block, they learn the associations between stimuli by observing pairs of connected stimuli (states). Following the learning phase, we examine their knowledge of the graph in several ways: 1) testing memory of the associations between pairs of pictures 2) navigation on the graph; starting from a source picture, participants repeatedly choose between two of the picture's neighbours until reaching the target with the aim to do so in the smallest number of steps. 3) Participants/agent are asked to report which of two pictures is closer to a target picture (without feedback). Using this task, we asked whether our participants/agent can infer the structural form of the underlying graph and transfer this knowledge to better accomplish the task. To test for transfer of structural knowledge, we conducted two behavioural experiments. In each experiment, we divided participants into two different groups. One group was exposed to graphs with the same structural form (but different images) on both days. The second group was exposed to graphs with different structural form (and images) on each day (Figure 1). This design allows us to control for all effects that are independent of the structure of the graphs as the task is independent of the structural form and its identity is not explicitly observed by the participants.

Inferring and transferring graph structure

In order to understand better the problem we consider model of this task. The task of the agent is to learn the graph. In a HMM, this task is separated into learning the distribution of sensory state associated with each graph node (termed the emission matrix, *B*), and the probability that each graph node leads to every other (termed the transition matrix, *A*). Therefore, the representation of the transition structure is naturally disentangled from the sensory information. Using these two matrices, the agent can estimate the distances (number of links) between two pictures.

If the agent has previous experience of the exact transition matrix, then the problem changes from learning a complete graph to (1) inferring which previously experienced transition matrix is relevant to the current problem, and (2) learning the emission matrix (as before). Inferring the transition matrix rather than learning it allows inference of links that were never observed. To perform structure inference, the agent estimates the probability of the sequence of observations (*O*) given each candidate transition matrix, $p(O|A_x)$ and invert these probabilities using Bayes' rule to compute $p(A_x|O)$.

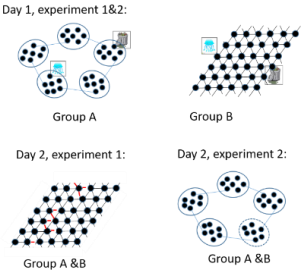

Figure 1: Graphs structures and experimental design.

Approximating the transition matrix using basis sets

Representing the transition matrix itself does not allow generalization over tasks that share the same structural form but differ in particularities such as graph size. To overcome this problem we approximate the transition matrices using basis sets (U_{st}) such that $A \sim f(U_{st} S_{st} {U_{st}}^T)$, *st* is a particular structural form and *S* is a diagonal matrix of weights (figure 2). An important feature of the basis sets is that the basis vectors can be stretched and compressed to adjust for different graph size (Freguson & Mahadevan). Each basis set can be controlled by parameters (θ) that can be inferred according to Bayes rule: $p(\theta|St, O) \propto p(O|\theta, St)p(\theta)$. The parameters can be the size of the graph, the number of clusters, the number of probable connecting nodes etc. Here, for simplicity, we assumed the number of nodes of both graphs is known and the number of nodes in each cluster should be equal.

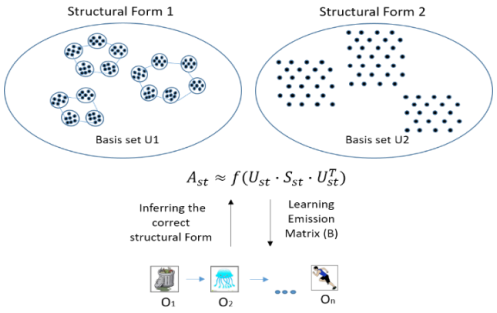

Figure 2: Inferring structure using basis sets

Basis sets definition

Using basis sets as condensed representation of structural form is inspired by the Hippocampal formation. The entorhinal cortex contains grid cells,

which have hexagonal firing patterns that are similar to the pattern of the transition matrix eigenvectors of a hexagonal graph (Hafting et al., Stachenfeld et al.). Furthermore, these cells are active in any environment with translational and rotational invariant transition structure, even if these tasks or environments are not spatial (Constantinescu et al.). We therefore hypothesised that grid cells are the basis set for all environments with this particular structural form and chose our basis set for hexagonal graph as the most (8) informative eigenvectors of the transition matrix

There are structural forms that inherently contain nodes that have different structural properties. Their fast identification can have beneficial effects on behaviour. For example, our graphs with underlying community structure contains nodes that connect two communities (connecting nodes). Identifying them is crucial for fast navigation on the graph. Thus, we chose as a basis set, vectors that assign each node to a community and vectors of connecting nodes assignment. Using this basis sets representations our model was able to infer the correct underlying structure.

Inferring unobserved trajectories

To test inference of unobserved links, the model learned the graph by sampling pairs of adjacent states while some of the links are never shown (red edges figure 1 lower, right panel). The model was still able to infer the correct structural form (figure 3, upper left panel). Further, the model had to choose between two states with the same number of observed links to the target state, the state that is closer to the target state on the complete graph. Indeed, our model is able to infer these unobserved links better than chance (Figure 3, lower left panel – here the model started with higher prior probability for hexagonal structure).

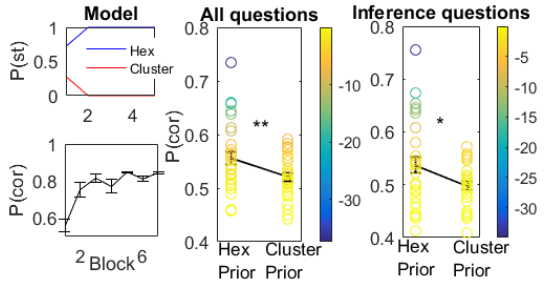

Figure 3: Inference of unobserved links (Hexagonal graph. 30 participants in each group)

People can use structural knowledge to infer unobserved trajectories

Now we ask whether humans could also use prior knowledge of the underlying graph structure to infer the existence of transitions that were never observed. We performed graph-learning experiments where participants learned three large graphs (36 nodes with degree of 6, Figures 1). We tested whether participants can infer the underlying graph structure and apply this knowledge to a new graph with new stimuli. Participants were segregated into two groups. They performed the task during two successive days (Figure 1). During the first day, one group learned two graphs with hexagonal structure while the second group learned two graphs with an underlying community structure. On that day, the graphs were learnt by observing a sequence of pictures that are taken from a random walk on the graphs.

We hypothesised that participants reach the second day with prior expectations over the underlying structural forms, as they associated the experienced graph statistics with our task. Participants who learned hexagonal graphs during the first day should expect hexagonal graph on the second day, while participants who previously learned graphs with underlying community structure should expect to learn again a graph with a community structure. We therefore asked whether participants can infer the underlying structural form during the first day and then use it as a prior knowledge during the second day. Notably, if they do, they will be able to infer the existence of transitions they have never observed (as in the model). Therefore, in this experiment, both groups of participants learnt hexagonal graph on the second day by observing pairs of adjacent pictures. Importantly, however, here the pairs were sampled randomly (i.e. neighbouring pairs were not sampled in succession) and many pairs were omitted. That is, many transitions were never explicitly observed by the subjects (depicted in figure 1 – red lines). We aimed to test whether subjects could use structural knowledge from the first day to infer the existence of these unobserved transitions.

To examine participants' ability to infer the existence of a link that was never observed explicitly, participants had to indicate which of two pictures is closer to a target picture; no feedback was given for this type of questions (more than 200 questions for each participant). Participants had to choose between two pictures with the same number of observed links to the target picture, the picture that has the smaller number of links to the target picture on the full graph. Only participants who were able to complete 'missing links' using knowledge of the underlying graph structure could answer these questions correctly. Indeed, participants who had experienced the hexagonal structure on different graphs the previous day, perform significantly better than control participants who had experienced graphs with underlying community structure (Figure 3, middle: all questions, right: 'missing links' questions only). These results indicate that similar to our model, participants extract sophisticated structural knowledge of the problem that generalises across different sensory realisations. They are able to transfer knowledge from one day to the other, and

use this knowledge to guide their decisions and infer unobserved trajectories.

Looking at figure 3, it can be seen that some participants performance is on chance level while some participants are very good at performing the task with p-value < 10^-10. We therefore concluded that while some participants have not transferred structural knowledge, the participants who succeed in doing so, do so consistently.

Using structural knowledge to set advantageous policies.

Not only can structural knowledge be used to infer unobserved transitions, it can also be used to direct advantageous policies. For example, while learning a graph with community structure, agents with no structural knowledge will spend large periods of time trapped in a single community. A simple policy of "prefer connecting nodes" overcomes this problem. In our model, the identity of these connecting nodes is easily recovered during the learning of the emission matrix (figure 4a: upper panel- structure inference, lower panel- connecting node inference).

To check whether participants are able to infer the existence of community structure and use a prior over the structural forms to inform their behaviour, we constructed the second experiment. In this experiment, participants were also segregated into two groups. As before, one group learned two hexagonal graphs and the other group learned two graphs with community structure during the first day. However now, both groups learned from random walk and navigate on a community-structured graph during the second day (Figure 1). Participants that learned graphs with underlying community structure on the first day performed better on the second-day navigation task (Figure 4c – number of steps to the target is shorter, $D_{t=0}$, is the initial distance between starting picture and the target). Importantly, however, they also learned the associations between pictures faster. While learning the associations participants determined their own learning pace by choosing when to observe the next picture. Participants that expected a graph with underlying community structure spend less time on learning each pair of pictures than participants that expected Hexagonal graph (Figure 4b). One possibility is that, instead of learning the individual pairwise associations, subjects simply inferred the community structure and assigned each node to the current community.

We further examined participants' choices during navigation to show that participants with the correct prior over the structural form infer nodes type better than participants with the wrong structural prior. During the navigation task, participants had to choose between two pictures or skip and sample a new pictures pair. We examined participants' choices during all trials in which they chose one of the pictures and one of them was a connecting node while the other was not. Participants who had the correct prior chose connecting nodes significantly more than participants who had the wrong prior ($p< 0.01$, Figure 4d). More than that, they chose connecting node more frequently, even if this choice was the wrong choice ($p< 0.01$, it took them far away from their target, Figure 4d). These results imply that participants are able to infer connecting node identity and use this knowledge during task. Please note that we are comparing two groups of participants performing the same task. Participants with a wrong prior even sample the state space more as they navigate longer on the graph. Therefore, assigning higher values for connecting nodes without identifying their identity cannot account for this effect.

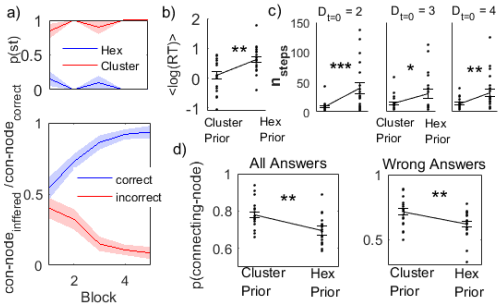

Figure 4: Policy transfer: Learning graph with community structure. (20 participants in each group)

Conclusions

Inference and transfer of structural knowledge can be achieved by basis sets representations for structural knowledge. Using structural knowledge the model and participants are able to: (1) Infer the existence of unobserved links. (2) Infer important task states. (3) Exploit behavioural policy that is tailored to a particular structural form.

References

Kemp C & Tenenbaum JB. (2008). The discovery of structural form. PNAS, 105(31) 10687-10692.

Girvan M & Newman M.E.J, (2002). Community structure in social and biological networks. *PNAS*. 99(12) 7821-7826.

Freguson I. & Mahadevan M. (2008). Proto- Transfer learning in Markov Decision Processes using spectral methods. *University of Massachusetts Amherst,Technical Report* (TR-08-23).

Hafting T, Fyhn M, Molden S, Moser MB, Moser EI. (2005). Microstructure of a spatial map in the entorhinal cortex. *Nature*.436. 801-806.

Stachenfeld K, Botvinick M, Gershman S. (2017). The hippocampus as a predictive map. *Nature Neuroscience* (20) 1643-1653.

Constantinescu AO, O'Reilly J, Behrens TEJ (2016). Organizing conceptual knowledge in humans with gridlike code. *Science* (352) 1464-1468.