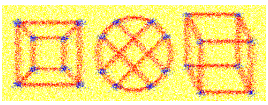
Journal of Graph Algorithms and Applications
http://jgaa.info/ vol. 0, no. 0, pp. 0–0 (0)
DOI: 10.7155/jgaa.00467

Visual Similarity Perception of Directed Acyclic Graphs: A Study on Influencing Factors and Similarity Judgment Strategies

Kathrin Ballweg[1] *Margit Pohl*[2]
Günter Wallner[2] *Tatiana von Landesberger*[1]

[1]Technische Universität Darmstadt, Darmstadt, Germany
[2]Vienna University of Technology, Vienna, Austria

Abstract

Visual comparison of directed acyclic graphs (DAGs) is commonly encountered in various disciplines (e.g., finance, biology). Still, knowledge about humans' perception of their similarity is currently quite limited. By similarity perception, we mean how humans perceive commonalities and differences of DAGs and herewith come to a similarity judgment. To fill this gap, we strive to identify factors influencing the DAG similarity perception. Therefore, we conducted a card sorting study employing a quantitative and qualitative analysis approach to identify (1) groups of DAGs the participants perceived as similar and (2) the reasons behind their groupings. We also did an extended analysis of our collected data to (1) reveal specifics of the influencing factors and (2) investigate which strategies are employed to come to a similarity judgment. Our results suggest that DAG similarity perception is mainly influenced by the number of levels, the number of nodes on a level, and the overall shape of the DAG. We also identified three strategies used by the participants to form groups of similar DAGs: divide and conquer, respecting the entire dataset and considering the factors one after the other, and considering a single factor. Factor specifics are, e.g., that humans on average consider four factors while judging the similarity of DAGs. Building an understanding of these processes may inform the design of comparative visualizations and strategies for interacting with them. The interaction strategies must allow the user to apply her similarity judgment strategy to the data. The considered factors bear information on, e.g., which factors are overlooked by humans and thus need to be highlighted by the visualization.

Submitted:	Reviewed:	Revised:	Accepted:	Final:
November 2017	January 2018	March 2018	April 2018	April 2018

Published:

Article type: Regular paper

Communicated by: F. Frati and K.-L. Ma

This work was financially supported by the DFG (LA 3001/2-1) and the FWF (I 2703-N31).

E-mail addresses: kathrin.ballweg@gris.tu-darmstadt.de (Kathrin Ballweg) margit@igw.tuwien.ac.at (Margit Pohl) guenter.wallner@tuwien.ac.at (Günter Wallner) tatiana.von.landesberger@gris.tu-darmstadt.de (Tatiana von Landesberger)

1 Introduction

Visual comparison of directed acyclic graphs (DAGs) is a task encountered in various disciplines, e.g., in finance, biology, or social network analysis. The task is strongly influenced by the human perception of similarity since comparison builds upon making similarity judgments. In spite of the numerous occurrences of this task and recent papers surveying visual graph comparison techniques [4, 17], knowledge about the human perception of graph similarity – especially for DAGs – is quite limited.

Only a few investigations address the comparison of graphs. Gleicher et al. [17] identified basic types of techniques for visual comparison (juxtaposition, superposition, and explicit encoding). Tominski et al. [54] explicitly deal with the comparison of large node-link diagrams in superposition.
Some interesting insights can be gained from the literature on dynamic graphs showing the evolution of node-link diagrams over time [4]. Others discuss the extension of these techniques with highlighting of commonalities and differences [1, 3, 6, 21]. However, none of these papers deal with the issue of similarity perception within the context of graph comparison.
Research on graph readability is related since the DAGs need to be well perceivable to compare them. Examples include studies on edge crossings and mental map preservation (e.g., [28, 44, 46, 47]).
The research investigating the comparison of other visualization types is also interesting. Pandey et al. [40] conducted an experiment to study the similarity perception of scatterplots. So, their work inspired our methodology.

To the best of our knowledge, there is no research focusing on how humans perceive the similarity of DAGs. We are especially interested in the *factors which influence the perception of similarity* (possibly, number of nodes/edges, edge crossings, etc.). We deem the knowledge about the influencing factors important for the generation of future actionable guidelines for comparative visualizations. For instance, based on the knowledge about which of the factors bearing comparison-relevant information are overlooked by humans, we can formulate guidelines on to be visually highlighted factors.
Towards this end, we conducted a study with small, unlabeled synthetic DAGs and used card sorting as our methodology. Card sorting studies are based on the comparison of multiple data items to each other. For DAGs this could be necessary when a financial analyst wants to compare multiple simulation runs of contagion effects in a network [57]. Biologists have to compare multiple DAGs when they analyze phylogenetic trees [20] or when they analyze multiple runs of mutation simulations [32]. We decided for these DAGs to keep the number of factors to be tested manageable. Especially because of the currently limited knowledge about graph similarity perception, we consider the manageability as crucial. However, because of our systematic procedure, the study scope can be easily extended in the future. The DAGs are represented as node-link diagrams. We address two research questions (RQs): (1) *Which groups do the participants form?*, and (2) *Which factors did the participants consider to judge the similarity?*

Moreover, we also investigated: (1) the *specifics of the factors influencing human similarity perception* and (2) *how the participants combine these factors to come to a final similarity judgment expressed by the grouping*, i.e., strategies they employ to solve the task of judging the similarity. Strategies are defined as sequences of processes for solving a task [31]. The factors influencing human similarity perception are elements of such strategies. Identifying such strategies aids the selection of appropriate interactions for comparative visualization systems, thus the interaction must allow the user to apply her strategy to the data. From the analysis of the factors' specifics, we can learn further details about the humans' mental model of similarity, e.g., the limit on similarity perception-influencing factors humans consider for their judgment.

Our results indicate that the similarity perception of DAGs in visual comparison is consistent and well objectifiable with graph theoretical and visual properties. *Visual factors* are factors which result from visualizing the DAGs as node-link diagrams, e.g., edge crossings. *Graph theoretical factors* result from the data structure per se, e.g., the depth of the DAG. The factors mainly influencing human similarity perception of DAGs are the *number of levels*, the *number of nodes on a specific level*, and the overall *shape*. The extension of our analysis revealed several specifics of the influencing factors – e.g., participants considered on average four factors and there is no recognizable tendency whether participants rather use visual or graph theoretical factors. These are relevant details on the human mental model regarding judging the similarity of DAGs. Furthermore, the analysis revealed that the participants adopt three different strategies: *divide and conquer*, *respecting the entire dataset considering the factors one after the other*, and *considering one single factor*. We provide supplementary material – including our study material (dataset, task sheets, etc.), our collected data, and our analysis results – on our website[1].

The remainder of this paper is structured as follows: In the next section, we review related work. In Section 3, we outline our study design, including the research questions, the dataset employed, and the study procedure. In Section 4, we discuss the analysis and results of which groups participants form *(RQ1)* and which factors they consider while judging the similarity *(RQ2)*. Section 5 reports detailed results on the used factors and the strategies participants employed to form their groups and consequently to solve the task of judging similarity. In Section 6 we summarize and discuss our results as well as outline future work. Finally, we draw conclusions from our work in Section 7 based on the elaborations in Section 6.

[1] http://www.gris.tu-darmstadt.de/research/vissearch/projects/DAGSimilarityPerception/index.html

2 Related Work

There exists an extensive body of research in perceptual psychology and pattern recognition on similarity judgments and dissimilarity measures (see [30, 41] for an overview). In the following, we will concentrate on work dealing with graphs and other types of visualizations.

Graph Visualization and Visual Comparison Techniques. Several recent surveys deal with graph visualization techniques and visual comparison techniques (e.g., [4, 17, 19, 56, 58]). The basic techniques, that is, juxtaposition, superposition, and explicit encoding – following Gleicher et al.'s [16, 17] classification – are sometimes enriched by emphasizing the commonalities and differences between graphs [10, 21]. Some highlight similar parts [3, 6, 21], while others emphasize differences by collapsing the identical parts [1]. The enrichment, that is, emphasizing the commonalities or differences, usually relies on a similarity function. In this respect, Gao et al. [14] provide an overview of research done on graph edit distances, a mathematical way to measure the similarity between pairwise graphs. However, it is still unknown whether the criteria on which existing similarity functions are based correspond to the criteria used by humans when visually comparing two or more node-link diagrams. Tominski et al. [54] proposed interaction techniques which aid users in doing comparison tasks and which were inspired by the real-world behavior of people when comparing information printed on paper. Getting a better understanding of the perceived differences and commonalities is likely to result in better visualization and interaction techniques. We can learn from this, for instance, which differences and/or commonalities are overlooked by humans and consequently need to be highlighted by the visualization system.

Graph Readability. Moreover, the existing body of work dealing with perceptual and cognitive aspects focuses primarily on the readability of single graphs. Several factors, including graph aesthetics (edge crossings [28, 44, 45], layout [11, 25, 33, 34, 39], graph design [22, 51], and graph semantics or content knowledge [29, 39, 46]) have been identified to be important for graph readability. Huang et al. [23], concerned with sociograms, note that good readability is not enough to effectively communicate network structures, emphasizing that the spatial arrangement of the nodes also influences viewers in perceiving the structure of social networks.

Visual Comparison of Node-Link Diagrams. While perceptual aspects of single graphs have been thoroughly investigated, literature dealing with perceptual aspects when comparing node-link visualizations is considerably more scarce. Notable papers in this space are the work of Bach et al. [3] and Ghani et al. [15] who are both concerned with dynamic graphs (cf. Beck et al. [4] for an overview). The work of Archambault et al. [2] and Bridgeman et al. [7] is also noteworthy. While Archambault et al. evaluated the effectiveness of difference maps which show changes between time slices of dynamic graphs, Bridgeman et al. were concerned with how the extent of the mental map preservation between two time slices can be measured (metrics) and how the suitability of the metrics can be evaluated. While we are not necessarily concerned with dynamic graphs,

these works are nonetheless relevant in our context as dynamic graphs are often analyzed by using discrete time-slices. In our previous work [59], we provided an overview of methodological challenges when dealing with the investigation of graph comparison and described a first preliminary study targeted towards identifying factors which influence the recognition of graph differences in very small star-shaped node-link diagrams. The work presented in this paper can be viewed as a continuation of these efforts.

Visual Comparison of Other Visualization Types. Beyond the perception of node-link diagrams, literature is currently also quite limited when it comes to the similarity perception of other visualization types. This sentiment is shared by Pandey et al. [40] who investigated how human observers judge the similarity of scatterplots. Our quantitative analysis as presented in this paper is partly based on the methodology put forward by Pandey et al. [40]. Fuchs et al. [13] looked into how contours affect the recognition of data similarity in star glyphs. Likewise, Klippel et al. [27] investigated the similarity judgments of star glyphs using a methodology similar to that of Pandey et al. [40] and ours: Participants were shown various visualizations which they then had to group according to their perceived similarity.

Human Strategies. In addition to the similarity perception-influencing factors, we also investigated how humans reached their similarity judgements. There has been some research concerning problem-solving strategies in various domains, but there is still fairly little research in the area of visualization concerning these issues. Essential research has been conducted by Newell and Simon [38], who point out that the some problem spaces can be very large so that heuristics must be used to decrease the number of solutions that need to be considered. Specific strategies have, for example, been investigated in the context of solving logical reasoning problems [37] and graph comprehension [12]. In graph comprehension, researchers typically distinguish between identifying data (finding data points) and going beyond the data (reasoning and drawing inferences). Based on that, Trickett and Trafton [55] developed a more general model of graph comprehension. Mirel [36] discussed the issue of strategies in the context of Human-Computer Interaction. She argues that in this context, high-level tasks should be investigated because only herewith it is possible to understand complex problem-solving activities. In information visualization, interaction strategies have been investigated to find out how users work with visualizations and how they make sense of the data presented in these visualizations [42, 50]. In addition to such domain-specific strategies, there are also general problem-solving strategies (e.g., means-end analysis). In general, there is evidence from diverse domains indicating that humans adopt various strategies to solve problems and that the investigation of these strategies can help to design systems in a way to support users to work more efficiently. The users' efficacy is increased by a system designed in that manner; thus, i.a. the chosen interaction techniques are based on the users' strategies. The system's interaction techniques, in turn, are the means by which the user can apply his task solving strategy.

3 Study Methodology

In this section, we present our study methodology. We were strongly inspired by the work of Pandey et al. [40] about the human similarity perception of large sets of scatterplots since we share the research questions for different data types. Pandey et al. substantiate that the methodological principle of card sorting produces valuable results for this type of research questions. For advantages, drawbacks, and the suitability of card sorting for our research questions, see Section 3.4.

3.1 Research Questions

Our superordinate research question (RQ) is: *What factors influence the human similarity perception of DAGs?* We first have to know the factors influencing the similarity judgment. Once we know the influencing factors, we can, for instance, research the specific degree of influence of a single factor as well as the interplay between the factors. To analyze our superordinate RQ, we formulate two subordinate ones:

- RQ1: *Which groups do the participants form?*
- RQ2: *Which factors did the participants consider to judge the similarity?*

We designed our study procedure for *RQ1* and *RQ2* (cf. Section 3.4). Moreover, we deemed it also very important to understand (1) the *specifics of the factors influencing human similarity perception* and (2) the strategies *how the participants combine these factors to come to a final similarity judgment.* While our study was not specifically designed for these questions, we were able to answer them by processing the data collected for *RQ2*. We elaborate on this extended analysis in Section 5.

3.2 Dataset

Creating an appropriate study dataset is challenging due to the large number of possible variations of the data properties [59]. Therefore, we were forced to limit the number of DAGs. Our object of study was 69 small (6 to 9 nodes), unlabeled, synthetic DAGs visualized as node-link diagrams (cf. Table 1). We decided on a traditional hierarchical node-link diagram layout with the root placed on top (), since Burch et al. [8] found that this layout type outperforms other types such as orthogonal or radial layouts. In the following, we will use the term DAG to also refer to its embedding.

We decided to have *synthetic* and *small* DAGs to keep the number of factors to be tested manageable and to evaluate the factors' influence systematically. The size of our DAGs is realistic. They are comparable to cascades in finance and biology (e.g., [32, 57]) and directed acyclic word graphs [52]. Because of our study and data creation methodology, it is easy feasible to systematically extend our results with further studies. Especially since knowledge about human graph

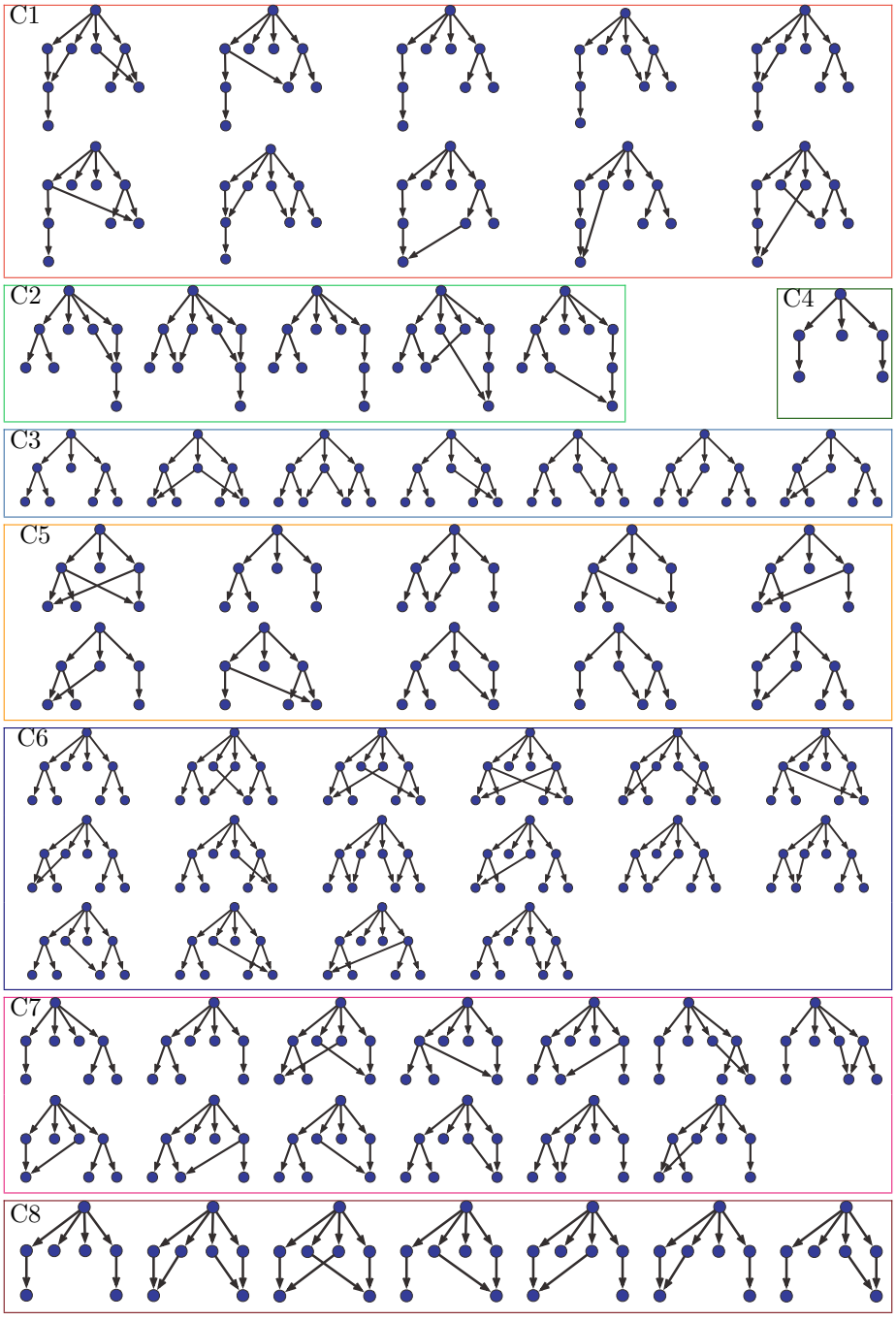

Table 1: Schematic representation of the 69 DAGs used in our study grouped by the clusters identified through hierarchical clustering. The color code is the same as in Figure 2 and 3.

similarity perception is currently quite limited, we consider the manageability of the problem crucial.

When creating the DAGs, we considered *known factors influencing graph readability* (e.g., edge crossings) and *characteristics of DAGs from real-world datasets* (e.g., a node may be the child of more than one parent node). We deem *factors of graph readability* important for visual graph comparison since to be able to visually compare DAGs it is necessary that they are well perceivable. We consider *properties of real DAGs* important for our studies since they influence the visual appearance of the DAGs. More importantly, considering properties of real DAGs strengthens the *realism* of our synthetic data and consequently the *transferability* of our results to real-world use cases.

To create our dataset, we started with DAG $G0$, as depicted in Figure 1. $G0$ is symmetric since it is easier to break symmetry than to introduce symmetry. Using $G0$, we cover *symmetric* and *asymmetric* DAGs. Covering symmetric and asymmetric DAGs is important since humans are sensitive to symmetry [60]. $G0$ is single-rooted since this is typical for various real-world DAG datasets; e.g., cascades. To test *node* and *edge changes (addition of node(s) and edge(s))* we had a two-stage DAG creation process:

1.: We created the base graphs $G1$ - $G6$ and their horizontal *reflections* by adding one, two, and three nodes.
We ensured that the addition of the node(s) is done in the *inner as well as the outer areas of* $G0$ (cf. Figure 1 - Base graphs). It is crucial to have changes in the inner as well as the outer areas as changes in the inner area are, presumably, harder to spot. Inner changes may get embedded in the already existing DAG and may, therefore, be less salient. The reflections of $G2$, $G4$, and $G6$ produce variation in the visual layout of the DAGs (cf. Figure 1). The ability to test the impact of *isomorphism* is a beneficial byproduct.

2.: We created all possible DAGs resulting from adding one and two edges (cf. Figure 1 - Alternatives).
We used our custom-made *GraphCreator*. *GraphCreator* creates all possible DAGs resulting from a specific change of a DAG, e.g., adding one edge to $G1$ - $G6$ and their reflections. Herewith, we ensure that we have the maximal possible variation from which we then sample our study dataset. We did not optimize the layout after a DAG change – e.g., resolving edge crossings – to avoid confounding effects by destroying the mental map.

Down-sampling – i.e., selecting a sub-set of the created DAGs – is necessary since the visual comparison of DAGs is a quite cognitive demanding task for the participants. For the down-sampling we considered the following factors:

- ***Edge crossing*:** *Edge crossing* is a prominent factor in graph readability [28, 44], so we presume that it also plays a role in visual graph comparison.

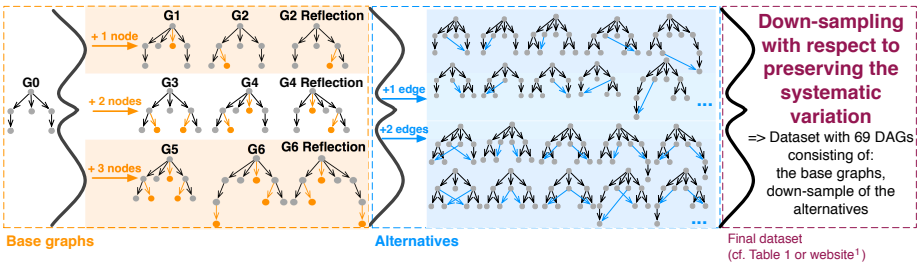

Figure 1: Dataset creation: (I) Base graph creation by adding 1, 2, and 3 nodes to $G0$, ensuring that the added node is placed at the *inner as well as the outer areas of* $G0$, (II) creation of all possible alternatives by adding one and two edges to the base graphs, (III) down-sampling of the alternatives considering these factors: *edge crossing, visual layout, more than one parent node has the same child node, long connections – typically across more than one level, changes are at the inner as well as the outer areas of the respective base graph* (cf. Section 3.2).

- ***Visual layout*:** The visual layout of DAGs does not contain any analytically relevant information about the DAGs' data structure and properties. However, it still has a significant impact on graph readability which is why we deem it important to test its influence on visual graph comparison [11, 23].

- ***More than one parent node has the same child node*** () and ***long connections – typically across more than one level*** () as *characteristics from real-world datasets*: By visually inspecting DAGs from real-world datasets we found these two frequently occurring properties. Parent nodes which have the same child node, for instance, occur in biological or financial cascades. Long edges occur, for instance, in directed acyclic word graphs in natural language processing. Due to their frequency, we considered these two properties as down-sampling factors.

We did the down-sampling under the constraint of preserving the systematic variation (cf. Figure 1 – Final dataset) and by ensuring that changes take place at the *inner as well as the outer areas of the respective base graph.* Our dataset is shown in Table 1 and downloadable here: website.

3.3 Participants

We recruited 20 volunteers (13 male, seven female, between 20 and 60 years). We had no prerequisite of having experience with DAGs. This way our results are not limited to experienced users. In our opinion, it is more likely that experienced users know which factors really bear relevant information for the comparison task whereas misconceptions are more likely for inexperienced users.

We are convinced that if we want to understand the human similarity perception and as a consequence improve comparative visualizations, *we need a varying range of expertise with DAGs*. To achieve this, we recruited participants with a diverse educational level (vocational training, undergraduate, graduate, post-graduate) and from various disciplines.

3.4 Study Procedure

The card sorting session was identical for every participant: after welcoming the participant, the experimenter informed the participant that their data will be anonymized, only used for study purposes and that they can abort the study at any time since they participate voluntarily. The participants had to acknowledge this in a consent form. Afterward, the experimenter handed over the study material and explained the task. Each session took approximately one hour.

Task. We asked the participants to group 69 DAGs with respect to their perceived similarity – multiple occurrences of a single DAG in different groups were allowed. Furthermore, we asked them to tag each group with the factors they used to build them. Finally, participants had to judge the easiness of forming the respective group (*"How difficult or easy was it for you to create this group?"*) and their confidence in the group's consistency (*"How doubtful or confident are you about the consistency of the DAGs in the group, i.e., would you create the same group again if you did this task again?"*). The questions regarding easiness and confidence were judged on a five-point Likert scale ("1 = very difficult/doubtful, 2 = difficult/doubtful, 3 = neutral, 4 = easy/confident, 5 = very easy/confident").

The formed groups provided the data needed to answer *RQ1* while the participants' group tags provided the data to answer *RQ2*. The easiness and confidence judgments provided information on the reliability of the formed groups and their tags. A high easiness score means that the grouping is solid, thus, due to a perceived easy assignment, it is less likely that a participant assigned a DAG randomly. A high confidence score means that the grouping is robust since it is highly probable that it would look similar in case the task was repeated.

For the task formulation, we kept the one from Pandey et al. [40] since it exactly captured what we wanted to ask our participants. Moreover, the formulation was already pretested and successful in Pandey et al.'s study.

Card Sorting Methodology. Card sorting is a well-known methodology in psychology and human-computer interaction for externalizing mental models humans have about the environment they live in. Wood and Wood [61] define card sorting as follows: *As the name implies, the method originally consisted of researchers writing labels representing concepts (either abstract or concrete) on cards, and then asking participants to sort (categorize) the cards into piles that were similar in some ways.* Humans group objects according to their perceived similarity into different categories. In this way, card sorting helps to uncover

the structure of mental models. There are different methods to conduct card sorting. Researchers generally distinguish between open vs. closed sorting tasks and between paper-based and computer-supported card sorting [18]. In closed card sorting, participants have to sort the cards according to a given scheme, whereas in open card sorting, the participants develop a scheme themselves. The procedures for card sorting tasks sometimes differ considerably. Sometimes, the cards that have been assigned to a category are placed in a pile [61], so that participants do not shuffle them around on a canvas. Especially in computerized card sorting, it is often not possible to see all cards from which to choose at the same time [9, 40], which forces the study participants to compare the cards in memory. We used an open, paper-based card sorting since literature indicates that the paper-based approach yields more consistent results than the computerized one [18]. To avoid confronting our participants with huge numbers of duplicate cards, we had the participants group the DAG IDs in lieu of physical cards. These duplicates would have been necessary if we had used physical cards since we allowed multiple occurrences of a single DAG in different groups. Possible drawbacks of the card duplicates could have included priming the participants regarding having to put the DAGs into more than one group or confusing the participants with a huge number of cards and duplicates.

Study Setup and Materials. We used an empty meeting room with good lighting for conducting the study. Each participant received the task sheet, the data sheet, sheets for building the groups, and sheets for tagging each group with the group building factors as well as for judging the easiness and the confidence. The data sheet consisted of the 69 randomly positioned DAGs. We decided to present our dataset on paper so that the participants could see all data items at the same time. The order of the data items was kept the same for all participants to control for which DAGs could be seen together and which had to be compared in memory (cf. Paragraph *"Card Sorting Methodology."*). The participants had to write down the DAGs' IDs which belong to a group and give each group a unique identifier. Furthermore, they had to write down the tags as well as their easiness and confidence judgment together with the unique group identifier. The materials are available on our website[1].

4 Analysis and Results

We did a quantitative and a qualitative analysis. The qualitative analysis provided the factors the participants tagged their formed groups with (*RQ2*). The quantitative analysis resulted in the perceptual consensus over all participants' groupings (*RQ1*). Moreover, it served as a verification of the participants' self-reported factors extracted in the qualitative analysis. Therefore, the quantitative analysis also contributes to *RQ2*.

4.1 Quantitative Analysis (RQ1, RQ2)

To answer *RQ1*, we needed to find the consensus among the individual participant's groupings. We did this by calculating the perceptual consensus of the perceived similarity. It is based on the assumption that the perceived similarity is expressed by whether two DAGs are (not) occurring in the same group. To ensure the reliability of the consensus grouping and the following analysis, we also calculated a perceptual consensus of the easiness and confidence for each consensus cluster based on the participants' individual judgments.

Based on the perceptual consensus grouping – the answer to *RQ1* – we could analyze:

1. *whether the similarity perception is objectifiable with graph theoretical or visual properties.*
   We understand objectifiability as whether it is possible to describe and distinguish the clusters based on the graph theoretical and visual properties of the DAGs belonging to the clusters. Graph theoretical properties result from the data structure per se, and visual properties are properties which result from visualizing the DAGs as node-link diagrams. Examples for these two factor types are edge crossings – visual – and the depth of the DAG – graph theoretical.

2. *whether the similarity perception of humans is consistent across individual people.*
   We define consistency as whether individual people consider similar factors for their similarity judgment.

In the objectifiability of human similarity perception also resides information on the set of influencing factors, the overlap of this set with the set of already known factors – e.g., those from graph readability – and, potentially, heretofore unknown factors. A consistent and objectifiable human similarity perception would mean that we would be able to model it. The model, in turn, would be the perception aware analog to the already existing mathematical similarity functions. To analyze the consistency and objectifiability, we performed a property analysis of the clusters. A high consistency regarding (1) the properties of the DAGs within each consensus cluster and (2) the properties distinguishing the consensus clusters would speak for a consistent human similarity perception regarding the influencing factors. Objectifiable clusters, i.e., clusters based on identifiable factors which seem not arbitrary, would speak for an objectifiable human similarity perception. In other words, the similarity judgment can be described by identifiable factors – e.g., visual or graph theoretical factors. This analysis also allows us to check for the presence of the bias 'saying one is influenced by the one thing and actually being influenced by another' in the self-reported factors. The mitigation potential resides in the perceived similarity consensus encapsulating what the participants really did.

We did the perceptual consensus calculations and the property analysis over all participants with complete data (16), i.e., participants who assigned each

DAG to at least one group. We had to exclude 4 participants who had forgotten to group some DAGs because the perceptual consensus calculation cannot deal with ungrouped data items.

Analysis. To build the perceptual consensus for the participants' similarity judgments, we calculated a pairwise perceptual distance between each pair of DAGs, based on the number of occurrences of each DAG pair in the same group and on the number of individual occurrences (for details cf. [40]). The perceptual distance calculation resulted in a 69×69 perceptual distance matrix (PDM). Like Pandey et al. [40] we did a hierarchical clustering, in our case with average linkage. We evaluated the correct number of clusters with the mean/median of the number of participant-formed groups and with the gap statistic [53]. The mean/median indicate the average number of participant-built groups and thus served as a reasonable estimator for the number of clusters. The gap statistic, like the individual groupings and similarity, employs the cluster similarity which made it another reasonable estimator. The hierarchical clustering result is the consensus grouping of all DAGs based on the perceived similarity consensus contained in the PDM.

For the clusters' property analysis we determined various properties for each graph. Based on this we determined the dominating properties of the clusters as well as the properties separating the clusters. Examples of the employed properties are: *depth*, *visual symmetry*, *visual leaning (left: , right:)*, *edge crossing – number and existence*, *edge length*, *number of nodes on a specific level*, and the *existence* and the *number of nodes having more than one parent node*. In case we did not find any dominant properties, we would look for new, heretofore unknown properties not present in our predefined list.

The calculation of the perceptual consensus for the easiness and confidence of each consensus cluster is based on the idea that each DAG inherits the easiness and confidence score of each participant group it belongs to (also called *individual easiness and confidence judgments*). Based on these *individual easiness and confidence judgments* we calculated an easiness and confidence score for each DAG (also called *DAG easiness and confidence score*). Based on the *DAG easiness and confidence score*s of the DAGs belonging to a respective consensus cluster we calculated the perceptual consensus for the easiness and confidence of each cluster. For the formulas, please refer to [40].

Results. The gap statistic indicated that the data creates eight clusters. Both the mean and median of the number of built groups supported the indicated eight clusters ($mean = 7.6, median = 8.0, STD = 2.6$). So, we decided to cut the dendrogram into eight clusters. Figure 2 shows the resulting dendrogram and the resulting clusters – marked using colored boxes. Table 1 shows the hierarchical clustering result with the visualized DAGs. The clusters are marked with the same colors which were used for the dendrogram. The easiness and confidence scores of all hierarchical clusters are around 4.0 (cf. Table 2). This means that the participants on average found their groups *easy* to build and

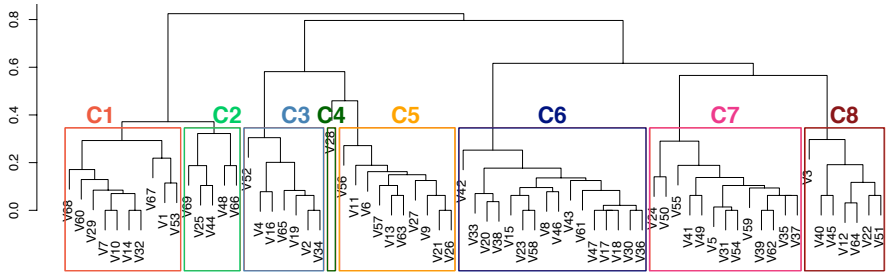

Figure 2: Dendrogram resulting from hierarchical clustering with average linkage. The resulting eight clusters (C1-C8) are marked using colored boxes.

were *confident* they would look similar if they repeated the task. Consequently, this means that the consensus grouping is solid and robust, i.e., *our collected data is reliable*.

The multi-dimensional scaling (MDS) plot, shown in Figure 3 illustrates that the eight identified color-coded clusters form three larger clusters – I: C1 and C2; II: C3, C4, and C5; III: C6, C7, and C8. The following discussion is structured along these three larger clusters. The properties which distinguish the clusters best are the *depth* of the DAGs, the *number of nodes on a specific level* of the DAGs, and the *visual leaning* of the DAGs. Table 2 summarizes the properties of the clusters.

Clusters C1 and C2 are identical in depth and number of nodes on each of their four levels. However, they are separated by the visual leaning. While the DAGs of C1 are left-skewed ( ), those of C2 are right-skewed ( ).

Clusters C3, C4, and C5 have identical depth (3) as well as three nodes on the second level. The number of nodes on the third level separates these clusters. The depth separates the clusters C3, C4, C5 (II) from C1, C2 (I).

Clusters C6, C7, and C8 have identical depth (3) and four nodes on the second level. The number of nodes on the third level separates them. The number of nodes on the second level separates C6, C7, C8 (III) from C3, C4, C5 (II). C6, C7, C8 (III) and C1, C2 (I) are separated by depth.

The visual leaning separating the clusters C1 and C2 suggests that not the reflection of $G6$ (cf. Figure 1) itself was apparent to the participants but rather a property which changed, that is, the leaning (cf. Section 3.2). This is contrary to the clusters C5 and C7. Cluster C5 clearly shows that neither the reflection of $G2$ (cf. Figure 1) itself nor a changed property mattered. It seems that purely the number of nodes dominates over, e.g., node position (2 left, 1 right vs. reflected: 1 left, 2 right). C7 shows that also the reflection of $G4$ itself (cf. Figure 1) or a changed property, e.g., node position, did not matter. It is remarkable that sometimes the impact of isomorphism resp. the visual layout is recognized based on a property that changed because of that and sometimes it seems to be dominated by another factor. All in all, this clearly shows that humans do not tend to discover that two graphs having different visual layout

C	$\overline{\text{Ease}}$	$\overline{\text{Conf}}$	DAG Properties
C1	4.3	4.2	• *depth*: 4 • *number of nodes on level 2*: 4; *on level 3:* 3; *on level 4:* 1 • *leaning*: left
C2	4.4	4.3	• *depth*: 4 • *number of nodes on level 2*: 4; *on level 3:* 3; *on level 4:* 1 • *leaning*: right
C3	4.1	4.0	• *depth*: 3 • *number of nodes on level 2*: 3; *on level 3:* 4
C4	4.1	4.1	• *depth*: 3 • *number of nodes on level 2*: 3; *on level 3:* 2
C5	3.6	3.8	• *depth*: 3 • *number of nodes on level 2*: 3; *on level 3:* 3
C6	3.7	3.7	• *depth*: 3 • *number of nodes on level 2*: 4; *on level 3:* 4
C7	3.6	3.8	• *depth*: 3 • *number of nodes on level 2*: 4; *on level 3:* 3
C8	3.8	3.9	• *depth*: 3 • *number of nodes on level 2*: 4; *on level 3:* 2

Table 2: Properties of the DAGs in the clusters C1-C8 along with average easiness ($\overline{\text{Ease}}$) and confidence ($\overline{\text{Conf}}$) values for each cluster.

have the same structure and are thus isomorphic; i.e. identical graphs. However, recognizing the structure would be really relevant for an analytical comparison thus the relevant information resides in the fact that the structure is the same and not that the visual appearance – e.g., the skewness – is different.

Interestingly, *edges* and *edge crossings* – important factors of graph reading and graph aesthetics – seem not to matter to the participants. The DAGs of C3 and C5 in Table 1 clearly show that the edges had no influence on the similarity judgment of the participants. Otherwise, DAGs with such different structure would not have been grouped. C7 shows that the participants also did not really care about edge crossings.

To conclude, we consider the hierarchical clusters to have high consistency regarding graph theoretical and visual DAG properties. They are also well objectifiable with these properties.

4.2 Qualitative Analysis (RQ2)

We performed a thematic analysis of the participants' tags to reveal the factors they considered. We also analyzed the factors' importance based on the number of mentions of a specific factor. For this analysis, we used the data of all 20 participants since the thematic analysis is not that reactive to the participants

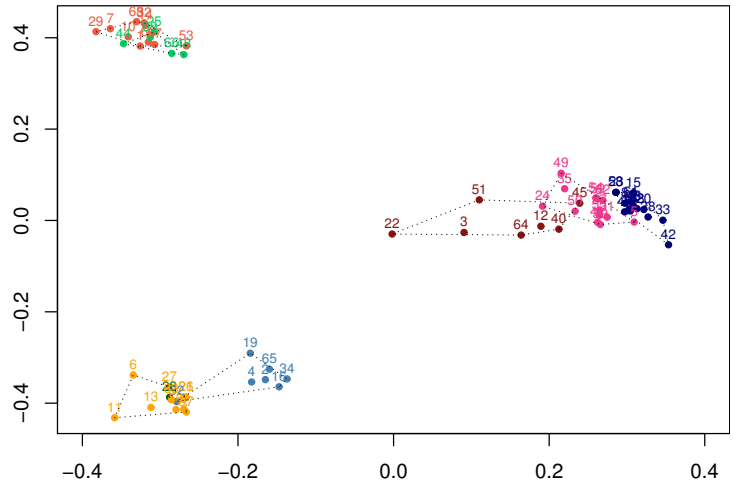

Figure 3: Multi-dimensional scaling (MDS) plot of the perceptual distance matrix (PDM). The clusters resulting from the hierarchical clustering are color-coded and surrounded by their convex hull. (Dendrogram (cf. Figure 2), Table 1, and MDS plot use the same color code for the respective cluster.

forgetting to group some data items. Since the four participants had forgotten less than ten DAGs (10 DAGs= 14.5% of the entire dataset), we could be certain that the tags provided by the participants still result from the to be grouped dataset and are not dependent on whether a DAG was grouped or not.

Analysis. First, we literally transcribed the participants' tags by noting each tag together with how the participant used it, e.g., in a hierarchical manner. Additionally, we collected the following data for the tags (henceforth called factors) of each participant:

- *factor type* – visual, graph theoretical, no type
- *combined vs. single factors* – e.g., number of levels vs. number of levels *and* number of nodes
- *number of distinct considered factors*
- *number of values per factor* – e.g., number of edge crossings = 1, 2 and 3 → number of values = >1 value per factor

We deemed the factor type as important since the graph theoretical properties are those which contain the information relevant for comparison insights due to them describing the DAGs per se. From graph readability research, we already know that visual factors – e.g., edge crossing – have significant influence. However, knowing these for visual comparison is beneficial for controlling their

Factors used by at least 20% of the participants	
number of levels	12
number of nodes on a specific level	9
shape	9
arm/branch (DAG sub-shape)	4
edge crossing	4
child node(s) with > 1 parent node	4
leaning	4
Factors used by less than 20% of the participants	
one parent node	3
visual symmetry (entire DAG)	3
number of nodes in the entire DAG	2
node position/layout	2
level style	2
graph type	2
DAG appears nearly full	2
leave nodes on higher level than the lowest level	2
number of edges to level $ID + 1$	2
visually approximated number of nodes in branch	1
one root-like node	1
cycle	1
long edges	1
outlier subgraph (self-defined)	1
visual symmetry (edges)	1
hierarchy violations (self-defined)	1
isomorphic	1
balance	1
"Other"	1
"Not classifiable"	1

Table 3: Number of mentions of the factors used by the participants to form the groups (■ : graph theoretical, ■ : visual, ■ : no type). Multiple mentions of the same factor by the same participant were excluded. When a participant used a combined factor it was a combination of a subset of the 27 factors listed here.

influence. We introduced the category 'no type' for factors which do not denote specific properties of a DAG. We collected the other data as meta-information on the factors the participants used to learn more about the participants' usage of the factors.
To unify the participants' wording we used an open coding procedure. Four coders abstracted the participants individual wording to more general factor denoting concepts. After the individual coding phase, the coders achieved a coding consensus by discussing their individual coding results.

For the analysis of the factor importance we counted for every factor how many participants used it. Multiple mentions o the same factor by one participant were not considered.

Results. The individual literal transcriptions can be found on our website[1]. Table 3 shows the factors considered by the participants together with how often a factor was named. In total, our participants used 27 distinct factors. Ten of them can be considered as graph theoretical factors (■) and 15 as visual factors (■). Two of the used factors are neither graph theoretical nor visual (■). The participants used descriptive wording such as *"others"* or *"not classifiable"* for these factors. Both of these factors had the purpose of distinguishing the DAGs within the group tagged with such a factor from the other groups.
Just five out of 20 participants used a combined factor and only two out of these five used more than one combined factor. The most frequently combined factor was *number of nodes on a specific level* (five times). The combined factors were constructed from the 27 distinct factors by using conjunctions like *or* or *and* ; e.g., number of nodes on the second level = 3 *and* number of nodes on the third level = 4.
Seven of the 27 factors were used by at least 20% of the participants (cf. Table 3, top). We will focus on these seven. For the other 20 factors, please refer to Table 3, bottom.

The most important factors according to usage frequency were: *number of levels* (i.e., depth of the DAG), *number of nodes on a specific level*, *shape*, *arm/branch* (≡ *DAG sub-shape)*, *edge crossing*, *child node(s) with* > 1 *parent node*, *visual leaning (left:*  *, right:* *)*.
The factor *shape* is basically the convex hull of the DAG (). Regarding *shape*, it is interesting to note that we could observe a correlation of *shape* with the *number of nodes on a specific level*. Participants, for instance, denoted a DAG such as  as "narrow/small pyramid" and a DAG such as  as "wide/large pyramid". However, it is clear that this coherence is also influenced by the DAGs' layout. *Arm/branch* refers to the *shape of a DAG's sub-graph* ().
Edge crossing deals with crossings of the visualized edges (). The participants considered different types of edge crossings, e.g., the mere presence of edge crossing or (un)resolvable edge crossings.
The factor *child node(s) with* > 1 *parent node* relates to the number of nodes which are the parent of another node ( , ). Again, we could observe that participants used different types of these factors; e.g., the mere existence of nodes with greater one parent node or the number of nodes in a DAG which have greater one parent node.

Interestingly, the extracted factors also substantiate that edges and edge crossings are immaterial to humans comparing DAGs. This confirms the findings of Section 4.1. The factor *edge crossing* is one of the least used of the most important factors. Other edge related factors, e.g., the visual edge length (*long edges*), were used just once (cf. Table 3, bottom). Various individual groupings also support the absence of recognizable influence of edge related factors, e.g.:  (factor: one parent left).

5 Extended Analysis

In addition to understanding the individual factors, we also deem it important to understand:

(1) *the specifics of the factors*; e.g., the average number of factors the participants consider while judging the similarity.

(2) *the strategies the participants employ to manage the dataset and come to a final similarity judgment*; e.g., grouping the dataset according to one factor and then grouping the resulting groups into further sub-groups.

Understanding the specifics of the factors helps us to learn the details of the human mental model; e.g., the average number of used factors over all participants tells us about the limit on similarity perception-influencing factors that humans consider for their judgment. Detailed knowledge about the human mental model is valuable for future perception-aware mathematical similarity functions; thus it tells us, i.a., how many factors such a function should consider. Understanding these strategies helps to offer useful interactions with comparative visualizations since the interaction is the means by which the users can apply their task-solving strategies to the data. To gain initial insights regarding these subsequent questions, we performed this extended analysis.

5.1 Factor Specifics Analysis

We assume the following specifics to be relevant:

(1) the limit on similarity perception-influencing factors humans consider for their judgment,

(2) whether graph theoretical or visual factors are more dominant respectively more frequently used.
Graph theoretical factors result from the data structure per se, and visual factors result from visualizing the DAGs as node-link diagrams. Examples include edge crossings (visual) and the depth of the DAG (graph theoretical).

(3) how humans think about the factors they use.

Therefore, we analyzed the following for the transcription data of our thematic analysis (cf. Section 4.2, Analysis):

(1) the average number of factors the participants used,

(2) tendencies regarding the factor type – e.g., is one factor type used more often than others,

(3) the value range of the factors – boolean (true/false), a single value per factor or more than one value per factor but not a boolean value; henceforth denoted as *boolean*, *1 value*, and > 1 *value*.

If they choose a boolean value range for their factor, they more likely think about the existence or non-existence of the property the factor denotes. If they attribute concrete values to the factor, they more likely do an analytical analysis promising more detailed knowledge about and, therefore, more precise insights based on the to be compared data.

Analysis. First, we collected the total number of distinct factors each participant used (cf. Table 4 column #). In addition to that, we collected information on how many factors of each type – graph theoretical, visual, no type (cf. Section 4.2) – each participant used (cf. Table 4 columns graph th., visual, no type). We then calculated descriptive statistics for this data: mean, median, standard deviation.
We also analyzed how many participants used just one factor type – graph theoretical, visual, no type – and determined which factor type it was. Since it can happen that participants used just one factor type but more than a single factor, we analyzed how many participants used just a single factor. For the single factor, we were also interested in the factor's type.

Second, we collected the distinct value range types the participants used(cf. Table 4 boolean, 1 value, > 1 value). These were *boolean*, *1 value*, > 1 *value*. A value range of a factor is *boolean*, if it is only concerned with whether a property is present or not; e.g., *There are nodes on the graph which only have one predecessor* ($\equiv$ *one parent node*) (cf. Figure 4 **(2)**, factor (4)). A factor has a *1 value* value range, if the participant focuses on one specific value and the value is not boolean; e.g., *number of layers* $= 4$ (cf. Figure 4 **(2)**, factor (2)). It would have been possible to interpret the *1 value* value range as *boolean* since it also depends on whether the factor has the specific value or not, but we kept it as a separate type due to the participants, in this case, respecting a specific value whereas for the boolean factors they were just concerned with the (non-)existence. We consider a factor to have a > 1 *value* value range, if the participant attributed more than one concrete value to the factor and if these values are not boolean. An example of this is the factor *number of nodes on level 2* $= 3, 4$ of participant 15 (cf. Figure 4 **(2)**, factor (1), (2)). We then collected the total number of each distinct value range type for each participant. Subsequently, we calculated the sum, mean, median and standard deviation as descriptive quantitative values respectively statistics for this data.

As the final step in this analysis, we determined how often each value range type was used for each of the 27 distinct factors we identified with our thematic analysis (cf. Section 4.2).

Results. The participants used on average 4 factors ($median = 4.0, mean = 3.8, STD = 1.7$). Two of them are graph theoretical factors and two are visual factors (graph theoretical – $mean = 2.0, median = 2.0, STD = 1.6$; visual – $median = 1.5, mean = 1.7, STD = 1.1$; paired t-test over all participants and their number of used visual and graph theoretical factors: $t(19) = 0.6227, p = 0.5409 \rightarrow$ no statistically significant difference). Thus, there is no clearly visi-

	boolean	**1 value**	**>1 value**	**Σ**	**graph th.**	**visual**	**no type**	**#†**
P01	1	0	4	5	4	1	0	5
P02	4	1	1	6	3	3	0	6
P03*	3	0	1	4	1	2	0	3
P04	0	0	1	1	0	1	0	1
P05	0	0	3	3	2	1	0	3
P06	2	2	2	6	5	1	0	6
P07	5	0	0	5	1	4	0	5
P08	2	0	3	5	3	2	0	5
P09	0	0	1	1	0	1	0	1
P10	4	0	1	5	4	1	0	5
P11	0	0	3	3	3	0	0	3
P12	0	0	2	2	1	1	0	2
P13*	2	0	3	5	0	3	1	4
P14	1	0	3	4	2	1	1	4
P15	5	1	1	7	5	2	0	7
P16	4	0	0	4	1	3	0	4
P17	0	0	2	2	0	2	0	2
P18	0	0	2	2	2	0	0	2
P19	0	0	3	3	1	2	0	3
P20	3	0	2	5	2	3	0	5
Σ	36	4	38	78	-	-	-	-
mean	1.8	0.2	1.9	-	2.0	1.7	1.0	3.8
median	1.5	0.0	2.0	-	2.0	1.5	1.0	4.0

*Participant used factor with two distinct value ranges, therefore she was considered for both value range types ⇒ sum (Σ) > #

†number of distinct factors

Table 4: Number of distinct factors (#), number of factor types – graph theoretical (graph th.) ( ), visual ( ), and no type ( ) and number of value range types – *boolean* ( ), *1 value* ( ), > 1 *value* ( ) – per participant

ble tendency of whether a person would rather use visual or graph theoretical factors, whereas regarding all 27 distinct factors we could observe a clear dominance of visual factors ($15 \times$ ■) over graph theoretical factors ($10 \times$ ■) (cf. Section 4.2, Table 3). Just two participants used a factor which is neither graph theoretical nor visual. Both of these factors had the purpose of distinguishing the DAGs within this group from the other groups. Six participants used just one factor type. Four of those six participants just used visual factors and two just used graph theoretical factors. Two of our participants used even just a single factor and for both it was a visual factor (*shape*).

In total, a *boolean* value range and a > 1 *value* value range were used approximately equally often; 36 (*boolean*) resp. 38 (> 1 *value*) times (cf. Table 4). So, one could think the existence-[2] and the analytically-coined[3] mental models are approximately equally distributed in the humans' mind. The mean and the median for these two value ranges support that hypothesis – *boolean*: $mean = 1.8, median = 1.5, STD = 1.8$, > 1 *value*: $mean = 1.9, median = 2.0, STD = 1.1$. However, when we look at the value ranges' frequency per distinct factor (cf. Table 5), the aforementioned facts and possible hypothesis are put into perspective. Here, we see that the existence-coined mental model is well-distributed over almost all factors whereas the analytically-coined mental model is used for only 10 factors. This is a crucial insight since the analytical information, relevant for a comparison, generally lies less in the existence of a property than in the concrete values of a property. Interestingly, the usage of the > 1 *value* value range tends to accumulate for the factors which influenced at least 20% of our participants (cf. Table 5). The *1 value* value range seems to be more of an outlier – it was used only four times (cf. Table 4 and 5).

5.2 Strategy Analysis

How the participants combined their similarity perception-influencing factors as tags for their similarity groups reflects the strategy they employed to come to a final similarity judgment. A strategy is defined as a sequence of processes to solve a task [31]. In the context of similarity judgment, this means a sequence of employing several influencing factors to come to a final similarity judgment in a specific manner, e.g., hierarchical or one after the other. We did a qualitative content analysis [49] of the transcription data to reveal the strategies for completing the task of similarity judgment.

As part of the strategy analysis, we were also interested in:

(1) *whether participants who employ different strategies employ the same or different factors.*

(2) *the dominant factors for the specific strategies.*
Here, dominant means whether there are factors used recognizably more often than others by participants employing a specific strategy.

[2] only thinking about the (non-)existence of a factor expressing that with *boolean* value ranges

[3] considering the concrete values of a factor expressing that with *(>)1 value* value ranges

		strategies							value range		
		divide&conquer									
27 distinct factors used by the participants		**level 1**	**level 2**	**level 3**	**level 4**	**level 5**	**sequence†**	**single factors**	**boolean**	**1 value**	**>1 value**
Factors used by at least 20% of the participants											
number of levels	■	7	1	0	0	0	5	0	0	2	10
number of nodes on a specific level	■	0	4	4	1	0	3	0	0	1	8
shape	■	1	1	0	1	0	6	2	2	0	7
arm/branch (DAG sub-shape)	■	0	0	0	0	0	4	0	4	0	0
edge crossing	■	0	1	0	0	0	3	0	4	0	0
child node(s) with > 1 parent node	■	0	1	0	0	0	3	0	3	0	1
leaning	■	0	3	0	0	0	1	0	0	0	4
Factors used by less than 20% of the participants											
one parent node	■	0	0	0	0	0	3	0	3	0	0
visual symmetry (entire DAG)	■	0	0	1	0	0	2	0	3	0	0
number of nodes in the entire DAG	■	0	1	0	0	0	1	0	0	0	2
node position/layout	■	0	1	0	0	0	1	0	1	0	2
level style	■	0	2	0	0	0	0	0	0	0	2
graph type	■	0	1	0	0	0	1	0	2	0	0
DAG appears nearly full	■	1	0	0	0	1	1	0	2	0	0
leave nodes on higher level than the lowest level	■	0	0	0	1	0	1	0	2	0	0
number of edges to level $ID + 1$	■	0	1	1	0	0	1	0	0	1	1
visually approximated number of nodes in branch	■	0	0	0	0	0	1	0	1	0	0
one root-like node	■	1	0	0	0	0	0	0	1	0	0
cycle	■	0	0	0	0	0	1	0	1	0	0
long edges	■	0	0	0	0	0	1	0	1	0	0
outlier subgraph (self-defined)	■	0	0	0	0	0	1	0	1	0	0
visual symmetry (edges)	■	0	0	0	0	0	1	0	1	0	0
hierarchy violations (self-defined)	■	1	0	0	0	0	0	0	0	0	1
isomorphic	■	0	0	0	0	0	1	0	1	0	0
balance	■	0	0	1	0	0	0	0	1	0	0
"Other"	■	0	0	0	0	0	1	0	1	0	0
"Not classifiable"	■	0	0	0	0	0	1	0	1	0	0

†respecting entire dataset and considering factors one after the other
level 1 = lowest level, level 5 = highest level

Table 5: Participants' usage of the 27 identified distinct factors (■ : graph theoretical, ■ : visual, ■ : no type) together with the three identified strategies: *divide and conquer* (split into the strategy hierarchy levels), *respecting the entire dataset and considering factors one after the other*, *considering a single factor*. Additionally, the frequency of the value ranges – *boolean*, *1 value*, > 1 *value* – that the participants chose for the 27 distinct factors is shown.

(3) *whether we can gain insights on the importance of the factors relative to each other.*
We understand the relative factor importance as something equivalent to the ranking of visual variables introduced by Bertin [5].

To address these questions we analyzed:

(1) which of the 27 distinct factors, identified in Section 4.2, are used in combination with which strategy,

(2) the usage frequency per factor per strategy, and

(3) the usage frequency per factor over all strategies for all 27 distinct factors.

Analysis. The participants' strategies were analyzed according to qualitative content analysis [49]. This method is especially used for studying verbal material, but also for other types of documents. It is a systematic method based on a system of categories – a coding scheme – which can be developed either before the process of analysis (top-down) or during this process (bottom-up) [48]. We used the bottom-up approach because there is no previous research in this area and the investigation of the strategies is exploratory. The elements of the strategies are the factors identified in the qualitative analysis described in Section 4.2. For every participant we did the following: Based on the transcriptions made during our qualitative analysis and the subset of the 27 distinct factors the participant was influenced by while judging the DAGs' similarity (cf. Section 4.2), we derived how the participant combined the factors to come to a final similarity judgment by objectifying the participant's groups with this subset of factors. By "how", we mean the order and the structure of the factors. Order means which factor was used first and which factors followed in which order. Structure means whether the participant used the factors for instance in a hierarchical manner or just one after the other. The derivation is done by respecting all the groups of the participant individually and testing the factors' order and structure combinations to find the order which produces exactly the groups of the participant.
We recorded the results as an ordered list and as a pictogram (cf. Figure 4 **(Ordered List)**, **(Pictogram)**). The ordered list represents the order and the structure of the factors used by the participant. The pictogram represents how the participant dealt with the dataset and the structure of the grouped dataset. The pictogram shows the entire dataset as a black square and the groups resulting from a specific factor as colored squares. A different color denotes that the groups result from different factors.

For naming the strategies, we used the results of the afore-explained analysis procedure. The central aspects for the naming were the structure of the participants' factors and how they dealt with the dataset; e.g., partitioning it and dealing with one sub-dataset at a time or always respecting the entire dataset.

To investigate our three more in-depth questions regarding the strategies, we analyzed which of the 27 distinct factors is used together with which strategy. For a strategy with a hierarchical factor structure, we analyzed which of

the identified distinct factors is used on which strategy hierarchy level by the participants. The hierarchical factor structure is represented in the **(Ordered List)** of Figure 4 by a change in the (sub-)bullet point style. The first hierarchy level uses numerical values in brackets, e.g., `(1)`. The second hierarchy level is identified by small Roman numbers, e.g., `ii.`. The third level uses again numerical values, e.g., `1.`. As a consequence, we could quantify which factor is used how often in combination with which strategy. For this analysis, we used the same data as for the identification of the strategies.

Results. We could observe that the participants used three distinct strategies to judge the similarity of DAGs: *divide and conquer*, *respecting the entire dataset and considering the factors one after the other*, and *considering a single factor*. Eleven participants chose one factor, grouped the entire dataset according to it, and then grouped the resulting groups into further sub-groups (*divide-and-conquer* strategy, cf. Figure 4 **(1)**). There were nine participants who always *respected the entire dataset considering the factors one after the other* (cf. Figure 4 **(2)**). For this strategy, the groups will stay the same if the factors are applied in a permuted order since the entire dataset is always considered. Some of the participants chose all their factors in advance. Others chose their factors in an ad hoc fashion; meaning, after having grouped the dataset according to a factor they thought about the next. Finally, there were two participants who did their grouping by considering just *one single factor* (cf. Figure 4 **(3)**). Participants employing the *divide and conquer* strategy used strategy hierarchies with an average depth of three ($mean = 2.8, median = 3.0, STD = 0.9$). Participants who *respected the entire dataset considering one factor after the other* employed on average five factors ($mean = 5.1, median = 5.0, STD = 2.1$). The two participants who *considered only one single factor* used the factor *shape*. One of these two participants used four different values of *shape* to do her grouping, and the other, eight.

In addition to the strategies we could observe the following details: There were two participants who did a combination of the *divide and conquer* strategy and the strategy of *respecting the entire dataset and considering the factors one after the other*. We counted these two participants towards both strategies. Three participants performed their *divide and conquer* strategy with the same factor per strategy hierarchy level, in other words, they varied factors across hierarchy levels but not within. This is contrary to the behavior of the rest of the participants who employed the *divide and conquer* strategy. These participants had a strong tendency to vary their factors across and within strategy hierarchy levels, especially within the second level. An example for this is participant P1; shown in Figure 4 **(1)**. P1 used just one factor for her first and third strategy hierarchy level – hierarchy level 1: *(1) number of levels*, hierarchy level 3: *1. and 2. number of nodes on a specific level* (cf. Figure 4 **(1)**). However, on her second hierarchy level she used three factors: *child node(s) with* > 1 *parent node*, *graph type*, and *number of nodes on a specific level* (cf. Figure 4 **(1)**, factors i., ii., iii.). Two of the three participants, who conducted their *divide*

and conquer strategy with the same factor per strategy hierarchy level, varied the value range type, cf. Section 5.1, within the strategy hierarchy levels as P1 did it on the third strategy hierarchy level. Just one participant kept the factors and value range types the same. Two of the 11 participants employing the *divide and conquer* strategy used non-discriminating factors at a certain strategy hierarchy level. Non-discriminating factors do not divide the entire dataset or the already existing groups into further sub-groups; for instance:  factor 1 discriminates the dataset, however, factor 2 does not further discriminate the dataset than the already existing groups. One of these two participants employed three non-discriminating factors and the other one two.

Table 5 summarizes the results for our three more in-depth questions regarding the strategies. These were: (1) *Are the same factors, or rather different ones, used across the strategies?*, (2) *Are there dominant factors regarding a specific strategy?*, (3) *Are we able to gain initial insights on the factors' importance relative to each other?*

From Table 5 it is evident that the factors used across the three strategies are comparable. It is not the case that one of the three strategies completely disregards some factors which, in turn, another strategy frequently respects. However, it also becomes apparent from Table 5 that more of the 27 factors are used in the *respecting the entire dataset and considering the factors one after the other* strategy than in the others. This observation can be related back to the tendency of participants who employed this strategy to use more factors; on average five compared to the three of the *divide and conquer* strategy.

The dominant factors of all three strategies reflect those we identified in our qualitative and quantitative analysis: *number of levels*, *number of nodes on a specific level*, *shape*, *arm/branch* ($\equiv$ *DAG sub-shape)*, *edge crossing*, *child node(s) with* > 1 *parent node*, *leaning*, *one parent node* (cf. Section 4).

The insights regarding a relative importance ranking of the factors are just very circumstantial. As can be seen from Table 5, the factor *number of levels* is most frequently used as the factor of first strategy hierarchy level. On the second and third strategy hierarchy level, the most frequent factors are *number of nodes on a specific level* and *leaning*. This relates well to the most frequent factors of the *respecting the entire dataset and considering the factors one after the other* strategy. The most frequent factors of this strategy are: *number of levels*, *number of nodes on a specific level*, *shape*, *arm/branch* ($\equiv$ *DAG sub-shape)*, *edge crossing*, *child node(s) with* > 1 *parent node*, *one parent node*. However, the frequency of the most frequent factors does not differ in total that much from the less frequent factors, therefore we are convinced that more thorough investigations with user studies specifically designed for this purpose are necessary to reveal the relative importance of the factors.

6 Result Summary, Discussion, and Future Work

We conducted a card sorting study to identify the factors influencing the similarity perception of DAGs. Herewith, we mitigate the present knowledge gap

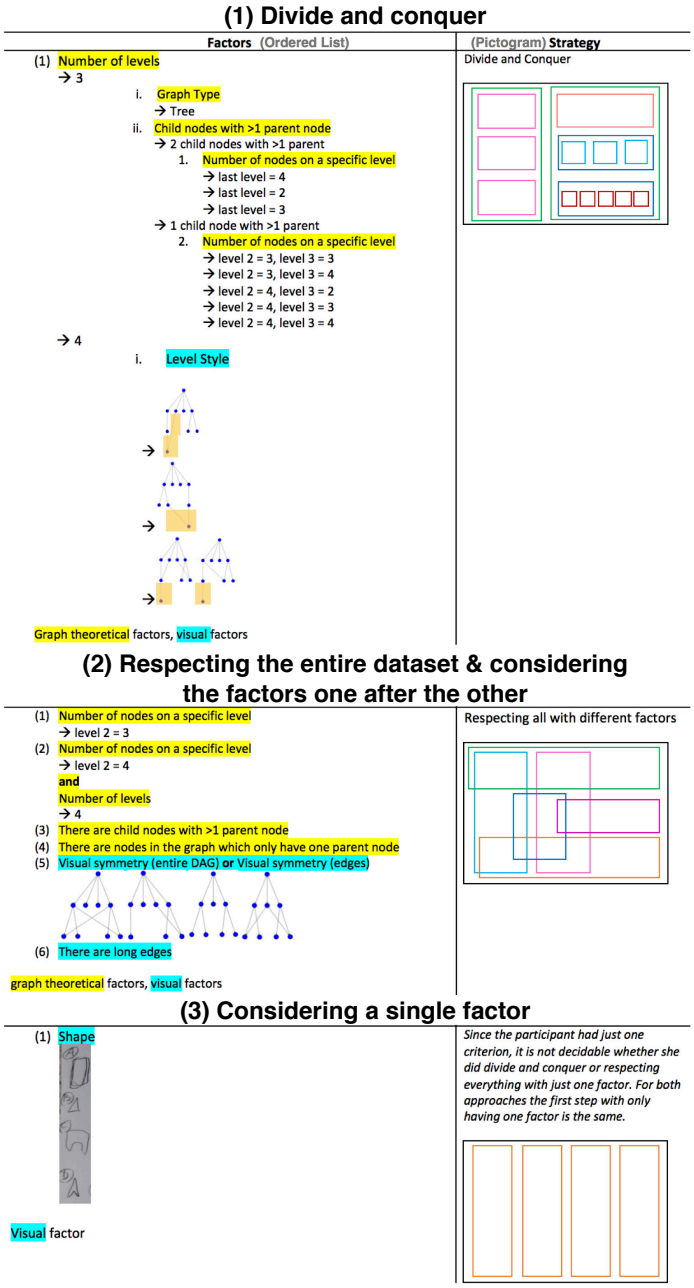

Figure 4: The three distinct strategies the participants employed to solve the task of judging the similarity presented on the example of three individual participants: (1) *divide and conquer*, (2) *respecting the entire dataset and considering the factors one after the other*, and (3) *considering a single factor*.

regarding this topic despite the vast presence of visual graph comparison tasks in various disciplines, e.g., finance, biology.

6.1 Research Questions: Formed Similarity Groups (RQ1) & Influencing Factors (RQ2)

Result Summary. Both the results of our quantitative (*RQ1*, *RQ2*) and qualitative (*RQ2*) analysis point to *similar* factors which seem to dominantly influence similarity perception of DAGs, namely the *number of levels (depth)*, the *number of nodes on specific levels*, and *shape-related aspects* such as the visual leaning of a DAG. The strong influence of shape is remarkable as in our case the spatial arrangement did not convey any additional information. This resulted in cases where structurally identical DAGs were assigned to different groups due to one being left-skewed or right-skewed. Being skewed to the left or right mainly played a role for the 4-level DAGs (cf. C1 and C2), most likely because it had a stronger influence on the overall shape than in the 3-level cases. Nevertheless, this observation supports previous results which found evidence that perception of graphs is sensitive to its spatial layout ([23, 34]). Surprisingly, edge crossings – an important factor concerning the readability of graphs [43] – contrary to our expectations did not seem to have a strong influence on perceived DAG similarity. This lack of influence is, for example, evident in the clusters C5 and C6 where no distinction between DAGs with and without edge crossings was made (cf. Table 1). In the participants' statements, we found evidence that they did not subconsciously resolve the edge crossing and therefore did not mention edge factors. On the contrary, the edges were not in the focus of the participants.

Discussion. The fixed order of our data items did not lead to order-influenced groupings. The individual groupings and the consensus grouping are well objectifiable with DAG properties and do not show signs for groupings influenced by the order of the DAGs on the data sheet. The quantitative analysis shows the objectifiability of the consensus grouping (cf. Section 4.1). We analyzed the individual groupings by checking the objectifiability of grouped consecutive data items (see our website[1] for details). We considered the analysis of grouped consecutive data items as a valid analysis instrument since grouping the DAGs based on their ID sequences is the most apparent option for an order-influenced grouping. We analyzed the participants' groupings for simple sequences (successor ID = predecessor ID +1, +2, +3 and +4). Since we did not find such simple order-influenced sequences, we can be sufficiently certain that more complex order-influenced sequences are also not present and that the individual groupings as well as our similarity consensus grouping result are based on reasonable factors of the participants. For the complete analysis, please refer to our supplementary material.

Future Work. In future work, it will be necessary to investigate how the identified factors and their importance varies across different graph sizes. It is, for instance, reasonable to assume that, for larger graphs, factors concerning details of a graph (e.g., number of parent nodes, number of nodes on a specific

layer) decrease in importance while factors concerning the overall appearance (e.g., shape) increase. Regardless of that, our study provides first results which can contribute to the design of comparative visualizations. Moreover, a better understanding of the factors which drive humans' similarity judgment may also be used towards developing perception-based graph similarity measures. Current notions of graph similarity such as graph isomorphism and edit distance (cf. [14]), descriptive statistics of graph structure measures such as degree distribution or diameter, or iterative approaches which assess the similarity of the neighborhood of nodes (e.g., [24, 26, 35]) rely purely on graph theoretical properties.

6.2 Extended Analysis: Factor Specifics & Strategy Analysis

Result Summary. From our extended analysis, we found that the averaged limit of factors humans consider for their similarity judgment is four factors and that there is no clear tendency whether the human similarity perception is driven by visual or graph theoretical factors if we focus on participants individually. This reveals an interesting divergence from the results of our analysis in Section 4. There, we could see a clear dominance of visual factors – 15 visual factors (■) and ten graph theoretical factors (■) (cf. Table 3). We further found initial insights that humans rather consider the pure existence of a similarity influencing factor than its concrete manifestations (cf. Table 5). Regarding an analytical comparison, this habit promises less detailed and precise knowledge about the data and consequently less detailed and precise insights based on the data. At this point, we were not able to discover the factors' relative importance to each other. Nevertheless, we found that the participants used three distinct strategies to solve the task of judging the similarity: *divide and conquer*, *respecting the entire dataset and considering the factors one after the other*, and *considering a single factor*.

Discussion. These results provide further relevant details for future perception-aware similarity measures. Perception aware similarity measures can draw, e.g., the following information from our results: the number of to be considered factors, the type – visual or graph theoretical – of factors which should be considered, the factors' value range, and which factors should be chosen from the list of potential factors with respect to the weighting factor of the respective factors. Moreover, these results help to improve comparative visualizations, e.g., by showing us which factors are rather overlooked by humans and consequently should be highlighted in comparative visualizations. Furthermore, they inform the choice and development of useful techniques for interacting with the comparative visualizations since the interaction is the means by which the users should be able to apply their task-solving strategies to the data.

Future Work. Still, at the moment these are initial or even circumstantial insights. Therefore, we aim at verifying and detailing the afore-discussed aspects in our future work. An example is the three similarity judgment strategies. Although our analysis result clearly reveals these three strategies, we think

more thorough investigations are necessary to verify these results. One way would be a user study specifically designed for the investigation of similarity judgment strategies.

7 Conclusion

To conclude, we consider the similarity perception of DAGs in visual comparison across people as consistent and well objectifiable using graph theoretical or visual properties. We find the objectifiability substantiated by our quantitative and qualitative analysis. Also, we find that humans employ a structured approach – a strategy – to judge the similarity of DAGs. Furthermore, we find that the specifics of the used factors – e.g., their value ranges – provide pivotal detail on the mental model of humans regarding the perceived similarity of DAGs.

Acknowledgments

We thank our participants who gave us their precious time.

References

[1] D. Archambault. Structural differences between two graphs through hierarchies. In *Proc. GI*, pages 87–94. Canadian Information Processing Society, 2009.

[2] D. Archambault, H. C. Purchase, and B. Pinaud. Difference map readability for dynamic graphs. In U. Brandes and S. Cornelsen, editors, *Graph Drawing: 18th International Symposium, GD 2010*, pages 50–61. Springer, 2011. doi:10.1007/978-3-642-18469-7_5.

[3] B. Bach, E. Pietriga, and J.-D. Fekete. Graphdiaries: Animated transitions and temporal navigation for dynamic networks. *IEEE Trans. Vis. Comput. Graphics*, 20(5):740–754, 2014. doi:10.1109/Tvcg.2013.254.

[4] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. The state of the art in visualizing dynamic graphs. In *Proc. EuroVis - STARs*, 2014.

[5] J. Bertin. *Semiology of Graphics*. University of Wisconsin Press, 1983.

[6] S. Bremm, T. Von Landesberger, M. Heß, T. Schreck, P. Weil, and K. Hamacher. Interactive visual comparison of multiple trees. In *Proc. IEEE VAST*, pages 31–40, 2011. doi:10.1109/VAST.2011.6102439.

[7] S. Bridgeman and R. Tamassia. Difference metrics for interactive orthogonal graph drawing algorithms. In S. H. Whitesides, editor, *Graph Drawing*, pages 57–71, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. doi:10.1007/3-540-37623-2_5.

[8] M. Burch, N. Konevtsova, J. Heinrich, M. Hoeferlin, and D. Weiskopf. Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2440–2448, Dec 2011. doi:10.1109/TVCG.2011.193.

[9] B. S. Chaparro, V. D. Hinkle, and S. K. Riley. The usability of computerized card sorting: A comparison of three applications by researchers and end users. *J. Usability Stud.*, 4(1):31–48, 2008.

[10] C. M. Collins and S. Carpendale. VisLink: Revealing relationships amongst visualizations. *IEEE Trans. Vis. Comput. Graphics*, 13(6):1192–1199, 2007. doi:10.1109/TVCG.2007.70611.

[11] T. Dwyer, B. Lee, D. Fisher, K. I. Quinn, P. Isenberg, G. Robertson, and C. North. A comparison of user-generated and automatic graph layouts. *IEEE Trans. Vis. Comput. Graphics*, 15(6):961–968, 2009. doi:10.1109/TVCG.2009.109.

[12] S. N. Friel, F. R. Curcio, and G. W. Bright. Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in mathematics Education*, 32(2):124–158, 2001. doi:10.2307/749671.

[13] J. Fuchs, P. Isenberg, A. Bezerianos, F. Fischer, and E. Bertini. The influence of contour on similarity perception of star glyphs. *IEEE Trans. Vis. Comput. Graphics*, 20(12):2251–2260, 2014. doi:10.1109/TVCG.2014.2346426.

[14] X. Gao, B. Xiao, D. Tao, and X. Li. A survey of graph edit distance. *Pattern Anal. Appl.*, 13(1):113–129, 2010. doi:10.1007/s10044-008-0141-y.

[15] S. Ghani, N. Elmqvist, and J. S. Yi. Perception of animated node-link diagrams for dynamic graphs. *Comput. Graph. Forum*, 31(3pt3):1205–1214, 2012. doi:10.1111/j.1467-8659.2012.03113.x.

[16] M. Gleicher. Considerations for visualizing comparison. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):413–423, 2018. doi:10.1109/TVCG.2017.2744199.

[17] M. Gleicher, D. Albers, R. Walker, I. Jusufi, C. D. Hansen, and J. C. Roberts. Visual comparison for information visualization. *Inf. Vis.*, 10(4):289–309, 2011. doi:10.1177/1473871611416549.

[18] G. Greve. Different or alike? Comparing computer-based and paper-based card sorting. *International J. of Strategic Innovative Marketing*, 1(1):27–36, 2014. doi:10.1080/01652176.2014.949390.

[19] S. Hadlak, H. Schumann, and H.-J. Schulz. A survey of multi-faceted graph visualization. In *Proc. EuroVis - STARs*, 2015.

[20] M. Hess, S. Bremm, S. Weissgraeber, K. Hamacher, M. Goesele, J. Wiemeyer, and T. von Landesberger. Visual exploration of parameter influence on phylogenetic trees. *IEEE Computer Graphics and Applications*, 34(2):48–56, 2014. doi:10.1109/MCG.2014.2.

[21] D. Holten and J. J. Van Wijk. Visual comparison of hierarchically organized data. *Comput. Graph. Forum*, 27(3):759–766, 2008. doi:10.1111/j.1467-8659.2008.01205.x.

[22] D. Holten and J. J. van Wijk. A user study on visualizing directed edges in graphs. In *Proc. CHI*, pages 2299–2308, 2009. doi:10.1145/1518701.1519054.

[23] W. Huang, S.-H. Hong, and P. Eades. Layout effects on sociogram perception. In P. Healy and N. S. Nikolov, editors, *Graph Drawing: 13th International Symposium, GD 2005*, pages 262–273. Springer, 2006. doi:10.1007/11618058_24.

[24] G. Jeh and J. Widom. SimRank: A measure of structural-context similarity. In *Proc. KDD*, pages 538–543, 2002. doi:10.1145/775047.775126.

[25] S. Kieffer, T. Dwyer, K. Marriott, and M. Wybrow. HOLA: Human-like orthogonal network layout. *IEEE Trans. Vis. Comput. Graphics*, 22(1):349–358, 2016. doi:10.1109/TVCG.2015.2467451.

[26] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999. doi:10.1145/324133.324140.

[27] A. Klippel, F. Hardisty, and C. Weaver. Star plots: How shape characteristics influence classification tasks. *Cartogr. Geogr. Inf. Sci.*, 36(2):149–163, 2009. doi:10.1559/152304009788188808.

[28] S. G. Kobourov, S. Pupyrev, and B. Saket. Are crossings important for drawing large graphs? In C. Duncan and A. Symvonis, editors, *Graph Drawing: 22nd International Symposium, GD 2014*, pages 234–245. Springer, 2014. doi:10.1007/978-3-662-45803-7_20.

[29] C. Körner. Concepts and misconceptions in comprehension of hierarchical graphs. *Learn. Instr.*, 15(4):281–296, 2005. doi:10.1016/j.learninstruc.2005.07.003.

[30] R. l. Goldstone and J. Y. Son. Similarity. In K. J. Holyoak and R. G. Morrison, editors, *The Cambridge Handbook of Thinking and Reasoning*. Cambridge University Press, 2005.

[31] P. Lemaire and L. Fabre. Strategic aspects of human cognition: Implications for understanding human reasoning. In M. Roberts and E. Newton, editors, *Methods of though: Individual differences in reasoning strategies*, pages 11–56. Psychology Press, New York, 2005.

[32] O. Lenz, F. Keul, S. Bremm, K. Hamacher, and T. von Landesberger. Visual analysis of patterns in multiple amino acid mutation graphs. In *Proc. IEEE VAST*, pages 93–102, 2014. doi:10.1109/VAST.2014.7042485.

[33] F. McGee and J. Dingliana. An empirical study on the impact of edge bundling on user comprehension of graphs. In *Proc. AVI*, pages 620–627, 2012. doi:10.1145/2254556.2254670.

[34] C. McGrath, J. Blythe, and D. Krackhardt. The effect of spatial arrangement on judgments and errors in interpreting graphs. *Soc. Networks*, 19(3):223–242, 1997. doi:10.1016/S0378-8733(96)00299-7.

[35] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity flooding: A versatile graph matching algorithm and its application to schema matching. In *Proc. ICDE*, pages 117–128, 2002. doi:10.1109/ICDE.2002.994702.

[36] B. Mirel. *Interaction Design for Complex Problem Solving: Developing Useful and Usable Software*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2004.

[37] B. J. Morris and C. D. Schunn. Rethinking logical reasoning skills from a strategy perspective. In E. Newton and M. Roberts, editors, *Methods of thought. Individual differences in reasoning strategies*, pages 2–38. Psychology Press, New York, 2005.

[38] A. Newell and H. A. Simon. *Human Problem Solving*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1972.

[39] L. R. Novick. The importance of both diagrammatic conventions and domain-specific knowledge for diagram literacy in science: The hierarchy as an illustrative case. In D. Barker-Plummer, R. Cox, and N. Swoboda, editors, *Diagrammatic Representation and Inference*, pages 1–11. Springer, 2006. doi:10.1007/11783183_1.

[40] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. CHI*, pages 3659–3669, 2016. doi:10.1145/2858036.2858155.

[41] E. Pekalska and R. P. W. Duin. *The dissimilarity representation for pattern recognition: Foundations and applications*. World Scientific Publishing Co., Inc., River Edge, NJ, USA, 2005.

[42] M. Pohl, G. Wallner, and S. Kriglstein. Using lag-sequential analysis for understanding interaction sequences in visualizations. *International Journal of Human-Computer Studies*, 96(Supplement C):54–66, 2016. doi:10.1016/j.ijhcs.2016.07.006.

[43] H. Purchase. Which aesthetic has the greatest effect on human understanding? In G. DiBattista, editor, *Graph Drawing: 5th International Symposium, GD 1997*, pages 248–261. Springer, 1997. doi:10.1007/3-540-63938-1_67.

[44] H. C. Purchase. Metrics for graph drawing aesthetics. *Journal of Vis. Languages & Computing*, 13(5):501–516, 2002. doi:10.1006/S1045-926x(02)00016-2.

[45] H. C. Purchase, E. Hoggan, and C. Görg. How important is the "mental map"? – an empirical investigation of a dynamic graph layout algorithm. In M. Kaufmann and D. Wagner, editors, *Graph Drawing: 14th International Symposium, GD 2006*, pages 184–195. Springer, 2007. doi:10.1007/978-3-540-70904-6_19.

[46] H. C. Purchase, M. McGill, L. Colpoys, and D. Carrington. Graph drawing aesthetics and the comprehension of UML class diagrams: An empirical study. In *Proceedings of the 2001 Asia-Pacific Symposium on Information Visualisation*, pages 129–137, 2001.

[47] H. C. Purchase, C. Pilcher, and B. Plimmer. Graph drawing aesthetics – created by users, not algorithms. *IEEE Trans. Vis. Comput. Graphics*, 18(1):81–92, 2012. doi:10.1109/TVCG.2010.269.

[48] J. Saldaña. *The Coding Manual for Qualitative Researchers*. SAGE Publications, 2 edition, 2012.

[49] M. Schreier. *Qualitative Content Analysis in Practice*. SAGE Publications, 2012.

[50] K. Sedig, P. Parsons, H.-N. Liang, and J. Morey. Supporting sense-making of complex objects with visualizations: Visibility and complementarity of interactions. *Informatics*, 3(4):20, 2016. doi:10.3390/informatics3040020.

[51] M. Tennekes and E. de Jonge. Tree colors: color schemes for tree-structured data. *IEEE Trans. Vis. Comput. Graphics*, 20(12):2072–2081, 2014. doi:10.1109/TVCG.2014.2346277.

[52] S. Thornley, R. Marshall, S. Wells, and R. Jackson. Using directed acyclic graphs for investigating causal paths for cardiovascular disease. *J. Biometrics Biostatistics*, 4:182, 2013. doi:10.4172/2155-6180.1000182.

[53] R. Tibshirani, G. Walther, and T. Hastie. Estimating the number of clusters in a data set via the gap statistic. *J. R. Stat. Soc. Series B. Stat. Methodol.*, 63(2):411–423, 2001. doi:10.1111/1467-9868.00293.

[54] C. Tominski, C. Forsell, and J. Johansson. Interaction support for visual comparison inspired by natural behavior. *IEEE Trans. Vis. Comput. Graphics*, 18(12):2719–2728, 2012. doi:10.1109/TVCG.2012.237.

[55] S. B. Trickett and J. G. Trafton. Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In D. Barker-Plummer, R. Cox, and N. Swoboda, editors, *Diagrammatic Representation and Inference*, pages 286–300. Springer Berlin Heidelberg, 2006.

[56] C. Vehlow, F. Beck, and D. Weiskopf. The state of the art in visualizing group structures in graphs. In *Proc. EuroVis - STARs*, 2015.

[57] T. von Landesberger, S. Diel, S. Bremm, and D. W. Fellner. Visual analysis of contagion in networks. *Inf. Vis.*, 14(2):93–110, 2015. doi:10.1177/1473871613487087.

[58] T. von Landesberger, A. Kuijper, T. Schreck, J. Kohlhammer, J. van Wijk, J.-D. Fekete, and D. Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Comput. Graph. Forum*, 30(6):1719–1749, 2011. doi:10.1111/j.1467-8659.2011.01898.x.

[59] T. von Landesberger, M. Pohl, G. Wallner, M. Distler, and K. Ballweg. Investigating graph similarity perception: A preliminary study and methodological challenges. In *Proc. VISIGRAPP*, pages 241–250, 2017. doi:10.5220/0006137202410250.

[60] E. Welch and S. Kobourov. Measuring symmetry in drawings of graphs. *Comput. Graph. Forum*, 36(3):341–351, 2017. doi:10.1111/cgf.13192.

[61] J. R. Wood and L. E. Wood. Card sorting: Current practices and beyond. *J. Usability Stud.*, 4(1):1–6, 2008.