Searching for rewards in graph-structured spaces

Max-Planck-Institut für Bildungsforschung
Max Planck Institute for Human Development

Charley M. Wu[1] (cwu@mpib-berlin.mpg.de), Eric Schulz[2], & Samuel J. Gershman[2]
[1]Center for Adaptive Rationality, Max Planck Institute for Human Development
[2]Department of Psychology, Harvard University

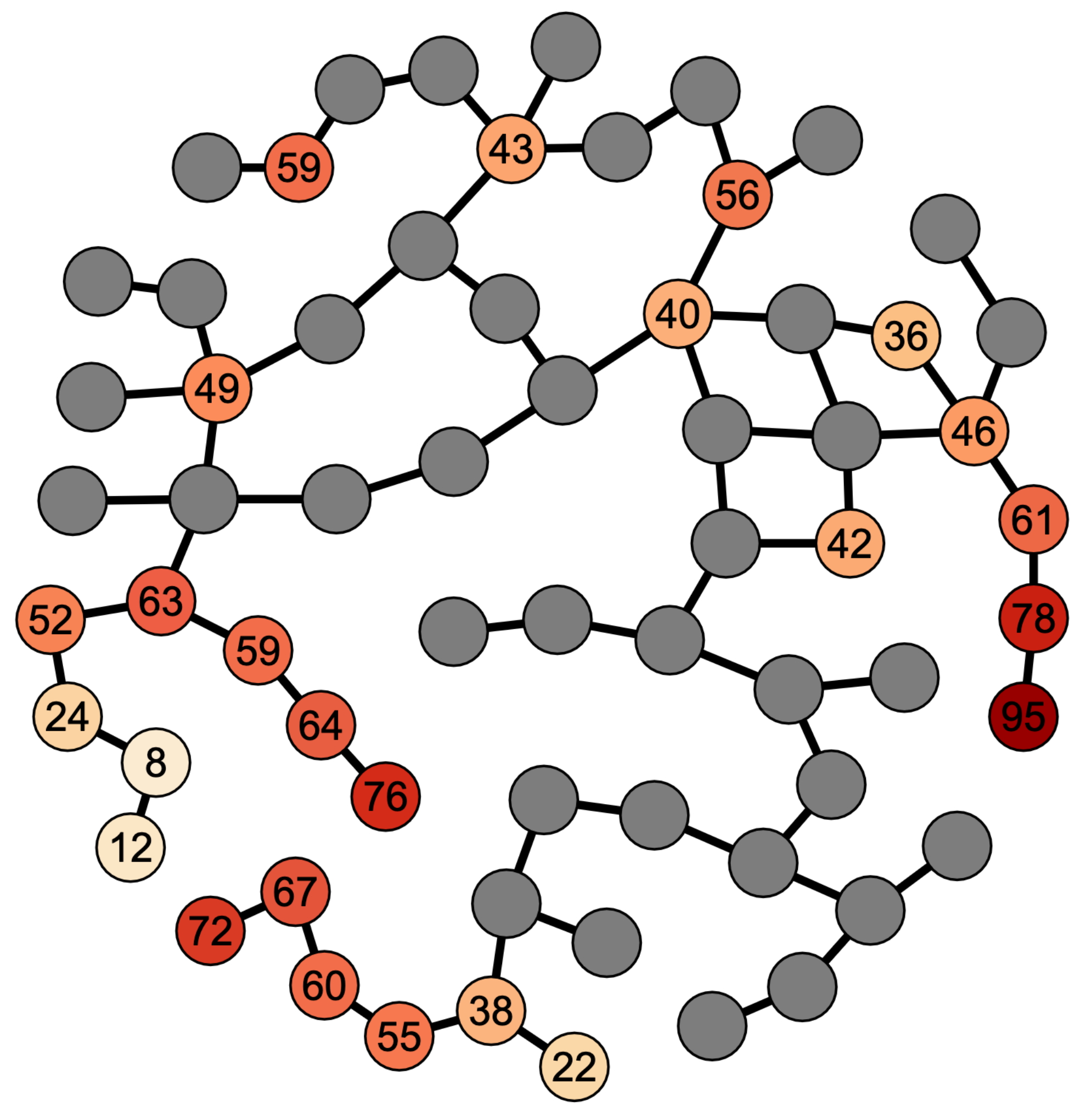
Introduction

From social networks to subway maps, many environments can be described using graph structures, where relationships are defined by connectivity rather than their singular features.

How do people generalize and explore structured spaces?

Graph Bandit Task

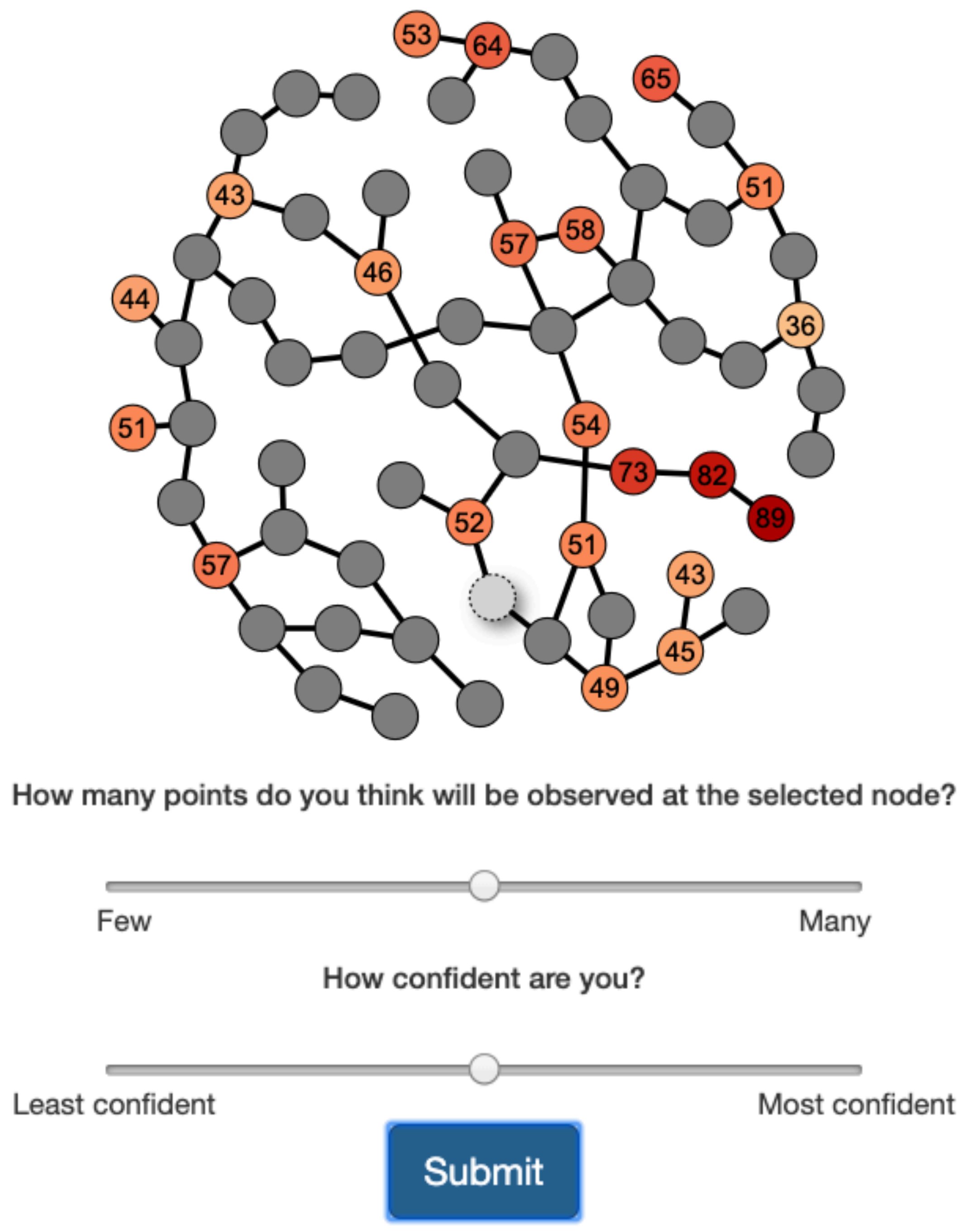
Each node is a reward-generating arm of the bandit with Gaussian noise. Expected rewards are correlated along the graph structure (i.e., highly connected nodes have similar rewards).



Participants were instructed to earn as many points as possible within a horizon of 25 clicks. Nodes begin grey, but reveal a numerical value and color (darker for higher numbers) when clicked.

Bonus Round Judgments

The 10th round was a "bonus round", where after 20 clicks we randomly selected 10 unrevealed nodes and asked participants to judge the expected number of points and rate their confidence

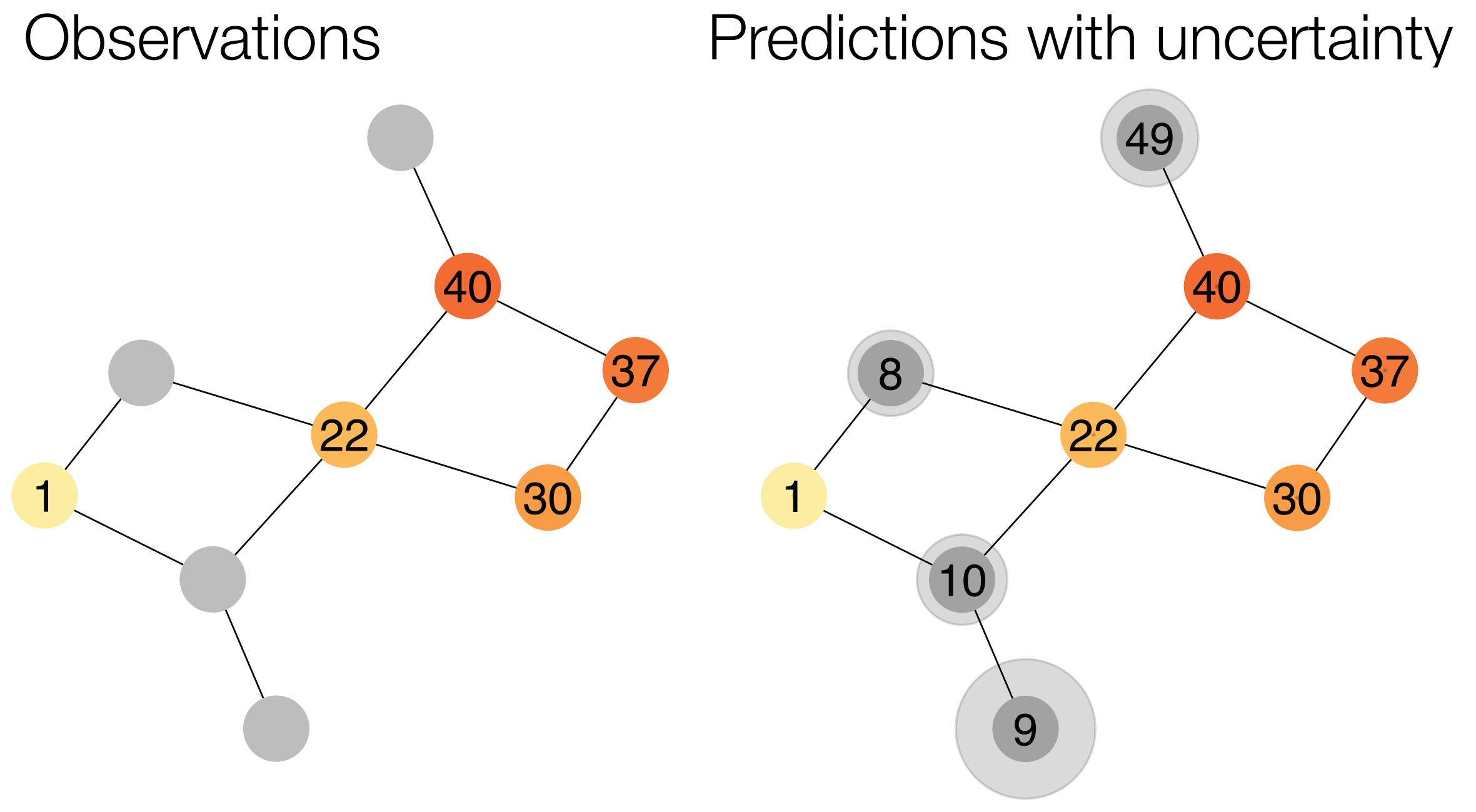


How many points do you think will be observed at the selected node?

How confident are you?

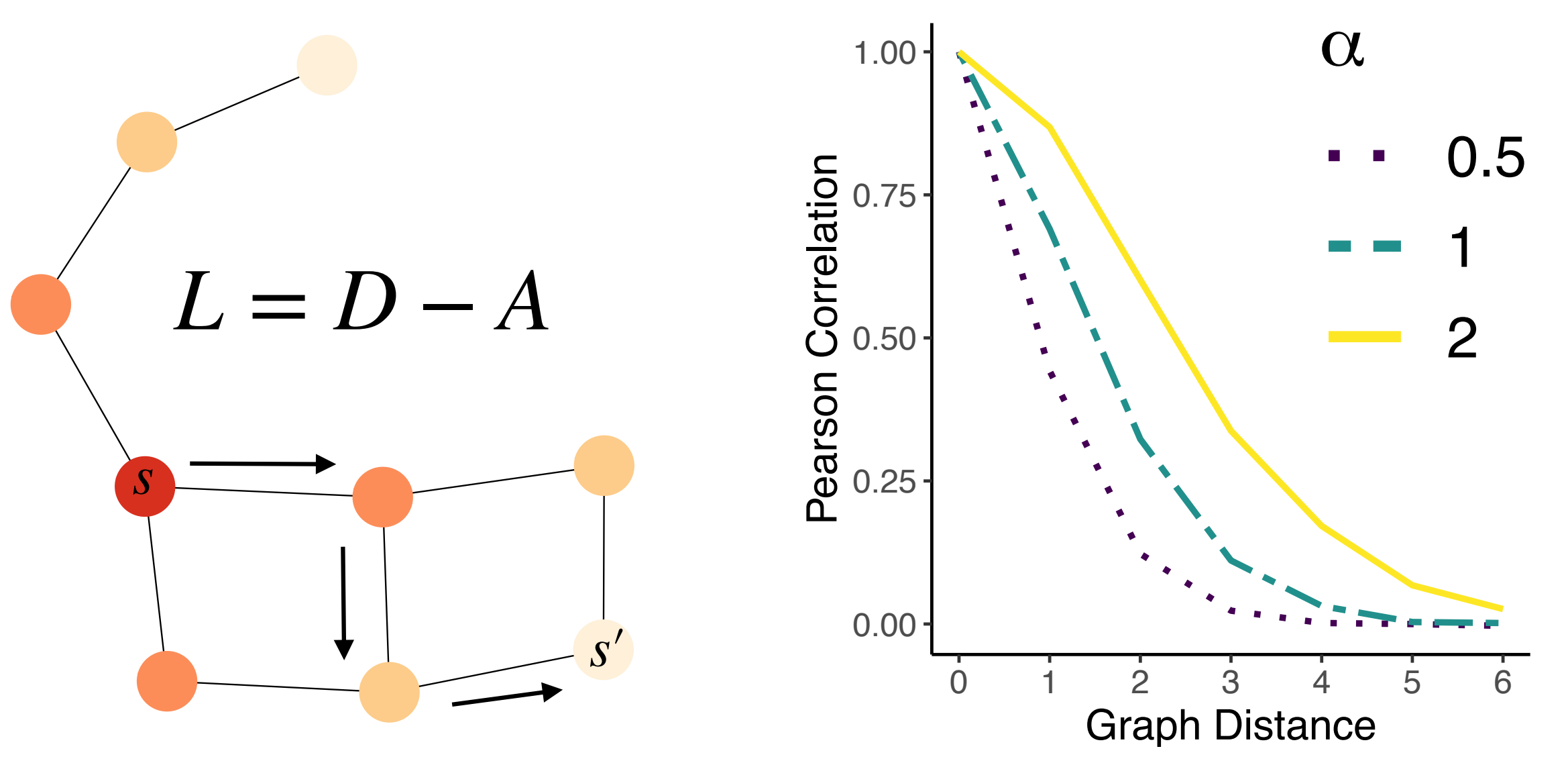
Models

The **Gaussian Process (GP)** model learns a value function across the input space, and thus generalizes about unobserved nodes.



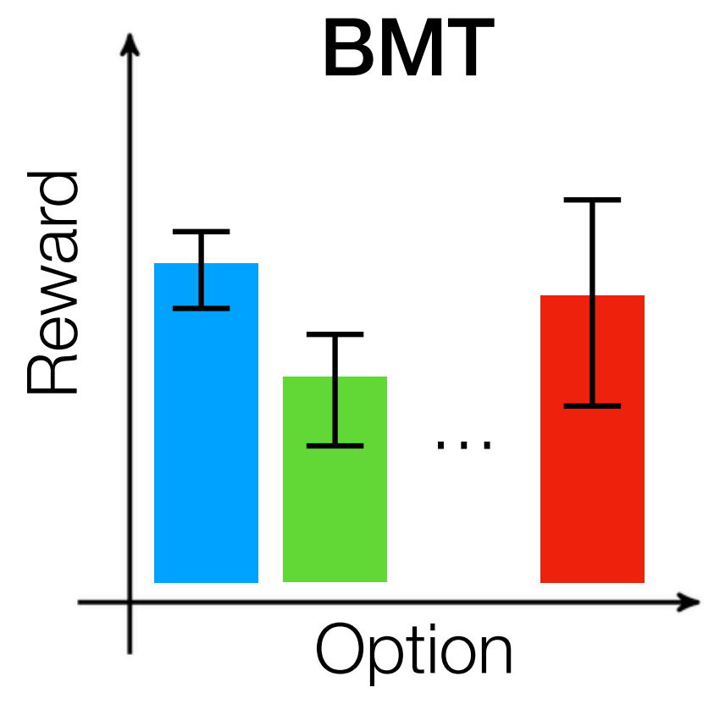
Generalization is defined using the *diffusion kernel* (DF) as a similarity metric[1], capturing the connectivity structure of the graph.

$$k_{DF}(s,s') = \exp(\alpha L)$$



The **Bayesian Mean Tracker (BMT)** is a Bayesian variant of a Rescorla Wagner model, and learns each option independently without generalization:

$$P(r_{j,t}|\mathscr{D}_{t-1}) = \mathscr{N}(m_{j,t}, v_{j,t})$$
$$m_{j,t} = m_{j,t-1} + G_{j,t}(y_t - m_{j,t-1})$$
$$v_{j,t} = [1 - G_{j,t}]v_{j,t-1}$$
$$G_{j,t} = v_{j,t-1}/(v_{j,t-1} + \theta_\epsilon^2)$$



The **Successor Representation (SR)** generalizes based on the similarity of successor states[2]:

$$V(s) = \sum_{s' \in S} M(s,s')R(s')$$

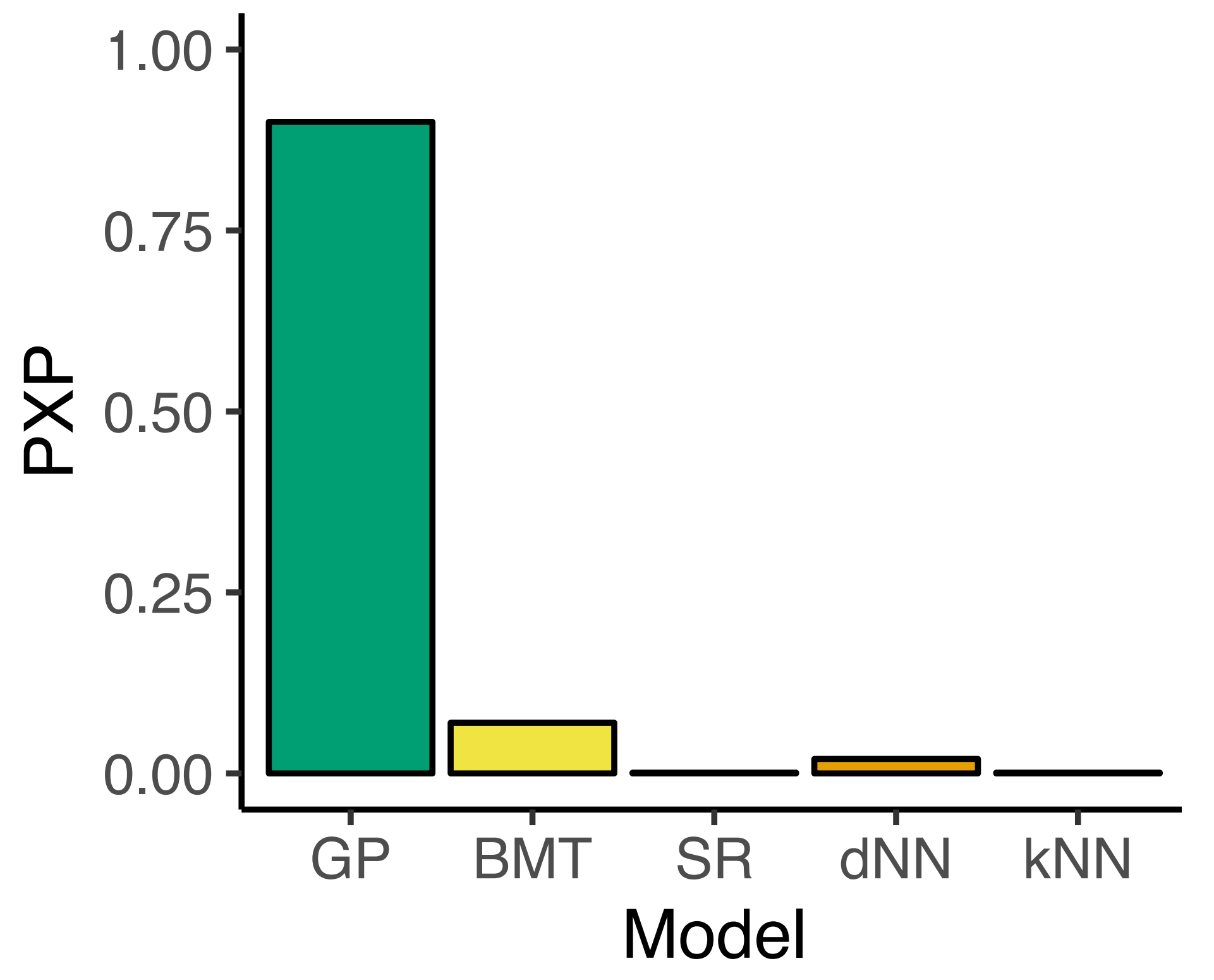
M(s,s') can be computed in closed form by assuming a random walk process:

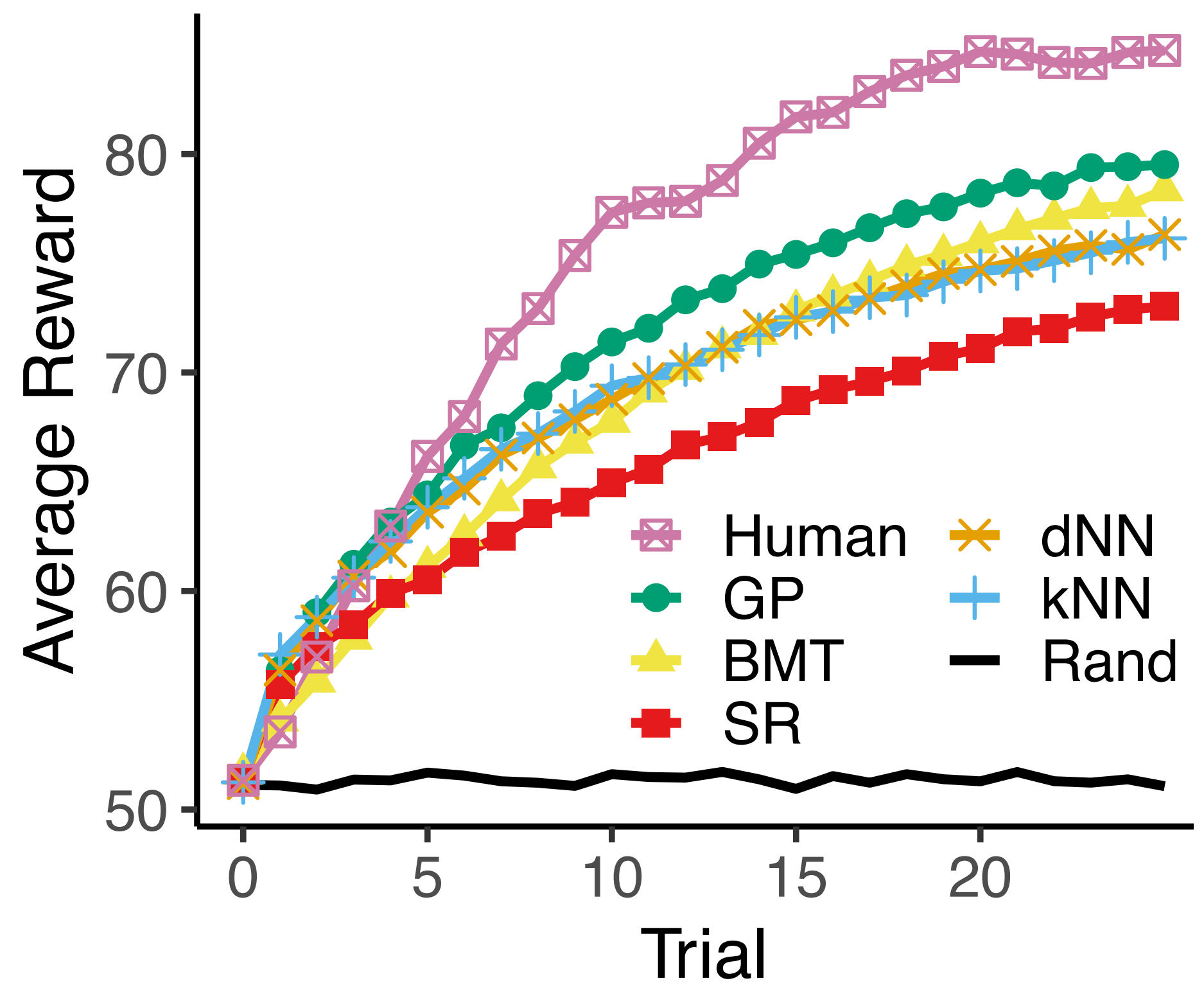
$$M(s,s') = (I - \gamma T)^{-1} \text{ where } T = I - D^{-1}L$$

While there are equivalencies to the diffusion kernel[3], the SR only produces point estimates of reward. Thus, it has no access to *directed exploration* strategies implemented by the GP (i.e., upper confidence bound sampling)

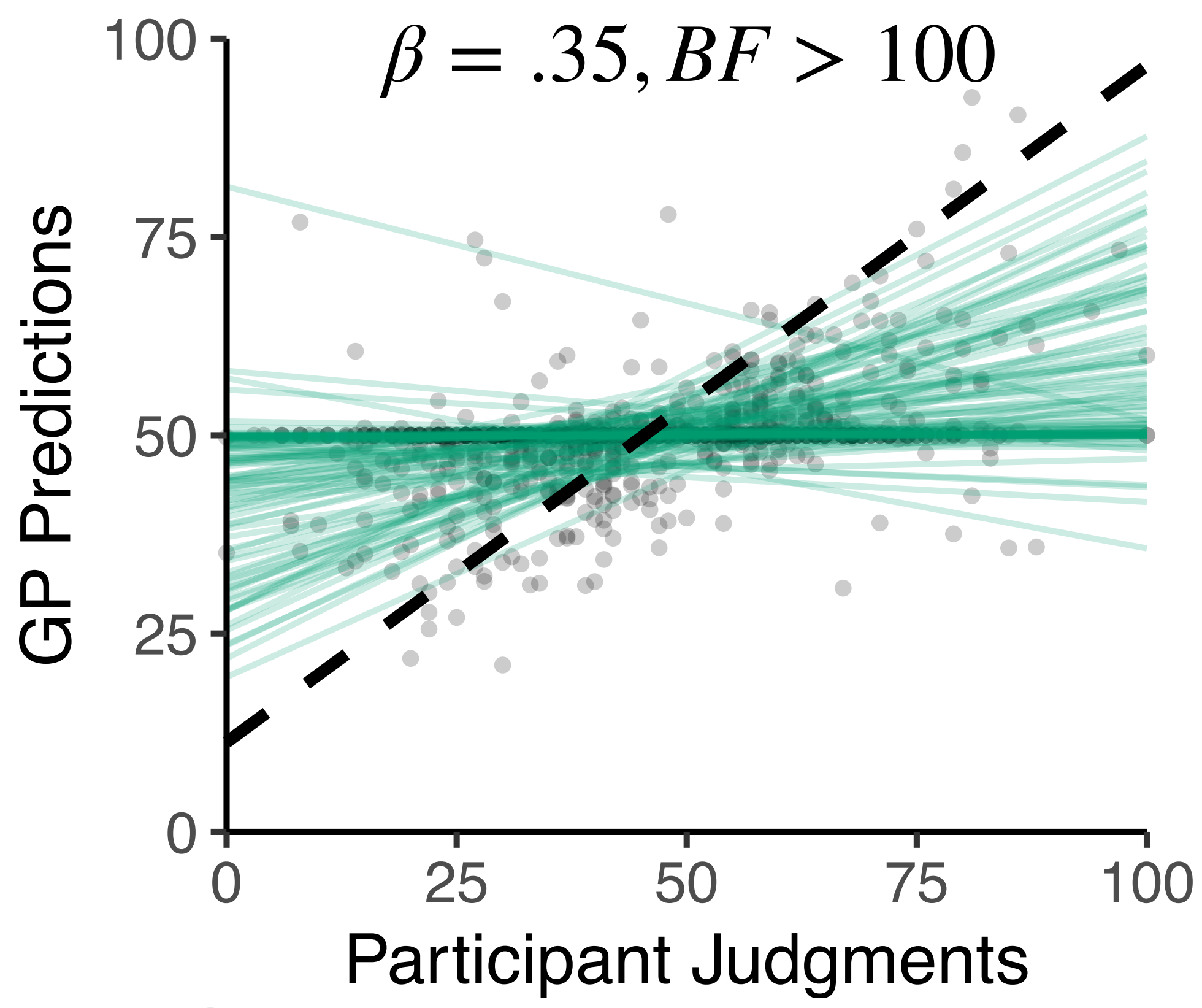
***d*-Nearest Neighbors (dNN)** and ***k*-Nearest Neighbors (kNN)** are heuristics that make predictions by averaging observed nodes within distance *d* or the *k* nearest nodes

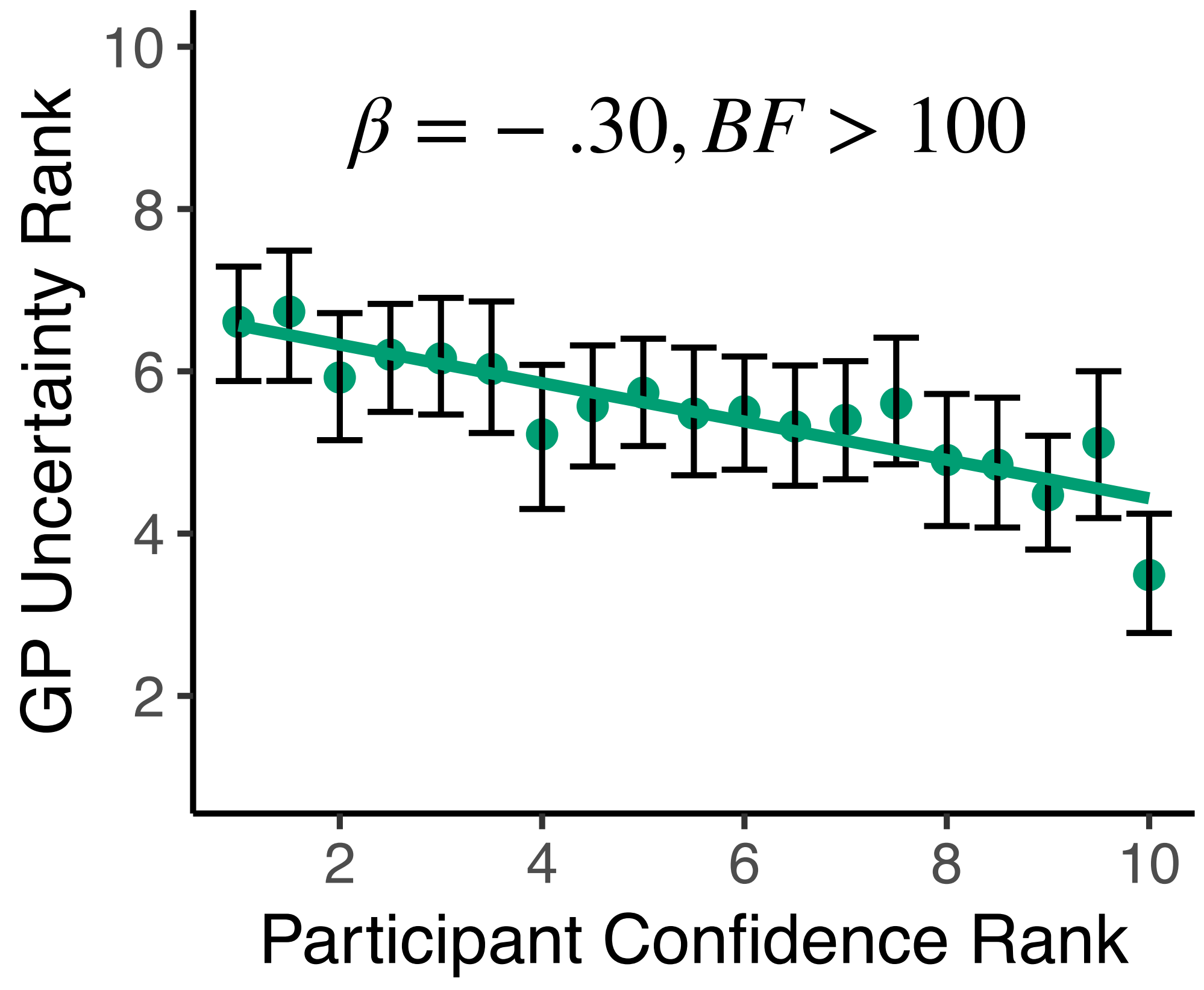
Results





Judgments & Confidence





Conclusion

We studied human learning and search in structured environments using a graph bandit task. The GP using a diffusion kernel provided the best predictions of choices, judgments, and confidence. This model unifies psychological models of function learning with the SR, building a bridge between different theories of generalization.

References:

[1] Kondor, R. I., & Lafferty, J. (2002). Diffusion kernels on graphs and other discrete input spaces. In *Proceedings of the 19th International Conference on Machine Learning* (pp. 315– 322).

[2] Dayan, P. (1993). Improving generalization for temporal difference learning: The successor representation. *Neural Computation*, 5(4), 613–624.

[3] Stachenfeld, K. L., Botvinick, M., & Gershman, S. J. (2014). Design principles of the hippocampal cognitive map. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in Neural Information Processing Systems 27* (pp. 2528–2536).