Cite this article: Ren D., Marusich L. R., O'Donovan J., Bakdash J. Z., Schaffer J., Cassenti D. N., Kase S. E., Roy H., Lin W. S., and Höllerer T. (2019). Understanding node-link and matrix visualizations of networks: A large-scale online experiment. *Network Science* 7, 242–264. https://doi.org/10.1017/nws.2019.6

Understanding Node-Link and Matrix Visualizations of Networks: A Large-Scale Online Experiment

Donghao Ren[1], Laura R. Marusich[2], John O'Donovan[1], Jonathan Z. Bakdash[3],[4], James Schaffer[1], [5], Daniel N. Cassenti[6], Sue E. Kase[6], Heather Roy[6], Wanyi (Sabrina) Lin[7], and Tobias Höllerer[1]

[1]University of California, Santa Barbara
[2]U.S. Army Research Laboratory South at the University of Texas at Arlington
[3]U.S. Army Research Laboratory South at the University of Texas at Dallas
[4]Texas A&M–Commerce
[5]U.S. Army Research Laboratory West
[6]U.S. Army Research Laboratory
[7]IBM Thomas J. Watson Research Center. The author is currently affiliated with BOSCH Center for Artificial Intelligence.

Abstract

We investigated human understanding of different network visualizations in a large-scale online experiment. Three types of network visualizations were examined: node-link and two different sorting variants of matrix representations on a representative social network of either 20 or 50 nodes. Understanding of the network was quantified using task time and accuracy metrics on questions that were derived from an established task taxonomy. The sample size in our experiment was more than an order of magnitude larger ($N = 600$) than in previous research, leading to high statistical power and thus more precise estimation of detailed effects. Specifically, high statistical power allowed us to consider modern interaction capabilities (e.g., mouse-over highlighting and pen-stroke annotations) as part of the evaluated visualizations, and to evaluate overall learning rates as well as ambient (implicit) learning. Findings indicate that participant understanding was best for the node-link visualization, with higher accuracy and faster task times than the two matrix visualizations. However, this performance advantage for node-link over matrix was attenuated in the larger (50-node) network. Analysis of participant learning indicated a large initial difference in task time between the node-link and matrix visualizations, with matrix performance steadily approaching that of the node-link visualization over the course of the experiment. We also report an exploratory analysis of participants' mouse movements and annotations. This research is fully reproducible as the web-based module and results have been made publicly available at: `https://osf.io/qct84/`.

Introduction

Understanding graph data is an important task in many domains, including law enforcement (Baccara & Bar-Isaac, 2008; McIllwain, 1999; Sparrow, 1991) (e.g., identifying key players in organized crime networks), public health (Blanchet & James, 2011; Newman, 2002) (e.g., detecting and mitigating outbreaks of disease), and military intelligence analysis (Bohannon, 2009; Krebs, 2002; MacCalman *et al.*, 2013) (e.g., uncovering the structure of terrorist organizations or identifying key leaders to engage within host nations). The widespread use of social networks, and the use of network visualizations in mass media (Sullivan, 1987; Lankow *et al.*, 2012), has caused the public to become familiar with general network concepts, yet interactive network visualization has not yet found its way into the most prevalent web search and social network sites. Effective visualization is helpful for understanding and interpreting most types of data, but for graph data it is critical. For example, a variety of tools exist within the field of intelligence analysis to help synthesize information from disparate intelligence sources, identify entities and the various relationships between them, and visualize this network to allow the analyst to identify patterns and draw inferences (Hall *et al.*, 2015). These tools typically rely on one of two types of graph visualization: 1) node-link representations, in which each node represents an entity and connecting lines represent the relationships between them, and 2) matrix-based representations, where each row/column corresponds to an entity and filled squares represent relationships between entities (see Figure 1 for an example of each visualization). In this work we investigate how these two visualization types affect human comprehension of graph data, and specifically networks relevant to the field of intelligence analysis. We focus here on the low-level graph comprehension tasks that are fundamental to intelligence analysis, but that do not require the domain expertise of the intelligence community. Because these tasks can be performed by the general population with only a small amount of training, we were able to use Amazon Mechanical Turk (Mturk) to recruit a large number ($N = 600$) of participants.

Knowledge about human performance on low-level graph analysis tasks, given one of these representations, is important for designing the next generation of visualization tools, which could potentially be embedded in social and information computing platforms (Bostandjiev *et al.*, 2011). While this research topic has received considerable attention over a decade ago (Ghoniem *et al.*, 2004, 2005), modern interactive graph visualization capabilities provide new opportunities for evaluation. Specifically, we partially replicate and build upon behavioral experiments by Ghoniem *et al.* (2004, 2005) using a much larger sample size. While previous work has compared both node-link and matrix visualizations for speed and accuracy, these studies have relatively small numbers of participants. The $N = 600$ study presented here brings improved statistical power that may better highlight performance differences between node-link and matrix visualizations for a variety of different analytical tasks. Last, our work emphasizes the importance and value of reproducible research (Peng, 2011). We share the code for the web-based study module used to run the experiment, the participant data, and the code for statistical analysis: `https://osf.io/qct84/`.

1 Related Work

Many studies that compared visualization types for graph data have typically focused on the efficiency of layout algorithms or aesthetic differences. Battista *et al.* (1998) provided a review of traditional approaches, and a more recent review is given by Von Landesberger *et al.* (2011). Less common are experiments that assess the impact of visualization type on human performance. The majority of earlier research used a limited subset of graph understanding tasks (e.g., path finding, importance of nodes, and number of nodes) with

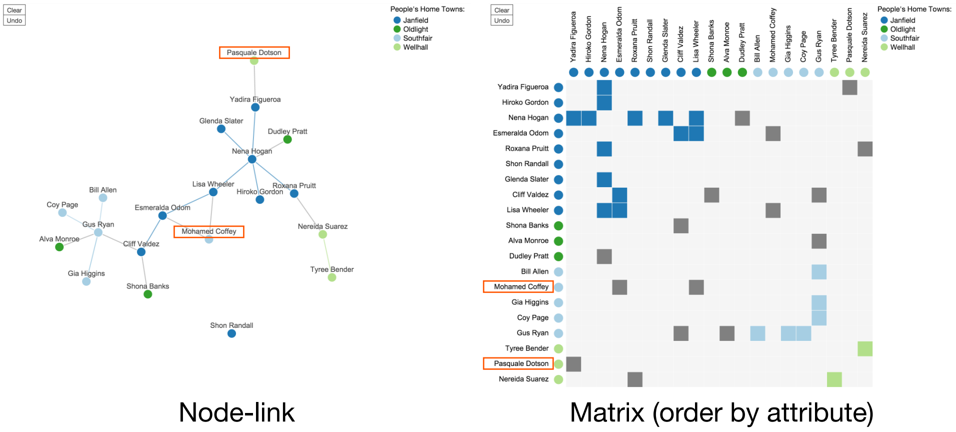


Fig. 1. Example tasks in our study: *"Which people lie along the shortest path between Pasquale Dotson and Mohamed Coffey?" "How many direct connections are there between people in Janfield and people in Oldlight?"* Participants were asked to use one of the visualizations to complete the tasks.

randomly generated networks. A description of previous behavioral research in graph visualization is provided below.

In the context of path-finding tasks, Purchase (1998) has examined human performance on node-link visualizations using eight different layout algorithms and a single graph of 17 nodes and 29 edges. Results indicated very similar performances across the eight layout algorithms, with somewhat lower accuracy on only one of the layouts. Purchase concluded that many of the efficiency and aesthetic considerations that go into the creation of layout algorithms do not necessarily have a corresponding impact on human readability, at least on small graphs similar to the one used in their experiment.

Similarly, McGrath *et al.* (1997) assessed node importance for connectivity with five different spatial arrangements of the node-link representation, using a single graph of 12 nodes. The graph was presented as a network of communications among people, and participants were asked to rate a subset of the nodes on prominence and importance as a bridge within the network. For this relatively higher-level interpretation task, results indicated that spatial arrangement does affect participants' ratings, although the researchers acknowledge that the best spatial arrangement is likely task dependent.

Ghoniem *et al.* (2004, 2005) performed studies of graph readability that compared node-link and matrix representations with random graphs. In their work, graphs varied in both size (20, 50, or 100 nodes) and edge density (0.2, 0.4, or 0.6). Readability was assessed using a variety of relatively simple tasks (e.g., estimating the total number of nodes and edges, finding a specific node, finding a specific edge connecting two nodes, identifying the node with the most edges, or identifying a path between two nodes). At the smallest graph sizes, performance was best when using the node-link visualization. However, as size and density increased, the authors found indications of improved performance when using matrix visualization over node-link on tasks other than path-finding. Okoe & Jianu (2015) replicated this experiment with 112 participants from MTurk. Our work considers a wider variety of network-relevant search and identification tasks.

While node-link and matrix graph representations have been combined in visual analytics tasks (Wong *et al.*, 2006) and compared for tasks involving weighted graphs (Alper *et al.*, 2013; Chang *et al.*, 2017), and novel compound graph visualizations have been derived and evaluated (Henry *et al.*, 2007; Henry & Fekete, 2007), to our knowledge, only one study compares human performance differences between node-link and matrix visualizations specifically in the context of intelligence analysis (Berardi *et al.*, 2013). In this experiment, participants (all U.S. Air Force intelligence analysts) were provided either a node-link or matrix visualization of a social network of 34 actors along with betweenness and closeness centrality measures for each actor. Their task was to identify leaders and clusters within the network. Both response time and the accuracy of responses indicated better performance in the node-link visualization.

In summary, previous work comparing human performance in matrix and node-link representations suggests that node-link may be more advantageous overall when networks are relatively small ($\sim$ 30 nodes or fewer) and on tasks that require tracing a path between two entities that are not directly connected. These studies indicated decreased human performance for node-link visualizations when networks were large and densely connected due to overlapping edges, which can make interpretation more difficult.

Prior studies have typically been conducted in small-scale laboratory environments that use either a single graph or a set of randomly generated graphs. One substantial limitation to these studies is the low statistical power associated with small sample sizes, which leads to potentially spurious and/or inflated statistical significance and effect sizes (Open Science Collaboration, 2015; Bakker *et al.*, 2012). Second, the "representativeness" of the networks used in some prior research is another possible limitation. Depending on the generating process, random networks often do not display characteristics common to real-world networks of people, including long-tailed degree distributions, triadic closure, and homophily, for example. While other studies, such as TopicNets (Gretarsson *et al.*, 2012) have focused on human understanding of relatively large social graphs through interactive tools, we focus our comparative study at a smaller scale (20 to 50 node range), following Henry and Fekete's hypothesis that larger graph sizes would introduce additional navigation and scrolling issues (Henry & Fekete, 2007), and Ghoniem *et al.*'s hypothesis that (static) node-link representations do not perform as well on large graphs with high density clusters, typical of social networks.

One additional factor that might affect the decision to utilize node-link or matrix visualizations is the idea of visualization insight (Yi *et al.*, 2008; Chang *et al.*, 2009). While certain visualizations might be more or less useful for answering particular types of targeted questions, there may be a less tangible "understanding" of underlying data relationships and distributions that arises from the experience of interacting with a visualization, and different visualizations can produce very different interaction experiences. For instance, a recent study of Twitter sense-making demonstrated that interface design could significantly affect a user's performance on a post-study estimation questionnaire (Schaffer *et al.*, 2015). Understanding visualization insight or bias is important because long-term memory is critical in the pattern recognition process (Eysenck & Keane, 2013), and insights might come at any time, not just during an analysis session (North, 2006). One might expect that node-link visualizations lead to more insight, due to their more intuitive presentation.

In the work presented here, our goal was to conduct a large-scale experiment with sufficient power to tease apart effects of visualization, network size, and task type on human performance. We evaluated participants on a wide range of questions that included lower-level readability tasks as well as higher-level understanding and interpretation tasks. In addition, we focused on realistic social networks that would be relevant to the domain of intelligence analysis. Finally, we collected not only task performance data but also data on interaction with the interface (e.g., mouse movement, annotation) in order to more fully understand how people complete graph-understanding tasks.

2 Research Question

The overarching research question is how human understanding differs when using the node-link and matrix visualizations to perform basic analytical tasks. Specifically, we evaluate and compare understanding in terms of node-, connectivity-, and attribute-level tasks. Human understanding, or readability, of graph representations for task completion was quantified using two objective measures of performance: response time and accuracy.

3 Study Design

We employed a mixed design with two between-participant factors (visualization type and graph size) and one within-participant factor (task type). Below, we describe the task types completed by each participant, the graph data used in the study, and the layout of each visualization type.

3.1 Experiment Task

This experiment was designed to assess graph understanding through a range of tasks that would be relevant to intelligence analysis without requiring domain expertise. Following the taxonomy of tasks common in graph analysis devised by Lee *et al.* (2006), we developed a list of questions that fell into one of five task categories: Adjacency, Accessibility, Common Connection, Connectivity (shortest path), and Attribute. These questions were presented in a random order to each participant and are listed below.

Adjacency (AD)

- (AD1) Name all of the people *Person AD1-A*[1] is directly connected to.
- (AD2) Do *Person AD2-A* and *Person AD2-B* have a direct (one-hop) connection?
- (AD3) Which person has the most direct connections?
- (Discarded[2]) How many people have no connection?

[1] AD1-A, AD2-A, etc., are placeholders for person names, which are randomly generated in "FirstName LastName" format.

[2] This question was phrased differently in different graph sizes, thus we discarded it from analysis.

Accessibility (AC)

- (AC1) Is *Person AC1-A* connected to *Person AC1-B* in two hops or less?
- (AC2) Is *Person AC2-A* connected to *Person AC2-B* in three hops or less?
- (AC3) How many people are exactly two hops away from *Person AC3*?
- (AC4) Name all the people who are exactly two hops away from *Person AC4*.

Common Connection (CC)

- (CC1) How many people are directly connected to both *Person CC1-A* and *Person CC1-B*?
- (CC2) Name all the people directly connected to both *Person CC2-A* and *Person CC2-B*.

Connectivity (CO)

- (CO1) What is the minimum number of hops between *Person CO1-A* and *Person CO1-B*?
- (CO2) Which people lie along the shortest path between *Person CO2-A* and *Person CO2-B*?
- (CO3) If *Person CO3-A* was removed from the network, would *Person CO3-B* and *Person CO3-C* still have a connection between them (either directly or through other people)?

Attribute (AT)

- (AT1) How many people are in *Town AT1*?
- (AT2) Name all the people in *Town AT2*.
- (AT3) How many direct connections are there between two people that are both in *Town AT3*?
- (AT4) How many direct connections are there between people in *Town AT4-A* and people in *Town AT4-B*?

In addition, we assessed each participant's implicit learning about the graph they encountered with several overview questions presented in a post-study questionnaire.

3.2 Graphs

Compared with other domains, intelligence analysts typically work with fairly small networks. When they encounter data sets with large numbers of entities, they will prune the network to a manageable number of key actors (see examples in Army (2006)). Consequently, we limited our assessment to a graph of 20 nodes (a typical size for intelligence analysts) and a graph of 50 nodes (approaching the upper limit of what intelligence analysts would generally work with). Each participant saw only one of these two graphs, described in detail below.

The 20-node network of people was adapted from the Kandahar Human Terrain Dataset and Scenario originally developed for the Army field exercise Command, Control, Communications, Computers, Intelligence, Surveillance and Reconnaissance (C4ISR) and Network Modernization Event 13 (E13). C4ISR E13 is an annual exercise offering researchers a military-relevant venue to assess, evaluate, and validate emerging technologies. The Kandahar data set consists of text data simulating blogs, news story feeds, tweets, and intelligence reports. These data sources contain an embedded network of 20 individuals of different tribal membership representing friendly, insurgent, and criminal elements operating near the Kandahar region of Afghanistan. The Kandahar data set was initially used in an experiment focused on visualization approaches for displaying measures of sentiment as well as an information-based problem-solving game currently under development (Kase *et al.*, 2015). For this experiment, the Kandahar 20-node network of individuals was extracted, with the four tribal affiliations used as a node attribute. The extracted graph has 20 vertices and 19 edges, with density 0.10.

The 50-node network of people was adapted from the synthetic Ali Baba data set originally developed at the National Security Agency (NSA) Jaworowski & Pavlak (2003). The Ali Baba data set consists of 752 simulated human intelligence (HUMINT) reports tracking the activities of a suspected terrorist network operating in England. The original version of the data set was cleaned and modified, and ground truth was determined, enabling its use for testing text extraction and social network analysis technologies, Mittrick *et al.* (2012). The Ali Baba data set provides rich social relational data contained within the text messages, making it an ideal data set for testing social network and information extraction techniques. The primary plot of the Ali Baba data set focuses around a fictitious terrorist group, the Ali Baba Group. The primary plot is woven within several distractor plots. The Ali Baba Group consists of approximately nine core members operating in England; however, the group members do not all interact directly — some interact with the many periphery characters. For this experiment, a 50-node network of individuals was extracted from the Ali Baba data set. The first-mentioned location (seven in total) of each individual was used as a node attribute. The extracted graph has 50 vertices and 71 edges, with density 0.06.

In both graphs, we randomly assigned a new name to each node so that the first three letters of first names and last names are unique. Similarly, the attribute category on each graph was relabeled as "people's home towns," and we generated random town names to avoid any familiarity of existing towns.

3.3 Visualizations

The layout of the node-link visualization was generated by the force-directed algorithm in D3. While there is no inherent ordering to nodes in the node-link visualization, node order is a key aspect in the matrix visualization and may affect performance on various tasks. Consequently, we assessed two basic layouts for the matrix representation in our online experiment – one ordered by attribute, and one ordered by degree. Degree ordering is considered as a baseline ordering. Despite the obvious benefit to the degree counting task (AD3), it falls shorts on revealing the community structure. Attribute ordering is chosen because of its direct connection to the graph attribute. In the graph we use, the attribute

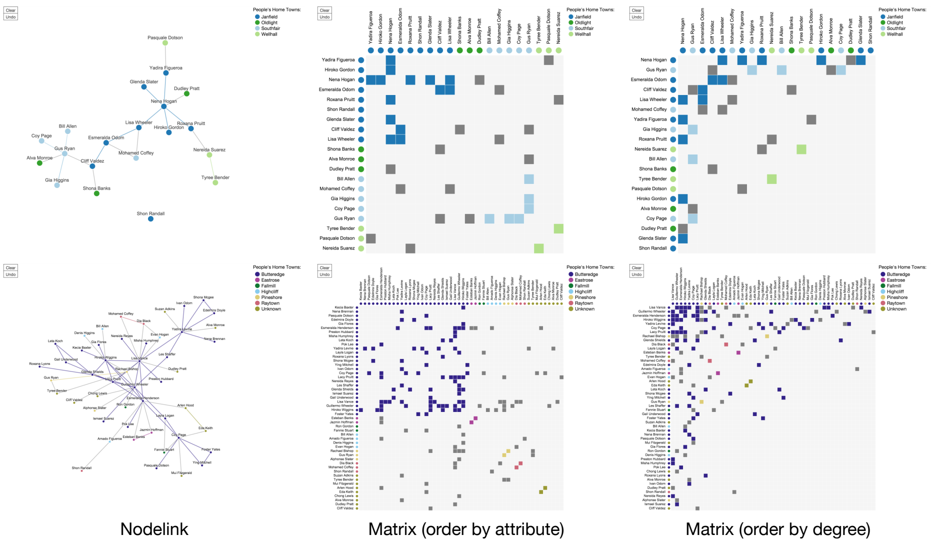


Fig. 2. Data sets and visualizations for the study. Two representational graphs from analytical tasks (small and large) were used to evaluate three visualizations (node-link, matrix order by attribute, matrix order by degree).

is related to the community structure, and it is hypothesized that ordering the matrix by attribute would be beneficial to the attribute-related tasks. As a result, each participant encountered one of three visualization types in this study: node-link, matrix-attribute, and matrix-degree (see Figure 2).

Across all visualization types, the color palette, node size, and font size for the same graph were identical. Within the same visualization type, the layout was the same for all participants. We enforced a minimal window size (1150 × 600 pixels) for the study. For users with higher resolution displays, the graph automatically scaled to accommodate the available screen space. Note we used a color-blind safe palette and we did not assess color-blindness.

In order to allow straightforward comparison between the visualization types, we did not allow participants to customize (e.g., repositioning nodes in node-link, reordering nodes in matrix, etc.) the visualization. However, we did allow a limited number of interactions that could be consistently applied across all conditions. These were designed to ease some of the practical challenges associated with the task, such as locating a specific node on the display, and to keep overall participation time within a reasonable range. The interactions are described below and can be viewed at `https://osf.io/qct84/`.

Mouse-over Interactions *Mouse-over node:* When the user mouses over a name or node in the node-link visualization, that node and its label are highlighted in red, the line weights of its edges are increased, and the labels and nodes of its neighbors are highlighted in orange. In the matrix visualizations, the node and its label are highlighted in red, the row and column corresponding to the node are shaded, and the labels and nodes of its neighbors are highlighted in orange. *Mouse-over edge:* When the user mouses over an edge in the node-link visualization, the line weight of that edge increases, the two connected nodes and their labels are highlighted in red, and all neighbors of those two nodes are highlighted in orange. In the matrix visualizations, the rows and columns of the two connected nodes are shaded, the two connected nodes and their labels are highlighted in red, and the nodes and labels of all neighbors of those two nodes are highlighted in orange. *Mouse-over legend:* When the user mouses over an attribute category in the legend, all nodes that belong to that category (i.e., all people in the specified town) are highlighted in black, and the corresponding node labels appear in bold weight.

Annotation In addition to mouse-over interactions, we supported annotations. Henry & Fekete (2007) argue that such interactive highlighting is useful, and also applied it in their comparative study. Dragging the mouse over the visualization results in a semi-transparent stroke. Two buttons at the top-left corners of the visualizations allow the user to clear all annotations or undo the last stroke.

Highlight Node from Question In questions that refer to specific nodes (e.g., "Name all of the people **Person A** is directly connected to"), the name is made bold within the question text. When participants hover the mouse over the name, the corresponding node is highlighted in the visualization (node in node-link, and row/column in the matrix, see Figure 1).

4 Methodology

This experiment was approved by the University of California at Santa Barbara's Institutional Review Board (IRB) and the U.S. Army Research Laboratory's IRB.

Participants We predetermined the number of participants to be 600 before running the study to attain greater than 80% statistical power to detect medium effect sizes for ANOVA. This included each main effect for the two between factors (visualization types and graph sizes), the main effects for the within factors (17 task type questions and each of the five categories of questions), and the two within-between interactions (visualization type x task type and graph size x task type). There were 100 participants for each of the six between-subject factor combinations (3 visualization types × 2 graph sizes). We randomized the order of the 17 task type questions to minimize the possible impact of task order.

The study system was run on MTurk with multiple batches until we obtained 600 successful completions. Use of MTurk enabled us to collect a much larger sample of participants than would be feasible in most laboratory studies. In addition, MTurk enables researchers to collect data from a diverse international population of English speakers, instead of a more uniform population of Western undergraduate students (Buhrmester *et al.*, 2011). While we lose the benefits of controlled environments that are possible in supervised studies, we benefited from increased statistical power and a more diverse and thus generalizable sample. There is also some evidence. There is also some evidence to suggest that MTurk participants may pay more attention to instructions than undergraduate students participating in typical lab studies (Hauser & Schwarz, 2015). In this study, we required that workers had a history of at least 50 approved Human Intelligence Tasks

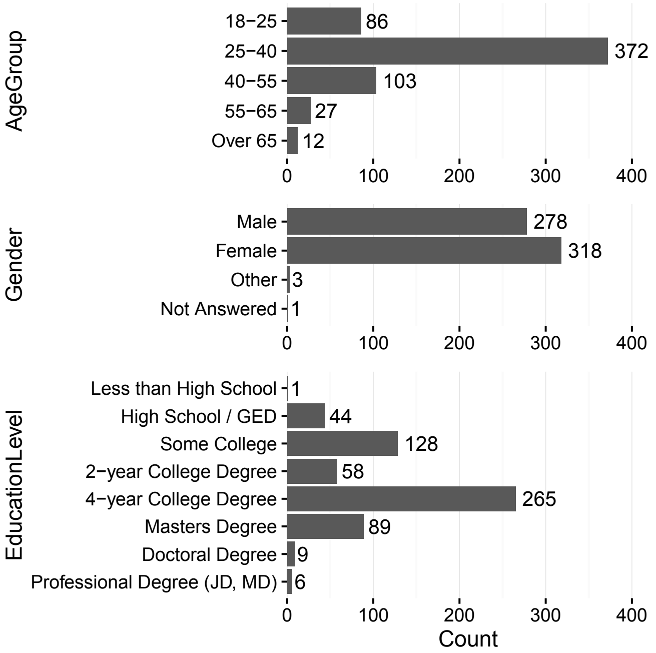


Fig. 3. Participant demographics.

(HITs). We did not use location-based filtering, but did ask about self-reported English language proficiency in the pre-study questionnaire.

We imposed several other prerequisites on eligibility for the study: 1) participants were required to use the Chrome browser, with a minimum window size of 1150×600; 2) they were also required to use a keyboard and mouse for interaction input (no mobile devices or touch-screen interaction); and 3) zoom level of the browser was set to zero.

On MTurk we advertised that participants would receive a one dollar base payment for completing the study with an additional bonus amount up to one dollar based on their performance. During the course of the study the current bonus payment earned was displayed on the screen, with increments added with every correct answer. After the study, all participants were ultimately paid the full amount of two dollars regardless of their actual performance.

We ran the study on MTurk in multiple batches until we had collected 100 valid sessions in each experimental condition (600 in total). A valid session was defined as follows: 1) The participant had not previously attempted the study, 2) the participant completed the entire study and the results were successfully submitted to our server, and 3) the participant correctly answered the pre-study attention check questions (described below). Participants were primarily between the ages of 25 and 40, with 281 reported males and 316 reported females (see Figure 3).

Procedure Upon accepting the task on MTurk, participants were redirected to our study. Participants read and agreed to a consent form and were checked that they met the study prerequisites. At this point, each participant was randomly assigned to a visualization type and graph size condition and asked to complete the following four main phases of the study: pre-study questionnaire, training, tasks, and post-study questionnaire. Here we describe these four phases.

Pre-study Questionnaire All participants completed an initial questionnaire that collected basic demographic information, as well as information about general familiarity with graph data, data visualization, and social networking. This questionnaire also included two attention check items to filter out participants who responded randomly instead of engaging with the task.

Training Each participant completed a self-paced training tutorial that explained the assigned visualization type using a very small example network. As part of the tutorial, participants were required to complete at least one instance of each of the interaction capabilities available (i.e., mouse-over, annotation, highlighting node from question). The training also included a small set of questions about the example data set, which participants were required to answer correctly to proceed to the next phase of the study. Participants that did not complete the training tutorial were replaced (see Table 5 for a summary of training completion rates).

Task Questions Immediately following the training, the participant began the experimental task questions. The presentation order of each of the 17 questions was randomized for each participant. Only the question text was initially visible on the display at the start of each question presentation. When participants pressed a "Start" button under the question task, the network visualization and response fields appeared, and the internal timer began for that question. We supported auto completion for questions that require entering the name(s) of person(s). When participants entered and confirmed their response, the screen cleared and then displayed the correctness of the previously answered question and the current bonus payment, as well as the text of the next question. We instructed participants to "go as fast as you can without making mistakes." We did not impose time limits, but we did allow participants to skip the current question if they could not determine the answer by clicking a "Cannot Answer" button. We added this option to minimize participant frustration and drop-outs.

Post-study Insight Questionnaire After completing the task questions, participants were redirected to a post-study questionnaire, where they self-rated their performance and the difficulty of the previous task. We also asked several implicit learning questions, listed below, designed to assess participants' learning and retention of general aspects of the structure of the graph they had viewed. Participants had no foreknowledge of these questions during the previous set of tasks.

- (I1): How would you describe the distribution of connections in the network? (Choose one from four descriptions)
- (I2): How would you describe the connections between and within towns? (Choose one from three descriptions)
- (I3): Please rank these (four) people from the network you just saw in order of how many direct connections they had. (Drag to rank)

- (I4): Please rank which (of the four) locations had the most people. (Drag to rank)
- (I5): Please estimate how many people were in the network you just saw. (Enter a number)

After completing this final questionnaire, participants received a completion code to be entered on MTurk for payment.

5 Analysis of Results

Here we describe our procedures for analyzing the data collected in our study. We first assess performance on the primary task using completion time and accuracy. We also examine accuracy on the post-study insight questions, summarize participants' self-ratings of task difficulty as well as their performance, report overall completion rates across the visualization conditions, and briefly explore differences in interface interaction across conditions.

Participants' task times were strongly positively skewed, which is typically the case with response time data. To meet assumptions of normality, the log transform of task time (Cohen *et al.*, 2003) was used in all analyses. One participant had an accuracy of 0 and mean response time under 3 seconds, indicating that the task was not genuinely attempted; we excluded this participant.

5.1 Primary Task Performance: Task Time and Accuracy

We first assessed whether participants' task time and accuracy were related measures of performance on the primary task — namely, whether a speed/accuracy tradeoff was present. We averaged task time and accuracy for each participant and conducted an ordinary least squares regression on this data, ignoring condition. The association between task time and accuracy approached zero, t = -1.748, $p = 0.081$, $R^2 = 0.005$, see Figure 4. Because of the high-powered design with a large sample size, the regression had a marginally significant p-value. However, the minuscule effect size indicates that task time and accuracy were close to linearly independence. Thus, they represent two distinct measures of performance for network understanding, and this justifies separate analysis of each measure.

Task Time

We conducted the analysis of task time with a Linear Mixed Effects (LME) Model in R using the lme4 (Bates *et al.*, 2015) package. An LME model was used instead of ANOVA because task types were unbalanced; the number of questions in each category varied. In addition, when possible, modeling the "stimuli" (task type) as a random effect permits stronger inferences about the generalizability of the results (Baayen *et al.*, 2008).

The formula of the model was the following:

$$\log(\texttt{TaskTime}) \sim \texttt{Visualization} \times \texttt{Size} \times \texttt{TaskType} + (\texttt{TaskType}|\texttt{ParticipantID})$$

Visualization, size, and task type were specified as fixed effects, and task type and participant as random effects. The described model clearly had a better relative fit than

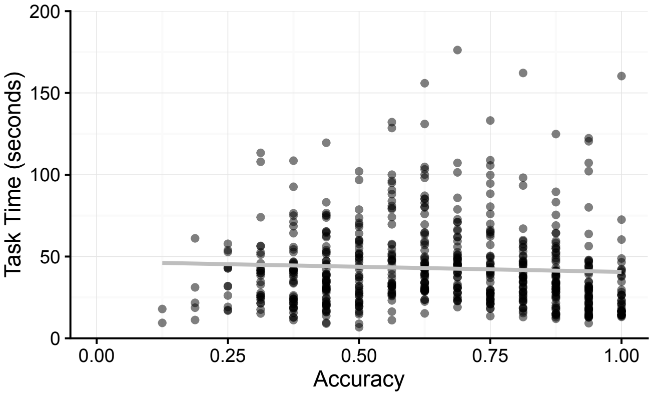


Fig. 4. Ordinary least squares regression: log(`TaskTime`) $\sim$ `Accuracy`. Each dot represents the average accuracy and (raw) task time for a participant. Three points above 200 seconds were truncated.

models with fewer parameters based on multiple fit indices, which penalize for the number of parameters, see Burnham & Anderson (2003).

The absolute fit of the model was assessed using two pseudo-R^2 values (Nakagawa & Schielzeth, 2013). For only fixed effects, the marginal pseudo-R^2 was 0.128. For fixed and random effects, the conditional pseudo-R^2 was 0.394. ANOVA (with the LME model) results for the fixed effects are shown in Table 1.

We find a significant main effect of Visualization (Nodelink vs. MatrixGroup vs. MatrixDegree) on task time (see Figure 5). A Tukey post-hoc test indicated that participants responded more quickly in the Nodelink condition than in the MatrixGroup and MatrixDegree conditions (taking 23% and 19% less time, respectively). We did not find any significant difference between the two matrix conditions.

We also found a significant main effect of graph size on task time, with participants generating longer task times in the large graph condition than in the small condition. Interestingly, we also find a significant interaction between visualization type and graph size upon task time. The advantage of node-link over the two matrix visualizations is more pronounced when graph size is small than when it is large. It seems that the increased difficulty associated with the large graph size attenuates the speed advantage of the node-link visualization.

The within-subjects TaskType variable also had a significant effect on task time (see bottom panel of Figure 5). Participants were slower to answer questions from the Common Connection and Connectivity categories than those from the other three categories. We did find a significant interaction between Visualization and TaskType, indicating that the relative advantage of node-link depended on the type of question participants encountered. In the two slowest task types, Common Connection and Connectivity, node-link greatly outperformed the two matrix visualizations. This suggests that the matrix representation is particularly difficult when used to perform these types of tasks.

In the other three task categories, the node-link advantage was smaller or even reversed. In the Attribute category of questions, participants in the MatrixGroup condition (where nodes were sorted by attribute) actually responded slightly more quickly than those in the Nodelink condition, but the effect was not significant ($p = 0.358$).

	F	**p-value**	
Visualization	15.579	< 0.001	***
Size	85.139	< 0.001	***
TaskType	181.372	< 0.001	***
Vis.:Size	6.274	0.002	**
Vis.:TaskType	13.694	< 0.001	***
Size:TaskType	41.927	< 0.001	***
Vis.:Size:TaskType	1.440	0.175	

Table 1. Task Time: Type III Satterthwaite tests. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

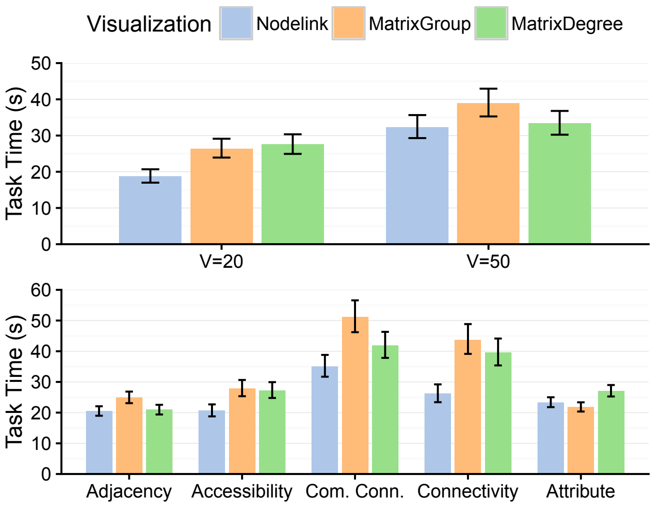


Fig. 5. Least-square means and 95% confidence intervals of $\log(\texttt{TaskTime})$ (plotted with a linear scale).

For most task types, there was no evidence of a meaningful difference in task time between the two matrix visualizations. However, participants in the MatrixDegree condition do exhibit faster task times on Adjacency questions than those in the MatrixGroup condition ($p = 0.004$, MatrixGroup took 19% more time on average). MatrixGroup, however, outperformed MatrixDegree within Attribute Questions (took 19% less time, $p < 0.001$).

Consequently, we examined task time from another perspective using an exploratory analysis of a new variable: path search radius. We define path search radius as the number of hops the participant has to search to complete the task. For attribute tasks, the radius is 0 for single nodes and 1 for connections between towns; for adjacency/common connection tasks, the radius is 1; for accessibility tasks, the radius is 1, 2, or 3, depending on the number of hops the task asks; and for connectivity tasks, since the number of hops is not specified, we consider the radius as "multiple".

We conducted the analysis with a similiar linear mixed model: $\log(\texttt{TaskTime}) \sim \texttt{Visualization} \times \texttt{Size} \times \texttt{PathSearchRadius} + (\texttt{PathSearchRadius}|\texttt{ParticipantID})$. For only fixed effects, the marginal pseudo-R^2 was 0.108. For fixed and random effects, the conditional pseudo-R^2 was 0.365.

We found a main effect of PathSearchRadius ($F = 144.289, p < 0.001$), an interaction effect between PathSearchRadius and Visualization ($F = 8.093, p < 0.001$) as well as an interaction between PathSearchRadius and Size ($F = 21.857, p < 0.001$). We found that the node-link visualization outperformed the matrix visualizations more in tasks that involve larger path search radii (see Figure 6 top).

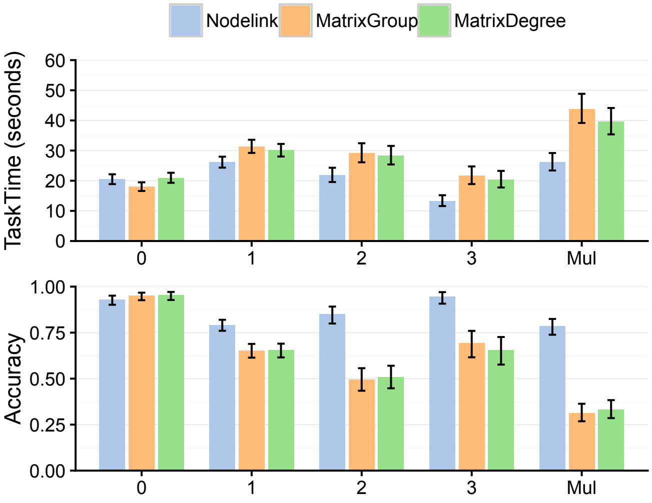


Fig. 6. PathSearchRadius: Least-square means and 95% confidence intervals of $\texttt{Accuracy}$. Analyzed with an LME model: $\log(\texttt{TaskTime}) \sim \texttt{Visualization} \times \texttt{Size} \times \texttt{PathSearchRadius} + (1|\texttt{ParticipantID})$ and a similar GLMM model.

Accuracy

We employed a generalized linear mixed model (GLMM) with the logistic function to evaluate accuracy. A GLMM is conceptually similar to an LME model, but is better suited for non-normal dependent measures such as binary outcomes (Bolker *et al.*, 2009). The formula used was $\texttt{IsCorrect} \sim \texttt{Visualization} \times \texttt{Size} \times \texttt{TaskType} + (1|\texttt{ParticipantID})$. Note that $\texttt{TaskType}$ was not specified as a random effect. This model was selected because it clearly provided the best fit based on indices of relative model fit. For accuracy, there may not have been sufficient variability in TaskType to specify it as a random effect.

The absolute fit of the accuracy model was assessed with the same methods performed with the task time model. For only fixed effects, the marginal pseudo-R^2 was 0.256; double the value of the task time model. For fixed and random effects, the conditional pseudo-R^2

Task Time (s)

	F	**p-value**	
Visualization	16.179	< 0.001	***
Size	56.331	< 0.001	***
PathSearchRadius	144.289	< 0.001	***
Vis.:Size	3.794	0.023	*
Vis.:PathSearchRadius	8.093	< 0.001	***
Size:PathSearchRadius	21.857	< 0.001	***
Vis.:Size:PathSearchRadius	2.486	0.011	*

Accuracy

	χ^2	**p-value**	
Visualization	9.3533	0.009	**
Size	1.3890	0.239	
PathSearchRadius	16.8297	0.002	**
Vis.:Size	10.5002	0.005	**
Vis.:PathSearchRadius	119.5751	< 0.001	***
Size:PathSearchRadius	51.6708	< 0.001	***
Vis.:Size:PathSearchRadius	40.3748	< 0.001	***

Table 2. Analysis with path search radius.

was 0.408, which was comparable to the task time model. Again, this indicates the selected model accounted for a large amount of variability in the data. ANOVA results for the fixed effects are shown in Table 3.

	χ^2	**p-value**	
Visualization	10.560	0.005	**
Size	34.339	< 0.001	***
TaskType	64.881	< 0.001	***
Vis.:Size	10.688	0.005	**
Vis.:TaskType	95.936	< 0.001	***
Size:TaskType	25.710	< 0.001	***
Vis.:Size:TaskType	20.055	0.010	*

Table 3. `Accuracy`: Type III Wald chi-square tests. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1.

Comparing visualization and graph size, we found many effects on participants' accuracy were similar to those on their task time. For example, participants exhibited significantly higher accuracy overall in node-link than matrix conditions, with no evidence of accuracy differences between the two matrix visualizations (see Figure 7). Additionally, participants were significantly more accurate when working with the small graph size than with the large graph size. There was a significant interaction between Visualization and

Size: the accuracy advantage of node-link was more pronounced when graph size was small than when it was large (again in correspondence with the task time results).

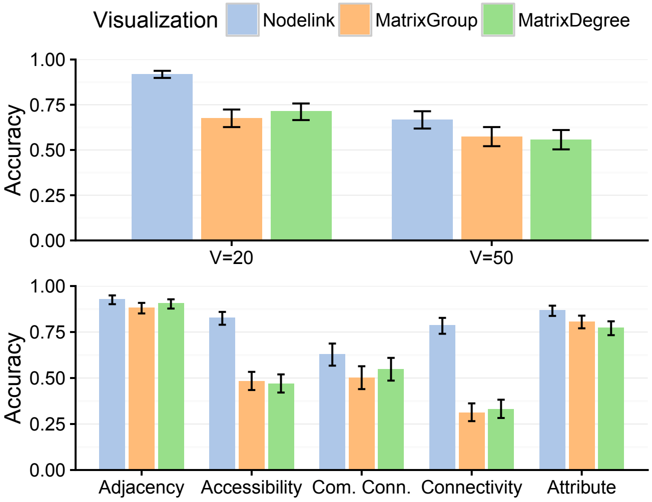


Fig. 7. Accuracy: Least-square means and 95% confidence intervals of `Accuracy`.

Accuracy was also significantly affected by TaskType, with higher accuracy in Adjacency and Attribute questions than the other three categories of question (see bottom panel of Figure 7). This implies the Adjacency and Attribute questions were much easier than the others, however, there was a significant interaction effect between Visualization and TaskType on participants' accuracy. In the three task types with the lowest accuracies (Accessibility, Common Connection, and Connectivity), the difference between node-link and matrix was much larger than it was in the other two categories. These three challenging task types caused a substantial decrement in accuracy in both MatrixDegree and MatrixGroup conditions, but only small decrements in Nodelink. In other words, the matrix visualization made the difficult TaskTypes even more difficult.

We also analyzed the path search radius in terms of accuracy. Similar to task time, we found a main effect of PathSearchRadius ($p < 0.001$), an interaction effect between PathSearchRadius and Visualization ($p < 0.001$), as well as an interaction between PathSearchRadius and Size ($p < 0.001$). We also found that the node-link visualization tended to outperform the matrix visualizations more in tasks that involve larger path search radiuses (see Figure 6 bottom).

5.2 Insight Bias

Participants completed five questions assessing insight in the post-study questionnaire, during which the graph visualization was no longer visible. For the first four questions (two multiple choice and two ranking-based questions) we did not find evidence for differences in accuracy across the three visualization types. On question I5, which required participants to estimate the size of the previously seen graph, we found that participants

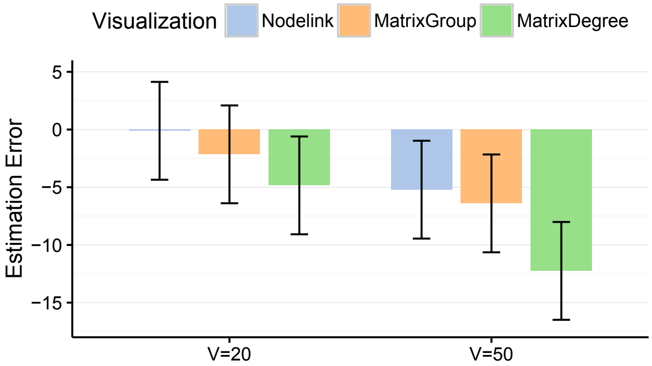


Fig. 8. (I5): Estimating the size of the graph using linear model: $\texttt{EstimationError} \sim \texttt{Visualization} \times \texttt{Size}$. Y-axis is estimation error (estimated size minus actual size). Error bars are 95% confidence intervals.

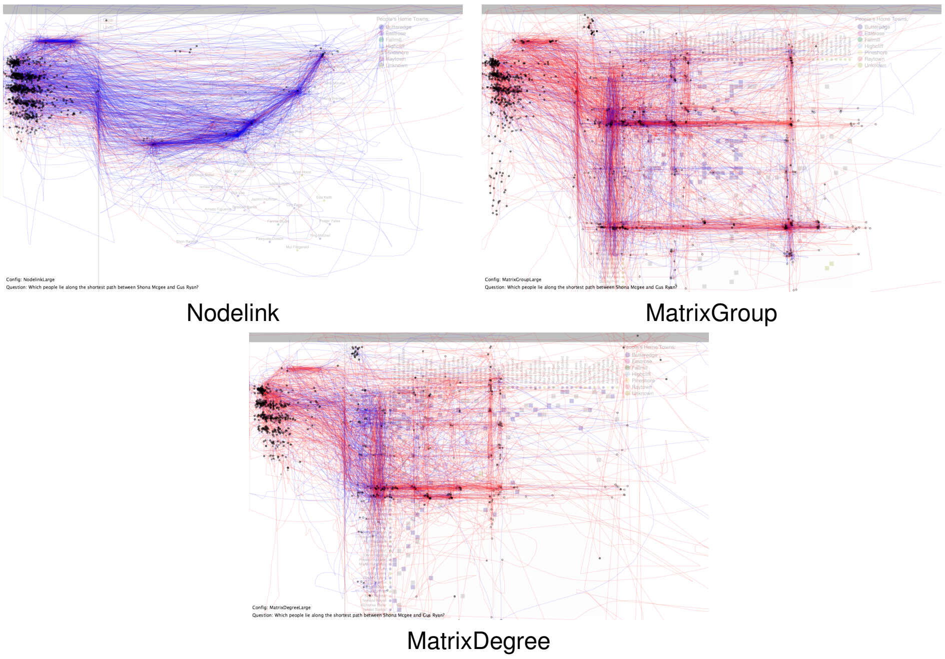
tended to underestimate the size of the graph in all conditions (see Figure 8). The size of the underestimation was affected by visualization type ($F = 3.969, p = 0.019$), with the most severe underestimation occurring in the MatrixDegree condition (with $\texttt{NodeLink}$ - $\texttt{MatrixDegree} = 5.9$ nodes). When the matrix is ordered by degree, its edges are more condensed towards the top-left corner, which may generate the illusion of a smaller network.

5.3 Interaction and Annotations

We collected mouse interaction data and annotations drawn for each participant on each task (see Figure 9 for an example). In this paper, we present an overall analysis of this data; detailed analysis and prediction are left for future work. We conducted an analysis on average mouse *moving* speed — total mouse trajectory length divided by total time the mouse was moving. We performed an analysis using the following model: $\log(\texttt{Speed}) \sim \texttt{Visualization} \times \texttt{Size} + (1|\texttt{ParticipantID})$. We found a significant main effect of visualization ($F = 104.823, p < 0.001$) and size ($F = 37.929, p < 0.001$). Details are shown in Figure 10. The mouse movement speed for node-link was faster than both matrix conditions ($p < 0.001$).

5.4 Learning Over Time

We assessed changes in task time and accuracy over the duration of the experiment to determine if participants exhibited implicit learning, and if learning rates differed among the three visualization types. Because questions were presented in a random order for each participant, time was represented using Question Index (the order of questions). For task time, we used the following model: $\log(\texttt{TaskTime}) \sim \texttt{Visualization} \times \texttt{Size} \times \texttt{QuestionIndex} + (1|\texttt{ParticipantID})$, treating question index as a continuous variable. We found a significant effect of QuestionIndex ($p < 0.001$), suggesting an overall reduction in task time over time in all three visualizations. However, we also found a significant


Config: NodelinkLarge
Question: Which people lie along the shortest path between Shona Mcgee and Gus Ryan?
Config: MatrixGroupLarge
Question: Which people lie along the shortest path between Shona Mcgee and Gus Ryan?
Config: MatrixDegreeLarge
Question: Which people lie along the shortest path between Shona Mcgee and Gus Ryan?

Nodelink

MatrixGroup

MatrixDegree

Fig. 9. Mouse trajectories and annotations for question CO2: "Which people lie along the shortest path between Person CO2-A and Person CO2-B?" in the large graph. Blue/red: trajectories of correct/incorrect answers.

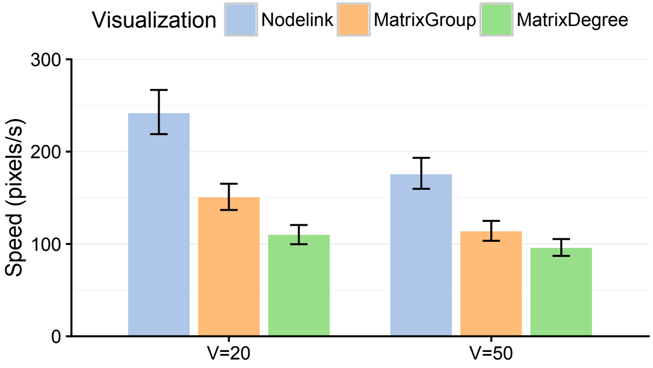


Fig. 10. Trajectory analysis. Participants move their mouse faster in the node-link conditions.

interaction between QuestionIndex and Visualization ($p = 0.043$). As shown in Figure 11 (top), task times decreased more rapidly in the matrix visualizations than in node-link. The large initial difference between node-link and the two matrix conditions diminished over the experiment; time for all three visualizations was comparable at the end.

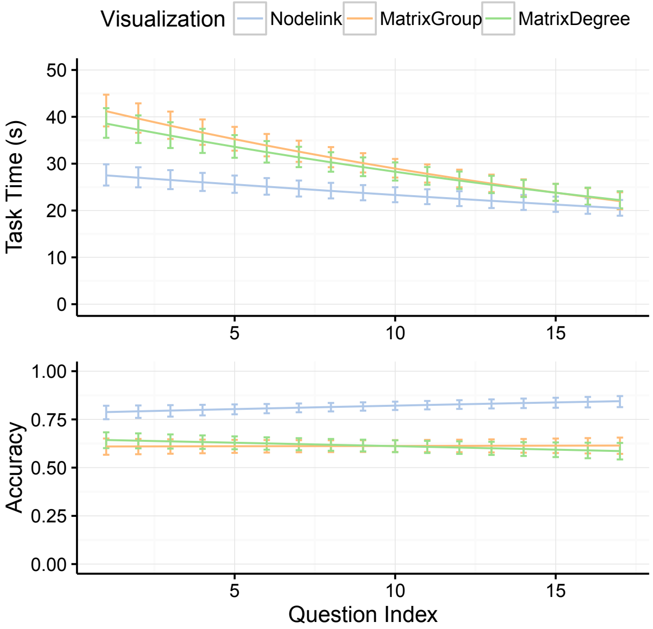


Fig. 11. Implicit Learning for Task Time and Accuracy over the duration of the experiment. Error bars represent 95% confidence intervals. The lines represent the fixed or overall effects for learning based on individual learning curves.

Using the corresponding model for accuracy, we did not find a significant effect of QuestionIndex for the three visualization types. Accuracy performance in node-link was superior to the two matrix conditions throughout the experiment (see Figure 11 bottom).

This result suggests that the slower task times associated with the matrix visualizations, when compared with node-link, may be overcome with practice. However, we did not observe a similar improvement in accuracy over the course of the study.

5.5 Self-Ratings

The post-study questionnaire included three items that required participants to provide self-ratings about the difficulty of the primary task, participants' performance on the primary task, and participants' performance on the ambient learning questions. Participants used a five-item Likert scale to record their responses to each of these items. We ran a linear model for both performance and difficulty (see Table 4).

Self-assigned Difficulty and Performance Visualization type affected participants' ratings of both the difficulty of the task ($F = 60.056, p < 0.001$)—estimated means and standard errors are shown in Table 4—and their performance on the task ($F = 50.659, p < 0.001$), with participants reporting less difficulty and better performance in the node-link visualization than either of the matrix visualizations. Although visualization also had an

Visualization	**Performance**	**Difficulty**	**AL Performance**
Node-link	3.78 (0.06)	3.45 (0.07)	2.61 (0.06)
MatrixGroup	2.97 (0.06)	2.55 (0.07)	2.40 (0.06)
MatrixDegree	3.08 (0.06)	2.59 (0.07)	2.48 (0.06)

Table 4. Means and standard errors for self-ratings. Performance (1 = Very Poor), difficulty (1 = Very Difficult), and ambient learning performance (1 = Very Poor).

influence on participants' ratings of their performance on the ambient learning questions ($F = 3.186, p = 0.042$), the size of the effect was much smaller.

Demographics Aligned with previous psychology research, task time (i.e., processing speed) increases as age increases (Salthouse, 1996). However, we did not find a significant impact of age on accuracy. We also did not find indications that age or gender interacts with the visualization and graph size.

Self-reported Familiarity As suggested by previous psychology research, individuals can be "unskilled and unaware" because less skilled individuals may be ignorant to their own lack of knowledge (Kruger & Dunning, 1999). Consistent with this finding, for accuracy, we found that people who reported "very familiar" with the visualizations performed worse than those who reported "not familiar" ($p = 0.019$)

5.6 Completion Rate

Because MTurk participants have the option to exit the study before completing it, we assessed completion rates across the different conditions to determine if there were any differences and to guide the design of future studies. In the Nodelink condition, 243 participants started the training (not counting multiple attempts from one participant), 98.8% of them completed training and 94.2% of them completed the tasks. However, in the matrix conditions, only 84.7% and 78.4% of the 589 participants who started training completed the training and tasks, respectively. The difference in completion rate between the matrix degree ordering and group ordering is less than 2%. Although the differences in attrition for training and completion were modest to slight, its selectivity may have had effects on the results. However, if anything, the lower self-selectivity in the matrix conditions should have inflated performance. Details of completion rate are shown in Table 5. The lower task completion rate on the matrix conditions aligns with our earlier findings that participants had, on average, more difficulty with the task questions in the two matrix visualizations than in the node-link visualization. The lower training completion rate in the matrix conditions additionally indicates that participants had more difficulty with the training material when it referenced a matrix view than when it referenced a node-link view.

	Node-link		**MatrixGroup**		**MatrixDegree**	
	20	50	20	50	20	50
S. Training	243		299		290	
C. Training	240 (99%)		251 (84%)		248 (86%)	
S. Tasks	110	130	115	136	126	122
C. Tasks	107	122	112	120	116	114
C. Tasks %	97%	94%	97%	88%	92%	93%
Summary	229 (94%)		232 (78%)		230 (79%)	

Table 5. Completion rate. Total number of participants in each condition who started training, completed training, started tasks, completed tasks, and summary % (completed tasks vs. started training).

6 Discussion

In this section, we discuss the results of our study and compare them with previous studies (Ghoniem *et al.*, 2004, 2005; Henry & Fekete, 2007).

Our study suggests that node-link is good in terms of both time and accuracy for most of the tasks we have tested. The only exceptions, where the matrix is better in terms of time, are those specific to the sorting of the matrix: 1) When sorted by degree, the matrix is better at finding the most connected node. In task AD3 (i.e., "mostConnected"), for the $V = 50$ graph, MatrixDegree was significantly faster than both node-link and MatrixGroup ($p < 0.001$ for both), and more accurate ($p < 0.001, 0.002$, respectively), but we did not find significant differences for the $V = 20$ graph. Node-link is more affected by the graph size in this task, which aligns with Henry & Fekete (2007)'s finding on this task. 2) When sorted by attribute (i.e., MatrixGroup), matrix is better for attribute-based tasks. In AT1 ("countTown"), node-link performed worse in accuracy than the MatrixGroup and MatrixDegree ($p = 0.047, 0.003$, respectively) in for $V = 20$, but we did not find significant differences for $V = 50$; however, in $V = 50$, MatrixGroup performed faster than node-link ($p < 0.001$). Other than these specific tasks, we did not find the matrix conditions superior to the node-link. The different results can be explained by the following reason: We tested the visualizations on social graphs that typically arise from analytical tasks, while Ghoniem *et al.* conducted their experiment on random graphs. The results on random graphs do not necessarily generalize to the type of graphs we experimented with. This is because social graphs typically have an unbalanced distribution of degrees (few nodes have more connections, and most nodes have few connections), which can benefit the node-link visualization.

While we did not experiment with "nodeCount" and "edgeCount" directly from the study, we asked the participants to estimate the node count based on their impression (i.e., in the post-study questionnaire without the visualization). We found that they tend to underestimate the node count in all conditions, and more severely in the matrix with degree sorting. This differs from Ghoniem *et al.*'s results, because in random graphs with name sorting, the matrix looks like a square (see Figure 1B in Ghoniem *et al.* (2005), as an example). In contrast, our matrix edges are more distributed towards the top. Therefore, the sorting of the matrix should be chosen carefully to avoid bias when the impression of graph size is important.

We found a significantly higher mouse-movement speed with the node-link visualization. Participants often followed edges in the node-link representation and annotated nodes, whereas, for matrix, participants followed rows or columns and annotated node labels, matrix cells (edges) and entire rows/columns (nodes). The higher speed could just be an intrinsic result of these behavioral patterns, but could also indicate higher confidence in navigation for the node-link case. This is left for future work to tease out.

Previous research has indicated that matrix visualizations have a higher learning curve. In our analysis of performance over time, although we found that people become faster with the matrix visualizations, we found no indication that they become more accurate. As one might expect, the higher learning curve of the matrix visualization was reflected in our insight measurements as an estimation error on the size of the graph. This effect is also reflected as slightly decreased user confidence (Table 4).

7 Limitations

One potential concern with our methodology involves the higher dropout rates observed in the matrix conditions as compared to the node-link condition. This introduces a possible confound, in which our sample of participants that did not drop out in the two matrix visualizations may have been, on average, more persistent, intelligent, and/or motivated than the sample of participants in the node-link visualization. Additionally, we collected our data on MTurk in multiple discrete batches, with random condition assignment. Over the course of data collection, we adjusted the participant assignment probabilities based on completion rates from previous batches. As a result, participants in the matrix conditions were disproportionately gathered in later batches. To avoid this phenomenon in future studies, we will use pilot studies to determine estimated completion rates across conditions, and then use a fixed assignment probability in later data collection.

A second limitation of the study concerns the relatively small scale and static nature of the graphs that were evaluated, compared against Ghoniem *et al.* (2005) and Henry & Fekete (2007), for example. Our motivation for this design was multifaceted. First, our aim was to use examples that were simulated, but were representative of real world intelligence analysis scenarios, and were verified by expert analysts. Second, due to the effort required to gather human interaction data at scale, we opted for a simpler data set to allow for mining of meaningful interaction patterns and work-flows in the data (see Figure 9, for example). We are conducting a follow-up analysis to apply machine learning techniques over the corpus of interaction patterns and resulting performance data, to gain a better understanding of efficient (and inefficient) tasks in the different graph representations. We believe that the results of the analysis will contribute to training of information analysts who use these views in a broad range of domains and applications. However, to further generalize the results in this paper, we are also planning a follow-up study using systematically varied graph size and density, with randomized layout and connectivity.

8 Conclusion

In this paper, we conducted a large-scale study ($N = 600$) to compare the node-link, matrix (ordered by degree), and matrix (ordered by group) representations. In these approximately one-hour-long studies, participants used one of the three graph representations to answer a broad range of questions covering a variety of graph tasks. Objective and subjective performance metrics were recorded for each, and a detailed statistical analysis was performed, followed by a discussion of the key results. To summarize, results show that node-link representations produced a better implicit understanding of the data, with higher response accuracy and faster completion times than the matrix representations (23% and 19% less time than MatrixDegree and MatrixGroup, respectively). This result is consistent with prior work (Henry & Fekete, 2007; Ghoniem *et al.*, 2005). For the larger social network graph, the performance difference between matrix and node-link representations was reduced, with node-link still performing best overall. Participants had better overall performance in terms of speed and accuracy on the smaller graph. A statistical analysis did not reveal a significant relation between time spent on analysis and accuracy of responses. Our large sample size also allowed for comparison of learning rates across the duration of the experiment. Learning results indicate a large initial difference in task time between the node-link and matrix visualizations, with matrix performance steadily approaching that of the node-link visualization over the course of the experiment. We also collected information on participants' mouse movements and annotations, and found that the node-link representation led to significantly higher mouse movement speeds during question answering.

To conclude, our study results show that node-link representations are more efficient analysis tools than matrix representations for the majority of smaller-scale social network graph analysis tasks, with the exception of those that require explicit sorting of nodes.

9 Acknowledgments

We thank all the participants from Amazon Mechanical Turk who participated in the study. This material is based in part upon work supported by the U.S. Army Research Laboratory (ARL) for D.R., J.O.D., J.S., W.L., and T.H. under Cooperative Agreement No. W911NF-09-2-0053 (Network Science Collaborative Technology Alliance), the ARL Senior Fellowship Program for L.R.M. administered by the Oak Ridge Associated Universities under Cooperative Agreement No. W911NF-17-2-0003, and NSF grant IIS-1748392 for D.R. and T.H. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of ARL or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation herein.

References

Alper, Basak, Bach, Benjamin, Henry Riche, Nathalie, Isenberg, Tobias, & Fekete, Jean-Daniel. (2013). Weighted graph comparison techniques for brain connectivity analysis. *Pages 483–492 of: Proceedings of the sigchi conference on human factors in computing systems*. ACM.

Army, U. S. (2006). *Field manual 2-22.3: Human intelligence collector operations.*

Baayen, R Harald, Davidson, Douglas J, & Bates, Douglas M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language*, **59**(4), 390–412.

Baccara, Mariagiovanna, & Bar-Isaac, Heski. (2008). How to organize crime. *The review of economic studies*, **75**(4), 1039–1067.

Bakker, Marjan, van Dijk, Annette, & Wicherts, Jelte M. (2012). The rules of the game called psychological science. *Perspectives on psychological science*, **7**(6), 543–554.

Bates, Douglas, Mchler, Martin, Bolker, Ben, & Walker, Steve. (2015). Fitting linear mixed-effects models using lme4. *Journal of statistical software*, **67**(1), 1–48.

Battista, Giuseppe Di, Eades, Peter, Tamassia, Roberto, & Tollis, Ioannis G. (1998). *Graph drawing: Algorithms for the visualization of graphs*. 1st edn. Upper Saddle River, NJ, USA: Prentice Hall PTR.

Berardi, Christopher W, Solovey, Erin Treacy, & Cummings, Mary L. (2013). Investigating the efficacy of network visualizations for intelligence tasks. *Pages 278–283 of: Intelligence and security informatics (isi), 2013 ieee international conference on*. IEEE.

Blanchet, Karl, & James, Philip. (2011). How to do (or not to do)a social network analysis in health systems research. *Health policy and planning*.

Bohannon, John. (2009). Counterterrorism's new tool: 'metanetwork' analysis. *Science*, **325**(5939), 409–411.

Bolker, Benjamin M, Brooks, Mollie E, Clark, Connie J, Geange, Shane W, Poulsen, John R, Stevens, M Henry H, & White, Jada-Simone S. (2009). Generalized linear mixed models: a practical guide for ecology and evolution. *Trends in ecology & evolution*, **24**(3), 127–135.

Bostandjiev, Svetlin, O'Donovan, John, Hall, Christopher, Gretarsson, Brynjar, & Hollerer, Tobias. (2011). Wigipedia: A tool for improving structured data in wikipedia. *Pages 328–335 of: Proceedings of the 2011 ieee fifth international conference on semantic computing*. ICSC '11. Washington, DC, USA: IEEE Computer Society.

Buhrmester, Michael, Kwang, Tracy, & Gosling, Samuel D. (2011). Amazon's mechanical turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science*, **6**(1), 3–5.

Burnham, Kenneth P, & Anderson, David R. (2003). *Model selection and multimodel inference: a practical information-theoretic approach*. Springer Science & Business Media.

Chang, Chunlei, Bach, Benjamin, Dwyer, Tim, & Marriott, Kim. (2017). Evaluating perceptually complementary views for network exploration tasks. *Pages 1397–1407 of: Proceedings of the 2017 chi conference on human factors in computing systems*. ACM.

Chang, Remco, Ziemkiewicz, Caroline, Green, Tera Marie, & Ribarsky, William. (2009). Defining insight for visual analytics. *Computer graphics and applications, ieee*, **29**(2), 14–17.

Cohen, Jacob, Cohen, Patricia, & West, Stephen. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences*. 3 edn. Mahwah, NJ: Lawrence Erlbaum Associates.

Eysenck, Michael W, & Keane, Mark T. (2013). *Cognitive psychology: A student's handbook*. Psychology press.

Ghoniem, Mohammad, Fekete, Jean-Daniel, & Castagliola, Philippe. (2004). A comparison of the readability of graphs using node-link and matrix-based representations. *Pages 17–24 of: Proceedings - ieee symposium on information visualization, info vis.*

Ghoniem, Mohammad, Fekete, Jean-Daniel, & Castagliola, Philippe. (2005). On the readability of graphs using node-link and matrix-based representations: a controlled experiment and statistical analysis. *Information visualization*, **4**(2), 114–135.

Gretarsson, Brynjar, O'Donovan, John, Bostandjiev, Svetlin, Höllerer, Tobias, Asuncion, Arthur, Newman, David, & Smyth, Padhraic. (2012). Topicnets: Visual analysis of large text corpora with topic modeling. *Acm trans. intell. syst. technol.*, **3**(2), 23:1–23:26.

Hall, David L., Graham, Jake, & Catherman, Emily. (2015). A survey of tools and resources for the next generation analyst. vol. 9499.

Hauser, David J, & Schwarz, Norbert. (2015). Attentive turkers: Mturk participants perform better on online attention checks than do subject pool participants. *Behavior research methods*, 1–8.

Henry, N., Fekete, J.-D., & McGuffin, M. J. (2007). Nodetrix: a hybrid visualization of social networks. *Ieee transactions on visualization and computer graphics*, **13**(6), 1302–1309.

Henry, Nathalie, & Fekete, Jean-Daniel. (2007). Matlink: Enhanced matrix visualization for analyzing social networks. *Pages 288–302 of: Proceedings of the 11th ifip tc 13 international conference on human-computer interaction - volume part ii.* INTERACT'07. Berlin, Heidelberg: Springer-Verlag.

Jaworowski, M, & Pavlak, S. (2003). *Ali baba dataset ground truth.* U.S. National Security Agency: Fort Meade, MD.

Kase, Sue E., Roy, Heather, & Cassenti, Daniel N. (2015). Visualizing approaches for displaying measures of sentiment. vol. 9499.

Krebs, Valdis E. (2002). Mapping networks of terrorist cells. *Connections*, **24**(3), 43–52.

Kruger, Justin, & Dunning, David. (1999). Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments. *Journal of personality and social psychology*, **77**(6), 1121.

Lankow, Jason, Ritchie, Josh, & Crooks, Ross. (2012). *Infographics: The power of visual storytelling*. Wiley.

Lee, Bongshin, Plaisant, Catherine, Parr, Cynthia Sims, Fekete, Jean-Daniel, & Henry, Nathalie. (2006). Task taxonomy for graph visualization. *Pages 1–5 of: Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM.

MacCalman, Molly, MacCalman, Alexander, & Wilson, Greg. (2013). Visualizing social networks to inform tactical engagement strategies that will influence the human domain. *Small wars journal*, **9**(8).

McGrath, Cathleen, Blythe, Jim, & Krackhardt, David. (1997). The effect of spatial arrangement on judgments and errors in interpreting graphs. *Social networks*, **19**(3), 223 – 242.

McIllwain, Jeffrey Scott. (1999). Organized crime: A social network approach. *Crime, law and social change*, **32**(4), 301–323.

Mittrick, M, Roy, H, Kase, S, & Bowman, E. (2012). *Refinement of the ali baba data set.* US Army Research Laboratory, ARL-TN-0476.

Nakagawa, Shinichi, & Schielzeth, Holger. (2013). A general and simple method for obtaining r2 from generalized linear mixed-effects models. *Methods in ecology and evolution*, **4**(2), 133–142.

Newman, Mark EJ. (2002). Spread of epidemic disease on networks. *Physical review e*, **66**(1), 016128.

North, Chris. (2006). Toward measuring visualization insight. *Computer graphics and applications, ieee*, **26**(3), 6–9.

Okoe, Mershack, & Jianu, Radu. (2015). Graphunit: Evaluating interactive graph visualizations using crowdsourcing. *Pages 451–460 of: Computer graphics forum*, vol. 34. Wiley Online Library.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, **349**(6251), aac4716–aac4716.

Peng, Roger D. (2011). Reproducible research in computational science. *Science*, **334**(6060), 1226–1227.

Purchase, Helen C. (1998). Performance of layout algorithms: Comprehension, not computation. *Journal of visual languages & computing*, **9**(6), 647 – 657.

Salthouse, Timothy A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological review*, **103**(3), 403.

Schaffer, James, Giridhar, Prasanna, Jones, Debra, Höllerer, Tobias, Abdelzaher, Tarek, & O'Donovan, John. (2015). Getting the message?: A study of explanation interfaces for microblog data analysis. *Pages 345–356 of: Proceedings of the 20th international conference on intelligent user interfaces*. ACM.

Sparrow, Malcolm K. (1991). The application of network analysis to criminal intelligence: An assessment of the prospects. *Social networks*, **13**(3), 251–274.

Sullivan, Peter. (1987). *Newspaper graphics*. Darmstadt, Germany: IFRA.

Von Landesberger, Tatiana, Kuijper, Arjan, Schreck, Tobias, Kohlhammer, Jörn, van Wijk, Jarke J, Fekete, Jean-Daniel, & Fellner, Dieter W. (2011). Visual analysis of large graphs: state-of-the-art and future research challenges. *Pages 1719–1749 of: Computer graphics forum*, vol. 30. Wiley Online Library.

Wong, Pak Chung, Foote, Harlan, Mackey, Patrick, Perrine, Ken, & Chin Jr., George. (2006). Generating graphs for visual analytics through interactive sketching. *Ieee transactions on visualization and computer graphics*, **12**(6), 1386–1398.

Yi, Ji Soo, Kang, Youn-ah, Stasko, John T, & Jacko, Julie A. (2008). Understanding and characterizing insights: how do people gain insights using information visualization? *Page 4 of: Proceedings of the 2008 workshop on beyond time and errors: novel evaluation methods for information visualization*. ACM.