Comparative Evaluation of Node-Link and Sankey Diagrams for the Cyber Security Domain

Rotem Blinder, Ofer Biller, Adir Even, Oded Sofer, Noam Tractinsky, Joel Lanir, Peter Bak

Comparative Evaluation of Node-Link and Sankey Diagrams for the Cyber Security Domain

Rotem Blinder[1], Ofer Biller[2], Adir Even[1], Oded Sofer[2], Noam Tractinsky[1], Joel Lanir[3], and Peter Bak[4]

[1] Ben-Gurion University of the Negev, Beer Sheva, Israel
[2] IBM Security, Beer Sheva, Israel
[3] University of Haifa, Haifa, Israel
[4] IBM research Haifa lab, Haifa, Israel

Abstract. Visualization tools are critical components of cyber security systems allowing analyzers to better understand, detect and prevent security breaches. Security administrators need to understand which users accessed the database and what operations were performed in order to detect irregularities. The current work compares the Sankey diagram with the more commonly used node-link diagram as an alternative visualization technique for cyber security tasks in a controlled experiment. The results indicate, that the Sankey tool showed a consistent advantage in task completion time and was more effective (measured by the percent of correct answers) in synoptic tasks, while the Node-link diagram was more effective in basic, elementary tasks. Further results revealed that performance had only a small effect on user satisfaction and preferences. Our results suggest that the Sankey tool may be a viable option for cyber security visualization tools and strengthens the need to provide personalized visualization tools based on user preferences.

Keywords: Cyber Security · Visualization · Sankey diagram.

1 Introduction

The growing threats to cyber security have motivated the search for solutions that detect, prevent, and minimize the damage associated with security breaches and cyber-attacks on data resources and information systems. Visualizing cyber security-related data suggest using the perceptual capabilities of humans in order to complement machine analysis and enable better analytical support in understanding this complex data. Studies show that effective visualization tools can help security analysts identify hostile activity and analyze its characteristics, thereby significantly increasing the safety level of data [**?**,**?**,**?**].

The design of effective cyber security visualization tools depends on the type of data collected, the tasks users need to perform using the visualization, and the design decisions of the visualization solutions that aim to meet these requirements. One of the most common tasks in cyber security is trying to understand database access [**?**,**?**]. Modern database servers log users' activity to allow automatic or manual detection of violations either in real-time or on log history. Administrators need to understand which

users accessed what table, and what type of operations were performed. However, this may not be a simple task as users are usually described by their IP-address, user name, operation system and other attributes, and database access is described by different database systems and views that reference multiple tables. System administrators are left with the difficult task of looking for irregularities and possible security violations within this data.

One of the most common visualizations used in cyber security, and especially when analyzing database access, is the node-link diagram [**?**]. Node-link diagrams, usually layed out using a force-directed alrorithm (as was done in our study), enable the projection of the complex interlinking structure of the users and databases access graph onto a two-dimensional screen by applying the right layout algorithms [**?**]. While the node-link diagram is widely used in cyber security it does have some disadvantages. The readability of node-link diagrams has been investigated and found to be often limited and too complex, especially when the number of nodes and links increase [**?**].

The Sankey diagram is a type of flow chart in which the width of the stream reflects the quantity of the flow [**?**]. Similar to a node-link diagram, Sankey diagrams show a directional relationship between different entities. However, the largest difference is that Sankey diagrams are constrained in their layout, grouping the nodes into layers displayed from left to right. In some versions of Sankey diagrams, the nodes can be grouped into semantic groups that depict the layers of the chart. The layout constraints, in form of clustering, has been proven to provide an advantage to graph readability for a number of tasks [**?**].

We posit that for database access analysis, Sankey diagrams can be a better choice than node-link diagrams. In order to assess the possible use of the Sankey diagram for cyber security visualizations, we compared its use with the more traditional node-link diagram by conducting an empirical quantitative user study on a large number of participants. We used real-world security data, asking participants to complete a set of tasks following a formal task taxonomy. We complemented the quantitative analysis with interviews with domain experts. Results indicate that the Sankey diagram was more effective (measured in completion accuracy) in general, synoptic tasks, while the node-link diagram was more effective in more basic, elementary tasks. In terms of user efficiency (measured by task completion time) results show that the Sankey diagram was overall more efficient than the node-link diagram. Finally, results suggest that performance had only a small effect on user preferences. We discuss the implications of these results and provide guidelines for the design of cyber security visualization tools.

2 Related Work

Cyber security visualization is a well-established research field. Previous efforts created many tools and techniques to support and improve cyber security tasks. Moreover, multiple surveys provide comprehensive reviews and more details on existing visualization techniques and systems for the cyber security domain [**?**,**?**,**?**,**?**]. However, while many tools and techniques exist, very few works have performed usability studies with users, and evaluations if they exist, are usually done per system in an ad-hoc and unsystem-

atic way [?,?]. There is a clear lack of emperical evaluations that aim to add theoretical knowledge to the field [?].

The node-link diagram is often used in the cyber security domain for the visualization of packet traces, intrusion alerts and database access [?]. The visual language of node-link diagrams can help to observe global patterns of connectivity [?], spot the presence of unexpected connections, and study trivial correlations between topology and the properties of nodes and edges through visual features. The topic of network and graph visualization is well-studied and has become a commodity in cyber security applications[?,?]. A general overview of node-link diagrams is beyond the scope of this paper. We refer readers to some of the available surveys in this field for in-depth information [?,?,?].

The Sankey diagram is a counterpart to this visualization. It depicts a flow from one set of values to another. The elements being connected are called nodes and the connections are called links. Node height and link width usually denote the volume of the flow. Sankeys are best used to show a many-to-many mapping between two domains or multiple paths through a set of stages. The interactive Sankey diagram allows selection, rearrangement, and filtering to select a specific category, and to see the associated inflows and outflows [?,?]. The Sankey technique is widely used in other domains, such as energy or water management, health-related applications and event sequence data analysis [?,?,?]. Although it is rarely used in the cyber security domain, some commercial systems, such as the IBM Security Guardium system, have started using it for various tasks. We propose that Sankey diagrams can be useful in depicting the *flow of information* from users to database tables and vice versa when monitoring and detecting anomalies in database access.

Evaluating visualization techniques for applicability is a major challenge and an important research direction in general [?,?]. Practices and guidelines for conducting valid and repeatable empirical evaluation have been proposed in [?,?,?]. Specifically, for graphs and networks, Huang [?] provides a comprehensive overview of measuring the effectiveness of graphs under different conditions of cognitive load. Usability studies involving Sankey and Node-link diagrams were performed in [?]. Their work focused on users' ability to create such diagrams programmatically using the *Prefuse* framework in an efficient way. Specifically, the Sankey diagram has been proved efficient in contrast to other visualization frameworks in [?]. In addition, the Sankey diagram was used as the main tool in the *Outflow* system for investigating event sequence data [?]. A user evaluation showed that users were able to learn how to use the diagram easily with little training and perform a range of tasks both accurately and rapidly.

Despite efforts to evaluate and compare many information visualization techniques, we did not find a systematic evaluation of performance between node-link and Sankey diagrams. The current study focuses on this issue, given the practical importance of such a comparison for the development of visualizations for cyber security systems.

3 Method

We postulate that performance and subjective evaluations depend on the type of visualization tool used and that these effects could be mitigated by the type of tasks in which

the users engage. We thus conducted a controlled laboratory experiment to test the effects of the two visualizations (Sankey and node-link) on user effectiveness, efficiency, satisfaction, and preference. To complement the controlled experiment, we also conducted interviews with security analysts, asking their opinion on the two visualization methods in relation to the task of understanding user access to a database.

3.1 Data preparation

We extracted real log files from a large data security platform of database access information in a large organization containing user information, details of the database accessed, and a timestamp. To create the visualizations, we processed, cleaned, and summarized these information sources in the following form:

Who performed the activity? This includes the database *user*, and the *IP address* of the source, among many other related attributes (which were not included in this research).

What activity was performed? This indicates the type of activity; (*verb*) such as selection, modification, or others. There was a very limited variance in the data on activity types, the most frequent activity being "selection" and then "execution" for the period in which we investigated the data.

On what was this activity performed? Contains the database system, the *database* and the table or view that was accessed.

When was it performed? This shows the time of the activity, which was only used for filtering purposes. We filtered the data, limiting the time span to one specific hour of database access information.

How many of these activities were performed by the user? This was computed by counting the access requests within the selected timeframe. This was aggregated over the time span.

These numbers and settings reflect a real-world scenario, and were used in the empirical evaluation.

3.2 Visual Design

We encoded the above information using two different techniques: node-link diagrams and Sankey diagrams. Care was taken to ensure that the same information is represented using only different channels and marks.

Node-Link Diagram Figure **??** represents a one hour time span for activity overview using the node-link diagram. To construct the node-link diagram, objects of the information layers are encoded as symbols (IP as a computer with "IP" on its screen, database as a disk-symbol, user as a person with a database symbol, and tables as grid-icons). Lines show the connection between the objects. Line thickness and symbol size encode the number of database transactions conducted. The type of activity, which we refer to as "verb", is depicted as a separate node type with its own icon. For interaction, we

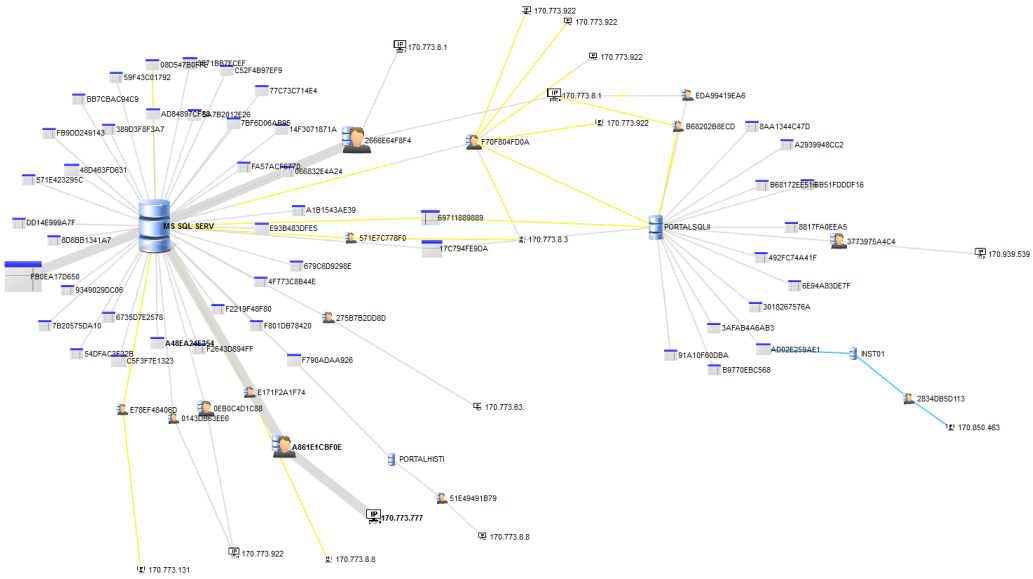

Fig. 1. Node-link diagram shows four selected layers of information: IP address, user, database, and database-table

supported selection and tooltips. When an object was selected, all corresponding connections are highlighted, and unselected objects fade out. When an object is hovered over, a tooltip including the name and number of transactions is presented.

The resulting visual encoding reflect the data and lead to a comprehensive network of activities in the system. The view simultaneously shows the topology of activities (who accesses what database), and specific details of each user's access patterns. As there are alternative encodings possible, we verified these with security domain experts, who confirmed that this reflects the common state of node-link diagrams in security systems. Study participants were able to investigate the activities of database users by selecting an icon and consecutively highlighting all corresponding connections. For demonstration purposes, Figure **??** shows activities on an MS SQL server with two major users (connected with thick lines to the server) and one high frequency table access (also connected to the server).

Sankey Diagram Figure **??** shows the Sankey diagram created on the database access information. The Sankey diagram uses a horizontal positioning for the four information layers; IP, users, database, and tables in a left-to-right order (the same layers as in the node-link diagram). Objects corresponding to one of these information layers are placed in a vertical position. Information layers are given a label on the horizontal position. Objects are represented as rectangular nodes, and connections between nodes as splines. The height of the node and the width of the lines encode the number of transactions.

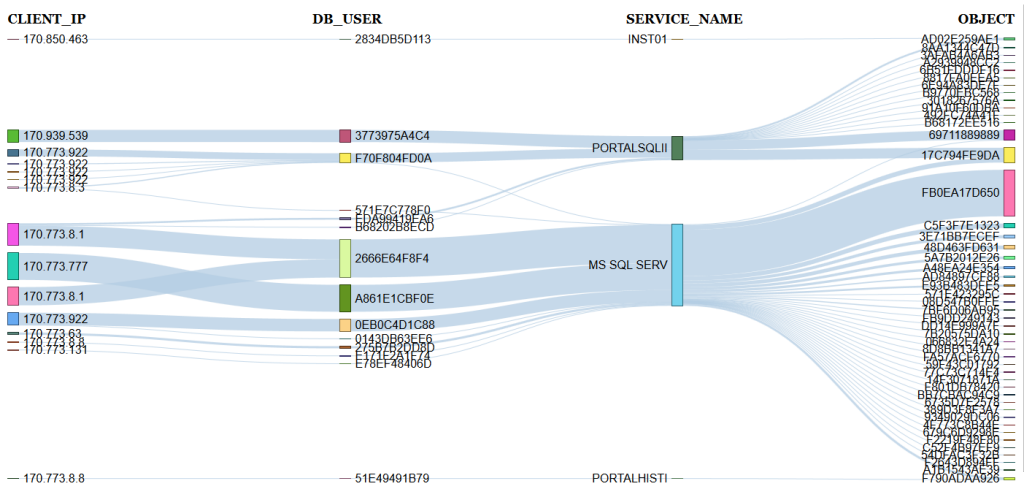

Fig. 2. Sankey diagram shows four selected layers of information: IP address, user, database, and database-table

Color distinguishes nodes from each other within a layer. The activity type (verb), was added as one of the information layers, connecting the database objects with the users. For interaction, selection and tooltips were used exactly the same way as in the node-link diagram.

The resulting image in Figure **??** shows the *flows* of data from the IP addresses and the users to the databases and tables. Study participants could point out databases or tables that are used more frequently than others, and select corresponding users with either high or low transaction counts.

Compared to the node-link diagram, the Sankey diagram has a much more constrained layout, due to the horizontal fixed positions of the information layers. As a result, in the Sankey diagram, users have to search horizontally for an information layer, and then vertically for a particular object. In contrast, in the node-link diagram, objects can appear at any position in the display and can only be recognized by the icons.

In node-link diagrams, users of the real-world systems could usually reposition items and select different information layers. For the Sankey diagram, users of real-world systems could usually change the horizontal position of the information layers. However, to avoid confounding, the software in the experiment allowed participants to only select and hover over objects in both chart types.

3.3 Participants

We had 135 third-year undergraduate engineering students participate in the experiment. All participants were enrolled in a database class and received course credit for their participation.

3.4 Procedure and Design

Participants were assigned randomly to one of two groups: treatment and control. In the treatment group, 77 participants used the two visualization tools mentioned above to address 14 tasks. The control group was used to validate the benefits of the two visualization tools compared to the use of a standard spreadsheet. Thus, in the control group, 58 participants used an Excel worksheet with the raw data to perform the same tasks.

At the start of the experiment, the participants were given a written description of the experimental purpose and signed a consent form. The experimenter then introduced and demonstrated the two visualization tools. Next, participants performed the tasks using two sets of structurally equivalent tasks in two consecutive blocks. In each block, the participants interacted with one of the two visualization tools (Sankey diagram or Node-Link diagram). The order in which the visualization tools were used was counterbalanced.

Each block began with four training tasks, to acquaint the participants with the visualization tool and the tasks. Next, they were presented with 14 experimental tasks. Participants were asked to work as quickly and accurately as possible. They answered each task by selecting from a predefined list of alternative answers. After choosing an answer, the participant pressed the "Next" button to move to the next task. Completion time and the selected answer for each task were recorded. At the end of the second session, participants responded to items asking about their satisfaction with each tool (using a 1 to 5 Likert scale) and indicated which of the tools they preferred. Each experimental block (i.e., working with one visualization tool) lasted between 30 and 40 minutes.

The control group received the same data sets and the same training and experimental tasks as the two visualization groups. The control group performed the tasks using the raw data set in an Excel worksheet, without the aid of a visualization tool.

All sessions were conducted in a quiet lab equipped with an Intel Core i5-4570 3.2 GHz computers and 24" monitors with a resolution of 1920x1080 pixels.

3.5 Experimental Tasks

We classified the experimental tasks according to the model proposed by Andrienko et al. [**?**], distinguising between *elementary* and *synoptic* tasks. Elementary tasks are defined as simple, basic tasks that usually require a single or only few basic operations (such as identify, locate or compare) to complete. Synoptic tasks are more general, more complex and usually require multiple operations. Each task question had a different number of response options varying from 3 to 10 options. We created two structurally equivalent task sets, each for use with a different visualization tool. The 14 tasks included 8 elementary tasks and 6 synoptic tasks. Our data analysis concentrated on this low level classification. Table **??** presents the tasks and provides additional information about other task attributes according the classification of Andrienko et al. [**?**].

Table 1. List of experimental tasks. The same tasks were used for both visualization tools, with different attribute values for each tool.

	Task	Task Type	Other attributes
1	What was the number of transactions of database "Grades" at 15:00?	Elementary	Ov, L
2	Which ClientIPs did the User "Yotam" use at 15:00?	Elementary	Ov, L
3	Which User had the highest number of transactions at 15:00?	Elementary	Ov, IL
4	When was the lowest number of transactions of database "Students" ?	Elementary	Ov, IL
5	Did ClientIP "773.922.841" use more Verbs than ClientIP "773.922.858" at 15:00? How many more?	Elementary	Ov, C
6	For DB "Grades" and "Students", which performed more diverse activities of different Users at 15:00?	Elementary	Ou, C
7	Mark 2 Verbs on which number of transactions of DBUser "Aviv" was higher than DBUser "Nimrod" at 15:00.	Elementary	Ov, RS
8	Find the time (hour) which DBUser "Nimrod" used less than 3 Client IPs	Elementary	Ou, RS
9	What was Database "Lecturers" trend between 15:00 - 16:00?	Synoptic	Ou, PI
10	For 16:00. Which user used the database (DB Name), Verb, and Client IP that no other User used?	Synoptic	OU, PI
11	What is the most common Verb on database "Grades" at 15:00?	Synoptic	Ov, BC
12	Which User used the most diverse DBNames at 15:00?	Synoptic	Ou, BC
13	For Users "Shlomi" and ""Yotam", which one has the largest growth rate of transactions between 13:00 and 14:00?	Synoptic	Ou, RS
14	Which Verb increased the most from 16:00-17:00?	Synoptic	Ou, RS

Legend: BC=behavior comparison, C=comparison, IL=inverse lookup, L=lookup, Ou=Outliers; Ov=Overview, PI= pattern identification, RS=relation seeking, RS=relation seeking.

3.6 Datasets

The source of data for the experiments was a cyber security system installed at a large company, with data gathered during a working day in 2016. The description given to participants in the experiment was that the data belonged to students in a "Databases" course, who check their personal data in the university information system. The students access the system's databases and carry out various activities. Each access includes the student's username ('User') and receives a 'ClientIP'. Other data included the name of the action performed by the user ('Verb'), for example- Select, Execute, Update, Truncate, Create, If, and Delete. The data also showed the database 'DBName' used by the students, for example- Grades, Students, Lecturers, Courses, Faculties, and Departments. The attribute values were replaced to match the cover story. For example, the 'ServiceName' "MS SQL SERVER" was changed to "Grades", the 'DBUser' "F70F804FD0A" was replaced by "John". To reduce carry over due to task familiarity between the two experimental blocks, we used different values for the attributes in each block. For example, the 'User' named "John" in the first block was presented with another name in the second block.

3.7 Expert interviews

To complement the results of the controlled experiment, we conducted semi-structured interviews with seven database administrators working in a big software company. We used a list of set questions that were elaborated on according to each interview. We asked their opinion on the suitability of the two visualization methods in relation to database access security tasks. Each expert was asked to work with both the Sankey and the node-link diagram on several tasks using a real-world dataset. The dataset shown to the experts was not the same as used in the quantitative experiment, but rather was one that was not constrained by the needs of a formal user study (e.g. larger, and more

representative of a real system). Tasks included identification and pattern definition for Users, Databases and Verbs separately, and in a pair-wise combination. Experts were asked to verbalize their thoughts (Thinkaloud) when completing the tasks, and were interviewed at the end of the session regarding their opinions.

4 Results

All participants completed the assignments successfully. The distribution of correct answers ranged from 17 to 28 (best possible result) with an average of 24.8 and a median of 25. The minimal completion time of all tasks combined was 794sec and the maximal time was 2,332sec, with a mean of 1,383sec and a median of 1,353sec.

4.1 Data Cleaning

The criterion for discarding outlier data was set in terms of task completion times. Outliers were defined as answers whose task completion times were 10 times smaller or greater than the sample's median completion time on that specific task. We found 7 such cases, distributed over 4 individuals. We set those times to missing values. In addition, examination of individual tasks identified 1 specific task in which performance measures differed greatly between the 2 visualization tools. The task (Task 12, see Table 1), was the only task in our battery that was classified as a combination of behavior comparison and outlier detection according to the low-level task taxonomy of [**?**]. It took much longer to complete using the Sankey tool (mean=108.9, median=103.8, SD=55. vs. mean=59.4, median=50.3, SD=28.8 in node-link) and answers were considerably less accurate (M=.57, SD=.50 in Sankey vs. M=.88, SD=32 , in node-link). Both differences were highly significant (paired-sample t(75)=6.88, $p < .001$ for completion time and t(76)=5.03, $p < .001$ for correctness). Due to the clear advantage of node-link in performing this task, we considered it separately from the other 13 tasks.

4.2 Main Analysis

Table 2 summarizes the experimental groups and the associated demographics. We analyzed the data using R Studio 1.1.383.

Table 2. Experimental groups and demographic data

Group	Sample Size (M/F)	Age Mean/SD
Sankey first	38 (11/27)	24.7/1.2
Node Link first	39 (11/28)	24.9/1.2
Control (Excel)	58 (33/25)	25.8/1.1
Overall	135 (55/80)	24.8/1.7

We first examined the potential effects of the demographic variables. Age was very weakly correlated with the three dependent variables ($r < .1$ for all variables). Separate

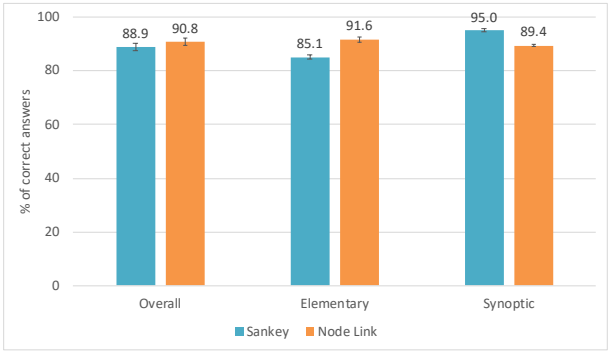

Fig. 3. Average effectiveness scores (percent correct answers with standard-error) of all tasks, elementary tasks only (8 tasks) and synoptic tasks only (5 tasks).

t-tests for differences between males and females on all three dependent variables were insignificant ($p > .47$ in all tests). Therefore, we did not consider those control variables in further analyses.

Effectiveness and Efficiency Compared to the Excel Baseline We performed a one-way ANOVA with three levels (Sankey, Node-Link, Excel) for effectiveness and efficiency results. Both analyses were significant (F(2,209)=15.31, $p < .001$ for effectiveness, F(2,209)=296, $p < .001$ for efficiency). Post-hoc analyses (Tukey HSD) revealed that, on average and over all tasks, the Excel group performed substantially lower on both measures. This finding established the superiority of the visualization tools over the default format. Therefore, in the subsequent analyses we focused on comparing the two visualization tools.

Effectiveness and Efficiency without Task 12 Figures **??** and **??** present the overall effectiveness and efficiency results, as well as results broken down by task type (elementary vs. synoptic) in each visualization tool. We analyzed the data using separate two-way (visualization tool and task type) within-subjects analyses of variance with effectiveness and efficiency as dependent variables. The analysis of the overall effectiveness score (percent of correct answers) found no difference between the groups (F(1,76) = .115, p=.12). There was a main effect for Task Type. Synoptic tasks had more correct answers than elementary tasks (F(1,76)=8.53, p =.005). However, this result was qualified by a significant Tool x Task Type interaction (F(1,76) = 28.10, $p < .001$). The interaction stemmed from a higher percentage of correct answers to the elementary tasks in node-link (paired-sample t(76)=4.27, $p < .001$) and a higher percentage of correct answers to the synoptic tasks in Sankey (t(76)=3.31, p= .001).

A two-way within-subjects analyses of variance with efficiency (task completion time) as the dependent variable found the main effects to be visualization tool and task type (F(1,76)=.12.82, p=.001 and F(1,76)=15.43, $p < .001$, respectively). There was no interaction effect (F(1,76)=1.92,p=.17). Participants answered more quickly with

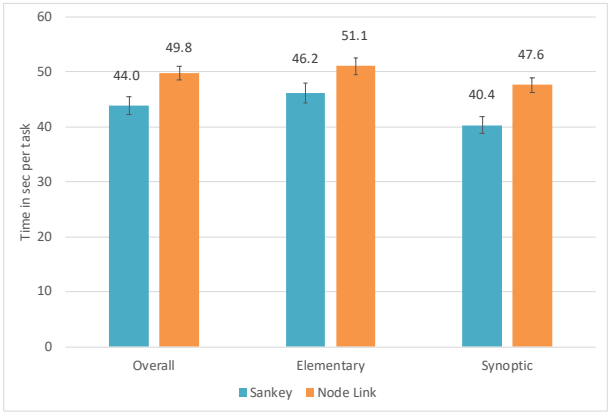

Fig. 4. Average efficiency (time in seconds with standard-error) of all tasks, elementary tasks only (8 tasks) and synoptic tasks only (5 tasks).

Sankey than in node-link on both task types. In addition, synoptic tasks were answered more quickly than elementary tasks.

Subjective Evaluation and Preference There was no difference in participant satisfaction from each tool (M=3.73, SD=.91 for Sankey, M=3.90, SD=.95 for node-link; paired sample t(76)=.94, p=.35). However, when asked which of the two tools they preferred, 50 participants (65%) preferred the node-link tool compared to 27 who preferred the Sankey tool. Regardless, there were only low correlations between the participants' achievements in the experiment and their tool of choice.

Figure **??** describes the relationships between performance measures, user satisfaction, and user preference. The data plotted are from the 77 individuals who participated in the experiment. Circles filled with orange denote participants who preferred the node-link tool; circles filled with blue denote those who preferred the Sankey tool. The circles' outline (stroke) denote differences in satisfaction, whereas the size of the circles represents the magnitude of the difference. Larger circles represent larger differences in satisfaction score. For example, Participant #5, just to the right and above the center, preferred the node-link tool, despite reporting considerably more satisfaction with the Sankey tool. Participant #16, just to the left and below the center, showed the same preference and satisfaction pattern.

The x-axis in Figure **??** presents effectiveness differences between the two visualization tools (Sankey correct – Node-Link correct). Positive values (right half) denote participants whose effectiveness using Sankey was better than their effectiveness using Node-link. The y-axis denotes differences in efficiency, expressed as Node-Link completion time – Sankey completion times. Positive numbers (upper half) denote that using Sankey was more efficient (took less time). The values on this axis are the differences in seconds divided by 100, for simplicity of presentation. The two participants (#5 and #16) discussed earlier (with more satisfaction for the Sankey, but preference fo

the node-link diagram) show very different performance patterns: #5 is more effective and more efficient with the Sankey, the other #16 with the node-link diagram.

The resulting matrix can be interpreted as follows. Quadrant II denotes participants who performed better on both aspects (effectiveness and efficiency) using *Sankey*. Quadrant IV denotes participants who performed better on both aspects using *node-link*. Quadrants I and III include users with performance trade-offs. In Quadrant I participants were more effective using node-link but more efficient using Sankey, whereas Quadrant III includes participants with the opposite type of tradeoff. For example, Participant #31 at the top of Quadrant I performed more effectively using node-link but was faster using Sankey. Participant #53 on the right-hand side of Quadrant III was more effective using Sankey but faster using node-link.

To test which factors affected the participants' evaluations, we conducted separate regression analyses for the two satisfaction items. In each model, the predictors were effectiveness (number of correct answers) and efficiency (average task completion time) of the two visualization tools. The results (Table **??**) were very similar in terms of the explained variance (about 10% for each tool) and the fact that the only significant predictor was the effectiveness score of that tool.

Table 3. A regression model to predict user satisfaction with the visualization tool

	DV= Node-Link Satisfaction			DV = Sankey Satisfaction		
Predictors	Beta	t	sig	Beta	T	sig
Sankey correct answers	-.204	-1.723	.089	.323	2.772	.007
Sankey time per task	-.042	-.353	.725	.049	.421	.675
Node Link correct answers	.300	2.550	.013	-.184	-1.588	.117
Node Link time per task	.086	.731	.467	.120	1.042	.301

A logistic regression with effectiveness, efficiency, and satisfaction scores on both tools as predictors correctly classified 83% of the participants' preferences (Table **??**). The model's Cox & Snell's R2 was .384. The only significant predictors in the model were the two satisfaction items (Table **??**).

4.3 Expert Interviews

The expert opinions elicited through the interviews showed a slight overall preference for the Sankey diagram. However, preference of tool was mostly dependent on the user task. When entities (Users, Databases and Verbs) had to be investigated on their own, experts stated that this was harder to perform with the Node-Link diagram, mostly due to the spread-out layout which sometimes caused entities "to be all over the place". As

Table 4. Classification table for the logistic regression analysis

		Predicted Choice		%
		NL	SK	Correct
Observed Choice	NL	44	6	88.0
	SK	7	20	74.1
Overall				83.1

Table 5. Logistic regression model of predictors of preferred visualization

	B	S.E.	Wald	df	Sig.	Exp(B)
Sankey correct	.311	.349	.794	1	.373	1.365
Node-Link correct	-.278	.291	.913	1	.339	.757
Sankey time per task	.023	.024	.898	1	.343	1.023
Node-Link time per task	-.031	.032	.923	1	.337	.970
Sankey satisfaction	1.135	.473	5.757	1	.016	3.110
Node-Link satisfaction	-.996	.387	6.615	1	.010	.369
Constant	-1.079	4.480	.058	1	.810	.340

one expert said: "It is hard finding the users, they are all placed in different positions". For these type of tasks, the constrained layout of the Sankey diagram seemed to be an advantage. However, For finding groups of Users connected to Databases, experts thought that the node-link diagram has a clear advantage since they were grouped in the layout closer together. Experts found it very intuitive that "close proximity indicates stronger connections". In the Sankey diagram this is more difficult as connecting lines need to be visually highlighted one user at the time. For comparison tasks between entities of the same type, both visualizations "require additional manual work" and there was no clear preference for either of the techniques. Finally, for tasks involving Databases and Verbs only, some of the experts expressed preference for the node-link diagram, where color coding helped the association between the entities, even though they stated that much effort needs to be put into this task using both types of visualizations.

5 Discussion

We conducted a systematic experimental comparison of two visualization solutions for the cyber security domain, specifically, for the analysis of database-related activities. The visualizations represent various design trade-offs that facilitate or hamper users' decision making in different types of tasks. Consequently, our research model postulated that the type of tasks in which the users engage could moderate the effects of the visualization tools on user performance. Thus, the participants in the main part of the experiment completed 14 well-defined tasks that were classified into 2 main types, based on [**?**] high-level classification of tasks to elementary and synoptic. In the first analysis, we compared the performance of participants who were aided by the visualization tools

to the performance of participants who viewed the data using a spreadsheet. Finally, we complemented the controlled experiment with interviews with seven domain experts.

Using the data from 135 participants in a between-groups design, the results first demonstrate that visualization tools are superior to the spreadsheet presentation of the database access data, in terms of both effectiveness and efficiency, confirming the benefit of visualizations as an analysis tool over the use of a spreadsheet. Subsequent analyses concentrated on the results of the within-subjects part of the experiment, in which 77 of the participants used 2 visualization tools. We compared the tools in terms of their effectiveness, efficiency, and user satisfaction and preference. During the analyses we found exceptional user performance data on a task that combined synoptic behavior comparison and outlier detection. We will discuss this task separately following a discussion of the results of the other 13 tasks and the implications of those results.

5.1 Effectiveness of the Visualization Tools is Contingent on Task

The analyses of the effectiveness data demonstrate the importance of considering the moderating effect of task type when evaluating the performance of visualization tools. This was also emphsized by the experts in their interviews. Without considering task type, the study's results would suggest that the two visualization tools provide the same degree of support for the cyber security context studied in this project. However, our analysis indicates that the node-link diagram helped users complete the elementary tasks more correctly relative to the Sankey diagram. At the same time, synoptic tasks were answered more correctly using the Sankey diagram.

A possible explanation for the moderating effect of task type is that the node-link diagram provides a semantic organization of the layout, bringing related objects closer together and pushing unrelated objects farther away. As a result, finding related objects, as required in elementary tasks, may benefit from this type of layout. In addition, the line-widths in the Sankey correspond to node sizes in a more explicit manner, thus it supports tasks requiring comparison better than node-link diagrams, where nodes and lines have different scales, and thus may be more suitable for synoptic tasks.

5.2 Efficiency of Visualization Tools

The results analysis revealed that, on average, using the Sankey diagram resulted in shorter task completion times. This was the case for both the elementary and the synoptic tasks, and thus suggests an inherent advantage to the Sankey diagram in terms of speed. On the one hand, this advantage represents speed-accuracy tradeoff for elementary tasks. Users performed faster with Sankey but more accurately with node-link. On the other hand, it represents a clear advantage for using Sankey when users engage in synoptic tasks; performance is both more accurate and faster.

From a practical perspective, these findings call for the incorporation of Sankey diagrams in support of database administrators who are interested in understanding database-access activities. Our conjecture about the reasons behind these findings is that the Sankey diagram provides constraints and superimposes a kind of organization to the layout by the horizontal positioning of the information layers. In contrast, location and orientation of nodes and links may change substantially in the node-link diagram.

Thus, the greater structure of the Sankey diagram improves familiarity and consistency, which can lead to faster performance when conducting any of the task types.

These findings are especially important given the ubiquity of node-link diagrams in cyber security systems. Our research suggests the possibility that at least certain types of cyber security tasks can be better handled by other types of visualizations. In our study, the Sankey visualization provided more effective support for users engaged in synoptic tasks and a higher overall efficiency. Considering different user goals (e.g., exploration rather than detection) or different task classifications (e.g.,[**?**]) suggests that additional visualization tools could also be beneficial for cyber security experts.

5.3 Subjective Evaluation of the Visualization Tools

User evaluation of the visualization tools revealed several interesting findings. First, although users expressed their satisfaction only after using both tools, their satisfaction was only correlated with the effectiveness of the tool for which a satisfaction score was given. In other words, performance on the other tool did not play a role in the satisfaction score, nor did the completion times of the evaluated tool. Second, the predictors used in our regression model explained only a small portion of the variance of the satisfaction score (about 10%). This finding may point to the existence of other factors affecting satisfaction, e.g., learnability and ease of use [**?**] or aesthetics[**?**]. Third, although the majority of users (about two-thirds) preferred the node-link tool, there was no difference in user satisfaction between the two tools. The logistic regression findings suggest that the only predictors for preference were user satisfaction with both tools. Performance measures had no effect on preference. Thus, user preference may result from a complex combination of factors, of which performance may not be the most important. Figure **??** provides a detailed view of user preferences, given satisfaction scores and performance measures in both systems. It can be argued that this figure portrays a story of diversity. Diversity in terms of effectiveness and efficiency, in terms of whether these performance aspects are traded-off against each other, and in terms of user satisfaction and preferences. The observed diversity in this study provides support for recent calls for the personalization of visualization tools [**?**].

In more general terms, the idea that performance depends on how support tools are commensurate with task demands is not new. Early research on decision support systems identified the importance of such a contingency view [**?**]. Later research provided evidence for the need to match the support tools to the task at hand [**?**,**?**]. As [**?**] suggests, "task-technology fit, when decomposed into its more detailed components, could be the basis for a strong diagnostic tool to evaluate whether information systems and services in a given organization are meeting their needs".

In this context, it is worth mentioning that user performance with the node-link diagram dominated their performance using the Sankey diagram for one specific task, Task 12. The task, "Which User used the most diverse DBNames at 15:00?" is classified as a synoptic task that involves behavior comparison and outlier detection. Our retrospective analysis of this task suggests that while using Sankey, users had a hard time completing this task because they needed to consult two diagram axes that were on the opposite sides of the screen. The axis representing the user was on the left of the screen, whereas the axis representing the database was on the right of the screen. Using

node-link, on the other hand, highlighting of a specific node causes unrelated values to fade out, leaving a relatively clear view of the relevant values of the associated entities. The immediate implication of this finding is that tasks of this type are better performed using node-link. However, it is also possible to conceive an adaptation of the Sankey diagram to the context of the task, such that remote axes can be brought closer by the user. While such a solution is more complex and requires greater expertise by the users, it is nonetheless feasible. In fact, it is likely desirable in a personalized system or if the Sankey diagram is chosen as the only visualization tool for the cyber security system.

6 Limitations

Experimental work usually requires the researchers to consider multiple design trade-offs. In the following, we list the limitations of our study in light of the design decisions we made and their potential threats to the validity of the findings.

Our study used students as participants, which may reduce the external validity of the findings. The reason for using students was mainly due to the difficulty of arranging a large sample of professionals for the controlled user study. To mitigate this effect, we framed the experimental scenario as one that the participants were familiar with (i.e., the university environment). They were also familiar with database essentials and aware of data security issues given the university scenario. We note that the tasks themselves were not trivial and the participants treated them seriously, taking on average close to 50 seconds to complete a task. Finally, from the perspective of isolating the net effects of the visualization tools and the experimental tasks on user performance and preferences, using participants who are not already involved in data security operations alleviates the confounding effects of previous experience (e.g., in using the familiar node-link diagram in cyber security systems or being previously engaged in similar or identical tasks).

Another limitation is the fact that tasks were classified and analyzed in our research only according to the highest level of classification in [**?**]. Tasks were also identified in terms of lower-level classifications; however, due to limitations on sample size and length of experimental session, we decided not to expand the number of tasks and thus did not include lower-level classifications as independent factors in the experimental designs. Moreover, other task classifications exist, which can also be used in the domain of this research. As an initial investigation, we used relatively short tasks based on a formal task taxonomy, rather than open-ended domain-based tasks. However, the tasks that we used in the study are sub-tasks that are used when investigating security breaches. Future studies will investigate domain-based tasks as well as examine these issues in the field, in real-world settings. Finally, the questions were multiple choice type questions with varied amount of answers. This may give rise to chance findings (on average, slightly below 0.25 chance to get the answer by guessing). However, we note that this is common in such experiments and the chance is divided equally between conditions.

It is possible that giving the participants feedback on their tasks would have made their subjective assessments of the visualization tools more reflective of their performance. However, such explicit feedback is rarely available in the real world, and thus

we opted not to include it. Given the discrepancy between performance and subjective measures, it would be useful to study how much of this discrepancy can be attributed to lack of feedback on performance and how much is due to other aspects influencing users' subjective evaluations.

We have used a force-directed layout for our node-link representation. However, there are other possible layout options to represent node-link diagrams. Using the force-directed layout was motivated by the popularity of this technique by the literature and commonly available tools. Unfortunately, the comparison of different layout algorithms is beyond the scope of the current effort, but should be considered in future research. Finally, the question of scalability of visualization techniques would have posed a significant complexity to our empirical setting, and would have prolonged the experiment for the participants. Therefore, we fixed the amount of data to a level typical for small- and medium-size enterprises. The effect of scalability on user performance in the visualization of security systems is a crucial research question, and is left to be investigated in future research.

7 Design Recommendations

The objective of this study was to compare two visualization systems in the cyber security context of database-activity monitoring in terms of their performance and users' subjective evaluations. The experiment's data included some clear and statistically significant results that can be used to devise design guidelines. Although appropriate scientific caution should be applied regarding the generalization of these guidelines beyond the study's cyber security context, we believe these guidelines can apply to other contexts that use tasks with a similar structure to those we used. We recommend the following design guidelines, taking into consideration the limitations described above:

- For elementary tasks, the node-link diagram produces more effective (i.e., correct) responses than the Sankey diagram.
- For synoptic tasks, the Sankey diagram produces more effective and more efficient responses than the node-link diagram. Thus, our results unequivocally support the use of Sankey for synoptic tasks.
- If efficiency (speed of completing tasks) is an important criterion, then the Sankey diagram is preferred over the node-link diagram. This result was statistically significant across both task types. Still, designers should consider the effectiveness-efficiency tradeoff when it comes to elementary tasks.
- For the special case of tasks that require synoptic behavior comparison of outliers, node-link was clearly the superior tool.
- Users preferred the node-link diagram over the Sankey diagram by a ratio of 2:1. However, user preference and satisfaction did not closely match performance, indicating that factors other than preference may be influencing satisfaction.
- Given users' diversity in performance and preference, and given that task type moderates the effects of visualization type on performance, we recommend that designers consider supporting users with more than one visualization method. Furthermore, designers should consider giving users the means to switch between methods as a function of the task and of their preference, either by user control, or by utilizing user-adapted techniques [**?**,**?**]

References

1. Albo, Y., Lanir, J., Bak, P., Rafaeli, S.: Off the radar: Comparative evaluation of radial visualization solutions for composite indicators. IEEE transactions on visualization and computer graphics **22**(1), 569–578 (2016)
2. Amar, R., Eagan, J., Stasko, J.: Low-level components of analytic activity in information visualization. In: Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on. pp. 111–117. IEEE (2005)
3. Andrienko, N., Andrienko, G., Gatalsky, P.: Exploratory spatio-temporal visualization: an analytical review. Journal of Visual Languages & Computing **14**(6), 503–541 (2003)
4. Ball, R., Fink, G.A., North, C.: Home-centric visualization of network traffic for security administration. In: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security. pp. 55–64. ACM (2004)
5. Best, D.M., Endert, A., Kidwell, D.: 7 key challenges for visualization in cyber network defense. In: Proceedings of the Eleventh Workshop on Visualization for Cyber Security. pp. 33–40. ACM (2014)
6. Best, D.M., Endert, A., Kidwell, D.: 7 key challenges for visualization in cyber network defense. In: Proceedings of the Eleventh Workshop on Visualization for Cyber Security. pp. 33–40. VizSec '14, ACM, New York, NY, USA (2014). https://doi.org/10.1145/2671491.2671497, http://doi.acm.org/10.1145/2671491.2671497
7. Boyandin, I., Bertini, E., Lalanne, D.: A qualitative study on the exploration of temporal changes in flow maps with animation and small-multiples. In: Computer Graphics Forum. vol. 31, pp. 1005–1014. Wiley Online Library (2012)
8. Carpendale, S.: Evaluating information visualizations. In: Information visualization, pp. 19–45. Springer (2008)
9. Chen, C., Yu, Y.: Empirical studies of information visualization: a meta-analysis. International Journal of Human-Computer Studies **53**(5), 851–866 (2000)
10. Conati, C., Carenini, G., Toker, D., Lallé, S.: Towards user-adaptive information visualization. In: AAAI. pp. 4100–4106 (2015)
11. Dickson, G.W., Senn, J.A., Chervany, N.L.: Research in management information systems: The minnesota experiments. Management science **23**(9), 913–934 (1977)
12. Elam, J.J., Mead, M.: Can software influence creativity? Information Systems Research **1**(1), 1–22 (1990)
13. Ferebee, D., Dasgupta, D.: Security visualization survey. In: Proceedings of the 12th Colloquium for Information Systems Security Education University of Texas. p. 124. Citeseer (2008)
14. Fink, G.A., North, C.L., Endert, A., Rose, S.: Visualizing cyber security: Usable workspaces. In: Visualization for Cyber Security, 2009. VizSec 2009. 6th International Workshop on. pp. 45–56. IEEE (2009)
15. Ghoniem, M., Fekete, J.D., Castagliola, P.: A comparison of the readability of graphs using node-link and matrix-based representations. In: IEEE Symposium on Information Visualization. pp. 17–24 (2004). https://doi.org/10.1109/INFVIS.2004.1
16. Girardin, L., Brodbeck, D.: A visual approach for monitoring logs. In: LISA. vol. 98, pp. 299–308 (1998)
17. Goodhue, D.L., Thompson, R.L.: Task-technology fit and individual performance. MIS quarterly pp. 213–236 (1995)
18. Hascoët, M., Dragicevic, P.: Interactive graph matching and visual comparison of graphs and clustered graphs. In: Proceedings of the International Working Conference on Advanced Visual Interfaces. pp. 522–529. ACM (2012)

19. Heer, J., Card, S.K., Landay, J.A.: Prefuse: a toolkit for interactive information visualization. In: Proceedings of the SIGCHI conference on Human factors in computing systems. pp. 421–430. ACM (2005)
20. Herman, I., Melançon, G., Marshall, M.S.: Graph visualization and navigation in information visualization: A survey. IEEE Transactions on visualization and computer graphics **6**(1), 24–43 (2000)
21. Hoekstra, R., Groth, P.: Prov-o-viz-understanding the role of activities in provenance. In: International Provenance and Annotation Workshop. pp. 215–220. Springer (2014)
22. Huanga, W.: Measuring effectiveness of graph visualizations: A cognitive load perspective. Information Visualization **8**, 139–152 (2009)
23. Kamra, A., Terzi, E., Bertino, E.: Detecting anomalous access patterns in relational databases. The VLDB Journal, The International Journal on Very Large Data Bases **17**(5), 1063–1077 (2008)
24. Lam, H., Bertini, E., Isenberg, P., Plaisant, C., Carpendale, M.S.T.: Empirical studies in information visualization: Seven scenarios. IEEE Transactions on Visualization and Computer Graphics **18**, 1520–1536 (2012)
25. Liu, S., Cui, W., Wu, Y., Liu, M.: A survey on information visualization: recent advances and challenges. The Visual Computer **30**(12), 1373–1393 (2014)
26. Ma, K.L.: Cyber security through visualization. In: Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60. pp. 3–7. Australian Computer Society, Inc. (2006)
27. Perer, A., Wang, F.: Frequence: Interactive mining and visualization of temporal frequent event sequences. In: Proceedings of the 19th international conference on Intelligent User Interfaces. pp. 153–162. ACM (2014)
28. Riehmann, P., Hanfler, M., Froehlich, B.: Interactive sankey diagrams. In: Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on. pp. 233–240. IEEE (2005)
29. Schmidt, M.: The sankey diagram in energy and material flow management. Journal of industrial ecology **12**(1), 82–94 (2008)
30. Shiravi, H., Shiravi, A., Ghorbani, A.A.: A survey of visualization systems for network security. IEEE Transactions on visualization and computer graphics **18**(8), 1313–1329 (2012)
31. Staheli, D., Yu, T., Crouser, R.J., Damodaran, S., Nam, K., O'Gwynn, D., McKenna, S., Harrison, L.: Visualization evaluation for cyber security: Trends and future directions. In: Proceedings of the Eleventh Workshop on Visualization for Cyber Security. pp. 49–56. ACM (2014)
32. Toker, D., Conati, C., Steichen, B., Carenini, G.: Individual user characteristics and information visualization: connecting the dots through eye tracking. In: proceedings of the SIGCHI Conference on Human Factors in Computing Systems. pp. 295–304. ACM (2013)
33. Tractinsky, N.: Visual aesthetics. The Encyclopedia of Human-Computer Interaction, 2nd Ed. (2013)
34. Wagner, M., Fischer, F., Luh, R., Haberson, A., Rind, A., Keim, D.A., Aigner, W., Borgo, R., Ganovelli, F., Viola, I.: A survey of visualization systems for malware analysis. In: EG Conference on Visualization (EuroVis)-STARs. pp. 105–125 (2015)
35. Wongsuphasawat, K., Gotz, D.: Exploring flow, factors, and outcomes of temporal event sequences with the outflow visualization. IEEE Transactions on Visualization and Computer Graphics **18**(12), 2659–2668 (2012)
36. Yin, X., Yurcik, W., Treaster, M., Li, Y., Lakkaraju, K.: Visflowconnect: netflow visualizations of link relationships for security situational awareness. In: Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security. pp. 26–34. ACM (2004)

37. Zhao, J., Liu, Z., Dontcheva, M., Hertzmann, A., Wilson, A.: Matrixwave: Visual comparison of event sequence data. In: Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems. pp. 259–268. ACM (2015)

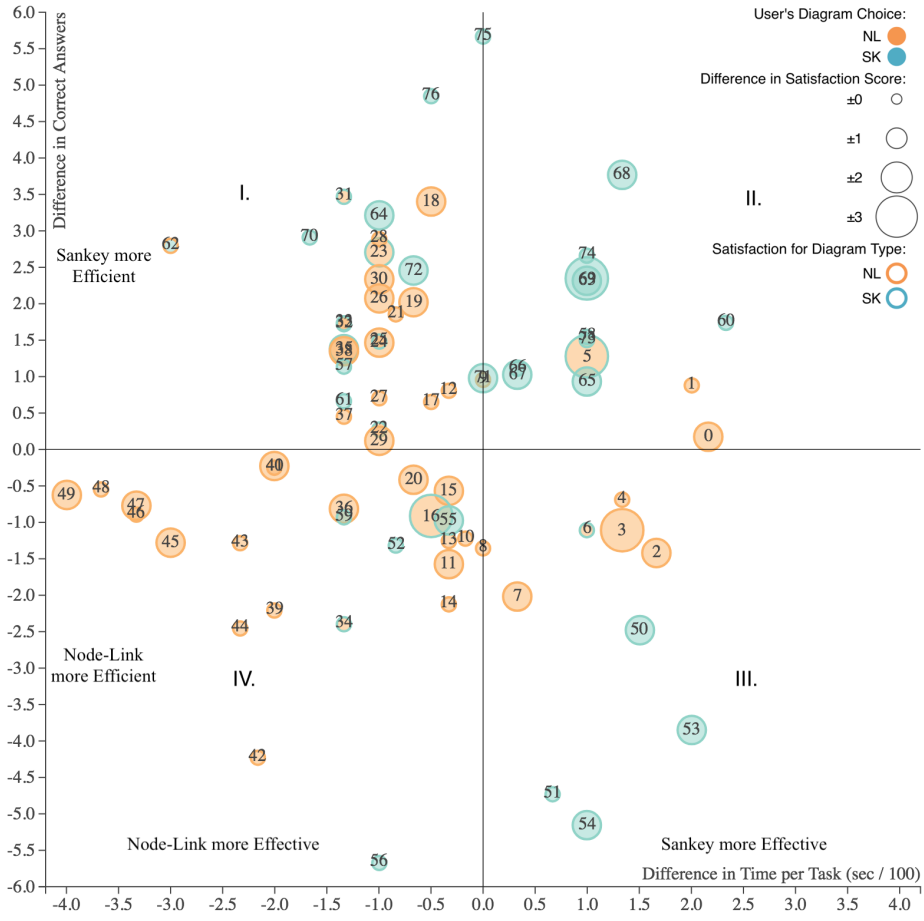

Fig. 5. Participant preference of a diagram (node-link or Sankey) is indicated by the colored circles on the scatter-plot. Differences in effectiveness (number of correct answers) are mapped to the x-axis, efficiency (average completion time in seconds/100) is mapped on the y-axis, and differences in satisfaction scores are plotted for each participant (labeled by the numbers) as the size of the circles.