Computational Brain & Behavior
https://doi.org/10.1007/s42113-018-0003-7

Estimating Semantic Networks of Groups and Individuals from Fluency Data

Jeffrey C. Zemla[1] · **Joseph L. Austerweil**[1]

Abstract

One popular and classic theory of how the mind encodes knowledge is an associative semantic network, where concepts and associations between concepts correspond to nodes and edges, respectively. A major issue in semantic network research is that there is no consensus among researchers as to the best method for estimating the network of an individual or group. We propose a novel method (U-INVITE) for estimating semantic networks from semantic fluency data (listing items from a category) based on a censored random walk model of memory retrieval. We compare this method to several other methods in the literature for estimating networks from semantic fluency data. In simulations, we find that U-INVITE can recover semantic networks with low error rates given only a moderate amount of data. U-INVITE is the only known method derived from a psychologically plausible process model of memory retrieval and one of two known methods that we found to be consistent estimators of this process: if semantic memory retrieval is consistent with this process, the procedure will eventually estimate the true network (given enough data). We conduct the first exploration of different methods for estimating psychologically valid semantic networks by comparing people's similarity judgments of edges estimated by each network estimation method. To encourage best practices, we discuss the merits of each network estimation technique, provide a flow chart that assists with choosing an appropriate method, and supply code for others to employ these techniques on their own data.

Keywords Knowledge representation · Semantic networks · Methodology · Fluency · Bayesian modeling

Introduction

How do people encode, store, and retrieve knowledge? Psychologists have examined memory retrieval using different tasks as a means of understanding mental representation (Tulving 1972). Using different methodologies, researchers have explored how concepts are organized within a specialized type of memory, known as semantic memory (Quillan 1966). Although significant progress has been made, placing constraints on models of semantic memory, basic questions concerning the nature of mental representation in semantic memory remain unresolved (Griffiths et al. 2007; Johnson-Laird et al. 2015; Jones et al. 1984, in press; Tversky 1977; Tversky and Hutchinson 1986).

Researchers have proposed different, plausible representations for semantic memory. One popular type of representation is a spatial representation, which represents each concept as a point in Euclidean space. Similarity of concepts (and other psychological functions) are estimated from the distance between the two corresponding points (Attneave 1950). Advances in computing power, mathematical tools, and our psychological understanding of memory have resulted in more sophisticated spatial models of memory in recent decades (e.g., Jones and Mewhort 2007; Landauer and Dumais 1997). Spatial models are grounded in psychological theories, such as the Bayesian framework for generalization (Shepard 1987; Tenenbaum and Griffiths 2001).

Spatial models are used often in psychology, though frequently the decision to use this semantic representation is motivated by convenience, rather than psychological validity. In many ways, this is a reasonable choice: the

✉ Jeffrey C. Zemla
zemla@wisc.edu

Joseph L. Austerweil
austerweil@wisc.edu

[1] Department of Psychology, University of Wisconsin-Madison, 1202 West Johnson Street, Madison, WI 53706, USA

Published online: 06 June 2018

underlying computational components of many spatial model estimation techniques (such as multidimensional scaling) are well understood and widely used, making interpretation of these models attractive (Kruskal and Wish 1978; Shepard 1964, 1980).

In addition, the research community has created best practices (Davison et al. 2010), tools, and code for estimating semantic spaces in a domain (e.g., Busing et al. 1997; Dennis 2007) which has resulted in their widespread use. Semantic spaces can be derived from a variety of data sources, including semantic fluency data (Chan et al. 1993), text corpora (Landauer and Dumais 1997), paired similarity ratings (Dry and Storms 2009), and triadic comparisons (Lee et al. 2016), making spatial representations broadly applicable. Regardless of the input data source or the particular spatial model employed, the basic principle remains: semantic concepts are represented by points in multidimensional space, and association strength is designated by the distance between two points.

In parallel, semantic networks have been proposed as an alternative to spatial memory representations (Collins and Loftus 1975; De Deyne and Storms 2008; Sattath and Tversky 1977; Steyvers and Tenenbaum 2005). Unlike spatial models, semantic networks represent semantic memory as a structured network in which concepts (nodes) are connected to semantically similar concepts by edges.[1] With recent computational advances in network science (Albert and Barabási 2002; Watts 2004), there has been a resurgence of interest and use of semantic networks in the study of semantic memory (Baronchelli et al. 2013; Falk and Bassett 2017).

Across the semantic network literature, there is little consistency in how the semantic networks of individuals or groups are estimated. Arguably, the most common approach is to estimate group-level networks from free association data, in which a large group of participants are given a set of cue words and are asked to respond with the first word that comes to mind. A network is formed by connecting each cue-response pair with an edge. As a result, networks generated from free association data will include a set of experimenter selected cue words.

The free association task places few constraints on participant responses, which can make it difficult to infer category-level semantic networks (e.g., one that contains only animals) from the data. Often, responses to cue words are semantically related but span categories (e.g., "paw" as a response to "dog"). However, responses may also be semantically unrelated (De Deyne and Storms 2008), such as word completions ("fish" as a response to "star") or rhymes ("yarn" as a response to "barn"). Since responses are largely unconstrained, it can take a substantial amount of free association data to estimate a category-level semantic network. This is especially true for generating individual-level networks (see Morais et al. 2013).

To estimate the semantic organization of a particular domain in an efficient manner, many researchers have begun to use semantic fluency data to estimate group-level networks. In the semantic fluency task (Bousfield and Sedgewick 1944; Henley 1969), participants generate as many items from a given category as they can in a fixed period of time (e.g., "list as many animals as you can in three minutes"). A characteristic pattern seen in semantic fluency data is clustered responses: participants tend to list semantically similar items in close proximity to each other (Gruenewald and Lockhead 1980; Troyer et al. 1997). For example, a participant may list the animals "*dog, cat, hamster*" in sequence because all of these animals belong to a common sub-category *pets*. This tendency to cluster items together tells us something about how concepts in a category are mentally organized. For example, in the animal category, sequences tend to be clustered based on a common environment (e.g., household pets, African animals) and taxonomy (e.g., canine, amphibians) more so than other possible relations such as visual features (e.g., red animals, small animals) or causal relations (e.g., predator-prey).

Responses that frequently occur near each other (across many fluency lists) are likely to be semantically related, and a semantic network can be constructed from these inferred associations. To date, however, there are no best practices for how a researcher should estimate networks from fluency data, no easy tools for estimating networks from fluency data, no empirical evaluation of network estimation techniques, and no methods derived from psychological models of the task. The goal of this article is to take a step towards addressing these gaps in our current understanding of how to estimate semantic networks from fluency data.

One process that generates clustered responses and mimics semantic fluency data is a random walk on a semantic network (Abbott et al. 2015; Goñi et al. 2010; Zemla and Austerweil 2017). In this article, we develop and validate a novel method for estimating semantic networks that reflects this process. Our model assumes that fluency data is generated by a censored random walk on a semantic network and estimates the most likely network to have generated these data. Building on a related technique (Jun et al. 2015), we call this model U-INVITE (Unweighted Initial Visit Emitting random walk).

[1] Formally, there may exist a function that converts between equivalent semantic spaces and networks (Anderson 1978). However, the choice of representation will usually be psychologically important. For instance, one representation may provide a more efficient coding than another. Likewise, different representations may require different retrieval processes to reproduce the same human data. That is, though the two representations may encode the same information, the parsimony and psychological plausibility of the representations may differ (Jones et al. 2015).

We validate this model in two ways. First, we show that an artificial network can be efficiently recovered from simulated fluency data generated by a censored random walk. Second, we compare how networks estimated by U-INVITE from human semantic fluency data compared to networks estimated by other methods using the same fluency data. We generate seven semantic networks for the animal category using seven techniques: U-INVITE, a hierarchical Bayesian version of U-INVITE designed to capture individual variation appropriately, Pathfinder Networks (PN), Community Networks (CN), Naïve Random Walk (NRW) networks, First Edge (FE) networks, and Correlation-Based Networks (CBN). We then evaluate the semantic similarity of edges in each network using a separate group of participants who rated the similarity of animal pairs. We conclude with a discussion of advantages and disadvantages of each network estimation technique, and we provide a flow chart (Fig. 10) to help researchers choose a network estimation technique that is most appropriate for their research goals.

The goal of this article is to advance our theoretical and practical understanding of methods for estimating semantic networks from human fluency data and propose best practices. We do not attempt to adjudicate whether semantic spaces or networks are better representations of semantic memory. Techniques for estimating semantic spaces are relatively mature and have been theoretical and empirically validated. In contrast, we have little understanding of the theoretical and empirical validity of methods for estimating semantic networks. Without a reliable understanding of how to estimate semantic networks, the debate between semantic spaces or networks will remain unanswerable; there will always remain a lingering question as to whether the method used to estimate a semantic network is appropriate, muddying any comparison between semantic networks and spatial models. The results and conclusions of this article are steps towards providing the necessary theoretical and empirical understanding of techniques for estimating semantic networks.

Semantic Networks

A semantic network is a representation of memory that describes the organization of declarative facts and knowledge in the mind. A network consists of a set of nodes and a set of edges. Each node in the network denotes a concept in semantic memory, such as *fish* or *purple*. Edges in a network are used to connect a pair of nodes that are semantically similar. For example, an edge might connect *plane* and *car* because both are vehicles. The set of nodes directly connected to a node by an edge are called its neighbors.

Often, the semantic networks used in psychology are agnostic as to the specific semantic relation between nodes that edges denote. Some possible semantic relations include causal relations (e.g., *Causes(moon, tides)*), featural similarity (e.g., *Similar_Color(sky, water)*), subordinate or superordinate relations (e.g, *Is_A(oak,tree)*), or temporal co-occurrence (e.g., *Precedes(Rain, Lightning)*). While some semantic networks make the relation type between nodes explicit (e.g., Stella et al. 2017), many leave it implicit (e.g., Steyvers and Tenenbaum 2005).

While the only essential features of a semantic network are a set of nodes and a set of edges, many networks have additional features. For example, edges in a network can be directed or undirected. A directed edge can be used to indicate that a semantic relation is not symmetric; for instance edges can be used to describe category membership (e.g., a directed edge may indicate a *hammer* is a type of *tool*, but not vice versa). Edges may also be weighted or unweighted to denote the strength of similarity between two concepts. For example, *ketchup* and *mustard* are strongly related (both are condiments) and might have a high edge weight, while *ketchup* and *taco* are weakly related (both are foods) and might have a low edge weight.

Alternatively, some aspects of directionality and weightedness can be implicitly coded in the network structure, rather than explicitly defined. For example, *fruit* may have many neighbors, whereas *apple* may have relatively few. As such, simple models of lexical retrieval based on spreading activation or random walks would generally predict that *fruit* should be primed more when given *apple* than vice versa. Similarly, random walk models of retrieval predict asymmetry in path lengths between two nodes, without having the need to explicitly assign parameters related to edge weight and direction. So, it is possible for unweighted, undirected networks with a simple retrieval process to capture some asymmetries in priming and similarity rating that are typically used to argue for the necessity of weighted, directed networks. An in-depth analysis is needed to assess whether weighted, directed networks are needed to capture human semantic retrieval and judgments.

Regardless of whether this information is coded directly in the representation, Goñi et al. (2011) notes that inter-rater agreement of category boundaries (i.e., whether two items are part of the same category) is quite high. That is, even if similarity in a semantic network is graded, people share similar intuitions about how similar two items must be to considered semantically similar. As such, semantic networks that are unweighted or undirected are not necessarily incompatible with an analogous network that is weighted and directed.

When estimating networks from semantic fluency data, it may be preferable to estimate unweighted and undirected networks. Estimating directionality and weightedness in a

network can vastly increase the number of parameters of the network. This makes network estimation difficult when data is sparse, which is typical of semantic fluency data. Estimating networks from a small amount of data can lead to noisy parameter estimation. For instance, Jun et al. (2015) estimates undirected, weighted networks from fluency data but typically performs poorly with small data sets (e.g., less than 100 lists). Additionally, the networks estimated by Jun et al. (2015) are fully connected (in which all node pairs are connected by a weighted edge), which are not typical of those used in psychology, and bear some resemblance to a low-dimensional spatial model.

Semantic networks are often defined in part by their topology. Much of the early research in semantic networks neither proposed nor identified any particular network topology as being necessary for representing semantic memory (Collins and Loftus 1975), though recent work has identified that network topology is critical to modeling behavior in memory retrieval tasks. Semantic networks tend to be small-world like (Morais et al. 2013; Steyvers and Tenenbaum 2005) and/or scale-free (De Deyne and Storms 2008; Steyvers and Tenenbaum 2005). In a small-world network, a small number of "hub" nodes have a very large number of neighbors, while most nodes in the network have a small number of neighbors. This topology allows rapid retrieval of information within a network with minimal wiring costs (i.e., relative few edges; Bassett and Bullmore 2006). A network's topology has important implications for how information is retrieved. For most proposed retrieval processes, a network's topology can have a dramatic influence on the way memory is searched (i.e., the order in which concepts are explored) and the time needed to locate information within a network.

Semantic Fluency Task

A popular way to estimate semantic networks for a domain of interest is using the semantic fluency task (e.g., Goñi et al. 2011). In the semantic fluency task, participants are asked to name as many items from a given category (e.g., animals, tools, foods) as they can in a fixed length of time (usually one to three minutes). Participants are allowed to report items at any level of abstraction, such as the basic-level (e.g., *dog*) or subordinate-level (e.g., *poodle*). Participants are told not to repeat themselves, but usually given no guidance beyond this.

In the traditional paradigm, each participant generates only a single fluency list. However, in a variation called the repeated fluency task (Zemla et al. 2016), each participant performs the task multiple times, generating several fluency lists from the same category. Items can be repeated from list to list, but participants are still told to avoid repeating an item within a single list. In the repeated fluency task, one or more filler tasks is included in between lists to minimize repetitions and sequences resulting from list memory and recency effects. The filler task may be a semantic fluency task for a different category, or a completely unrelated task. Alternatively, repeated fluency data can be collected longitudinally, collecting lists from a single participant every few weeks (or longer).

Data from the semantic fluency task show stereotypical patterns. Healthy participants are generally good at avoiding intrusions (listing non-category items) and perseverations (repetitions) within a list, although the task is also widely used with patients who have semantic memory deficits, who tend to have intrusions and perseverations (Shindler et al. 1984). Participants also show typicality effects, in which common or prototypical category members—such as *dog* of the category animals—are reported more often and earlier than non-typical category members, such as *aardvark* (Quaranta et al. 2016).

Another hallmark pattern of semantic fluency data is clustering: participants tend to list sequences of items that belong to a common cluster (i.e., a sequence of *pets* when listing animals). Further, participants show a tendency to switch clusters at theoretically optimal times (Hills et al. 2012), consistent with the marginal value theorem from the optimal foraging literature (Charnov 1976). Consistent with this theory, participants are able to quickly list items when they begin a new cluster. However, when the time needed to retrieve the next item in that cluster exceeds the global average time needed to to retrieve a new category item (of any cluster), participants tend to switch to a new cluster—the optimal time to switch.

These patterns have been taken as evidence indicating that the mental representation for categories such as animals is patchy: semantically, similar clusters of animals are stored nearby in mental space. An ongoing debate surrounds what semantic representation and retrieval process best explains these patterns. Traditionally, semantic fluency data has been explained as the result of a two-stage retrieval process, where participants switch between a global cue (e.g., animals) and local cue (e.g., pets) in a spatial representation when searching for category members (Hills et al. 2012). However, it has also been shown that semantic fluency data can be explained by a single-stage retrieval process on a different semantic representation: a censored random walk on a semantic network (Abbott et al. 2015). Under this model, fluency data is generated by traversing edges at random in a semantic network. When a novel item is traversed, it is output (i.e., added to the fluency list). When the same item is traversed a subsequent time, the item is censored (i.e., not added to the fluency list).

The semantic networks used by Abbott et al. (2015) were pre-defined by the researchers (they showed two sensible

networks both produced patchy behavior). How should a researcher choose one semantic network or another? To investigate this question, we evaluate different network estimation methods given fluency lists of animals. Psychologists have studied human knowledge representation for animals extensively, making it an ideal candidate for evaluating different techniques. In the next section, we discuss different methods for estimating a semantic network for a domain.

Previous Work: Estimating Semantic Networks

Several approaches have been used previously to estimate semantic network representations from psychological data or corpora. One method for constructing a semantic network relies on structured databases. For example, a network can be formed by connecting synonyms in a thesaurus (Steyvers and Tenenbaum 2005); connecting related words in a lexical database such as WordNet (Miller 1995; Steyvers and Tenenbaum 2005); or using a network formed by hyperlinks in an interactive encyclopedia such as Wikipedia (Navigli and Ponzetto 2012). These databases have been tailored with human oversight. Another method for constructing networks is to analyze a large text corpus. For instance, one can estimate which concepts are related by counting word co-occurrences in an article, or using other related natural language processing techniques (Masucci et al. 2011).

Another approach to constructing semantic networks is to use psychological data from behavioral experiments in which participants retrieve items from memory or perform similarity judgments. In particular, the free association task is used frequently to construct networks (De Deyne and Storms 2008; Morais et al. 2013; Nelson et al. 2004). In the free association task, participants are given a cue word and asked to respond with the first word that comes to mind. For instance, if the cue word is *dog*, a participant might respond with *cat* or *wolf*. Typically, only a single response is solicited for each cue word, though soliciting multiple responses can result in more rich networks (De Deyne and Storms 2008). A network is constructed by forming an edge between each cue and response pair. Although this remains a useful method, it is unrestrained. Responses can (and do) cross category boundaries (e.g., an acceptable response to *dog* is *bone*), which can result in very broad networks. As a result, it can require a very large number of responses and participants to estimate a category representation (e.g., an animal semantic network), and the resultant network may underestimate the density of within-category associations. In addition, the set of cue words chosen by the experimenter can bias the content and topology of the network.

Aside from the free association task, the semantic fluency task is another widely used paradigm for collecting psychological data used to infer semantic networks. One reason for this is that it enables a researcher to focus on a particular category of interest. In the sections below, we describe several computational methods for estimating networks from fluency data.

First Edge

The First Edge (FE) method (Abrahao et al. 2013; Jun et al. 2015) estimates a network by inferring a single edge from each fluency list. The first and second response generated in each list are connected by an edge; the rest of the list is discarded and not used for inference.

Because the First Edge method estimates only a single edge from each list, it is not an efficient means of constructing semantic networks from fluency data. However, because the first pair of items in a fluency list is often strongly related, the model provides an important benchmark to compare with other methods. In addition, the First Edge method is a statistically consistent estimator of a semantic network from fluency data assumed to have been generated by the censored random walk. This means that given enough fluency lists generated by a censored random walk, the First Edge method will recover the true network.

Naïve Random Walk (NRW) and Thresholding Models

The Naïve Random Walk (Jun et al. 2015; Lerner et al. 2009) estimates a network under the assumption that each fluency list is generated by an *uncensored* random walk on a semantic network. That is, it assumes every time a node is traversed in the random walk, the node appears as a response in the fluency list, on both the first traversal of that node *and* on subsequent traversals. An edge is inferred between each pair of adjacent items in each fluency list. Because adjacent items are typically more similar than non-adjacent items, the Naïve Random Walk is a quick and effective method for estimating a semantic network.

A major limitation of the Naïve Random Walk is that edges are inferred in a binary fashion. A single co-occurrence of two items is sufficient to infer an edge in the network, and no amount of additional data will reject that edge. As such, the Naïve Random Walk can be efficient and effective when given a very small number of lists but is not effective at constructing networks from large amounts of data. Under the assumption that all possible item pairs can appear in fluency data, the Naïve Random Walk will result in a network that is fully connected when given enough data.

Often, there are some adjacent pairs of items in fluency data that are not semantically similar, either by chance or

because they span a cluster switch boundary. One method for pruning these edges is to apply a threshold procedure to a network constructed using a Naïve Random Walk (Lerner et al. 2009). A threshold model generalizes the Naïve Random Walk model by estimating an edge between any adjacent pair of items in a fluency list if that pair appears more than a fixed number of times (T_n) or fixed proportion of times (T_p) across all fluency lists. Higher threshold values result in sparser networks (fewer edges) whose edges typically reflect stronger similarity than edges inferred with lower thresholds.

Community Network

The Community Network (CN) approach (Goñi et al. 2011) extends the threshold model by providing a principled approach for filtering out spurious edges that emerge from large data sets. Unlike First Edge and Naïve Random Walk, the Community Network also allows inference of edges between non-adjacent items.

The Community Network method estimates a semantic network using a co-occurrence matrix C that has been generated from a set of fluency lists, where C_{ij} usually denotes the number of times items i and j appear within a span of w responses across all fluency lists. w is free parameter that denotes a window size (see Fig. 1). As the window size w is increased, the method can infer edges between items that appear further apart within a fluency list. The Community Network infers an edge between any pair of items i and j when C_{ij} co-occurrences are significantly unlikely to occur by chance alone, given the total number of fluency lists M and the frequency f_i of an item i in the data set. Because it is difficult to reliably estimate significance with a low number of co-occurrences, a lower-bound threshold T_n is sometimes applied to the co-occurrence matrix. For each pair of items i and j,

$$C_{ij} = \begin{cases} \text{\# of times } i \text{ and } j \text{ co-occur within window } w \text{ across all lists} & \text{if \# of co-occurences} > T_n, \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where T_n is typically 1. In other words, for an edge to be inferred between a pair of items, the pair must co-occur

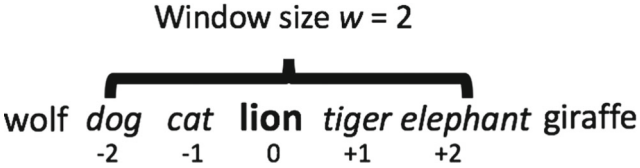

Fig. 1 Co-occurrences are calculated by how many times two items appear within an arbitrary window size w. Here, lion co-occurs with dog, cat, tiger, and elephant, but not wolf or giraffe

within window w more than once. The probability of two items i and j occurring in the same list within window size w by chance alone is given by:

$$P_{ij}^{linked} = P_{ij}^{list} P_{ij}^{(<W)} \quad (2)$$

where P_{ij}^{list} is the probability of i and j appearing in the same list:

$$P_{ij}^{list} = \frac{f_i f_j}{M^2} \quad (3)$$

and $P_{ij}^{(<w)}$ is the probability of i and j appearing within window size w:

$$P_{ij}^{(<w)} = \frac{2}{l(l-1)} \left(wl - \frac{w(w+1)}{2} \right) \quad (4)$$

where l is the mean length of the fluency lists. See Goñi et al. (2011) for more details.

For each pair of items i and j in the data, a binomial 95% confidence interval is computed using the Clopper-Pearson exact method (Clopper and Pearson 1934) given the number of fluency lists M and the number of co-occurrences C_{ij}. An edge is inferred for any item pair when P_{ij}^{linked} is less than the lower bound of the confidence interval, indicating the number of observed co-occurrences is greater than expected from chance alone. The size of this confidence interval can be adjusted to control the sparsity of the network; wider confidence intervals generate sparser networks but whose edges are likely to have greater semantic similarity.

Following Goñi et al. (2011), we use a threshold T_n of 1 and a window size w of 2 in the simulations below.

Pathfinder Network

Paulsen et al. (1996) propose one of the earliest methods for constructing semantic networks from fluency data. Following Chan et al. (1993),[2] they define a distance metric used to measure the similarity between any pair of items in a set of fluency lists. In contrast to many of the other techniques presented here, the Pathfinder method defines a distance score for every pair of items in the data set that decreases as two items appear closer together in a list and as they co-occur across multiple lists (indicating stronger semantic similarity). That is, there is no explicit restriction on how near two items must appear in a fluency list for them to be connected with an edge in a network.

[2]Chan et al. (1993) initially introduced the equations in this section for deriving a proximity matrix from fluency data, but they do not use Pathfinder to estimate semantic networks as we construe them. Conversely, Chan et al. (1995b) uses Pathfinder to estimate semantic networks from a proximity matrix, but use triadic comparison data rather than semantic fluency data to build a proximity matrix. To the best of our knowledge, Paulsen et al. (1996) are the first to perform both steps of this procedure.

They define

$$D_{ij} = \frac{M}{M_{ij}^2} \sum_{k=1}^{M_{ij}} \frac{d_{ijk}}{N_k} \tag{5}$$

where:

- D_{ij} is the computed distance between items i and j
- d_{ijk} is the distance between items i and j in list k (i.e., the number of items between i and j for list k, plus one)
- N_k is the total number of items in list k
- M_{ij} is the total number of lists that contain both items i and j
- M is the total number of lists

Following this procedure for every pair of responses in the data set yields an N by N proximity matrix, where N denotes the number of unique responses in the data set. Paulsen et al. (1996) construct a semantic network from this matrix using the Pathfinder method for generating unweighted, undirected networks from proximity matrices (Schvaneveldt 1990; Schvaneveldt et al. 1989).

In a Pathfinder network, only the path with the shortest distance between every pair of nodes in the proximity matrix is preserved.[3] The path with the shortest distance between nodes i and j may not contain a direct edge between i and j. Rather, a path between intermediary nodes may traverse the network with a shorter distance. A Pathfinder network is constrained by two parameters, q and r. q constrains the maximum number of steps that can occur in the shortest path between two nodes. When $q = N - 1$ there are no constraints on the maximum path length connecting two nodes. When calculating path lengths, the Minkowski distance between two nodes is used (a generalization of Euclidean and Manhattan distance), parameterized by exponent r. When $r = 1$, the Minkowski distance is identical to Manhattan distance, or the sum of all path segment lengths. When $r = \infty$, the distance metric is identical to Chebyshev distance, in which a path length is equal to the largest of those path segment lengths.

q and r are free parameters that influence the density of the network. As either parameter is increased, the density of the estimated network is decreased. When $q = N - 1$ and $r = \infty$ (as in Chan et al. 1995b; Razani et al. 2010), Pathfinder extracts the sparsest possible network (i.e., the fewest number of edges) from the proximity data. Other parameterizations can extract the densest possible network (e.g., Chan et al. 1995a) or something in between (e.g., Vinogradov et al. 2003).

We implement the Pathfinder estimation technique with parameters $q = N - 1$ and $r = \infty$ to extract the sparsest possible network from the data. This parameterization is computationally efficient compared to other parameterizations, and corresponds to the union of all minimum spanning trees (Quirin et al. 2008). A minimum spanning tree is a set of edges that connects all items in a network (N) while minimizing the total distance (i.e., $\sum_{i,j} D_{ij}$) of all edges in the network.

Correlation-Based Networks

Another method for constructing networks estimates edges between items based on how many lists the items co-occur in, regardless of their relative positions in that list (Borodkin et al. 2016; Kenett et al. 2013). Initially, an M by N matrix F is constructed, where M indicates the total number of lists and N indicates the total number of unique items produced in those lists. Each cell F_{ij} in the matrix takes on a binary value of 1 or 0 to indicate whether item i appeared in list j. Thus, the sum of the values in F is equal to the total number of responses given (excluding perseverations).

Pairwise Pearson correlations are computed for each column pair (i.e., each possible pair of items). A high correlation indicates that two items are likely to co-occur within any given list. The set of all item pairs, representing all possible edges in the network, are sorted by their correlations from high to low. A semantic network is constructed from this list by adding an edge for each pair (in descending order) to the network so long as the resultant network remains planar (i.e., the network can be drawn on a 2D plane such that no two edges intersect). This filtering procedure is equivalent to that of constructing a planar maximally filtered graph (Tumminello et al. 2005). Planar graphs have the nice property of being easy to visualize in two-dimensions, but it is unclear whether this restriction has any psychological validity.

This method is typically used to estimate group-level semantic networks, when the number of fluency lists is large. When only a small number of fluency lists are used (e.g., 3 to 5), the results can be unstable because pairwise correlations with small data sets are biased to extreme values, and are often undefined (i.e., if two items always or never co-occur in a list).[4] Also, there tend to be many "ties" (identical correlations) that are produced.

[3] Sometimes, two items do not co-occur in any list and thus the distance between them is infinite or undefined. In these cases, an edge between the two items is omitted. It is not clear from the literature if this is standard practice.

[4] To mitigate this problem, a thresholding procedure is sometimes applied (as in Kenett et al. 2013) to only include words that appear in at least T_n lists. In our simulations (Section "Model Validation"), we found that applying a small threshold of $T_n = 2$ produced worse results, and so we report only the model results with no threshold.

Our implementation omits an edge whenever a pairwise correlation is undefined and does not use any principled secondary criteria to further sort the edge list when ties occur.

U-INVITE

The network estimation methods described so far work under the assumption that items that co-occur frequently in the same list or in close proximity to each other are likely to be semantically related. This assumption is justified based on the finding that fluency data tends to be clustered. The estimated networks tend to be evaluated by qualitative inspection. However, none of the techniques presuppose psychologically plausible models for producing fluency data. Instead, they rely on generic statistical heuristics that identify clusters in the data.

Recent work in psychology has identified several potential process models (Abbott et al. 2015; Hills et al. 2012; Zemla and Austerweil 2017) that generate clustered fluency data. We propose a technique that capitalizes on these advances by inverting a process model used to simulate fluency data—specifically, the censored random walk model of search on a semantic network (Abbott et al. 2015).

We call this method U-INVITE, building on a previous model for constructing fully connected, weighted, and asymmetric semantic networks from fluency data generated by a censored random walk known as INVITE (Jun et al. 2015). In contrast to INVITE, U-INVITE produces networks with binary non-directional edge weights (0 or 1), analogous to other methods described above, and produces much sparser networks that are not fully connected.[5] This restriction facilitates estimating networks given small data sets by constraining the possible weights for an edge to two values: zero or one. It comes at the expense of being less expressive. This is an example of the bias-variance dilemma, where we have introduced bias to enable estimation from small sample sizes (Geman et al. 1992; Griffiths 2010).

Estimating Semantic Networks with U-INVITE

The generative model of U-INVITE (shown in Fig. 2) proposes that semantic fluency data can be modeled as a censored random walk on a semantic network (Abbott et al. 2015; Zemla and Austerweil 2017). Given a network, the initial item in a list is chosen stochastically based on the prior probability of a random walk encountering each item in the network (i.e., the first item is chosen proportionally to the number of neighbors connected to each node). Subsequent list items are produced by following a random walk on the network. A traditional random walk produces many duplicate responses, because it traverses over many nodes more than once. To avoid repetitions in a fluency list, an additional constraint is applied: after a node is traversed for the first time, subsequent traversals over that node are censored (unobserved). This censoring process is analogous to monitoring processes in lexical retrieval (e.g., Levelt et al. 1999). Because of this censoring process, two items that appear adjacent in a censored fluency list are not necessarily directly connected by an edge in the network (see Fig. 3). This results in fluency data that resembles the traditional cluster-and-switch phenomenon, because not all observed transitions in a fluency list share a semantic relation, as they would in an uncensored random walk.

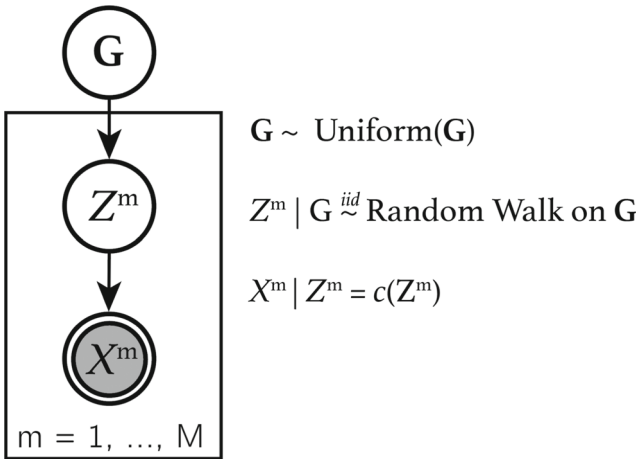

Fig. 2 Z denotes a set of independent uncensored random walks on network $\mathbf{G}$. A censoring function $c(x)$ is applied to each walk to produce a set of censored random walks X. This censoring function is deterministic, and X denotes the observed (fluency) data

Assuming we have a set X of M fluency lists, where each list has been generated by a censored random walk on a semantic network, our objective is to calculate the most likely semantic network $\mathbf{G}$ that could have produced those lists, $P(\mathbf{G} \mid X)$. Here, we represent $\mathbf{G}$ as an N by N binary and symmetric link matrix, where N denotes the total number of unique responses across all fluency lists. Assuming a uniform distribution over $\mathbf{G}$ and using Bayes' rule, the most likely semantic network $\mathbf{G}$ is the one that maximizes the likelihood of the data:

$$\mathbb{P}(X^1,\ldots,X^M|\mathbf{G}) = \prod_{m=1}^{M} \mathbb{P}\left(X_1^m\right) \prod_{k=1}^{N_m-1} \mathbb{P}\left(X_{k+1}^m|X_{1:k}^m\right) \quad (6)$$

where N_m denotes the number of items in the mth censored list, X^m denotes the mth censored list, and X_k^m denotes the

[5] Note that we refer to networks as connected if there exists a path between every pair of nodes in the network. The largest connected component is the largest subset of nodes in the network where there exists a path between every pair of nodes. We use fully connected to mean that every pair of nodes is connected with an edge.

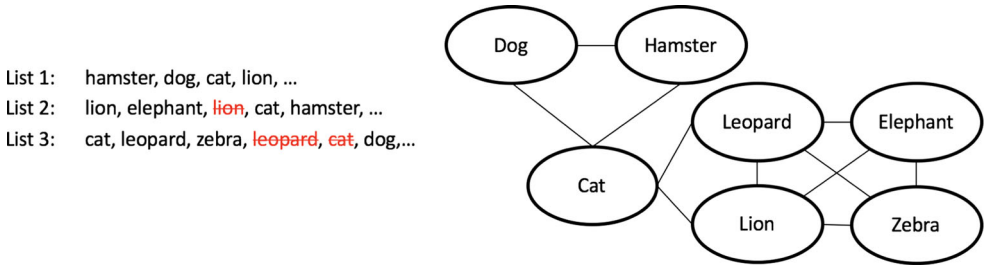

Fig. 3 Each list (left) denotes a random walk on a semantic network (right). Items that appear more than once in a random walk are censored (shown here in red/strikethrough) on all but their first occurrence. Censored random walks on a semantic network have been used to model human semantic fluency data

kth item from the mth censored list. That is, the likelihood of a semantic network is the product of all transition probabilities across all lists, multiplied by the product of all initial response probabilities (the probability of a random walk having started at the first items in each list). For a connected, undirected, unweighted network, the probability of starting list m with item X_1^m is proportional to the number of edges connected to that node.

$$\mathbb{P}\left(X_1^m\right) = \frac{\text{\# of neighbors of } X_1^m \text{ in } \mathbf{G}}{\text{Total \# of edges in } \mathbf{G}} \tag{7}$$

The key to U-INVITE is to calculate transition probabilities in the data by treating each transition as an independent absorbing walk over a state space of nodes, where previously traversed nodes (i.e., $X_{1:k}^m$ in Eq. 6) are treated as transient, and all other nodes are treated as absorbing nodes (Jun et al. 2015; Zemla et al. 2016). Specifically, a transition probability $\mathbb{P}(X_{k+1}^m|X_{1:k}^m)$ is equivalent to the probability of a walk starting at node X_k^m and being absorbed by node X_{k+1}^m where nodes $X_{1:k}^m$ are transient and all unobserved nodes are absorbing.

First, we translate the network $\mathbf{G}$ into a transition probability matrix $\mathbf{A}$ (of the same dimensions), where

$$\mathbf{A}_{ij} = \frac{\mathbf{G}_{ij}}{\sum_{k=1}^{N} \mathbf{G}_{ik}} \tag{8}$$

For each transition probability that we calculate, we reorder this transition matrix so that the rows and columns are arranged in list order. This forms a new matrix, $\mathbf{A}'$. Items that do not appear in a given list are excluded from $\mathbf{A}'$.[6] The matrix is then subdivided into quadrants:

$$\mathbf{A}' = \begin{pmatrix} Q & R \\ 0 & I \end{pmatrix} \tag{9}$$

Q denotes transitions between transient nodes and R denotes transitions from transient to absorbing nodes. To ensure the walk is absorbing, the lower two quadrants of the matrix are replaced with 0 (a matrix of zeros) and I (the identity matrix). We can then calculate a transition probability as:

$$\mathbb{P}\left(X_{k+1}^m|X_{1:k}^m\right) = \begin{cases} \sum_{i=1}^{k} E_{ki} R_{i1} & \text{if } E \text{ exists} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where E denotes the fundamental matrix of the Markov chain for this transition (Doyle and Snell 1984):

$$E = (I - Q)^{-1} \tag{11}$$

and E_{ij} is the expected number of times a Markov chain with transition matrix Q that starts at node i will visit node j before being absorbed.

Network Search Procedure

We use a stochastic search procedure to find the semantic network that maximizes the objective function given in Eq. 6. A network is initialized using a secondary method that ensures the fluency lists can be produced by a censored random walk with non-zero likelihood (e.g., a Naïve Random Walk, or a fully connected network). In the simulations below, we use a network initialization procedure that starts with a network generated from a modified Community Network technique. Typically, the Community Network technique results in a network that cannot have generated the data by a censored random walk (i.e., the likelihood of the data given this network is zero). We overcome this by iteratively inserting an edge between any two responses in the data set whose transition probability is zero, until the likelihood of the data is non-zero.

From this initial network, edges are toggled—either pruned from or added to the network—one at a time.[7] When

[6] Although these items are not explicitly encoded in $\mathbf{A}'$, they are implicitly accounted for because $\mathbf{A}'$ is derived from the transition matrix $\mathbf{A}$, which *does* contain every item. In removing some rows and columns from $\mathbf{A}$ to form $\mathbf{A}'$, the transition probability mass for any node in $\mathbf{A}'$ does not always sum to 1.

[7] Toggling multiple edges simultaneously is inefficient because often the resultant network produces the data with zero probability. In fact, even single edge changes can result in "impossible" networks because the space of possible networks is highly constrained if we assume a censored random walk as a generative process. This constraint has made it challenging to use other techniques that can approximate the posterior probability directly, such as Gibbs sampling (Geman and Geman 1987). We leave this problem for future research.

an edge change results in a network with a higher likelihood of producing the data, the change is accepted; otherwise, the edge change is reverted. This procedure continues until no possible edge change produces a better network. Although changing only one edge at a time is susceptible to local maxima, we have found that this procedure performs robustly given different network initializations for re-estimating a network from censored lists generated from that network.[8]

The order in which edges are toggled is non-random. We adopt a set of heuristics for choosing which edges to toggle in order to test the most promising edges first. We enumerate a list of all possible edges and non-edges in the network (excluding self-loops), and separate them into three groups:

1. Existing edges that can be pruned
2. Non-edges that can be added to form "triangles" (Newman 2009), making the network more clustered
3. All other non-edges

Typically, a moderate proportion of edges in (1) are pruned, a smaller number of edges in (2) are added, and relatively few edges in (3) are added. We cycle through all three of these groups in order to test the most probable edge changes first. If a group is completed and an edge change occurred, we start the process over with group (1).

Within each group, a utility value is attached to each node. The utility value for node i reflects the average likelihood of the data over all networks tested (in the current group) that modify an edge attached to that node:

$$U(i) = \frac{\sum_{g \in h_i} \mathbb{P}(X|g)}{|h_i|} \tag{12}$$

where h_i is the set of networks that have been tested in the current group where an edge connected to i has been modified and $|h_i|$ is the number of networks in h_i.

Instead of adding or removing edges at random within each group, nodes are ranked according to their utility. An edge between the two nodes with the highest utility is toggled. This heuristic assumes that the likelihood of the data depends on increasing or decreasing access to certain nodes. This is consistent with previous findings that semantic networks tend to be small-world-like (Steyvers and Tenenbaum 2005), where a small number of nodes are expected to have many edges and a large number of nodes have very few edges. The heuristic relies on an assumption that if the removal (or addition) of an edge connected to a node makes the data more likely, then the removal (or addition) of other edges connected to that node will have a similar effect on likelihood.

We arrived at this search procedure after trying a number of different possible estimation procedures (e.g., stochastic search that toggles a random edge on each step). This one outperformed the others on simulated data. If all possible edges are tested exhaustively, the heuristics do not have a noticeable impact on the estimated network. That is, toggling all possible edges in a random order typically results in a similar network. However, the process of testing all possible edges is computationally expensive (there are $2^{N(N-1)/2}$ possible networks), and using a heuristic to prioritize edge changes results in quicker convergence of the search procedure. Additionally, one could specify a tolerance value for the number of edges to be toggled before convergence, which can dramatically reduce the time needed to find an adequate network.[9]

Hierarchical Model

Like other fluency-based network estimation techniques, the accuracy of U-INVITE can suffer when only a small amount of data is used to estimate a network. When dealing with psychological data, it is typical to have only a small number of lists per participant, making individual network estimation difficult.

To aid estimation when given only a small amount of fluency data, we introduce a hierarchical component to the model that jointly estimates the networks of many participants together. We do so by assuming that there is a latent group-level network that generates each participant's network. This assumption has an effect during estimation such that each individual network influences the group network, and the group network serves as a prior when estimating individual networks. The generative model for this process is shown in Fig. 4. We call this model Hierarchical U-INVITE.

The probability of any edge $\mathbf{G}^s_{ij}$ in an individual network for participant s follows a *zero-inflated beta-binomial distribution*. In a standard beta-binomial distribution, the prior probability of an edge is proportional to the number of participants who have that edge in their individual network. Of all the individual networks that contain both i and j, α_{ij} denotes the number of these networks that do not have

[8] See the Supplementary Material for more details.

[9] For the simulations reported below, we use the above heuristic but do not set a tolerance value—so all possible edges are toggled exhaustively until the search reaches a maximum. In the Supplementary Material, we show that when a small tolerance value is used, the estimated network is slightly worse but computation time is dramatically reduced.

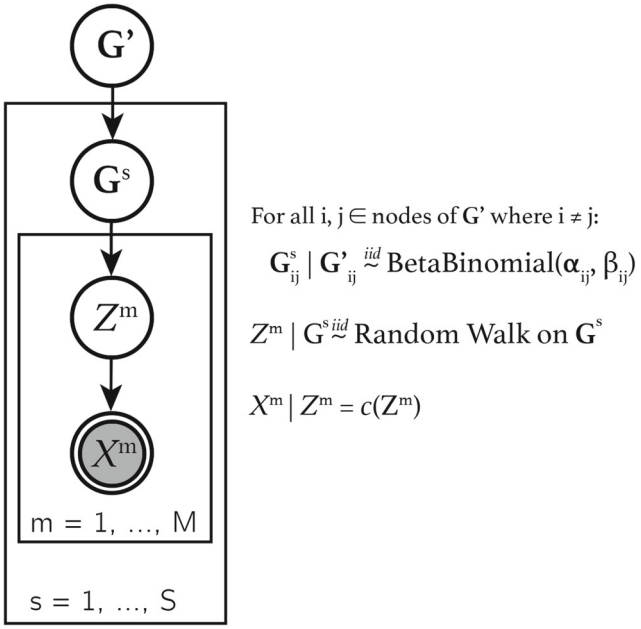

Fig. 4 Each individual semantic network $\mathbf{G}^s$ is generated from a group-level semantic network $\mathbf{G}'$. The probability of an edge being estimated in an individual network is dependent on the fluency data for that participant, as well as the proportion of other participants who have that edge in their network

an edge between i and j, while β_{ij} denotes the number of networks that do have an edge between i and j:

$$\alpha_{ij} = \sum_{s}^{S} \mathbb{1}_{\mathbf{G}^s_{ij}=0} \tag{13}$$

$$\beta_{ij} = \sum_{s}^{S} \mathbb{1}_{\mathbf{G}^s_{ij}=1} \tag{14}$$

where $\mathbb{1}_{G^s_{ij}=1}$ is an indicator function that is one when there is an edge between i and j is present in the network of the sth participant and zero when there is no such edge. S denotes the total number of participants.

In practice, using a standard beta-binomial distribution to infer edges for an individual's network results in a biased estimation. The reason is data sparsity: Z^m typically represents a very small number of possible random walks on a given network, traversing over a small fraction of the edges. The data are further diluted by the censoring process which makes only a portion of these transitions observable. As a result, the beta-binomial distribution often predicts a non-edge in a network not because the evidence favors a non-edge, but because of the lack of evidence for an edge.

As such, the observed counts over-inflate the number of networks that should not contain an edge (α_{ij}). A zero-inflated beta-binomial distribution compensates for this by introducing a prior, p_{hidden}, which denotes the proportion of the inferred non-edge counts (α) that result from data sparsity. The use of zero-inflated distributions is common in domains where an excess of "zeros" exist. For example, they are often used in modeling linguistic word count data (e.g., Jansche 2003) where most words do not occur in a given document, and in epidemiological data where a portion of the population is not at risk for a disease (e.g., Araujo et al. 2011; Böhning et al. 1999).

In the zero-inflated beta-binomial distribution, the proportion of observed non-edges is denoted by

$$\gamma_{ij} = p_{hidden}(\alpha_0 + \alpha_{ij}) \tag{15}$$

and the edge prior is denoted by

$$P\left(\mathbf{G}^s_{ij} | \gamma_{ij}, \beta_{ij}\right) = \frac{\beta_{ij} + \beta_0}{\gamma_{ij} + \beta_{ij} + \beta_0} \tag{16}$$

where β_0 and α_0 are free parameters that influence the edge prior when no information about an edge exists.

The *maximum a posteriori* group-level network can be estimated from the prior by inferring an edge whenever an edge prior $\mathbb{P}\left(G^s_{ij} | \gamma_{ij}, \beta_{ij}\right) > .5$, with the additional constraint that $\beta_{ij} > 1$. This additional constraint reduces spurious edges by ensuring that any edge in the group network exists in at least two individual networks (analogous to the threshold T_n used in the Community Network approach). We chose $p_{hidden} = .5$, $\beta_0 = 1$, and $\alpha_0 = 2$ so that the prior on any edge is $\mathbb{P}\left(\mathbf{G}^s_{ij} | \gamma_{ij}, \beta_{ij}\right) = .5$ when no data about that edge is available.[10]

Model Validation

To validate the different estimation methods, we conducted a large-scale simulation study in which we constructed an animal semantic network from the University of South Florida (USF) free association norms (Nelson et al. 2004) and simulated fluency data from this network using censored random walks. We compare how well each network estimation method is able to reconstruct the USF network as the amount of simulated data varies. The USF network is often assumed to be a "valid" semantic network, in that it accurately reflects mental associations between items in the network. This is because the free association task (used to build the USF network) explicitly asks for cue-response pairs to be associated, and most responses are semantically related. As such, the USF network serves as an important benchmark in semantic network research (Griffiths et al. 2007; Nematzadeh et al. 2017).

[10]These parameters were obtained from a grid search when initially developing the model, using USF network reconstruction (Section "Model Validation") as a benchmark. To avoid over-fitting, we selected $p_{hidden} = 0.5$, which performed well over a range of values for α_0 and β_0. We then set α_0 and β_0 to one such that they encode an uninformative (i.e., uniform) prior for edges with no data.

It is important to note, however, that because we are simulating data on this network, the psychological validity of the network is not crucial for validating the method as a whole. Nonetheless, the performance of a network estimation technique will depend on the topology of that network (Jun et al. 2015). The USF network is thought to have topological properties that resemble "valid" semantic networks, such as a small-world-ness (De Deyne and Storms 2008; Steyvers and Tenenbaum 2005), which makes it a better test case than alternatives (e.g., a random network).

The USF free association norms have been used extensively by psychologists as a standard for investigating the properties of semantic networks and semantic relatedness. Over three decades, the authors collected free response data from over 6,000 participants and 5,019 cue words. We constructed a semantic network from these norms using only a single category (animals) of cue-response pairs. We extracted every cue-response pair in the free-association data that consisted of two animals and adjoined each pair by an edge. From this, we used the largest connect component of the network as our animal semantic network. This network consists of 160 nodes (animals) and 393 undirected edges.

Fluency data were created for fifty simulated participants by a censored random walk on the unweighted and undirected USF semantic network. The initial item in each walk was chosen with probability proportional to the number of edges it has (Eq. 7). For each simulated participant, we generated three fluency lists of 35 animals each. As such, no simulated participant's data spanned the entire network. We varied how many simulated participants' data were used to estimate the group network, and compared the estimated network to the original USF semantic network. The data for each simulated participant is generated from the same network, and so the generative model employed here matches that of non-hierarchical U-INVITE; Hierarchical U-INVITE expects networks to vary between participants. These simulations were repeated ten times (resulting in ten simulated data sets of fifty participants each).

Results

From this simulated fluency data set, we estimated seven networks using each of the methods described earlier: Naïve Random Walk, First Edge, Community Network, Correlation-Based Network, Pathfinder, U-INVITE, and Hierarchical U-INVITE. Although each simulated participant produced three fluency lists, these lists were treated as independent for all non-hierarchical methods (i.e., there is no delineation of which list was generated by which participant in these methods).

Figure 5 shows the performance of each of the methods in reconstructing the USF network. We consider three metrics: (1) total cost (i.e., the number of edge changes needed to reach the USF network from the estimated network), (2) hits (i.e., the number of edges in the USF network successfully estimated by the method), and (3) false alarms (i.e., the number of edges estimated by the method that are not in the USF network).

All of the methods showed some ability to estimate the USF network. Given enough data, the non-hierarchical U-INVITE method performs the best and converges to the USF network, while the Hierarchical U-INVITE method trends in the same direction but does not converge after fifty participants. One reason for this may be that the hierarchical model assumes that each participant has some individual variation, but the data are really generated from one semantic network. In this simulation, all simulated participant networks are identical, which matches the generative process of the non-hierarchical version of U-INVITE.

Other techniques perform well or poorly in different situations. The Naïve Random Walk performs well with a small number of lists, while the Community Network performs well with a moderate number of lists. However both the Naïve Random Walk and Community Network techniques begin to do worse with a large amount of data, suggesting that they will likely never estimate the USF network even if additional data were provided to the model. In contrast, the remaining methods continue to improve (or plateau) as more data are provided.

The methods also differ in their sensitivity to hits and false alarms. The First Edge method is limited to adding at most one edge per list, but never infers a false edge. Both U-INVITE methods as well as the Pathfinder Network manage to keep false alarms relatively low. Naïve Random Walk, Community Network, and non-hierarchical U-INVITE are the quickest at correctly identifying correct edges (hits).

Overall, the non-hierarchical U-INVITE model was best able to reconstruct the USF network when given enough data. More generally, the most appropriate method to use may depend on several factors, including the amount of data available to fit a network, one's willingness to tolerate false alarms, and the data's adherence to the censored random walk model. In the next section, we evaluate how well each network estimation technique captures human similarity judgments.

Experiment: a Comparison of Network Estimation Techniques

Although these simulations validate U-INVITE as a consistent method for estimating networks from censored random walks, the psychological evidence in favor of U-INVITE is predicated upon the validity of the censored

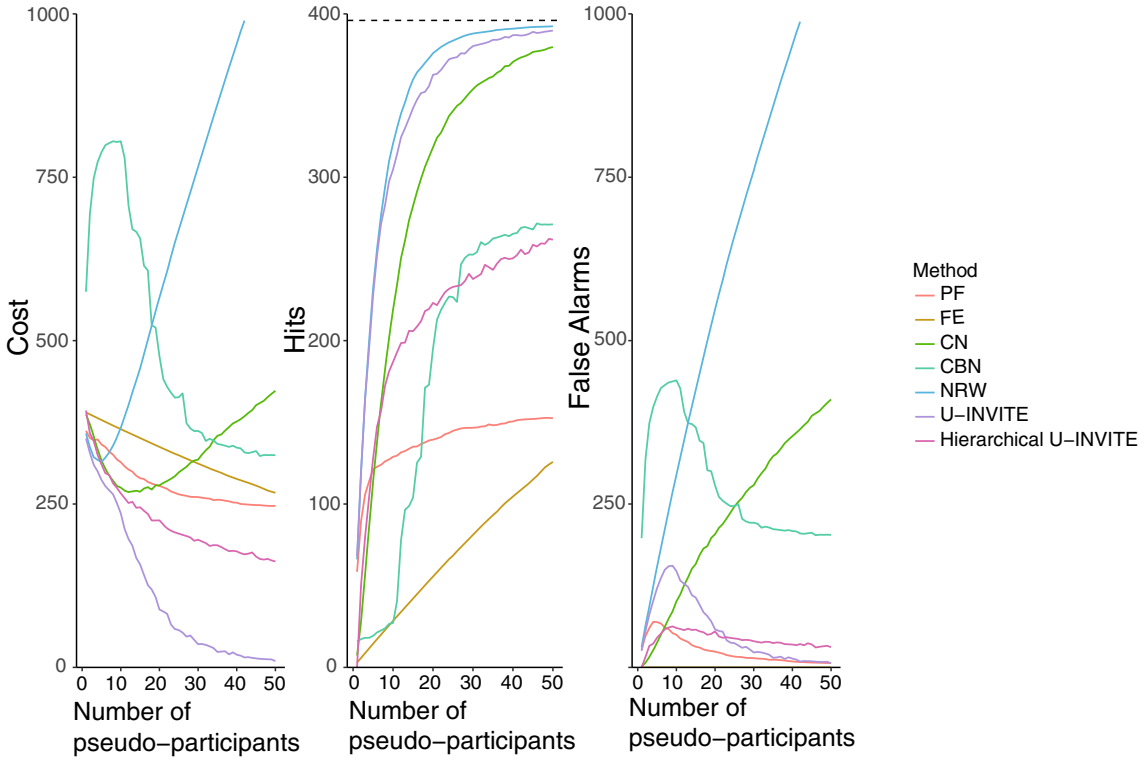

Fig. 5 (Left) The total cost (misses plus false alarms) for each network method is shown as the amount of data used to estimate the network is increased. (Center) The number of hits for each network method is shown. The dotted line represents 393 edges (the number of edges in the USF network). (Right) The number of false alarms for each network method is shown. Because the data are simulated using censored random walks, the First Edge method produces no false alarms. All results are averaged across ten simulated data sets (i.e., each data point is the average of ten networks)

random walk process model of memory retrieval. To address the psychological validity of the networks estimated by U-INVITE (and other methods), we estimated semantic networks from human fluency data using each of the methods described above and compared the similarity of edges in each estimated network using an independent group of raters.

Participants, Materials and Methods

Semantic Fluency Task Fifty participants (42% female, ages 18-62, median age 30) were recruited from Amazon Mechanical Turk to complete the repeated semantic fluency task. Participants completed three fluency lists for each of three categories (nine lists total): animals, tools, and food. The lists were pseudo-randomized so that each triad of lists contained one of each category, and that participants never completed the same category in back-to-back lists.

For each list, participants were given three minutes to name as many items from the category as possible, relying only on their own memory. They were also instructed to avoid repetitions within a list, but told that repetitions across lists were fine. A countdown timer was shown for each list as participants entered items into a text box. To avoid memory cueing from previous responses, a list of the participant's previous responses was not shown on the screen. Instead, each response faded from the screen after the participant hit Enter (the fade animation took 800ms), indicating that the item was recorded.

Only data from the animal category are analyzed here. The food and tools categories were used as filler tasks to minimize the effects of short-term memory and list recall, and were not analyzed. Spelling errors in the data were corrected, and responses were standardized (e.g., "hippo" to "hippopotamus," "giraffes" to "giraffe"). As is typical with healthy participants, intrusions (i.e., non-animals) were rare and appeared to be limited only to ambiguous items (e.g., do mythical creatures such as unicorns count as animals?). Intrusions were not removed from the fluency data. However, all perseverations were removed from the data as they cannot be generated under the U-INVITE generative model.[11] In total, 34 items (roughly .6% of all responses) were identified as perseverations (after

[11] The U-INVITE model described above describes a perfect censoring process. This does not need to be the case, and future work should explore whether including a faulty censoring process improves estimation.

Table 1 Each of the seven group networks are compared on several different network metrics

	PF	FE	CN	CBN	NRW	U-INVITE	Hier. U-INVITE
Connected network?	Yes	No	No	Yes	Yes	Yes	No
Number of nodes*	363	59	255	363	363	363	181
Number of edges	439	73	623	1083	2785	2663	224
Number of network components	1	4	4	1	1	1	7
Nodes in largest component	363	52	249	363	363	363	164
Network density*	.007	.043	.019	.016	.042	.045	.014
Average shortest path length**	7.21	4.17	4.38	5.63	2.73	2.76	8.01
Average node degree*	2.42	2.47	4.89	5.97	15.34	14.67	2.48
Clustering coefficient*	.003	.076	.311	.550	.228	.211	.160

*Metric excludes nodes with no edges

**Metric was calculated only on the largest component of the network

responses were standardized). On average, lists contained 35.8 responses, ranging from 6 to 68 responses per list.

Similarity Rating Task One hundred and one different participants (39% female, ages 21–60, median age 30) were recruited from Amazon Mechanical Turk to complete a similarity rating task.

With the semantic fluency data collected from all fifty participants, a group-level semantic network was generated using each of the methods described above. 3,905 edges (animal pairs) were estimated by at least one of these methods.[12] An additional 270 animal pairs were selected at random where no method had inferred an edge.[13] These pairs were randomized and split into twenty batches roughly equal in number. As such, approximately 6.5% of the animal pairs in each batch were pairs that were not an edge in any network. Each participant rated a single batch of animal pairs. Each pair of animals was presented sequentially, and participants were asked to rate the pair on a sliding scale from 1 ("not at all similar") to 100 ("very similar"). The ordering of the pairs was randomized, as was the ordering of animals within each pair.

Participants ticked a box to indicate if they did not recognize at least one animal in the pair. These similarity ratings (roughly 7% of the ratings) were removed prior to analysis. In addition, another 3% of the ratings were removed for having response times under 500 ms. After outlier removal, each animal pair had an average of 4.6 similarity ratings.

[12]Due to an error, we did not obtain similarity ratings for 4 of these edges.

[13]Originally, 200 non-edges were selected at random. Due to an error when estimating networks, some animal pairs originally identified as edges became non-edges.

Group Network Results

Group-level networks were constructed for each of the seven methods used for the USF simulations. Although each participant generated three fluency lists, each list was treated as independent (i.e., treated as if it came from a different participant) for all methods except Hierarchical U-INVITE.[14] Table 1 summarizes several properties of these networks.

In total, 363 animals were listed at least once across all fifty participants. However, not all of the networks consist of a single component with all of these items. Three networks (First Edge, Community Network, and Hierarchical U-INVITE) consist of multiple unconnected components. All of the networks are sparse (less than 5% of possible animal pairs were connected by an edge). The networks vary in how much they are clustered, with the Pathfinder Network showing very little clustering and Correlation-Based Network showing very high levels of clustering.

Figure 6 shows the average similarity rating of edges and non-edges in each network.[15] The 270 control non-edges (animal pairs chosen at random—not shown in the figure) received an average rating of 23.9, suggesting that all of the methods were capable of extracting semantically similar animal pairs as edges. The Hierarchical U-INVITE model

[14]The reason for this is that Hierarchical U-INVITE is the only method sensitive to which participant produced each list.

[15]Note that unlike the mean rating for edges, the mean rating of non-edges is based only on a sample of non-edges that were rated. In a network of 363 nodes there are 65,703 possible animal pairs in total. The estimated networks are sparse, so the vast majority of these pairs are non-edges. It is prohibitive for human raters to exhaustively rate these pairs. As such, the only non-edges that were rated are animal pairs that appear as an edge in at least one method, as well as the 270 control non-edges. In other words, each mean non-edge rating is derived from the 4,171 animal pair ratings collected, excluding all of the edges in that network.

performed the best, with an average edge similarity rating of 66.4.

However, it is difficult to rely solely on mean edge similarity ratings to compare methods, as the methods produce networks that vary in size (see Table 1 to compare network sizes). There is an inherent signal-to-noise trade-off in estimating network edges. A conservative method may estimate a small number of edges with high confidence, and these edges are likely to be highly similar. Conversely, a liberal method may estimate a large number of edges to capture as many semantic associations as possible, at the expense of including many edges that have low semantic similarity. Figure 6 also shows the average similarity rating of non-edges for each method. A lower average rating is desirable for non-edges, indicating that the animal pairs not estimated as edges are not semantically similar. The Naïve Random Walk and U-INVITE perform best here (with average non-edge similarity ratings of 22.9 and 23.1, respectively).

It should be noted that many of the network estimation techniques employed here could easily be parameterized to threshold a network, resulting in larger or smaller networks. Some techniques in fact have multiple ways to achieve this; for instance, the size of a Community Network can be reduced by increasing the width of the co-occurrence confidence interval, decreasing the window size w, or increasing the co-occurrence threshold T_n. Because many of these estimation techniques are highly flexible, we considered a single implementation of each model as typically used in the literature, noting any deviations in the model descriptions in the sections above. Some exceptions apply: the Naïve Random Walk and First Edge methods, as described, have zero free parameters and have a fixed network size. The Pathfinder implementation is parameterized to generate the sparsest possible network, though other parameterizations could increase the density of the network.

The ideal network estimation to use depends in part on one's tolerance for accepting bad edges or rejecting good edges. To compare different methods across a continuum of possible tolerance values, we computed a goodness score which weights the contribution of edges and non-edges:

$$GoodnessScore = \frac{c \sum_{edges} SimilarityScore + (1-c) \sum_{nonedges} (101 - SimilarityScore)}{c * NumEdges + (1-c) * NumNonEdges} \tag{17}$$

where c is a free parameter. When $c = 1$, the score reflects only the goodness of the edges in a network (i.e., Fig. 6 right). Contrastly, when $c = 0$, a high score indicates a tendency to reject bad edges only (i.e., Fig. 6 left). In Fig. 7, we plot each network's score as we vary c between 0 and 1.

When equal weight is given to edges and non-edges ($c = .5$), the Community Network performs the best, indicating that the method is effective at both estimating semantically similar edges while rejecting semantically dissimilar animal pairs. Note that since the networks are sparse, a weight

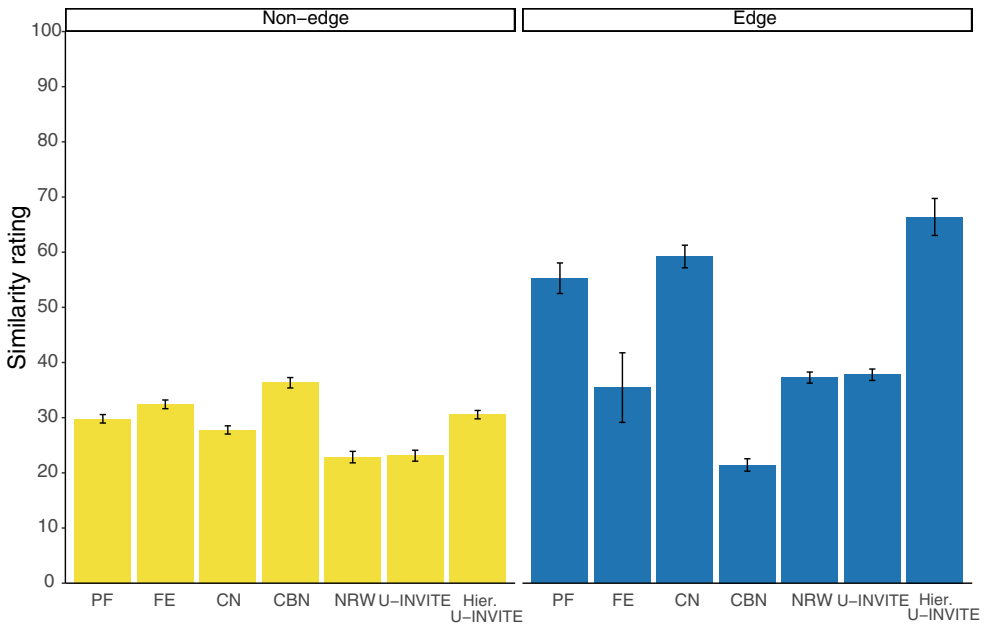

Fig. 6 The blue bars (right) show the average similarity rating across all edges in the network for each method. The yellow bars (left) show the average similarity rating for all non-edges in the network for each method. Error bars denote 95% confidence interval

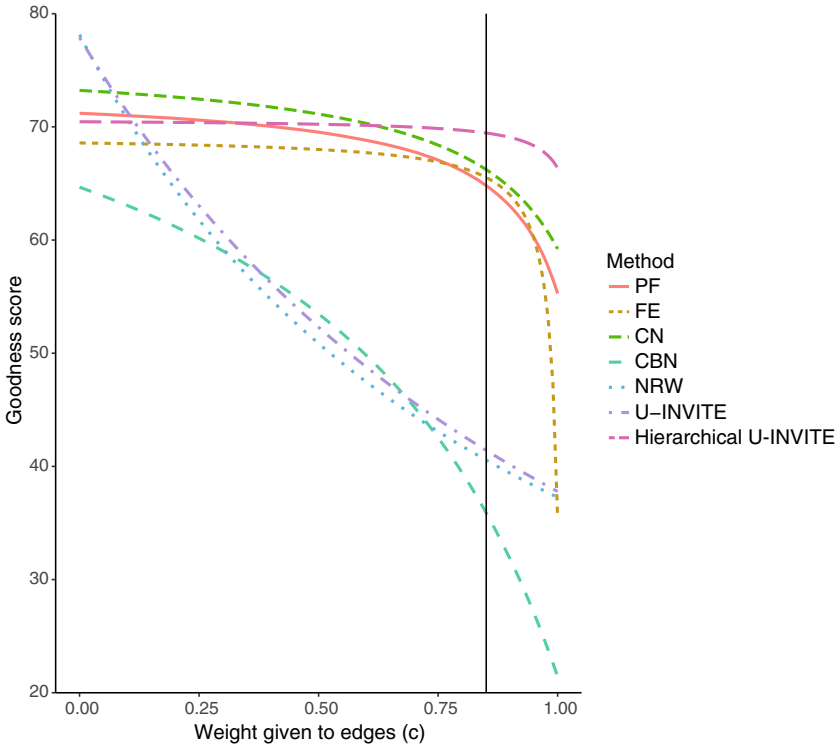

Fig. 7 The average goodness score for each network is shown, varying the weight c given to edges and non-edges. The solid vertical bar indicates a sparsity weighting ($c = .85$), which approximately equalizes the contribution of edges and non-edges. The sparsity weighting is the median sparsity across all seven networks

of $c = .5$ results in a score that is more influenced by non-edges than edges. That is, even though edges and non-edges are weighted equally, non-edges provide a larger contribution to the numerator of Eq. 17 because the total number of non-edges is higher than the total number of edges. Given this asymmetry, it may be more reasonable to weight edges more than non-edges. At these intermediate values ($.5 > c > 1$), the Hierarchical U-INVITE model typically performs best. One way to weight the importance of edges is inversely proportional to the sparsity of the network. This holds when the edge and non-edge portions of the denominator are equal, i.e., $c * NumEdges = (1 - c) * NumNonEdges$. This value is different for each network, as each network has a different level of sparsity. However, we determined the median sparsity weight to be $c = .85$ across the seven networks,[16] shown in Fig. 7 with a solid vertical line.

Another way to compare the estimated networks is to look at the proportion of edges in a given network that are rated above a fixed threshold value. For example, one might want to estimate a network in which most edges are expected to have a similarity rating of at least 50 out of 100. Figure 8 (left) shows the proportion of edges in each network that have a rating above a threshold value, varying the threshold on the x-axis from 0 to 100. The curves in Fig. 8 bear some resemblance to receiver operating characteristic (ROC) curves, commonly used in signal detection domains to assess the sensitivity of a model as a threshold of acceptibility is varied. In this case, the area under each curve (a popular ROC metric; e.g., see Bradley 1997) is equivalent to the mean semantic similarity rating for edges in that network.

The ordinal rankings of the network estimation techniques vary very little depending on the threshold value. Hierarchical U-INVITE consistently estimates networks with the highest proportion of edges rated above a threshold value (regardless of what that threshold is), whereas the Correlation-Based Network consistently estimates networks with a low proportion of edges above a threshold. This result suggests that some methods (such as the Correlation-Based Network) may consistently estimate too many edges with a low similarity rating, whereas other methods (such as Hierarchical U-INVITE) are more conservative but consistently estimate good edges. In Fig. 8 (right), we show the proportion of edges in each network estimated above a fixed threshold of 50.

[16]This sparsity weighting was determined based on the number of edges and non-edges *with ratings*, as opposed to the absolute number of edges and non-edges. This is done because non-edges without ratings are not included in the calculation of the goodness score.

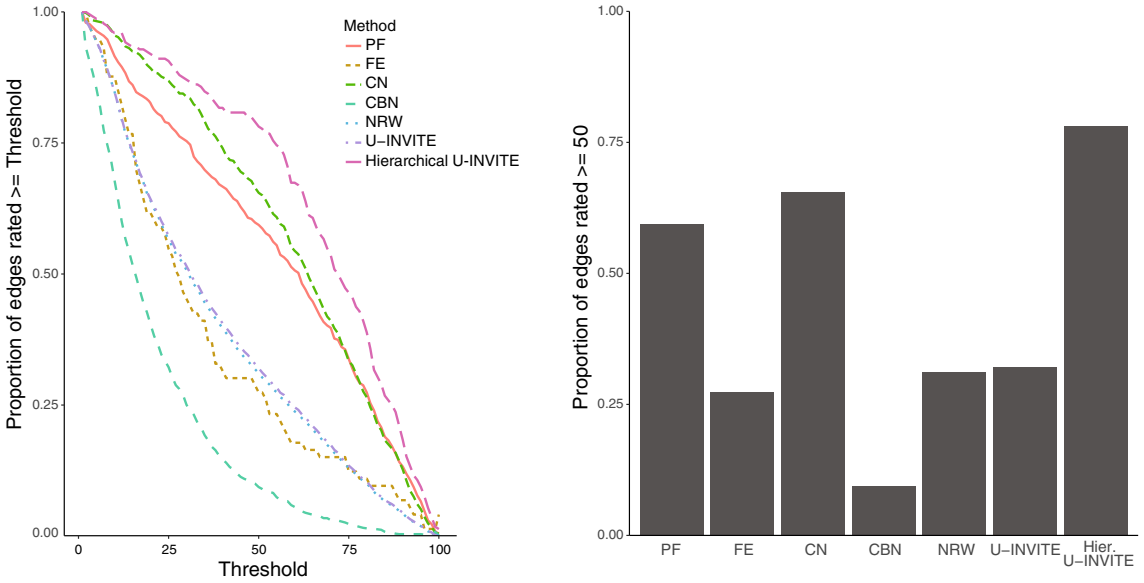

Fig. 8 (Left) The proportion of edges (*y*-axis) that remain in the network after edges with similarity ratings less than a threshold (*x*-axis) are removed. (Right) The proportion of edges that remain in each network after removing edges with a similarity rating less than 50

Individual Network Results

Though semantic networks are widely used in psychology, the vast majority of this research focuses on group-level semantic networks rather than explore individual variation in semantic representations (though see Morais et al. 2013 and Zemla et al. 2016). Typical experimental designs collect very few fluency lists per participant (often just a single list), and existing methods are not well suited for constructing networks from such a small amount of data. Methods such as the Correlation-Based Network can be unstable because a small number of data points can produce many undefined correlations and/or perfect correlations, both of which are problematic when converting the correlation matrix to a network. It also ignores sequential dependencies between items. Methods that rely on significance testing (such as the Community Network) are unlikely to produce many significant results with a small sample size, resulting in a network with very few edges.

The Hierarchical U-INVITE method assumes that semantic knowledge is shared across individuals. For example, if Peter thinks a horse is like a zebra then Mary probably will too, even if they may disagree on the strength of this similarity. By using a group prior, the Hierarchical U-INVITE method can estimate individual variation in semantic networks by determining where deviations from the group network would lead to better fits of an individual's data.

We constructed individual semantic networks for each of the fifty participants using each of the methods described above. Properties of these networks are described in Table 2. Analogous to the group-level network analysis, we computed the mean similarity rating of edges (and non-edges) in each network. All of the edges in the individual networks were present in at least one of the group semantic networks (thus, similarity ratings were available for every edge in every individual network—but, as with the group networks, ratings were not collected for most non-edges). To compare network estimation techniques, we then averaged the mean similarity ratings of networks across all participants. The results are shown in Fig. 9.

The Community Network method produced networks with the highest mean similarity rating (61.2) but, as expected, estimated very few edges per participant (about 20 edges per participant). These networks nearly always consisted of multiple unconnected components, and the method estimated an empty network (a network with no edges) for four participants. Similarly, the First Edge method also consistently produced extremely sparse networks (about 2 edges per participant) that were typically unconnected.

The Correlation-Based Network produced the largest networks, with roughly 127 edges per participant. However like the Community Network and First Edge methods, not all of the nodes in the networks had edges. Though the networks were large, the mean semantic similarity of the network's edges was lower than that of smaller networks generated by other methods, indicating a size-quality trade-off. The remaining methods (Naïve Random Walk, U-INVITE, Hierarchical U-INVITE, and Pathfinder) all produced networks that were roughly similar in mean edge similarity, with Pathfinder the highest at 46.2.

The networks showed little variance in the similarity ratings of non-edges. One explanation for this is that

Table 2 Network metrics were calculated first on each individual network, then averaged across all 50 participants for that method (standard deviation shown in parentheses)

	PF	FE	CN	CBN	NRW	U-INVITE	Hier. U-INVITE
Prop. of non-empty networks	1.0	1.0	.92	1.0	1.0	1.0	1.0
Prop. of connected networks	1.0	0.36	0.04	1.0	1.0	1.0	1.0
Number of nodes*	59.4 (18.5)	4.1 (1.5)	23.9 (14.2)	44.4 (19.8)	59.4 (18.5)	59.4 (18.5)	59.4 (18.5)
Number of edges	62.3 (19.6)	2.4 (0.8)	20.2 (15.1)	127.3 (59.4)	88.0 (26.7)	84.6 (26.3)	84.4 (27.1)
Number of network components	1.0 (0)	1.9 (0.8)	7.6 (3.5)	1.0 (0)	1.0 (0)	1.0 (0)	1.0 (0)
Nodes in largest component	59.4 (18.5)	2.4 (0.5)	4.9 (2.2)	44.4 (19.8)	59.4 (18.5)	59.4 (18.5)	59.4 (18.5)
Network density*	.04 (.01)	.5 (.32)	.14 (.2)	.16 (.08)	.06 (.02)	.06 (.02)	.06 (.02)
Average shortest path length**	10.3 (3.9)	1.1 (0.2)	1.6 (0.6)	2.7 (0.6)	5.2 (4.0)	5.3 (4.0)	5.2 (3.9)
Average node degree*	2.1 (.1)	1.1 (.1)	1.5 (.4)	5.7 (.2)	3.0 (.4)	2.9 (.4)	2.9 (.3)
Clustering coefficient*	.04 (.04)	0 (0)	.28 (.22)	.59 (.03)	.08 (.05)	.06 (.04)	.07 (.04)

All metrics calculated on non-empty networks only. *Metric excludes nodes with no edges. **Metric was calculated only on the largest component of the network

only a small sample of non-edges were rated (see Section "Participants, Materials and Methods"), and most of these non-edges appear in all networks. As such, the sets of non-edge ratings in each network largely overlapped with each other. Nonetheless, the Hierarchical U-INVITE method generated networks with the lowest mean non-edge similarity rating (29.4).

Limitations

The present work provides the first known attempt to compare many different methods for constructing semantic networks from fluency data. As such, there are limitations to be addressed by future work.

We collected and analyzed fluency data for the animal category, which is very common in the semantic fluency literature. Although the literature suggests that fluency data from other semantic categories exhibits similar structure (e.g., clustered responses), these network inference techniques should be validated on other semantic categories. Similarly, it is not certain whether these network inference methods would work equally well on non-semantic fluency data. The phonemic fluency task (listing words that begin with a particular letter or sound) is commonly used and may rely on different cognitive and neural mechanisms

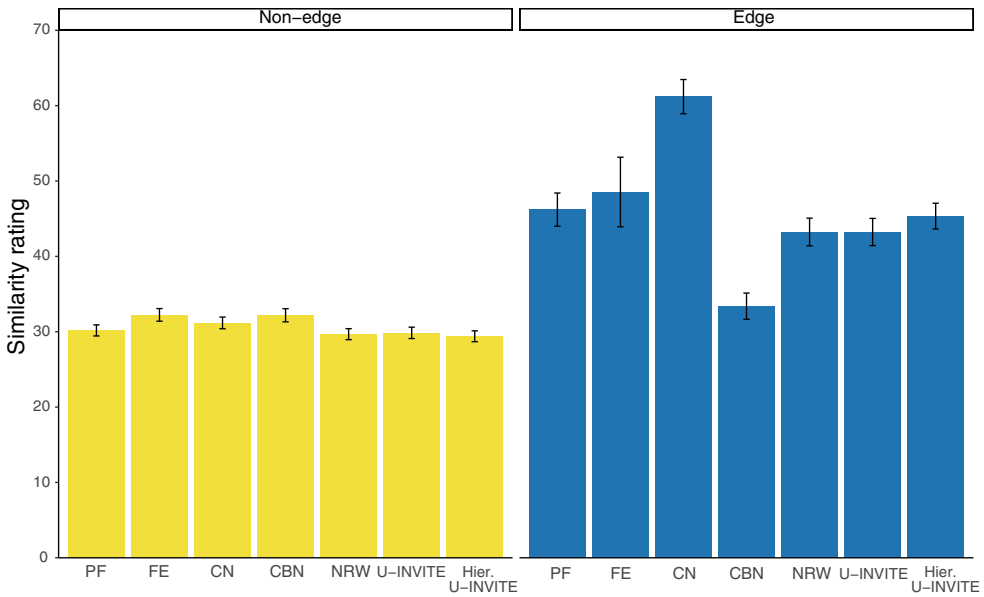

Fig. 9 The average similarity rating of edges and non-edges in the estimated individual networks is shown for each method. The Community Network (CN) produces networks with the highest mean similarity rating, though these networks are very sparse (sometimes containing no edges at all) and often contain multiple unconnected components. We found little difference in the average non-edge similarity rating across the methods. Error bars denote 95% confident intervals

(Troyer et al. 1998). These results call into question whether a censored random walk is appropriate for modeling non-semantic fluency data.

It is not clear how our findings would differ if we assume that mental search can stray from the target category. Abbott et al. (2015) proposed a censored random walk model in which search paths can navigate through non-category items which are then censored (e.g., a walk might traverse an indirect path from "dog" to "bone" to "dinosaur" when listing animals, with "bone" being censored). Any network inferred from semantic fluency data is necessarily limited to only the items that appear in those lists. This can pose a problem for network estimation techniques that expect associations to be direct, as indirect associations could result in spurious or weakly related edges.

We used a single paradigm to test all network inference methods: the repeated fluency task with three lists per individual and three minutes per list in a single session. Historically, participants are usually give one minute to produce one list; however, experiments vary in how many fluency lists are collected and the time limit provided for each list. We expect that the estimated networks would improve with more data (more lists or more time per list) and worsen with less data, but we have not provided a systematic comparison. The nature of the repeated fluency task may introduce artifacts in the data that might be explained by short-term list memory, rather than long-term semantic memory. For example, Zemla and Austerweil (2017) found that bigrams are commonly repeated across lists in a single session. These co-occurrences may lead to spurious edges in an estimated semantic network if these transitions are the result of short-term priming rather than long-term semantic association.[17] Similarly, primacy and recency effects may emerge from the repeated fluency task. In our data set, 16 of 50 participants started at least two lists with the same animal, and 11 participants started all three lists with the same animal. Future research should explore the interaction between short and long-term memory when modeling fluency data.

In Section "Model Validation," we found that these network inference techniques are successful at inferring networks when lists are produced by a censored random walk. It is less clear whether these techniques are equally successful when adherence to the random walk model of memory retrieval is not obeyed, or whether the advantages and disadvantages of each technique are robust to parameter changes for each model.

Finally, our results highlight several network inference techniques used to estimate unweighted and undirected networks. We do not know whether these techniques would be successful at inferring network structure if the data were generated from a weighted or directed network.

[17]Zemla and Austerweil (2017) discuss a generative model that accounts for these bigrams, though that model is not validated here.

Discussion

Although semantic networks are widely used in psychology, how to best construct a psychologically valid semantic network remains an open problem (Baronchelli et al. 2013). We developed novel non-hierarchical and hierarchical estimation techniques derived from a psychologically plausible model of memory retrieval. We evaluated five existing methods for constructing category-level networks from semantic fluency data and our own methods using validations via computer simulations and behavioral experiments.

Overall, based on the simulation and experimental results, we caution to advocate for a single network estimation technique. Rather, the best technique likely depends on the goal of the researcher. In Fig. 10, we show how to find an appropriate technique, given our results above. We suggest three key decisions that determine the best technique to use: (1) Do you want to generate a group-level network or individual-level networks? (2) Should the estimated network include a node for every response in the data and should the network be connected (i.e., should there be a path from every node to every other node?) This constraint is often important in psychological modeling; however, forcing this requirement can increase the number low quality (or spurious) edges. (3) Do you prefer to maximize edge similarity in part by estimating few edges (conservative), or minimize non-edge similarity, in part by estimating many edges (liberal)? Note that we only make this distinction for the group-level network, as we found virtually no difference in the quality of non-edges for individual networks across the different methods.

Aside from the factors illustrated in Fig. 10, a major difference in the estimation techniques is whether they adhere to a psychological process model of memory retrieval. To our knowledge, U-INVITE (both hierarchical and non-hierarchical versions) is the first proposed network inference method that is tied to a specific process model of how semantic fluency data is produced: the censored random walk (Abbott et al. 2015; Zemla et al. 2016). Methods that are not tied to a specific process sometimes result in more accurate networks when this process is misspecified. However, specification of a retrieval process allows researchers to test hypotheses about how impairments to the retrieval process affect the data that is produced. This may be particularly important given the wide use of the semantic fluency task in patients with semantic deficits such as those with Alzheimer's disease, Huntington's disease, and other neurodegenerative diseases—each of which may produce

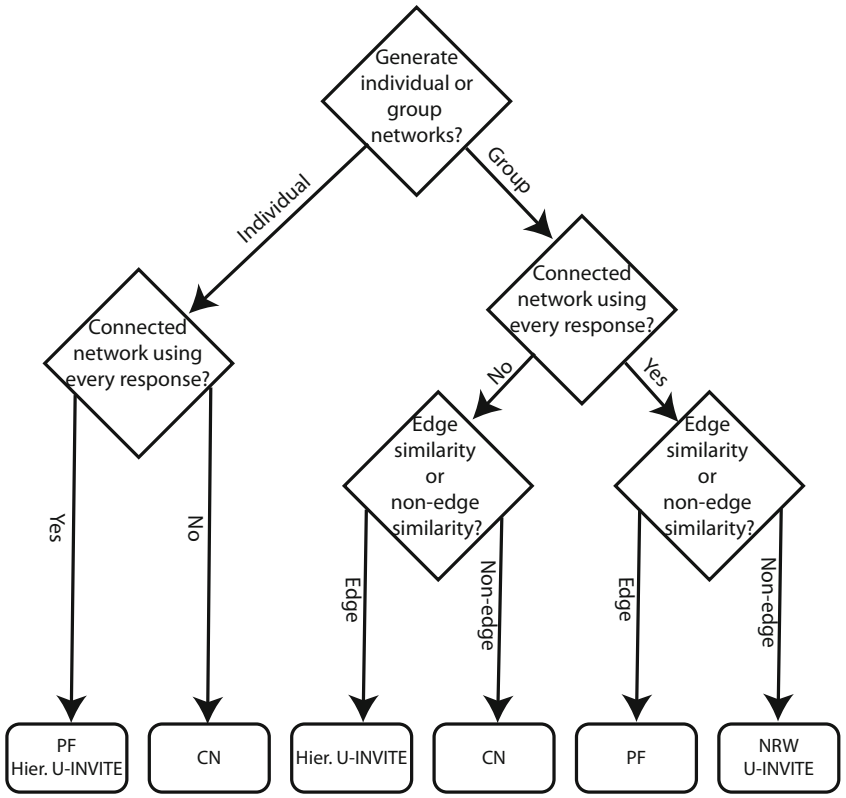

Fig. 10 Based on our experimental results above, this flow chart provides guidance for selecting an appropriate network estimation technique

differentially impaired profiles on the semantic fluency task (e.g., Randolph et al. 1993). Even for non-patient populations, experimental design choices (such as the amount of time between fluency lists) may influence aspects the retrieval process (Zemla and Austerweil 2017) and individual differences may affect semantic representation (Kenett et al. 2014). Modeling the retrieval process may provide greater insight into how semantic representations differs across groups (e.g., young and old, expert and non-expert, impaired and non-impaired).

From our simulation studies, we found that only two methods (U-INVITE and First Edge) are consistent estimators under the assumption that fluency data is produced by a censored random walk. In a simulated data set, we found that U-INVITE can reproduce the widely used USF semantic network (Nelson et al. 2004) with only a moderate amount of data. Consistency is a desirable theoretical property for statistical estimators as it means that the estimator will converge to the correct answer, given enough data.

We also compared the ability of all seven estimation techniques to produce psychologically valid semantic networks by comparing ratings of semantic similarity of edges in each network. The average ratings from each network were higher than the ratings of randomly selected animal pairs (the control non-edges), suggesting that all of the methods reviewed here are capable of estimating reasonable networks to some extent. Hierarchical U-INVITE produced a network with the highest average similarity rating for edges, although this may partially reflect a trade-off between network size and edge quality.

Overall, the most appropriate network inference method may depend on a number of factors, including the amount of data used to construct the network, tolerance for false alarms or misses, and the desire to construct individual or group level networks. Two methods that appear to perform particularly well across a range of tests (using both simulated and human fluency data) are the Community Network method and the Hierarchical U-INVITE method. In addition, both of these methods have sufficient flexibility to manipulate the number of estimated edges when signal-to-noise trade-offs are important.

The extant literature on semantic networks uses a variety of techniques to construct those networks. As such, the method employed constitutes a "researcher degree of freedom" that can significantly affect results. An understanding of the strengths and weaknesses of these methods allows researchers to choose an appropriate network estimation technique, and to fairly assess the results that follow from that choice. Our suggested best practices provide guidance to researchers, which we hope will encourage them to use the best method for their specific study and reduce researcher degree of freedom. This will increase confidence in network analyses, as it discourages

examining different network metrics while varying network estimation techniques until one finds a metric-estimation combination that yields a significant difference.

Acknowledgements Support for this research was provided by NIH R21AG0534676 and the Office of the VCGRE at UW-Madison with funding from the WARF. The authors would also like to thank Yoed Kenett and Kwang-Sung Jun for helpful discussion on an early version of U-INVITE.

Supplementary material All experiment data analyzed in this manuscript (semantic fluency data and the similarity rating data), derived semantic networks (individual and group networks generated for each method), R code to regenerate all figures in the manuscript, and additional supplementary material as noted throughout the manuscript is available online at https://osf.io/uy9jx/. In addition, a Python library created to generate networks using each of the described techniques is available at https://github.com/AusterweilLab/snafu-py.

References

Abbott, J., Austerweil, J., Griffiths, T. (2015). Random walks on semantic networks can resemble optimal foraging. *Psychological Review*, *122*(3), 558–569.

Abrahao, B., Chierichetti, F., Kleinberg, R., Panconesi, A. (2013). Trace complexity of network inference. In *Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 491–499). ACM.

Albert, R., & Barabási, A.L. (2002). Statistical mechanics of complex networks. *Reviews of Modern Physics*, *74*(1), 1–54.

Anderson, J.R. (1978). Arguments concerning representations for mental imagery. *Psychological Review*, *85*(4), 249–277.

Araujo, N.Bd., Barca, M.L., Engedal, K., Coutinho, E.S.F., Deslandes, A.C., Laks, J. (2011). Verbal fluency in Alzheimer's disease, Parkinson's disease, and major depression. *Clinics*, *66*(4), 623–627.

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

Baronchelli, A., Ferrer-i Cancho, R., Pastor-Satorras, R., Chater, N., Christiansen, M.H. (2013). Networks in cognitive science. *Trends in Cognitive Sciences*, *17*(7), 348–360.

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

Böhning, D., Dietz, E., Schlattmann, P., Mendonca, L., Kirchner, U. (1999). The zero-inflated poisson model and the decayed, missing and filled teeth index in dental epidemiology. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *162*(2), 195–209.

Borodkin, K., Kenett, Y.N., Faust, M., Mashal, N. (2016). When pumpkin is closer to onion than to squash: the structure of the second language lexicon. *Cognition*, *156*, 60–70.

Bousfield, W.A., & Sedgewick, C.H.W. (1944). An analysis of sequences of restricted associative responses. *The Journal of General Psychology*, *30*(2), 149–165.

Bradley, A.P. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, *30*(7), 1145–1159.

Busing, F., Commandeur, J.J., Heiser, W.J., Bandilla, W., Faulbaum, F. (1997). PROXSCAL: a multidimensional scaling program for individual differences scaling with constraints. *Softstat*, *97*, 67–74.

Chan, A.S., Butters, N., Paulsen, J.S., Salmon, D.P., Swenson, M.R., Maloney, L.T. (1993). An assessment of the semantic network in patients with Alzheimer's disease. *Journal of Cognitive Neuroscience*, *5*(2), 254–261.

Chan, A.S., Butters, N., Salmon, D.P., Johnson, S.A., Paulsen, J.S., Swenson, M.R. (1995a). Comparison of the semantic networks in patients with dementia and amnesia. *Neuropsychology*, *9*(2), 177–186.

Chan, A.S., Salmon, D.P., Butters, N., Johnson, S.A. (1995b). Semantic network abnormality predicts rate of cognitive decline in patients with probable Alzheimer's disease. *Journal of the International Neuropsychological Society*, *1*(3), 297–303.

Charnov, E.L. (1976). Optimal foraging: attack strategy of a mantid. *The American Naturalist*, *110*(971), 141–151.

Clopper, C.J., & Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, *26*(4), 404–413.

Collins, A.M., & Loftus, E.F. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*(6), 407–428.

Davison, M.L., Ding, C.S., Kim, S.K. (2010). Multidimensional scaling. In *The reviewer's guide to quantitative methods in the social sciences* (pp. 265–280). New York: Routledge.

De Deyne, S., & Storms, G. (2008). Word associations: network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231.

Dennis, S. (2007). How to use the LSA web site. *Handbook of latent semantic analysis* (pp. 57–70).

Doyle, P.G., & Snell, J.L. (1984). Random walks and electric networks. Mathematical Association of America.

Dry, M.J., & Storms, G. (2009). Similar but not the same: a comparison of the utility of directly rated and feature-based similarity measures for generating spatial models of conceptual data. *Behavior Research Methods*, *41*(3), 889–900.

Falk, E.B., & Bassett, D.S. (2017). Brain and social networks: fundamental building blocks of human experience. *Trends in Cognitive Sciences*, *21*(9), 674–690.

Geman, S., & Geman, D. (1987). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. In *Readings in computer vision* (pp. 564–584). Elsevier.

Geman, S., Bienenstock, E., Doursat, R. (1992). Neural networks and the bias-variance dilemma. *Neural Computation*, *4*, 1–58.

Goñi, J., Martincorena, I., Corominas-Murtra, B., Arrondo, G., Ardanza-Trevijano, S., Villoslada, P. (2010). Switcher-random-walks: a cognitive-inspired mechanism for network exploration. *International Journal of Bifurcation and Chaos*, *20*(03), 913–922.

Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., de Mendizábal, N.V., Corominas-Murtra, B., Bejarano, B., Ardanza-Trevijano, S., Peraita, H., Wall, D.P., et al. (2011). The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. *Cognitive Processing*, *12*(2), 183–196.

Griffiths, T.L. (2010). Bayesian models as tools for exploring inductive biases. In Banich M, & Caccamise, D. (Eds.) *Generalization of knowledge: multidisciplinary perspectives,*. New York: Psychology Press.

Griffiths, T.L., Steyvers, M., Tenenbaum, J.B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.

Gruenewald, P.J., & Lockhead, G.R. (1980). The free recall of category examples. *Journal of Experimental Psychology: Human Learning and Memory*, *6*(3), 225–240.

Henley, N.M. (1969). A psychological study of the semantics of animal terms. *Journal of Verbal Learning and Verbal Behavior*, *8*(2), 176–184.

Hills, T.T., Jones, M.N., Todd, P.M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440.

Jansche, M. (2003). Parametric models of linguistic count data. In *Proceedings of the 41st annual meeting on association for computational linguistics* (Vol. 1, pp. 288–295). Association for Computational Linguistics.

Johnson-Laird, P.N., Herrmann, D.J., Chaffin, R. (1984). Only connections: a critique of semantic networks. *Psychological Bulletin*, *96*(2), 292–315.

Jones, M.N., & Mewhort, D.J. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1.

Jones, M.N., Hills, T.T., Todd, P.M. (2015). Hidden processes in structural representations: a reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570–574.

Jones, M.N., Gruenenfelder, T.M., Recchia, G. (in press). In defense of spatial models of semantic representation. *New Ideas in Psychology*.

Jun, K.S., Zhu, X., Rogers, T.T., Yang, Z., et al. (2015). Human memory search as initial-visit emitting random walk. In *Advances in neural information processing systems* (pp. 1072–1080).

Kenett, Y.N., Wechsler-Kashi, D., Kenett, D.Y., Schwartz, R.G., Ben Jacob, E., Faust, M. (2013). Semantic organization in children with cochlear implants: computational analysis of verbal fluency. *Frontiers in Psychology*, *4*, 1–11.

Kenett, Y.N., Anaki, D., Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience*, *8*, 1–16.

Kruskal, J.B., & Wish, M. (1978). *Multidimensional scaling* Vol. 11. Beverly Hills: Sage Publications.

Landauer, T.K., & Dumais, S.T. (1997). A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*(2), 211–240.

Lee, M.D., Abramyan, M., Shankle, W.R. (2016). New methods, measures, and models for analyzing memory impairment using triadic comparisons. *Behavior Research Methods*, *48*(4), 1492–1507.

Lerner, A.J., Ogrocki, P.K., Thomas, P.J. (2009). Network graph analysis of category fluency testing. *Cognitive and Behavioral Neurology*, *22*(1), 45–52.

Levelt, W.J., Roelofs, A., Meyer, A.S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–38.

Masucci, A.P., Kalampokis, A., Eguíluz, V.M., Hernández-García, E. (2011). Wikipedia information flow analysis reveals the scale-free architecture of the semantic space. *PloS one*, *6*(2), e17,333.

Miller, G.A. (1995). WordNet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Morais, A.S., Olsson, H., Schooler, L.J. (2013). Mapping the structure of semantic memory. *Cognitive Science*, *37*(1), 125–145.

Navigli, R., & Ponzetto, S.P. (2012). BabelNet: the automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, *193*, 217–250.

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.

Nematzadeh, A., Meylan, S.C., Griffiths, T.L. (2017). Evaluating vector-space models of word representation, or, the unreasonable effectiveness of counting words near other words. In *Proceedings of the 39th annual meeting of the cognitive science society* (pp. 859–864).

Newman, M.E. (2009). Random graphs with clustering. *Physical Review Letters*, *103*(5), 1–5.

Paulsen, J.S., Romero, R., Chan, A., Davis, A.V., Heaton, R.K., Jeste, D.V. (1996). Impairment of the semantic network in schizophrenia. *Psychiatry Research*, *63*(2), 109–121.

Quaranta, D., Caprara, A., Piccininni, C., Vita, M.G., Gainotti, G., Marra, C. (2016). Standardization, clinical validation, and typicality norms of a new test assessing semantic verbal fluency. *Archives of Clinical Neuropsychology*, *31*(5), 434–445.

Quillan, M.R. (1966). *Semantic memory. Tech. rep*. Cambridge: Bolt Beranek and Newman Inc.

Quirin, A., Cordón, O., Guerrero-Bote, V.P., Vargas-Quesada, B., Moya-Anegón, F. (2008). A quick MST-based algorithm to obtain Pathfinder networks (∞, n-1). *Journal of the Association for Information Science and Technology*, *59*(12), 1912–1924.

Randolph, C., Braun, A.R., Goldberg, T.E., Chase, T.N. (1993). Semantic fluency in Alzheimer's, Parkinson's, and Huntington's disease: dissociation of storage and retrieval failures. *Neuropsychology*, *7*(1), 82–88.

Razani, J., Chan, A., Nordin, S., Murphy, C. (2010). Semantic networks for odors and colors in Alzheimer's disease. *Neuropsychology*, *24*(3), 291–299.

Sattath, S., & Tversky, A. (1977). Additive similarity trees. *Psychometrika*, *42*(3), 319–345.

Schvaneveldt, R.W. (1990). *Pathfinder associative networks: studies in knowledge organization*. Westport: Ablex Publishing.

Schvaneveldt, R.W., Durso, F.T., Dearholt, D.W. (1989). Network structures in proximity data. *Psychology of Learning and Motivation*, *24*, 249–284.

Shepard, R.N. (1964). Attention and the metric structure of the stimulus space. *Journal of Mathematical Psychology*, *1*(1), 54–87.

Shepard, R.N. (1980). Multidimensional scaling, tree-fitting, and clustering. *Science*, *210*(4468), 390–398.

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

Shindler, A.G., Caplan, L.R., Hier, D.B. (1984). Intrusions and perseverations. *Brain and Language*, *23*(1), 148–158.

Stella, M., Beckage, N.M., Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, 1–10.

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.

Tenenbaum, J.B., & Griffiths, T.L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*, 629–641.

Troyer, A.K., Moscovitch, M., Winocur, G. (1997). Clustering and switching as two components of verbal fluency: evidence from younger and older healthy adults. *Neuropsychology*, *11*(1), 138–146.

Troyer, A.K., Moscovitch, M., Winocur, G., Alexander, M.P., Stuss, D. (1998). Clustering and switching on verbal fluency: the effects of focal frontal-and temporal-lobe lesions. *Neuropsychologia*, *36*(6), 499–504.

Tulving, E. (1972). Episodic and semantic memory. In Tulving, E., & Donaldson, W. (Eds.) *Organization of memory* (chap 10. pp. 382–402). New York: Academic Press.

Tumminello, M., Aste, T., Di Matteo, T., Mantegna, R.N. (2005). A tool for filtering information in complex systems. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(30), 10:421–10:426.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327–352.

Tversky, A., & Hutchinson, J. (1986). Nearest neighbor analysis of psychological spaces. *Psychological Review*, *93*(1), 3–22.

Vinogradov, S., Kirkland, J., Poole, J.H., Drexler, M., Ober, B.A., Shenaut, G.K. (2003). Both processing speed and semantic memory organization predict verbal fluency in schizophrenia. *Schizophrenia Research*, *59*(2), 269–275.

Watts, D.J. (2004). The "new" science of networks. *Annual Review of Sociology*, *30*, 243–270.

Zemla, J.C., & Austerweil, J.L. (2017). Modeling semantic fluency data as search on a semantic network. In *Proceedings of the 38th annual meeting of the cognitive science society*.

Zemla, J.C., Kenett, Y.N., Jun, K.S., Austerweil, J.L. (2016). U-INVITE: estimating individual semantic networks from fluency data. In *Proceedings of the 38th annual meeting of the cognitive science society* (pp. 1907–1912).