Graphs as a Fundamental Data Structure in Human Cognition

Matt Wetzel

June 16, 2020

Contents

1 Introduction

The ability to generalize knowledge about the objects and events we've experienced to the objects and events we've yet to experience is a fundamentally important aspect of human cognition. This ability is likely mediated (in part) by our capacity to group objects and events into categories for later classification and inference. A great deal of the research on human category learning has typically relied on sequential presentation of stimuli (sampled from an experimenter-defined set), often defined by very few, easily identifiable features (Kurtz, 2015). This paradigm is particularly useful for testing computational theories of cognition (Wills and Pothos, 2012), given that the simple, abstract stimuli leveraged by category learning researchers can be represented as feature vectors (a data structure that's easy to integrate into computational algorithms). However, featural information is not the only information available to learners when making classification and inference decisions. How features, stimuli, and categories relate to each other can be very useful for human learners. If computational theories of categorization aim to elucidate the role of relational information during classification and inference, then they may benefit from a type of data structure that explicitly captures relations.

1.1 Graph Theory & Applications

A graph is a type of data structure defined by a set of **nodes** and **edges** (Newman, 2003). A node can represent any kind of object or concept, while an edge represents some way in which 2 nodes are related. An edge can represent either a one-way connection from one node to another (**directed edge**), or a bidirectional connection between 2 nodes (**undirected edge**)[1]. The **degree** of a node is defined by its number of edges. The **neighborhood** of a given node is the set of nodes it connects to. A set of nodes that all share a connection is referred to as a **clique**. The **distance** between

[1]The directionality of an edge or graph depends on the type of system being represented.

two nodes is the number of edges that spans the path that connects them. The **density** of a graph typically describes total number of existing edges[2]. Graph structured reasoning typically encompasses utilizing these properties of graphs to accomplish a few key objectives: inference & classification of *nodes*, inference & classification of *edges*[3], and inference, classification, & generation of entire *graphs* (Z. Wu et al., 2020).

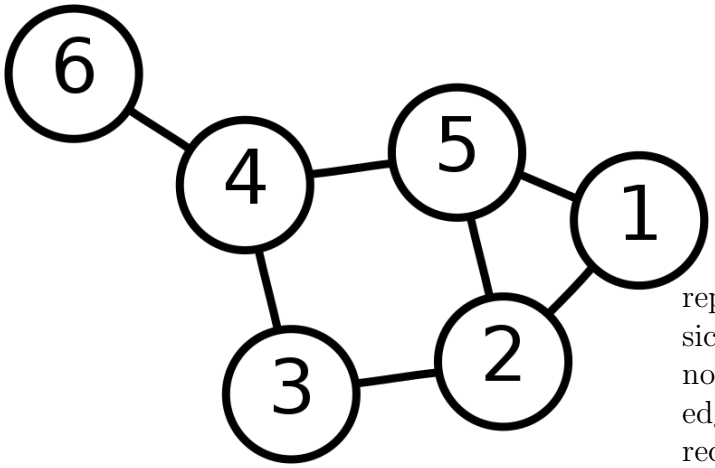

Figure 1: a visual representation of a basic graph with numbered nodes and undirected edges (image taken directly from: *wikipedia*).

Graphs have been used to describe and study a large variety of real world systems. For example, representing the Internet as a collection of websites (nodes) and links (edges) is foundational to modern search engines (Newman, 2003). A fundamental, century-old doctrine of neuroscience describes information processing in the brain as a directed graph of neurons connected via synapses (Bullock et al., 2005). Graphs are a key data structure used to describe communities of users on social media platforms, where individuals are represented as nodes and their associations/communications are represented as edges (Arya and Worring, 2018). Additionally, scientific publications can be represented as a graph where papers (nodes) connect to other publications via citations (edges). Given the vast number of structured learning domains that exist, reasoning and statistical analysis of graphs has recently seen a surge of interest in the engineering and machine learning community (Bronstein, Bruna, LeCun, Szlam, and

[2]While this terminology paragraph might seem arbitrary, the types of reasoning that can be performed on a graph are often constrained by these various properties. If humans perceive and represent their environment as graphs, then this terminology should be useful for describing human reasoning as well

[3]commonly referred to as *link completion*

Vandergheynst, 2017).

Despite the ubiquitousness of graphs in the natural world, reasoning on graphs is inherently difficult. Unlike sensory data (e.g., sight, sound) or symbolic descriptions (e.g., hand-coded feature vectors), graphical data has no guaranteed, consistent structure that can be exploited for computation (Bronstein et al., 2017; Z. Wu et al., 2020). However, the structural variability of graphs that make reasoning difficult are the very thing that make them powerful for learning and inference. For example, if I wanted to make inferences about an individual, I can leverage knowledge about that individual's friends and associates without knowing anything about the individual at all. Across many domains, the structural properties of graphs provide a powerful source of information for learning within a given domain, and even across unlearned domains as well (Graves et al., 2016).

1.2 Core Argument of This Paper

Premise 1: It's intuitively reasonable to assume that some of the objectives of a category learning system are to correctly classify and infer the properties of objects and events (Markman and Ross, 2003). Computational models of category learning typically accomplish these goals by learning some mapping between exemplars of a category and category labels, typically via exposure to one isolated item at a time (Kruschke, 1992; Kurtz, 2015; Love, Medin, and Gureckis, 2004; Nosofsky, 2011).

Premise 2: Critically, machine learning researchers and engineers have highlighted that many real world systems can be described as a graph of interconnected objects and relations. Importantly, the qualitative and statistical properties that describe graphs provide a powerful medium for classification and inference regarding objects and their relationships, as well as the construction of entire graphs themselves (Bronstein et al., 2017; Schlichtkrull et al., 2018).

Argument: If the objectives of category learning are classification and inference in real world domains, and graphical representations provide a powerful medium for

those objectives, then it's relatively straightforward to speculate that humans leverage graph structured representations for category learning. If so, it would be reasonable to assume that **(a)** humans might be sensitive to the various properties of graph structured learning domains, and **(b)** cognitive representations might be structured graphically themselves. The first section of this paper will highlight key research across different branches of psychology that address whether either of those assumptions are plausible. Then, the implications of graphical representations on two critical phenomena within the category learning literature (similarity & unidimensional bias) will be discussed. Finally, this paper will explore the argument that many aspects of cognition can be framed as computational operations on graph structured representations.

1.3 Are Humans Sensitive to Graph Structured Learning Domains?

Language learning has been a very rich testing ground for testing humans' sensitivity to graph structured learning problems, typically referred to as *Network Science* (Lynn and Bassett, 2019). Network science has provided an explanatory framework for human language at many different levels, including language acquisition via statistical learning (Lynn and Bassett, 2019; Saffran, Aslin, and Newport, 1996), word recognition & the mental lexicon (Chan and Vitevitch, 2009; Vitevitch, 2008), and speech production (Chan and Vitevitch, 2010). Language and speech provide a particularly interesting network structure: both the syllables within words and words within sentences[4] are constrained by a probabilistic, temporal order. The temporal relationship shared by the discrete units of language can be represented as a graph of transition probabilities.

In a foundational study that laid the groundwork for the utilization of network science in language, Saffran, Newport, Aslin, et al. (1996) demonstrated that young adults were sensitive to the transition probabilities between phonemes in a made-up language. With the goal of studying the mechanisms of sound segmentation in lan-

[4] and even the ideas and topics that words convey

guage, Saffran, Newport, Aslin, et al. (1996) exposed subjects long audio streams of unsegmented speech sounds that utilized 12 unique syllables clustered to form 6 unique 3-syllable words. Because the audio stream was unsegmented, there were no pauses to indicate when one word stops or another began. Critically, the probability that one syllable would transition from another were embedded in the 6 words used in the audio stream, such that within-word syllable transitions were more likely than between-word syllable transitions. At test, subjects decided which of two possible word choices belonged to the made-up language. The key finding was that subjects were able to recognize the original words from the language when compared to words with 3-syllable combinations that were statistically unlikely (i.e., "nonword" foils with average syllable transition probabilities of 0). In addition, subjects could even recognize the difference between original words and pseudo-word foils (where only one syllable transition was statistically unlikely), albeit with worse performance relative to nonwords. Even further, Saffran, Aslin, and Newport (1996) replicated this phenomena in 8-month old infants (using a novelty-preference paradigm for assessing sensitivity to statistically likely or unlikely syllable combinations)[5].

Saffran & colleagues' work demonstrated that humans are sensitive to the transition probabilities of syllables in language, hypothesizing that humans can leverage these statistics (among other cues) to segment otherwise continuous speech signals into discrete linguistic units (e.g., words)[6]. Similar findings have been extended with visual presentation of shapes and scenes as well (Brady and Oliva, 2008; Turk-Browne, Jungé, and Scholl, 2005). While the notable result from these demonstrations is that humans leverage transition probabilities during learning, the serial presentation paradigm itself has a very interesting property: the collective body of stimuli and their transitions form the basis of a graph. Interestingly, a host of recent findings suggest that humans

[5]Under the assumption that infants spend longer time listening to novel stimuli, Saffran, Aslin, and Newport (1996) found that infants spent more time "listening" to nonwords than real words in the made-up language

[6]the ability to segment information from a continuous medium into discrete signals might also relate more generally to human categorization as well

are not only sensitive to transition probabilities of stimulus presentations, but also sensitive the structural properties of transition graphs as well (Kahn, Karuza, Vettel, and Bassett, 2018; Karuza, Kahn, and Bassett, 2019; Karuza, Kahn, Thompson-Schill, and Bassett, 2017; Lynn and Bassett, 2019).

Kahn et al. (2018) provide a notable demonstration of humans' sensitivity to the structure of transition graphs. In their experiments, subjects were instructed to press sequences of keys on a keyboard. The task was self-paced, and subjects were given a signal on each trial regarding which key to press next. The probability of one key transitioning to another was defined by a particularly structured graph representing keys as nodes and transitions as edges. Critically, subjects were trained on 3 different types of graphs: a modular graph[7], a lattice[8], and a random graph with no modularity or structure. One critical finding was that the modular graph was much easier to learn than the lattice or random graphs (which themselves did not differ in learnability). Additionally, Kahn et al. (2018) found that subjects' reaction times were much slower when crossing edges between modules[9] in the modular graph structure (referred to as a *cross-cluster surprise effect*). This is interesting given that each edge had an equally likely transition probability; Kahn et al. (2018) attribute these distinctly slower reaction times as evidence of subjects' sensitivity to higher-order structure in the transition graph.

In a similar demonstration, Karuza et al. (2019) found slower reaction times at cross-cluster boundaries across different graphs that varied in their community structure. Subjects were shown sequences of visual stimuli[10], and asked to identify whether each stimulus was presented at it's normal orientation (which was established via a prior training phase). Importantly, the order of stimulus presentation was defined by the graph representing each visual stimulus as a node and the transition to a new stim-

[7]Interestingly, many natural graphs in the real world have a modular structure (Kahn et al., 2018).
[8]A lattice is a structured graph with no modularity.
[9]or, densely connected clusters
[10]The stimuli used were images of realistic looking 3D objects that were difficult to identify (Horst and Hout, 2016).

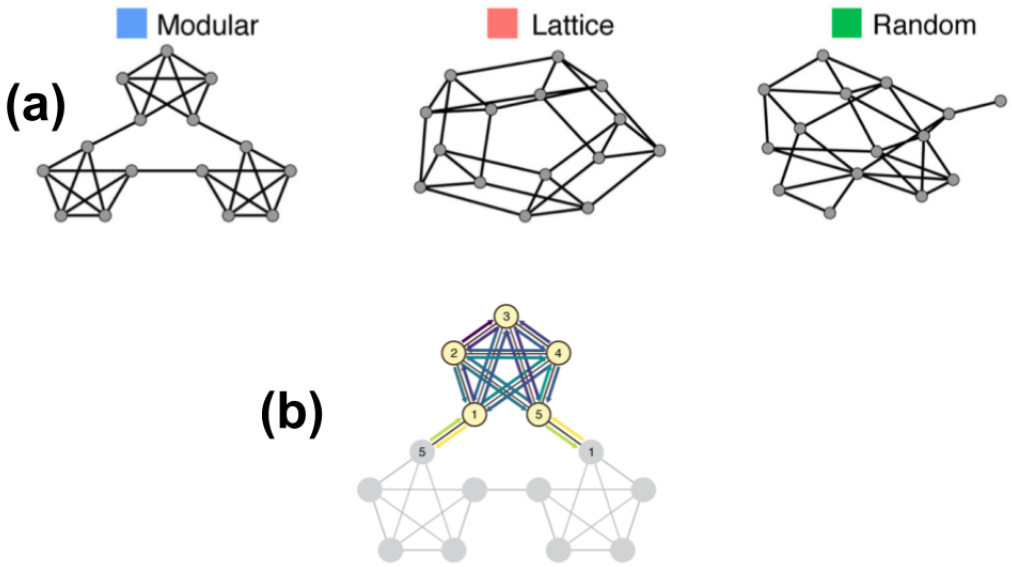

Figure 2: **(a)** Graph structures used in Kahn, Karuza, Vettel, and Bassett (2018). Note the three distinct clusters in the modular graph. **(b)** Reaction times for different key transitions (edges) in the modular graph (yellow indicating longer RTs). Taken directly from *Kahn, Karuza, Vettel, and Bassett (2018)*.

ulus as an edge. Karuza et al. (2019) replicated the cross-cluster surprisal effect in a variety of graphs with different numbers of local communities[11]. This work highlights that humans are sensitive to boundary conditions in a variety of graph structured learning problems, across multiple cognitive domains. In a similar preparation, Karuza et al. (2017) found that the particular *path* used to sample nodes and their transitions mediated subjects' sensitivity to cluster boundaries, clarifying that subjects knowledge of network structure reflects the way in which the network is traversed.

These demonstrations are compelling evidence that humans are sensitive to both local and global properties of network structured learning domains. However, these preparations all utilize a very particular type of experimenter-defined graph where edges are defined by stimulus transitions (Kahn et al., 2018; Karuza et al., 2019; Karuza et al., 2017). More evidence is needed to support whether humans are sensitive to complex learning domains in nature, where the number of nodes and edges are vast and the structural statistics are very complex. Phonological relationships between English

[11] ranging from 2-6 communities each

words provide an interesting real-world domain that has been leveraged to explore humans' sensitivity to highly complex networks in nature (Chan and Vitevitch, 2009, 2010; Vitevitch, 2008; Vitevitch, Chan, and Roodenrys, 2012).

An average English-speaking adult is estimated to have an awareness of around 17,000 words (Goulden, Nation, and Read, 1990). In a recent analysis, Vitevitch (2008) attempted to elucidate the undirected graphical structure that describes the phonological similarity between words[12]. In addition to other descriptive network statistics like average path length and degree distribution[13], Vitevitch (2008) identified that the global *clustering coefficient* of the phonological graph of English words is much higher than what would be expected from a randomly connected graph. The global *clustering coefficient* describes the degree to which a typical node's neighbors are also connected to each other[14].

Critically, the clustering coefficient of a single node has a meaningful influence on lexical processing. Chan and Vitevitch (2009) found that vocally-presented words with a low clustering coefficient were much easier to identify through background noise than words with a high clustering coefficient[15]. Utilizing a picture-naming task, Chan and Vitevitch (2010) found that subjects took longer to produce words with high clustering coefficients. Additionally, Chan and Vitevitch (2010) reanalyzed data from a prior study of speech errors (Fay and Cutler, 1977) and found that errors were more likely to occur for words with high clustering coefficients. Even further, Vitevitch et al. (2012) found that words with high clustering coefficients were more accurately remembered in a recognition task, and led to higher false-alarm rates in a false-memory task. Together, these results suggest that phonological network structure of the English language has important consequences on speech recognition, speech production, and mem-

[12] A given word is *phonological neighbor* to another word if it can subsume the form of that word by the addition or removal of a single phoneme; phonological similarity has been implicated in a number of phenomena in the language learning literature (Vitevitch, 2008)

[13] i.e., average neighborhood size of each node

[14] Interestingly, graphs with a high clustering coefficient are empirically easier to navigate (Porter, 2012), and have many other useful properties (Vitevitch, 2008)

[15] Chan and Vitevitch (2009) hypothesize that this confusability might arise from within-cluster interference during the spreading-activation of representations in the mental lexicon

ory.

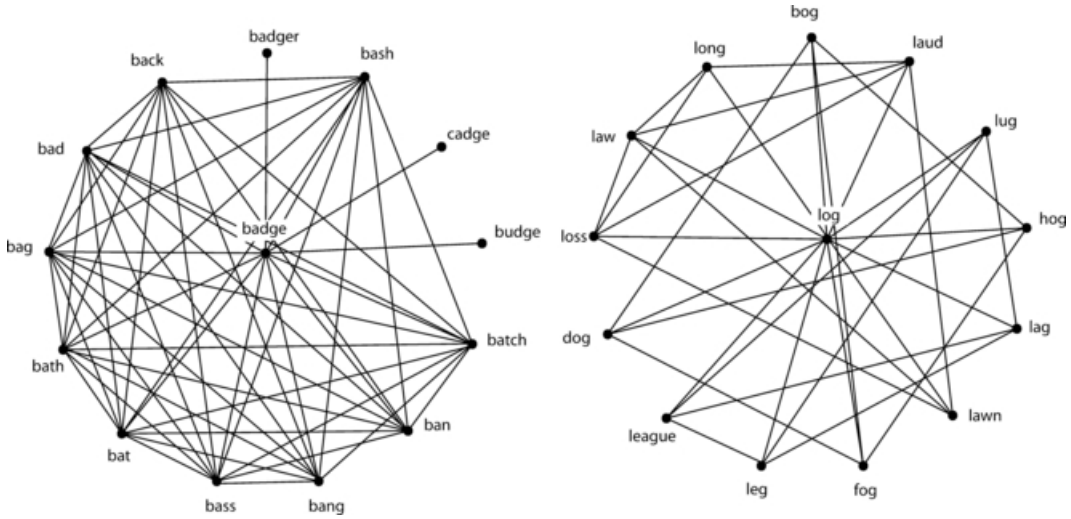

Figure 3: Comparison of the neighborhood structure of a word with a high clustering coefficient (left) and a low clustering coefficient (right); Taken directly from *Chan and Vitevitch (2010)*.

While there is an increasing amount of evidence that humans are sensitive to the network structure present in various learning domains (Chan and Vitevitch, 2009, 2010; Kahn et al., 2018; Karuza et al., 2019; Karuza et al., 2017; Saffran, Newport, Aslin, et al., 1996; Vitevitch et al., 2012), these studies alone don't directly address whether cognitive representations embody network structures themselves (though they are certainly suggestive). The next section will explore the psychological plausibility that cognitive representations leverage graphical structure.

1.4 Are Graphical Representations Psychologically Plausible?

The idea that cognitive representations leverage graphical structure dates back to the well-known *spreading activation* theory (Collins and Loftus, 1975; Collins, Quillian, et al., 1969). Building off of Quillian (1967)'s symbolic framework of representing semantic meaning and inheritance in computers, Collins, Quillian, et al. (1969) proposed that human semantic memory is structured as a conceptual hierarchy[16], where superordinate categories branch into subordinate categories (see figure 1.4). In their seminal

[16]Note that a hierarchy is a type of acyclic graph known as a *tree*

demonstration, Collins, Quillian, et al. (1969) had subjects determine the truthfulness of certain assertions about the properties of objects/concepts; the properties of each assertion were relevant to different "levels" of the conceptual hierarchy. Importantly, the path distance between the target object and its given property was predictive of the speed at which subjects responded.[17]. Collins, Quillian, et al. (1969) took this as evidence that human semantic memory is represented as a hierarchical graph[18] (or, a tree).

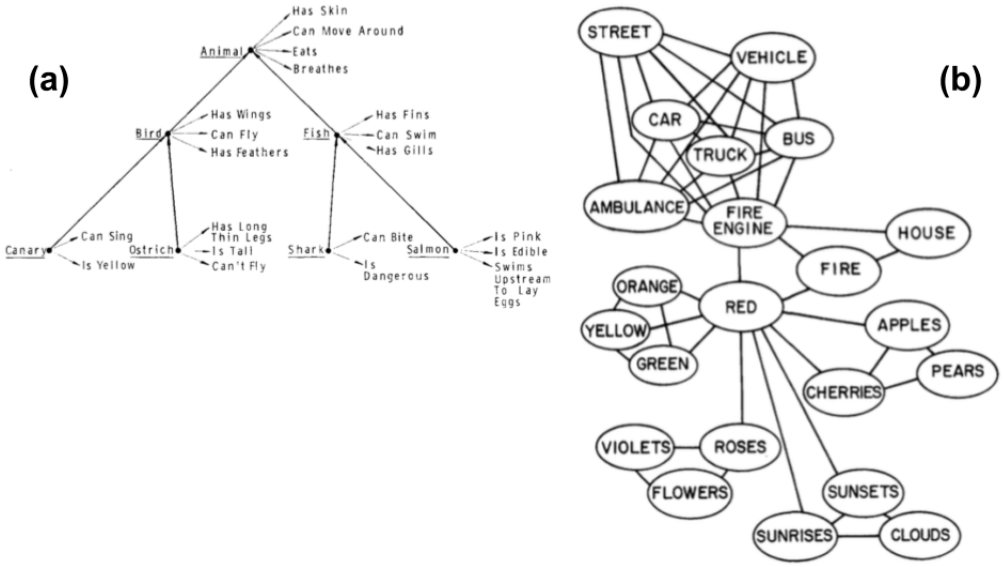

Figure 4: **(a)** Conceptual hierarchical as a model of human semantic representations proposed by Collins, Quillian, et al. (1969). **(b)** Extended version proposed by Collins and Loftus (1975) that relaxes the hierarchical and acyclic assumptions. Taken directly from *Collins, Quillian, et al. (1969) & Collins and Loftus (1975)*.

While certainly compelling, Collins, Quillian, et al. (1969) demonstration lacks predictive value as a model of cognitive representations. Given that the structure of a semantic network can be modified to "fit" any set of experimental results, *spreading-activation theory* lacked falsifiability (O'connor, Cree, and McRae, 2009). In an interesting demonstration, Steyvers and Tenenbaum (2005) help resolve this lack of falsifiability by tethering the hypothetical structure of semantic networks to a large cor-

[17]For example, two sentences might be: "A canary is an animal" or "A canary is a bird". Subjects should respond faster to the former sentence, since the property *bird* is closer to *canary* along the conceptual hierarchy than *animal*.

[18]This assumption was later relaxed in Collins and Loftus (1975)

pus of word association norms from Nelson, McEvoy, and Schreiber (2004)[19]. Steyvers and Tenenbaum (2005) performed a network analysis on the graph of word association norms, where each word (node) is connected according to the likelihood that it will trigger another word in a free association task. Steyvers and Tenenbaum (2005) found that the graph of word association norms embody a similar structure to many graphs in nature (including the phonological similarity network analyzed by Vitevitch, 2008), and laid the groundwork for integrating corpus analysis into the study of human semantic processing. In similar work, Abbott, Austerweil, and Griffiths, 2012 found that the results of a free association task could be predicted by a random walk algorithm on along a semantic network[20].

The findings that graph traversal algorithms can explain human semantic associations is promising evidence that cognitive representations leverage graphical structure (Abbott et al., 2012; Griffiths et al., 2007; Steyvers and Tenenbaum, 2005). However, independent of what cognitive representations actually are, an important constraint for any model of representations is that it be feasible to implement using neural hardware. Are graphical cognitive representations neurologically plausible? Though individual networks of neurons in the brain are an example of a graphical system (Bullock et al., 2005; Newman, 2003), whether neural activity manifests as graphical cognitive representations is an open question. However, a number of studies using a variety of neurophysiological methodologies have begun to explore this question (Bassett et al., 2010; Garvert, Dolan, and Behrens, 2017; Schapiro, Rogers, Cordova, Turk-Browne, and Botvinick, 2013).

Representational similarity analysis (RSA; Kriegeskorte, Mur, and Bandettini, 2008) is a methodological framework for exploring cognitive representations via the pairwise similarity of objects / stimuli on some psychological or physiological metric (e.g., similarity ratings or fMRI activity). A basic assumption of RSA is that the rep-

[19] In Nelson et al. (2004), around 149 students were shown a word, and asked to provide the first associated word that they thought of in response.

[20] Interestingly, Griffiths, Steyvers, and Firl (2007) found that Google's Page Rank algorithm for navigating the Web could also be leveraged to explain free association data in humans.

resentation of two objects are similar if they elicit a similar measurement on some metric of choice. Leveraging this basic assumption, Garvert et al. (2017) had subjects sequentially view images of objects while recording brain activity via fMRI. The order in which stimuli were selected was based on a random walk through a transition graph (similar to Kahn et al., 2018; Karuza et al., 2019). Garvert et al. (2017) found that the pairwise similarity between objects activations in of the entorhinal cortex (a mediator in the hippocampal memory system; Witter, 2011) was predicted by the path length between objects in the transition graph. In a similar preparation, Schapiro et al. (2013) found that community structure in a modular graph was also predictive of pairwise representational similarity in the Inferior Frontal Gyrus, the Anterior Temporal Lobe, and Superior Temporal Gyrus (using the same modular graph from Kahn et al., 2018). Interestingly, each of those brain regions is implicated in processing semantic meaning (Schapiro et al., 2013). In a more macro-level analysis of brain functioning, Bassett et al. (2010) found evidence of modularity in the graph representing the covariability in bold responses of various brain regions.

Graphical models provide an efficient way of storing the relatively massive amount of semantic and conceptual knowledge that humans possess (Abbott et al., 2012; Collins, Quillian, et al., 1969); in addition, a number of studies highlight the plausibility of graphical representations in human cognition (Bassett et al., 2010; Garvert et al., 2017; Steyvers and Tenenbaum, 2005) and that humans are sensitive to the network structure of various learning domains (Chan and Vitevitch, 2009, 2010; Kahn et al., 2018; Karuza et al., 2019; Karuza et al., 2017; Vitevitch et al., 2012). Given the productiveness of applying graph theory and network science to the study of human language and semantic processing, the next section will highlight relatively unaddressed implications of graphical representations on human category learning.

2 Implications of Graphical Representations on Category Learning

Graph structured representations have an interesting history in the category learning literature (going beyond Collins, Quillian, et al. (1969)'s assertion of a graphical representation of semantic meaning and category hierarchies). Early theorists recognized the importance of hierarchies in natural categories (Collins, Quillian, et al., 1969; Mervis and Rosch, 1981; Palmeri, 1999; though see Sloman, 1998), which embody a very particular type of graphical structure (a tree). Additionally, many leading models of category learning are instantiated as a graph using the connectionist framework[21] of cognition (Kruschke, 1992; Kurtz, 2007; Rogers and McClelland, 2004). Even the classic exemplar (Nosofsky, 2011) and prototype (Minda and Smith, 2001) theories of category learning can be interpreted as directed graphical models mapping features to probability distributions (Danks, 2007). Despite many theories independently invoking the idea that category representations are graphical in nature, there is very little work connecting category learning and network science. The next section will explore whether the language of network science can address any fundamental, ongoing issues in the category learning literature.

2.1 Unidimensional Bias as *Small World* Regularization

Traditional, artificial classification learning (TACL) has been a widely used paradigm for studying human category learning in the lab (Kurtz, 2015; Markman and Ross, 2003; Nosofsky, Gluck, Palmeri, McKinley, and Glauthier, 1994). TACL experiments typically involve repetitively exposing subjects to visual stimuli (one at a time); on each trial, subjects guess the category label of the presented stimulus (given a set of options). The category labels of the stimuli are predefined by the experimenter, of-

[21] Connectionist models are a particular class of graphical models that leverage distributed representations (Griffiths, Chater, Kemp, Perfors, and Tenenbaum, 2010; McClelland, Rumelhart, Group, et al., 1987)

ten specified in a way that helps disentangle different theoretical accounts. Stimuli are often composed of very simple objects with highly identifiable features, which can easily be represented as a feature vector of discrete or continuous values. The TACL paradigm has been a very productive avenue for advancing computational theories of category learning, uncovering a number of important phenomena that have become theoretical benchmarks in the modeling endeavor (Kurtz, 2015; Nosofsky et al., 1994; Wills and Pothos, 2012).

One notable benchmark phenomenon is the unidimensional bias; that is, the well-replicated finding that humans are especially adept at learning categories defined by a single feature (Kruschke, 1993; Nosofsky, 2011; Shepard, Hovland, and Jenkins, 1961). In addition, when asked to freely sort a set of objects into separate categories, humans tend to prefer sorting categories on the basis of single feature (Ahn and Medin, 1992; Ashby and Alfonso-Reese, 1995; Medin, Wattenmaker, and Hampson, 1987; Milton, Wills, and Hodgson, 2009). In a foundational study, Medin et al. (1987) gave subjects an array of stimuli and asked them to sort the stimuli into 2 equal groups. The stimuli varied according to 5 relevant, binary features (that could take on 2 possible values/states). Importantly, the stimuli were selected by the experimenters from a 2-category structure defined by *family resemblance*, where each stimulus has most features in common with one of 2 *prototypes*[22]. Across several experiments, subjects (who were unaware of the actual category structure that stimuli were sampled from) were much more likely to sort categories into groups based on a single dimension[23].

Why might the unidimensional bias be theoretically unintuitive, despite consistent replication in the category learning literature? Many theorists have argued that a "ideal" category structure should group exemplars so that members of the same category are featurally similar, while members from opposing categories are featurally dissimilar (Medin et al., 1987; Rosch and Mervis, 1975). The *family resemblance* structure

[22]The *Family Resemblance* category structure is notable in that it seems to underlie many categories in the natural world (Rosch and Mervis, 1975)

[23]In fact, some of Medin et al. (1987)'s experiments failed to show any behavior besides unidimensional sorting

is ideal in that it accomplishes that goal (in addition to being prevalent in natural categories). On the contrary, a unidimensional category does not *seem* to consistently meet this optimization goal; it might produce exemplars that are identical on the basis of one feature, but would allow exemplars to be radically different when considering other features[24]. Many successful accounts of category learning leverage selective attention as a mechanism to prioritize a unidimensional bias (Kruschke, 1992; Love et al., 2004; Nosofsky, 2011); selective attention is likely a critical factor in why some theoretical accounts succeed relative to others.

There are various explanations as to why unidimensional bias might be so prevalent despite it's clash with historic theories about what defines an "ideal" category structure. Some researchers have argued that focusing on a single dimension requires less effort than integrating all features into a similarity computation (Wills, Milton, Longmore, Hester, and Robinson, 2013). Another theoretical suggestion comes from Feldman (2000), who demonstrated that a family resemblance category structure in the domain of binary-featured stimuli has a higher degree of *boolean complexity* than a unidimensional category structure[25]. Both of these accounts are compelling (and neither are mutually exclusive). However, the next section will propose another (also not mutually exclusive) explanation of why the unidimensional bias might persist in human cognition (using a well known phenomena in the network science literature).

Small World[26] networks are formally defined as a type of sparsely connected graph where the average path distance between nodes grows less rapidly as the number of nodes in the network increases (Porter, 2012). Importantly, small world networks have a number of interesting statistical properties that commonly describe many real world graphs in nature, such as:

[24] For example, *apple* and *firetruck* both belong to the category of *red objects*, but share very few features in common.

[25] *Boolean complexity* can be defined as the shortest possible mapping between the set of binary features and a dichotomous category label that can be realized by a boolean circuit (Feldman, 2000).

[26] The term *small world* is in reference to the fact that it's unintuitively easy to find a short, associative path connecting two human beings despite the fact that the entire social network defining human communication is incredibly vast and complex (Newman, 2003).

- the WorldWideWeb (Newman, 2003),
- the human brain (Bassett and Bullmore, 2006[27],
- the graph representing phonological similarity of English words (Vitevitch, 2008, as described earlier),
- semantic networks generated from free recall experiments (Nelson et al., 2004; Steyvers and Tenenbaum, 2005).

Small world networks can be characterized as being sparsely connected, with a relatively smaller number of densely connected nodes acting as "hubs" connecting the rest of the nodes in the network (Steyvers and Tenenbaum, 2005). Another important feature of small world networks is that they are relatively easy to navigate, particularly when the navigator has limited access to anything beyond local information (Newman, 2003). While it might seem like a stretch to say that this "easy to navigate" property has any implications to human cognition, recall that Abbott et al. (2012) demonstrated that human free association appeared particularly similar to a random walk along a graph of semantic associations.

How might small world networks relate to the unidimensional bias in category learning? In a category learning experiment, stimuli are defined by a set of features that are each predictive (or not predictive) of some category label. Many category learning models can be functionally described as building probability density functions mapping features to category labels (Ashby and Alfonso-Reese, 1995), reducing the problem of category learning to feature-label association. If the goal of the human categorization system is to produce isolated mappings between features and labels, it seems surprising that subjects might prefer the "less-than-deal" unidimensional category structure. However, note that the features and labels used in category learning studies bear close resemblance to the conceptual "nodes" researchers invoke when de-

[27] though see Hilgetag and Goulas (2016) and Papo, Zanin, Martinez, and Buldu (2016) for limitations to the small world model of neuroanatomy

scribing graphical models of human semantic representations (Collins and Loftus, 1975; Collins, Quillian, et al., 1969; Love et al., 2004; see figure 1.4). If the goal of a categorization system is to construct and refine the complex semantic networks historically invoked by cognitive scientists (Abbott et al., 2012; Collins and Loftus, 1975; Collins, Quillian, et al., 1969; Love and Sloman, 1995; Sloman, 1998; Steyvers and Tenenbaum, 2005), then a unidimensional "bias" might serve a very functional purpose: regularizing sparsity in a graphical representation of our environment.

Steyvers and Tenenbaum (2005) found that the semantic network describing free association norms resembled a small world network, which are characterized as being sparsely connected. This sparse connectivity might make navigation through semantic networks more efficient (Abbott et al., 2012; Newman, 2003). How might a unidimensional bias enforce sparsity? Steyvers and Tenenbaum (2005) developed a recursive algorithm for generating the types of small world networks commonly found in nature. The algorithm starts with an initial set of fully connected nodes, incrementally adding new nodes with M connections to a randomly sampled neighborhood[28]. Steyvers and Tenenbaum (2005) found that a relatively low M (of 11) can produce relatively large networks[29] that retain some of the key elements of small world networks.

Each iteration of the Steyvers and Tenenbaum (2005) algorithm seems intuitively analogous to the structure of a typical category learning experiment: one node (a category label) is probabilistically associated with a neighborhood of other nodes (the set of critical features present in each stimulus). If the mechanisms of category learning we typically isolate in the laboratory reflect some generative process of building a probabilistic network of knowledge about things and their features, then a unidimensional prior on probabilistic associations might lead to more efficient knowledge retrieval later on. As a proof-of-concept, large networks (with 200 nodes) were generated using the network building algorithm of Steyvers and Tenenbaum (2005); the M parameter was adjusted for each. The result of this simulation highlight that the denser networks pro-

[28] M will influence how sparse the final network is
[29] 5018 nodes in total

duced node degree distributions that deviate from what's expected in a small world network (figure 2.1). While a unidimensional prior might be less-than-ideal in an isolated laboratory experiment, it might be a necessary trade-off for efficient navigation in representational space[30].

2.2 Similarity, Graph Embeddings, and Path Length

Similarity is a commonly invoked construct in many domains of cognition (Goldstone, 1994b), particularly for theories of categorization (Goldstone and Son, 2012; Kurtz, 2015; Nosofsky, 2011; Shepard, 1987). An intuitive assumption about categories is that objects from the same category should be similar in some way. Many category learning models leverage a spatial interpretation of similarity, where the similarity between two stimuli (or between a stimulus and a category representation) is determined by their distance in some psychologically-transformed space (Kruschke, 1992; Minda and Smith, 2001; Nosofsky, 2011). While category learning models that rely on spatial similarity metrics have provided relatively successful accounts of human learning and generalization, spatial interpretations of similarity make a number of critical assumptions that violate numerous empirical findings (Holyoak and Gordon, 1983; Polk, Behensky, Gonzalez, and Smith, 2002; Tversky, 1977; Tversky and Gati, 1982. Additionally, similarity judgments in category learning models often only consider category representations and category exemplars in isolation, neglecting the impact of relational knowledge on similarity perception (Goldstone, Medin, and Gentner, 1991; Markman and Gentner, 1993). The next section will contrast spatial and relational approaches to similarity, and highlight the strengths & weaknesses of graphical representations for addressing unresolved issues in similarity research.

***Spatial Metaphor of Similarity*:** A widely prevalent (though controversial) assumption in psychology is that stimuli can be represented as vectors of values defined

[30]Speculatively, reduction in graph density through the unidimensional bias may also lead to conceptual representations with less boolean complexity as well (Feldman, 2000).

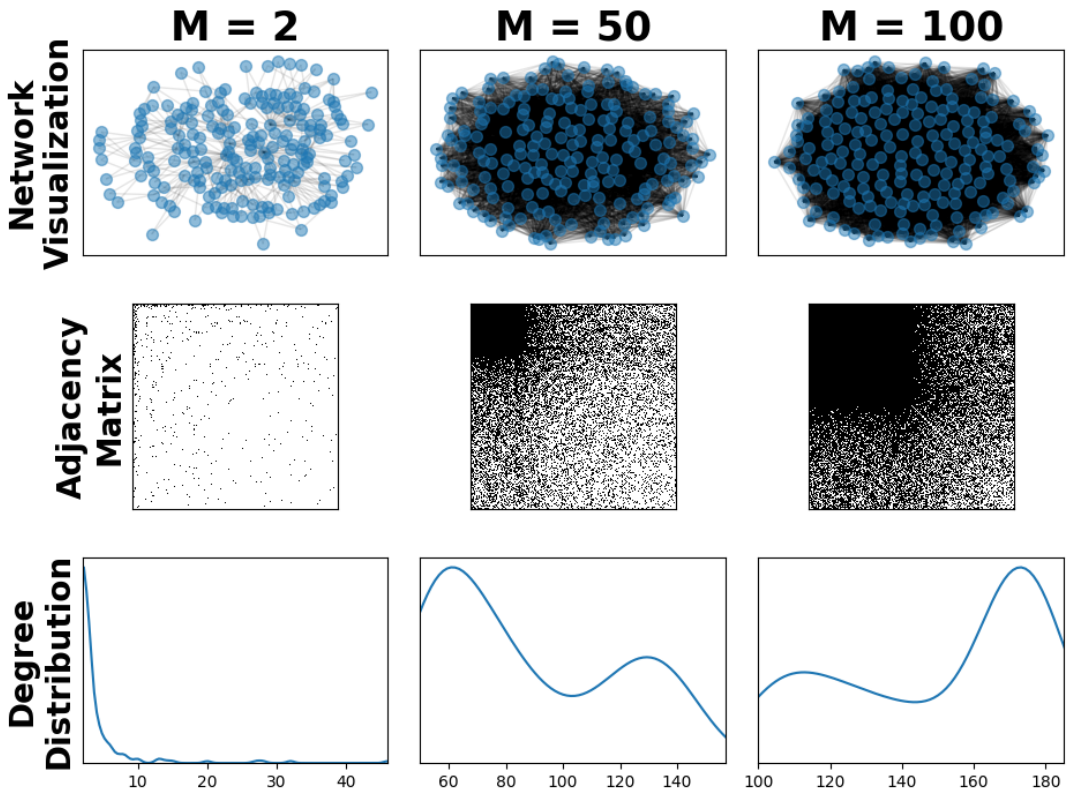

Figure 5: Results from a simulation showing the network, adjacency matrix, and node degree distribution for 3 examples of a network built using the algorithm from Steyvers and Tenenbaum (2005). Note that in all instances, the total number of nodes is held constant.

by their features (Kurtz, 2015; Shepard, 1987)[31]. There are number of phenomena that make this assumption seem viable. For example, Guttman and Kalish (1956) trained pigeons to produce pecking responses during the presentation of a color stimulus. After training, Guttman and Kalish (1956) varied the wavelength of the color stimulus and found that the frequency of pecking responses decreased monotonically as wavelength deviated from the value originally seen during training [32]. At a physiological level, the firing rates of individual neurons in the animal brain can be predicted by continuous functions mapping firing rates to stimulus dimensions (Butts and Goldman, 2006; Kang, Shapley, and Sompolinsky, 2004). Shepard (1987) extended this idea by suggesting that the probability of stimulus generalization decreases exponentially as dissimilarity increases (i.e., a monotonic generalization gradient). Additionally, Shepard (1987) relaxed any assumptions about true stimulus features and suggested that the exponential generalization gradient spanned some psychological space that could be inferred using a validated behavioral measurement (e.g., similarity ratings).

The exponential decay function suggested by Shepard (1987) has been leveraged by many successful models of human category learning, which typically utilize a similarity metric that compares the similarity of objects to some set of reference points. These *reference points* might consist of previously experienced exemplars (Kruschke, 1992; Nosofsky, 2011), category-specific centroids (Minda and Smith, 2001), or category-independent centroids that capture densely populated areas of feature space (Love et al., 2004; Rosseel, 2002). Critically, these models rely on stimulus descriptions that only consider stimulus features and associated category labels. This approach has been successful in adequately describing a number of empirical phenomena in the categorization literature (Kruschke, 1993; Nosofsky et al., 1994). However, models leveraging similarity in feature space explicitly neglect any information about how exemplars and categories relate to each other or their given context. This shortcoming is critical given

[31] For example, the representation of the squares in figure 2.2 can be represented as a vector describing their size and shading

[32] The training wavelength in Guttman and Kalish (1956)'s experiment can be represented as a single point in a 1-dimensional, continuous space

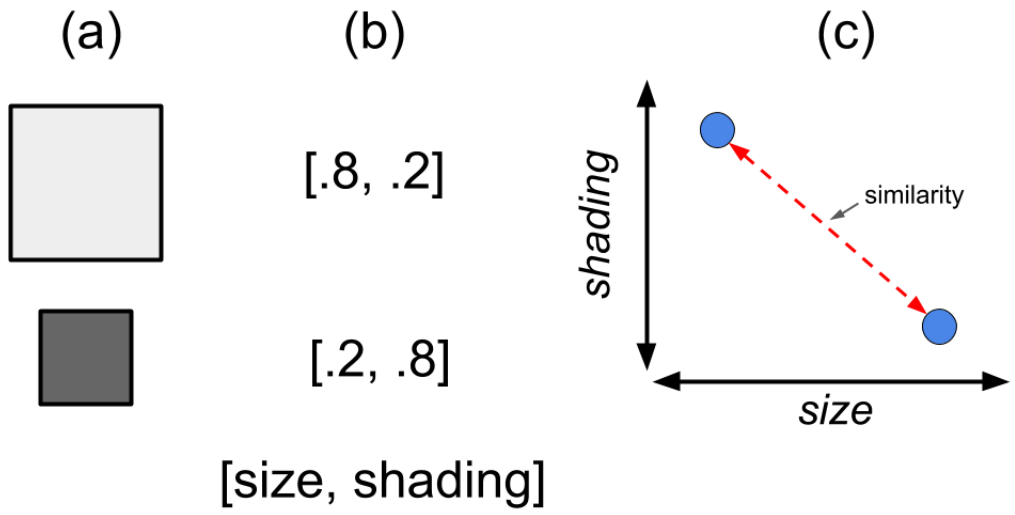

Figure 6: (a) Example of two stimuli, each with 2 core features (size & shading). (b) Those same stimuli represented as feature vectors. (c) The two stimuli plotted in a continuous space.

many empirical demonstrations that suggest an important role of relational knowledge in similarity perception (Goldstone, 1994a; Goldstone et al., 1991; Markman and Gentner, 1993).

***Relational Knowledge and Similarity Perception*:** In category learning models described thus far, stimuli (or, exemplars) are defined by their *features* (or, attributes). In an effort to highlight the dissociation between features and relations, Goldstone et al. (1991) asked subjects to pick which of 2 base stimuli was more similar to a target stimulus (figure 2.2). Critically, the base stimuli differed in the number of shared features or shared relations with the target stimulus. Across a series of experiments using a variety of stimulus sets, Goldstone et al. (1991) found that (a) subjects were more likely to choose the relationally similarity stimulus over the featurally similar stimulus, and (b) subjects choices were influenced by the number of existing relational or featural matches in the set. That is, subjects were more likely to choose the featurally similar stimulus when the number of pre-existing featural matches was high, and more likely to choose the relationally similar stimulus when the number of pre-existing featural matches was low. In addition to showing the role of relations in

similarity perception, Goldstone et al. (1991)'s results suggest that the collective dynamics of features and relations impact similarity judgments in distinct ways.

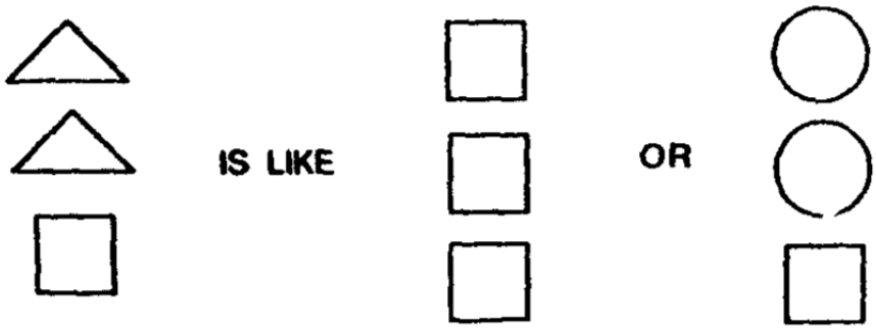

Figure 7: example of the task in Goldstone, Medin, and Gentner (1991) (taken directly from Goldstone, Medin, and Gentner, 1991)

In a different preparation, Markman and Gentner (1993) showed subjects pairs of scenes involving a set of related objects (figure 2.2). Subjects engaged in a *one-shot mapping* tasks, where they were directed to find the item from the target scene that "goes with" the item from the base scene. Importantly, the target object was present in both scenes (though its relational role was different in each). If subjects were considering only the featural attributes of the scenes, then the obvious choice is to choose the same object as it appears in both scenes. However, Markman and Gentner (1993) found many instances where subjects mapped objects based on their relational roles in the scene (despite those objects having relatively dissimilar features). The likelihood for relational responding increased when subjects were asked to rate the similarity of the two scenes[33] before the one-shot mapping task.

The importance of relational roles in similarity perception was explored further by Goldstone (1994a), who asked subjects to rate the similarity of two objects composed of features with identical roles (e.g., wings, tail, head, body). Goldstone (1994a) found that similarity ratings were strongest when objects shared properties on features with the same role (e.g., both objects have red tails) than when the objects shared properties on features with different roles (e.g., object A has a red tail and object B has a red

[33] on a scale from 1-9

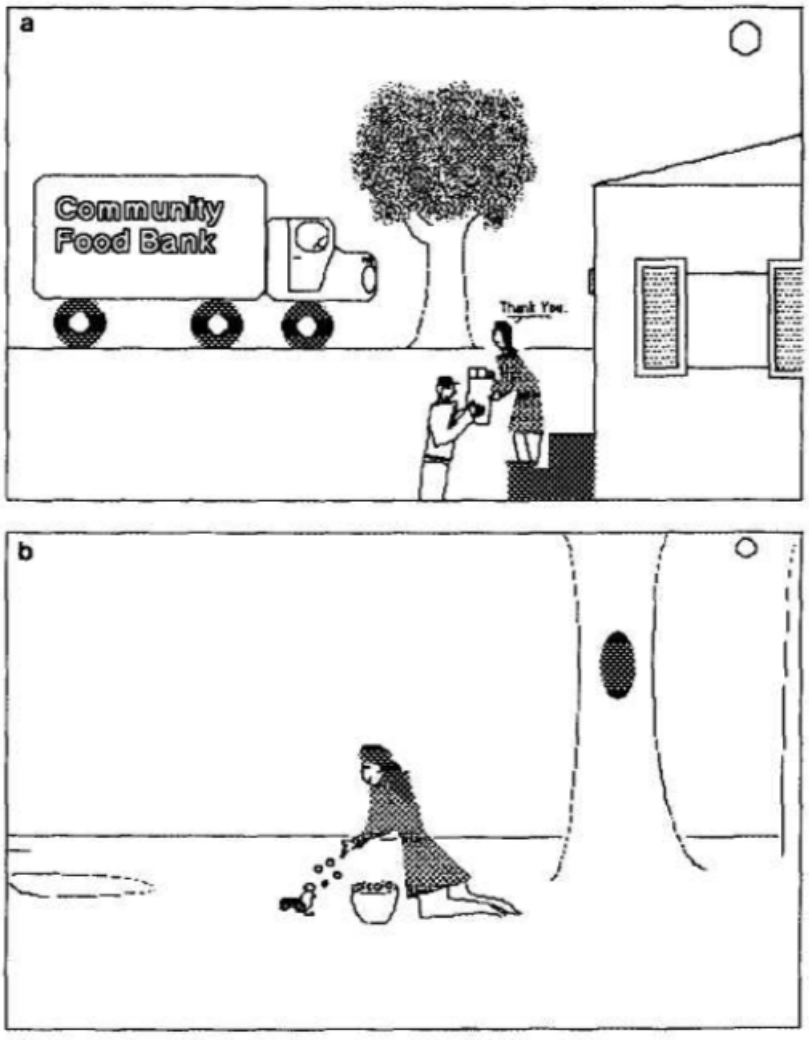

Figure 8: example of the stimulus preparation used in Markman and Gentner (1993) (taken directly from Markman and Gentner, 1993)

body)[34]. These results highlight that the structural information defining a stimulus influences similarity perception in a way that's difficult to account for when stimuli are strictly represented as feature vectors in isolation.

***Spatial Similarity and Graph Embeddings*:** In important question is whether findings from the relational similarity literature can be reconciled with findings that exponential distance is highly predictive of categorization and generalization behavior in human learners (Kruschke, 1992; Love et al., 2004; Nosofsky et al., 1994; Shepard, 1987). Typically, models leveraging spatial similarity computations treat stimuli

[34]The property used in this example (color:red) was not an actual property used by Goldstone (1994a)

as points in the space defined by their features (often modified using a multidimensional scaling algorithm that distorts the space to better reflect some other psychological property[35]; Shepard, 1962). Connectionist models deviate from this approach by transforming stimuli into a new representational space that optimizes the model's ability to perform some task[36]. In both of these distinct approaches, a stimulus's position in psychological space is determined primarily by its feature vector. There is no explicit impact of relational information present in the domain. The reconciliation of spatial and relational theories of similarity might benefit from a representational space that *explicitly* considers relational knowledge about a given stimulus.

Graph embedding algorithms aim to embed stimuli from a graph onto a vector space that optimally organizes stimuli based on the interaction between their attributes and the attributes of their neighbors[37] (Kipf and Welling, 2016; Veličković et al., 2017; Z. Wu et al., 2020). Geometric deep learning algorithms that generate graph embedding spaces are ideal for a number of reasons:

- they utilize connectionist-style learning algorithm (which have had widespread success as explanatory models in psychology and neuroscience; Khaligh-Razavi and Kriegeskorte, 2014; Lake et al., 2015; Long et al., 2018; Peterson et al., 2016),
- the embedding space is constrained by the task of the model (which can be flexible), and
- distance between stimuli in a graph embedding space reflect relational and object-specific properties of the network being learned.

Once a sufficiently useful graph embedding space is learned, it is relatively straightforward to apply Shepard (1987)'s inverse exponential generalization gradient to the

[35] for example, pairwise similarity ratings

[36] Interestingly, the representational space of some variants of connectionist models predict a number of interesting behavioral and physiological phenomena of human beings (Khaligh-Razavi and Kriegeskorte, 2014; Lake, Zaremba, Fergus, and Gureckis, 2015; Long, Yu, and Konkle, 2018; Peterson, Abbott, and Griffiths, 2016)

[37] or in many cases, their extended neighborhoods as well

space where stimuli are embedded. This might provide a way of integrating relational and spatial theories of similarity; though it is an open question whether distance in a graph embedding space has any correspondence to psychology of similarity perception. In addition, there are many different variants of geometric deep learning algorithms, each with particular properties that shape the representational space being learned (Kipf and Welling, 2016; Veličković et al., 2017; Z. Wu et al., 2020). While this over-flexibility might limit the usefulness geometric deep learning algorithms as models of human similarity perception, psychologist and machine learning researchers might both mutually benefit from investigating which graph embedding spaces most closely align with human behavior (e.g., similarity ratings and generalization).

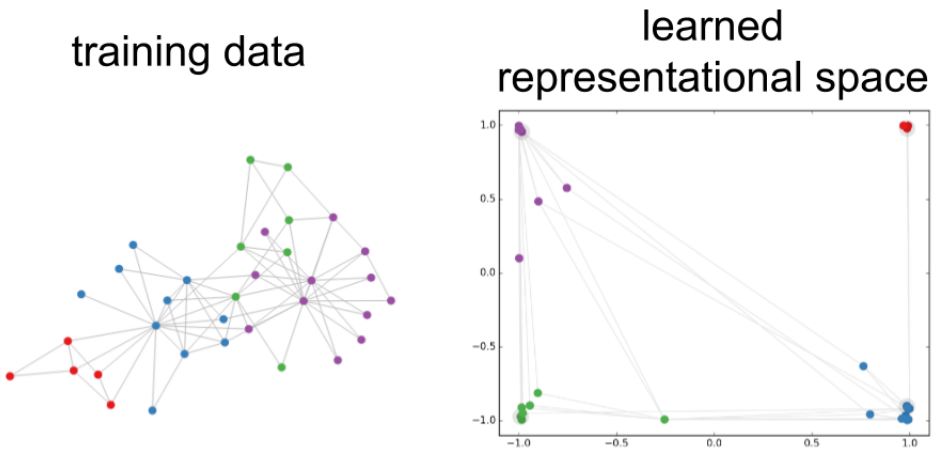

Figure 9: Example of a graph embedding spaced learned by a Graph Convolutional Network (Kipf and Welling, 2016). Taken directly from Kipf and Welling (2016).

***Further Limitations of the Spatial Metaphor: Symmetry and the Triangle Inequality*:** Graph embedding spaces provide one preliminary avenue towards allowing spatial theories of similarity to leverage relational properties of a learning domain. However, there are other historical limitations with the spatial metaphor of similarity that are also present in graph embedding spaces: *symmetry* and the *triangle inequality problem*. Symmetry refers to the counterintuitive phenomenon where similarity

judgments between 2 stimuli vary depending on presentation[38] (Holyoak and Gordon, 1983; Polk et al., 2002; Tversky, 1977; though see Nosofsky, 1991). For example, Tversky (1977) asked subjects to rate the similarity between a less and a more prominent country. Tversky (1977) found that subjects gave higher similarity ratings when the less prominent country was presented first[39]. This demonstration and others (Holyoak and Gordon, 1983; Polk et al., 2002) highlight a clear symmetry violation in similarity ratings that deviates from a purely spatial interpretation of similarity perception[40].

The second limitation was raised by Tversky and Gati (1982), who found conditions where human similarity perception violates a fundamental assumption of geometric spaces. More specifically, a geometric representational space requires that the distance between points A and C must always be less than the distance between points A and B plus the distance between B and C; referred to as the *triangle inequality* ($dist(A,B)+dist(B,C) > dist(A,C)$; see figure 2.2.a). Tversky (1977) provides an interesting set-theoretical alternative that doesn't make the same geometric assumptions of spatial similarity models. Tversky (1977)'s *featural model* suggests that the similarity between 2 stimuli i and j reflects 3 important components: (1) shared features between the stimuli, (2) exclusive features of i, and (3) exclusive features of j (see figure 2.2.b; Tversky (1977)); given by the equation:

$$similarity(i,j) = \theta f(i \cap j) - \alpha f(i-j) - \beta f(i-j)$$

where θ, α, β are weights on some function f. In Tversky (1977)'s model, there are no inherent assumptions requiring similarity ratings to follow the normal axioms of geometry (e.g., symmetry, the triangle inequality). Interestingly (and of relevance to this

[38] which would be unusual if stimuli are represented as points in some vector space; the similarity between points A and B should be exactly the same as the similarity between points B and A

[39] It is worth noting that many demonstrations of asymmetry utilize this type of sequential presentation (Polk et al., 2002; Tversky, 1977).

[40] Interestingly, Polk et al. (2002) also found asymmetries between the number of processing cycles needed for a recurrent neural network to shift from one stimulus representation to another, which (a) might suggest an important role of recurrent computation in similarity perception, and (b) arguably lends support for *transformational* theories of similarity (Imai, 1977; Wiener-Ehrlich, Bart, and Millward, 1980).

paper), Tversky (1977)'s featural model can be reformulated in the language of network science. For example, if stimuli and featural attributes are represented as connected nodes, than Tversky (1977)'s featural model can be framed as a computation on the resulting bipartite[41] graph. If we remove the weights and functions that transform each of the 3 components in Tversky (1977)'s model[42], then the resulting similarity function can be described as the number of shared neighbors minus the number of distinct neighbors.

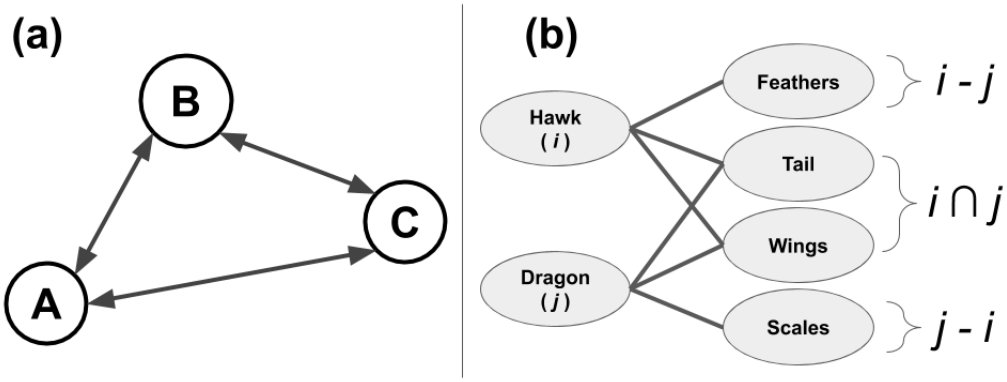

Figure 10: **(a)** 3 stimuli plotted in geometric space. **(b)** Tversky (1977)'s featural model of the similarity between two items i and j, visualized as a bipartite graph.

It's an open question whether there are benefits to framing Tversky (1977)'s feature model through the lens of network science beyond an appeal to Occam's razor. However, *shared features* in a graph corresponds with another network science metric that has had notable consideration in the semantic similarity literature: path distance between two nodes (Collins and Loftus, 1975; Kenett, Levi, Anaki, and Faust, 2017; Lee, Kim, and Lee, 1993; Rada, Mili, Bicknell, and Blettner, 1989). For example, Rada et al. (1989) suggested that the average shortest path length between two nodes could accurately predict human semantic judgments. Additionally, Kenett et al. (2017) found that path length predicted reaction times in discrete semantic relatedness judgments. These results, as well as the empirical success of Tversky (1977)'s featural model (Tversky and Gati, 1982), leave open the possibility that network statistics might underlie

[41] wikipedia's description of bipartite graphs
[42] which may arguably be too drastic of a change to the theory to make this exercise meaningful

various phenomena in the similarity perception literature[43].

***Subsection Conclusion*:** The concept of similarity has received over a century of attention from psychologists (Attneave, 1950; Goldstone and Son, 2012; James, 1890/1950; Tversky, 1977), but has been notoriously difficult to resolve. Similarity appears to be multifaceted, evident by the empirical success of many distinct theoretical camps (Goldstone et al., 1991; Nosofsky et al., 1994; Shepard, 1987; Tversky, 1977). There may be key differences in the nuances of particular experimental preparations that predict when one theoretical model of similarity will succeed over another. For example, TACL experiments typically utilize isolated stimuli that collectively vary on a restricted set of shared features[44] (Kurtz, 2015; Nosofsky et al., 1994). The success of the spatial metaphor in this domain might be contingent on the typical lack of disjunctive features and relational statistics that would otherwise influence similarity perception[45]. Additionally, violations of the spatial metaphor are typically observed via sequential presentation of stimuli (as opposed to pairwise comparison), and typically use relatively complex, conceptual stimuli (Holyoak and Gordon, 1983; Tversky, 1977)[46]. The apparent inconsistencies in the spatial metaphor might also reflect the type of similarity space that theorists commonly invoke[47] (e.g., city-block, euclidean). Beyond differences in methodological preparations, network science and graph statistics may also help explain similarity perception in the attribute category learning and relational similarity literature as well.

[43] However, the spatial versus featural debate seems to have been reignited by the finding that distributional semantic models (which closely resemble the spatial metaphor of similarity) are also predictive of similarity judgments (Bhatia, Richie, and Zou, 2019; Günther, Dudschig, and Kaup, 2016; though see De Deyne, Perfors, and Navarro, 2016), leaving this issue unresolved in the present literature.

[44] e.g., size, shading, color

[45] i.e., the second and third component of Tversky (1977)'s featural model reduce to zero

[46] But not in Polk et al. (2002), who used color patches (which is arguably closer to stimuli used in a standard TACL experiment)

[47] Jäkel, Schölkopf, and Wichmann (2008) suggest that Hilbert spaces might be better suited to serve the spatial metaphor of similarity

3 Graphs, Cognition, and Directions in Graph Learning Research

Graph theory and network science might be useful in explaining many empirical observations in the category learning literature, such as category hierarchies (Mervis and Rosch, 1981), exemplar and prototype effects (Danks, 2014), attention and unidimensional bias, and similarity perception. As discussed earlier in this paper, graphs have also been implicated language (Chan and Vitevitch, 2009, 2010; Vitevitch, 2008; Vitevitch et al., 2012), motor sequences (Kahn et al., 2018), and event segmentation (Karuza et al., 2019; Karuza et al., 2017). The next section will briefly highlight other domains in which graph theory and network science have been (and could be) invoked in explanatory models of cognition. Then, recent computational approaches to graph learning will be discussed.

3.1 Graphs and Cognition

***Causal Cognition*:** The ability for humans to recognize causal relationships between objects and events in their environment has been suggested as a hallmark of human cognition (Penn, Holyoak, and Povinelli, 2008). Internal mental or conceptual models are often invoked as the mediating construct of our causal cognitive abilities (Barrett, Abdi, Murphy, and Gallagher, 1993; Carey, 1985; Murphy and Medin, 1985; Springer and Keil, 1989). While the notion of a mental model is somewhat flexible[48], many researchers have suggested that our causal understanding of the world is grounding in probabilistic graphical models (Danks, 2014; Griffiths et al., 2010; Holyoak and Cheng, 2011). Probabilistic graphical models (or, Bayesian networks) are directed, acyclic[49] graphs (DAG) that represent objects or events as nodes and causal relationships as one-directional, weighted edges. Many theoretical explanations of causal cog-

[48] and arguably difficult to falsify

[49] Acyclic refers to the constraint that no child node can be connected to its parent (i.e., no cycles); which arguably prevents circular reasoning

nition posit that the goal of the learner is to estimate useful parameters that best describe a graphical model of the learner's environment (Danks, 2014).

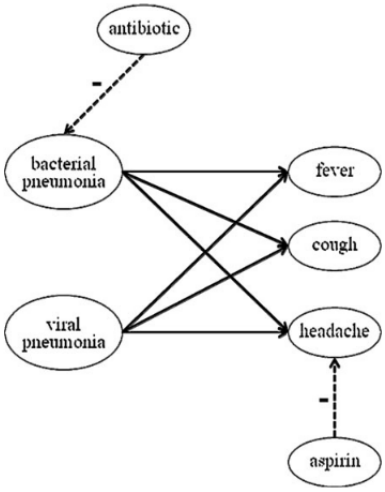

Figure 11: Example of a Bayesian network describing the causal connection between pathogens and experienced symptoms (taken directly from Holyoak and Cheng, 2011)

Unlike connectionist networks, probabilistic graphical models require an explicitly defined modular structure prior to parameter estimation[50] (Griffiths et al., 2010). However, because the latent structure of probabilistic graphical models are typically specified by the experimenter (McClelland et al., 2010), there is a large explanatory gap in how information from a noisy perceptual experience is mapped onto the relevant elements of a graphical representation (i.e., the binding problem). In contrast, connectionist networks excel at learning complex mappings between raw perceptual data and some desired behavior (He, Zhang, Ren, and Sun, 2015; LeCun, Bengio, and Hinton, 2015). Interestingly, probabilistic graphical models can be fluidly integrated with connectionist algorithms to map real-world, complex perceptual data onto the latent variables of a graphical model using variational inference techniques (M. J. Johnson, Duvenaud, Wiltschko, Adams, and Datta, 2016), potentially bridging the gap between emergent (McClelland et al., 2010) and structured (Griffiths et al., 2010) approaches to

[50] which researchers have argued makes Bayesian networks considerably more interpretable than connectionist networks (Griffiths et al., 2010)

explaining semantic and causal cognition[51]. It may also explain how simple, associative learning mechanisms can produce a representation of conditional relationships between objects and events.

***Graphs and Analogical Reasoning*:** Analogical reasoning is also considered to be a hallmark of human intelligence (Penn et al., 2008), its importance ranging from normal cognitive behavior to systematic scientific discovery (Black and Hesse, 1966; Gentner and Holyoak, 1997; Gust, Krumnack, Kühnberger, and Schwering, 2008). A common theme in many theories of analogical reasoning is the invocation of structured representations that explicitly isolate object relations from object features (Doumas, Hummel, and Sandhofer, 2008; Forbus, Ferguson, Lovett, and Gentner, 2017; Gentner, 1983; Hummel and Holyoak, 1997; Larkey and Love, 2003)[52]. In her seminal paper on analogical reasoning, Gentner (1983) describes knowledge representation as a propositional network: *relations* and *attributes* serve as predicates that leverage *features*, *objects*, and *relations*[53] as arguments. This use of predicate calculus[54] to describe structured representations is a common element of many distinct computational models of analogical reasoning (Doumas et al., 2008; Forbus et al., 2017; Hummel and Holyoak, 1997; Larkey and Love, 2003). In contrast with the approach offered by Gentner (1983), Kemp, Griffiths, and Tenenbaum (2004) suggest that relational systems themselves can be represented as graphs where objects/entities are represented by nodes and relations are represented by edges. Kemp, Tenenbaum, Niyogi, and Griffiths (2010) demonstrate how probabilistic models can leverage this representational style for theory discovery. Both of these approaches leverage networks as the fundamental data structure of knowledge representation, suggesting that network science may play a key role

[51] which may reduce down to a question of where certain aspects of cognition reside on a scale ranging from completely distributed to completely modular (given that both connectionist and Bayesian networks are both different types of graphs; Griffiths et al., 2010)

[52] This isolation is supported by evidence of a dissociation between relational and featural similarity judgments of human subjects (Goldstone et al., 1991)

[53] Gentner (1983) defines higher-order relations as relational predicates that take other relations as arguments

[54] See Rensink (2004) for a demonstration on how predicate calculus (or, first-order logic) can be represented as a graph

in explaining many hallmark characteristics of human intelligence.

***Object Perception and Scene Grammars*:** Graphical structured representations have a history of being invoked in theories of object perception and scene understanding (Biederman, 1987; Hummel and Biederman, 1992; Pylyshyn, 1973). For example, Pylyshyn (1973) argued that mental imagery of objects and scenes were mediated primarily by propositional representations[55], leveraging the symbolic, recursive mechanisms that (Pylyshyn argues) mediate language (Chomsky, 1965). Biederman (1987)'s very influential *Recognition-by-components* theory suggests that wholistic object perception is grounded in a relational hierarchy of simple geometric components (or, *geons*). Han and Zhu (2008) utilize hierarchical, graphical representations of objects in scenes (which they refer to as *attribute grammars*) for parsing simple shapes from natural images. The theoretical notion that visual understanding leverages the same fundamental representational structure as language, causal reasoning, and analogy is particularly attractive: it allows various aspects of cognition from a variety of learning domains to be explained by a relatively restricted set of computational principles (Griffiths et al., 2010)[56].

While a graphical explanation of object and scene understanding is compelling, it faces the same fundamental problem of mapping objects from a continuous perceptual stream into their respective roles in structured representation. Typically, this task is handled using hand-coded representations generated by the experimenter, which Chalmers, French, and Hofstadter (1992) argue leaves many important questions unanswered. Both probabilistic graphical models and connectionist models have historically suffered from this limitation (Hinton, 1981; Hummel et al., 2004; Torralba, Tenenbaum, and Salakhutdinov, 2011). However, deep learning algorithms might provide a tentative solution. Deep learning models are particularly useful for learning statistically useful visual features that can be leveraged as constituent elements in Bayesian probabilis-

[55] Recall the close connection between propositional logic and graphs (Rensink, 2004)

[56] Griffiths et al. (2010), who originally suggested graphical models as the common representational structure of vision, language, and causal reasoning in 2010, also argued that Bayesian probabilistic networks were a good candidate for this unification endeavor

tic models (M. J. Johnson et al., 2016; Torralba et al., 2011). Santoro et al. (2017) used features from a convolutional network (CNN) as objects in a relational learning model that achieved relatively high performance on the CLEVR[57] dataset (J. Johnson et al., 2017). Given that deep learning is undergoing preliminary success as an explanatory model of human visual perception (Khaligh-Razavi and Kriegeskorte, 2014; Long et al., 2018; Peterson et al., 2016), deep learning models might be the tentative, psychologically-plausible[58] bridge between sensory perception and structured representations of objects and scenes (though see Baker, Lu, Erlikhman, and Kellman, 2018, Baker, Erlikhman, Kellman, and Lu, 2018, and Erdogan and Jacobs, 2016 for key explanatory limitations of deep learning models in their present form)[59].

3.2 Methodological Paradigms for Studying Graph Learning in Humans

How humans behave in graph structured learning domains is an emerging field of research, already garnishing some promising results (Kahn et al., 2018; Karuza et al., 2017; Lynn and Bassett, 2019; C. M. Wu, Schulz, and Gershman, 2020). However, the number of available paradigms is somewhat limited. Much of the work suggesting human sensitivity to network structure relies on the sequential presentation of stimuli in isolation, where the structural relations are defined by the transition probability from one stimulus to another (Kahn et al., 2018; Karuza et al., 2017; Zurn and Bassett, 2020). While work on transition graphs are a promising avenue for integrating network science and human cognition, it arguably embodies a particular type of homogeneous relation that might motivate distinct learning strategies relative to other types of relations (such as semantic or spatial)[60].

[57] which tasks a model with answering questions about the spatial relations between simple geometric objects

[58] relatively speaking

[59] However, note that these demonstrations do not consider the recent work leveraging deep networks for structured representations (discussed in this section), leaving open the question of the explanatory value of deep learning in cognitive science

[60] Additionally, the relations present in semantic networks are often very heterogeneous

Another interesting question in the graph learning literature is the impact of manipulating the statistical relationships between stimulus features in addition to network topology. Early steps in answering this question have recently been taken by C. M. Wu et al. (2020). In one experiment, C. M. Wu et al. (2020) had subjects view an entire graph where each node was defined by a single feature (number of passengers at a train station). Subjects were asked to guess the feature value of an unlabeled node, ideally using information about other connected nodes in the graph to guide their judgments. Importantly, C. M. Wu et al. (2020) found that subjects were able to improve their judgments about the functional relationship between node features as the task progressed. In contrast with work using transition probabilities, subjects in C. M. Wu et al. (2020)'s experiment were shown the entire graph structure on each trial. The use of entire graphs as stimulus presentation provides an interesting addition to the presently limited number of paradigms where subjects are explicitly given local information alone[61].

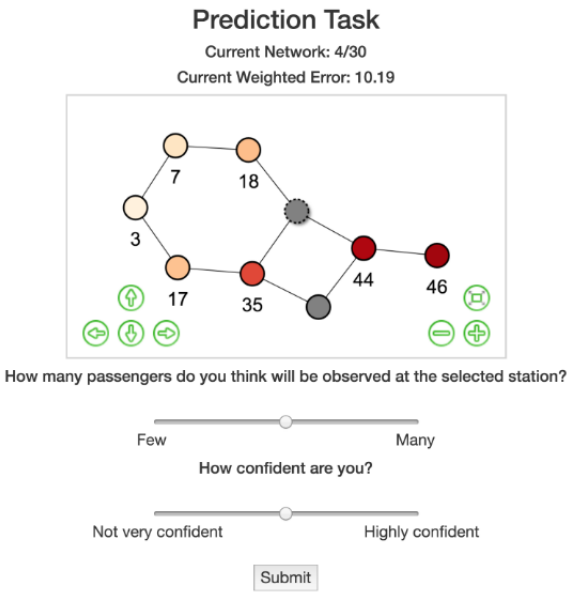

Figure 12: Example of the preparation used in C. M. Wu, Schulz, and Gershman (2020)'s graph learning experiment (taken directly from C. M. Wu, Schulz, and Gershman, 2020)

[61] In a sense, these two paradigms elicit graphical inference in opposite directions (global to local versus local to global)

Though research in network science seems to be progressing rapidly, there are still many unanswered questions regarding how network structure influences human cognition, particularly in the domain of category learning. Given evidence that English-words and semantic concepts embody a network structure (Steyvers and Tenenbaum, 2005; Vitevitch, 2008), and the potential implication of network structure in perceptual segmentation (Saffran, Aslin, and Newport, 1996), it would be interesting to investigate whether humans leverage network structure during the unsupervised formation of categorical knowledge. This might be addressable using a modification to the traditional paradigm for classification learning. Rather than in isolation, stimuli could be presented in pairs and subjects could be tasked with learning about the presence of a relationship between the co-presented stimuli (instead of a category label). Whether subjects leverage network structure for unsupervised categorization could be addressed using a free sort task after learning (Medin et al., 1987).

3.3 Connectionist Approaches to Graph Learning

Learning and inference of graph structured data has become an increasingly popular topic in the machine learning literature (Bronstein et al., 2017). Connectionist models have achieved relative success at tackling very difficult graph learning problems (Graves et al., 2016; Kipf and Welling, 2016; Schlichtkrull et al., 2018). Given connectionism's rich history in psychology (Hummel et al., 2004; McClelland et al., 2010; McClelland et al., 1987)[62], it might be interesting to explore whether geometric deep learning algorithms have any predictive value as models of human reasoning in graph structured domains. At the very least, they suggest that the associative, error-driven mechanisms of traditional connectionist models are a powerful tool for learning and inference on structured and unstructured data. Though as stated earlier, there are many distinct implementations of geometric deep learning algorithms, some of which

[62]Prior to the recent interest in geometric deep learning, there have been many other attempts to integrate structured representations into connectionist models (Pollack, 1990; Smolensky, 1990)

may provide more promising avenues for cognitive scientists than others.

In particular, many variants of geometric deep learning algorithms rely on an aggregation step that collapses over nodes in a neighborhood during training (Z. Wu et al., 2020). This aggregation step places considerable constraints on the types of local neighborhood structures that geometric deep learning algorithms can differentiate (Xu, Hu, Leskovec, and Jegelka, 2018), and prevents models from capitalizing on intra-neighborhood variability. This limitation is circumvented by Veličković et al. (2017), who leveraged the attention mechanisms that have seen recent success in natural language processing (using a model they call *Graph Attention Networks*, or GATs). GATs are a particularly appealing avenue for modeling human reasoning, given that they utilize similar computational mechanisms for both relational reasoning and language comprehension. Another recent approach in applying connectionist learning algorithms to graph structured learning domains comes from work on memory-augmented networks (Graves et al., 2016). Using a combination of memory and attention, Graves et al. (2016) found that a recurrent network trained on randomly generated graphs could efficiently navigate other graph structured learning domains (such as graph representations of subway tunnels or family trees). Interestingly, Graves et al. (2016) suggest preliminary parallels between their model and hippocampus-mediated human memory.

Figure 13: Example of the learning domains that Graves et al. (2016)'s *differentiable neural computer* was able to (arguably) generalize between (taken directly from Graves et al., 2016)

4 Conclusions

A number of demonstrations have provided preliminary evidence that humans are sensitive to the network characteristics of graph structured learning domains (Kahn et al., 2018; Karuza et al., 2019; Lynn and Bassett, 2019; C. M. Wu et al., 2020). Graphs have also been implicated as a representational structure across many different literatures[63] (Abbott et al., 2012; Gentner, 1983; Griffiths et al., 2010; Vitevitch, 2008). Additionally, many puzzling phenomena in the category learning literature seem reasonably intuitive when viewed through the lens of network science. The widespread use of network structures in psychological theories and the recent application of graphs for tackling very difficult cognitive challenges (Graves et al., 2016; Hamrick et al., 2018; Han and Zhu, 2008; Z. Wu et al., 2020) suggest that graphs might be a fundamentally important, cross-domain data structure for cognitive representations.

[63] either as large scale relational systems or as predicate networks

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2012). Human memory search as a random walk in a semantic network. In *Nips* (pp. 3050–3058).

Ahn, W.-K., & Medin, D. L. (1992). A two-stage model of category construction. *Cognitive Science*, *16*(1), 81–121.

Arya, D., & Worring, M. (2018). Exploiting relational information in social networks using geometric deep learning on hypergraphs. In *Proceedings of the 2018 acm on international conference on multimedia retrieval* (pp. 117–125).

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Attneave, F. (1950). Dimensions of similarity. *The American journal of psychology*, *63*(4), 516–556.

Baker, N., Erlikhman, G., Kellman, P. J., & Lu, H. (2018). Deep convolutional networks do not perceive illusory contours. In *Cogsci*.

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS computational biology*, *14*(12), e1006613.

Barrett, S. E., Abdi, H., Murphy, G. L., & Gallagher, J. M. (1993). Theory-based correlations and their role in children's concepts. *Child Development*, *64*(6), 1595–1616.

Bassett, D. S., & Bullmore, E. (2006). Small-world brain networks. *The neuroscientist*, *12*(6), 512–523.

Bassett, D. S., Greenfield, D. L., Meyer-Lindenberg, A., Weinberger, D. R., Moore, S. W., & Bullmore, E. T. (2010). Efficient physical embedding of topologically complex information processing networks in brains and computer circuits. *PLoS computational biology*, *6*(4).

Bhatia, S., Richie, R., & Zou, W. (2019). Distributed semantic representations for modeling human judgment. *Current opinion in behavioral sciences*, *29*, 31–36.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological review*, *94*(2), 115.

Black, M., & Hesse, M. (1966). Models and analogies in science. South Bend, University of Notre Dame Press.

Brady, T. F., & Oliva, A. (2008). Statistical learning using real-world scenes: Extracting categorical regularities without conscious intent. *Psychological science*, *19*(7), 678–685.

Bronstein, M. M., Bruna, J., LeCun, Y., Szlam, A., & Vandergheynst, P. (2017). Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, *34*(4), 18–42.

Bullock, T. H., Bennett, M. V., Johnston, D., Josephson, R., Marder, E., & Fields, R. D. (2005). The neuron doctrine, redux. *Science*, *310*(5749), 791–793.

Butts, D. A., & Goldman, M. S. (2006). Tuning curves, neuronal variability, and sensory coding. *PLoS biology*, *4*(4), e92.

Carey, S. (1985). *Conceptual change in childhood*. MIT press.

Chalmers, D. J., French, R. M., & Hofstadter, D. R. (1992). High-level perception, representation, and analogy: A critique of artificial intelligence methodology. *Journal of Experimental & Theoretical Artificial Intelligence*, *4*(3), 185–211.

Chan, K. Y., & Vitevitch, M. S. (2009). The influence of the phonological neighborhood clustering coefficient on spoken word recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(6), 1934.

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. *Cognitive science*, *34*(4), 685–697.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological review*, *82*(6), 407.

Collins, A. M., Quillian, M. R. et al. (1969). Retrieval time from semantic memory.

Danks, D. (2007). Theory unification and graphical models in human categorization. *Causal learning: Psychology, philosophy, and computation*, 173–189.

Danks, D. (2014). *Unifying the mind: Cognitive representations as graphical models*. Mit Press.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of coling 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 1861–1870).

Doumas, L. A., Hummel, J. E., & Sandhofer, C. M. (2008). A theory of the discovery and predication of relational concepts. *Psychological review*, *115*(1), 1.

Erdogan, G., & Jacobs, R. A. (2016). A 3d shape inference model matches human visual object similarity judgments better than deep convolutional neural networks. In *Cogsci*.

Fay, D., & Cutler, A. (1977). Malapropisms and the structure of the mental lexicon. *Linguistic inquiry*, *8*(3), 505–520.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Forbus, K. D., Ferguson, R. W., Lovett, A., & Gentner, D. (2017). Extending sme to handle large-scale cognitive modeling. *Cognitive Science*, *41*(5), 1152–1201.

Garvert, M. M., Dolan, R. J., & Behrens, T. E. (2017). A map of abstract relational knowledge in the human hippocampal–entorhinal cortex. *Elife*, *6*, e17086.

Gentner, D. (1983). Structure-mapping: A theoretical framework for analogy. *Cognitive science*, *7*(2), 155–170.

Gentner, D., & Holyoak, K. J. (1997). Reasoning and learning by analogy: Introduction. *American psychologist*, *52*(1), 32.

Goldstone, R. L. (1994a). Similarity, interactive activation, and mapping. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 3.

Goldstone, R. L. (1994b). The role of similarity in categorization: Providing a groundwork. *Cognition*, *52*(2), 125–157.

Goldstone, R. L., Medin, D. L., & Gentner, D. (1991). Relational similarity and the nonindependence of features in similarity judgments. *Cognitive psychology*, *23*(2), 222–262.

Goldstone, R. L., & Son, J. Y. (2012). *Similarity.* Oxford University Press.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied linguistics*, *11*(4), 341–363.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... Agapiou, J., et al. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471–476.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in cognitive sciences*, *14*(8), 357–364.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with pagerank. *Psychological Science*, *18*(12), 1069–1076.

Günther, F., Dudschig, C., & Kaup, B. (2016). Latent semantic analysis cosines as a cognitive similarity measure: Evidence from priming studies. *The Quarterly Journal of Experimental Psychology*, *69*(4), 626–653.

Gust, H., Krumnack, U., Kühnberger, K.-U., & Schwering, A. (2008). Analogical reasoning: A core of cognition. *KI*, *22*(1), 8–12.

Guttman, N., & Kalish, H. (1956). Discriminability and stimulus generalization./. exp. *Psychol*, *51*, 79–88.

Hamrick, J. B., Allen, K. R., Bapst, V., Zhu, T., McKee, K. R., Tenenbaum, J. B., & Battaglia, P. W. (2018). Relational inductive bias for physical construction in humans and machines. *arXiv preprint arXiv:1806.01203*.

Han, F., & Zhu, S.-C. (2008). Bottom-up/top-down image parsing with attribute grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *31*(1), 59–73.

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the ieee international conference on computer vision* (pp. 1026–1034).

Hilgetag, C. C., & Goulas, A. (2016). Is the brain really a small-world network? *Brain Structure and Function*, *221*(4), 2361–2366.

Hinton, G. F. (1981). Shape rfprfsfntatton in parallel systems.

Holyoak, K. J., & Cheng, P. W. (2011). Causal learning and inference as a rational process: The new synthesis. *Annual review of psychology*, *62*, 135–163.

Holyoak, K. J., & Gordon, P. C. (1983). Social reference points. *Journal of Personality and Social Psychology*, *44*(5), 881.

Horst, J. S., & Hout, M. C. (2016). The novel object and unusual name (noun) database: A collection of novel images for use in experimental research. *Behavior research methods*, *48*(4), 1393–1409.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological review*, *99*(3), 480.

Hummel, J. E., & Holyoak, K. J. (1997). Distributed representations of structure: A theory of analogical access and mapping. *Psychological review*, *104*(3), 427.

Hummel, J. E., Holyoak, K. J., Green, C., Doumas, L. A., Devnich, D., Kittur, A., & Kalar, D. J. (2004). A solution to the binding problem for compositional connectionism. In *Compositional connectionism in cognitive science: Papers from the aaai fall symposium, ed. sd levy & r. gayler* (pp. 31–34).

Imai, S. (1977). Pattern similarity and cognitive transformations. *Acta Psychologica*, *41*(6), 433–447.

Jäkel, F., Schölkopf, B., & Wichmann, F. A. (2008). Similarity, kernels, and the triangle inequality. *Journal of Mathematical Psychology*, *52*(5), 297–303.

James, W. (1890/1950). *The principles of psychology*. Cosimo, Inc.

Johnson, J., Hariharan, B., van der Maaten, L., Fei-Fei, L., Lawrence Zitnick, C., & Girshick, R. (2017). Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 2901–2910).

Johnson, M. J., Duvenaud, D. K., Wiltschko, A., Adams, R. P., & Datta, S. R. (2016). Composing graphical models with neural networks for structured representations and fast inference. In *Advances in neural information processing systems* (pp. 2946–2954).

Kahn, A. E., Karuza, E. A., Vettel, J. M., & Bassett, D. S. (2018). Network constraints on learnability of probabilistic motor sequences. *Nature human behaviour*, *2*(12), 936–947.

Kang, K., Shapley, R. M., & Sompolinsky, H. (2004). Information tuning of populations of neurons in primary visual cortex. *Journal of neuroscience*, *24*(15), 3726–3735.

Karuza, E. A., Kahn, A. E., & Bassett, D. S. (2019). Human sensitivity to community structure is robust to topological variation. *Complexity*, *2019*.

Karuza, E. A., Kahn, A. E., Thompson-Schill, S. L., & Bassett, D. S. (2017). Process reveals structure: How a network is traversed mediates expectations about its architecture. *Scientific reports*, *7*(1), 1–9.

Kemp, C., Griffiths, T. L., & Tenenbaum, J. B. (2004). Discovering latent classes in relational data.

Kemp, C., Tenenbaum, J. B., Niyogi, S., & Griffiths, T. L. (2010). A probabilistic model of theory formation. *Cognition*, *114*(2), 165–196.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *43*(9), 1470.

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, *10*(11), e1003915.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in systems neuroscience*, *2*, 4.

Kruschke, J. K. (1992). Alcove: An exemplar-based connectionist model of category learning. *Psychological review*, *99*(1), 22.

Kruschke, J. K. (1993). Human category learning: Implications for backpropagation models. *Connection Science*, *5*(1), 3–36.

Kurtz, K. J. (2007). The divergent autoencoder (diva) model of category learning. *Psychonomic Bulletin & Review*, *14*(4), 560–576.

Kurtz, K. J. (2015). Human category learning: Toward a broader explanatory account. In *Psychology of learning and motivation* (Vol. 63, pp. 77–114). Elsevier.

Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep neural networks predict category typicality ratings for images. In *Cogsci*.

Larkey, L. B., & Love, B. C. (2003). Cab: Connectionist analogy builder. *Cognitive Science*, *27*(5), 781–794.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

Lee, J. H., Kim, M. H., & Lee, Y. J. (1993). Information retrieval based on conceptual distance in is-a hierarchies. *Journal of documentation*.

Long, B., Yu, C.-P., & Konkle, T. (2018). Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proceedings of the National Academy of Sciences*, *115*(38), E9015–E9024.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). Sustain: A network model of category learning. *Psychological review*, *111*(2), 309.

Love, B. C., & Sloman, S. A. (1995). Mutability and the determinants of conceptual transformability. In *Proc. 17th annu. conf. cogn. sci. soc* (pp. 65–59).

Lynn, C. W., & Bassett, D. S. (2019). Graph learning: How humans infer and represent networks. *arXiv preprint arXiv:1909.07186*.

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive psychology*, *25*(4), 431–467.

Markman, A. B., & Ross, B. H. (2003). Category use and category learning. *Psychological bulletin*, *129*(4), 592.

McClelland, J. L., Botvinick, M. M., Noelle, D. C., Plaut, D. C., Rogers, T. T., Seidenberg, M. S., & Smith, L. B. (2010). Letting structure emerge: Connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, *14*(8), 348–356.

McClelland, J. L., Rumelhart, D. E., Group, P. R., et al. (1987). *Parallel distributed processing*. MIT press Cambridge, MA:

Medin, D. L., Wattenmaker, W. D., & Hampson, S. E. (1987). Family resemblance, conceptual cohesiveness, and category construction. *Cognitive psychology*, *19*(2), 242–279.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual review of psychology*, *32*(1), 89–115.

Milton, F., Wills, A. J., & Hodgson, T. L. (2009). The neural basis of overall similarity and single-dimension sorting. *Neuroimage*, *46*(1), 319–326.

Minda, J. P., & Smith, J. D. (2001). Prototypes in category learning: The effects of category size, category structure, and stimulus complexity. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*(3), 775.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological review*, *92*(3), 289.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Newman, M. E. (2003). The structure and function of complex networks. *SIAM review*, *45*(2), 167–256.

Nosofsky, R. M. (1991). Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, *23*(1), 94–140.

Nosofsky, R. M. (2011). The generalized context model: An exemplar model of classification. *Formal approaches in categorization*, 18–39.

Nosofsky, R. M., Gluck, M. A., Palmeri, T. J., McKinley, S. C., & Glauthier, P. (1994). Comparing modes of rule-based classification learning: A replication and extension of shepard, hovland, and jenkins (1961). *Memory & cognition*, *22*(3), 352–369.

O'connor, C. M., Cree, G. S., & McRae, K. (2009). Conceptual hierarchies in a flat attractor network: Dynamics of learning and computations. *Cognitive science*, *33*(4), 665–708.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin & Review*, *6*(3), 495–503.

Papo, D., Zanin, M., Martinez, J. H., & Buldu, J. M. (2016). Beware of the small-world neuroscientist. *Frontiers in human neuroscience*, *10*, 96.

Penn, D. C., Holyoak, K. J., & Povinelli, D. J. (2008). Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences*, *31*(2), 109–130.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2016). Adapting deep network features to capture psychological representations. *arXiv preprint arXiv:1608.02164*.

Polk, T. A., Behensky, C., Gonzalez, R., & Smith, E. E. (2002). Rating the similarity of simple perceptual stimuli: Asymmetries induced by manipulating exposure frequency. *Cognition*, *82*(3), B75–B88.

Pollack, J. B. (1990). Recursive distributed representations. *Artificial Intelligence*, *46*(1-2), 77–105.

Porter, M. A. (2012). Small-world network. *Scholarpedia*, *7*(2), 1739.

Pylyshyn, Z. W. (1973). What the mind's eye tells the mind's brain: A critique of mental imagery. *Psychological bulletin*, *80*(1), 1.

Quillian, M. R. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. *Behavioral science*, *12*(5), 410–430.

Rada, R., Mili, H., Bicknell, E., & Blettner, M. (1989). Development and application of a metric on semantic nets. *IEEE transactions on systems, man, and cybernetics*, *19*(1), 17–30.

Rensink, A. (2004). Representing first-order logic using graphs. In *International conference on graph transformation* (pp. 319–335). Springer.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic cognition: A parallel distributed processing approach*. MIT press.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive psychology*, *7*(4), 573–605.

Rosseel, Y. (2002). Mixture models of categorization. *Journal of Mathematical Psychology*, *46*(2), 178–210.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928.

Saffran, J. R., Newport, E. L., Aslin, R. N., et al. (1996). Word segmentation: The role of distributional cues. *Journal of memory and language*, *35*(4), 606–621.

Santoro, A., Raposo, D., Barrett, D. G., Malinowski, M., Pascanu, R., Battaglia, P., & Lillicrap, T. (2017). A simple neural network module for relational reasoning. In *Advances in neural information processing systems* (pp. 4967–4976).

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature neuroscience*, *16*(4), 486.

Schlichtkrull, M., Kipf, T. N., Bloem, P., Van Den Berg, R., Titov, I., & Welling, M. (2018). Modeling relational data with graph convolutional networks. In *European semantic web conference* (pp. 593–607). Springer.

Shepard, R. N. (1962). The analysis of proximities: Multidimensional scaling with an unknown distance function. ii. *Psychometrika*, *27*(3), 219–246.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, *237*(4820), 1317–1323.

Shepard, R. N., Hovland, C. I., & Jenkins, H. M. (1961). Learning and memorization of classifications. *Psychological monographs: General and applied*, *75*(13), 1.

Sloman, S. A. (1998). Categorical inference is not a tree: The myth of inheritance hierarchies. *Cognitive Psychology*, *35*(1), 1–33.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial intelligence*, *46*(1-2), 159–216.

Springer, K., & Keil, F. C. (1989). On the development of biologically specific beliefs: The case of inheritance. *Child development*, 637–648.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, *29*(1), 41–78.

Torralba, A., Tenenbaum, J. B., & Salakhutdinov, R. R. (2011). Learning to learn with compound hd models. In *Advances in neural information processing systems* (pp. 2061–2069).

Turk-Browne, N. B., Jungé, J. A., & Scholl, B. J. (2005). The automaticity of visual statistical learning. *Journal of Experimental Psychology: General*, *134*(4), 552.

Tversky, A. (1977). Features of similarity. *Psychological review*, *84*(4), 327.

Tversky, A., & Gati, I. (1982). Similarity, separability, and the triangle inequality. *Psychological review*, *89*(2), 123.

Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv preprint arXiv:1710.10903*.

Vitevitch, M. S. (2008). What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. *Journal of memory and language*, *67*(1), 30–44.

Wiener-Ehrlich, W. K., Bart, W. M., & Millward, R. (1980). An analysis of generative representation systems. *Journal of Mathematical Psychology*, *21*(3), 219–246.

Wills, A. J., Milton, F., Longmore, C. A., Hester, S., & Robinson, J. (2013). Is overall similarity classification less effortful than single-dimension classification? *The Quarterly Journal of Experimental Psychology*, *66*(2), 299–318.

Wills, A. J., & Pothos, E. M. (2012). On the adequacy of current empirical evaluations of formal models of categorization. *Psychological bulletin*, *138*(1), 102.

Witter, M. (2011). Entorhinal cortex. *Scholarpedia*, *6*(10), 4380.

Wu, C. M., Schulz, E., & Gershman, S. J. (2020). Inference and search on graph-structured spaces. *bioRxiv*.

Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Philip, S. Y. (2020). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Xu, K., Hu, W., Leskovec, J., & Jegelka, S. (2018). How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*.

Zurn, P., & Bassett, D. S. (2020). Network architectures supporting learnability. *Philosophical Transactions of the Royal Society B*, *375*(1796), 20190323.

Appendices

A Details of Network Generation Examples

Steyvers and Tenenbaum (2005)'s Growing Network Model: To examine the impact of density on the properties of large scale networks, the network generation algorithm from Steyvers and Tenenbaum (2005) was used to generate a large set of 24 networks of node size 200. The algorithm starts off by generating a fully connected set of M nodes. After initialization, a new node is added to the network. The new node is connected to M neighbors of an already existing node i. The existing node i is selected from a distribution:

$$P_i = \frac{k_i}{\sum k_n}$$

where k_i is the degree of a random node and $\sum k_n$ is the sum of the degrees of all nodes in the network. The M connections between the new node and the neighborhood of node i are sampled uniformly. This repeats until the algorithm reaches the desired network size.