Behavior Research Methods
https://doi.org/10.3758/s13428-023-02271-6

ORIGINAL MANUSCRIPT

Local but not global graph theoretic measures of semantic networks generalize across tasks

Maria M. Robinson[1] · Isabella C. DeStefano[1] · Edward Vul[1] · Timothy F. Brady[1]

Accepted: 9 October 2023

Abstract

"Dogs" are connected to "cats" in our minds, and "backyard" to "outdoors." Does the structure of this semantic knowledge differ across people? Network-based approaches are a popular representational scheme for thinking about how relations between different concepts are organized. Recent research uses graph theoretic analyses to examine individual differences in semantic networks for simple concepts and how they relate to other higher-level cognitive processes, such as creativity. However, it remains ambiguous whether individual differences captured via network analyses reflect true differences in measures of the structure of semantic knowledge, or differences in how people strategically approach semantic relatedness tasks. To test this, we examine the reliability of local and global metrics of semantic networks for simple concepts across different semantic relatedness tasks. In four experiments, we find that both weighted and unweighted graph theoretic representations reliably capture individual differences in local measures of semantic networks (e.g., how related *pot* is to *pan* versus *lion*). In contrast, we find that metrics of global structural properties of semantic networks, such as the average clustering coefficient and shortest path length, are less robust across tasks and may not provide reliable individual difference measures of how people represent simple concepts. We discuss the implications of these results and offer recommendations for researchers who seek to apply graph theoretic analyses in the study of individual differences in semantic memory.

Keywords Semantic networks · Individual differences · Models of memory · Network science · Graph theory

Fundamental to human cognition is the capacity to construct and extract meaning from our experience, and there is an extensive body of theoretical and computational modeling literature aimed at understanding semantic knowledge structures and how they are acquired and used across a variety of tasks (e.g., Borge-Holthoefer & Arenas, 2010; Griffiths et al., 2007; Jones et al., 2015; Kemp & Tenenbaum, 2008; Kumar, 2021; Kumar et al., 2021; Landauer & Dumais, 1997; Rogers & McClelland, 2004). Network-based representation schemes, where semantic concepts are represented as a network of interconnected nodes, are one of the most general representations of semantic knowledge structure (e.g., Kemp & Tenenbaum, 2008) and have been used for decades to represent the structure of semantic memory (e.g., Collins & Loftus, 1975). Since graphs are a useful representational scheme for understanding human semantic knowledge (e.g., Baronchelli, et al., 2013), graph theoretic analyses designed to quantify the local and global structure of graphs have been applied to model human semantic network structure (Kenett & Hills, 2022; Siew et al., 2019).

Much of the work using graph analytic approaches has been applied to aggregate data, for instance, to examine how semantic networks vary as a function of different conceptual spaces (e.g., Steyvers & Tenenbaum, 2005). More recently, however, graph theoretic analyses have been used to model individual differences in the properties of semantic networks to draw inferences about the relationship between semantic memory and higher-level cognition (e.g., Bieth et al., 2021; for review see Kenett & Faust, 2019). This line of research challenges a caricatured view of semantic memory as a passive storage system of general verbal and semantic knowledge that is invariant across both individuals and fluctuations in processing (a view that could be inferred from the classic work of Tulving, 1972, for example). Indeed, much converging evidence indicates that processing of concepts, like processing of episodic events, is sensitive to variations in external context (e.g., Howard et al., 2011) as well as in

✉ Maria M. Robinson
mrobinson@ucsd.edu

[1] Department of Psychology, University of California, San Diego, CA, USA

Published online: 28 November 2023

the internal states of the agent (e.g., Lin & Murphy, 2001; Rayner & Frazier, 1989). The study of individual differences in semantic networks, therefore, provides critical insight into the flexibility of semantic memory, as well as how higher-level cognitive processes may interact with semantic representations and their organization.

A major unanswered question in current studies that examine individual differences in semantic networks concerns the reliability of the basic metrics of semantic network structure within individuals. This answer remains unexplored because, within each study, only a single method is used for collecting semantic relatedness judgments. Given that no single method or analytic approach is guaranteed to yield a direct and pure measure of latent cognitive processes or representations (e.g., Falmagne & Narens, 1983), evidence that performance on a single semantic relatedness task predicts performance on tasks designed to measure other cognitive processes is inherently ambiguous. This point is particularly relevant in the study of individual differences, because evidence for reliable effects at the aggregate need not entail reliable effects at the level of individuals. A prime example of this is in the study of cognitive control, where the subtraction method is used to measure executive control via "congruency effects" (e.g., the difference in response times when people respond to a stimulus that is flanked by response-congruent versus response-incongruent stimuli; Eriksen & Eriksen, 1974). Recent research demonstrates that while congruency effects are reliable at the aggregate, they may not be reliable at the level of individuals, presumably because reliable group effects entail small interindividual differences (Hedge et al., 2018). Likewise, in the study of semantic memory, evidence for reliable group-level differences in semantic network topology (De Deyne et al., 2013; Kenett et al., 2017; Kumar et al., 2022; Steyvers & Tenenbaum, 2005) does not entail reliable variations in network topology within individuals. In this context, it is possible that existing evidence for individual differences in performance on a single semantic relatedness task reflect variations in how individuals strategically approach the particular semantic relatedness tasks, rather than latent differences in knowledge representation schemes. For instance, two people may have a different response criterion for calling something "highly similar" on a given Likert scale, yielding networks with different densities, even though their actual knowledge representations are identical.

Our goal is to fill this research gap by directly examining whether robust individual differences exist in concept representations across structurally different semantic relatedness tasks. To this end, we use two different semantic similarity judgment tasks, a Likert similarity judgment task and an adaptive version of a spatial arrangement task (Kriegeskorte & Mur, 2012). Critically, these two tasks have both been validated as measures of semantic relatedness in prior work (e.g., Richie et al., 2020), but place different demands on participants for reporting on relatedness judgments and involve different methods for translating relatedness judgments to proximity scores[1]. As such, if we find that these tasks both capture stable individual differences in performance, this will provide strong evidence that individual differences in semantic judgments are invariant across measurement approaches and may reflect true variations in semantic memory structure for simple concepts.

To preview our results, we do indeed find that network-based representations preserve individual differences in concept representation schemes across the two tasks. Specifically, we find that local metrics, which consider the semantic content of specific concepts and their relations within the network, are generalizable across semantic relatedness tasks (e.g., people who think dogs are similar to cats show evidence of this in both tasks). In contrast, we find that common global metrics of network topology that capture the overall structure of semantic networks are less reliable across these two tasks (e.g., people who seem to think all concepts are similar, and thus have a very connected semantic network when this network is derived from one task, do not show evidence of this same structural property when the other task is used). We discuss the broader implications of these results for theorizing about and measuring individual differences in concept representation schemes.

Prior work on individual differences in processing of semantic relatedness

In this section, we review recent work that applies graph theoretic approaches to the study of individual differences in semantic memory to predict other indices of global function, such as creativity (for a more general review on the application of graph theoretic approaches to semantic memory see, e.g., Kumar et al., 2021; Siew et al., 2019). We focus on this seminal work because it provides a major step towards integrating network science into the study of individual differences in semantic memory structure. Accordingly, our goal is to build on this line of research by directly and systematically examining the reliability of core local and global metrics of network topology.

In this domain, analyzing higher-level properties of semantic networks involves converting each individual's semantic relatedness data into a graph, in which nodes represent concepts and edges represent the connections between

[1] By proximity scores we mean semantic similarity—where higher values indicate greater similarity between concepts, or semantic distance—where lower values indicate greater similarity between concepts.

them. In some studies, researchers study individual differences while analyzing aggregate data by splitting individuals into groups based on some characteristic of interest (e.g., levels of creativity) and constructing semantic networks from each group's data. Researchers typically apply graph theoretic analyses to examine topological properties of these graphs, such as the degree to which nodes in the network cluster together. These metrics can then be related to performance on other tasks, such as divergent thinking tasks to measure creativity. Intuitively, topological metrics might reflect general aspects of semantic reasoning by conveying ability to traverse conceptual spaces in such tasks; for instance, semantic networks that are more closely connected allow for fluid and efficient association between concepts.

As an example of the kind of analyses typically used in this literature to make general conclusions about creativity and other constructs, consider an influential study by Kenett et al. (2014). The authors classified high- and low-creativity individuals using a battery of creativity tasks and explored whether there were systematic differences in network properties between the two groups. Semantic networks for high- and low- creativity individuals were constructed from aggregate data obtained from a free association task (Rubinsten et al., 2005; also see Wulf et al., 2022a). Global properties of these networks, such as the degree to which nodes (concepts) in the network cluster with one another, were quantified and indicated that low-creativity individuals exhibited more connectivity within clusters of concepts, and less connectivity across these clusters, as indexed by higher network modularity. Furthermore, concepts within networks of low-creativity individuals were reported to be more spread out, as indexed by larger average shortest path length.

Although this work provides provisional support for the view that graph theoretic measures can capture important individual differences in semantic memory organization, it suffers from important limitations, many of which are acknowledged by the authors. First, group differences in network modularity and average path length were small in magnitude. Second, networks were not constructed at the level of individuals but from aggregate data within each creativity group, so it is unclear how consistent these differences are at the level of individual subjects within groups; perhaps these differences are driven by only a few individuals in each sample, or subtle differences at the group level reflect aggregation artifacts (e.g., Estes, 1956). Third, networks from both groups were constrained to have 96 words from the free association task. This approach ensures that measures of network structure are unaffected by network size, however, this constraint may result in the loss of meaningful data in individual differences from the free association task.

Cosgrove et al. (2021) applied a similar methodological approach and group-level analysis to examine structural differences in semantic networks between older and younger adults. These authors reported group differences in the clustering coefficient, average path length, and network modularity. Differences between groups in these metrics tended to be, again, quite small, and were not reliable across different samples. In addition to some of the limitations listed above, the authors highlight another potential problem of using free association tasks to construct semantic networks: it remains unclear whether these group differences reflect differences in semantic network organization or other global processes, such as executive function, that may underpin recall processes in free association tasks. Together, studies using aggregate-based analyses and free association tasks to construct networks at the level of groups are limited and thus make it difficult to draw strong inferences regarding individual differences in network topology.

Several recent papers analyzed semantic networks at the level of individuals using semantic relatedness judgments, therefore addressing limitations of aggregate analyses and retrieval-based semantic tasks. For instance, Benedek et al. (2017) examined whether individual differences in semantic network structure related to creative thinking while controlling for intelligence. Instead of using a free association task to construct semantic networks, these researchers used a similarity judgment task in which participants were shown a pair of words and were instructed to judge their semantic relatedness using a continuous slider scale; analyses were also conducted at the level of individuals rather than groups. However, results in this study depended on how networks were constructed. In particular, these researchers examined three ways of transforming similarity data into graphs: (1) using a fixed edge number approach in which each individual had the same number of edges in an unweighted graph, (2) using a fixed minimum relatedness approach in which they used a fixed similarity cutoff to construct unweighted graphs, and (3) constructing weighted graphs from the raw proximity matrix. These authors found statistically significant individual differences in graph topology only when using the fixed minimum relatedness thresholding approach, and not the other two filtering approaches. These results suggest that associations between measures of network topology and other indices of global function may not be robust across different methods for constructing networks.

More broadly, this result highlights that relatedness judgments on a given semantic relatedness task are jointly determined by individuals' latent concept representations as well as strategic differences in how people approach the similarity relatedness task, and the use of a single semantic relatedness task confounds individual differences in semantic similarity judgments and task response strategies. This follows because individuals may have the same latent representations of semantic relatedness between concepts but differ in how they map their judgments to similarity scales (for relevant critiques in other domains: Liddell & Kruschke, 2018;

Malmberg, 2002; Regenwetter et al., 2019). Such differences in response policies will affect the edge density of their network, which would also affect each of the reviewed metrics of network topology. This critique applies to several other recent papers that use only rating-based semantic relatedness tasks to examine the relationship between individual differences in semantic network structure and indices of creativity (e.g., Kennet et al., 2019; He et al., 2021; Ovando-Tellez et al., 2022a, b). Another limitation of this work, which is acknowledged by the authors, is that correlations between indices of semantic network structure and other indices of global function tend to be low (e.g., Cosgrove, et al., 2023; Marko & Riečanský, 2021), some studies use small sample sizes ($n < 10$) to measure individual differences (e.g., Morais et al., 2013; Wulff et al., 2022a, b), and these analyses are typically uncorrected for a false discovery rate (e.g., Bernard et al., 2019).

In short, while these studies have made a major contribution to the study of semantic memory by using network science to quantify individual differences and raise critical questions about how individual differences in representations of simple concepts may relate to other cognitive processes, they also raise critical measurement questions about the reliability of network topology metrics within individuals. Accordingly, our goal is to build on this line of research by directly examining whether local and global measures of individual differences in semantic organization reflect latent differences in semantic structure or differences in how participants approach semantic relatedness tasks. The contribution of this work is to identify any potential boundary conditions in using these measures, and provide recommendations on how these constraints can be used to guide methodological and theoretical work in the study of individual differences in this domain. Finally, we underscore that the limitations we identify in these studies are not unique to this research domain, but reflect a deeper challenge faced by all behavioral researchers when measuring individual differences (e.g., Hedge et al., 2018), as well as unobservable constructs, which is validating their choice of measurement (for related discussions see literature on "the problem of coordination," e.g., Kellen et al., 2021; Van Fraassen, 2008, and "meaningfulness," e.g., Falmagne & Narens, 1983; Roberts, 1985).

Current work

We approach measurement validation by examining whether experimental outcomes are invariant across different measurement instruments that are designed to probe the same processes (e.g., Van Fraassen, 2008). Specifically, in four separate experiments we examine whether (1) network-based representations of different kinds preserve individual differences in semantic relatedness[2] judgments and (2) whether graph theoretic metrics of semantic network properties reliably capture individual differences.

We focus on widely used semantic similarity tasks as opposed to free association tasks, because the former provide an efficient way of collecting semantic relatedness judgments and, as reviewed, they do not require recall-based processes, which may confound semantic organization with other processes, such as executive control (e.g., Taconnat et al., 2010). To this end, we use two structurally different methods for collecting similarity relatedness judgments: (1) a rating similarity task in which participants are instructed to judge the semantic similarity of word pairs using a numerical rating scale and (2) an adaptive version of a spatial multi-arrangement task in which participants are instructed to spatially arrange words inside a 2D arena based on their semantic relatedness (Kriegeskorte & Mur, 2012). More precisely, in the spatial multi-arrangement task, participants are instructed to place similar (dissimilar) words closer together (further apart), such that the spatial physical distance between words reflects their relative distance in semantic space. On each trial of the multi-arrangement task, participants are presented with a subset of words, and the latent distance structure is inferred on a trial-by-trial basis by rescaling the redundant distance information across trials.

Critically, both the pairwise rating task and multi-arrangement tasks have been validated in prior work as measures of semantic relatedness (e.g., Charest, et al., 2014; Kriegeskorte & Mur, 2012; Majewska et al., 2021; Richie et al., 2020), but provide different response methods for measuring semantic proximity judgments, as well as different methods for computing proximity scores. Therefore, shared variance across these tasks is less likely to reflect strategic differences in how people approach the tasks, such as in how people sample ratings on Likert scales, or correlations in nuisance variance such as those related to trial-by-trial motor errors.

Finally, we examine the robustness of our results across different filtering approaches used to construct semantic networks, samples of participants ($n = 70$ in each of four experiments), and simple concepts. We also examine the effects of scale granularity by varying the number of response options on the pairwise rating scale across experiments. Specifically, in Experiments 1a and 2a we used a six-point Likert scale, whereas in Experiments 1b and 2b we used a 100-point slider scale. We considered both scales because each has potential strengths and limitations. On the one hand, the six-point Likert scale has

[2] We do not make the traditional distinction between associative and semantic relations because their definitions are overlapping, as noted in, e.g., Kumar (2021).

fewer response options, and prior work suggests that this may reduce decision noise in subjective rating judgment tasks (Benjamin et al., 2013). On the other hand, the fine-grained 100-point slider scale may provide a more sensitive measure of subtle differences in individual semantic relatedness judgments.

Local and global properties of semantic networks

We examine whether local and global measures of semantic networks are stable within a person across tasks. Local metrics are sensitive to the content of concepts because they are measured from specific semantic units within the network, and how related they are to one another. In other words, local metrics cannot be abstracted away from one domain and applied to another, because they reflect relatedness between specific concepts. The most holistic measure of local representations is the set of all relations between all concepts therein, as characterized by the complete adjacency matrix of the semantic network. It is also possible to simplify the content of a semantic network by distilling it to the relative importance of all the concepts within the network by calculating the eigenvector centrality of each concept in the network.

In contrast, global metrics abstract away from the specific concepts in the network and consider the overall topology of the semantic network. We consider two measures of the global structure of the network: the degree to which concepts in the network cluster together (measured by the average clustering coefficient), and the overall interconnectedness of the network (measured by the average shortest path length). For each of these measures we ask whether it is consistent within an individual across semantic relatedness tasks for the same simple stimuli, to assess whether that level of description of the semantic network captures stable individual differences. Throughout our analysis, we focus on these basic rather than composite measures of network topology—such as, “small-worldness” or “Network Portrait Divergence” (Bagrow & Bollt, 2019)—in order to isolate which properties of network topology are reliable.

To our knowledge, this analysis is the first to examine the robustness of metrics of semantic network structure across different empirical measurements of similarity. If metrics of network topology are recoverable across these tasks, it would indicate that these graph theoretic measures may capture substantive variance in semantic network content and structure. In contrast, finding that we cannot recover (some of) these metrics would highlight important boundary conditions that can guide future research on individual differences in semantic memory.

Weighted and unweighted representations of semantic networks

Finally, in order to test the robustness of our results across filtering approaches, we consider different ways of constructing semantic networks from human similarity judgments. Each of the approaches for constructing networks falls within the class of associative network modeling, where concepts (words) and the relations between them are represented as semantic units which are connected to each other through associative links (Collins & Loftus, 1975). First, we consider weighted, fully connected networks which have been analyzed in prior work (e.g., Benedek, et al., 2017; Kennet et al., 2019; He et al., 2021). A fully connected semantic network using weighted edges assumes that every word is related to every other word, but to varying quantitative degrees. The resulting network is described by a full, scalar similarity/adjacency matrix obtained directly from the semantic relatedness tasks. The weighted semantic network involves minimal transformations of raw human behavior and preserves the continuous nature of the similarity judgment data, and may thus have an advantage in detecting fine-grained individual differences. We also analyze sparse weighted networks constructed with the Pathfinder algorithm (Schvaneveldt et al., 1989), which prunes networks by preserving only the shortest possible weighted paths between nodes, conditioned on the data. A possible advantage of analyzing sparse weighted networks is that potential spurious edges are removed, which may reduce noise and lead to more robust measures of semantic network structure.

Second, we analyze several variants of unweighted, binary networks. Such a network is described by a binary adjacency matrix which can be obtained by thresholding the similarity matrices. Such sparse binary semantic networks have been used to model individual differences in semantic structure (e.g., Kenett & Faust, 2019), and might also have an advantage by discarding nuisance variation in similarity judgments to extract just the raw relations. We considered several ways of thresholding networks. In one set of analyses, we construct unweighted binary networks with fixed edge densities of 33% and 50%. In another set of analyses, we construct unweighted binary networks using a similarity criterion. Specifically, based on prior work (e.g., Benedek, et al., 2017), we use a raw similarity cutoff, where we preserve links between nodes that are above a given similarity threshold (e.g., all ratings above “50” on the 100-point slider scale).

To summarize, the weighted and unweighted approaches differ in how they formalize connections between semantic units and can be seen as two extremes of representing the granularity of associations between concepts. We examine a wide range of approaches towards constructing each of these networks, based on existing filtering methods as well as prior work in this research domain. Our goal is to determine whether a particular filtering approach leads to more robust metrics of similarity across different relatedness tasks, so that it can be used in future work on individual differences in semantic network properties. We summarize all methods for constructing networks in further technical detail in the Methods section, and, for ease of exposition, we review details for computing each measure of network topology in the Results section.

Methods

Participants

In each experiment, we collected data until our final sample size was $n = 70$. This afforded 80% statistical power to detect a low to moderate, directional effect size of at least $\rho = 0.3$ (Faul et al., 2007). Participants were from the University of California, San Diego community and participated in exchange for course credit. Participants were at least 18 years old and provided informed consent. We excluded participants if they did not finish one of the two tasks (12, 5, 8, and 3 participants in Experiments 1a, 1b, 2a, and 2b, respectively). Of the remaining sample, we excluded participants who, at the end of the study, did not report following the instructions when completing the tasks (13, 11, 6, and 7 participants in Experiments 1a and 2a, respectively) or reported that they were non-native and/or non-fluent English speakers (8 and 5 participants in Experiments 1a and 2a, respectively). Seven and nine participants in Experiments 1 and 2 chose to opt out of prescreening, respectively. The study was completed online through the university's Sona system and was approved by the institutional review board. All data and analytic code are available on the Open Science Framework repository at https://osf.io/26wku/.

Procedure and materials

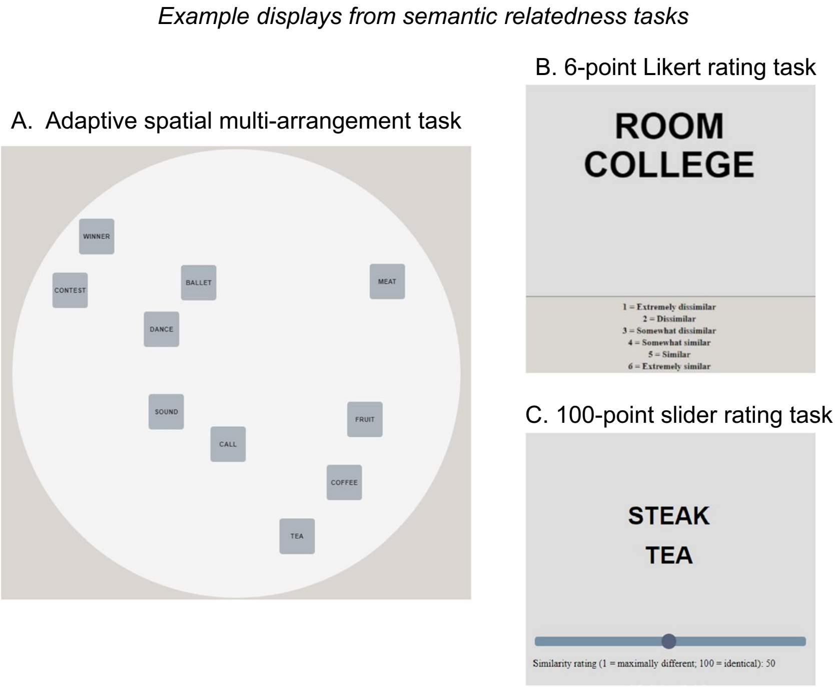
Participants completed an online version of the adaptive spatial multi-arrangement task (Kriegeskorte & Mur, 2012) and a six-point Likert semantic similarity judgment task (Experiments 1a and 2a) or 100-point slider semantic similarity judgment task (Experiments 1b and 2b).We include example instructions in Appendix 4. Experiments 1a and 1b used the same set of 20 words, and Experiments 2a and 2b used a different set of 20 words. Thus, the four experiments use four different independent samples and examined the effects of stimuli (Experiment 1 versus 2) and scale granularity in the pairwise rating task (Experiment a versus b).

All tasks were presented on a computer screen with the restriction that participants could not complete the experiment on a mobile device and that the browser window size was at least 800 × 775 pixels. After completing the experiment, participants were given a quality check prompt, which asked them to report whether they had followed instructions when completing both tasks. On average, it took participants 50 minutes to complete both tasks. Experiments were programmed in HTML/CSS/JavaScript. Code for the adaptive multi-arrangement task algorithm was cross-checked with MATLAB code provided by Kriegeskorte and Mur.

Spatial arrangement task (all experiments)

Figure 1A shows an example experimental trial on the adaptive spatial multi-arrangement task (Kriegeskorte & Mur, 2012; Majewska et al., 2021). On each trial, participants were shown a subset of words and instructed to use the entire circular arena to arrange words based on their semantic similarity, placing similar words closer together and dissimilar words further apart. The adaptive algorithm ensures repeated sampling of word pairs that are placed in close spatial proximity. This sampling procedure is based on the assumption that placement error is constant across trials and that the signal-to-noise ratio for semantic proximity judgments is proportional to the onscreen distance (because fine-grained differences in similarity may be indistinguishable from placement error for highly similar concepts). Therefore, words that are placed close to one another on a given trial will have a low signal-to-noise ratio. The algorithm "zooms in" on such word pairs on subsequent trials, permitting participants to make high-resolution similarity judgments for highly similar concepts. For instance, in the left panel of Fig. 1A, the three words in the lower left quadrant of the arena PICTURE, FRAME, and DOORWAY are relatively close to one another. Accordingly, on a subsequent trial, the algorithm zooms in on these words, meaning that the participant may only be shown these three words and be instructed to use the full circular arena to arrange them based on their similarity.

These repeated placements of specific word pairs are combined into a single rescaled distance matrix that ignores the onscreen distance of placements on specific trials. The adaptive version of the algorithm can be implemented until a criterion is reached, or until the experiment times out. In our experiment, participants were given a maximum of 35 minutes to complete the task. A major potential advantage of this method and analysis is that homing in on subsets of similar concepts has the potential to recover the high-dimensional structure of the similarity data (Kriegeskorte & Mur, 2012). For our purposes, another major advantage of using this method is that it places different processing demands than a Likert similarity task for making similarity judgments and involves different analysis and transformations of the similarity data than a Likert similarity task. We provide a full technical description of the algorithm in the


Example displays from semantic relatedness tasks
A. Adaptive spatial multi-arrangement task
WINNER
CONTEST
BALLET
MEAT
DANCE
SOUND
FRUIT
CALL
COFFEE
TEA
B. 6-point Likert rating task
ROOM
COLLEGE
C. 100-point slider rating task
STEAK
TEA

1 = Extremely dissimilar
2 = Dissimilar
3 = Somewhat dissimilar
4 = Somewhat similar
5 = Similar
6 = Extremely similar

Similarity rating (1 = maximally different; 100 = identical): 50

Fig. 1 Example experimental trials from the adaptive spatial multi-arrangement (**A**), Likert rating (**B**), and slider rating (**C**) tasks. The sample trial for the spatial multi-arrangement task shows a simulated arrangement of ten words

Appendix. In-depth technical details on and validation of the experimental procedure and algorithm are reported in Kriegeskorte and Mur (2012).

Six-point Likert pairwise similarity rating task (Experiments 1a and 2a)

Figure 1B shows an example trial on the Likert similarity judgment task. On each trial, participants were shown a word pair in the center of the screen. Participants were instructed to judge the similarity between the two words and use the number keypad to make one of six ratings: (1) extremely dissimilar, (2) dissimilar, (3) somewhat dissimilar, (4) somewhat similar, (5) similar, or (6) extremely similar. Participants self-advanced to the next trial by pressing the spacebar. Participants made judgments for all pairwise combinations of 20 words, yielding a total of 190 trials. All word pairs were presented randomly across trials.

100-Point slider pairwise similarity rating task (Experiments 1b and 2b)

Figure 1C shows an example trial on the 100-point slider similarity judgment task. On each trial, participants were shown a word pair in the center of the screen along with a 100-point slider scale which was anchored at the center. Participants were instructed to judge the similarity between the two words and use a continuous slider number keypad to make one of 100 similarity judgments ranging from "maximally different" (1) to "identical" (100). Participants self-advanced to the next trial by pressing the spacebar. Participants made judgments for all pairwise combinations of 20 words, yielding a total of 190 trials. All word pairs were presented randomly across trials.

Stimuli

All words were taken from the Nelson et al. (2004) word association norms database. The database was constructed using data from over 6000 participants; it consists of 5019 words and their associates obtained using a free association task. For each experiment, we selected 20 words. Experiments 1a and 1b used the same set of words, which were all simple nouns. We sampled words such that each had a matching, semantically related word. Specifically, one of the words was a cue word and the second word was its associate. Experiment 1 had the following words: BALLET, BANANA, CALL, COFFEE, CONTEST, DANCE, FRUIT, INDOOR, LION, MEAT, NOISE, PAN, PHONE, POT, OUTDOOR, SOUND, STEAK, TEA, TIGER, WINNER. To examine the generality of our results, in Experiments 2a and 2b we used a distinct set of 20 words from the Nelson et al. norms database. In these experiments we used simple nouns and verbs, and not all words on the list were cue–associate pairs. Experiment 2 had the following words: BLOCK, BOARD, CALENDAR, COLLEGE, DATE, DESK, DOORWAY, EXAMINE, FRAME, HOURS, LIBRARY, NOTES, PICTURE, READ, ROOM, SCHEDULE, SENTENCE, STRUCTURE, STUDY, SUBJECT.

Analysis

The overall approach for our analyses is to evaluate whether measures of the semantic network obtained from one person are consistent across tasks. To do so, we first extract the semantic similarity matrix across all tested concepts from each task, and then construct semantic graphs from these similarity matrices. We assess the consistency of various properties of the semantic graphs across tasks, within participants, by comparing these consistency scores to a null distribution obtained by permuting data across subjects.

Similarity matrices

All analyses start by calculating a semantic similarity matrix, measuring the apparent similarity of each pair of words. For both the six-point Likert and 100-point slider similarity ratings, we calculated the pairwise similarity ratings by min–max normalizing the raw rating scores because, on both scales, larger values indicate greater perceived similarity. For spatial multi-arrangement ratings, we calculated similarity matrices by min–max normalizing the negated distance matrices obtained from the spatial arrangement judgments (the calculation of the distance matrix for this task is explained in Appendix 1). Note that min–max normalization was used for convenience such that the similarity scales on the two tasks fell in the same range (0 to 1) and did not affect the results. Within each experiment, for each subject we thus obtain two semantic similarity matrices arising from the two (rating versus spatial arrangement) tasks.

Relatedness data averaged across participants For completeness and to demonstrate that we replicate findings from prior work (e.g., Richie et al., 2020) we begin by reporting descriptive statistics and associations between semantic relatedness data across the two tasks for all words when data are averaged across participants. In Experiment 1a, the average distance score was .07 ($\sigma = .015$; range: .017–.086) on the spatial arrangement task, the average similarity rating was 2.7 ($\sigma = 1.17$; range: 1.34–5.7) on the six-point Likert scale, and the correlation between these unconverted ratings across participants for all words was ($r(68) = -0.86$, $p < .001$)[3]. In Experiment 1b, the average distance score was .07 ($\sigma = .015$; range: .016–.084) on the spatial arrangement task, the average similarity rating was 35.4 ($\sigma = 23.3$; 6.5–93.2) on the 100-point slider scale, and the correlation between these ratings across participants was ($r(68) = -0.85$, $p < .001$). In Experiment 2a, the average distance score was .069 ($\sigma = .014$; range: .029–.088) on the spatial arrangement task, the average similarity rating was 3.23 ($\sigma = 1.07$; range: 1.36–5.54) on the six-point Likert scale, and the correlation between these ratings across participants was ($r(68) = -.90$, $p < .001$). In Experiment 2b, the average distance score was .069 ($\sigma = .013$; range: .034 - .09) on the spatial arrangement task, the average similarity rating was 44 ($\sigma = 19.1$) on the 100-point slider scale, and the correlation between these ratings across participants was ($r(68) = -0.89$, $p < .001$).

Constructing graphs from similarity matrices

For illustration, Fig. 2 shows the similarity matrices obtained from the Likert and spatial arrangement tasks for two participants in Experiment 1a, which are equivalent to the adjacency matrices of the weighted semantic graph. Figure 2 also shows the unweighted (binary) graph constructed with an edge density threshold of 33% for the two tasks. Appendix Figs. 7, 8, 9 and 10 shows example binarized graphs and adjacency matrices constructed from the aggregate data.

To construct *fully connected weighted* semantic graphs, we treat the similarity matrices directly as adjacency matrices; thus, the similarity matrices shown in the left panel of Fig. 2 correspond to the adjacency matrices of the fully connected weighted graphs they define. To construct sparse weighted semantic graphs, we used the Pathfinder algorithm (Schvaneveldt, 2023), which preserves only the shortest path between nodes and discards remaining links. As discussed, weighted sparse graphs may help eliminate spurious node links and provide more robust measures of semantic network structure.

We constructed *unweighted* (binarized) graphs using two filtering approaches with different thresholds. The first approach transforms the similarity matrices into binary adjacency matrices by finding a similarity threshold that would achieve a desired edge density, resulting in the sparsely connected binary graphs. Specifically, we chose thresholds for similarities to achieve a fixed proportion of edges across participants and tasks. We evaluated both 33% and 50% edge densities to examine whether our results were reliable across different thresholding criteria. Given the 190 possible edges $(n(n-1))/2$ between the 20 words in our sample, this yielded thresholds of 64 and 95 total edges, respectively. Because the six-point Likert task used a coarse measure of similarity, we could not meet this fixed criterion exactly for each individual in the Likert task, so we chose individuals' thresholds such that the number of edges was as close as possible to 64 or 95. In Experiment 1a the average edge

[3] Correlations are expected to be negative in the raw data because the spatial arrangement task measures semantic relatedness in terms of distance, and the rating tasks measure semantic relatedness in terms of similarity.

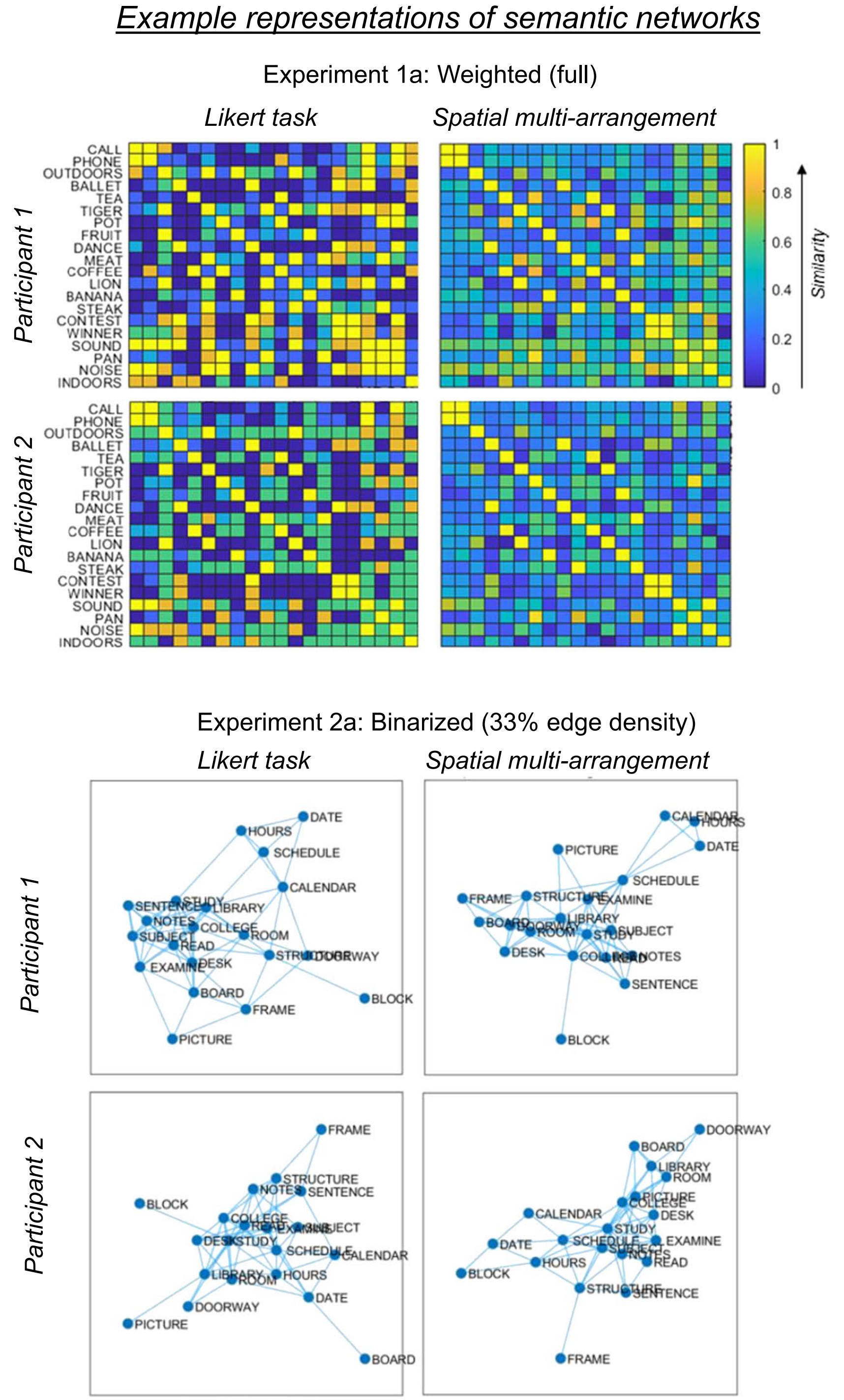


Fig. 2 Example similarity matrices, and thus weighted-graph adjacency matrices, obtained for two example participants from the Likert and spatial arrangement tasks (left), and the sparse unweighted graph that arises from thresholding these adjacency matrices to 33% edge density for unweighted graph analyses (right)

density across participants was 64 (σ = 17; range: 25–164) with a 33% edge density threshold and 87 (σ = 21; range: 25–164) with a 50% edge density threshold. In Experiment 1b the average edge density across participants was 63 (σ = 4; range: 38–68) with a 33% edge density threshold and 92 (σ = 11; range: 46–109) with a 50% edge density threshold. In Experiment 2a the average edge density across participants was 63 (σ = 11; range: 37–92) with a 33% edge density threshold and 92 (σ = 13; range: 62–117) with a 50% edge density threshold. In Experiment 2b the average edge density across participants was 62 (σ = 5; range: 32–67) with a 33% edge density threshold and 96 (σ = 12; range: 75–142) with a 50% edge density threshold. Note that graphs constructed using the 33% and 50% edge density thresholds are overlapping within participants for the same words, with the exception that graphs constructed using the 50% criterion are more interconnected graphs on average.

We then thresholded the continuous similarity data from the spatial arrangement task such that we obtained the same number of edges in the adjacency matrices from the spatial arrangement task as we had in the Likert task for each participant. This procedure ensured that the number of edges was the same within each individual across the two tasks and, accordingly, that individual differences in graph properties were not obfuscated due to differences in scale granularity in Experiments 1a and 2a. For permutation analyses, we randomly paired the Likert data from one subject with the spatial arrangement task data from another subject, and we repeated this procedure for each "synthetic subject" to ensure that the permutation analysis also had matched edge densities for the two tasks.

The second filtering approach for constructing binarized networks transforms the similarity matrices into binary adjacency matrices by using a similarity instead of a link-density threshold. Based on prior work (Benedek, et al., 2017) we used a mid-scale cutoff from each task's (dis)similarity scales. For data from slider and Likert pairwise similarity tasks, we linked nodes for word pairs that participants assigned a similarity rating greater than "50" or "3," respectively. From data from the spatial multi-arrangement task, we linked all word pairs that were below the average of the median scores of each participant within each experiment.

Permutation analyses

For any measure of the similarity of semantic structure within individuals, across tasks, we must construct a null hypothesis corresponding to an absence of stable individual differences, but which respects the possibility of consistent semantics for the whole group. To test for the presence of individual differences in semantic network structure, we used permutation analyses by pairing the data of one subject from the Likert (Experiments 1a and 2a) or slider pairwise similarity task (Experiments 1b and 2b) with the spatial arrangement task data from another subject. For instance, Tim's adjacency matrix on the Likert task might be paired with Isabella's adjacency matrix on the spatial arrangement task, while Tim's spatial arrangement data and Isabella's Likert data are paired with other subjects. On each permutation, all the subject–task measurements are used exactly once, but paired with the corresponding task of another subject, yielding the same number of *synthetic subjects* that we had in the unpermuted data.

Any across-task similarity measures we calculate on such synthetic subjects correspond to samples of the null hypothesis wherein there are no stable individual differences in semantic structure across participants. We repeated this procedure 10,000 times with different permutations of the data to construct a null distribution. We can then obtain a p-value against this null hypothesis by evaluating what proportion of null samples obtained via permutation showed similarity across tasks equal to or greater than the unpermuted data, thus asking whether the consistency of a particular property of the semantic graph is more consistent within individual, across tasks, than we would expect by chance. Effectively, this analysis gets at the following question: *Is semantic network structure more similar across tasks when using the same person's data versus when using a different person's data?*

Results

For both weighted and unweighted semantic graphs we test for stable individual differences in the *content* of the semantic network, as well as the *structure* of the semantic network. The full content of the semantic network amounts to the overall pattern of relations among words, as captured by the full 20 × 20 adjacency matrix describing the full network. The semantic content of the network may be roughly summarized by considering just the relative importance of all 20 nodes, characterized by a vector of all their node centralities. Both content-based measures of the semantic structure evaluate whether individual differences in semantic relatedness judgements are robust across the two tasks.

In contrast, measures of the *structure* of the semantic network abstract away from the specific nodes (concepts) over which the network is defined and extract overall measures of network connectivity. We focus on two such structural measures: the degree to which concepts are clustered (clustering coefficient), and the efficiency of conceptual interconnections (as measured by average path length between concepts). Both of these structural measures disregard the specific nodes but seek to characterize some more general property of an individual's semantic network, which might capture some higher-level property of an individual's high-level semantics.

All reported p-values from our permutation tests are corrected to control for multiple comparisons using a Bonferroni correction. To increase our ability to detect individual differences, we define family-wise comparisons by counting the number of tests within each filtering method for constructing graphs, which results in four tests. In addition to permutation p-values, we report descriptive statistics from the permutation analyses (denoted with the "*Perm*" subscript) and the absolute value of the z-score standardized effect size from the permutation analyses (denoted as $|d_{Perm}|$).

Consistency of local metrics of semantic networks

We examine whether local metrics of semantic networks capture individual differences that are consistent across tasks. We evaluate two measures: (1) the overall set of node relations, as characterized by the full $n \times n$ adjacency matrix, and (2) the relative semantic importance of nodes, as captured by the n-unit vector of node centralities. Insofar as these measures are consistent across (6- or 100-point) rating and spatial arrangement tasks, it indicates that there are individual differences in the semantic structure among these concepts that are stable across tasks.

Overall: Complete adjacency matrix

First, we ask whether the overall semantic similarity structure as captured in the full adjacency matrix was consistent for a given individual across the rating and spatial arrangement tasks. We quantified the consistency of the links across networks by measuring the Euclidean distance between their adjacency matrices. In each experiment there are 20 words, and each 20×20 similarity/adjacency matrix can be treated as a 1×190 similarity vector. Below, we represent the adjacency matrix in vector form for convenience because the upper and lower triangular matrices of adjacency matrices for undirected graphs are redundant. Equation 1 shows the formula for the Euclidean distance (d) between the similarity vector obtained from the Likert rating task (R) and similarity vector from the spatial multi-arrangement task (S),

$$d_{LT} = \sqrt{\sum_{i=1}^{190} (R_i - S_i)^2}. \tag{1}$$

Smaller distances between adjacency matrices indicate greater consistency of the two matrices.

Weighted graphs The overall structure of semantic relations was more similar across tasks from the same subject than for randomly paired subjects in all experiments for both full weighted graphs and sparse weighted graphs. In Experiment 1a, the average distance between Likert and spatial arrangement task adjacency matrices was 4.32 and 4.56 for full weighted and sparse weighted matrices, respectively. For full ($\underline{X}_{Perm}$ = 4.66) and sparse ($\underline{X}_{Perm}$ = 4.93) weighted matrices, these distances were significantly smaller within subjects than in the permuted sample ($|d_{Perm}|$ = 8.0 and $|d_{Perm}|$ = 10.7, respectively; ps < .001). In Experiment 1b, the average distance between slider and spatial arrangement task adjacency matrices was 4.11 and 4.15 for full weighted and sparse weighted matrices, respectively. For full ($\underline{X}_{Perm}$ = 4.5) and sparse ($\underline{X}_{Perm}$ = 4.54) weighted matrices, these distances were significantly smaller within subjects than in the permuted sample ($|d_{Perm}|$ = 7.4 and $|d_{Perm}|$ = 12.3, respectively; ps < .001). In Experiment 2a, the average distance between Likert and spatial arrangement task adjacency matrices was 4.32 and 5.43 for full weighted and sparse weighted matrices, respectively. Again, for full ($\underline{X}_{Perm}$ = 4.79) and sparse ($\underline{X}_{Perm}$ = 6.04) weighted matrices, these distances were significantly smaller within subjects than in the permuted sample ($|d_{Perm}|$ = 7.9 and $|d_{Perm}|$ = 11.9, ps < .001). Finally, in Experiment 2b, the average distance between slider and spatial arrangement task adjacency matrices was 4.3 and 5.38 for full weighted and sparse weighted matrices, respectively. For full ($\underline{X}_{Perm}$ = 4.55) and sparse ($\underline{X}_{Perm}$ = 5.74) weighted matrices, these distances were significantly smaller within subjects than in the permuted sample ($|d_{Perm}|$ = 5.2 and $|d_{Perm}|$ = 9.75, respectively; ps < .001).

Binarized graphs We also found that the structure of semantic relations was more similar across tasks from the same subject than for randomly paired subjects in all four experiments for unweighted graphs with both 33% and 50% edge densities, and unweighted graphs with a fixed mid-scale threshold. In Experiment 1a, the average distance between unweighted graph matrices was 7.02 with edge density of 33% ($\underline{X}_{Perm}$ = 7.4), 7.76 with edge density of 50% ($\underline{X}_{Perm}$ = 8.09), and 8.15 with a fixed mid-scale threshold ($\underline{X}_{Perm}$ = 8.54). Each of these distances within subjects was significantly smaller than in the permuted sample ($|d_{Perm}|$ = 7.97, $|d_{Perm}|$ = 6.62, and $|d_{Perm}|$ = 8.66, respectively; ps < .001). We found the same pattern of results in Experiment 1b, where the average distance between unweighted graph matrices within the same subject was 7.07 when edge density was 33% ($\underline{X}_{Perm}$ = 7.54), was 7.97 when edge density was 50% ($\underline{X}_{Perm}$ = 8.34), and 8.14 when we used a fixed mid-scale threshold ($\underline{X}_{Perm}$ = 8.51). Again, each of these distances within subjects was significantly smaller than in the permuted sample ($|d_{Perm}|$ = 10.96, $|d_{Perm}|$ = 8.46, and $|d_{Perm}|$ = 8.42, respectively; ps < .001). Likewise, in Experiment 2a, the average distance between unweighted graph matrices was 6.99 within the same subject when edge density was 33% ($\underline{X}_{Perm}$ = 7.56), 7.90 within the same subject when edge density was 50% ($\underline{X}_{Perm}$ = 8.37), and 7.88 when we used a fixed mid-scale threshold ($\underline{X}_{Perm}$ = 8.31). For each threshold, these distances were significantly smaller within subjects than in the permuted sample ($|d_{Perm}|$ = 12.88, $|d_{Perm}|$ = 10.14, and $|d_{Perm}|$ = 8.98, respectively; ps < .001). Finally, in Experiment 2b, the average distance between unweighted graph matrices was 7.3 when edge density was 33% ($\underline{X}_{Perm}$ = 7.69), 8.18 when edge density was 50% ($\underline{X}_{Perm}$ = 8.49), and 8.19 when we used a fixed mid-scale threshold ($\underline{X}_{Perm}$ = 8.57). Again, for each threshold these average distances were significantly

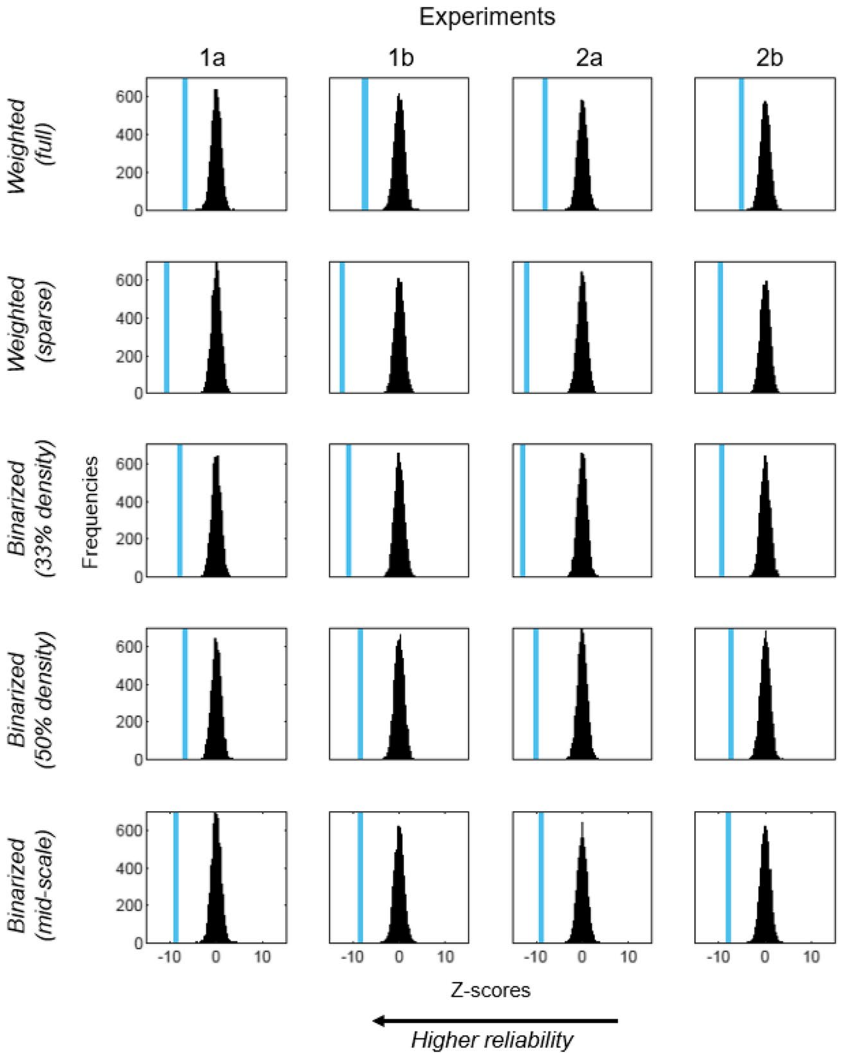


Fig. 3 Permutation results for comparison of distances between adjacency matrices constructed from the two relatedness tasks. The distribution (black) shows *z*-scored means obtained from permuting adjacency matrices of each task across 70 participants and averaging the obtained distance scores across the permuted sample. This procedure was repeated 10,000 times to obtain a null distribution of sample means. The line is the z-scored mean of distances between matrices in the unpermuted sample where the correspondence between participants is preserved across tasks. Blue indicates that the unpermuted mean of distances was statistically smaller than in the null distribution

smaller within subjects than in the permuted sample ($|d_{Perm}| = 9.5$, $|d_{Perm}| = 7.42$, and $|d_{Perm}| = 8.01$, respectively; $ps < .001$).

Summary Figure 3 shows results of all permutation analyses. Altogether, across all four experiments, and regardless of whether semantic network was represented as a weighted or binary graph, regardless of what filtering procedure we used to construct graphs, and regardless of how we collected data in the pairwise semantic relatedness task and which stimuli we used, the distance between adjacency matrices was significantly smaller when comparisons were made within the same individual than when made across individuals. These results were consistent when we used the Pearson correlation instead of Euclidean distance to quantify relatedness between matrices. This indicates that there are stable individual differences in local metrics of semantic networks that generalize across tasks.

Node importance: Eigenvector centrality

Next, we examine the consistency of local metrics of semantic networks across tasks in terms of the relative node importance by calculating "eigenvector centrality" for each node (as in Bieth et al., 2021). This measure collapses the 20×20 adjacency matrix onto a vector of length 20, and thus summarizes the full semantic content of the network in terms of a single scalar property of each concept: how important is it compared to the other concepts? Eigenvector centrality of a given concept within a semantic network captures the degree to which a given concept is related to other interconnected concepts: nodes are more central when they are more heavily connected to other nodes, and when they are heavily connected to other heavily connected nodes. For directed graphs, the eigenvector centrality is related to the PageRank algorithm, which can be used to characterize the relative importance of websites given their link structure (Page et al., 1997), or words, given free association data like the Nelson norms (e.g., Griffiths et al., 2007). Nodes with higher centrality scores have more importance because they are linked to more sources of information (other semantic units) and can be reached with greater efficiency within the network.

Eigenvector centrality (c^E) is related to the eigenvector decomposition of the adjacency matrix of the graph (Latora et al., 2017):

$$Ac^E = \lambda c^E, \tag{2}$$

where A is the similarity adjacency matrix for the semantic graph, and λ and c^E are its (largest) eigenvalue and the absolute value of its corresponding eigenvector, respectively. Technically, the eigenvector centrality score of each node is proportional to the sum of the centralities of its neighboring nodes.

As in the other analyses, to evaluate the consistency of the centralities of all concepts across tasks, we calculated the correlation between eigenvector centrality scores of the 20 words from the two tasks for each participant, and then averaged these correlations across subjects. We compared this average correlation to the distribution of average correlations obtained from 10,000 permutation iterations.

Weighted graphs For all experiments we found that correlations of centralities across tasks were greater within subjects than in the permuted samples. In Experiment 1a, the average correlation across participants between node centrality scores from the Likert and spatial arrangement tasks for the same subject was $r_{Avg} = .21$ for the full weighted matrix ($r_{Avg,Perm} = 07$) and $r_{Avg} = .13$ for the sparse matrix ($r_{Avg,Perm} = -.015$). In both cases, averaged correlations were higher within subjects than in the permuted samples ($|d_{Perm}| = 4.87$, and $|d_{Perm}| = 3.29$, respectively; $ps < .002$). In Experiment 1b, the average correlation between node centrality scores from the slider and spatial arrangement tasks for the same subject was $r_{Avg} = .26$ for the full weighted matrix ($r_{Avg,Perm} = .055$) and $r_{Avg} = .18$ for the sparse matrix ($r_{Avg,Perm} = .004$). Again, these average correlations were significantly higher within subjects than in the permuted samples ($|d_{Perm}| = 6.39$, and $|d_{Perm}| = 4.26$, respectively; $ps < .001$). In Experiment 2a, the average correlation between node centrality scores from the Likert and spatial arrangement tasks for the same subject was $r_{Avg} = .58$ for the full weighted matrix ($r_{Avg,Perm} = .46$) and $r_{Avg} = .57$ for the sparse matrix ($r_{Avg,Perm} = .46$). These average correlations were higher with participants than in the permuted samples ($|d_{Perm}| = 8.01$, and $|d_{Perm}| = 5.94$, respectively; $ps < .001$). Finally, in Experiment 2b, the average correlation between node centrality scores from the slider and spatial arrangement for the same subject was $r_{Avg} = .58$ for the full weighted matrix ($r_{Avg,Perm} = .49$) and $r_{Avg} = .59$ for the sparse matrix ($r_{Avg,Perm} = .50$). Again, these average correlations were higher within subjects than in the permuted samples ($|d_{Perm}| = 6.31$, and $|d_{Perm}| = 5.84$, respectively; $ps < .001$).

Binarized graphs We also found that node centrality scores were more consistent across tasks from the same subject than for randomly paired subjects in all four experiments for unweighted graphs with both 33% and 50% edge densities, and unweighted graphs with a mid-scale threshold. In Experiment 1a, the average node centrality score correlation between the two semantic relatedness tasks was $r_{Avg} = .16$ when edge density was 33% ($r_{Avg,Perm} = .037$), $r_{Avg} = .21$ when edge density was 50% ($r_{Avg,Perm} = .08$), and $r_{Avg} = .22$ when we used a fixed mid-scale threshold ($r_{Avg,Perm} = .07$). For all three thresholds, these average correlations were larger within subjects than in the permuted samples ($|d_{Perm}| = 3.2$, $|d_{Perm}| = 3.98$ and $|d_{Perm}| = 4.62$, respectively; $ps < .001$). We found the same pattern of results in Experiment 1b, where the average correlation in node centrality scores across the two semantic relatedness tasks was $r_{Avg} = .21$ when edge density was 33% ($r_{Avg,Perm} = .01$), $r_{Avg} = .22$ when edge density was 50% ($r_{Avg,Perm} = .07$), and $r_{Avg} = .17$ for the same subject when we used a fixed mid-scale threshold ($r_{Avg,Perm} = .05$). For all three thresholds, these average correlations were larger within subjects than in the permuted samples ($|d_{Perm}| = 5.0$, $|d_{Perm}| = 5.01$ and $|d_{Perm}| = 3.68$, respectively; $ps < .001$). Likewise, in Experiment 2a, the average correlation of node centrality scores was $r_{Avg} = .6$ when edge density was 33% ($r_{Avg,Perm} = .49$), $r_{Avg} = .54$ when edge density was 50% ($r_{Avg,Perm} = .43$), and $r_{Avg} = .24$ when we used a fixed mid-scale threshold ($r_{Avg,Perm} = .07$). For all three thresholds, these average correlations were larger within subjects than in the permuted samples ($|d_{Perm}| = 6.42$, $|d_{Perm}| = 6.58$ and $|d_{Perm}| = 4.97$, respectively; $ps < .001$). Finally, in Experiment 2b, the average correlation of node centrality scores was $r_{Avg} = .59$ when edge density was 33% ($r_{Avg,Perm} = .497$), $r_{Avg} = .51$ when edge density was 50% ($r_{Avg,Perm} = .45$), and $r_{Avg} = .52$ when we used a fixed mid-scale threshold ($r_{Avg,Perm} = .45$). For all three thresholds, these average correlations were larger within subjects than in the permuted samples ($|d_{Perm}| = 5.83$, $|d_{Perm}| = 3.66$ and $|d_{Perm}| = 4.55$, respectively; $ps < .001$).

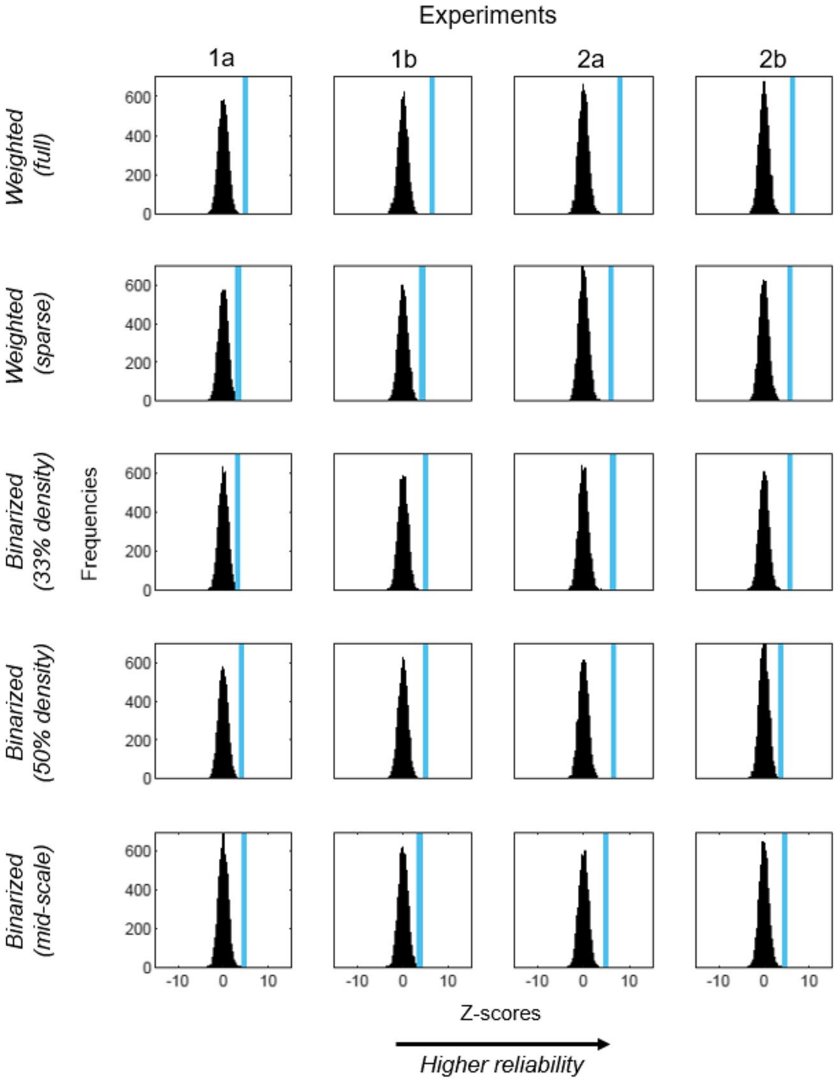


Fig. 4 Permutation results for comparison of correlations in centrality scores for 20 words obtained from the two relatedness tasks. The distribution (black) shows *z*-scored average correlation coefficients obtained from permuting centrality scores of each task across 70 participants and averaging the obtained correlations across the permuted sample. This procedure was repeated 10,000 times to obtain a null distribution of sample average correlations. The line is the *z*-scored mean correlation between centrality scores obtained in the unpermuted sample, where the correspondence between participants is preserved across tasks. Blue indicates that the unpermuted mean was statistically greater than the null distribution

Summary Figure 4 shows results of all permutation analyses. Once again, across all four experiments, regardless of whether semantic network was represented as a weighted or binary graph, regardless of what filtering procedure we used to construct graphs, and regardless of how we collected data in the pairwise semantic relatedness task and which stimuli we used, node centrality scores across the two semantic relatedness tasks were more robust when comparisons were made within the same individual than when made across individuals. This provides converging evidence that there are stable individual differences in local metrics of semantic networks that generalize across tasks.

Consistency of global metrics of semantic networks

So far, our analyses suggest that there are reliable individual differences in local metrics of semantic networks, indicating that people have consistent idiosyncrasies to their semantics that show up across tasks. As reviewed in the introduction, however, researchers often are interested not in individual differences in the semantic associations of the set of stimuli investigated, but seek to characterize individual differences in some global topological property of the semantic space. For instance, perhaps some people have greater separation in their semantic representation between different content domains, yielding an overall greater tendency for concepts to cluster. Or perhaps others have a greater degree of interconnection among all concepts. Insofar as there are stable individual differences in such global properties of the overall structure of the semantic network, they might be used to predict individual differences in a broad range of tasks beyond the specific concepts whose similarity had been measured, like creativity (Benedek, et al., 2017).

We focused on two global properties that characterize distinct aspects of the semantic network and have previously been used to characterize the structure of semantic networks: (1) the extent to which concepts are clustered, as measured by the average clustering coefficient, and (2) the interconnectedness of the semantic network, as measured by the average shortest path length (e.g., Benedek et al., 2017; Griffiths et al., 2007).

Before presenting our findings, we point out that measures of average clustering coefficients and average path length are correlated with the number of edges in a graph. This poses a potential problem for analyzing binary graphs with an edge density threshold that were constructed from the Likert pairwise semantic relatedness tasks, because this scale is coarse and not all participants can meet an exact cutoff of 33% or 50% edge densities. To correct for this in Experiments 1a and 2a, we quantified the associations for average shortest path length and clustering coefficient using the partial correlation coefficient, controlling for individual differences in the number of edges across people's graphs. This provides an indirect way of examining whether these measures capture individual differences in network topology that are not due to potential differences in response policies. For each comparison, we report the significance of this partial correlation coefficient. Under conditions where the partial correlation coefficient is statistically significant, we also report the results of the permutation test, which tells us whether the measure preserves individual differences in the data.

Clustering: Average clustering coefficient

The average clustering coefficient captures the degree to which concepts are closely connected within clusters with more distant connections between clusters. The local clustering coefficient (C_j) of a particular node j quantifies the degree to which that node's neighbors (adjacent nodes) are connected to one another and can be defined so long as the target node has two or more neighbors. Insofar as concepts are clustered, then neighbors of a node will be more heavily interconnected. In an unweighted graph, the local clustering coefficient for a node j is given by the formula (Watts & Strogatz, 1998)

$$C_j = \frac{v_j}{n_j(n_j - 1)/2}, \tag{3}$$

where the numerator v_j is the total number of edges present between the n_j neighbors of node j, and the expression in the denominator is the maximum number of possible edges between those neighbors. If a node has fewer than two neighbors, $n_j< 2$, the denominator is 0, and the local clustering coefficient is undefined. For weighted graphs, the local clustering coefficient is given by the formula

$$C_j = \frac{\sum_i \sum_k w_{ji} w_{ik} w_{kj}}{\sum_i \sum_{k \neq i} w_{ji} w_{jk}}, \tag{4}$$

where the numerator is triads of nodes (constructed with respect to node j) weighted by the strength of connections between nodes within the triad, and the denominator is the sum of node–pair connections with respect to node j (within the triad), which sets the upper bound on the product of triad weights in the numerator (Kalna & Higham, 2006; Zhang & Horvath, 2005). More intuitively, this formula describes the weighted average connection strength between two neighbors of a node, weighted by the product of the node's connection to those two neighbors. This has the property that as edge weights approach the unweighted, binary limits of 1 and 0, the weighted clustering coefficient approaches the unweighted clustering coefficient. The average clustering coefficient is the average of the local clustering coefficients across all nodes in the graph.

Weighted graphs In Experiment 1a, the average clustering coefficient for full weighted graphs was .45 and .44 in the Likert task and spatial arrangement task, respectively, and the average clustering coefficient for sparse weighted graphs was .39 and .43 in the Likert task and spatial arrangement task, respectively. The correlation between clustering coefficients across tasks was not statistically significant in the unpermuted sample for full weighted graphs ($r(68) = .208$, $p = 0.084$) or for sparse weighted graphs ($r(68) = .103$, $p = .39$). In Experiment 1b, the average clustering coefficient for full weighted graphs was .45 and .44 in the slider task and spatial arrangement task, respectively, and the average clustering coefficient for sparse weighted graphs was .34 and .45 in the slider task and spatial arrangement task, respectively. The correlation between clustering coefficients across tasks was statistically significant for full weighted graphs ($r(68) = .301$, $p = .011$) and statistically more robust within the same subject than in the permuted sample ($r_{Perm} = -.004; |d_{Perm}| = 2.54$; $p = .024$). The correlation between clustering coefficients across tasks in Experiment

1b was not statistically significant for sparse weighted graphs ($r(68) = .053$, $p = .663$). Likewise, in Experiment 2a, the average clustering coefficient for full weighted graphs was .52 and .52 in the Likert task and spatial arrangement task, respectively, and the average clustering coefficient for sparse weighted graphs was .44 and .46 in the Likert task and spatial arrangement task, respectively. The correlation between clustering coefficients across tasks was statistically significant for full weighted graphs ($r(68) = .243$, $p = 0.043$), but it was not statistically more consistent within subjects than in the permuted sample after correcting for multiple comparisons ($r_{Perm} = -.02; \left|d_{Perm}\right| = 2.018$; $p = .084$). The correlation between clustering coefficients across tasks in Experiment 2a was statistically significant for sparse weighted graphs ($r(68) = .472$, $p < .001$) and was significantly more consistent within subjects than in the permuted sample ($r_{Perm} = -.01; \left|d_{Perm}\right| = 3.99$; $p < .001$). Finally, in Experiment 2b, the average clustering coefficient for full weighted graphs was .51 and .54 in the slider task and spatial arrangement task, respectively, and the average clustering coefficient for sparse weighted graphs was .34 and .45 in the slider task and spatial arrangement task, respectively. The correlation between clustering coefficients across tasks was not statistically significant for full weighted graphs ($r(68) = -.011$, $p = .92$) or for sparse weighted graphs ($r(68) = .167$, $p = .17$).

Binarized graphs In Experiment 1a, the average unweighted clustering coefficient was .54 and .62 when using the 33% edge density thresholding criterion in the Likert task and spatial arrangement task, respectively; was .61 and .64 when using the 50% edge density thresholding criterion in the Likert and spatial arrangement task, respectively; and was .58 and .65 in the Likert and spatial arrangement task when using a fixed mid-scale threshold, respectively. The partial correlation coefficient between clustering coefficients across tasks was significant in the 33% edge density threshold (r_P $(68) = .31$, $p = .011$) and was significantly more robust within people than in the permuted sample ($r_{P,Perm} = .035; \left|d_{Perm}\right| = 2.25$; $p = .0452$). The partial correlation between clustering coefficients was not significant with the 50% edge density thresholding criterion (r_P $(68) = .23$, $p = .06$), and the correlation was not significant with the fixed mid-point criterion ($r(68) = .10$, $p = .4$). In Experiment 1b, the average unweighted clustering coefficient was .53 and .61 when using the 33% edge density thresholding criterion in the slider task and spatial arrangement task, respectively; was .63 and .65 when using the 50% edge density thresholding criterion in the slider and spatial arrangement task, respectively; and was .59 and .65 in the Likert and spatial arrangement task when using a fixed mid-scale threshold, respectively. The correlation between clustering coefficients across tasks was not significant in the 33% edge density threshold ($r(68) = .12$, $p = .18$). The correlation between clustering coefficients was significant with the 50% edge density threshold ($r(68) = .28$, $p = .021$) and was more robust within people than in the permuted sample ($r_{Perm} = -.004; \left|d_{Perm}\right| = 2.27$; $p = .046$). The correlation between clustering coefficients was not significant with the fixed mid-point criterion ($r(68) = .07$, $p = .57$). In Experiment 2a, the average clustering coefficient was .53 and .63 when using the 33% thresholding criterion in the Likert task and spatial arrangement task, respectively; was .61 and .67 when using the 50% thresholding criterion in the Likert and spatial arrangement task, respectively; and was .58 and .64 with the fixed mid-point criterion, respectively. The partial correlation between clustering coefficients across tasks was significant in the unpermuted sample for both the 33% (r_P $(68) = .47$, $p < .001$; $r_{P,Perm} = -.01$) and 50% (r_P $68) = .61$, $p < .001$; $r_{P,Perm} = -.02$) thresholding criteria. For both thresholds, partial correlations were more robust within people than in the permuted sample ($\left|d_{Perm}\right| = 3.96$ and $\left|d_{Perm}\right| = 4.89$, respectively; $ps < .001$). The correlation between clustering coefficients was not significant when using a fixed mid-point threshold ($r(68) = .092$, $p = .45$). Finally, in Experiment 2b, the average clustering coefficient was .51 and .61 when using the 33% thresholding criterion in the slider task and spatial arrangement task, respectively; was .62 and .66 when using the 50% thresholding criterion in the slider and spatial arrangement task, respectively; and was .58 and .66 with the fixed mid-point criterion, respectively. The correlation between clustering coefficients across tasks was not significant with the 33% edge density threshold ($r(68) = -.004$, $p = .97$). The correlation was significant with a 50% edge density threshold ($r(68) = .28$, $p = .019$) and was more consistent within people than in the permuted sample ($r_{Perm} = -.01; \left|d_{Perm}\right| = 2.30$; $p = .04$). The correlation between clustering coefficients was not significant when using a fixed mid-point threshold ($r(68) = -.115$, $p = .35$).

Summary Figure 5 shows results of all permutation analyses. Together, we do not find robust evidence that measures of semantic network clustering reliably capture individual differences across the two semantic relatedness tasks. While we do find some statistically significant effects in some experiments (e.g., Experiment 2a), the associations tend to be low to moderate and are not robust across each of the filtering criteria. More importantly, we do not replicate these with small variations in experimental design, such as change in stimuli or change in the granularity of the pairwise semantic relatedness task.

Interconnectedness: Average shortest path length

The efficiency and interconnectedness of a graph corresponds to how easily one can traverse from one node to

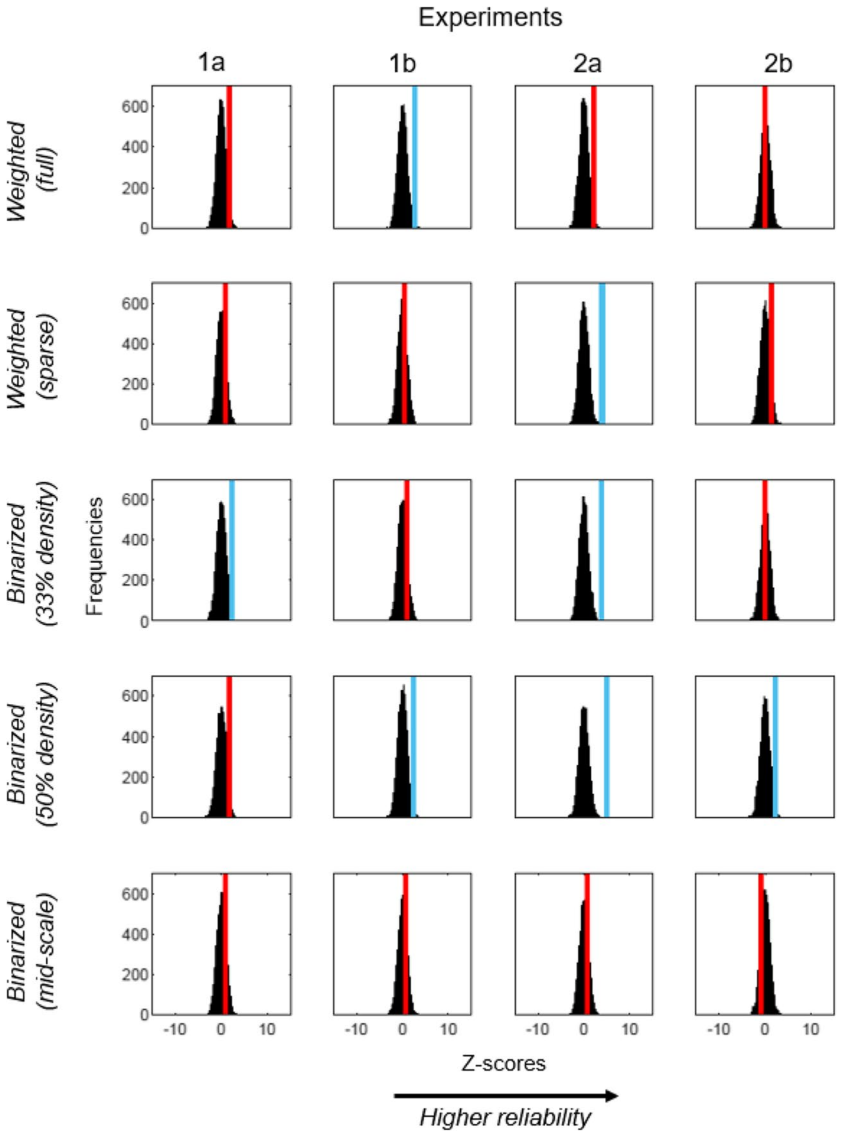


Fig. 5 Permutation results for comparison of correlation clustering coefficients of graphs across participants. The distribution (black) shows *z*-scored correlation coefficients obtained from permuting clustering coefficients computed from graphs of each task across 70 participants. This procedure was repeated 10,000 times to obtain a null distribution of sample correlations. The line is the *z*-scored correlation between clustering coefficients obtained in the unpermuted sample, where the correspondence between participants is preserved across tasks. Blue and red lines indicate that the unpermuted mean was either statistically greater than or not statistically different from the null distribution, respectively

another. While there are several possible measures of this efficiency, the simplest is the average shortest path length. In an unweighted, sparse graph, the shortest path length between nodes *i* and *j* refers to the minimum number of links that need to be traversed from node *j* to reach node *i* (Barabasi, 2016). The average shortest path length is computed by averaging the shortest path length over all $n^*(n-1)/2$ pairs of nodes, yielding one number corresponding to the overall interconnectedness of the network that captures how spread out semantic units are within the network or, conversely, how "efficiently" information traverses the network. Is this measure of overall semantic network connectivity stable across tasks for a given subject?

The average shortest path length (L) is given in Eq. 5,

$$L = \frac{1}{\left(\frac{N(N-1)}{2}\right)} \sum\nolimits_{i \neq j} d_{i,j} \tag{5}$$

where $d_{i,j}$ is the shortest distance between nodes i and j, N is the number of distinct nodes in the network, and the expression in the denominator is the number of unique node pairs in an undirected graph.

For unweighted, binary graphs, the distance of a particular path between two nodes is defined as the total number of edges along that path. Thus, the shortest path between two nodes (d_{ij}) is the number of edges in the path between nodes *i* and *j* that has the fewest edges. For some tasks and participants in our sample, semantic networks were disconnected, so the shortest path was undefined for some node pairs; we excluded these node pairs from that person's average.

For weighted graphs, all nodes are connected, but with edges of varying weight. We adopt the conventional approach of defining distance for a weighted edge to be 1/weight, and thus the length of a path is defined as the sum of the distances along that path. The shortest path between nodes *i* and *j* (d_{ij}) is defined as the path with the smallest sum of distances. Note that in our case, edge weights are given by scaled similarities, and thus edge distances are proportional to the semantic proximity between word pairs.

Weighted graphs In Experiment 1a, the average shortest path length for full weighted graphs was .73 and 1.07 in the Likert task and spatial arrangement task, respectively, and the averaging shortest path length score for sparse weighted graphs was .82 and .69 in the Likert task and spatial arrangement task, respectively. The correlation between average shortest path length scores across tasks was not statistically significant in the unpermuted sample for full weighted graphs ($r(68) = .11$, $p = 0.37$) or for sparse weighted graphs ($r(68) = .204$, $p = .09$). In Experiment 1b, the average shortest path length for full weighted graphs was .36 and .52 in the slider task and spatial arrangement task, respectively, and the average shortest path length for sparse weighted graphs was .57 and .69 in the slider and spatial arrangement task, respectively. The correlation between average shortest path length across tasks was not statistically significant for full weighted graphs ($r(68) = .23$, $p = .06$), but it was significant for sparse weighted graphs ($r(68) = .53$, $p < .001$), and was more robust within people than in the permuted sample ($r_{Perm} = -.01; \left|d_{Perm}\right| = 4.46; p < .001$). In Experiment 2a the shortest path length for full weighted graphs was .5 and .88 in the Likert task and spatial arrangement task, respectively, and the averaging shortest path length for sparse weighted graphs was .67 and .54 in the Likert task and spatial arrangement task, respectively. The correlation between average shortest path length across the two tasks was not statistically significant for full weighted graphs ($r(68) = .14$, $p = 0.26$) or for sparse weighted graphs ($r(68) = -.032$, $p = 0.79$). Finally, in Experiment 2b the average shortest path length for full weighted graphs was .29 and .41 in the slider and spatial arrangement task, respectively, and the average shortest path length for sparse weighted graphs was .44 and .51 in the slider task and spatial arrangement task, respectively. The correlation between average shortest path length scores across tasks was not statistically significant for full weighted graphs ($r(68) = -.014$, $p = .9$) or for sparse weighted graphs ($r(68) = -.076$, $p = .54$).

Binarized graphs In Experiment 1a, the average shortest path length was 3.61 and 3.75 when using the 33% thresholding criterion in the Likert task and spatial arrangement task, respectively; was 3.12 and 3.16 when using the 50% edge density thresholding criterion in the Likert and spatial arrangement task, respectively; and was 1.65 and 1.53 in the Likert and spatial arrangement task when using a fixed mid-scale threshold, respectively. The partial correlation coefficient between average path length scores across tasks was significant with the 33% edge density threshold (r_P $(68) = .305$, $p = .01$), but not more robust within people than in the permuted sample after correcting for multiple comparisons ($r_{P,Perm} = .13; \left|d_{Perm}\right| = .65; p = 1$). The partial correlation between average path length across the two tasks was not significant in the right direction with a 50% edge density threshold ($r(68) = -.256$, $p = .03$). The correlation in average path length was not significant with a fixed mid-point threshold ($r(68) = -.155$, $p = .2$). In Experiment 1b, the average path length was 1.83 and 1.91 when using the 33% edge density thresholding criterion in the slider and spatial arrangement task, respectively; was 1.55 and 1.56 when using the 50% edge density thresholding criterion in the slider and spatial arrangement task, respectively; and was 1.73 and 1.53 in the slider and spatial arrangement task when using a fixed mid-scale threshold, respectively. The correlation between average path length scores across tasks was not significant in the 33% edge density threshold ($r(68) = .08$, $p = .51$). The correlation between average path length scores was significant with the 50% edge density threshold ($r(68) = .478$, $p < .001$) but was not more robust within people than in the permuted sample ($r_{Perm} = .40; \left|d_{Perm}\right| = .33; p = 1$). The correlation between average shortest path length scores was not significant with the fixed mid-point criterion ($r(68) = .055$, $p = .65$). In Experiment 2a, the average shortest path length was 3.65 and 3.76 when using the 33% thresholding criterion in the Likert task and spatial arrangement task, respectively; was 3.07 and 3.11 when using the 50% thresholding criterion in the Likert and spatial arrangement task, respectively; and was 1.65 and 1.59 with the fixed mid-point criterion, respectively. The partial correlation between average path length scores across tasks was not significant in the unpermuted sample with the 33% edge density threshold ($r(68) = .16$, $p = .19$), was not significant with the 50% edge density threshold ($r(68) = .19$, $p = .11$), and was not significant when using a fixed mid-point threshold ($r(68) = .043$, $p = .72$). Finally, in Experiment 2b, the average shortest path length was 1.89 and 1.91 when using the 33% thresholding criterion in the slider task and spatial arrangement task, respectively; was 1.51 and 1.53 when using the 50% thresholding criterion in the slider and spatial

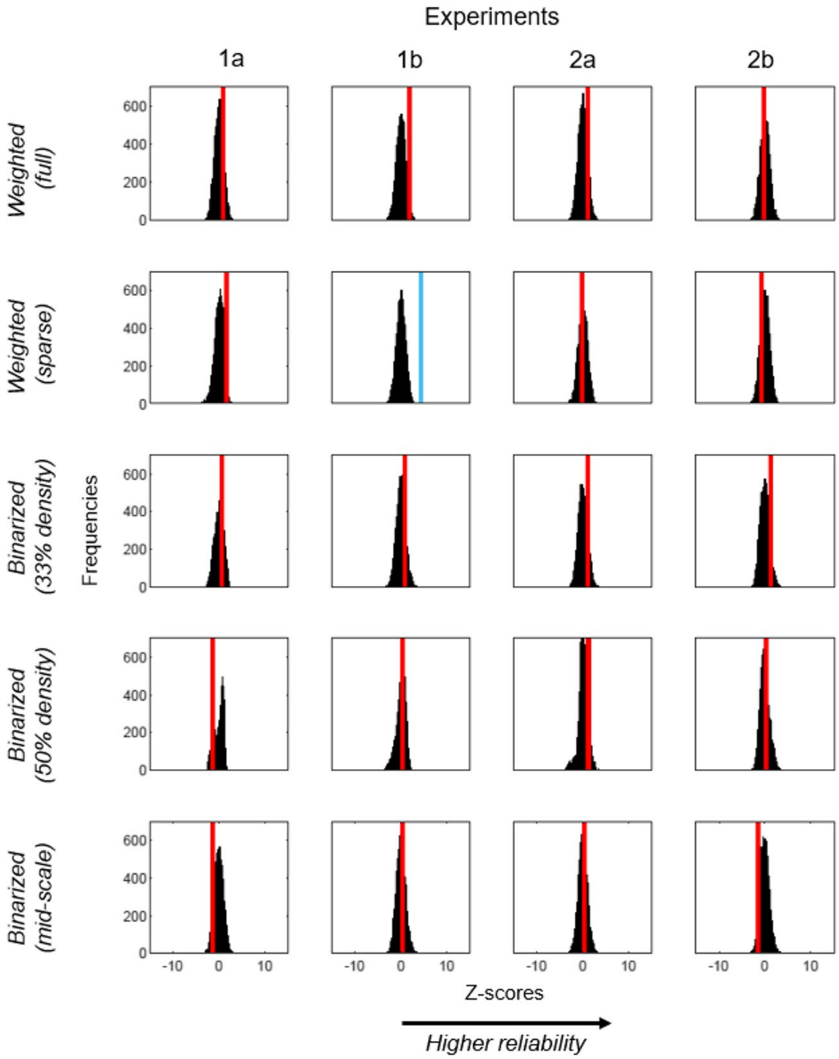


Fig. 6 Permutation results for comparison of correlations in average path length of graphs across participants. The distribution (black) shows *z*-scored correlations obtained from permuting average path length scores of each task across 70 participants. This procedure was repeated 10,000 times to obtain a null distribution of sample correlations. The line is the *z*-scored correlation of average path length across tasks obtained in the unpermuted sample, where the correspondence between participants is preserved across tasks. Blue and red lines indicate that the unpermuted mean was either statistically greater than or not statistically different from the null distribution, respectively

arrangement task, respectively; and was 1.64 and 1.57 with the fixed mid-point criterion, respectively. The partial correlation between average path length scores was not significant with the 33% edge density threshold ($r(68) = .209$, $p = .083$) or the 50% edge density threshold ($r(68) = .067$, $p = .58$), or with a fixed mid-point threshold ($r(68) = -.16$, $p = .183$).

Summary Figure 6 shows results of all permutation analyses. Together, once again, we do not find robust evidence that a measure of semantic network efficiency (average path length) reliably captures individual differences across the two semantic relatedness tasks. In general, associations of this metric across tasks tend to be low to moderate and are not robust across each of the filtering criteria. We also fail to replicate these with small variations in experimental design, such as change in stimuli or change in the granularity of the pairwise semantic relatedness task.

Summary across experiments and analyses The collective set of results is summarized in Table 1. Together, these analyses suggest that, unlike metrics of semantic network content, metrics of network structure are not as reliable across different experiments and ways of representing the data. With weighted and binarized graphs we did not find a reliable association in the average clustering of graphs or average shortest path length across experiments, or thresholding criteria. Furthermore, we found that associations between these metrics across the two semantic relatedness tasks were

Table 1 Summary of core findings from all experiments. Descriptive statistics are given from within sample analyses. X_D denotes the mean Euclidean distance between similarity matrices in the unpermuted sample, r denotes the Pearson correlation coefficient across tasks for the same stimuli and participants, and r_{Avg} denotes the average of correlation coefficients within participants (see Results section for additional details). All p-values are from the permutation analysis and are corrected with Bonferroni correction for multiple (four) comparisons. Blue boxes denote significant p-values. In the Summary column, ✓ indicates that a given measure reliably captures individual differences in semantic network properties across each of the filtering criteria, — indicates that a given metric captured individual differences, but not reliably across filtering criteria, and ✗ indicates that the metrics did not capture individual differences for any of the thresholding criteria. As can be seen, content-based but not structure-based metrics tended to be reliable across tasks, filtering methods, and network representation structures

Experiment	Network properties	Network representation scheme					Summary
		Weighted (full)	Weighted (sparse)	Binarized (33% edge density)	Binarized (50% edge density)	Binarized (mid-scale)	
1a	Node similarity	$X_D = 4.32$; $p < .001$	$X_D = 4.56$; $p < .001$	$X_D = 7.02$; $p < .001$	$X_D = 7.76$; $p < .001$	$X_D = 8.15$; $p < .001$	✔
	Node centrality	$r_{Avg} = .21$; $p < .001$	$r_{Avg} = .13$; $p = .0016$	$r_{Avg} = .16$; $p = .0032$	$r_{Avg} = .21$; $p < .001$	$r_{Avg} = .22$; $p < .001$	✔
	Network clustering	$r = .21$; $p = .15$	$r = .10$; $p = .12$	$r_P = .31$; $p = .045$	$r_P = .23$; $p = .18$	$r = .10$; $p = .77$	—
	Network mean path length	$r = .11$; $p = .66$	$r = .20$; $p = .76$	$r_P = .31$; $p = 1$	$r_P = -.26$; $p = 1$	$r = -.16$; $p = 1$	✘
1b	Node similarity	$X_D = 4.11$; $p < .001$	$X_D = 4.15$; $p < .001$	$X_D = 7.07$; $p < .001$	$X_D = 7.97$; $p < .001$	$X_D = 8.14$; $p < .001$	✔
	Node centrality	$r_{Avg} = .26$; $p < .001$	$r_{Avg} = .18$; $p < .001$	$r_{Avg} = .21$; $p < .001$	$r_{Avg} = .22$; $p < .001$	$r_{Avg} = .17$; $p = .0012$	✔
	Network clustering	$r = .30$; $p = .024$	$r = .05$; $p = 1$	$r = .12$; $p = .66$	$r = .28$; $p = .046$	$r = .07$; $p = 1$	—
	Network mean path length	$r = .23$; $p = .11$	$r = .53$; $p < .001$	$r = .08$; $p = .87$	$r = .48$; $p = 1$	$r = .06$; $p = 1$	—
2a	Node similarity	$X_D = 4.34$; $p < .001$	$X_D = 5.43$; $p < .001$	$X_D = 7.00$; $p < .001$	$X_D = 7.9$; $p < .001$	$X_D = 7.88$; $p < .001$	✔
	Node centrality	$r_{Avg} = .58$; $p < .001$	$r_{Avg} = .57$; $p < .001$	$r_{Avg} = .6$; $p < .001$	$r_{Avg} = .54$; $p < .001$	$r_{Avg} = .24$; $p < .001$	✔
	Network clustering	$r = .24$; $p = .09$	$r = .47$; $p < .001$	$r_P = .47$; $p < .001$	$r_P = .61$; $p < .001$	$r = .09$; $p = .89$	—
	Network mean path length	$r = .13$ $p = .55$	$r = -.03$; $p = 1$	$r_P = .16$; $p = .54$	$r_P = .19$; $p = .39$	$r = .04$; $p = 1$	✘
2b	Node similarity	$X_D = 4.3$; $p < .001$	$X_D = 5.38$; $p < .001$	$X_D = 7.30$; $p < .001$	$X_D = 8.18$; $p < .001$	$X_D = 8.19$; $p < .001$	✔
	Node centrality	$r_{Avg} = .58$; $p < .001$	$r_{Avg} = .59$; $p < .001$	$r_{Avg} = .59$; $p < .001$	$r_{Avg} = .51$; $p < .001$	$r_{Avg} = .52$; $p < .001$	✔
	Network clustering	$r = -.011$; $p = 1$	$r = .17$; $p = .32$	$r = -.004$; $p = 1$	$r = .28$; $p = .04$	$r = -.12$; $p = 1$	—
	Network mean path length	$r = -.014$; $p = 1$	$r = -.08$; $p = 1$	$r = .21$; $p = .44$	$r = .07$; $p = 1$	$r = -.16$; $p = 1$	✘

in the low to moderate range. We discuss the implications of these results in greater detail in the general discussion.

Summary of core findings from all experiments. Descriptive statistics are given from within sample analyses. X_D denotes the mean Euclidean distance between similarity matrices in the unpermuted sample, r denotes the Pearson correlation coefficient across tasks for the same stimuli and participants, and r_{Avg} denotes the average of correlation coefficients within participants (see Results section for additional details). All p-values are from the permutation analysis and are corrected with Bonferroni correction for multiple (four) comparisons. Blue boxes denote significant p-values. In the Summary column, ✓ indicates that a given measure reliably captures individual differences in semantic network properties across each of the filtering criteria, — indicates that a given metric captured individual differences, but not reliably across filtering criteria, and ✗ indicates that the metrics did not capture individual differences for any of the thresholding criteria. As can be seen, content-based but not structure-based metrics tended to be reliable across tasks, filtering methods, and network representation structures

General discussion

The goal of this study was to directly test whether there are individual differences in concept representation schemes by examining whether individual differences in semantic relatedness judgments are invariant across different measurement approaches. Across four experiments with different stimuli, methods for collecting semantic relatedness judgments, and samples of participants, we found evidence for such individual differences. We found that we could predict some individual differences in the properties of semantic networks across structurally different semantic relatedness tasks using simple concepts. Our work has both theoretical and methodological implications for conceptualizing and measuring individual differences in semantic representation schemes. In particular, we lay the groundwork for conducting studies of individual differences in semantic representations by demonstrating that individual differences in some properties of semantic network structure are not simply driven by how individuals approach the tasks. Our research also highlights important measurement issues and boundary conditions for this research, which we discuss next.

Quantifying individual differences in semantic networks

Local properties of semantic networks

We found that we could detect individual differences in the semantic relatedness between pairs of words across qualitatively different semantic similarity tasks using both binarized and weighted representations of the semantic graph. This finding is important because binarizing the data involves potentially losing meaningful information about individuals' semantic judgments, which may be otherwise preserved in the continuous data. Our finding that binarizing the data does not hurt predictive accuracy, therefore, indicates that substantive variation is preserved under this type of transformation. We underscore that these results cannot be driven by choice of a thresholding criterion, because we used the same thresholding criterion for each of the two tasks in both our unpermuted and permuted sample. More broadly, our results support the view that there are robust individual differences in semantic relatedness, and that these can be reliably captured with different network-based representation schemes.

Furthermore, we find that individual differences in measures of concept "importance" or centrality were recoverable across the tasks. In particular, the same concepts tended to be more highly interconnected with other concepts in the network across the two similarity tasks. Our results are aligned with recent findings from Wulff et al. (2022b), who also reported reliable individual differences in properties of semantic networks content across participants. Our work moves beyond this work by using a large sample size (70 versus 8 participants) and a wider range of methodological and analytic approaches.

Global properties of semantic networks

We also examine the reliability of global metrics of semantic network topology across different measurement approaches. Unlike local metrics that capture how specific semantic units relate to one another, global metrics generalize across the specific units and try to characterize some general properties of the semantic graph.

We find that global metrics like network clustering and efficiency (average path length) are not particularly reliable across different task structures and methods of representing semantic networks. When binarizing the data, we find that the average clustering coefficient may be recoverable, but our ability to do so was not robust across different thresholding criteria or experiments. Likewise, for weighted full and sparse graphs, we did not recover the clustering coefficient reliably across experiments. For average shortest path length, we detected reliable individual differences only in Experiment 1b when analyzing weighted sparse graphs. In all other experiments, we found no statistical association in average shortest path length across the two tasks. Finally, when analyzing the clustering of graphs or the average shortest path length of graphs, we found that associations between these metrics across the two tasks were low to moderate at best.

In short, we find that we cannot recover these global metrics as reliably as local measures. Furthermore, if individual differences in global indices of network topology for simple concepts exist, these measures are expected to generalize across specific concepts: that is, for the average shortest path length to be a useful measure of someone's idiosyncratic semantic network structure, we would expect that it would be stable across semantic networks defined over different sets of concepts. We find that they are not even stable across different semantic relatedness tasks that use the same words. These results suggest that there are reliable individual differences in specific semantic associations between concepts—as captured by local metrics of semantic networks—but overall semantic network structure is highly sensitive to the method and items used to collect similarity judgments that define the network and, therefore, may not be a particularly stable individual difference.

Implications for research linking semantic network structure to other tasks

Our experiments yield an estimate of shared variance across different semantic relatedness tasks for simple concepts. This estimate sets an upper bound on the amount of shared variance between a single semantic relatedness task and tasks that probe other cognitive (e.g., creativity) processes: if a semantic network measure is not reliable across elicitation methods, it cannot reliably predict other cognitive processes. Therefore, our results suggest that individual differences in local semantic network structure may be reliably recovered and corroborate the use of similarity-based distances to quantify individual differences in semantic networks (e.g., Reilly et al., 2023).

In contrast, measures of global structure of semantic networks are relatively unreliable, as indexed by the lack of consistent associations we find between these metrics across two semantic relatedness tasks, in which we would expect associations between these metrics to be maximal. Furthermore, in instances when we do find associations between these metrics across the two tasks, we use filtering criteria that were either not used in prior work or failed to show associations between semantic network structure and indices of global function in other studies. For instance, in Experiment 2a, we find significant correlations between clustering coefficients across the two semantic relatedness tasks when analyzing binarized graphs with a fixed edge density. In contrast, Benedek, et al., (2017) failed to find associations between the structure of semantic networks in indices of creativity when using a fixed edge density filtering procedure. Therefore, even in cases when we do find associations between these metrics, these associations do not appear to be replicated in work that links these measures of network topology to other indices of global function.

Our failure to find that individual differences in global network structure are fully robust across different semantic relatedness tasks could reflect aspects of our methodological design, such as our sample size or choice of stimuli. However, we note that our sample size is either much larger (e.g., Cosgrove et al., 2023, $n = 26$ per group) or close to that used in prior studies (e.g., Benedek, et al., $n=79$ in the final analysis), and we chose our sample size in a way that permitted us to detect the upper bound for a low correlation. Finally, across four experiments we did find consistent evidence for robust associations across tasks, both at the aggregate and for local measures of individual differences. Collectively, this indicates that our findings are likely not driven by low power, but that for simple concepts and standard (undergraduate) samples of participants, individual differences in global, but not local, semantic network structure are weak if present. Nevertheless, one important future direction is to apply our validation approach with diverse samples that vary in key demographics, such as age or expertise, where there may be greater variation across people in latent knowledge structures (as suggested by Cosgrove et al., 2023).

With regard to our choice of stimuli, there are currently no predefined, theory-based criteria for choosing word stimuli to measure individual differences in semantic networks in the general population. We selected concrete, simple concepts from a validated word corpus, and therefore our criteria for choosing words are aligned closely to those reported in prior work (e.g., Benedek, et al., 2019). However, one fruitful direction for future research is to examine possible effects of concept spaces on researchers' ability to recover individual differences in metrics of network topology. A core (implicit) assumption in current research in this domain is that properties of network topology are stable across stimuli spaces. For this reason, we chose simple, concrete concepts in all our studies. However, it is possible that robust individual differences in network topology do exist but only when measuring specific stimuli spaces. This prediction aligns with evidence that knowledge structures vary by concept domains (e.g., Kemp & Tenenbaum, 2008). Therefore, closely examining the effects of psycholinguistic properties of words, concept spaces and their interactions with expertise, as well as effects of overall inter-concept relatedness, is an important direction for pinning down whether and how individual differences in semantic network topology vary in other domains.

Another potential concern may be that our semantic relatedness tasks place different demands on participants, making it difficult to find associations across them. For instance, the pairwise similarity judgment task involves making judgments about two words whereas the spatial multi-arrangement task involves making judgments about more than two words. However, our goal is to examine whether measures of latent differences in semantic network structure can be measured reliably across such nuisance variations in task structure. As such, introducing task differences of this type is precisely the point of our studies because it gives us an opportunity to isolate latent differences in semantic network structure from strategic differences in how people approach the tasks. Furthermore, both tasks have been vetted in prior work (e.g., Richie et al., 2020) as well as our own (see Appendix 3 for extended discussion) and, by design, are *more* like one another than (e.g., creativity) tasks that have been linked to semantic relatedness judgments in this literature. Finally, these concerns would not explain why we managed to find strong associations between these tasks at the aggregate, as well as reliable individual differences in local but not global metrics of semantic networks. In short, our results indicate that, given our stimuli and characteristics of our sample, global metrics are only weakly correlated across different semantic relatedness tasks at best.

It is possible that global metrics of network topology are less robust because they are highly sensitive to choice of thresholding criteria and concept spaces. It is also possible that global properties of semantic network structure arise in generative memory tasks; however, a major hurdle in using such tasks is to isolate individual differences in semantic memory organization from possible individual differences in executive function. Another possibility is that most of the variance in structural properties of individuals' semantic networks is driven by the demands of the semantic relatedness tasks rather than latent structural properties of semantic networks. For instance, Likert scale ratings of similarity might capture stable individual differences in rating biases, and when such individually biased ratings are converted into networks, they might yield individual differences in overall network density; these in turn might be correlated with another task. However, the stable individual difference driving this effect would be the overall bias in Likert judgments, rather than a task-stable feature of semantic networks. A conceptually related finding was recently reported in the neural computational modeling domain. Domhof et al. (2021) demonstrated that different approaches to brain parcellation, that is, the reduction of the dimensionality of brain networks to distinct regions, can substantially change graph theoretic measures of network topology at the level of individual subjects, including clustering and efficiency metrics. These results indicate that global graph theoretic measures of network structure at the level of individuals may be highly susceptible to peripheral differences in researchers' analytic choices (e.g., Simmons et al., 2011), meaning that in their current application, these metrics may not provide reliable measures of interindividual variation in brain network connectivity. These results align with ours, because we also did not find reliable associations between these metrics in semantic memory, even when varying several core aspects of our methodological and analytic approach, such as stimuli, rating scale granularity, and filtering criteria.

Based on our results and the results of Domhof et al. (2021), we also underscore that peripheral analytic decisions can substantially affect the results of structural graph theoretic analyses. In the current context, we find that differences in thresholding criteria can change researchers' conclusions about the presence or absence of individual differences in metrics of network topology. Therefore, to reduce researchers' degrees of freedom (e.g., Wicherts et al., 2016), we advocate that researchers either use principled a priori criteria for thresholding graphs or ensure that their results are robust across different thresholding criteria.

Another major recommendation for researchers who seek to quantify structural properties of semantic networks at the level of individuals is to collect data from multiple different semantic relatedness tasks and check whether associations between semantic network properties and other cognitive processes are reliable regardless of how semantic judgments are collected. More generally, since no method is a pure measure of latent processes, we suggest that researchers conduct replications of studies that apply graph theoretic analyses to study individual differences, using diverse techniques for collecting semantic relatedness data and various stimuli sets. Inferences about how structural network properties relate to individual differences from a single study and method should be made with caution.

Conclusion

Our work contributes to a growing research area on the application of graph theoretic analyses for understanding semantic network structure. We show that different graph-based representations may preserve individual differences in semantic memory organization for loosely related concepts. We also show which metrics of network topology are robust and which are not, given our methodology. We find that local measures such as node centrality scores are robust across different experiments and network filtering approaches. In contrast, global metrics of network structure such as average clustering coefficient and shortest path length are less reliable. Our paper serves as a springboard for refining methodology in this growing research domain, and we suggest several future directions that can help elucidate the robustness of these measures under other testing conditions. We recommend that researchers who apply these metrics to study individual differences either preregister predefined filtering criteria to avoid excess degrees of freedom or demonstrate that their results are robust across analytic decisions, as well as ensure that their results are consistent across different methods for collecting semantic relatedness data.

Appendices

Appendix 1: Technical description of the adaptive multi-arrangement task and algorithm

The adaptive multi-arrangement task requires participants to arrange items in an arena based on their semantic relatedness. On subsequent trials, the algorithm samples two items for which there is already some evidence. This ensures that dissimilarity matrices are aligned across trials. Once a pair of items is sampled, additional items are sampled if they improve "trial efficiency." Trial efficiency is the ratio between "utility benefit" and "trial cost." Trial cost is defined as the additional cost of evaluating a given pair of items and is simply the number of items sampled for a given trial (n) raised to a power X, that is, n^X. We used an exponent of 1.2 under the assumption that the time it takes items is super-linear but sub-quadratic. Trial benefit is the additional utility gained if a given pair of items is included on a trial. In this context, trial benefit is the sum of evidence utility, which is calculated using the exponential saturation function $1 - e^{-w*d}$, where w is the current evidence weight for a given word pair, which is simply the onscreen distance of that item squared. This definition of evidence utility assumes that the dissimilarity signal-to-noise ratio is proportional to the onscreen distances, such that smaller distances have a smaller signal-to-noise ratio. The evidence utility exponent d was set to 10, which is the default value used by Kriegeskorte and Mur (2012). For this formula, evidence utility is arbitrarily close to 1 as w approaches .5. For this reason, .5 is used as a criterion for terminating the algorithm; that is, once each item pair has an evidence value of .5, or times out (after 35 minutes), the experiment ends.

Since the algorithm "zooms in" on subclusters of items on different trials, a scaling factor needs to be defined that rescales and combines distances of each arrangement in a way that ignores the on-screen distance for that specific arrangement. This is implemented iteratively. A reference dissimilarity matrix is used to calculate the rescaling factor on each trial. For 20 words, the reference dissimilarity matrix can be seen as a vector of 190 values, and the rescaled matrix is this vector normalized. The values in this vector are the average of the onscreen distances weighted by their evidence utility obtained from previous trials. These values are used to rescale dissimilarity vectors for all item pairs obtained on the current and previous trials. Specifically, after the reference dissimilarity vector is normalized, entries from dissimilarity estimates from all trials are set to equal the values in the normalized reference vector. Then a new reference matrix is calculated using the evidence-weighted average of the rescaled distances. This is repeated iteratively until the root mean square of the deviations between the reference matrix from the previous and current iteration is arbitrarily close to 0.

Appendix 2: Example graph representations

Graphs

Each figure below shows binary graphs of two randomly sampled participants and the average data from Experiment 1a (Fig. 7) and 2a (Fig. 8), as well as the similarity adjacency matrices of two different sample participants and average data from Experiment 1a (Fig. 9) and 2a (Fig. 10).

Appendix 3: Secondary methodological contribution

Validation of the adaptive multi-arrangement task for detecting individual differences

A secondary methodological contribution of our research is that we are the first to show that the adaptive version of the spatial multi-arrangement task can be used to predict individual differences in semantic processing for words. Previous work by Kriegeskorte and Mur (2012) demonstrated that this task and algorithm can be used to recover the high-dimensional structure of similarity judgments for visual stimuli, and follow-up work by Charest et al. (2014) demonstrated that the task correlates with individual differences in neural representations of visual stimuli, e.g., real-world objects. Furthermore, recent work by Majewska et al. (2021) applied the adaptive spatial multi-arrangement task to a large-scale data set with verb stimuli and demonstrated that it can provide a fine-grained measure of subclasses of semantic concepts, although the authors did not examine its potential to capture individual differences. Finally, Richie et al. (2020) recently demonstrated that a non-adaptive version of the algorithm (Goldstone, 1994), which does not involve "zooming in" on clusters of objects, can be used to recover high-dimensional structures of words. Richie et al. (2020) also found that performance on this task correlates with performance on binary similarity judgment; however, they did not demonstrate that this measure is sensitive to individual differences and did not compare the robustness of different modeling approaches in their capacity to capture such individual differences. In short, our work contributes to a line of research on validating the adaptive version of the spatial multi-arrangement task. While prior work has shown its potential to recover the high-dimensional structure of similarity judgments and its relative efficiency, we show that it can also be used to recover individual differences in similarity judgments for concepts with different modeling approaches.

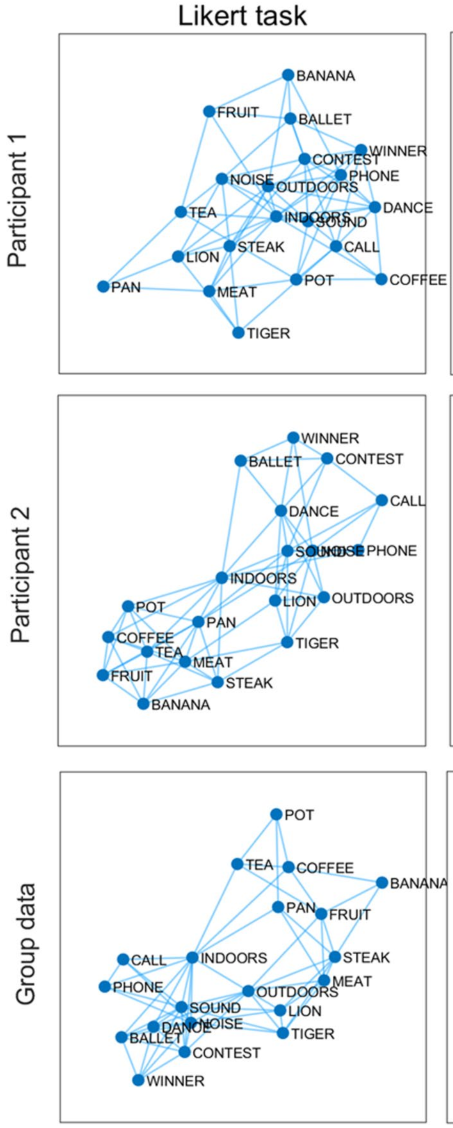


Fig. 7 Example graphs from Experiment 1 of two participants (first two upper panels) and graph constructed from average data (lower panel)

Appendix 4: Example of experiment instructions

Instructions for spatial multi-arrangement task

In this study you will complete two tasks. This is the first session and this task will take approximately 35–40 minutes to complete. The first part of the experiment is called the Word Arrangement task. It is explained below. The second part of the experiment will be described to you during the second session of the study. The Word Arrangement task requires you to arrange 20 words according to their similarity. Specifically, you will use the mouse to click on a word and drag it into a circular arena. You should use the relative distance between words to indicate how similar you think each words is relative to other words. In other words, similar objects are placed closer together; dissimilar objects are placed far apart. In the current context, the objects are blocks with words inside of them, and the distance from the center of two blocks represents their dissimilarity. If you were to place two blocks such that they completely overlap with one another, that would mean you consider the words to be identical. Consider this example...<Image of example arrangement> You will not be shown all of the 20 words at once, but will be shown subsets of the 20 (up to 10) words on each trial. Often, words will repeat across trials so that we can obtain similarity judgments between all words, and/or get more precise similarity judgements for specific words pairs. Thus, on some trials, you will see many words (up to 10), and on other trials you will see fewer words (as

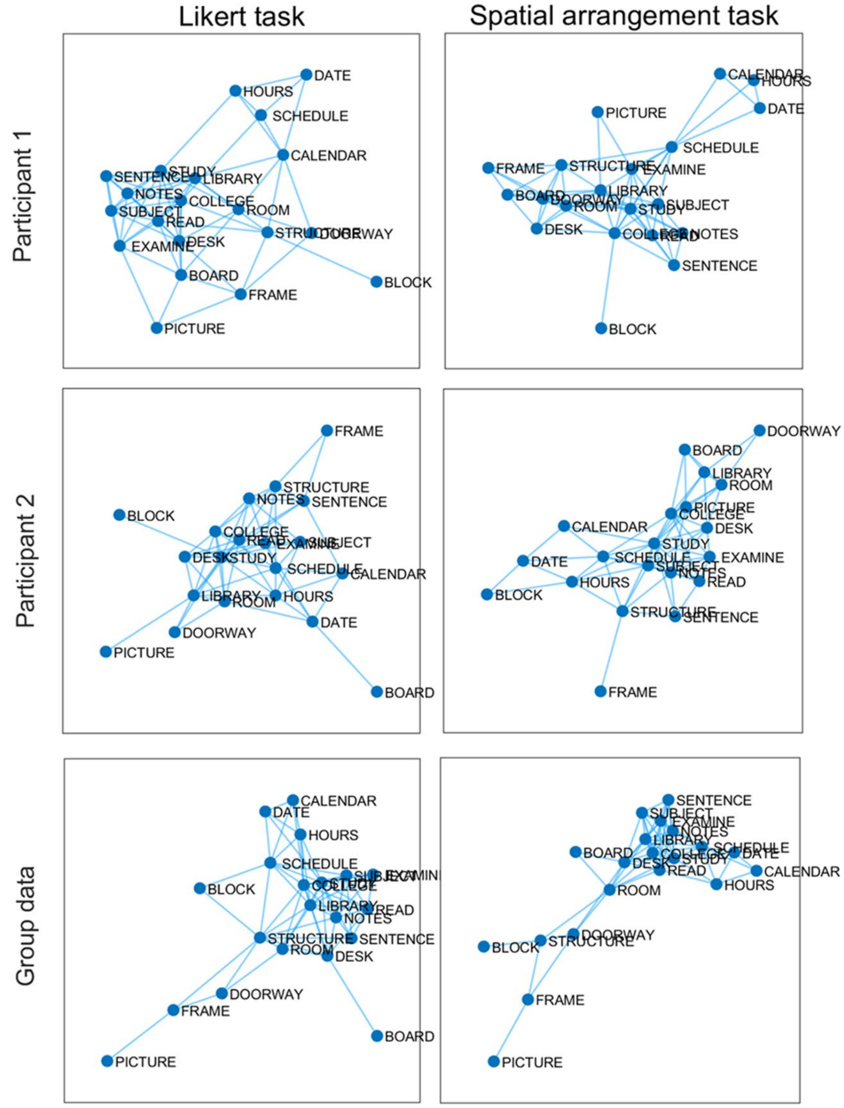


Fig. 8 Example graphs from Experiment 2 of two participants (first two upper panels) and graph constructed from average data (lower panel)

few as 3). It does not matter how many words you see; you should use all of the space available to you in the circle to arrange the words and communicate the dissimilarity between words given to you on a given trial. For instance, if you see 10 words, you should arrange words that are relatively similar to one another closer together, and words that are less similar further apart. As an example, consider the example array below. The words “Dracula,” “vampire,” and “cape” are relatively close to each other because this person considered these words to be more similar to one another than the others. On a different trial, however, the algorithm <b>zooms in</b> on the three words “Dracula,” “vampire,” and “cape.” This allows us to collect more precise measurements of your judgments of similarity. Therefore, if you see fewer (3 words), you should still use ALL of the space in the arena to precisely communicate the relative similarity of those three words. As an example, consider the example array below, which now just has the words “Dracula,” “vampire,” and “cape,” This person considered “Dracula” and “vampire” to be more similar to one another, so they are placed closer together, and are further apart from the word “cape.” Again, it is important to note that even if these words are similar to one another, all of the space in the circle is used to communicate the relative similarity between these three words. Finally, note from this example that it does not matter where on the circle you place words—that is, whether they are on the top, left, right, or bottom of the circle. What matters is the distance between each word, which is a measure of how similar you think those words are to each other. If you need to reset an arrangement, you can right-click the “START OVER” button on the bottom left-hand side of

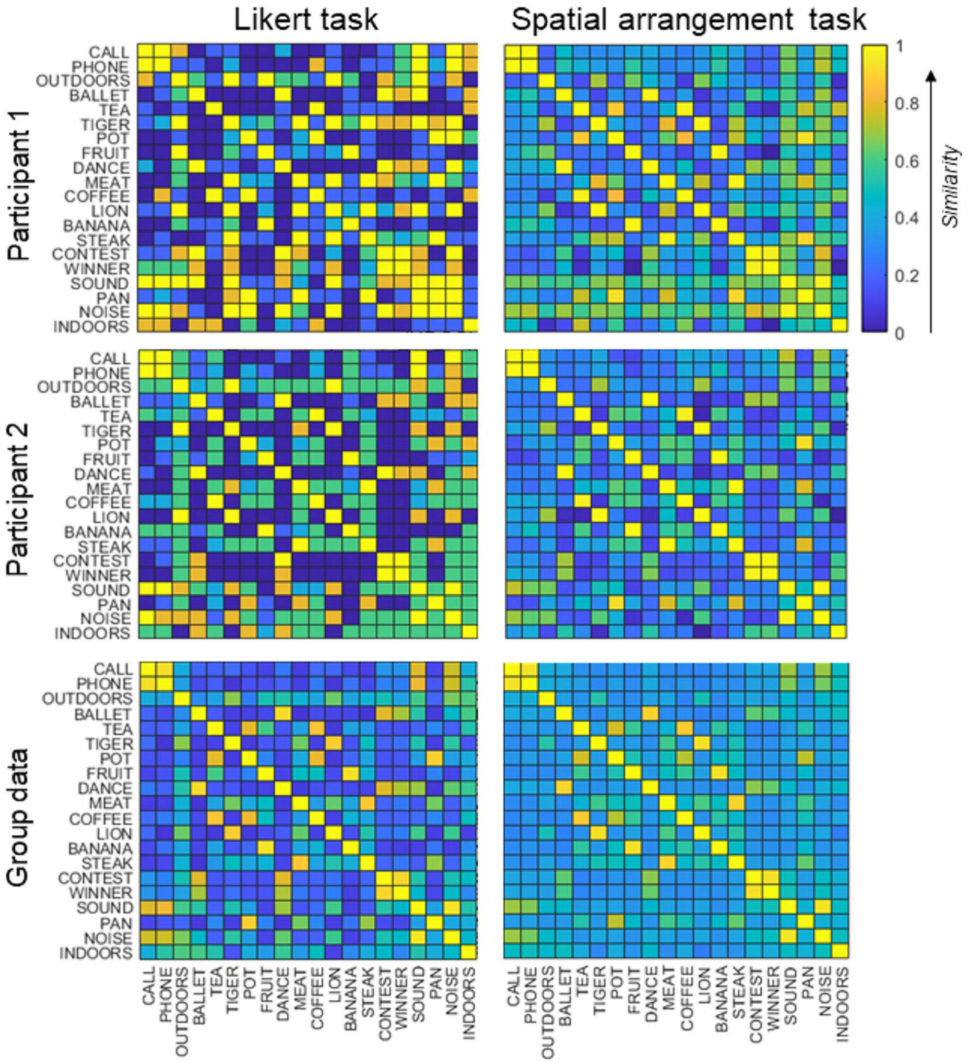


Fig. 9 Heatmaps of adjacency matrices (0=similar, 1=similar) from Experiment 1 of two participants (first two upper panels) and average data (lower panel)

the screen. Once you are ready to advance, you should click the "NEXT TRIAL" button on the bottom right hand side of the screen. The program uses an adaptive algorithm that is based on the precision and consistency with which you make your judgments. Therefore, you will see words repeatedly on different trials. The more you think about your judgments on each trial, the faster the experiment will end. If you arrange words randomly on each trial, the algorithm will not reach criterion, and it will take longer to complete the task. Therefore, you should try to make your arrangements precise and consistent, rather than speeding through. This will ensure that the algorithm reaches criterion faster, and ends this task. Keep in mind that there is no wrong way to arrange the words. This is a method for measuring your subjective judgment of similarity between each of the words, so there is no wrong answer as long as you are not doing the arrangements randomly and follow the instructions given above.

Instructions for relatedness ranking task (100-point slider scale). Note that instructions for the Likert 6-point rating task are identical with the exception that they refer to the 6-point rather than 100-point scale

You are done with the first part of the experiment and ready to start the second part of the experiment. In this part of the task you will be asked to judge similarity between two words in a different way. This part of the experiment will last 20–25 minutes. Specifically, you will be shown a pair of words at a time, and asked to rate how similar you think the two words are on a scale from 1 (maximally different) to 100 (identical). To report on your similarity

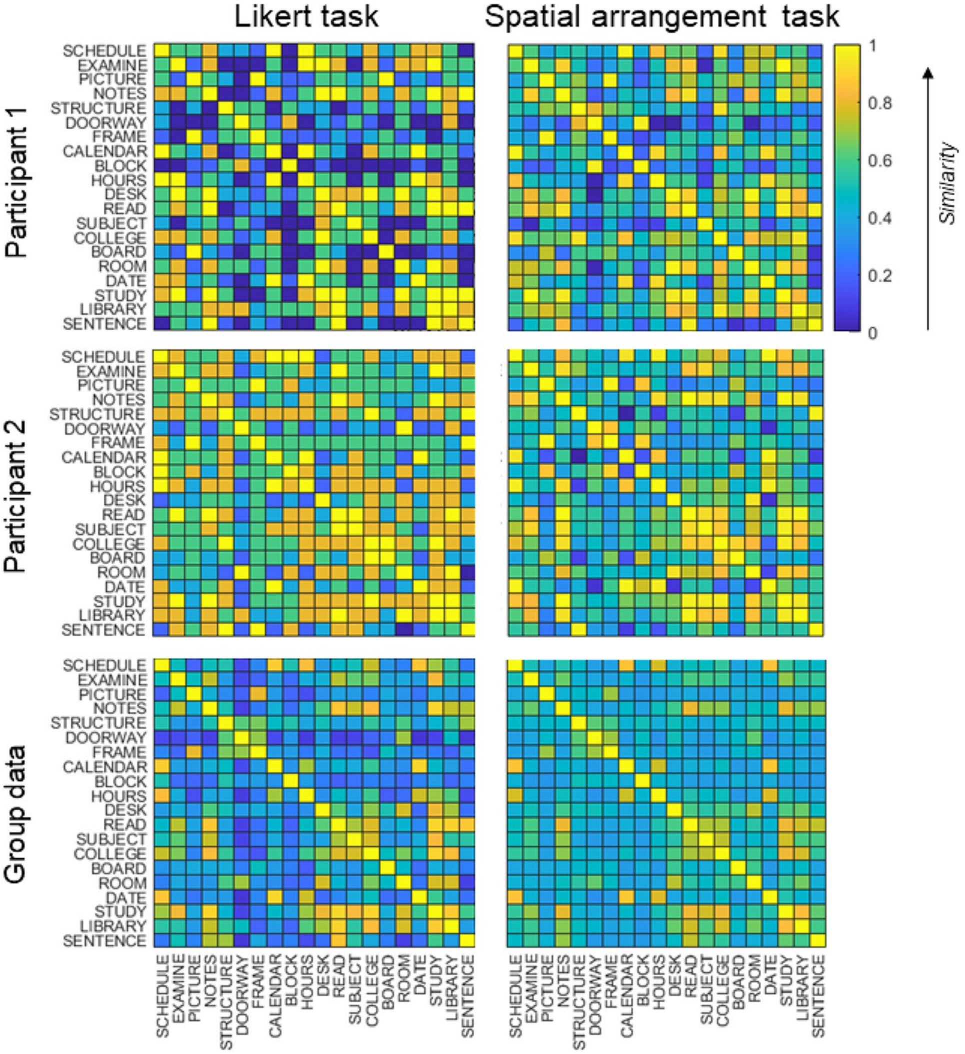


Fig. 10 Heatmaps of adjacency matrices (0=similar, 1=similar) from Experiment 2 of two participants (first two upper panels) and average data (lower panel)

judgments, you will use a sliding scale with your mouse. Please think about each rating carefully and try to make your ratings as precise as possible using values of the scale that seem to best match your judgment. For instance, I may think that the words “bread” and “baguette” are extremely similar, so I would give them a rating of 90. I may think that “bread” and “butter” are similar, but less similar than baguette, so I would give them a rating of 75. I may think that “bread” and “knife” are somewhat similar, though less similar than “bread” and “butter” so I would give them rating of 60. Similarly, I may think that “bread” and “doctor” are extremely dissimilar, so I would give them a rating of 1. Note that these are just examples to illustrate how different values of the scale relate to different similarity judgments, and you should choose values that seem best to you.

Data Availability All data and analytic code are available on the Open Science Framework repository at https://osf.io/26wku/.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Bagrow, J. P., & Bollt, E. M. (2019). An information-theoretic, all-scales approach to comparing networks. *Applied Network Science, 4*(1), 1–15.

Barabasi, A. L. (2016). *Communities* (pp. 321–377). Network Science.

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.

Benedek, M., Kenett, Y. N., Umdasch, K., Anaki, D., Faust, M., & Neubauer, A. C. (2017). How semantic memory structure and intelligence contribute to creative thought: a network science approach. *Thinking & Reasoning, 23*(2), 158–183.

Benjamin, A. S., Tullis, J. G., & Lee, J. H. (2013). Criterion noise in ratings-based recognition: evidence from the effects of response scale length on recognition accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(5), 1601.

Bieth, T., Kenett, Y., Ovando-Tellez, M., Lopez-Persem, A., Lacaux, C., Oudiette, D., & Volle, E. (2021). Dynamic changes in semantic memory structure support successful problem-solving.

Borge-Holthoefer, J., & Arenas, A. (2010). Categorizing words through semantic memory navigation. *The European Physical Journal B, 74*, 265–270.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences, 111*(40), 14565–14570.

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

Cosgrove, A. L., Beaty, R. E., Diaz, M. T., & Kenett, Y. N. (2023). Age differences in semantic network structure: Acquiring knowledge shapes semantic memory. *Psychology and aging*.

Cosgrove, A. L., Kenett, Y. N., Beaty, R. E., & Diaz, M. T. (2021). Quantifying flexibility in thought: The resiliency of semantic networks differs across the lifespan. *Cognition, 211*, 104631.

De Deyne, S., Navarro, D. J., & Storms, G. (2013). Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behavior research methods, 45*, 480–498.

DeStefano, I., Vul, E., & Brady, T. F. (2020). Influences of both prior knowledge and recent history on visual working memory. In: *Proceedings of the Annual Conference of the Cognitive Science Society*.

Domhof, J. W., Jung, K., Eickhoff, S. B., & Popovych, O. V. (2021). Parcellation-induced variation of empirical and simulated brain connectomes at group and subject levels. *Network Neuroscience, 5*(3), 798–830.

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception & psychophysics, 16*(1), 143–149.

Estes, W. K. (1956). The problem of inference from curves based on group data. *Psychological Bulletin, 53*(2), 134.

Falmagne, J. C., & Narens, L. (1983). Scales and meaningfulness of quantitative laws. *Synthese*, 287–325.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G* Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior research methods, 39*(2), 175–191.

Goldstone, R. (1994). An efficient method for obtaining similarity data. *Behavior Research Methods, Instruments, & Computers, 26*(4), 381–386.

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

He, L., Kenett, Y. N., Zhuang, K., Liu, C., Zeng, R., Yan, T., Huo, T., & Qiu, J. (2021). The relation between semantic memory structure, associative abilities, and verbal and figural creativity. *Thinking & Reasoning, 27*(2), 268–293. https://doi.org/10.1080/13546783.2020.1819415.

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior research methods, 50*, 1166–1186.

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2011). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science, 3*(1), 48–73.

Jones, M. N., Willits, J., Dennis, S., & Jones, M. (2015). Models of semantic memory. *Oxford handbook of mathematical and computational psychology*, 232–254.

Kalna, G., & Higham, D. J. (2006). Clustering coefficients for weighted networks. In *Symposium on network analysis in natural sciences and engineering* (p. 45).

Kellen, D., Davis-Stober, C. P., Dunn, J. C., & Kalish, M. L. (2021). The problem of coordination and the pursuit of structural constraints in psychology. *Perspectives on Psychological Science, 16*(4), 767–778.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687–10692.

Kenett, Y. N., & Faust, M. (2019). A semantic network cartography of the creative mind. *Trends in cognitive sciences, 23*(4), 271–274.

Kenett, Y. N., & Hills, T. T. (2022). Editors' introduction to networks of the mind: How can network science elucidate our understanding of cognition? *Topics in Cognitive Science, 14*(1), 45–53.

Kenett, Y., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8*, 407.

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

Kenett, Y. N., Ovando-Tellez, M., Benedek, M., & Volle, E. (2019). Building Individual Semantic Networks and Exploring their Relationships with Creativity. *Proc Natl Aca Sci, 41*.

Kriegeskorte, N., & Mur, M. (2012). Inverse MDS: Inferring dissimilarity structure from multiple item arrangements. *Frontiers in Psychology, 3*, 245.

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review, 28*, 40–80.

Kumar, A. A., Steyvers, M., & Balota, D. A. (2022). A critical review of network-based and distributional approaches to semantic memory structure and processes. *Topics in Cognitive Science, 14*(1), 54–77.

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.

Latora, V., Nicosia, V., & Russo, G. (2017). *Complex networks: principles, methods and applications*. Cambridge University Press.

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: What could possibly go wrong? *Journal of Experimental Social Psychology, 79*, 328–348.

Lin, E. L., & Murphy, G. L. (2001). Thematic relations in adults' concepts. *Journal of Experimental Psychology: General, 130*(1), 3.

Majewska, O., McCarthy, D., van den Bosch, J. J., Kriegeskorte, N., Vulić, I., & Korhonen, A. (2021). Semantic Data Set Construction from Human Clustering and Spatial Arrangement. *Computational Linguistics, 47*(1), 69–116.

Malmberg, K. J. (2002). On the form of ROCs constructed from confidence ratings. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(2), 380.

Marko, M., & Riečanský, I. (2021). The structure of semantic representation shapes controlled semantic retrieval. *Memory, 29*(4), 538–546.

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

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.

Ovando-Tellez, M., Benedek, M., Kenett, Y. N., Hills, T., Bouanane, S., Bernard, M., Belo, J., Bieth, T., & Volle, E. (2022). An investigation of the cognitive and neural correlates of semantic memory search related to creative ability. *Communications Biology, 5*(1), 604.

Ovando-Tellez, M., Kenett, Y. N., Benedek, M., Bernard, M., Belo, J., Beranger, B., Bieth, T., & Volle, E. (2022). Brain connectivity–based prediction of real-life creativity is mediated by semantic memory structure. *Science Advances, 8*(5), eabl4294.

Page, L., Brin, S., Motwani, R., & Winograd, T. (1997). *PageRank: Bringing order to the web* (72nd ed.). Stanford Digital Libraries Working Paper.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15*(5), 779.

Regenwetter, M., Hsu, Y. F., & Kuklinski, J. H. (2019). Towards meaningful inferences from attitudinal thermometer ratings. *Decision, 6*(4), 381.

Reilly, J., Finley, A. M., Litovsky, C. P., & Kenett, Y. N. (2023). Bigram semantic distance as an index of continuous semantic flow in natural language: Theory, tools, and applications. *Journal of Experimental Psychology: General*.

Richie, R., White, B., Bhatia, S., & Hout, M. C. (2020). The spatial arrangement method of measuring similarity can capture high-dimensional semantic structures. *Behavior Research Methods, 52*(5), 1906–1928.

Roberts, F. S. (1985). Applications of the theory of meaningfulness to psychology. *Journal of Mathematical Psychology, 29*(3), 311–332.

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

Rubinsten, O., Anaki, D., Henik, A., Drori, S., & Faran, Y. (2005). Free association norms in the Hebrew language. *Word norms in Hebrew*, 17–34.

Schvaneveldt, R. W., Durso, F. T., & Dearholt, D. W. (1989). Network structures in proximity data. In G. Bower (Ed.), *The psychology of learning and motivation: Advances in research and theory* (24th ed., pp. 249–284). Academic Press.

Schvaneveldt, R. (2023). Pathfinder Networks (https://www.mathworks.com/matlabcentral/fileexchange/59378-pathfinder-networks), MATLAB Central File Exchange. Retrieved April 18, 2023.

Siew, C. S., Wulff, D. U., Beckage, N. M., Kenett, Y. N. (2019). Cognitive network science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, *2019*.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science., 22*, 1359–1366. https://doi.org/10.1177/0956797611417632

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.

Taconnat, L., Baudouin, A., Fay, S., Raz, N., Bouazzaoui, B., El-Hage, W., ... & Ergis, A. M. (2010). Episodic memory and organizational strategy in free recall in unipolar depression: The role of cognitive support and executive functions. *Journal of Clinical and Experimental Neuropsychology*, 32(7), 719–727.

Tulving, E. (1972). Episodic and Semantic Memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381–403). Academic Press.

Van Fraassen, B. C. (2008). *The empirical stance*. Yale University Press.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440–442.

Wicherts, J. M., Veldkamp, C. L., Augusteijn, H. E., Bakker, M., Van Aert, R., & Van Assen, M. A. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology, 7*, 1832.

Wulff, D. U., De Deyne, S., Aeschbach, S., & Mata, R. (2022). Using network science to understand the aging lexicon: Linking individuals' experience, semantic networks, and cognitive performance. *Topics in Cognitive Science, 14*(1), 93–110.

Wulff, D. U., Aeschbach, S., De Deyne, S., Mata, R. (2022a). Data from the MySWOW proof-of-concept study: Linking individual semantic networks and cognitive performance. *Journal of Open Psychology Data*, *10*(1).

Zhang, B., & Horvath, S. (2005). A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, *4*(1).

Open science statement All data and analysis code are available at: https://osf.io/26wku/.

This work is not preregistered.