Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach

Carolina Nobre[1], **Dylan Wootton**[1], **Lane Harrison**[2], **Alexander Lex**[1]
[1]University of Utah, [2]Worcester Polytechnic Institute
[1]Salt Lake City, UT, USA, [2]Worcester, MA, USA
cnobre@sci.utah.edu, wootton.dylan@gmail.com, ltharrison@wpi.edu, alex@sci.utah.edu

ABSTRACT
Visualizing multivariate networks is challenging because of the trade-offs necessary for effectively encoding network topology and encoding the attributes associated with nodes and edges. A large number of multivariate network visualization techniques exist, yet there is little empirical guidance on their respective strengths and weaknesses. In this paper, we describe a crowdsourced experiment, comparing node-link diagrams with on-node encoding and adjacency matrices with juxtaposed tables. We find that node-link diagrams are best suited for tasks that require close integration between the network topology and a few attributes. Adjacency matrices perform well for tasks related to clusters and when many attributes need to be considered. We also reflect on our method of using validated designs for empirically evaluating complex, interactive visualizations in a crowdsourced setting. We highlight the importance of training, compensation, and provenance tracking.

Author Keywords
Multivariate networks visualization; crowdsourced evaluation.

CCS Concepts
•Human-centered computing → Empirical studies in visualization;

INTRODUCTION
Multivariate networks (MVNs) are networks comprised of the network's topology in the form of nodes and links, and attributes about those nodes and links. Most real-life networks are multivariate: a social network has node attributes such as the age, name, and institutional affiliations of individuals; and edge attributes such as the types and frequencies of interactions. In many analysis cases, it is necessary to see both the topology and attributes simultaneously [27]. A recent survey by Nobre et al. [24] discusses 11 types of multivariate network visualizations (MVNVs) and gives recommendations of when to use which visualization technique. These recommendations, however, are mostly based on visualization best practices and rely to only a small degree on empirical data, perpetuating the notion of visualization as an "empire built on sand" [21]. How do we know whether a particular visual encoding or an interaction technique is better than the many alternatives available? The visualization literature largely relies on two approaches: controlled experiments that measure correctness and time on simplified tasks for well-defined stimuli [22] and evaluations with experts, such as insight-based evaluation [32, 28] and case studies [34]. Carpendale [6] discusses the conflict between evaluating visualizations with real users and real datasets, with high ecological validity, and controlled experiments that require recruiting a large enough participant sample to draw quantitative conclusions. Running controlled experiments requires abstraction and simplification of real-life tasks and careful variation of a small number of design factors [22].

In this paper, we attempt to answer questions about the merits of two multivariate network visualization techniques by pushing the limits of controlled experiments for complex, interactive visualization techniques. We compare two MVNVs: node-link diagrams (NL) with on-node encoding and adjacency matrices (AM) with a juxtaposed tabular visualization. Testing these two conditions required the design and implementation of two functional prototypes. We designed and implemented these techniques based on existing guidelines, and followed a multi-stage testing and piloting process. We also elicited feedback from experts on network visualization to validate and refine the designs. The aim of these activities was to ensure that the two conditions in the experiment resemble visualizations that might be used in practice. We postulate eight hypotheses and evaluate them with a set of 16 tasks, which we derive from Nobre et al.'s task analysis [24]. We also report on insights generated in an open-ended task.

Our contributions are twofold: We provide the first set of empirical evidence on the performance of two important MVNV visualization techniques for different tasks, and we develop and reflect on an approach for controlled experiments using complex, validated visualization techniques.

RELATED WORK
Here we provide an analysis of prior evaluations of network visualization techniques, followed by a discussion of the current landscape of crowdsourced evaluations of interactive visualization techniques. For prior work regarding MVNVs, we refer to a recent survey [24] and to Kerren et al.'s book [19].

This is the authors' preprint version of this paper. Please cite the following reference:
Carolina Nobre, Dylan Wootton, Lane Harrison, Alexander Lex. Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), to appear, 2020.

Network Evaluation Studies

There is a long history of work that has evaluated the merits of different network representations [38]. However, most of this work focuses on network topology and treats node and edge attributes only cursorily, if at all. We limit our discussion to larger studies that evaluate NLs and/or AMs.

Several studies compare NL and AM representations [14, 18, 7, 26, 30]. Ghoniem et al. [14] assessed the performance of both approaches using seven topology-based tasks on graphs of sizes between 20 and 100 nodes and densities from 20% to 60% of all possible edges. Interactivity was limited to selecting nodes and links. They found that NL outperformed AM in small and sparse graphs, but for larger or more dense graphs, AM produced more accurate results. The exception is path-based tasks, where node-link diagrams outperformed the adjacency matrix regardless of network size or density. In a follow-up study, Keller et al. [18] evaluated six tasks on a domain-specific, directed network, using both NL and AM representations. They confirmed the results of Ghoniem et al. for the investigated network types. Okoe et al. [26] also reproduced the Ghoniem et al. study for larger networks (up to 330 nodes), but for much sparser graphs with densities of 1.6% to 3.2% of edges. They used a more diverse set of tasks in a large, crowdsourced study. Interactions included panning and zooming, moving nodes in the NL condition, and highlighting nodes and incident edges. Color was used to encode clusters that were detected algorithmically. Their work confirms earlier findings on each technique and reveals that adjacency matrices perform better on cluster tasks.

Ren et al. [30] compared NL with two sorting variants of AMs for networks of 20 and 50 nodes in a large, crowdsourced study. They found that NL resulted in higher accuracy and faster response time, but that this difference diminished as participants became familiar with the visualizations.

Three studies evaluated approaches for visualizing two or more edge attributes. Alper et al. [3] found that for tasks involving the comparison of weighted graphs, AMs outperformed NLs. Abuthawabeh et al. [1] ran a user study comparing AMs with a technique that depicts multiple types of edges in separate, parallel, node-link diagrams. They found that participants identified the same graph structures with both visualizations. Schoeffel et al. [33] encoded edge attributes as bars directly on the edges in NLs. Their study on a small network (10 nodes, 10-15 edges, up to five attributes) revealed that this can be useful for small graphs with no edge crossings.

The existing work on network evaluation studies thus far has mostly focused on the topology of the network. No studies currently consider node attributes beyond a simple, topology-derived attribute, such as cluster membership or node degree, which means we currently have no guidance based on empirical data about which network visualization technique to use when both node attributes and network topology are relevant to an analysis.

Evaluation of Interactive Visualization Systems

User studies are among the most common forms of evaluation within the field of information visualization [22]. Lam et al. [22] have categorized user studies that are aimed at evaluating user performance into two types: (1) understanding the limits of visual perception and cognition for specific visual encodings, and (2) assessing how one visualization or interaction technique compares to another. User studies can be carried out either in a controlled lab setting or by using a crowdsourcing platform. Although lab studies afford the most control over participant selection, participant attention, training, and the testing environment, they also incur a high cost of recruiting users, as well as an inherently limited participant pool [13]. Crowdsourcing platforms offer a potential solution to this problem by providing access to a much larger group of participants. The existing user performance work on evaluating interactive visualizations is characterized primarily by validating a new approach comparing it with existing ones [4, 2], which is almost exclusively carried out in a lab setting with a smaller number of participants. Even though crowdsourced studies have been frequently used for performance evaluations of the perceptual type (e.g., [35, 16, 15]), they are relatively scarce for evaluating interactive visualization techniques (e.g., [26, 11]). This scarcity may stem from the perceived challenges of using a remote group of non-expert participants to evaluate an interactive system — a topic we investigate in this work. Additionally, very little work has compared two or more existing **complex, interactive** techniques.

CHALLENGES

The visualization community has embraced a wide set of quantitative and qualitative evaluation methodologies, ranging from quantitative experiments conducted in a lab or on crowdsourcing platforms, to qualitative studies, to insight based evaluation, to case studies [22]. In this paper, we conduct a crowdsourced study with two complex, interactive visualization techniques. Developing such techniques requires many design decisions. How can we know that any effects we observe are not confounded by any of these design decisions? Carpendale [6] describes three factors to consider: generalizability (can a study be applied to other people and situations), precision (can a study be definite about the measurement and can it account for the factors), and realism (is the context of the study like the context in which it will be used). However, current evaluation methodologies typically cannot satisfy all three simultaneously. A common approach to designing a quantitative study is to carefully modify selected factors or variables, so that the effect of each factor can be isolated. This approach leads to studies that are precise, but frequently not realistic. The need to isolate factors and variables significantly limits the progress we can make with one study and reduces realism. In this paper, we take a different approach: we compare two techniques that are distinct in many ways simultaneously. Although this approach poses a threat to generalizability, we mitigate this threat by employing a rigorous design process following existing evidence and our own expertise, and by validating our designs in a multi-staged qualitative process. Being able to compare two complex techniques increases realism since we can include factors that a real system would have. By simultaneously designing both techniques, we can better account for confounding variables than if we compared two existing systems that were designed separately.

A related challenge is that in a crowdsourced study, we have to use novices as proxies for experts. Complex and interactive visualization techniques designed for experts are frequently not immediately intuitive and require learning before they can be useful. Obviously, it would be best to validate a system with the user group it was designed for. In practice, however, recruiting researchers knowledgeable about network analysis is difficult, and becomes de facto impossible when a study aims to increase precision by increasing the number of participants. In our study, we attempt to educate participants about a visualization technique through an extensive training program (by crowdsourcing standards). We argue that careful training makes it possible to study interactive visualization techniques that would otherwise be suitable only for experts.

VISUALIZATION AND INTERACTION DESIGN

Several approaches are available for encoding both the topology and the attributes of a network [24]. For this study, we chose two common encodings: a node-link diagram with on-node/on-edge encoding (NL), and an adjacency matrix with embedded edge encoding and a juxtaposed table for the node attributes (AM). We chose these two techniques because (a) they are the most common generic network visualization techniques, and (b) several previous studies have compared NL and AM for topology-centric tasks [14, 26, 25, 38, 30], which allows us to use the findings of these studies to design our techniques based on these empirical recommendations.

When designing the techniques, we made choices that we justify in this section. We adopted the following design principles: (1) Use the most perceptually efficient encoding available for data, given the affordances of the technique. (2) Follow common practice in network visualization systems, such as Cytoscape of Gephi. (3) Provide a set of common, technique-specific interactions. In order to ensure our design decisions for each visualization technique were appropriate for experimentation, we used a multi-stage validation approach with expert evaluations and three rounds of pilots. Both designs can be viewed at `https://vdl.sci.utah.edu/mvnv-study/`, or by running the supplementary code.

For the feedback from visualization experts, we followed the heuristic evaluation method proposed by Wall et. al. [36] and added a set of custom questions tailored to our design decisions. We ran a pilot with one expert to evaluate our survey, which led to design recommendations we implemented but also to issues identified in our survey. The final survey had 69 questions that consisted of ratings on a 7-point Likert scale (high marks are good) and free-text questions asking for design suggestions and criticism. We then asked 10 experts to explore the two techniques with a set of four representative tasks and then to fill out the survey. While responses were helpful, we received only two responses within two weeks and one more response after we had already conducted the main study. The pilot and the two timely responses yielded multiple suggestions for improvements, which we incorporated, or, if not, discuss in this paper. Both the survey and the results are included in the supplemental material. Overall, experts rated the version of the tool they saw (before we implemented their suggestions for improvements) critically. The average rating for the heuristic evaluation question for the Node-Link diagram was 3.85; the average rating for the custom questions was 4.47. For the matrix, the average heuristic score was 5.69; the custom questions were scored at an average of 5.03. We attribute these low scores, particularly for the NL condition, partially to the difficulty of framing the questions. We attempted to get ratings for our design decisions, assuming the two techniques and the dataset as given, i.e., the best designed node-link diagram should get a perfect 7 on all questions. We found, however, that the experts mostly judged the techniques globally.

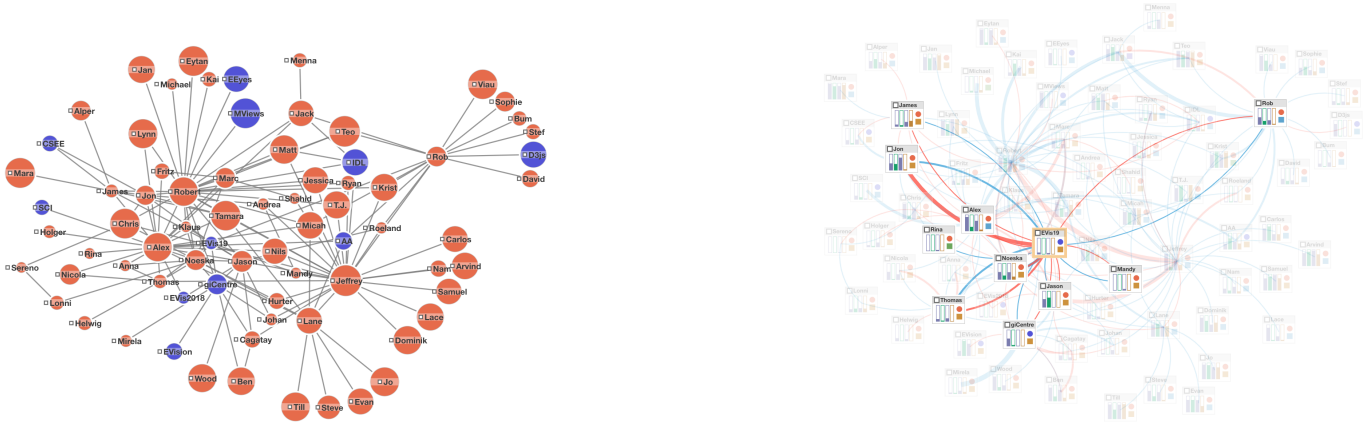
Node-Link Design

We studied node-link layouts with on-node encoding, as defined by Nobre et al. [24], since this is a widely used encoding and is supported by common network visualization tools. A key decision was to use traditional network layouts, including node positioning, exclusively for visualizing topological structure. We considered alternative layouts, such as positioning nodes by attribute values, as a set of separate techniques beyond the scope of our study. We experimented with many different layout algorithms and parameterizations. We calculated layouts using Cytoscape's implementation of the Prefuse force-directed layout, which we optimized manually.

We designed two versions of **node-attribute** encodings, shown in Figure 1, which we used depending on the number of attributes being encoded at once. Both versions always show all node labels. For conditions involving one numerical and/or one categorical attribute, we encoded the numerical attribute with the size of a circle and the categorical attribute as color, as shown in Figure 1a. Both of these choices are the highest ranked available visual channel for the respective data type [23, p. 95]. This design has the advantage that both a numerical and a categorical attribute can be visualized simultaneously with little interference, and global judgments of the individual attributes are easy. For conditions involving more than two attributes, we used nested bar charts for numerical values and colored glyphs for categorical values, as shown in Figure 1b. These encodings again use the most expressive visual channels available and support the comparison of multiple values within a node well. For this encoding, global judgments are more difficult due to the smaller mark size and the many marks present. As the nodes have to be bigger to accommodate multiple glyphs, the network topology is less apparent than in the simpler configuration. We include a legend showing the meaning of all visual encodings in a given configuration.

We visualize **edge attributes** using color and/or edge thickness for categorical and quantitative data respectively. Figure 1b shows two edge types with different edge weights. We use straight lines for edges for conditions with only a single type, but use separate, curved edges when edges of multiple types are present. There is conflicting evidence about the merits of curved edges [37, 17].

We limited **interactions** to a small set that we considered essential for the techniques. For both techniques, we provided a label-based search, highlighting of individual nodes, tooltips, and highlighting of neighbors of nodes. To highlight neighbors of a selected node, we faded all non-neighbors out, as shown in Figure 1b. We found the fade-out necessary



(a) On-node encoding with color and size.

(b) On-node encoding with nested bars and glyphs.

Figure 1: Node-link designs. (a) We use node size and color for conditions with one categorical and/or one quantitative attribute. (b) We use bar charts and colored glyphs for more than two attributes. Figure (b) also shows edge types (color) and edge weights (thickness), in addition to neighborhood highlighting: only neighbors of the selected node *EVis19* are rendered opaquely.

in the node-link diagram due to the high amount of visual clutter caused by crossing edges and colorful nodes. Previous work found that without neighbor highlighting, users quickly became frustrated [14, 26]. We also provided the ability to drag nodes, which is essential to disambiguate edge crossings.

Adjacency Matrix

Adjacency matrices are the second major class of network visualization. Network topology is encoded in the matrix, where filled-in cells indicate the presence of an edge. Matrices rely on seriation or sorting algorithms to reveal topological patterns of interest, although neighborhoods are also easily identified by scanning a row or column. A key benefit of matrices is their ability to visualize dense networks, since every possible edge is already represented on the screen.

Edge attributes are commonly encoded using brightness/saturation or color for quantitative and categorical data types. We use a gray color scale to encode numerical values. Multiple edges are more difficult to encode, since the space available for edges in a cell is small. Alper et al. [3] discuss various encodings and argue for nested rectangles of different color, where the brightness encodes a numerical attribute for two edge types with weights, an approach we adopt.

A simple but efficient way of encoding **node attributes** in adjacency matrices is using juxtaposed tables. We juxtapose a tabular visualization [29, 12] with the matrix, and keep the rows consistent between the matrix and the table [20, 31, 5], as shown in Figure 2. As a result, we can use highly efficient encodings for the attributes. We employ aligned bars for quantitative attributes, which use position and size redundantly, and spatial region and color for categorical data. As for the NL, we provide **interactions** such as highlighting, tool-tips, and node search. On cell/edge selection or hovers, we highlight the row and column at this intersection, and the equivalent row/column on the other side of the diagonal. We also highlight neighbors; however, in this case we use simple (green) color highlighting (see Figure 2, as the matrix does not suffer from cluttering. As a (more powerful) alternative to moving nodes, we introduce sorting. The matrix can be sorted interactively by node label, a seriation algorithm [10], by the neighborhood of a node, or by any attribute.

Discussion

It was our goal to develop two techniques that we can compare fairly by carefully designing each technique to use the best possible visual encodings and interactions. We would then be able to measure the inherent benefits and drawbacks of the techniques. Even though we believe that the visual encodings are largely equivalent, our expert evaluators argued that the interactions are not. In particular, interactive sorting in the adjacency matrix does not have an equally powerful counterpart in the node-link diagram. However, we argue that the set of interaction techniques we provide is consistent with what is commonly expected of the respective visualization techniques, and sorting/seriation is essential but also natural in matrix layouts. One expert suggested we also include sorting for nodes, resulting in attribute-driven positioning of the nodes. Even though this operation is reasonably equivalent to sorting in the adjacency matrix, it leads to significant clutter and overlap in the node-link diagram. Another suggestion from an expert was to provide a query and filter system, for example, using scented widgets in the legend. We decided against enabling filters and queries, as we would mostly be testing that query and filter system, not the inherent qualities of the alternative network visualization approaches.

STUDY DESIGN

We aim to investigate the strengths and weaknesses of AMs and NLs for a diverse set of tasks on multivariate networks. Our study used a between-subjects design in which each participant was randomly assigned to either the AM or NL condition.

Procedure

In each condition, participants were assigned 16 tasks, with the first 15 presented in random order, and a final free exploration task presented at the end of the study. We recruited participants on Prolific, a crowdsourcing platform with a research focus. Based on completion times of pilot experiments, each participant was paid $12.50 USD, for an estimated duration of 40 minutes, resulting in an hourly rate of about $15 USD. All participants viewed and agreed to an IRB-approved

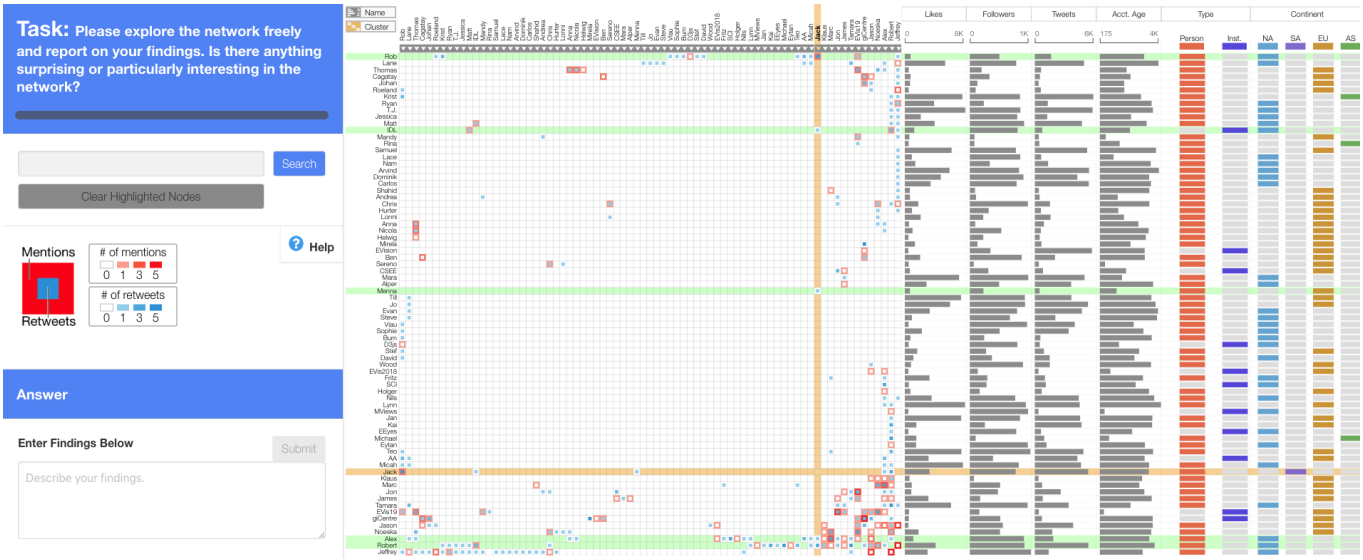


Figure 2: Adjacency matrix design. Two types of edge attributes are encoded using nested rectangles (red and blue). The color saturation encodes edge weights. Attributes are visualized in the juxtaposed table. The matrix is sorted by clusters, and the selected node's neighbors are highlighted in green. **This figure also shows the study interface** with the task instructions and the answer field, the search interface, and the legend.

consent form. To be eligible for the study, participants had to use a laptop or desktop device with a resolution of at least 1400x850 pixels available screen space in the browser. Our procedure consisted of five phases: *Passive Training*, *Active Training*, *Trials*, *Study*, and *Demographics and Feedback*. The full study for both conditions can be viewed at https://vdl.sci.utah.edu/mvnv-study/. **Passive training** was a video introduction to the dataset, as well as an explanation of the technique (NL or AM) with which the participant would be interacting. Participants had to watch the video before being allowed to continue. The training also introduced analysis strategies that are useful for tracing paths or identifying clusters. **Active Training** was achieved with a guided tour of the actual visualization and interaction mechanisms. Participants had to use interactions to be allowed to proceed. During **Trials**, participants had to correctly answer two tasks to proceed with the study. These tasks were meant to further train the user on the technique and to verify participants were attentive. During the **Study** itself, tasks were presented in random order to minimize learning effects across participants. Answers were given as a set of selected nodes, through a controlled or free text form, or as a combination of both. Each task was followed by a survey asking for confidence, perceived difficulty, and comments. The study interface is shown in Figure 2. A final form collected **Demographics and Feedback** on the study and the training.

Measures

Throughout the study, we collected a set of qualitative and quantitative measures. We captured the start and end time for each phase, as well as time browsed away from the study window, which allowed us to assess the average time spent on training and trials, as well as filter out participants who rushed through the study. During the *Trials* phase, we captured all submitted answers, including incorrect ones. During the *Study* phase, we collected time spent on each task, the submitted answer, the confidence in the submitted answer, and the perceived difficulty of the task; the latter two on a 7-point Likert scale. Through the free-response questions, we collected qualitative feedback for each task. The final *Demographics and Feedback* form includes free-response questions where users provided feedback on the training material and on the overall study. We collected rich provenance data of users interactions, including searching for node, dragging a node, (un)selecting a node, clearing selected nodes, sorting operations, and hovering on a node (which shows a tool-tip).

When calculating correctness, we used non-binary rules that map to a 0–1 scale. For example, we gave 0.5 points for an answer that contains the second-largest node, if the task asked for the largest. We provide details on our scoring method for each task in the supplementary material. When calculating time to completion, we subtracted time spent away from a tab. Although this approach does not ensure a participant paid attention in all cases, it does reduce outliers.

Data

As our dataset, we used a Twitter network of interactions during the EuroVis 2019 conference (collected by J. Guerra-Gomez, used with permission). We chose a Twitter network because we expected participants to be familiar with social networks. The network has 75 nodes, 143 edges, an average degree of 3.81, and a density of 0.16. We also created a smaller version of this network, which contains 25 nodes, 56 edges, an average degree of 4.48, and a density of 0.3. Following Ghoniem et al. [14], we define density as $d = \sqrt{e/n^2}$ where e is the number of edges and n the number of nodes. For the nodes, we had the attributes *names*, *# followers*, *# following (friends)*, *# tweets*, *# likes*, *account age in days*, *node type* (person or institution, categorical), and *continent of origin* (categorical). For edges, we used a categorical attribute/edge type (retweeted or mentioned), and a numerical attribute for each type that contained the number of each action (retweet or mentioned). We modified the source data in several ways:

	Task Name	Task Prompt
T01	Node Search on Attr.	Find the North American with the most tweets.
T02	Node Search on Attr. w/ Distractors	Find the European person or institution with the least likes.
T03	Node Search on Top. w/ Multiple Attrs.	Which person has many interactions in this network, several followers, but few tweets and likes in general?
T04	Neighbor Search on Attr.	Find all of Lane's European neighbors.
T05	Neighbor Search on Attr. w/ Distractors.	Find all of giCentre's North American neighbors.
T06	Neighbor Search on Edge Attr.	Who had the most mention interactions with Jeffrey?
T07	Neighbor Overview on Edge Attr.	Does Alex have more mention interactions with North American or European accounts? Who does he have the most mentions interactions with?
T08	Attr. of Common Neighbors.	Among all people who have interacted with both Jeffrey and Robert, who has the most followers?
T09	Edge Attr.	What is the most common form of interaction between Evis19 and Jon? How often has this happened?
T10	Node Attr. Comparison.	Select all of Noeska's neighbors who are people and have more friends than followers.
T11	Node Attr. Comparison (Small).	Select the people who have interacted with Thomas and have more friends than followers.
T12	Cluster and Attr. Estimation	Select all the people who are in a cluster with Alex and estimate their average number of followers.
T13	Attr. along Shortest Path	What is the institution on a shortest path between Lane and Rob? What is its continent of origin?
T14	Attr. along Shortest Path (Small).	What is the institution on a shortest path between Jason and Jon? What is its continent of origin?
T15	Attr. on Multiple Paths	Of the North Americans who are two interactions away from Sereno, who has been on twitter the longest?
T16	Free Explore	Explore the network freely and report on your findings. Is there anything surprising or interesting?

Table 1: List of tasks and instructions as given to participants. Refer to the supplementary material for details.

we simplified names to shorten them and clipped outliers (for example, an extreme number of followers) so that numerical values are easily comparable on a single linear scale. We also manually added account type and continent of origin. Although retweets and mentions are directed on Twitter, we simplified the network by treating them as undirected.

For network size, we chose the largest network that could reasonably be represented as a node-link or adjacency matrix on a standard-size display without zoom or panning, while still rendering multiple attributes. For example, a node-link diagram with on-node encoding can display only a limited number of nodes before occlusion and overlap render the technique inadequate [24]. By choosing a network that reaches these limits, we aim to draw conclusions about networks that are close to the "hardest" case that can be visualized with these techniques. Larger networks either require different approaches or must be filtered first. For tasks where we believed network size could have a technique-dependent effect on task performance, we included an equivalent task on a small network and a large one. For density, we consciously chose a sparse network, because we know from prior work that adjacency matrices outperform node-link diagrams in dense networks for most tasks [14]. If we can show that AM outperforms NL in sparse multivariate networks, we can generalize that they also outperform them in dense networks. We discuss the implications of our choices on generalizability in the limitations section.

Tasks

We created the tasks based on the two recent taxonomies for MVN tasks [24, 19]. Our tasks cover the main topological structures outlined in both taxonomies: single nodes, neighbors, clusters, and paths. The tasks are listed in Table 1. Details about each task are described in the supplement.

Hypothesis

Prior to running the study, we developed a set of hypotheses about how the two visualization approaches would compare for different types of tasks. We present the hypotheses below and later use them to frame and discuss our results.

Distractor Effects Hypothesis: In the AM, accuracy and time will be resilient to the number of distractors (node attributes that are visualized but that are not necessary for the task). Distractors will have a negative effect in terms of performance and time in the NL, since adding many attributes to the NL will make identifying topological structures more difficult, as the nodes get bigger, and the extraneous attributes will make it harder to isolate the attribute necessary for the task. We test the hypothesis with Tasks 1, 2, 4, and 5.

Attribute Sorting Hypothesis: The AM will perform better (accuracy and time) in tasks that benefit from sorting the matrix based on attributes, such as identifying an extreme node according to a numerical attribute (Tasks 1, 2).

Scalable Attributes Hypothesis: The AM is more scalable with respect to node attributes (Task 3). It will lead to faster and more accurate decisions when many attributes (>3) are present, except for tasks on topological structures ill-suited for adjacency matrices, such as paths.

Common Neighbor Hypothesis: The NL will lead to more accurate and faster responses for all tasks that are concerned with common neighbors of two or more nodes. Although there is weak evidence that show the AM to be advantageous for similar tasks [26], we believe that the known benefits of NL for path-finding will be relevant for tasks of this type (Task 8).

Within-Node Comparison Hypothesis: If the task involves comparing attributes of identical scale within individual nodes (Tasks 10, 11), the NL will lead to more accurate and faster responses. We believe this to be due to the layout of the bars, as the bars in the nodes use the same axis, whereas the attributes in the matrix use adjacent columns.

Cluster Hypothesis: The AM will lead to more accurate and faster results for tasks involving clusters (Task 12). We believe that clusters are difficult to spot in the node link diagram, especially since node size is fairly large to accommodate attributes.

Path Hypothesis: As we know from previous studies, matrices are ill-suited for path-based tasks [14, 18, 26]. Hence, this hypothesis states that the NL will perform more accurately and faster for all tasks related to paths (Tasks 13, 14, 15).

Insight Generation Hypothesis: When freely exploring the network (Task 16) with the AM, participants are more likely to have attribute-based insights. Conversely, the NL will lead to

more topology based insights, such as the presence or absence of connections among nodes.

Pilots, Analysis, and Experiment Planning

We conducted several tests and pilots to evaluate tasks, system usability, data collection modalities, measures, and our procedure. We estimated the number of participants required to uncover effects based on a pilot run on Prolific with 20 participants. We used a power analysis between the two different conditions to estimate the variance in our quantitative measures, which we combined with our observed means to estimate the number of trials required. Due to the limitations of null hypothesis significance testing, we base our analysis on best practices for fair statistical communication in HCI [9] by reporting confidence intervals and effect sizes. We compute 95% bootstrapped confidence intervals [8] and effect sizes using Cohen's d to indicate a standardized difference between two means. For each task, we display the accuracy and time results in the form of a violin plot, which approximates the density distribution of accuracy and time on task for all participants. We superimpose the mean value with a 95% confidence interval error bar to facilitate comparison. Compared to just reporting confidence intervals, violin plots have the advantage of making the distribution of the data salient. Although we include p-values from Mann-Whitney tests (given the non-normal distributions of time and accuracy data) in our figures and highlight Bonferroni-corrected significant results (we consider a corrected threshold of p=0.003), these are only a supplement to the analysis.

RESULTS

We recruited 322 participants for this study. Half the participants were assigned the node-link diagram (NL) and the other half the adjacency matrix (AM). After reviewing all submissions, we excluded the responses of 10 AM and 9 NL participants due to low-effort answers or incomplete submissions. Submissions were classified as low-effort when the participant completed the study in under 10 minutes and had an average accuracy of under 30%. This left us with 151 valid AM and 152 valid NL submissions. Here we present a comparison of task accuracy and time to completion between the two conditions. We group the tasks according to the hypothesis they were intended to investigate. Refer to Table 1 and the supplementary material for task descriptions, configurations, correct answers, and scoring methodology. The study data, results, and the analysis scripts are available at https://github.com/visdesignlab/mvnv-study-analysis. We also include links to each condition in the figure captions.

Distractor Hypothesis

To evaluate our hypothesis that encoding non-task-essential attributes, or *distractors*, would hinder performance in the NL but not in the AM representation, participants were given two pairs of tasks, one with distractors and one without. The first of these task pairs targets single nodes.The second of these task pairs targets node neighbors.

The accuracy and time for all four tasks are shown in Figure 3. For the no-distractor task T1, there were no significant differences in task accuracy between NL and AM, with both conditions showing high accuracy. We found a significant but small difference in time, where AM leads to faster response times. The addition of distractors in T2 led to a strong decrease in accuracy in NL ($M = 0.59$ $[0.51, 0.65]$), but to only a small degree in the AM condition ($M = 0.92$ $[0.87, 0.96]$), resulting in a notable difference in accuracy between the two conditions. Likewise, there is a strong and significant difference in time to completion ($M = 2.23$ $[0.71, 0.83]$ for NL, $M = 0.85$ $[0.79, 0.93]$ for AM). These data confirm the distractor hypothesis for single node targets. The distractor effect does not hold for neighborhood tasks. The no-distractor condition T4 shows slightly better accuracy with the NL than the AM condition. When distractors are added in T5, both conditions decreased in accuracy, but a significant difference appears favoring the NL condition (NL $M = 0.95$ $[0.89, 0.97]$, AM $M = 0.82$ $[0.75, 0.87]$). NL leads to faster responses.

From these results, we can conclude that NL and AM are roughly equivalent for global search tasks on the nodes with a single attribute that is encoded as the node size (NL) or a sorted bar chart (AM). However, if distractors are present, the NL condition has to accommodate the larger number of attributes with an encoding that is less amenable to global search — small nested bars — and strong differences appear in accuracy and time. In neighborhood tasks, where users first identify the node of interest through the search feature, this effect inverts, although performance in AM does not deteriorate quite as much as does that in NL in global search. We speculate that the reason for this is that neighborhood highlighting is more efficient in NL, and that this effect drowns out any benefits the attribute representation in the AM might have.

Attribute Sorting Hypothesis

Tasks 1 and 2 were also used to investigate the hypothesis that tasks that relied on global maximum or minimum attribute values would benefit from the sorting inherent to the table in the AM configuration. The results (shown in Figure 3a) indicate that when comparing bubbles in NL to sorted tables in the AM, no significant differences appear in task accuracy, yet a significant difference with a medium effect size ($d = 0.63$ $[0.38, 0.87]$) appears in time to completion. When comparing complex on-node encoding (nested bars) to sorted tables, as in T2, the effects are significantly and strongly in favor of sorted tables (accuracy: $d = -0.92$ $[-1.16, -0.65]$, time: $d = 1.46$ $[1, 26, 1.66]$). Hence, we can conclude that the attribute sorting hypothesis is correct.

Scalable Attribute Hypothesis

Task 3 is an advanced task designed to test the hypothesis that tasks that relied on several attributes and topology would perform better in AM condition. We speculated that the table would support easy comparison of multiple attributes across nodes, and that sorting on a single attribute could help. However, the results show that the NL condition was significantly more accurate than the AM condition, although both were of fairly low accuracy. The perceived difficulty was high, reported on average at about 5 in both conditions. Both conditions took approximately the same amount of time. Overall, this hypothesis is not supported. In retrospect, we believe that since this task required integrating information from many different attributes and topological features, participants had to resort to a serial search, i.e., scanning all nodes, to answer

it. In this case, the bars and sorting capability of the matrix did not offer a benefit.

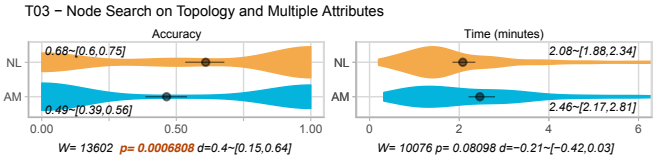


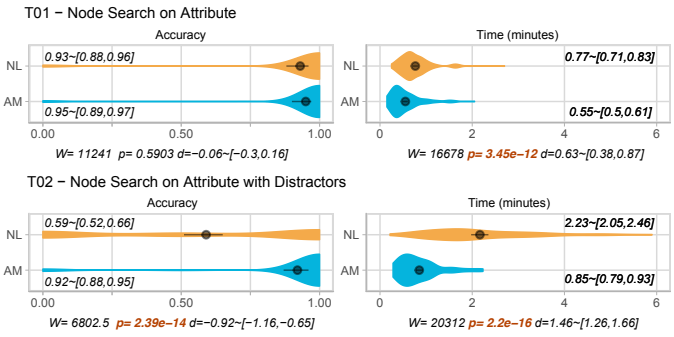
Figure 4: Results for Scalable Attribute Hypothesis. NL, AM.

Common Neighbor Hypothesis

Task 8 asks participants to select the common neighbor of two nodes with the most followers. User performance on this task (see supplement) does not reveal significant differences in accuracy or in time between NL and AM, and hence the hypothesis does not hold. Overall, this outcome was surprising to us. We expected that the known benefits of path finding for node-link diagrams would extend to common neighbor tasks. Although the correct node was not directly in between the nodes in the NL condition, NL layouts in general cannot guarantee this. Also note that our scoring gave 0.5 points to the common neighbor with the second highest number of followers. The violin plot indicates that very few participants selected the second-best answer in the AM condition, whereas this was a common response in the NL condition.

Within-Node Comparison Hypothesis

T10 and T11 test the hypothesis that the node-link diagram is better suited to comparing attributes of the same scale within a given node, on large and small networks respectively. Figure 5 shows a small but significant accuracy advantage of the NL over the AM for the large network. The smaller network (see supplement) task shows similar accuracies for both conditions. However, the NL condition resulted in much faster responses in both the large (NL $M = 0.87$ $[0.81, 0.96]$ vs AM $M = 1.79$ $[1.66, 1.93]$) and the small networks (NL $M = 0.71$ $[0.65, 0.83]$ vs AM $M = 1.25$ $[1.14, 1.41]$). We conclude that this hypothesis is supported, with the caveat that the task we used involved identifying neighbors of a target node. The results of Task 4 indicate that NL tends to have advantages when neighborhoods are involved. One possible reason is that a task on identifying a node globally would benefit from sorting strongly; i.e., we believe that the sorting hypothesis supersedes the node comparison hypothesis.



(a) Node search tasks without (T1) and with (T2) distractors.

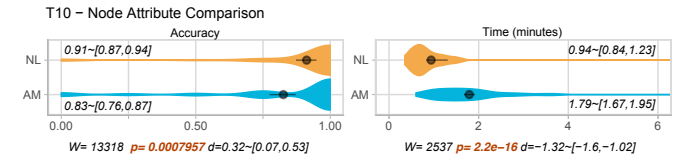


Figure 5: Results for Within-Node Hypoth. NL, AM

Cluster Hypothesis

To investigate our hypothesis that the AM would perform better on cluster tasks, T12 asks users to identify the nodes in a cluster, and then to average the number of followers in that cluster. Results (Figure 6) show significantly higher accuracies with a medium effect size for the AM condition than for the NL, and comparable time. Of note are the overall very low accuracies and the long time spent on task for both conditions. To investigate this result further, we also computed user accuracy for selecting the cluster separately from that for estimating the average attribute value. The results (see supplementary figures) indicate that users were equally able to estimate the average value of an attribute in the NL and the AM, but were better at selecting the cluster structure in the AM ($M = 0.2$ $[0.16, 0.26]$) than in the NL ($M = 0.04$ $[0.02, 0.07]$). Overall, we conclude that the hypothesis is supported.

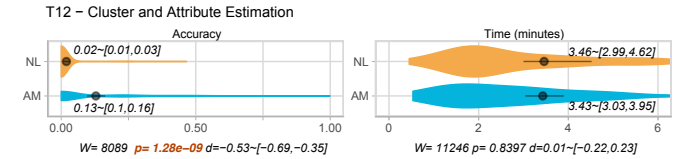
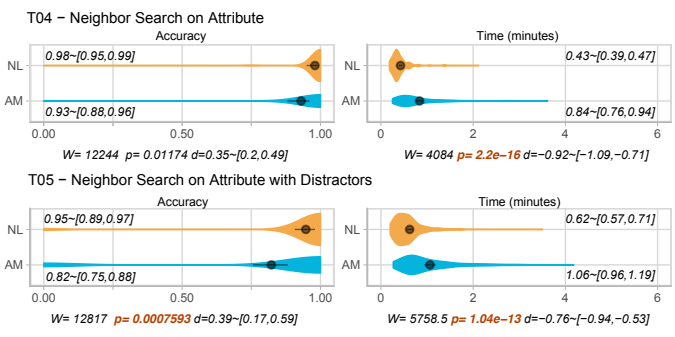


Figure 6: Results for Cluster Hypoth. NL, AM.

Path Hypothesis

Our path hypothesis postulated that the NL would outperform the AM for all path-related tasks. T13 and T14 test a path task on a large network and a small one, respectively. The results, shown in Figure 7 (see supplement for T14), confirm our hypothesis with the NL resulting in significantly higher accuracy for T13 and for T14, although the effect size is less pronounced in the small network. T15 is a more challenging path task and again shows a significantly higher accuracy with a large effect size in the NL vs. the AM. Participants also were more than a minute faster on average in the NL.



(b) Neighborhood search tasks without (T4) and with (T5) distractors.

Figure 3: Task results for the Distractor and Sorting Hypothesis. Stimuli: T1 NL, AM, T2 NL, AM, T4 NL, AM, T5 NL, AM.

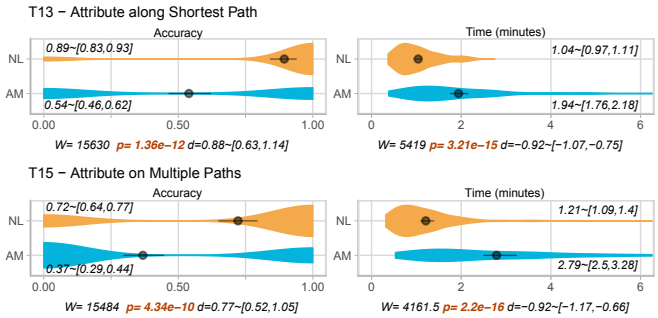


Figure 7: Results for Path Hyp. T13 NL, AM, T15 NL, AM.

Edge Attributes

Tasks 6, 7, and 9 investigate the performance of each condition on tasks that relied on edge attributes. T6 and T9 require the user to select and inspect a single neighbor based on edge attributes. T7 is an overview task, requiring the user to summarize edge attributes for all edges incident to a node. All three tasks show no significant difference in accuracy or time to completion between the two conditions, possibly as a result of the visualizations' ability to highlight neighbors, thereby significantly reducing the search space for completing these tasks. Results plots for these tasks can be found in the supplement.

Insight Generation Hypothesis

Task 16 instructed participants to freely explore the network and report on any insights they derived from their exploration. In order to analyze the responses, we performed a qualitative coding of a sample of 120 responses (60 of each condition), categorizing insights into a set of codes that were derived by an initial open coding of the data. A full list of codes and their frequencies is provided in the supplement.

Two types of insight were markedly more common in the **adjacency matrix** condition: overview and ranked attribute insights. Overview insights were those that derived from an assessment of the entire network, such as *"Institutions have much fewer tweets in general than a person's account."* Ranked attribute insights were those that identified maximum or minimum values for one or more attributes, such as *"MViews has the 2nd youngest account age; however it has the 4th largest follower count even though it also has the fewest tweets."* Across-node attribute comparisons were also more common in the AM, including insights such as *"Nodes with the most followers actually have much fewer tweets than those with far fewer followers"*.

Conversely, the **node-link diagram** favored topology-only, topology-attribute, and within-node-attribute comparison insights. Topology-only insights do not mention any attributes and included *"Steven, Evan, Jo, and Till have only ever had 1 interaction and they have all been with Lane."* Topology-attribute insights were more common in NL and refer to those that comment on both the structure of the network and one or more attributes. Observed examples include *"It does seem a bit odd that Jeffrey, Alex, and Rob have such large networks with their lower than average tweeting"*. Within-node attribute insights are those in which the comment relies on comparison of attributes for a given node, such as *"Jeffrey hasn't made many tweets (less than a thousand) yet has a lot of followers and a fairly long account age."*

We were positively surprised by the extent and the quality of the insights, the engagement of the participants, and the ability of both techniques to reveal insights of various types, albeit with different frequency. We saw the biggest differences in overview-attribute insights (much more in AM) and ranked-attribute insight (exclusively in AM). Consequently, we consider this hypothesis to be confirmed.

DISCUSSION

Overall, our results show that AMs are best suited for tasks on clusters, tasks that benefit from sorted attributes, and tasks that are performed in the presence of distractors and require scanning the entire network. NLs, on the other hand, are well suited for paths, tasks that rely on within-node comparisons, and tasks on neighbors given the ability of highlighting neighbors in the network. The disadvantage of the AM for performing path tasks is well known [26, 14, 19] and was once again confirmed in our study. Our results indicate that the suitability of AMs for cluster tasks, previously described by Okoe et al. [26], also holds for cluster tasks involving attributes. The cluster task in the present study incorporates attributes both in the cluster selection (only select nodes of type person) and in the attribute estimation component.

We could not find differences between the AM and NL in the ability to analyze edge attribute in three different tasks, which is in contrast to the results of Alper et al. [3], where AM performed better than NL. Our NL edge design was different in that we used curved, separated edges instead of straight, bundled edges. The discrepancy in our findings may be due to the differences in study design, our different NL design, or the dataset and types of tasks used in both studies. All tasks used in the Alper et al. study involved comparing parallel edges between pairs of nodes, so the linear juxtaposition of edges in their design versus the curved edges in the present study could conceivably affect performance. **Based on these results, we conclude that adjacency matrices are at least as good for edge attribute tasks as node-link diagrams, even for the very sparse graphs we tested.**

Okoe et al. [26] found weak evidence for AM to perform better in common neighbor tasks. Our findings do not support this, if the tasks also consider attributes. Both the NL and the AM performed about equally well. User insights from the free-explore task corroborate the results from other tasks. For example, they show that AMs are particularly well suited for tasks on extreme values for one or more attributes, which is also supported by Task 1. Generally, insights related to attributes were much more common in the adjacency matrix, and insights related to topology were slightly more common in the node-link diagram. This finding highlights the respective strengths and weaknesses of the techniques. We discuss implications for designing network visualizations in the supplement.

METHODOLOGY CONSIDERATIONS

Creating an evaluation methodology for comparing complex interactive systems with non-expert users resulted in a unique set of challenges and considerations that we outline and reflect on. First, we had to ensure participants were properly trained on the interactive techniques. We conducted two rounds of pilots to test and improve our training materials. Our first pilot

revealed that the passive training videos plus trials were insufficient to recall all available interactions, particularly for the adjacency matrix. As a result, we implemented active training, which required users to perform all available interactions correctly before starting the study. Results from the second pilot were much improved, as indicated by both user performance and participant feedback.

Another challenge is the long study duration when compared to typical crowdsourcing tasks. We found that by offering above-average hourly compensation and clearly specifying the duration of the study, participants were willing to invest the longer time to complete the study. We recruited all 322 participants who completed the study in under two hours. 423 people started the study but returned it before completion, usually within a few minutes. Feedback by participants indicates that our study was more involved than others on Prolific.

We logged all interactions with the system in order to track engagement and collect a rich dataset on how users leverage the features in each visualization to perform the tasks. We mostly used this data to verify the quality of each trial and to identify and reason about outliers. For example, we found that exceptionally long task completion time often correlates with participants browsing away from the tab showing the experiment. An initial analysis of this data reveals visualization-specific interaction patterns: For example, participants highlight neighbors more often in the NL, suggesting that neighbors are easier to identify in AM without highlighting. We leave a detailed analysis of this data to future work.

LIMITATIONS

Our comparison of two complex systems makes it impossible to identify individual factors that contribute to the performance of the techniques. For example, we do not know whether it is interactive sorting or the encoding in aligned bars that leads to attribute-related insights in the AM.

One limitation of our study is that we were not able to compare multiple different network types in terms of size and density. Instead, we focused on testing a broad set of tasks on networks with varying attribute configurations. Although we cannot make claims as to whether our results would generalize to networks with significantly different topological characteristics, we based our decision to test two size variants of a single network on prior studies that investigated the effect of network size on task accuracy. Ghoniem et al.'s [14] results indicate that for networks under 100 nodes, there was no effect between size and user accuracy for all tasks except for overview tasks such as estimating the number of nodes or links in a graph, where node-link diagrams performed better for small graphs. For density, Ghoniem et al. [14] found that AM outperforms NL for dense graphs in all tasks with the exception of path-finding, where NL representations were always better. By choosing a sparse network for this study, we were able to attribute performance differences between AM and NL to the controlled network attribute variations, and not to the inherent advantage of the AM over NL for dense networks.

REFLECTIONS

We believe that our study is unique in terms of the complexity of the techniques and the tasks that we evaluate. Reflecting on our approach, we ask whether such a study that cannot adequately separate individual factors is worth the effort. Aren't more qualitative approaches, like case studies, a better approach for these cases? We believe that a quantitative approach is appropriate because it can give definitive answers to questions about the relative merits of two widely used interactive visualization techniques, and that it is not feasible to isolate all factors and test them individually. We believe that our results are impervious to subtle changes in the visualization or interaction design. For example, we believe that if the embedded bar charts in the NL conditions were replaced with an alternative, equivalently powerful design, our results would still hold.

Second, should this method be the new gold standard for evaluating a novel technique? Absolutely not. It is important to use this approach only for techniques that have passed the test of time, and for which the design process is free of biases. We believe, however, that a broad class of published visualization techniques is amenable to our approach.

Reflecting on our validation process, we believe that the expert interviews helped us to design better techniques, yet we were frustrated by the low return rate and by our inability to have a discussion with the experts. We were largely unsuccessful framing our questions so that they were answered in the context of that technique. We believe that a dialog, potentially in a series of structured interviews, would have been more fruitful.

Regarding the use of crowdsourcing platforms for evaluating complex techniques, we found that users achieved remarkably high average accuracy (75.2% across both conditions). We attribute this (1) to our extensive training and (2) to our above average compensation, which we speculate motivated participants to pay attention and be willing to invest time. We have evidence for the former in the form of qualitative feedback on the study, where participants complimented the training. The latter is demonstrated by our high rates of useful comments on the open-explore tasks (80% of all answers contained insights). Given this data, we believe that the view of the visualization community on which kinds of studies can be run on a crowdsourcing platform might be too narrow, and that a broader range of systems is amenable to crowdsourced experiments.

CONCLUSION

Multivariate networks are a common and important data type. We present the first quantitative study of two complex, validated visualization techniques tailored to multivariate networks. Our results reveal that the AM outperforms the NL for tasks on clusters as well as configurations where non-task essential attributes are encoded. NL are bested suited for paths and tasks that rely on within-node comparisons. We also reflect on our approach of using advanced visualization designs in quantitative studies. We highlight the importance of including active training, as well as above-average compensation to ensure motivated users for a longer-than-average study.

ACKNOWLEDGEMENTS

We thank John Guerra-Gomez for providing the data and the experts who gave feedback on our visualization designs. This work was funded by NSF 1835904 and 1815587.

REFERENCES

[1] Ala Abuthawabeh, Fabian Beck, Dirk Zeckzer, and Stephan Diehl. 2013. Finding Structures in Multi-Type Code Couplings with Node-Link and Matrix Visualizations. In *Working Conference on Software Visualization (VISSOFT)*. IEEE, 1–10. DOI: http://dx.doi.org/10.1109/VISSOFT.2013.6650530

[2] Diane Lindwarm Alonso, Anne Rose, Catherine Plaisant, and Kent L. Norman. 1998. Viewing Personal History Records: A Comparison of Tabular Format and Graphical Presentation Using LifeLines. *Behaviour & Information Technology* 17, 5 (1998), 249–262. DOI: http://dx.doi.org/10.1080/014492998119328

[3] Basak Alper, Benjamin Bach, Nathalie Henry Riche, Tobias Isenberg, and Jean-Daniel Fekete. 2013. Weighted Graph Comparison Techniques for Brain Connectivity Analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 483–492. DOI: http://dx.doi.org/10.1145/2470654.2470724

[4] Juhee Bae and Benjamin Watson. 2011. Developing and Evaluating Quilts for the Depiction of Large Layered Graphs. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (2011), 2268–2275. DOI: http://dx.doi.org/10.1109/TVCG.2011.187

[5] Philip Berger, Heidrun Schumann, and Christian Tominski. 2019. Visually Exploring Relations Between Structure and Attributes in Multivariate Graphs. In *Information Visualisation (IV)*. IEEE, 261–268. DOI: http://dx.doi.org/10.1109/IV.2019.00051

[6] Sheelagh Carpendale. 2008. Evaluating Information Visualizations. In *Information Visualization: Human-Centered Issues and Perspectives*, John T. Stasko, Jean-Daniel Fekete, Chris North, and Chris North (Eds.). Springer, 19–45. DOI: http://dx.doi.org/10.1007/978-3-540-70956-5_2

[7] Chunlei Chang, Benjamin Bach, Tim Dwyer, and Kim Marriott. 2017. Evaluating Perceptually Complementary Views for Network Exploration Tasks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1397–1407. DOI: http://dx.doi.org/10.1145/3025453.3026024

[8] Geoff Cumming. 2013. *Understanding the New Statistics: Effect Sizes, Confidence Intervals, and Meta-Analysis*. Routledge.

[9] Pierre Dragicevic. 2016. Fair Statistical Communication in HCI. In *Modern Statistical Methods for HCI*. Springer, 291–330. DOI: http://dx.doi.org/10.1007/978-3-319-26633-6_13

[10] Jean-Daniel Fekete. 2015. *Reorder.Js: A JavaScript Library to Reorder Tables and Networks*. Technical Report.

[11] Mia Feng, Cheng Deng, Evan M. Peck, and Lane Harrison. 2018. The Effects of Adding Search Functionality to Interactive Visualizations on the Web. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 1–13. DOI: http://dx.doi.org/10.1145/3173574.3173711

[12] Katharina Furmanova, Samuel Gratzl, Holger Stitz, Thomas Zichner, Mirsolava Jaresova, Martin Ennemoser, Alexander Lex, and Marc Streit. 2019. Taggle: Scalable Visualization of Tabular Data through Aggregation. *Information Visualization* (2019). DOI: http://dx.doi.org/10.1177/1473871619878085

[13] Ujwal Gadiraju, Sebastian Möller, Martin Nöllenburg, Dietmar Saupe, Sebastian Egger-Lampl, Daniel Archambault, and Brian Fisher. 2017. Crowdsourcing Versus the Laboratory: Towards Human-Centered Experiments Using the Crowd. In *Evaluation in the Crowd. Crowdsourcing and Human-Centered Experiments*. Vol. 10264. Springer, 6–26. DOI: http://dx.doi.org/10.1007/978-3-319-66435-4_2

[14] Mohammad Ghoniem, Jean-Daniel Fekete, and Philippe Castagliola. 2005. On the Readability of Graphs Using Node-Link and Matrix-Based Representations: A Controlled Experiment and Statistical Analysis. *Information Visualization* 4, 2 (2005), 114–135. DOI: http://dx.doi.org/10.1057/palgrave.ivs.9500092

[15] Jeffrey Heer and Michael Bostock. 2010. Crowdsourcing Graphical Perception: Using Mechanical Turk to Assess Visualization Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 203–212. DOI:http://dx.doi.org/10.1145/1753326.1753357

[16] Jeffrey Heer and George Robertson. 2007. Animated Transitions in Statistical Data Graphics. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)* 13, 6 (2007), 1240–1247. DOI: http://dx.doi.org/10.1109/TVCG.2007.70539

[17] Weidong Huang, Peter Eades, Seok-Hee Hong, and Henry Been-Lirn Duh. 2016. Effects of Curves on Graph Perception. In *Pacific Visualization Symposium (PacificVis)*. IEEE, 199–203. DOI: http://dx.doi.org/10.1109/PACIFICVIS.2016.7465270

[18] René Keller, Claudia M. Eckert, and P. John Clarkson. 2006. Matrices or Node-Link Diagrams: Which Visual Representation Is Better for Visualising Connectivity Models? *Information Visualization* 5, 1 (2006), 62–76. DOI:http://dx.doi.org/10.1057/palgrave.ivs.9500116

[19] Andreas Kerren, Helen C. Purchase, and Matthew Ward (Eds.). 2014. *Multivariate Network Visualization*. Number 8380 in Lecture Notes in Computer Science. Springer.

[20] Ethan Kerzner, Alexander Lex, Crystal Lynn Sigulinsky, Timothy Urness, Brian William Jones, Robert E. Marc, and Miriah Meyer. 2017. Graffinity: Visualizing Connectivity in Large Graphs. *Computer Graphics Forum (EuroVis)* 36, 3 (2017), 251–260. DOI: http://dx.doi.org/10.1111/cgf.13184

[21] Robert Kosara. 2016. An Empire Built On Sand: Reexamining What We Think We Know About Visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization (BELIV '16)*. ACM, 162–168. DOI: http://dx.doi.org/10.1145/2993901.2993909

[22] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. 2012. Empirical Studies in Information Visualization: Seven Scenarios. *IEEE Transactions on Visualization and Computer Graphics* 18, 9 (2012), 1520–1536. DOI: http://dx.doi.org/10.1109/TVCG.2011.279

[23] Tamara Munzner. 2014. *Visualization Analysis and Design*. CRC Press, Taylor & Francis Group.

[24] Carolina Nobre, Miriah Meyer, Marc Streit, and Alexander Lex. 2019. The State of the Art in Visualizing Multivariate Networks. *Computer Graphics Forum (EuroVis)* 38, 3 (2019), 807–832. DOI: http://dx.doi.org/10.1111/cgf.13728

[25] Mershack Okoe, Radu Jianu, Stephen Kobourov, Radu Jianu, and Stephen Kobourov. 2018b. Revisited Experimental Comparison of Node-Link and Matrix Representations. In *Graph Drawing and Network Visualization*. Vol. 10692. Springer, 287–302. DOI: http://dx.doi.org/10.1007/978-3-319-73915-1_23

[26] Mershack Okoe, Radu Jianu, and Stephen G. Kobourov. 2018a. Node-Link or Adjacency Matrices: Old Question, New Insights. *IEEE Transactions on Visualization and Computer Graphics* (2018), 2940–2952. DOI: http://dx.doi.org/10.1109/TVCG.2018.2865940

[27] Christian Partl, Alexander Lex, Marc Streit, Denis Kalkofen, Karl Kashofer, and Dieter Schmalstieg. 2013. enRoute: Dynamic Path Extraction from Biological Pathway Maps for Exploring Heterogeneous Experimental Datasets. *BMC Bioinformatics* 14, Suppl 19 (2013), S3. DOI: http://dx.doi.org/10.1186/1471-2105-14-S19-S3

[28] Catherine Plaisant, Jean-Daniel Fekete, and Georges Grinstein. 2008. Promoting Insight-Based Evaluation of Visualizations: From Contest to Benchmark Repository. *IEEE Transactions on Visualization and Computer Graphics* 14, 1 (2008), 120–134. DOI: http://dx.doi.org/10.1109/TVCG.2007.70412

[29] Ramana Rao and Stuart K. Card. 1994. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus + Context Visualization for Tabular Information. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*. ACM, 318–322. DOI: http://dx.doi.org/10.1145/191666.191776

[30] Donghao Ren, Laura R. Marusich, John O'Donovan, Jonathan Z. Bakdash, James A. Schaffer, Daniel N. Cassenti, Sue E. Kase, Heather E. Roy, Wan-yi (Sabrina) Lin, and Tobias Höllerer. 2019. Understanding Node-Link and Matrix Visualizations of Networks: A Large-Scale Online Experiment. *Network Science* 7, 2 (2019), 242–264. DOI: http://dx.doi.org/10.1017/nws.2019.6

[31] Ilkin Safarli and Alexander Lex. 2019. TaMax: Visualizing Dense Multivariate Networks with Adjacency Matrices. In *Proceedings of the IEEE Information Visualization Conference (InfoVis Posters)*.

[32] Purvi Saraiya, Chris North, and Karen Duca. 2005. An Insight-Based Methodology for Evaluating Bioinformatics Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 11, 4 (2005), 443–456.

[33] Sebastian Schöffel, Johannes Schwank, Jan Stärz, and Achim Ebert. 2016. Multivariate Networks: A Novel Edge Visualization Approach for Graph-Based Visual Analysis Tasks. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '16)*. ACM, 2292–2298. DOI:http://dx.doi.org/10.1145/2851581.2892451

[34] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. 2012. Design Study Methodology: Reflections from the Trenches and the Stacks. *IEEE Transactions on Visualization and Computer Graphics (InfoVis)* 18, 12 (2012), 2431 –2440. DOI: http://dx.doi.org/10.1109/TVCG.2012.213

[35] Melanie Tory, David Sprague, Fuqu Wu, Wing Yan So, and Tamara Munzner. 2007. Spatialization Design: Comparing Points and Landscapes. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (2007), 1262–1269. DOI: http://dx.doi.org/10.1109/TVCG.2007.70596

[36] Emily Wall, Meeshu Agnihotri, Laura Matzen, Kristin Divis, Michael Haass, Alex Endert, and John Stasko. 2019. A Heuristic Approach to Value-Driven Evaluation of Visualizations. *IEEE Transactions on Visualization and Computer Graphics* 25, 1 (2019), 491–500. DOI: http://dx.doi.org/10.1109/TVCG.2018.2865146

[37] Kai Xu, Chris Rooney, Peter Passmore, Dong-Han Ham, and Phong H. Nguyen. 2012. A User Study on Curved Edges in Graph Visualization. *IEEE Transactions on Visualization and Computer Graphics* 18, 12 (2012), 2449–2456. DOI: http://dx.doi.org/10.1109/TVCG.2012.189

[38] Vahan Yoghourdjian, Daniel Archambault, Stephan Diehl, Tim Dwyer, Karsten Klein, Helen C. Purchase, and Hsiang-Yun Wu. 2018. Exploring the Limits of Complexity: A Survey of Empirical Studies on Graph Visualisation. *Visual Informatics* 2, 4 (2018), 264–282. DOI:http://dx.doi.org/10.1016/j.visinf.2018.12.006

Evaluating Multivariate Network Visualization Techniques Using a Validated Design and Crowdsourcing Approach

Supplementary Material

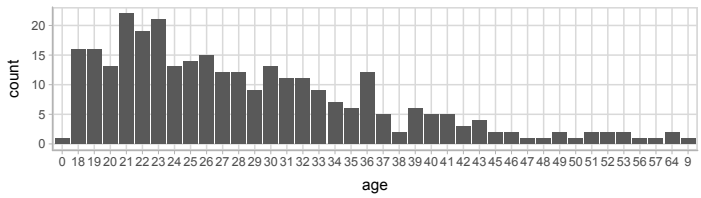
Carolina Nobre[1], Dylan Wootton[1], Lane Harrison[2], Alexander Lex[1]
[1]University of Utah, [2]Worcester Polytechnic Institute

ACM CHI 2020

Contents

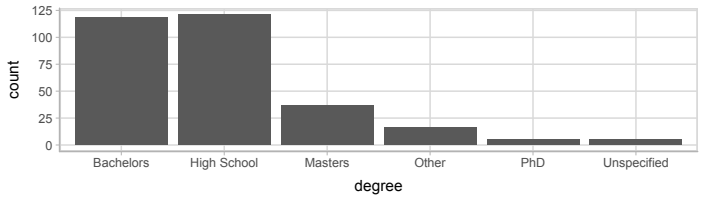
1 Participant Demographics

This section presents the demographic information for our participant pool, including distribution of age, sex, highest degree achieved, browser used, and self-assessed visualization proficiency.

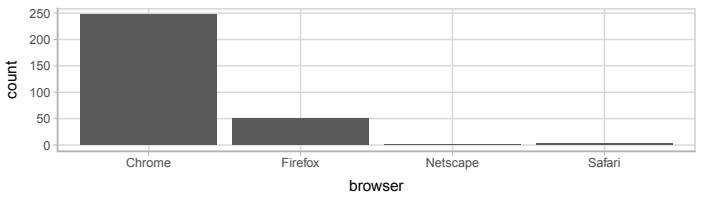


(a) Age Distribution

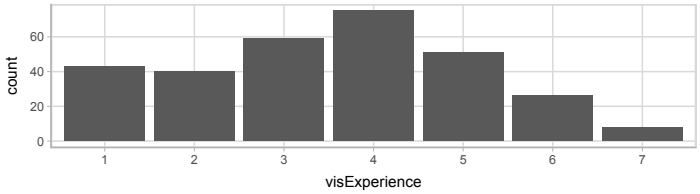
(b) Sex Distribution



(c) Degree Distribution



(d) Browser Distribution



(e) Vis Experience Distribution

Figure 1: Participant Demographics.

2 Tasks

The following section contains, for each task, the task prompt, which hypothesis it was meant to investigate, the visual configuration shown to the user depending on their assigned condition, how we scored for accuracy on that task, and results. We also provide links to the interactive visualizations with task instructions and stimuli.

The result plot caption contains the following statistical parameters: Wilcox Test (W), p-value (p), Cohen's d for Effect Size (d), and Median value + 95% confidence intervals (M).

	Task Name	Task Prompt	Properties	Topology Target	Hypothesis
T01	Node Search on Attribute	Find the North American with the most tweets.	Large 2NA	Single Node	Distractor, Attribute Sorting
T02	Node Search on Attribute with Distractors	Find the European person or institution with the least likes.	Large 6NA	Single Node	Distractor, Attribute Sorting
T03	Node Search on Topology and Multiple Attributes	Which person has many interactions (edges) in this network, several followers, but few tweets and likes in general?	Large 4NA	Single Node	Scalable Attributes
T04	Neighbor Search on Attribute.	Find all of Lane's European neighbors.	Large 1NA	Neighbors	Distractor
T05	Neighbor Search on Attribute with Distractors.	Find all of giCentre's North American neighbors.	Large 6NA	Neighbors	Distractor
T06	Neighbor Search on Edge Attribute.	Who had the most mention interactions with Jeffrey?	Large 2EA	Neighbors	Edge Attributes
T07	Neighbor Overview on Edge Attribute.	Does Alex have more mention interactions with North American or European accounts? Who does he have the most mentions interactions with?	Large 1NA2EA	Neighbors	Edge Attributes
T08	Attribute of Common Neighbors.	Among all people who have interacted with both Jeffrey and Robert, who has the most followers?	Large 1NA	Neighbors	Common Neighbor
T09	Edge Attributes.	What is the most common form of interaction between Evis19 and Jon? How often has this interaction happened?	Large 2EA	Neighbors	Edge Attribute
T10	Node Attribute Comparison.	Select all of Noeska's neighbors that are people and have more friends than followers.	Large 3NA	Neighbors	Within-node Comparison
T11	Node Attr. Comparison on Small Network.	Select the people who have interacted with Thomas and have more friends than followers.	Small 3NA	Neighbors	Within-node Comparison
T12	Cluster and Attribute Estimation	Select all the people who are in a cluster with Alex. Estimate the average number of followers among the selected people.	Large 1NA	Cluster	Cluster
T13	Attribute along Shortest Path	What is the institution on a shortest path between Lane and Rob? What is its continent of origin?	Large 2NA	Path	Path
T14	Attribute along Shortest Path on Small Network.	What is the institution on a shortest path between Jason and Jon? What is its continent of origin?	Small 2NA	Path	Path
T15	Attribute on Multiple Paths	Of the North Americans who are two interactions away from Sereno, who has been on twitter the longest?	Large 1NA	Paths	Path
T16	Free Explore	Please explore the network freely and report on your findings. Is there anything surprising or particularly interesting in the network?	Large 6NA	NA	Insight Generation

Table 1: Summary of tasks, the configrations, the topology target, and the associated hypothesis.

2.1 Task 1: Node Search on Attribute

Instruction: Find the North American with the most tweets.
Properties: Large Network, 2 Node Attributes, Topology Target: Single Node.
Hypothesis: Distractor Effect Hypothesis, Sorting Attribute Hypothesis.
Scoring: Full score for T.J. 0.5 points for the NA with the second most tweets (Arvind).
Links: NL, AM

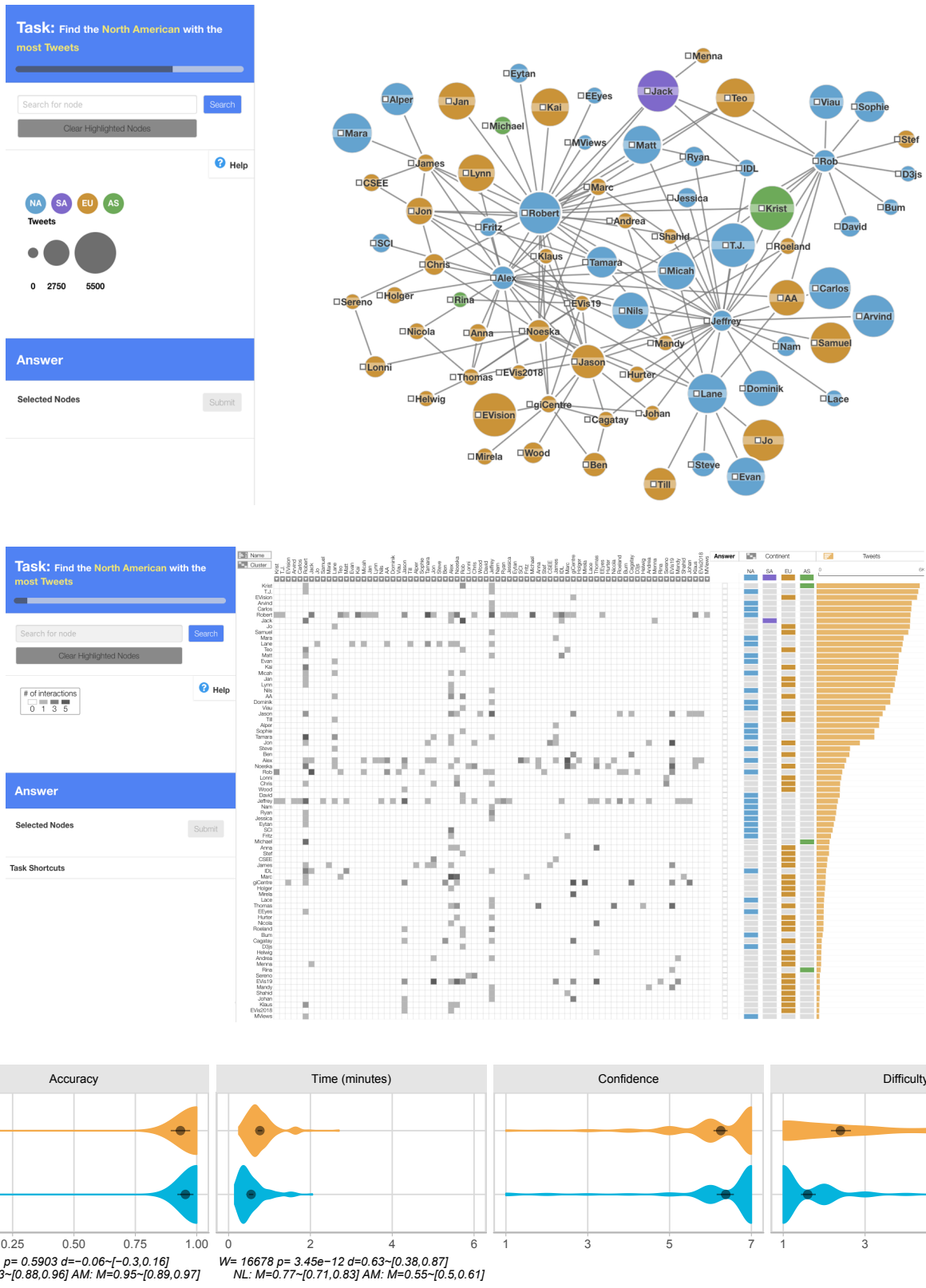


Figure 2: Task 1 configuration and results.

2.2 Task 2: Node Search on Attribute with Distractors

Instruction: Find the European person or institution with the least likes.
Properties: Large Network, 6 Node Attributes, Topology Target: Single Node.
Hypothesis: Distractor Effect Hypothesis, Sorting Attribute Hypothesis.
Scoring: .5 points for the two Europeans with the second least likes (Jason/Evision).
Links: NL, AM

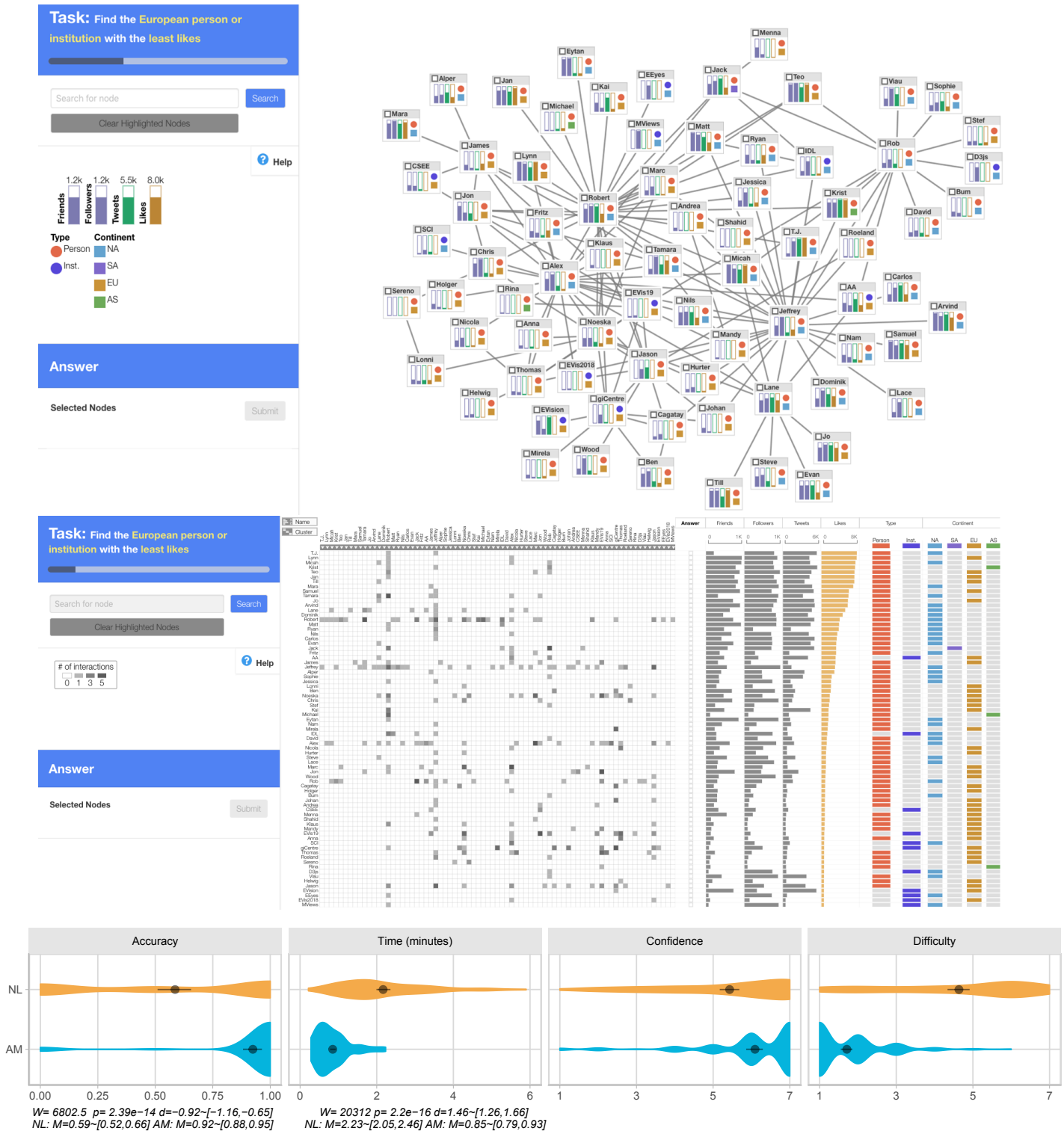


Figure 3: Task 2 configuration and results.

2.3 Task 3: Node Search on Topology and Multiple Attributes

Instruction: Which person has many interactions (edges) in this network, several followers, but few tweets and likes in general?
Properties: Large Network, 4 Node Attributes, Topology Target: Single Node.
Hypothesis: Scalable Attributes.
Scoring: This task didn't ask for a precise answer. We gave 1 point for Jeffrey and Alex, 0.5 points for Noeska and Rob.
Links: NL, AM

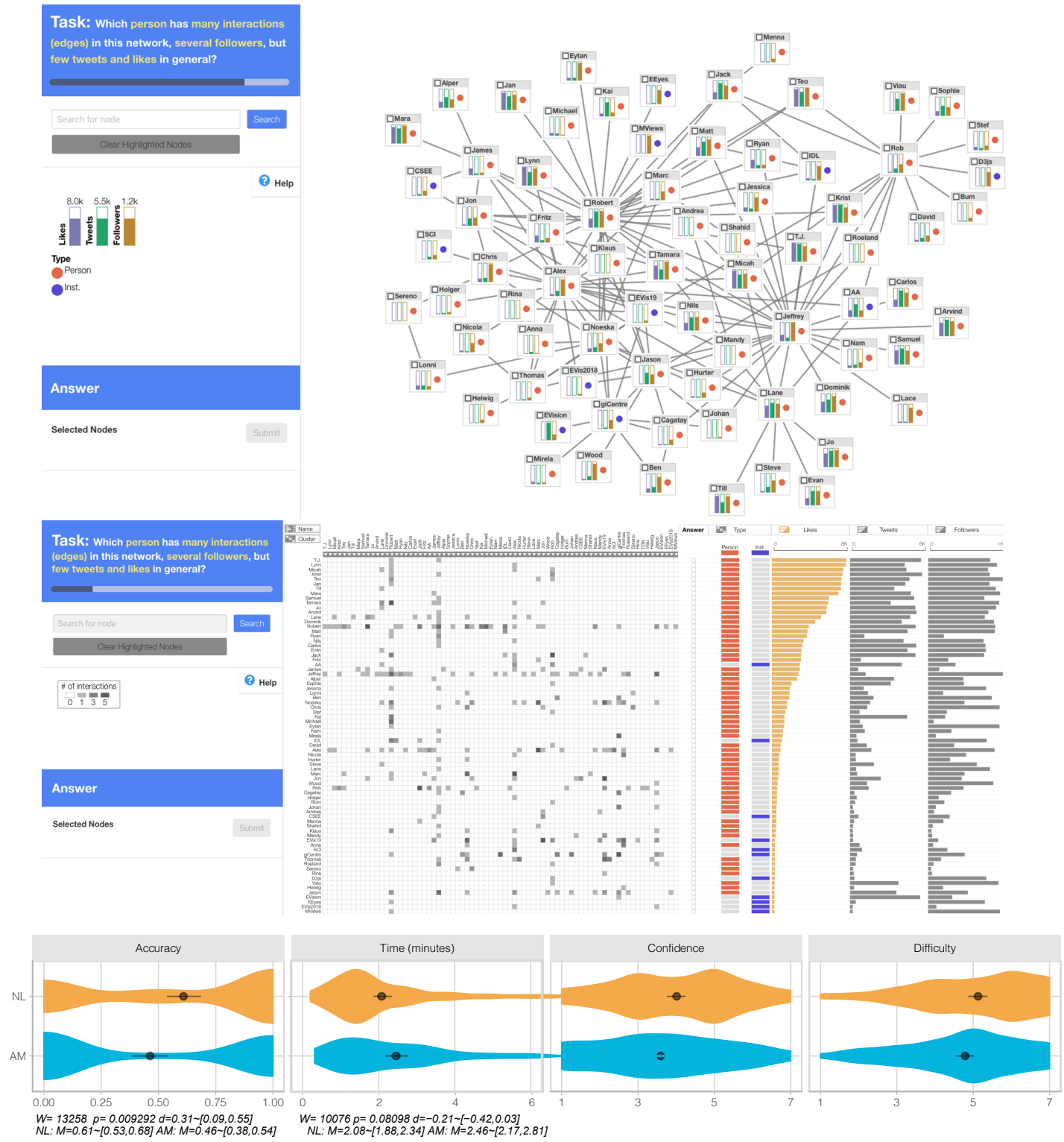


Figure 4: Task 3 configuration and results.

2.4 Task 4: Neighbor Search on Attribute

Instruction: Find all of Lane's European neighbors
Properties: Large Network, 1 Node Attribute, Topology Target: Neighbors.
Hypothesis: Distractor Hypothesis.
Scoring: Correct answer is AA, Noeska, Till, and Joe. 1/4 point for each correct answer. -1/4 point for each incorrect answer.
Links: NL, AM

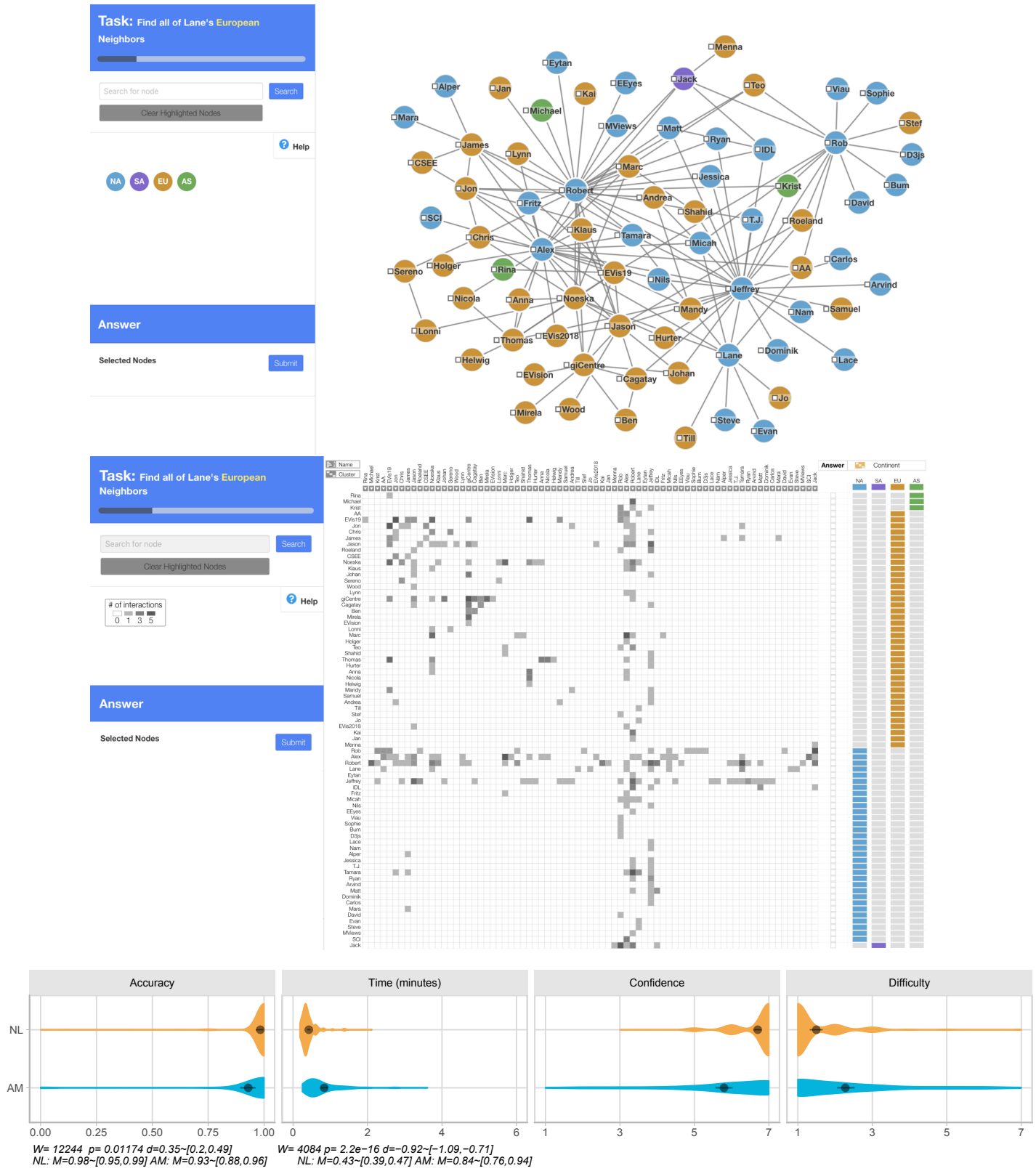


Figure 5: Task 4 configuration and results.

2.5 Task 5: Neighbor Search on Attribute with Distractors

Instruction: Find all of giCentre's North American neighbors
Properties: Large Network, 6 Node Attributes, Topology Target: Neighbors.
Hypothesis: Distractor Hypothesis.
Scoring: Full score for only Robert. 0 otherwise.
Links: NL, AM

Figure 6: Task 5 configuration and results.

2.6 Task 6: Neighbor Search on Edge Attribute.

Instruction: Who had the most mentions interactions with Jeffrey?
Properties: Large Network, 2 Edge Attributes, Topology Target: Neighbors.
Hypothesis: Edge Attributes.
Scoring: Full score for only Robert. 0 otherwise
Links: NL, AM

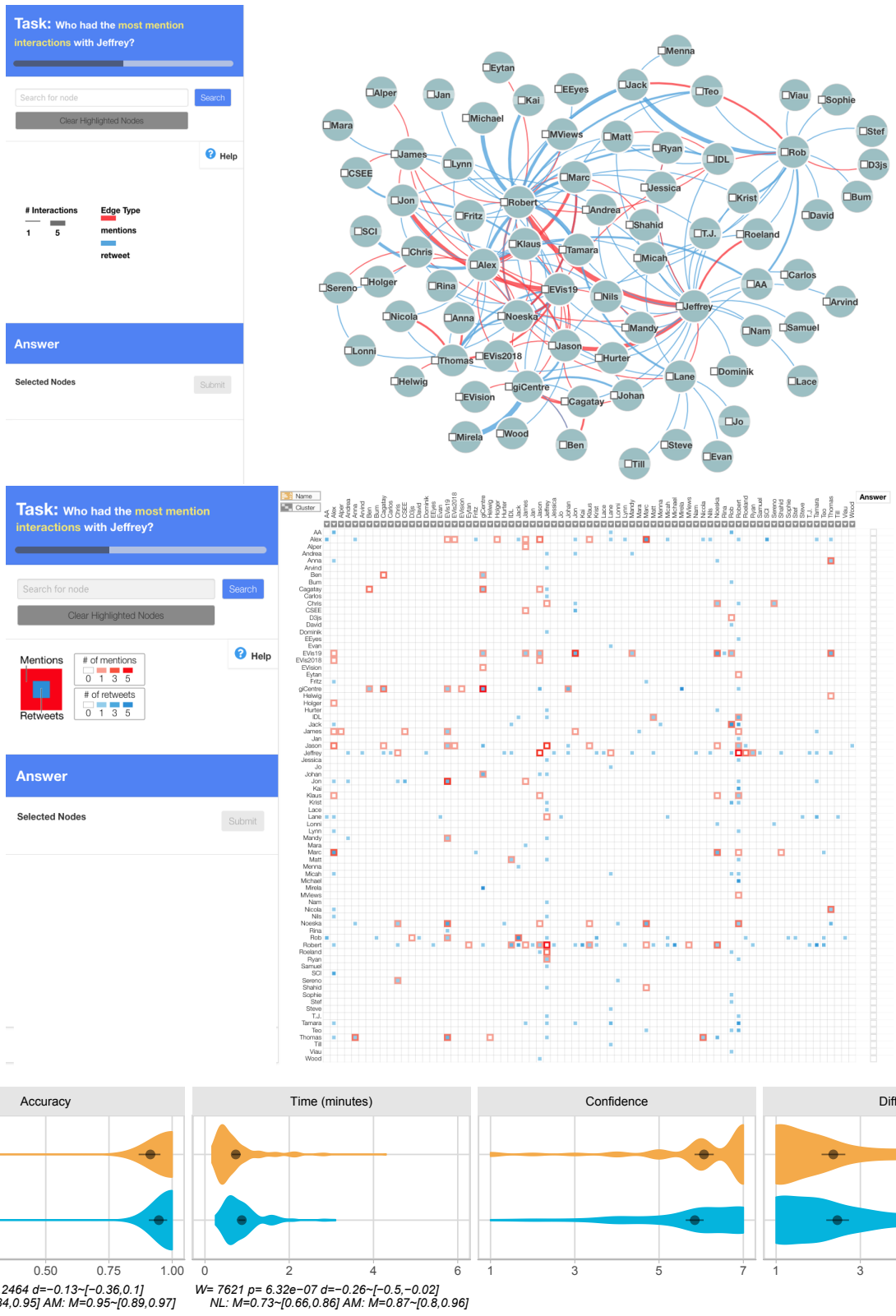


Figure 7: Task 6 configuration and results.

2.7 Task 7: Neighbor Overview on Edge Attribute

Instruction: Does Alex have more mention interactions with North American or European accounts? Who does he have the most mentions interactions with?
Properties: Large Network, 1 Node Attribute, 2 Edge Attributes, Topology Target: Neighbors.
Hypothesis: Edge Attributes
Scoring: European (worth .5 points). Marc (worth .5 points).
Links: NL, AM

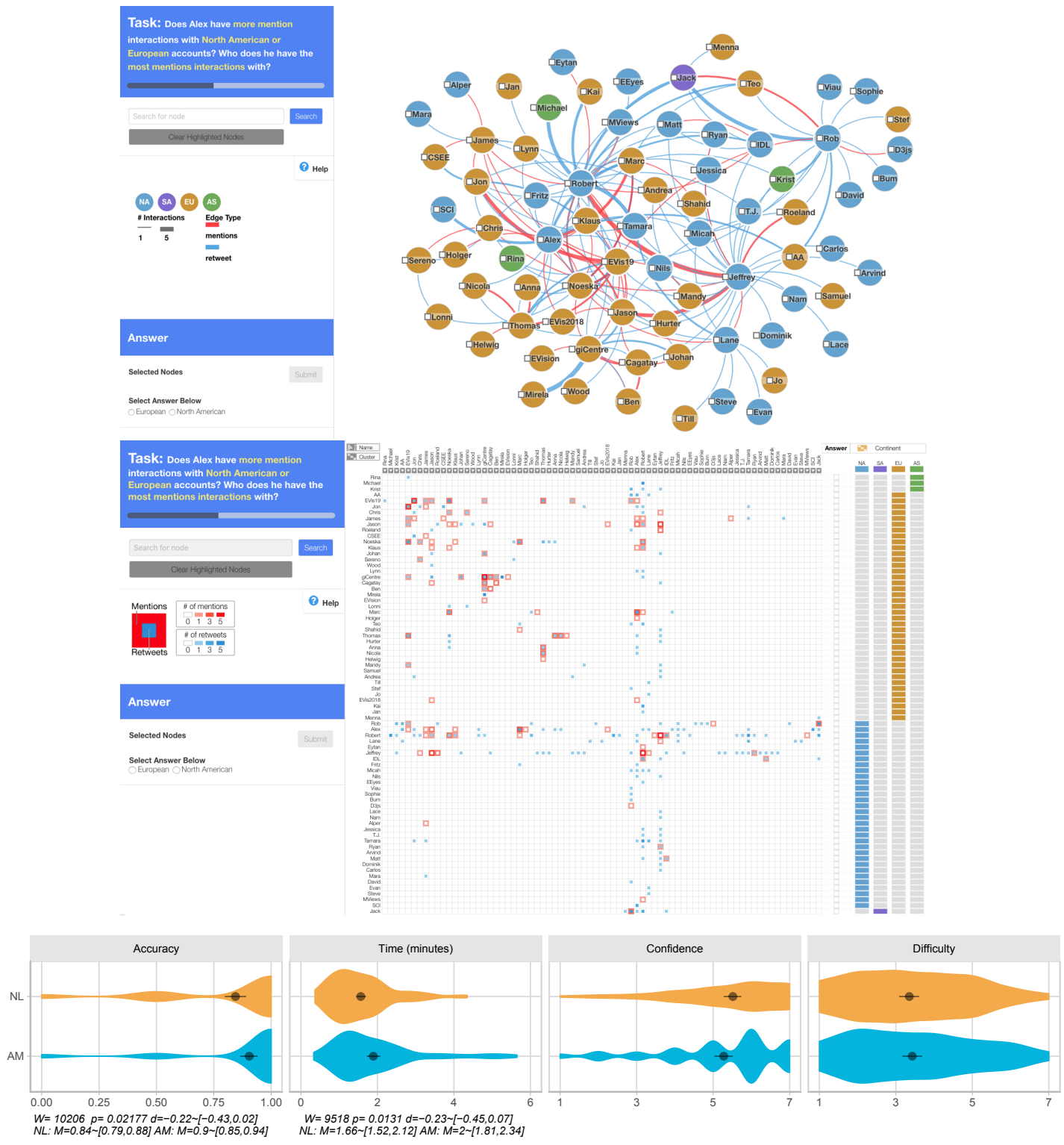


Figure 8: Task 7 configuration and results.

2.8 Task 8: Attribute of Common Neighbors

Instruction: Among all people who have interacted with both Jeffrey and Robert, who has the most followers?
Properties: Large Network, 1 Node Attribute, Topology Target: Neighbors.
Hypothesis: Common Neighbor Hypothesis
Scoring: Full score for Chris, .5 point for Tamara.
Links: NL, AM

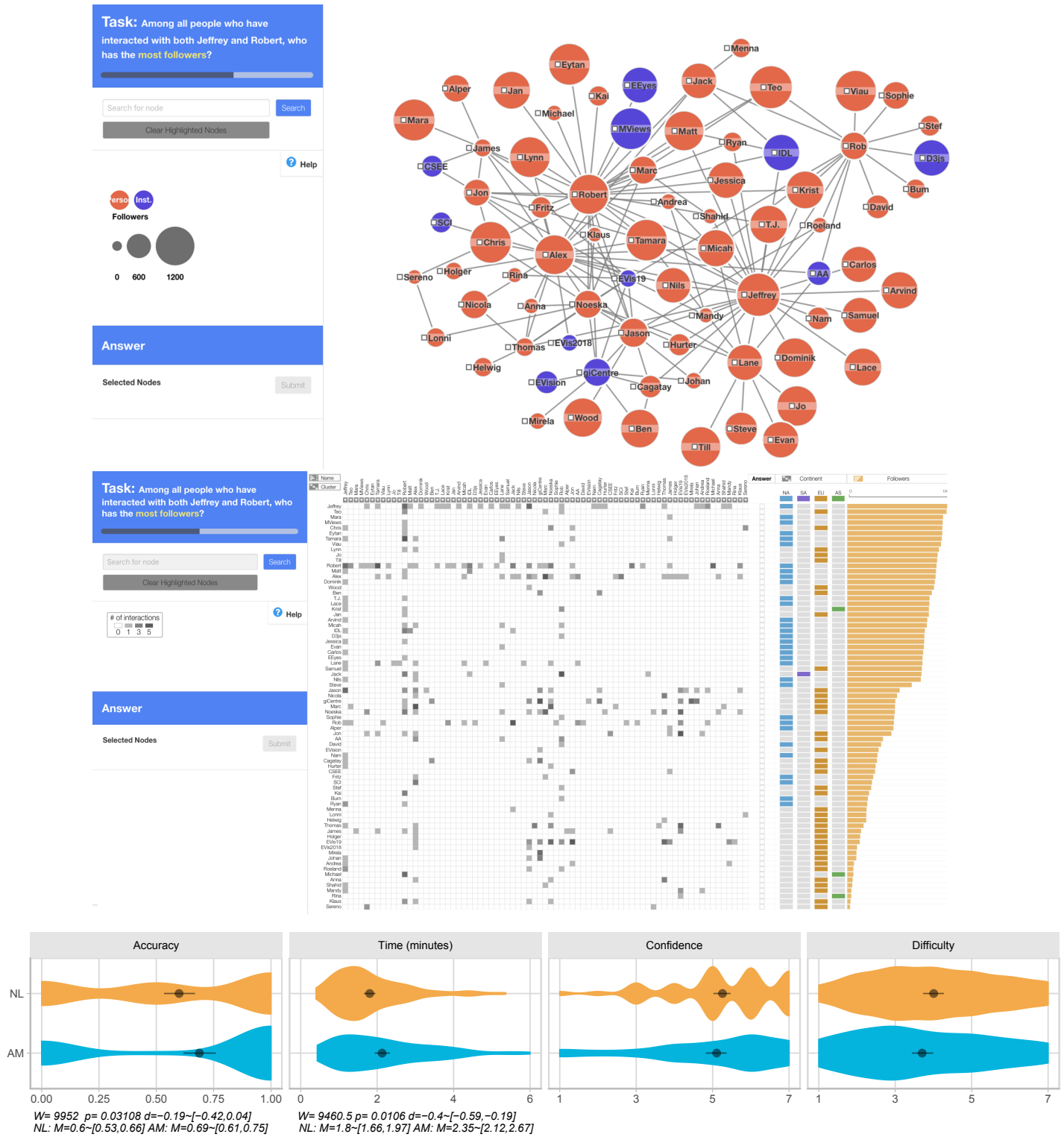


Figure 9: Task 8 configuration and results.

2.9 Task 9: Edge Attributes

Instruction: What is the most common form of interaction between Evis19 and Jon? How often has this interaction happened?
Properties: Large Network, 2 Edge Attributes, Topology Target: Neighbors.
Hypothesis: Edge Attributes
Scoring: Most common interaction is 'Mentions', worth .5 points. The number of times it has happened is 4, worth .5 points if part A was correct.
Links: NL, AM

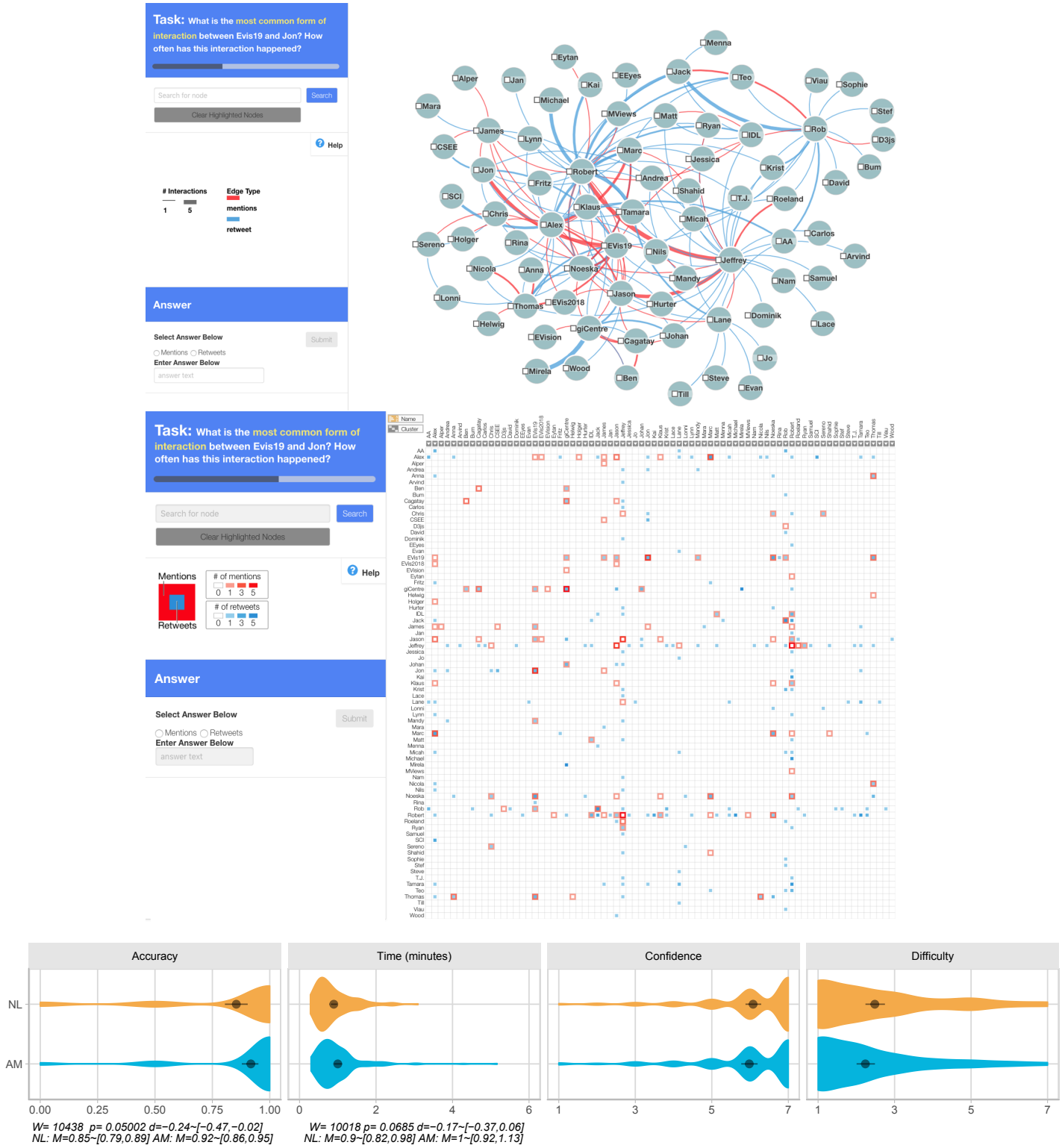


Figure 10: Task 9 configuration and results.

2.10 Task 10: Node Attribute Comparison

Instruction: Select all of Noeska's neighbors that are people and have more friends than followers.
Properties: Large Network, 3 Node Attributes, Topology Target: Neighbors.
Hypothesis: Within-Node Attribute Comparison Hypothesis
Scoring: Correct answers are Lonni, Thomas, Anna, and Klaus. 1/4 point for each correct answer. -1/4 point for each incorrect answer.
Links: NL, AM

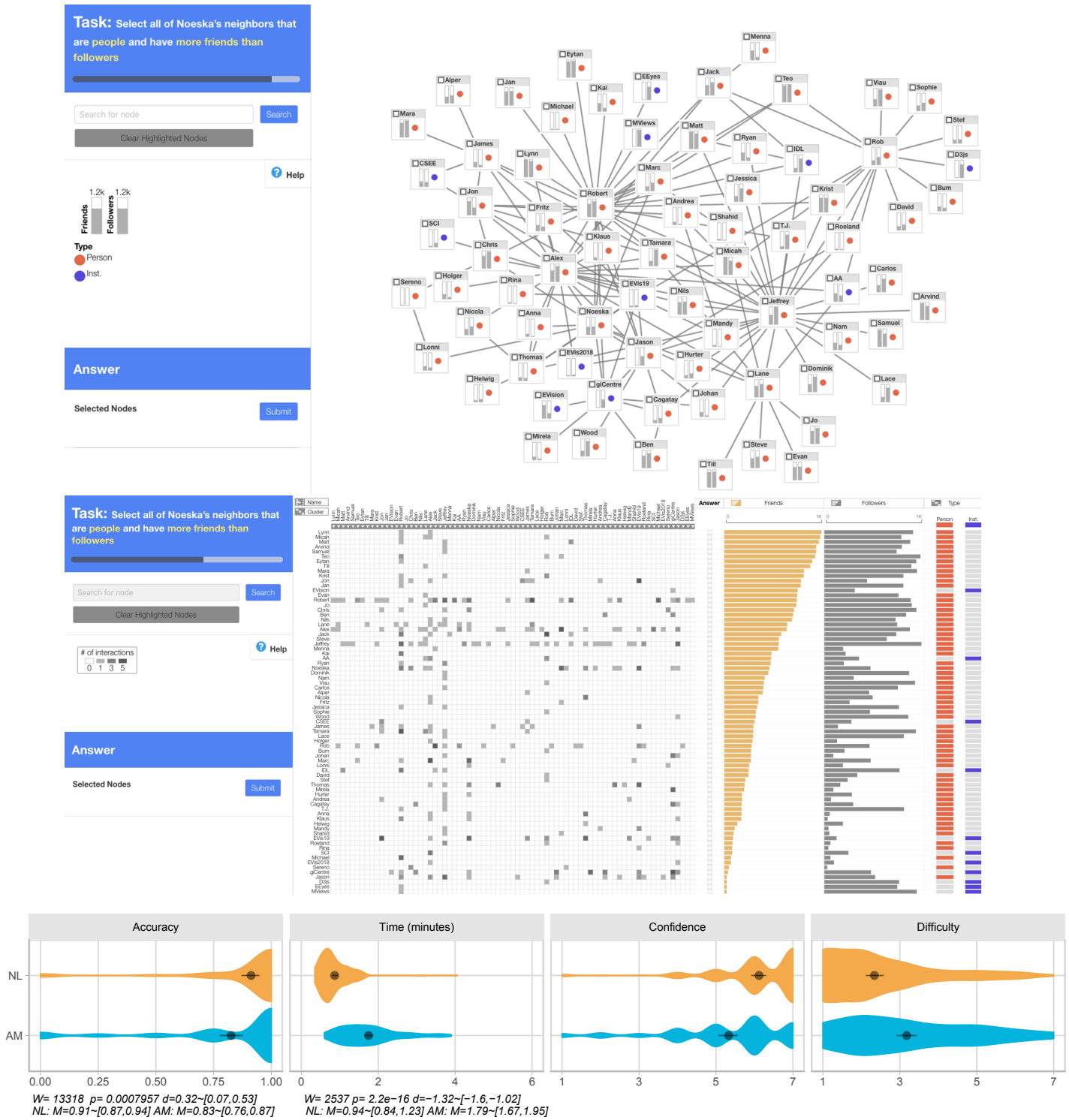


Figure 11: Task 10 configuration and results.

2.11 Task 11: Node Attribute Comparison on Small Network

Instruction: Select the people who have interacted with Thomas and have more friends than followers.
Properties: Small Network, 3 Node Attributes, Topology Target: Neighbors
Hypothesis: Within-node Attribute Comparison Hypothesis
Scoring: Correct answer is Anna. Full score for only Anna, 0 otherwise.
Links: NL, AM

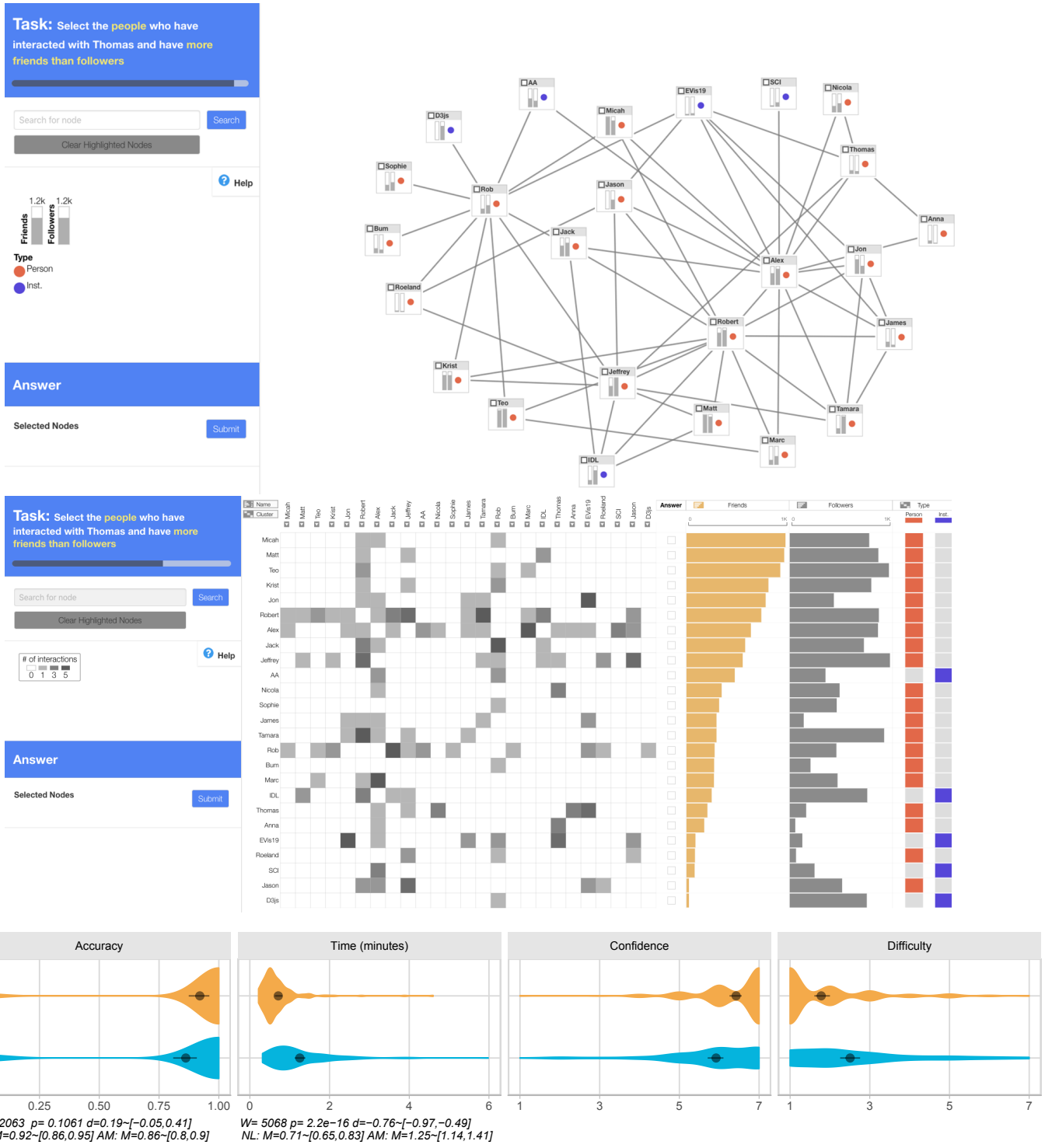


Figure 12: Task 11 configuration and results.

2.12 Task 12: Cluster and Attribute Estimation

Instruction: Select all the people who are in a cluster with Alex. Estimate the average number of followers among the selected people.
Properties: Large Network, 1 Node Attribute, Topology Target: Cluster.
Hypothesis: Cluster Hypothesis
Scoring: For subquestion 1, we determined clusters by two different methods: using a network clustering plugin to Cytoscape [2], and a seriation algorithm (optimal leaf clustering) [1] for the adjacency matrix. Based on these algorithmically defined clusters, we inspected the NL and AM to identify which nodes are distinctly in clusters in both visualizations and in the cluster results from the algorithms. Based on this, we defined a core cluster containing Alex, Robert, Noeska, and Jason. There are other members of the core cluster, but they are institutions, not people. The cluster score was then defined by the edit distance to the correct answer, with the following exceptions: As Alex was the node asked for in the question, including it doesn't get points, leaving it out doesn't incur a penalty. We also defined an extended cluster that contained people nodes that could reasonably be included in the cluster. This extended cluster includes Tamara, James, Jon, Marc and Klaus. These nodes were excluded from calculating the edit distance, i.e., including them did not incur a benefit or penalty.

For subquestion 2, the average number of followers of the cluster members, we averaged the number of followers for the nodes selected by the user in part A and computed the standard deviation of those values. The score was weighted from average -1/2 std dev to average +1/2 std dev, with full score given for the average, going to 0 at the extremes. The score for the combined task was the score for part A multiplied by the score for part B.
Links: NL, AM

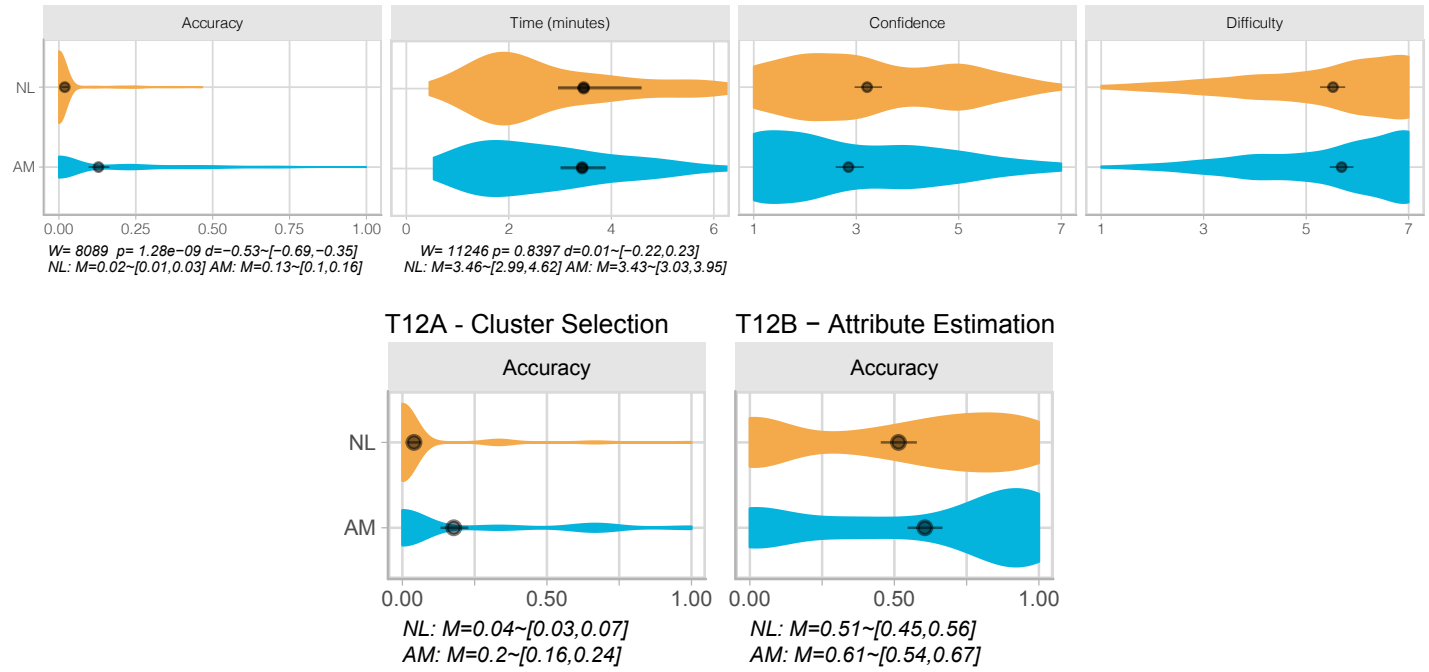


Figure 13: Task 12 results.

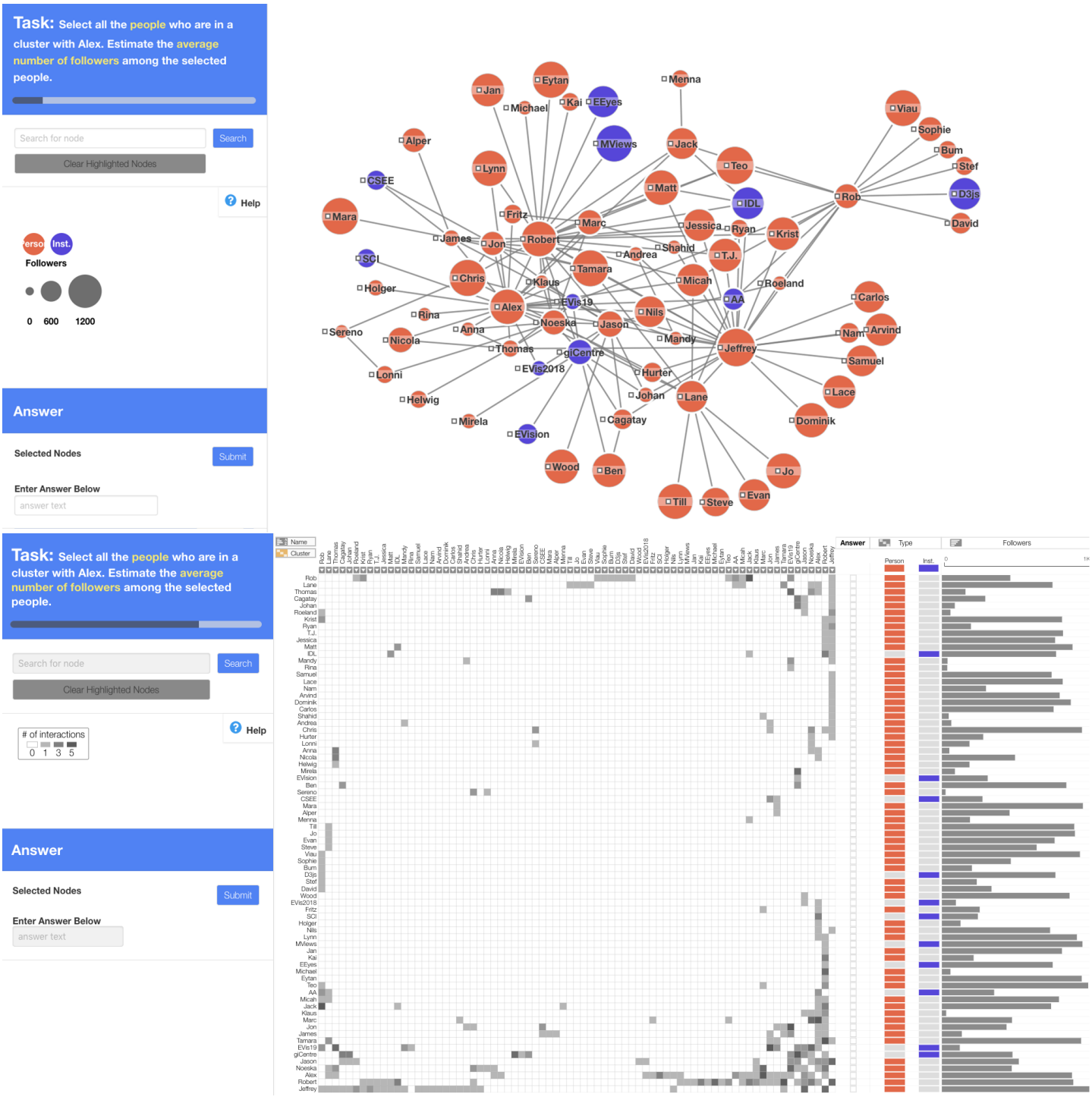


Figure 14: Task 12 configuration.

2.13 Task 13: Attribute Along Shortest Path

Instruction: What is the institution on a shortest path between Lane and Rob? What is its continent of origin?
Properties: Large Network, 2 Node Attributes, Topology Target: Paths.
Hypothesis: Path Hypothesis
Scoring: The answer to subquestion 1 is AA, for which the user is awarded .5 point. Anything else is 0. If the user got the first part right, the continent of origin is EU, worth another .5 points.
Links: NL, AM

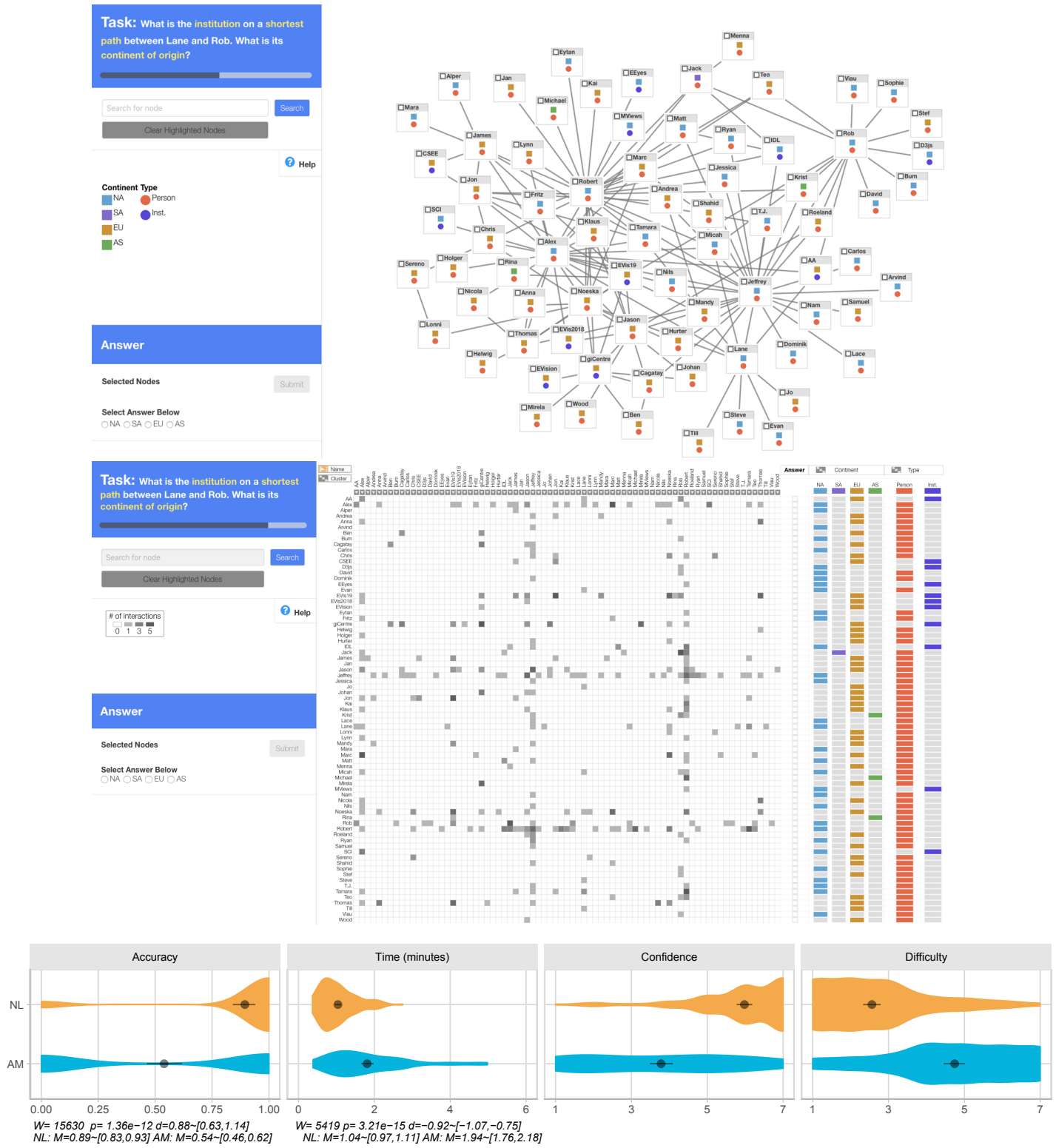


Figure 15: Task 13 configuration and results.

2.14 Task 14: Attribute Along Shortest Path on Small Network

Instruction: What is the institution on a shortest path between Jason and Job? What is its continent of origin?
Properties: Small Network, 2 Node Attributes, Topology Target: Paths
Hypothesis: Path Hypothesis
Scoring: The answer to subquestion 1 is EVis19, for which the participant is awarded .5 point. Anything else is 0. The answer to subquestion 2 is EU. If the participant got subquestion 1 right, and then got subquestion 2 right, they got another 0.5 points.
Links: NL, AM

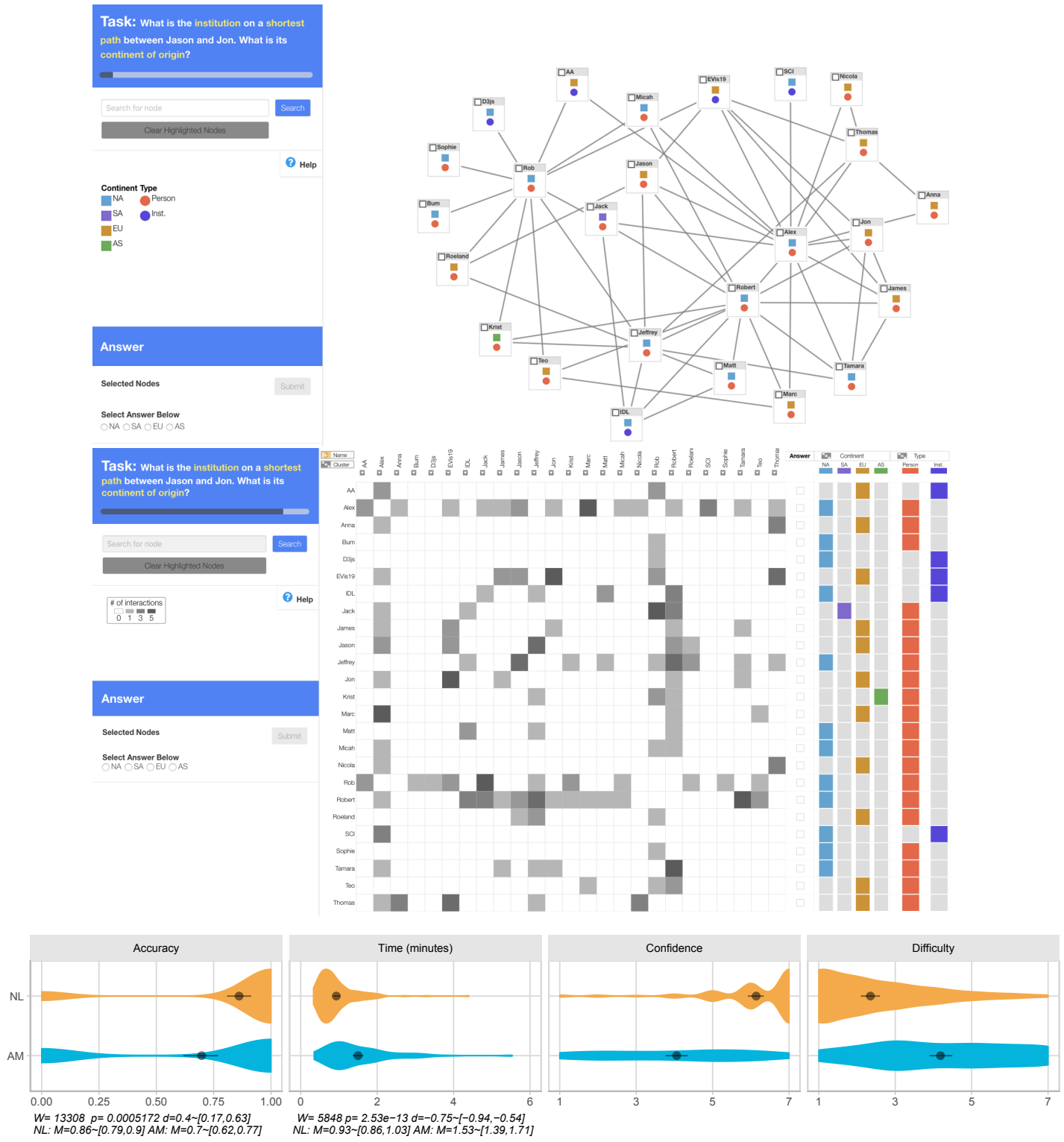


Figure 16: Task 14 configuration and results.

2.15 Task 15: Attribute on Multiple Paths

Instruction: Of the north americans who are two interactions aways from Sereno, who has been on twitter the longest?
Properties: Large Network, 1 Node Attribute, Topology Target: Paths
Hypothesis: Path Hypothesis
Scoring: The answer is Robert. 1 point Robert, 0 otherwise.
Links: NL, AM

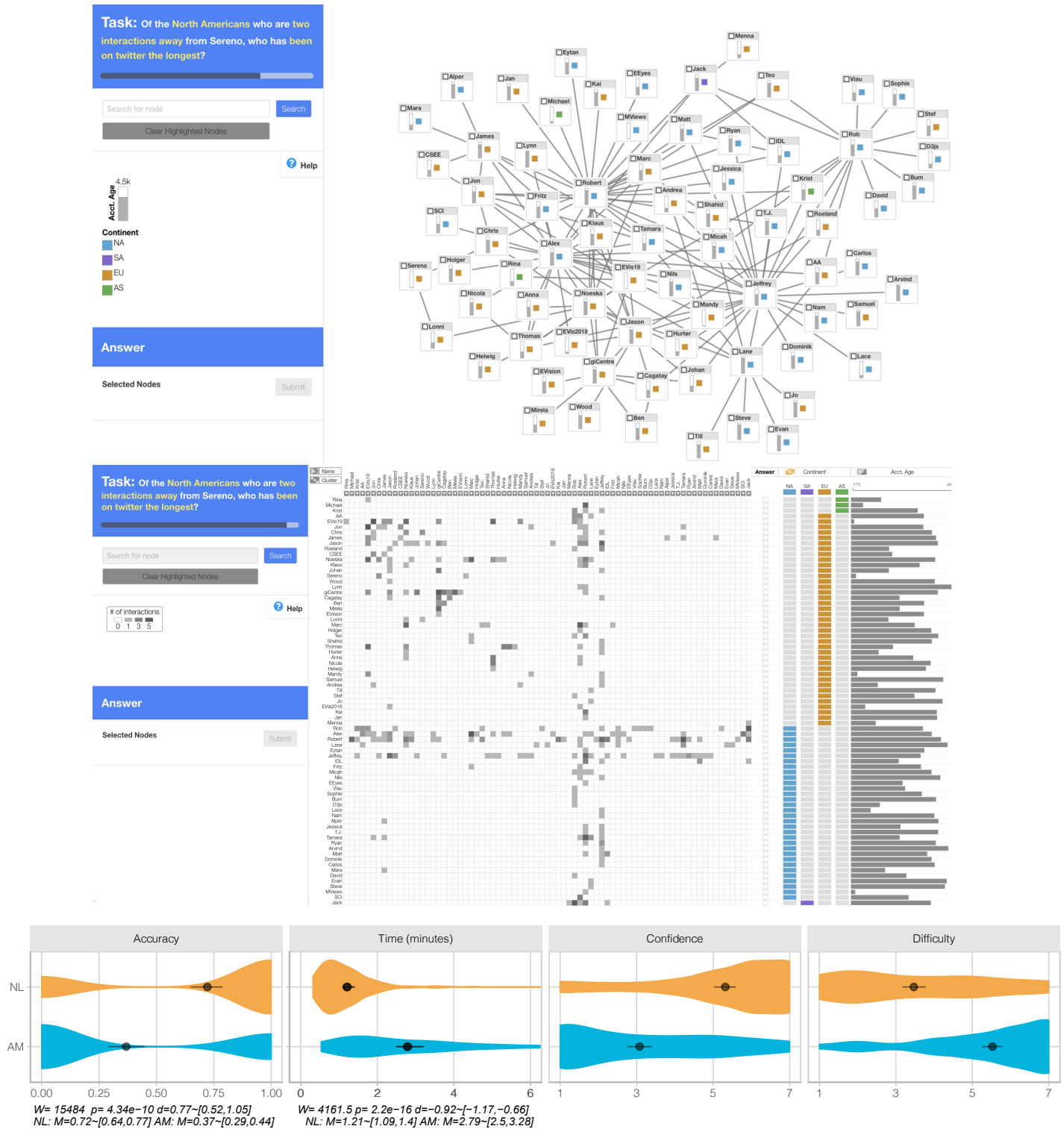


Figure 17: Task 15 configuration and results.

2.16 Task 16: Free Explore

Instruction: Please explore the network freely and report on your findings. Is there anything surprising or particularly interesting in the network?
Properties: Large Network, 6 Node Attributes, Topology Target: NA
Hypothesis: Insight Generation Hypothesis
Scoring: Qualitative coding based on answer types..
Links: NL, AM

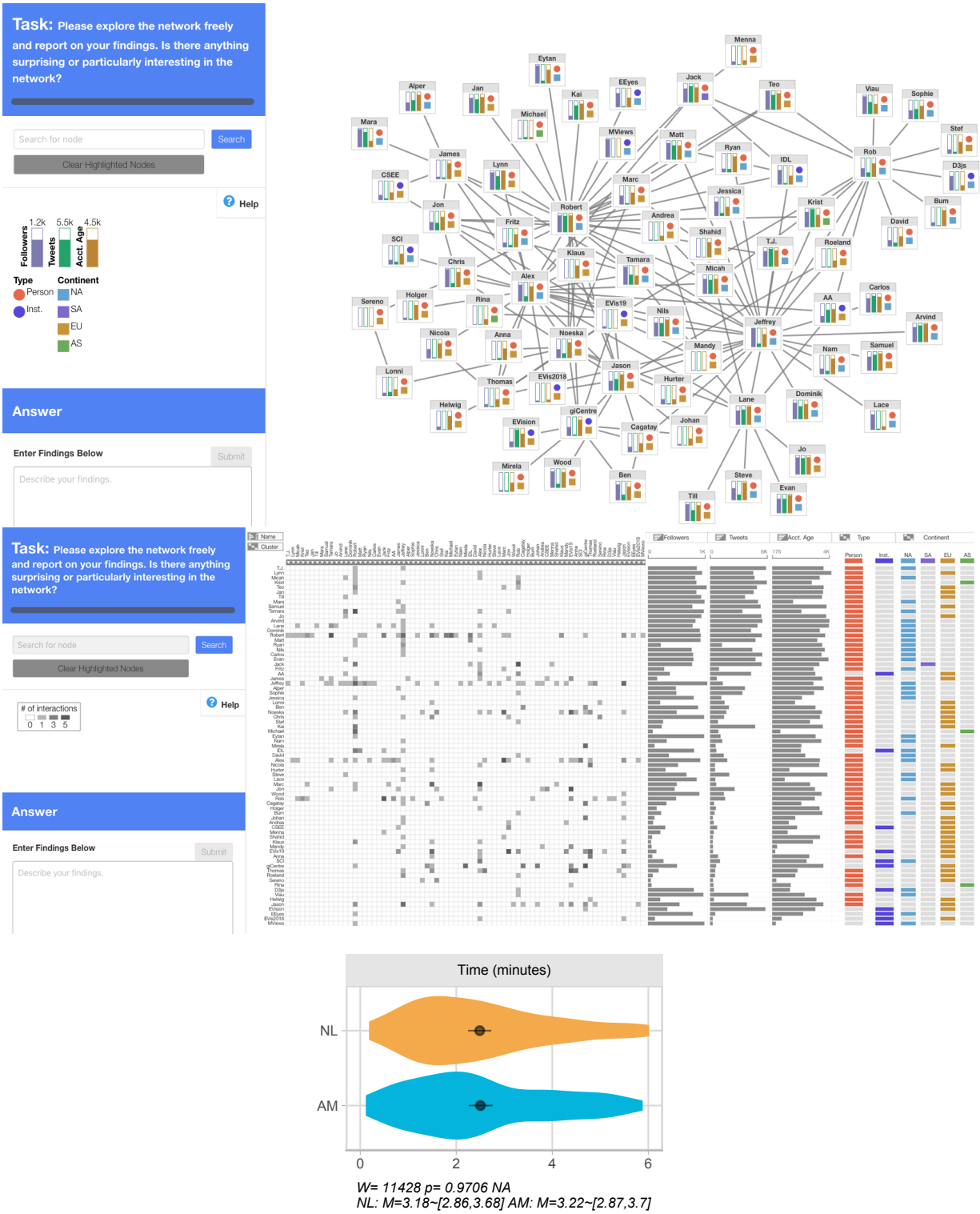


Figure 18: Task 16 configuration and results.

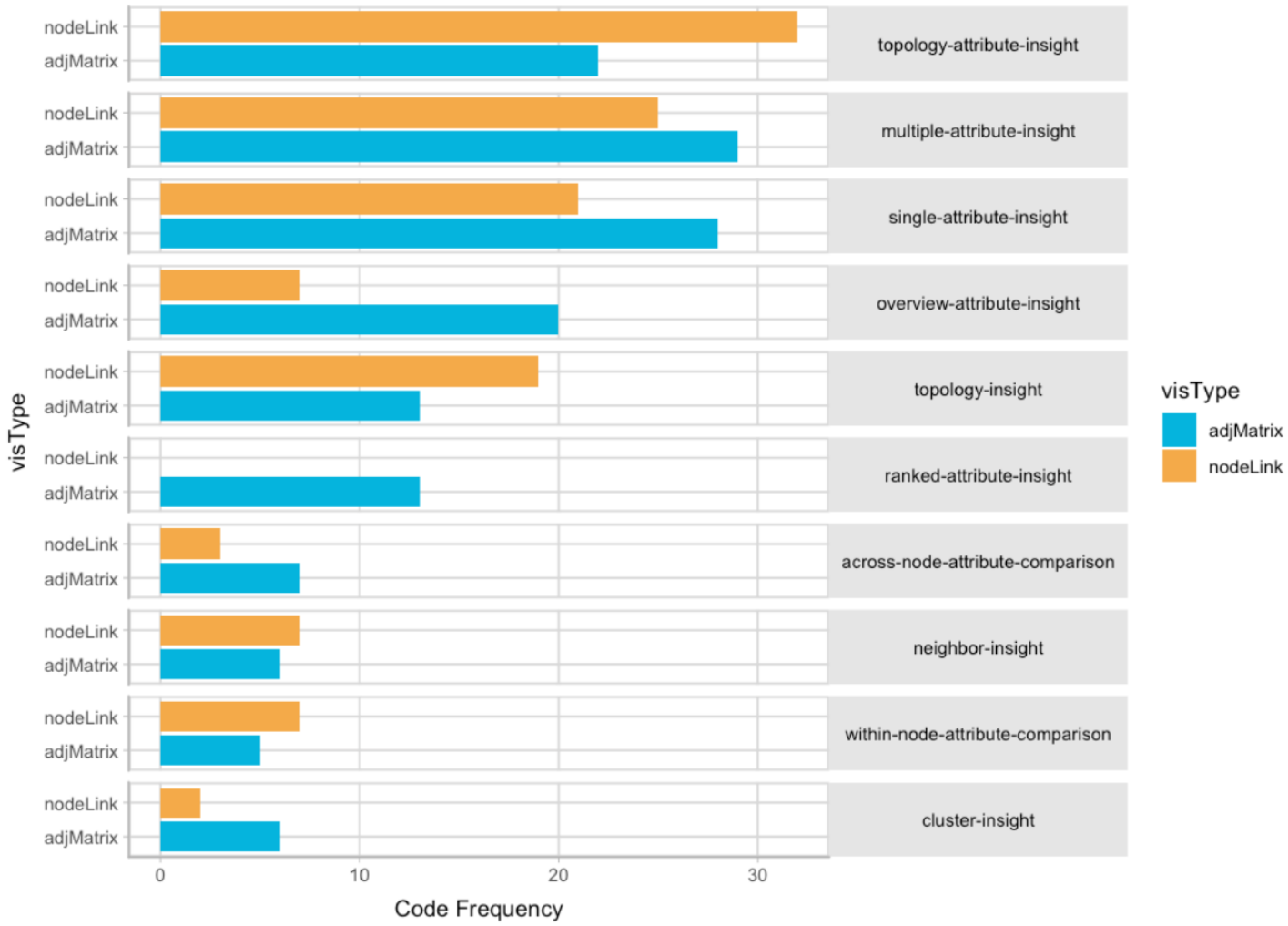


Figure 19: Frequencies of insights faceted by type of insight and condition.

3 Provenance

We tracked various types of interactions using a provenance framework. The following figures show the frequency of selected interactions. We used custom visualizations to inspect the provenance data, which is available at https://vdl.sci.utah.edu/mvnv-study-analysis/.

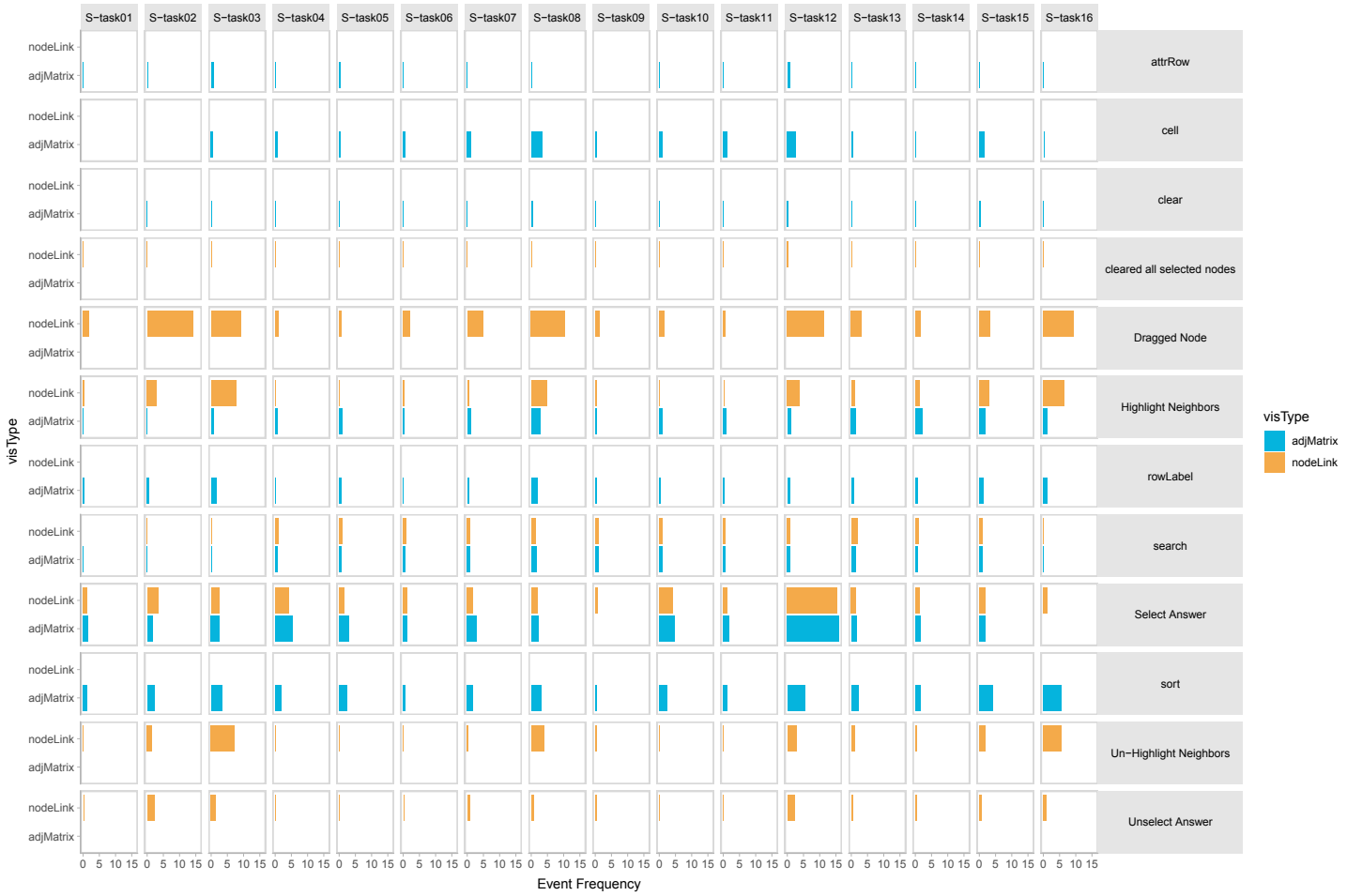


Figure 20: Interaction count per participant as a function of task and specific interaction type.

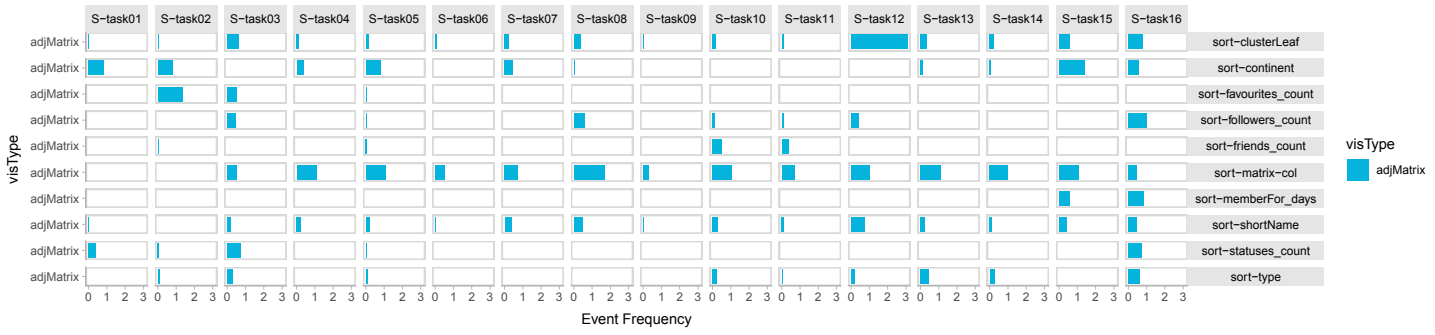


Figure 21: Interaction count per participant for sorting interactions in the Adjacency Matrix.

4 Design Guidelines

For a sparse network with few attributes, we recommend a design similar to our node-link diagram with bubbles. Global tasks on a single numerical attribute are about as well supported as in the AM. Most topology tasks are supported about equally well, with the exceptions of path-related tasks, which are much better supported in NL, and cluster tasks, which are better supported in AM. We believe it is important to provide the ability to selectively show attributes of interest in NL, so that bubble size can be leveraged and nested charts can be avoided. Interactions such as node dragging and neighborhood highlighting were extensively used, and a system should certainly provide them.

For sparse networks with attributes that need to be analyzed simultaneously, and any dense network, we recommend a design similar to our adjacency matrix. While performance on the task that had the most attributes but also considered topology (T3) was slightly more accurate with the NL, the AM performed well on global discovery tasks (e.g., T1) and resulted in a lot of overview insights. The AM is clearly the method of choice for discovering clusters or communities and characterizing their attributes, but is also competitive for neighborhood and common neighbour tasks. As far as interaction is concerned, sorting was extensively used in the matrix, and is clearly an important feature for any implementation.

References

[1] Jean-Daniel Fekete. 2015. Reorder.Js: A JavaScript Library to Reorder Tables and Networks. In *IEEE Symposium on Information Visualization*. Chicago, United States, 3.

[2] Min Li, Jian-er Chen, Jian-xin Wang, Bin Hu, and Gang Chen. 2008. Modifying the DPClus Algorithm for Identifying Protein Complexes Based on New Topological Structures. *BMC Bioinformatics* 9, 1 (Sept. 2008), 398. DOI:http://dx.doi.org/10.1186/1471-2105-9-398