This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics

Touch? Speech? or Touch and Speech? Investigating Multimodal Interaction for Visual Network Exploration and Analysis

Ayshwarya Saktheeswaran, Arjun Srinivasan, and John Stasko

Abstract—Interaction plays a vital role during visual network exploration as users need to engage with both elements in the view (e.g., nodes, links) and interface controls (e.g., sliders, dropdown menus). Particularly as the size and complexity of a network grow, interactive displays supporting multimodal input (e.g., touch, speech, pen, gaze) exhibit the potential to facilitate fluid interaction during visual network exploration and analysis. While multimodal interaction with network visualization seems like a promising idea, many open questions remain. For instance, do users actually prefer multimodal input over unimodal input, and if so, why? Does it enable them to interact more naturally, or does having multiple modes of input confuse users? To answer such questions, we conducted a qualitative user study in the context of a network visualization tool, comparing speech- and touch-based unimodal interfaces to a multimodal interface combining the two. Our results confirm that participants strongly prefer multimodal input over unimodal input attributing their preference to: 1) the freedom of expression, 2) the complementary nature of speech and touch, and 3) integrated interactions afforded by the combination of the two modalities. We also describe the interaction patterns participants employed to perform common network visualization operations and highlight themes for future multimodal network visualization systems to consider.

Index Terms—Multimodal Interaction; Network Visualizations; Natural Language Interfaces;

◆

1 INTRODUCTION

Network visualizations, often in the form of node-link diagrams, are useful for describing and exploring data relationships in many domains such as biology [1], the social sciences [2], and transportation planning [3], just to name a few. When visually exploring networks, people often need to focus on subgraphs of interest (e.g., by selecting specific nodes and links, filtering), investigate specific connections (e.g., finding adjacent nodes, following paths), and adjust the visual properties of the network (e.g. changing graphical encodings such as color and size). Given this multitude of tasks, interaction plays a vital role during visual network exploration as users need to engage with both elements in the view (e.g., nodes, links) and interface controls (e.g., sliders, dropdown menus).

With the growing size and complexity of networks, recent work has begun to examine more fluid and expressive platforms and interaction techniques for visual network exploration and analysis. Researchers have explored a number of settings including tabletops and vertical touchscreens (e.g., [4], [5], [6], [7]), AR/VR (e.g., [8], [9], [10], [11]), and large wall-sized displays (e.g., [12], [13], [14]), among others, facilitating interaction through a variety of input modalities such as touch, pen, gestures, and speech.

Given the diverse and complementary strengths and weaknesses of different input modalities, an emerging theme within visualization research has been to explore multimodal interfaces that combine two or more modes of input (e.g., [5], [15], [16], [17], [18]). In the context of network visualizations, such multimodal interfaces may enable a more fluid interaction experience [19], allowing people to perform common operations such as finding paths and changing visual encodings (e.g. through speech), while simultaneously interacting with and investigating different parts of the network (e.g. through touch). Although multimodal interaction with network visualization seems like a promising idea, many open questions persist. For instance, do users actually prefer multimodal input over unimodal input, and if so, why? Does it enable them to interact more naturally, or does having multiple modes of input confuse users? When employing multiple modalities, how do people interact with networks and perform common network visualization operations?

We conduct a qualitative user study with 18 participants to address such questions and investigate user interactions with a multimodal network visualization system. Ultimately, by understanding more about user interaction and preferences, we seek to help future designers build better multimodal visualization systems that are seeded by people's natural behaviors [20]. Along these lines, we focus on two increasingly popular input modalities that are ubiquitous across applications and devices—namely, touch and speech. To derive practical evidence of how people cope with system limitations and react when the system does not behave as expected, we perform this investigation using a working prototype of a speech- and touch-based network visualization tool, *Orko* [5]. We develop two unimodal (speech-only and touch-only) network visualization systems to enable comparison against the multimodal version of the Orko system. We split participants into three groups of six participants: one group interacted with both the touch-only version and the multimodal version, the second group

---

- *Ayshwarya Saktheeswaran, Arjun Srinivasan, and John Stasko are with Georgia Institute of Technology, Atlanta, GA, 30332.*
E-mail: {ayshwarya6, arjun010}@gatech.edu, stasko@cc.gatech.edu

Manuscript received December 20, 2019; revised January 23, 2020.

interacted with both the speech-only version and the multimodal version, and the third group only interacted with the multimodal version of the system. Collectively, based on our observations and participant feedback from the study, we make the following contributions:

- Verifying that people prefer speech- and touch-based multimodal input over unimodal input during visual network exploration, we identify specific factors explaining this preference including: the freedom of expression, the complementary nature of speech and touch, and integrated interaction experience afforded by the combination of the two modalities.
- Furthermore, to aid the design of future systems, we describe interaction patterns that participants employed to perform common network visualization operations (e.g., finding paths, filtering) and highlight promising areas for future work.

2 RELATED WORK

An underlying motivation for our work is based on visualization research themes highlighted in Lee et al.'s article [20] where the authors emphasize "going beyond the mouse and keyboard" as one of the key opportunities for visualization research. Along these lines, multiple systems and studies have explored how people interact with visualizations in post-WIMP settings such as mobile/tablets [21], [22], [23], [24], large interactive displays [14], [25], [26], and even virtual environments [9], [27], [28]. Given the widespread adoption of direct manipulation as an interaction technique in desktop-based visualization systems, a majority of these systems have explored the use of touch-based input for interaction [21], [22], [25], [26], [29], [30], [31]. Furthermore, based on its increasingly important role as part of our daily interaction with technology, natural language is another form of input that has gained increased interest within the visualization research community [17], [18], [32], [33], [34], [35], [36] and as part of commercial systems [37], [38], [39].

While existing systems have demonstrated that both touch and natural language exhibit potential to facilitate interaction with visualizations, recent work has conjectured that the combination of the two is perhaps even more promising [15]. For instance, with Orko, Srinivasan and Stasko [5] demonstrated how speech- and touch-based multimodal interaction can be used during visual network exploration and analysis. Kassel and Rohs [16] recently presented a tablet-based visualization system, Valletto, that allows users to specify visualizations through a combination of touch and speech-based input. The development of these systems and their preliminary studies demonstrate that multimodal input combining speech and touch is feasible for interacting with visualization systems. However, to ensure that we explore the potential of such interfaces to their fullest, we need to understand people's natural behavior, preferences, and expectations from such interfaces [20]. Prior studies have explored how people use pen and touch (e.g., [4], [40]), touch and proximity/spatial movement (e.g., [14], [41]), and touch and tangible objects (e.g., [42], [43]), among others in the context of visualization tools. Earlier studies in the broader HCI community also have investigated multimodal interaction involving speech [44], [45], [46] to better understand the benefits of multimodal input in terms of user performance metrics such as time and error. As part of our work, we focus on exploring what aspects of speech- and touch-based multimodal input makes it promising for interaction with network visualization tools.

Network visualizations have been extensively studied by the visualization community and many existing systems allow people to interactively explore networks by visualizing them using different layouts and representations. A complete review of network visualization systems can be found in survey reports such as [47], [48], [49]. More relevant to our work, however, are taxonomies that characterize key analytic tasks and operations people perform when interacting with network visualizations [50], [51], [52]. Specifically, we leverage these taxonomies to generate tasks for our user study so they are representative of what people might do when conducting visual network exploration and analysis in a realistic scenario. Furthermore, given the crucial role of interaction while visually exploring network-based data, many researchers have examined the use of different input modalities for interacting with network visualizations across a range of devices. For instance, Frisch et al. [4] demonstrated how people use pen and touch to edit node-link diagrams on tabletops. Schmidt et al. [7] and Thompson et al. [6] also have explored multi-touch interactions for facilitating interaction with network visualizations focusing on operations like selection and basic layout editing. More recent work has also begun to investigate gesture-based interaction with network visualizations in virtual reality [8], [9], [10]. Our findings contribute to this growing space of network visualization systems in post-WIMP settings by furthering our understanding of how people interact with networks using touch and speech.

3 STUDY SYSTEMS

We used the Orko system [5] as our test bed for the study. A detailed description of Orko's features and implementation can be found in [5][1]. Below we summarize the changes we made to the system for our study. These changes were based on Orko's preliminary user study findings [5] and a series of eight pilot sessions we conducted as part of our work.

3.1 Interfaces and Operations

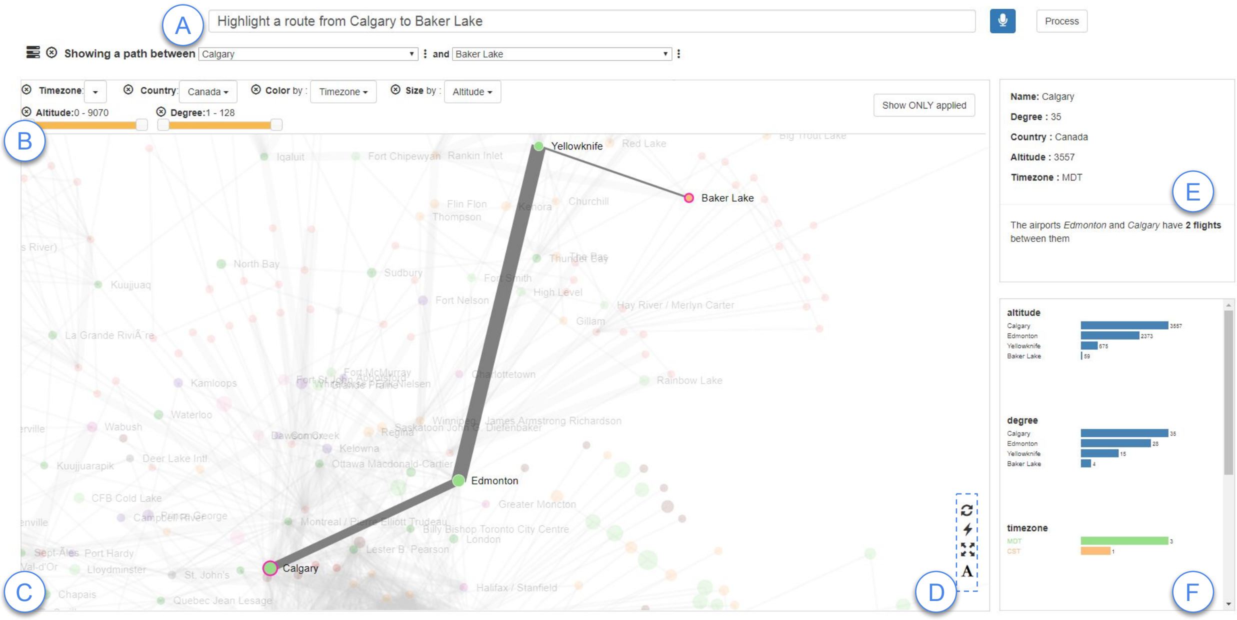
Multimodal interface. Figure 1 shows the final multimodal interface we used for our study. Similar to the original system, the interface consists of the speech input and feedback row (Figure 1A), the visualization canvas (Figure 1C), quick-access icons (Figure 1D), and details and summary panels (Figure 1E,F). To enhance its visibility, we repositioned the filters and encoding row (Figure 1B) placing it above the visualization canvas as opposed to below the canvas in the original system.

To enable comparison between unimodal and multimodal input, we developed two unimodal systems mimicking Orko's user interface components. Table 1 lists the operations supported across all three interfaces and how they could be performed using touch, speech, or a combination

---

[1] 1. We also provide a link to videos demonstrating the Orko system as supplementary material.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics



Highlight a route from Calgary to Baker Lake

Showing a path between Calgary and Baker Lake

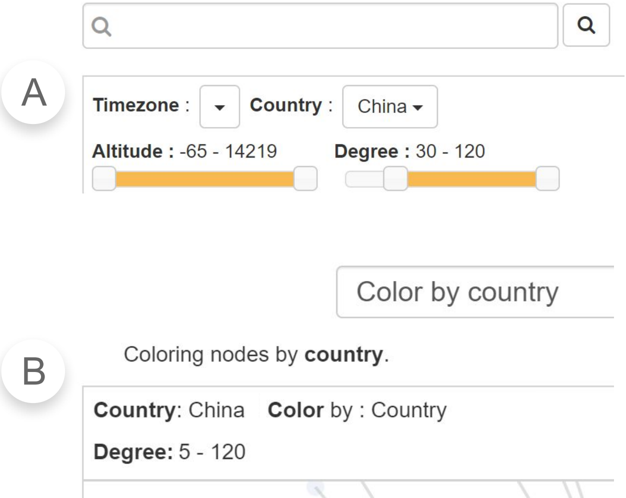
Timezone Country: Canada Color by: Timezone Size by: Altitude

Altitude: 0 - 9070 Degree: 1 - 128

Name: Calgary
Degree : 35
Country : Canada
Altitude : 3557
Timezone : MDT

The airports *Edmonton* and *Calgary* have **2 flights** between them

Fig. 1. User interface of the multimodal study interface. A) Speech input and feedback, B) Filters and encodings, C) Visualization canvas, D) Quick-access icons, E) Details panel, and F) Summary panel. In this case, the system is highlighting a path between Calgary and Baker Lake.



Timezone : Country : China

Altitude : -65 - 14219 Degree : 30 - 120

Color by country

Coloring nodes by **country**.

Country: China **Color** by : Country

Degree: 5 - 120

Fig. 2. Screenshots from the unimodal study systems displaying the input, filters, and encodings rows in the (A) touch-only and (B) speech-only interface.

of the two. In addition to the operations listed in Table 1, participants could also select nodes by tapping, drawing a lasso, or using speech (e.g. "*Select all airports in China*").

Touch-only interface. This interface functioned comparably to current network visualization tools, allowing people to select nodes and links and navigate the view with simple touch gestures, and adjust sliders and dropdown menus to filter points or change visual encodings. Additionally, we replaced the speech input and feedback row (Figure 1A) by a single input box that facilitated searching for nodes (Figure 2A). Entering search terms was supported through a virtual keyboard since this was a touch-only system.

Speech-only interface. This interface supported all operations listed in Table 1 through speech alone. Touch input was disabled throughout the interface. Visually, this interface had the same components as shown in Figure 1 with one key difference: since it was a voice-based system, participants could not directly manipulate the sliders and dropdown menus (Figure 2B). In other words, to adjust filters or apply visual encodings, participants always had to use voice commands (e.g., "*Change the timezone to CST*," "*Size by degree*").

3.2 Triggering and Interpreting Speech Commands

In the speech-only interface, participants could trigger speech recognition using the wake word "*System*" (similar to "*Alexa*," "*Ok Google*"). In the multimodal interface, in addition to using the wake word, participants could also tap the microphone button (🎙) next to the input box (Figure 1A) to trigger speech recognition. Similar to the original Orko system, all interfaces were implemented as web-based systems using Python, HTML, CSS, and JavaScript. The standard HTML5 webkit speech recognition API [53] was used to recognize speech input. To improve recognition accuracy, we also trained the recognizer with system keywords (e.g., 'find,' 'color,' 'path') and values in the loaded dataset.

We reused Orko's command interpreter [5] to respond to speech commands. At a high-level, the system employs a two-step process to interpret commands. First, an input command is matched against the pre-defined grammar patterns (e.g. `Color by [color]`) defined using the Artificial Intelligence Markup Language (AIML) [54] to identify the operations (e.g., *Find Path*, *Filter*) and attributes/values (e.g., *Calgary*, *altitude*). If the input command does not match a pre-defined pattern, the system then tokenizes the command string and compares it to the underlying lexicon (composed of attributes and values in the dataset, as well as keywords such as 'find,' 'path,' 'filter,' etc.) Based on

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics

TABLE 1
Operations supported in the study interfaces with their corresponding touch-only, speech-only, and multimodal interactions. Note that the speech commands shown are only examples and the systems supported a wider variety of phrasings. Speech commands preceded by “>” are examples of follow-up commands.

Operation	👆 **Touch**	🗣 **Speech**	👆 + 🗣 **Touch+Speech**
Find Nodes	Type label using virtual keyboard	*“Find Calgary airport”*, *“Show Canberra”*, *“Search for Auckland”*	🗣 *“Find Canberra”* + 👆 update node using query manipulation widgets (Figure 1A)
Find Connections	Double tap on a node	*“Show connections of Adelaide”* >*“How about Auckland”*, *“Find airports with connections to Wales”*	👆 Select nodes + 🗣 *“Show connections”*, 🗣 *“Show airports connecting to Auckland”* + 👆 Update node using query manipulation widgets
Find Paths	Long press on source node and tap on target node	*“Highlight a route from Normanton to Julia Creek”*, *“Find a path between Billings and Denver”*	👆 Select nodes + 🗣 *“Highlight path”*, 🗣 *“Show route between Calgary and Baker Lake”* + 👆 Update nodes using query manipulation widgets
Filter Nodes	Adjust dropdowns and sliders in filters & encodings row (Figure 1A)	*“Just show airports in Central Standard Time”*, *“Filter to show Canadian airports at an altitude of over 5000 feet”* >*“Focus on ones with degree more than 10”*	🗣 *“Filter by degree”* + 👆 Adjust degree slider, 🗣 *“Only highlight airports in Australia”* + 👆 Change country using dropdown
Change Visual Encodings	Adjust dropdowns in filters & encodings row	*“Color airports by timezone”*, *“Size nodes by degree”* >*“Now by altitude”*	🗣 *“Color by country”* + 👆 Change country using color dropdown, 🗣 *“Resize nodes”* + 👆 Select attribute from dropdown
Navigate	Two-finger pinch for zoom, one-finger drag on canvas for pan	*“Zoom out”*, *“Center graph”*, *“Zoom in more”*, *“Pan left”* >*“Some more”*, *“Move right”*	— (only supported through touch *or* speech)
Interface Actions	Tap quick-access icons (Figure 1D)	*“Refresh canvas”*, *“Show all node labels”*	👆 Select nodes + 🗣 *“Show labels”*

this comparison, the system identifies the operations and attributes/values using both syntactic (cosine similarity [55]) and semantic (Wu-Palmer similarity score [56]) similarity metrics. While we preserved the underlying architecture, to design the speech-only interface and support equivalence between interfaces, we extended the grammar and lexicon to support navigation operations (zoom/pan) through speech (e.g., “*zoom in*,” “*pan left*”).

4 STUDY

The ultimate objective of our study was to understand how people interact with network visualization tools using touch, speech, and a combination of the two. More specifically, we had three key goals when conducting the user study in the context of a network visualization tool:

RG1 Understand if and why multimodal interaction is preferred over unimodal interaction.

RG2 Understand if and how prior experience of interacting using one input modality impacts subsequent multimodal interaction.

RG3 Identify different input and interaction patterns people use for common operations during visual network exploration.

4.1 Methodology

We conducted a qualitative study where two groups of participants first interacted with either a speech-only or a touch-only interface followed by the multimodal interface. This allowed us to collect participant preferences and feedback to compare unimodal and multimodal interaction (**RG1**). As a baseline to see how people interact with the multimodal interface when they encounter it for the first time (without having worked with the speech- or touch-only version), we also included a third group of participants who only interacted with the multimodal version of Orko. Comparing the interactions of the first two groups with the third group allowed us to check if prior experience using the system with just one of the modalities resulted in any notable differences in terms of interaction behavior (**RG2**).

We considered different study designs including a three condition (touch, speech, multimodal) within-subjects study and a study where participants used unimodal touch or speech input and multimodal input in counterbalanced orders. However, a within-subjects study with three conditions would last over three hours ($\sim$60 min. per condition) which was impractical. In the second alternative, having participants use the unimodal system after the multimodal system would not allow us to assess the priming effects of an individual modality. In other words, if participants interacted with the multimodal interface first, they would already have experienced all the supported interactions, not allowing for any assessment based on prior experience using individual modalities.

4.2 Participants and Experimental Setup

We recruited 18 participants (P1-P18), ages 18-66, five females and 13 males. 14 participants were native English speakers and the remaining four participants self-reported as being fluent English speakers. We sent recruitment emails to university mailing lists and recruited participants on a “first come first serve” basis. Participants who only interacted with the multimodal system (P13-P18) received a $10 Amazon Gift Card as compensation whereas participants who interacted with both the unimodal and multimodal systems (P1-P12) received a $20 Amazon Gift Card as compensation.

In terms of their backgrounds, only eight participants said they had some prior experience of working with network visualization tools but 14/18 participants (except P2, P7, P8, P9) had some experience working with general visualization tools (e.g., Tableau, Excel). All participants had

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics

TABLE 2
Tasks used in the study.

Task	Unimodal (Asia Pacific Flight Network)	Multimodal (US-Canada Flight Network)
T1	Which of these airports have direct flights to both Auckland and Canberra: [Melbourne, Perth, Townsville, Adelaide, Queenstown].	Consider only one hop journeys from Hartsfield Jackson to Ted Stevens airport. Show that there are exactly 8 possible layover airports.
T2	Consider Auckland airport and the airports it has direct flights to: - Among these airports, show that Auckland has most flights to Sydney Kingsford Smith. - Show that among the Chinese airports that Auckland has direct flights to, Beijing Capital is the busiest airport. - Now assume you had to fly to Western Australia from Sydney Kingsford Smith airport. Name the most accessible airport in Western Australia that Sydney has a non-stop flight to.	Consider Edmonton airport and the airports it has direct flights to: - Of all these airports, show that Yellowknife is the airport it has the most flights to. - Consider airports in the United States that Edmonton has direct flights to. Show that Palm Springs is the least busy airport. - Now assume you had to fly from Edmonton to a city in the Central US region and you are traveling through Palm Springs. Name the airport in the Central US region that is most accessible from Palm Springs.
T3	Let us call airports that have direct flights to 55 or more airports as "popular" airports. - T/F: China has the most number of "popular" airports. - Now assume that you are traveling from Sydney Kingsford Smith airport to Domodedovo through one of these "popular" airports. T/F: you have to travel through either Thailand or China.	Show that there are only two Canadian airports that have direct flights to 40 or more other airports. T/F: Among all airports that have direct flights to both these Canadian airports, Denver is at the highest altitude.
T4	List the airports you would have to fly through when travelling from Normanton to Julia Creek.	Suppose you want to fly from Fairbanks to Wales. Find a set of airports through which you must fly.
T5	Assume you live in Brisbane and you want to go a high altitude location (>2100 feet) in Australia. Since there are are no direct flights to such locations, you would have to travel through at least one other airport when travelling from Brisbane. Name airport(s) you could fly through.	Say you are living in Billings and you want to go for a vacation to a high altitude location (>7000 feet). However, Billings Logan does not have direct flights to such locations, but it has a direct flight to an airport that does. Name that airport.
T6	Pick any two airports that have at least one direct international flight. Consider these two airports and the airports they have direct flights to. Now compare the two groups of airports with respect to different characteristics such as accessibility, altitude levels, variability in time zones, etc. You may also list any additional observations you make based on interacting with the network.	Pick any two airports that have at least one direct international flight. Consider these two airports and the airports they have direct flights to. Now compare the two groups of airports with respect to different characteristics such as accessibility, altitude levels, variability in time zones, etc. You may also list any additional observations you make based on interacting with the network.

prior experience working with touch-based devices including phones, tablets, and laptops. All but two participants (P2, P14) said they used speech-based systems (e.g., Siri, Alexa) frequently. None of the participants had any prior experience working with touch- or speech-based visualization systems. All participants interacted with the system running on Google's Chrome browser on a 55" Microsoft Perceptive Pixels device. The screen was set to a resolution of 1920 x 1080 pixels.

4.3 Dataset and Tasks

As the primary focus of our study was understanding user interactions, we had to ensure that the network selected for the study encouraged interaction with the visualization and allowed us to cover a wide variety of tasks including browsing, attribute-based filtering and reconfiguration, and group-level exploration [50], [52], [57]. Given this high-level goal, we wanted to select a dataset where: (1) nodes had both numerical and categorical attributes so participants could filter and change visual encodings and (2) the connections had an intrinsic meaning so the tasks could emulate real-world scenarios. Additionally, to avoid differences due to domain knowledge, we wanted a dataset from a domain that was familiar to all participants (i.e., participants knew what the different attributes meant). With these criteria in mind, we selected two undirected flight networks as our datsets for the study.

The first dataset contained 551 airports (nodes) in the Asia Pacific region and 2263 bidirectional flights between airports (links) whereas the second dataset contained 556 airports in United States and Canada and 2219 flights between those airports. Each airport in the dataset had four attributes including its *altitude*, *country*, *timezone*, and a derived attribute indicating number of airports it was connected to (*degree*). Participants explored the Asia Pacific network in the unimodal condition, and the US-Canada network in the multimodal condition.

Participants performed six tasks (T1-T6) with each dataset. Table 2 lists the tasks used during the study. These study tasks were generated based on existing network visualization task taxonomies [50], [51], [52] and included topology-level tasks, attribute-level tasks, browsing-tasks, and a group-level comparison task. For instance, in terms of Lee et al.'s taxonomy [50], T1 and T4 correspond to topology-based tasks (finding paths and connections), while T2, T3, and T5 involve a combination attribute-based tasks (filtering), browsing (following paths), and topology-based tasks. For P1-P12, tasks between the unimodal and multimodal conditions were designed such that they had a comparable level of difficulty. The order of tasks was randomized between conditions for each participant to prevent them from memorizing the operations they performed for a task.

To prevent participants from reading out the tasks as commands into the system as-is, we framed the tasks as a combination of scenario-based questions and jeopardy-style facts [18] that participants had to prove true/false. In other words, to "solve" a task, participants had to interact with the system and get to a point where the visualization either proved or disproved the given statement or highlighted the required sub-graph. For instance, consider the task "*Suppose you want to fly from Fairbanks to Wales. Find a set of airports through which you must fly.*" in Table 2. To solve this task, participants could either find the path between the Fairbanks and Wales airports or they could manually, incrementally explore connections out of one of these airports until they reached the other. In either case, since there were multiple correct answers, participants had to visually highlight or show the list of airports (path) that one would need to travel through.

4.4 Procedure

Sessions lasted between 50-60 minutes for participants who only interacted with the multimodal system and 125-135

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics

minutes for participants who interacted with both the unimodal and multimodal systems. The study procedure was as follows:

Consent and Background (3-5 min): Participants signed a consent form and answered a questionnaire describing their background with visualization tools and touch- and speech-based applications.

System Introduction (3-5 min): The experimenter introduced the system, describing the user interface and supported operations. For speech interaction, participants were only informed about the operations the system supported and were not given a detailed vocabulary or list of possible commands for each operation. Instead, participants were encouraged to interact with the systems as naturally as possible, using any commands they felt were appropriate in the context of the given datasets.

Practice (3-5 min): Participants tested the touch and speech input until they felt comfortable using them. In this phase, participants interacted with a network of 552 European soccer players (nodes) that were linked if they played for the same club or national team (6772 links). Each node had five attributes indicating the player salary, goals scored, field position, club, and country they represented.

Dataset and Task Introduction (3-5 min): Participants were given a description about the flight network dataset along with the six tasks printed on a sheet of paper.

Task Solving (30 min): Participants interacted with the system to solve the tasks. This phase was capped at 30 minutes. Participants were encouraged (but not mandated) to think aloud and interact with the experimenter, particularly when they felt the system functioned unexpectedly. To avoid prompting interactions or disrupting the participants' workflow, the experimenter did not intervene during the session and only responded when participants initiated the discussion.

Debrief (10-15 min): Participants filled out a post-session questionnaire and engaged in an interview describing their experience with the system.

Participants who performed tasks with two systems (P1-P12) were given a 15 minute break between the two sessions. After this break, except for the consent and background step, we followed the same procedure as with the first system. These participants were also asked to state and describe their preference between the unimodal and multimodal versions of the system during the debrief. We video recorded all participant interactions with the system and audio recorded all interviews.

4.5 Data Analysis

Two experimenters individually reviewed both the audio and video data collected during the study to identify themes in interaction patterns and participant feedback. The resulting themes were then collectively discussed and iteratively refined into groups of observations using an affinity diagramming approach. This helped us characterize subjective feedback and participant behavior to qualitatively answer the initial questions driving the study (**RG1**, **RG2**). Furthermore, we also performed closed coding of the session videos to categorize the different types of interactions performed during the study (**RG2**, **RG3**). For this analysis, we used the operations in Table 1 as our set of pre-established codes. For each attempt at performing an operation, we noted if a participant used speech, touch, or a combination of the two. For instance, if a participant filtered nodes using a single spoken query (e.g., "*Show airports located at over 2100 feet*"), we would count this as one speech-only interaction. Alternatively, to filter, one could also directly adjust the slider (touch-only) or use a combination of the two modalities ("*Filter by altitude*" + drag slider). The intended operations were generally apparent due to the 'think aloud' protocol, the design of the tasks, and by the participant's reaction to system's interpretation of their interaction. The closed coding was also performed by two experimenters individually and conflicting observations or mismatches in counts were collectively resolved. We also used the session videos to determine the task completion times.

5 RESULTS

Addressing our study goals, in this section, we describe our key findings corresponding to the preference for multimodal interaction (Section 5.2), the effect of priming users with one modality (Section 5.3), and the different input and interaction patterns employed by the participants (Section 5.4).

5.1 Task and Interaction Overview

11/12 participants (P1-P12) who interacted with the unimodal interface completed all six tasks, whereas one participant (P8) completed five. In terms of the correctness of task responses, four participants made errors: P3 and P7 answered one of the six tasks incorrectly and P8 and P9 responded incorrectly to two tasks. In the 18 sessions with the multimodal interface, all participants except P18 (who completed five tasks) completed all six tasks with only three participants (P3, P10, P15) making one error (each) while responding to the six study tasks. In terms of time, the task phase lasted, on average, 24 minutes with the touch-only interface, 23 minutes with the speech-only , and 21 minutes with the multimodal interface.

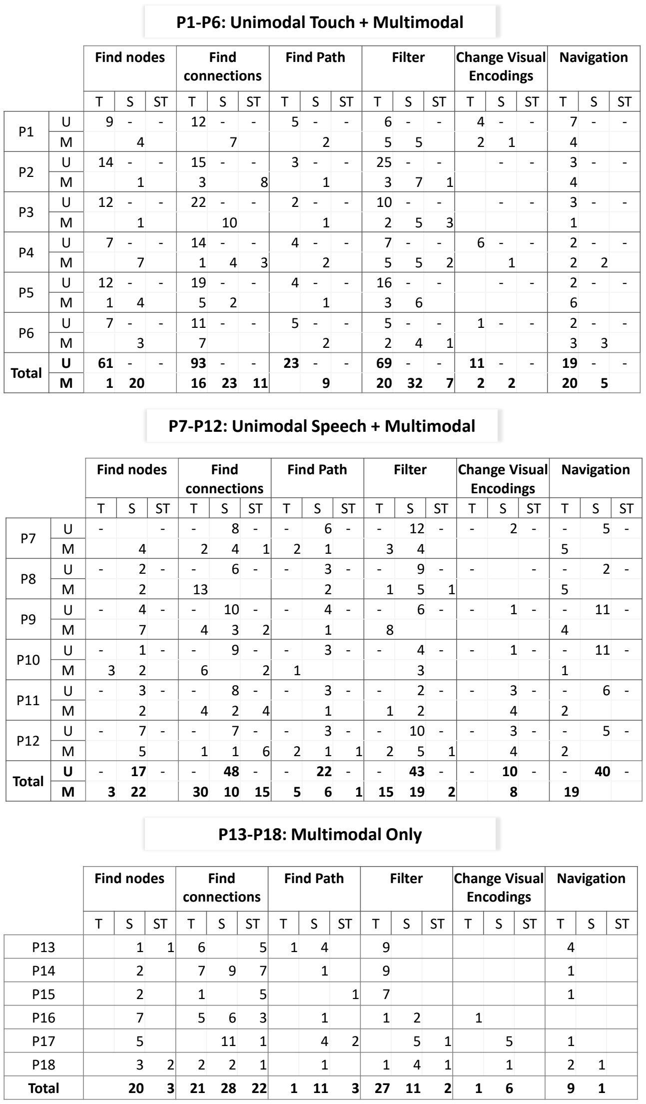
We recorded a total of 1052 interactions corresponding to the seven operations in Table 1 across the 18 participants and the two study interfaces. Table 3 shows the distribution of 945/1052 interactions for six operations (*Find Nodes*, *Find Connections*, *Find Path*, *Filter*, *Change Visual Encodings*, *Navigate*) that are common across network visualization systems. We exclude *interface actions* from Table 3 since these are generic tool-level operations (e.g. refreshing the canvas) and are not specific to network visualizations.

5.2 Preference for multimodal interaction

When asked which of the two systems they preferred, all 12 participants (P1-P12) who worked with both the unimodal and multimodal interfaces said that they preferred the multimodal system over the unimodal system. This was not surprising given similar findings in earlier studies [44], [46] and the simple fact that the multimodal system provided all capabilities that the unimodal system did. Hence, we

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics

TABLE 3
Distribution of 945 interactions used to perform six common network visualization operations during the study. **U**: Unimodal interface, **M**: Multimodal interface, **S**: Speech, **T**: Touch, **ST**: Multimodal interactions. A '-' indicates that a modality was not supported in a condition or that participants were not assigned to a condition.



P1-P6: Unimodal Touch + Multimodal

		Find nodes			Find connections			Find Path			Filter			Change Visual Encodings			Navigation		
		T	S	ST	T	S	ST	T	S	ST	T	S	ST	T	S	ST	T	S	ST
P1	U	9	-	-	12	-	-	5	-	-	6	-	-	4	-	-	7	-	-
	M		4			7			2		5	5		2	1		4		
P2	U	14	-	-	15	-	-	3	-	-	25	-	-		-	-	3	-	-
	M		1		3		8		1		3	7	1				4		
P3	U	12	-	-	22	-	-	2	-	-	10	-	-		-	-	3	-	-
	M		1			10			1		2	5	3				1		
P4	U	7	-	-	14	-	-	4	-	-	7	-	-	6	-	-	2	-	-
	M		7		1	4	3		2		5	5	2		1		2	2	
P5	U	12	-	-	19	-	-	4	-	-	16	-	-		-	-	2	-	-
	M	1	4		5	2			1		3	6					6		
P6	U	7	-	-	11	-	-	5	-	-	5	-	-	1	-	-	2	-	-
	M		3		7				2		2	4	1				3	3	
Total	**U**	**61**	-	-	**93**	-	-	**23**	-	-	**69**	-	-	**11**	-	-	**19**	-	-
	M	**1**	**20**		**16**	**23**	**11**		**9**		**20**	**32**	**7**	**2**	**2**		**20**	**5**	

P7-P12: Unimodal Speech + Multimodal

		Find nodes			Find connections			Find Path			Filter			Change Visual Encodings			Navigation		
		T	S	ST	T	S	ST	T	S	ST	T	S	ST	T	S	ST	T	S	ST
P7	U	-		-	-	8	-	-	6	-	-	12	-	-	2	-	-	5	-
	M		4		2	4	1	2	1		3	4					5		
P8	U	-	2	-	-	6	-	-	3	-	-	9	-	-		-	-	2	-
	M		2		13				2		1	5	1				5		
P9	U	-	4	-	-	10	-	-	4	-	-	6	-	-	1	-	-	11	-
	M		7		4	3	2		1		8						4		
P10	U	-	1	-	-	9	-	-	3	-	-	4	-	-	1	-	-	11	-
	M	3	2		6		2	1				3					1		
P11	U	-	3	-	-	8	-	-	3	-	-	2	-	-	3	-	-	6	-
	M		2		4	2	4		1		1	2			4		2		
P12	U	-	7	-	-	7	-	-	3	-	-	10	-	-	3	-	-	5	-
	M		5		1	1	6	2	1	1	2	5	1		4		2		
Total	**U**	-	**17**	-	-	**48**	-	-	**22**	-	-	**43**	-	-	**10**	-	-	**40**	-
	M	**3**	**22**		**30**	**10**	**15**	**5**	**6**	**1**	**15**	**19**	**2**		**8**		**19**		

P13-P18: Multimodal Only

	Find nodes			Find connections			Find Path			Filter			Change Visual Encodings			Navigation		
	T	S	ST	T	S	ST	T	S	ST	T	S	ST	T	S	ST	T	S	ST
P13		1	1	6		5	1	4		9						4		
P14		2		7	9	7		1		9						1		
P15		2		1		5			1	7						1		
P16		7		5	6	3		1		1	2		1					
P17		5			11	1		4	2		5	1		5		1		
P18		3	2	2	2	1		1		1	4	1		1		2	1	
Total		**20**	**3**	**21**	**28**	**22**	**1**	**11**	**3**	**27**	**11**	**2**	**1**	**6**		**9**	**1**	

were more interested in understanding what aspects of the combination of speech- and touch-based interaction with the system led participants to prefer it (**RG1**).

One hypothesis for why participants preferred the multimodal system, developed after reviewing their interaction counts in Table 3, was that in some cases, multimodal input allowed them to perform tasks with fewer interactions. However, tasks could be performed using multiple strategies through varied operations, each resulting in a different number of steps. Thus, basing the preference on interaction counts alone would be unjustified because we did not control for which strategy or operations participants used during a task. Instead, we coupled the participants' verbal comments and our observations of their interactions to identify three factors listed below that we believe led to their preference for multimodal interaction.

5.2.1 Freedom of expression

Out of the 945 interactions, 489 were performed in the context of the mulitmodal interface (P1-P6 M, P7-P12 M, and P13-P18 in Table 3). Among these, 233 (48%) used unimodal speech input, 190 (39%) involved unimodal touch input, and 66 (13%) used both modalities sequentially. Although only 13% of interactions involved sequential use of modalities, all participants used both modalities (individually or together) during at least three out of the six tasks in a session.

Interaction patterns also varied across participants for the same operation. For instance, observing the interactions for P13-P18 in Table 3, we can notice that P17 and P18 primarily used speech (individually or sequentially with touch) for filtering. On the other hand, P13-P16 primarily used touch to filter nodes. Interaction patterns varied even for individual participants across tasks. For instance, while performing the first task, P1 issued a unimodal speech command to find connections. However, during the second task, to find connections, he used speech and touch sequentially.

Participants also verbally commented on their preference for multimodal input over unimodal input in their post-session interviews. Participants said that having multiple modes of input gave them more freedom to try different ways to perform a task. For instance, highlighting the use of speech and touch for different operations, P8 said *"The combination is certainly better. Voice is great when I was asking questions or finding something I couldn't see. Touch let me directly interact."* Similarly, P2 said *"It (multimodal input) felt more natural. I really liked that I could choose what I wanted to do with my hands and what I wanted to say."* Stating multimodal interaction was more natural, P10 also said *"Working with this second system felt more natural. If I wanted to filter by something I could just say that but when I'd see something interesting I could touch it without having to say something and wait for the system to process it."*

The varied interaction patterns within and across participants coupled with the subjective comments highlight how the multimodal interface provided more freedom of expression, allowing participants to interact based on the task context or personal preferences.

5.2.2 Complementary nature of modalities

A popular hypothesis about multimodal interaction is that it allows users to offset the weaknesses of one modality with the strengths of another [15], [58]. Along these lines, when describing their experience with the multimodal system, 12 participants (P2, P4-6, P9-11, P13, P15-18) explicitly commented on the complementary nature of touch and speech and how it was a key advantage of multimodal input.

Participants found the ability to correct speech with touch very useful, with some participants even stating that the combination is vital to make effective use of speech. For example, P17 said *"I liked that I could correct with touch. Because it's not always going to be perfect right. Like the smart assistant on the phone sometimes gets the wrong thing but doesn't let me correct and just goes okay."* Talking about cases when the system populated the right filtering attribute but did not detect the right value, P2 said *"the system would bring the correct dropdown even if it didn't get the value right and then I could simply correct that."* In addition to correcting

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TVCG.2020.2970512, IEEE Transactions on Visualization and Computer Graphics

speech recognition and ambiguity errors with touch, participants also appreciated that they could leverage touch to modify existing queries. For instance, referring to the query manipulation dropdowns in the speech input feedback row (Figure 1A) , P18 said *"I liked that it allowed me to modify my command without having to say it again."*

On the other hand, participants also found the ability to use natural language when they either forgot a gesture or were unable to perform an operation using touch. P13, for instance, said *"I used voice when I didn't know how to do it with touch."* highlighting that speech can aid in overcoming memorability issues associated with touch gestures. Similarly, five participants (P4, P6, P13, P15, P18) used speech commands to navigate the view (i.e., zoom and pan) when they were in a dense region of the network and were unable to pinch or drag without touching the nodes on the canvas.

In addition to affirming the benefits of complementary interactions, such comments also highlight the value in further exploring elements like ambiguity widgets [18] to help users resolve challenges with speech and query manipulation widgets [5] that help users modify existing utterances.

5.2.3 Integrated interaction experience

Another theme among the participants' comments regarding the advantages of multimodal interaction alluded to the notion of *fluidity* as characterized by Elmqvist et al. [19]. For instance, referring to the ability of being able to apply filters while interacting with nodes on the view, P16 said *"Generally I prefer touch but here speech was good because then I don't have to look through filters and I can just say it while moving points."* Although he was initially skeptical about multimodal input, during his interview, P10 said *"It was somehow less complex even though more interactions were added."* suggesting that the combination of modalities helped reduce the overall cognitive load. The comparatively fluid nature of multimodal input also led to participants perceiving themselves as being faster with the task even though the overall task completion times were comparable across the study interfaces. For instance, P1 said *"Having the combination was a lot easier to work with. Instead of having to find nodes and then highlight connections, I could do it in one command and then continue to interact with the graph."*

Motivated by such comments, we further reviewed the session videos to better understand what specifically about multimodal input evoked the feeling of fluidity. Based on our review, we attribute the fluidity of interaction in the multimodal interface to speech and touch facilitating *integrated interactions* [40] that are defined as *"interactions where a person's hands, tools, actions, interactions, visual response, and feedback are in situ where the data is visualized. That is, to effect an interaction, a person's attention is not drawn away from the visual representations of data in which they are interested."* Examples of integrated interactions during the study included applying a filter using speech while dragging nodes—not having to take eyes off the nodes in focus, or the ability to find nodes using speech without having to divert attention in order to type on a virtual keyboard that occluded the underlying visualization, among others. While these are seemingly straightforward interactions in isolation, they illustrate that the modalities together allowed participants to stay in the flow of their analysis rather than divert their attention to other user interface elements.

5.3 Effects of priming users with speech or touch

One of our study goals also was to observe if participants interacted differently with the multimodal system when they had prior experience with one of the two modalities (**RG2**). More specifically, we were curious if participants would continue to use the same modality and not use multimodal input? Would participants rely more heavily on the modality they first experienced? Would interaction patterns for these participants notably differ from those who only interact with the multimodal system? We were interested in these questions as the findings could challenge the need for multimodal input altogether. For instance, one possible outcome was that participants who interacted with the unimodal touch system (P1-P6) would continue to use only touch in the multimodal setting and similarly P7-P12 would use only speech input in the multimodal setting. Such a finding would suggest that people resort to what they know, refraining from learning new ways to interact with a system. Alternatively, it could also imply that adding input modalities may have limited (or no) benefits when users know how to work with a system using a specific modality.

We observed that participants who had prior experience working with the unimodal system (P1-P12) interacted with the multimodal system comparably to participants (P13-P18) who worked only with the multimodal system. When we explicitly asked participants if their experience with the first system affected their behavior, participants said that they used both modalities subconsciously and did not think about it until we asked them to reflect on it. For instance, P12 said *"Now that I think of it, not consciously but I did use speech to mostly narrow down to a subset and then touch to do more detailed tasks."* In fact, perhaps the single most important aspect that decided which modality would be used was the operation being performed. For instance, consider participants P1-P6 and the *find path* operation in Table 3. In this case, when interacting with the multimodal interface, all participants switched to using only speech commands even though they had all previously performed the operation using touch. Furthermore, this interaction pattern of primarily using speech for finding paths is comparable to the participants in the other conditions (P7-P18), suggesting a general mapping between the modality and task. Combined with the comments from the previous section, these observations suggest that people naturally adapted to using a new modality that was more suited for an operation even if they were experienced at performing the same operation with a different mode of input.

5.4 Operations and interaction patterns

Figure 3 summarizes the number of participants who used speech-only (🗣), touch-only (👆), or combined speech and touch input (🗣 + 👆) for performing common network visualization operations (**RG3**). Note that a single participant may have performed more than one type of interaction for an operation (e.g., as shown in Table 3, for finding paths, P13 used both speech-only and touch-only

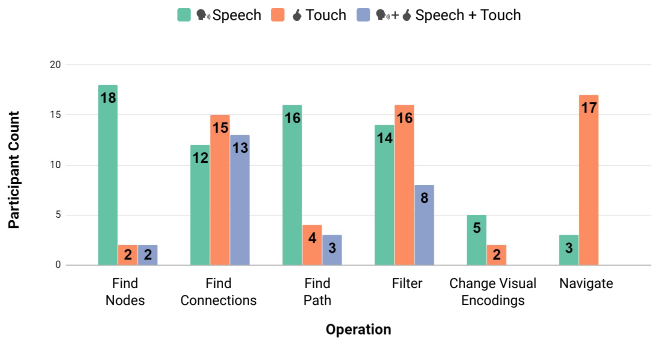


Fig. 3. Number of participants using different modes of input in the multimodal interface for each type of operation. Navigation was primarily performed using touch, whereas finding nodes and paths was largely performed through speech. Other operations had more variety in input patterns.

interactions). Our goal here is not to suggest one "best" input modality or interaction for an operation but rather to highlight the variety in patterns so future system designers can make more informed interaction design decisions. For this analysis, we only considered the 489 interactions with the multimodal interface when listing the actions since the unimodal interfaces did not give participants the option to choose their preferred style of input.

At a first glance, both Table 3 and Figure 3 suggest that participants largely performed operations using a single modality, infrequently *combining* modalities. However, it is important to also note that participants switched between modalities for different operations, using both modalities individually or together at some point during all sessions. Affirming to the myths of multimodal interaction [59], this switching between modalities highlights that the value of the multimodal interaction does not only come from modalities being combined but also stems from the availability of different modalities to perform varied operations amidst a task.

6 DISCUSSION

Our observations and participants' subjective feedback during the study also guided us to some higher-level findings and takeaways that we discuss below.

6.1 Dissecting the integration of speech and touch

Among the interactions that used both modalities, touch preceded speech in only 1/66 cases. This is in stark contrast to Oviatt et al.'s study investigating pen and speech-based interaction where 99% of the sequential multimodal constructions involved the use of pen before speech [46]. However, this pattern of speech preceding touch in visualization systems was also observed in a preliminary study with Orko [5]. Setlur et al. [17] also demonstrate such interactions in Eviza with queries like "*earthquakes with magnitudes between 4 and 6 here*" (where *here* is later specified through a lasso drawn using the mouse). We attribute this contrast between sequential use of modalities in recent studies of visualization systems to Oviatt et al.'s [46] seminal study with the QuickSet system to three factors. First, it may be a practical constraint arising from the modalities used and the study task. Specifically, Oviatt et al.'s study focused on the task of drawing on a map and used a digital pen as one of the input modalities. In contrast, recent studies of visualization systems (including ours) use either touch [5] or mouse [17], focusing on visual information seeking tasks. Our second hypothesis is that speech interfaces have become much more popular now and people more commonly treat speech as a primary input modality. The third reason for this behavior could be that in the context of a network visualization, the sequential integration of speech and touch best maps to Shneiderman's "*Overview first, zoom and filter, then details-on-demand*" mantra [60]. In other words, people get an overview by looking at the view, use speech to filter since it affords simultaneous specification of multiple filtering criteria, and then use touch to get details-on-demand due to the precision it affords.

6.2 Enhancing Discoverability

One hypothesis emerging from an input modality (or even multimodal input) being more natural for interaction is that it encourages users to explore a wider range of operations. We did not observe this behavior, however. In both the unimodal and the multimodal conditions, participants largely resorted to single step and fundamental network visualization operations (find, find connections, and path). Operations such as changing the color and size of nodes were also performed less frequently (7/18 sessions) even though they proved more effective than filtering for participants who leveraged them. However, the participants who did explore a wider range of operations were mostly those with prior experience using network visualization tools. While this can be attributed to the system implementation and tasks to an extent, it also highlights an important consideration regarding the system discoverability.

Discoverability applies to both aspects of discovering *what* operations are supported as well as *how* to perform them [61]. Particularly, since the system supported speech input, there was not always a one-to-one mapping between the GUI and possible operations (e.g. dropdowns for changing visual encodings showed up only when invoked via speech), potentially resulting in participants forgetting the operation. While the initial training and practice phase helped participants get acquainted with how to use the system, recollecting which operations could be performed during tasks was a common challenge. In fact, realizing he could have performed some initial tasks faster had he used the *find path* operation, during his interview, P13 said "*In fact, it would be helpful if I could tap on the nodes and the system could remind me of what I could do.*" hinting at the use of feedforward techniques [61], [62] to aid discoverability of speech input. To this end, one idea for future systems to explore may be to suggest contextually-relevant operations and corresponding speech commands based on the active state of the view and previously performed interactions (e.g., [63], [64], [65]). For example, if one issues a command to find two nodes, the system could suggest finding connections or finding the path as follow-up commands.

6.3 Exploring Proactive Behavior and Supplementing Visualizations with Textual Summaries

Recently, there has been a growing call for proactive system behavior in natural language interfaces (NLIs) for visualization [66], [67]. One minor way in which Orko incorporates such behavior is by dynamically reordering the charts in the summary container (Figure 1F) based on the user's most recent action [5]. During the study, all six participants in the speech-only condition (P7-P12) and five other participants (P1, P2, P14, P17, P18) explicitly commented on this behavior being helpful. Participants perceived the reordering of the summary charts as intelligent behavior and said that the charts often gave them answers for the questions they were thinking of posing next. For instance, P18 said *"I really liked the charts that came up on the right. They always seemed to be relevant to what I was thinking of at the time."* Based on these observations, an open research opportunity lies in exploring more proactive multimodal visualization interfaces that preempt user questions.

Given the availability of speech as an input modalitiy, unsurprisingly, participants expected the system to be more conversational and even "answer" questions. For instance, advocating for support for textual responses in addition to changes in the visualization, P4 said "*Working with the system for a while starts making you want to ask higher-level questions and get specific answers or summaries as opposed to just the visualization*." Based on such behavior, perhaps an interesting research opportunity is to explore multimodal network visualization systems that blend elements of question answering (QA) systems and also supplement visualizations with textual summaries. While recent work [68], [69] has begun exploring the idea of interactively linking text and network visualizations for presentation and storytelling, extending these ideas to support interactive network exploration is an open area for future work.

7 LIMITATIONS AND FUTURE WORK

Devices and Modalities. As common with laboratory experiments, our study had some limitations and constraints that must be considered when generalizing the results. We only considered speech- and touch-based input in the context of a single vertical display located at a touching distance from a user. Thus, building upon these results in different settings such as tablets or in AR/VR may require further testing. Similarly, considering additional modalities such as pen or gaze may also have a major impact on participants' interactions and is another factor that we did not consider in the presented study. We used Orko as our study interface since it was a minimalist system that supported core network visualization operations and was previously tested. However, changing the system interface or interactions (e.g., including more sophisticated multi-touch gestures as in [6], [7]) may impact the interactions and participant preferences.

Study Design and Datasets. As stated earlier, we did not counterbalance the order of the unimodal and the multimodal systems as participants interacting with the unimodal system after the multimodal system would already know and have experience with all the supported interactions. That said, reversing the order of the systems could allow understanding what aspects of multimodal interaction participants "missed" the most when working with the unimodal system. While this is fundamentally different from understanding the effects of priming participants with one modality (**RG2**), it is certainly an important extension to the current study. Furthermore, both the datasets used in the study were undirected, unipartite networks. Although the operations covered in the study are generalizable, formally verifying the results and understanding potential variations in interaction patterns for dynamic, multipartite, and/or directed networks is an open topic for future work.

Speech Recognition Errors. We used the standard speech recognition API [53] for web-based systems and trained it with the potential keywords (e.g., 'find', 'filter', 'path') and dataset-specific values to improve accuracy. Even so, there were 115 speech-to-text errors (excluded from Table 3 to avoid double counting interactions) across the 18 participants. Specifically, in the speech-only interface, across the six participants (P7-P12), there were 4-18 recognition errors (avg. 9), and in the multimodal interface, the number of errors ranged from 0-8 (avg. 3) across the 18 participants. These recognition errors led to some frustration among participants that may have impacted their interaction patterns. For example, after encountering recognition errors in the speech-only interface, P9 switched to using more touch interactions in the multimodal interface, saying that "*The voice recognition would have to be improved a lot before I can feel comfortable using voice alone to control the system*." While these are valid concerns, they are beyond the scope of our work and are imposed by the available technology. That said, such issues make the results more practically applicable by mirroring interactions with general voice user interfaces where incorrect recognition is the most common type of error [70]. In fact, these errors coupled with the earlier stated comments about complementing speech with touch further motivate the need to design multimodal systems that give users the ability to overcome errors of speech input or give them the freedom choose a different form of input.

8 CONCLUSION

We report a qualitative user study investigating how people interact with a network visualization tool using only touch, only speech, and a combination of the two. In addition to verifying that participants prefer multimodal input over unimodal input for visual network exploration, we discuss the different factors driving these preferences such as freedom of expression, the complementary nature of speech and touch, and integrated interactions afforded by the combination of the two modalities. We also report different interaction patterns participants employed to perform common network visualization operations, highlighting how people naturally adapt to new modalities that are more suited for an operation. We hope the observations from this study can help designers of future systems better understand user interaction preferences, ultimately resulting in the creation of multimodal visualization systems that are more expressive than current tools and that support a fluid interaction experience.

ACKNOWLEDGMENTS

This work was supported in part by the National Science Foundation grant IIS-1717111.

REFERENCES

[1] O. Mason and M. Verwoerd, “Graph theory and networks in biology,” *IET Systems Biology*, vol. 1, no. 2, pp. 89–119, 2007.

[2] J. Moody, “The structure of a social science collaboration network: Disciplinary cohesion from 1963 to 1999,” *American Sociological Review*, vol. 69, no. 2, pp. 213–238, 2004.

[3] T. L. Magnanti and R. T. Wong, “Network design and transportation planning: Models and algorithms,” *Transportation Science*, vol. 18, no. 1, pp. 1–55, 1984.

[4] M. Frisch, J. Heydekorn, and R. Dachselt, “Investigating multi-touch and pen gestures for diagram editing on interactive surfaces,” in *Proceedings of ITS '09*. ACM, 2009, pp. 149–156.

[5] A. Srinivasan and J. Stasko, “Orko: Facilitating multimodal interaction for visual exploration and analysis of networks,” *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 511–521, 2018.

[6] J. Thompson, A. Srinivasan, and J. Stasko, “Tangraphe: interactive exploration of network visualizations using single hand, multi-touch gestures,” in *Proceedings of AVI '18*. ACM, 2018, p. 43.

[7] S. Schmidt, M. A. Nacenta, R. Dachselt, and S. Carpendale, “A set of multi-touch graph interaction techniques,” in *Proceedings of ITS '10*. ACM, 2010, pp. 113–116.

[8] M. Cordeil, T. Dwyer, K. Klein, B. Laha, K. Marriott, and B. H. Thomas, “Immersive collaborative analysis of network connectivity: Cave-style or head-mounted display?” *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 441–450, 2017.

[9] C. Hurter, N. H. Riche, S. M. Drucker, M. Cordeil, R. Alligier, and R. Vuillemot, “Fiberclay: Sculpting three dimensional trajectories to reveal structural insights,” *IEEE transactions on visualization and computer graphics*, 2018.

[10] A. Drogemuller, A. Cunningham, J. Walsh, M. Cordeil, W. Ross, and B. Thomas, “Evaluating navigation techniques for 3d graph visualizations in virtual reality,” in *2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA)*. IEEE, 2018, pp. 1–10.

[11] T. Sun, Y. Ye, I. Fujishiro, and K.-L. Ma, “Collaborative visual analysis with multi-level information sharing using a wall-size display and see-through hmds,” in *Proceedings of PacificVis '19*. IEEE, 2019, pp. 11–20.

[12] U. Kister, K. Klamka, C. Tominski, and R. Dachselt, “Grasp: Combining spatially-aware mobile devices and a display wall for graph visualization and interaction,” in *Computer Graphics Forum*, vol. 36, no. 3. Wiley Online Library, 2017, pp. 503–514.

[13] A. Lee, D. Archambault, and M. Nacenta, “Dynamic network plaid: A tool for the analysis of dynamic networks,” in *Proceedings of CHI '19*. New York, NY, USA: ACM, 2019, pp. 130:1–130:14. [Online]. Available: http://doi.acm.org/10.1145/3290605.3300360

[14] R. Langner, U. Kister, and R. Dachselt, “Multiple coordinated views at large displays for multiple users: Empirical findings on user behavior, movements, and distances,” *IEEE transactions on visualization and computer graphics*, 2018.

[15] B. Lee, A. Srinivasan, J. Stasko, M. Tory, and V. Setlur, “Multimodal interaction for data visualization,” in *Proceedings of AVI '18*. ACM, 2018, p. 11.

[16] J.-F. Kassel and M. Rohs, “Valletto: A multimodal interface for ubiquitous visual analytics,” in *Extended Abstracts of CHI '18*. ACM, 2018, p. LBW005.

[17] V. Setlur, S. E. Battersby, M. Tory, R. Gossweiler, and A. X. Chang, “Eviza: A natural language interface for visual analysis,” in *Proceedings of UIST '16*. ACM, 2016, pp. 365–377.

[18] T. Gao, M. Dontcheva, E. Adar, Z. Liu, and K. G. Karahalios, “Datatone: Managing ambiguity in natural language interfaces for data visualization,” in *Proceedings of UIST '15*. ACM, 2015, pp. 489–500.

[19] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, “Fluid interaction for information visualization,” *Information Visualization*, vol. 10, no. 4, pp. 327–340, 2011.

[20] B. Lee, P. Isenberg, N. H. Riche, and S. Carpendale, “Beyond mouse and keyboard: Expanding design considerations for information visualization interactions,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2689–2698, 2012.

[21] S. M. Drucker, D. Fisher, R. Sadana, J. Herron *et al.*, “Touchviz: a case study comparing two interfaces for data analytics on tablets,” in *Proceedings of CHI '13*. ACM, 2013, pp. 2301–2310.

[22] R. Sadana and J. Stasko, “Designing and implementing an interactive scatterplot visualization for a tablet computer,” in *Proceedings of AVI '14*. ACM, 2014, pp. 265–272.

[23] B. Lee, M. Brehmer, P. Isenberg, E. K. Choe, R. Langner, and R. Dachselt, “Data visualization on mobile devices,” in *Extended Abstracts of CHI '18*. ACM, 2018, p. W07.

[24] M. Brehmer, B. Lee, P. Isenberg, and E. K. Choe, “Visualizing ranges over time on mobile phones: a task-based crowdsourced evaluation,” *IEEE transactions on visualization and computer graphics*, 2018.

[25] B. Lee, R. H. Kazi, and G. Smith, “Sketchstory: Telling more engaging stories with data through freeform sketching,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 19, no. 12, pp. 2416–2425, 2013.

[26] E. Zgraggen, R. Zeleznik, and S. M. Drucker, “Panoramicdata: Data analysis through pen & touch,” *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 2112–2121, 2014.

[27] M. Cordeil, A. Cunningham, T. Dwyer, B. H. Thomas, and K. Marriott, “Imaxes: Immersive axes as embodied affordances for interactive multivariate data visualisation,” in *Proceedings of UIST '17*. ACM, 2017, pp. 71–83.

[28] J. Wagner Filho, C. M. Freitas, and L. Nedel, “Virtualdesk: A comfortable and efficient immersive information visualization approach,” in *Computer Graphics Forum*, vol. 37, no. 3. Wiley Online Library, 2018, pp. 415–426.

[29] P. Isenberg and D. Fisher, “Collaborative brushing and linking for co-located visual analytics of document collections,” *Computer Graphics Forum*, vol. 28, no. 3, pp. 1031–1038, 2009.

[30] D. Baur, B. Lee, and S. Carpendale, “Touchwave: kinetic multi-touch manipulation for hierarchical stacked graphs,” in *Proceedings of ITS '12*. ACM, 2012, pp. 255–264.

[31] R. Sadana and J. Stasko, “Designing multiple coordinated visualizations for tablets,” in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 261–270.

[32] K. Cox, R. E. Grinter, S. L. Hibino, L. J. Jagadeesan, and D. Mantilla, “A multi-modal natural language interface to an information visualization environment,” *International Journal of Speech Technology*, vol. 4, no. 3-4, pp. 297–314, 2001.

[33] Y. Sun, J. Leigh, A. Johnson, and S. Lee, “Articulate: A semi-automated model for translating natural language queries into meaningful visualizations,” in *International Symposium on Smart Graphics*. Springer, 2010, pp. 184–195.

[34] J. Aurisano, A. Kumar, A. Gonzalez, J. Leigh, B. DiEugenio, and A. Johnson, “Articulate2: Toward a conversational interface for visual data exploration,” in *IEEE Visualization*, 2016.

[35] E. Hoque, V. Setlur, M. Tory, and I. Dykeman, “Applying pragmatics principles for interaction with visual analytics,” *IEEE transactions on visualization and computer graphics*, vol. 24, no. 1, pp. 309–318, 2018.

[36] B. Yu and C. T. Silva, “FlowSense: A Natural Language Interface for Visual Data Exploration within a Dataflow System,” *IEEE transactions on visualization and computer graphics*, vol. 26, no. 1, pp. 1–11, 2019.

[37] “IBM Watson Analytics,” http://www.ibm.com/analytics/watson-analytics/.

[38] “Microsoft Power BI,” https://powerbi.microsoft.com/en-us/.

[39] “Tableau Ask Data,” https://www.tableau.com/about/blog/2018/10/announcing-20191-beta-96449.

[40] J. Walny, B. Lee, P. Johns, N. H. Riche, and S. Carpendale, “Understanding pen and touch interaction for data exploration on interactive whiteboards,” *IEEE Transactions on Visualization and Computer Graphics*, vol. 18, no. 12, pp. 2779–2788, 2012.

[41] W. Büschel, P. Reipschläger, R. Langner, and R. Dachselt, “Investigating the use of spatial interaction for 3D data visualization on mobile devices,” in *Proceedings of ISS*. ACM, 2017, pp. 62–71.

[42] L. Besançon, P. Issartel, M. Ammi, and T. Isenberg, “Hybrid tactile/tangible interaction for 3D data exploration,” *IEEE transactions on visualization and computer graphics*, vol. 23, no. 1, pp. 881–890, 2016.

[43] L. Besançon, M. Sereno, M. Ammi, L. Yu, and T. Isenberg, “Hybrid Touch/Tangible Spatial 3D Data Selection,” *Computer Graphics Forum*, vol. 38, no. 3, pp. 553–567, 2019.

[44] R. Pausch and J. H. Leatherby, “An empirical study: Adding voice input to a graphical editor,” in *Journal of the American Voice Input/Output Society*. Citeseer, 1991.

[45] L. Karl, M. Pettey, and B. Shneiderman, “Speech-activated versus mouse-activated commands for word processing applications: An empirical evaluation,” *International Journal of Man-Machine Studies*, vol. 39, no. 4, pp. 667–687, 1993.

[46] S. Oviatt, A. DeAngeli, and K. Kuhn, “Integration and synchronization of input modes during multimodal human-computer interaction,” in *Referring Phenomena in a Multimedia Context and their Computational Treatment*. Association for Computational Linguistics, 1997, pp. 1–13.

[47] I. Herman, G. Melançon, and M. S. Marshall, “Graph visualization and navigation in information visualization: A survey,” *IEEE Transactions on visualization and computer graphics*, vol. 6, no. 1, pp. 24–43, 2000.

[48] T. Von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. J. van Wijk, J.-D. Fekete, and D. W. Fellner, “Visual analysis of large graphs: state-of-the-art and future research challenges,” in *Computer graphics forum*, vol. 30, no. 6. Wiley Online Library, 2011, pp. 1719–1749.

[49] F. Beck, M. Burch, S. Diehl, and D. Weiskopf, “A taxonomy and survey of dynamic graph visualization,” in *Computer Graphics Forum*, vol. 36, no. 1. Wiley Online Library, 2017, pp. 133–159.

[50] B. Lee, C. Plaisant, C. S. Parr, J.-D. Fekete, and N. Henry, “Task taxonomy for graph visualization,” in *Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization*. ACM, 2006, pp. 1–5.

[51] J. Pretorius, H. C. Purchase, and J. T. Stasko, “Tasks for multivariate network analysis,” in *Multivariate Network Visualization*. Springer, 2014, pp. 77–95.

[52] B. Saket, P. Simonetto, and S. Kobourov, “Group-level graph visualization taxonomy,” *arXiv preprint arXiv:1403.7421*, 2014.

[53] https://developer.mozilla.org/en-US/docs/Web/API/Web_Speech_API, 2019.

[54] N. Bush, R. Wallace, T. Ringate, A. Taylor, and J. Baer, “Artificial Intelligence Markup Language (AIML) Version 1.0. 1,” *ALICE AI Foundation Working Draft*, 2001.

[55] A. Singhal *et al.*, “Modern information retrieval: A brief overview,” *IEEE Data Eng. Bull.*, vol. 24, no. 4, pp. 35–43, 2001.

[56] Z. Wu and M. Palmer, “Verbs semantics and lexical selection,” in *Proceedings of the 32nd annual meeting on Association for Computational Linguistics*, 1994, pp. 133–138.

[57] V. Yoghourdjian, D. Archambault, S. Diehl, T. Dwyer, K. Klein, H. C. Purchase, and H.-Y. Wu, “Exploring the limits of complexity: A survey of empirical studies on graph visualisation,” *Visual Informatics*, vol. 2, no. 4, pp. 264–282, 2018.

[58] P. R. Cohen, M. Dalrymple, D. B. Moran, F. Pereira, and J. W. Sullivan, “Synergistic use of direct manipulation and natural language,” in *ACM SIGCHI Bulletin*, vol. 20, no. SI. ACM, 1989, pp. 227–233.

[59] S. Oviatt, “Ten myths of multimodal interaction,” *Communications of the ACM*, vol. 42, no. 11, pp. 74–81, 1999.

[60] B. Shneiderman, “The eyes have it: A task by data type taxonomy for information visualizations,” in *Visual Languages, 1996. Proceedings., IEEE Symposium on*. IEEE, 1996, pp. 336–343.

[61] D. Norman, *The design of everyday things: Revised and expanded edition*. Constellation, 2013.

[62] J. Vermeulen, K. Luyten, E. van den Hoven, and K. Coninx, “Crossing the bridge over Norman’s Gulf of Execution: revealing feedforward’s true identity,” in *Proceedings of CHI’ 13*. ACM, 2013, pp. 1931–1940.

[63] E. Corbett and A. Weber, “What can i say?: addressing user experience challenges of a mobile voice user interface for accessibility,” in *Proceedings of MobileHCI ’16*. ACM, 2016, pp. 72–82.

[64] A. Furqan, C. Myers, and J. Zhu, “Learnability through adaptive discovery tools in voice user interfaces,” in *Proceedings of the CHI ’17*. ACM, 2017, pp. 1617–1623.

[65] A. Srinivasan, M. Dontcheva, E. Adar, and S. Walker, “Discovering natural language commands in multimodal interfaces,” in *Proceedings of IUI ’19*. ACM, 2019, pp. 661–672.

[66] A. Srinivasan and J. Stasko, “Natural language interfaces for data analysis with visualization: Considering what has and could be asked,” in *Proceedings of EuroVis*, vol. 17, 2017, pp. 55–59.

[67] M. Tory and V. Setlur, “Do what i mean, not what i say! design considerations for supporting intent and context in analytical conversation,” in *IEEE Conference on Visual Analytics Science and Technology*, 2019.

[68] S. Latif and F. Beck, “Vis author profiles: Interactive descriptions of publication records combining text and visualization,” *IEEE transactions on visualization and computer graphics*, vol. 25, no. 1, pp. 152–161, 2018.

[69] S. Latif, K. Su, and F. Beck, “Authoring combined textual and visual descriptions of graph data,” in *EuroVis (Short Papers)*, 2019.

[70] C. Myers, A. Furqan, J. Nebolsky, K. Caro, and J. Zhu, “Patterns for how users overcome obstacles in voice user interfaces,” in *Proceedings of CHI ’18*. ACM, 2018, p. 6.

Ayshwarya Saktheeswaran is a Master’s student in Human-Computer Interaction at the Georgia Institute of Technology. Her research interests include information visualization, human-computer interaction, and educational technology.

Arjun Srinivasan is Ph.D. student in Computer Science at the Georgia Institute of Technology. His current research focuses on the design of intelligent and expressive visualization tools that combine multimodal input (e.g., speech and touch) and mixed-initiative interface techniques for human-data interaction.

John Stasko is a Regents Professor in the School of Interactive Computing and the Director of the Information Interfaces Research Group at the Georgia Institute of Technology. His research is in the areas of information visualization and visual analytics, approaching each from a human-computer interaction perspective. John was named an ACM Distinguished Scientist in 2011 and an IEEE Fellow in 2014. He received his PhD in Computer Science at Brown University in 1989.