Investigating Graph Similarity Perception: A Preliminary Study and Methodological Challenges

Tatiana von Landesberger[1], Margit Pohl[2], Günter Wallner[2], Martin Distler[1] and Kathrin Ballweg[1]

[1]*Interactive Graphics Systems Group (GRIS), Technische Universität Darmstadt, Fraunhoferstrasse 5, Darmstadt, Germany*

[2]*Institute for Design & Assessment of Technology, Vienna University of Technology, Argentinierstrasse 8, Vienna, Austria*

Keywords: Graph Comparison, Evaluation Methodology, Stimulus Development, Perception, Node-link Diagrams.

Abstract: Graphs have become an indispensable model for representing data in a multitude of domains, including biology, business, financing, and social network analysis. In many of these domains humans are repeatedly confronted with the need to visually compare node-link representations of graphs in order to identify their commonalities or differences. Yet, despite its importance little is known about how much visual differences affect users' perception of graph similarity. As a result, more systematic investigations addressing this issue are necessary. However, from a methodological point of view there are still many open questions regarding the investigation of graph comparisons. To this end, this paper provides an overview of methodological challenges, presents results of an explorative study conducted to identify individual factors influencing the recognition of graph differences, and discusses lessons learned from this study. Our considerations and results can serve as foundation for further studies in this area and can contribute to the comparability of these investigations.

1 INTRODUCTION

Humans often use graph visualizations to assess how (dis)similar graphs are (Gleicher et al., 2011; von Landesberger et al., 2009; Von Landesberger et al., 2011). Graphs play an important role in many areas, for example, in biology, in business, financing, or in the analysis of social networks. In all these areas network visualizations have to be compared, either different but similar diagrams or dynamic diagrams at different points in time. It is, therefore, an important question how network visualizations should be designed to support easy and efficient comparison, so that essential differences are not overlooked.

By now, various visualization techniques supporting visual graph comparison have been developed (Gleicher et al., 2011; Von Landesberger et al., 2011). Among the many different types of visualizations for representing graph structures we will focus on node-link diagrams in this paper due to their widespread use and their appropriateness for small graphs (see also Ghoniem et al., 2005). Graph mining has proposed dedicated graph similarity functions (Gao et al., 2010). User studies have revealed factors for good readability of individual graphs (Archambault and Purchase, 2012; Dwyer et al., 2009; Kieffer et al., 2016; Purchase, 2002). The latter investigations have shown the need to study cognitive aspects for creating effective individual graph visualizations. However, there exist only a few guidelines for designing good comparative visualizations (Gleicher et al., 2011). Little is known about how much visual differences affect users' graph similarity notion.

Therefore, systematic investigations addressing this issue are necessary. These investigations have to clarify, on the one hand, how graphs have to be designed to support comparison processes. On the other hand, methodological issues have to be addressed. From a methodological point of view, there are still many open questions regarding the investigation of graph comparisons. To this end, we conducted an explorative study to identify influencing factors for the assessment of differences of graphs by human users. In the context of this study we realized that there are still many unresolved methodological problems. There are methodologies of data collection and analysis from Cognitive Psychology and HCI which can be used (e.g., thinking aloud, observation, etc.), but there are still issues specific to the area of graph comprehension which need to be addressed. These issues are especially related to the development of a dataset and a set of appropriate stimuli (visualizations) for

von Landesberger T., Pohl M., Wallner G., Distler M. and Ballweg K.
Investigating Graph Similarity Perception: A Preliminary Study and Methodological Challenges.
DOI: 10.5220/0006137202410250
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 241-250
ISBN: 978-989-758-228-8

the investigation to ensure that the investigations are valid and yield results which are useful for designers. Based on the research questions, the dataset, the graph layout and structure, and the variations of elements of the graphs have to be planned carefully to cover the relevant issues in this context. This is a non-trivial question which is rarely discussed in the visualization community. A clarification of these issues can also aid further research in other areas of information visualization. In Section 2 we will provide an overview of these problems.

To address these issues we conducted a study of perceived similarities between pairs of small labeled networks shown as juxtaposed node-link diagrams (see Section 4). We analyzed users' ratings and statements during visual graph comparison. To avoid confounding effects, we constrained the study to small static comparisons. We present lessons learned from this study (see Section 6). The presented considerations for methodological issues can serve as a basis for further studies, enabling comparability among future investigations.

2 CHALLENGES

1. *Definition of an appropriate dataset:* The development of an appropriate dataset is a challenging endeavor. In research on perceived graph similarity, researchers have to decide between realistic and synthetic datasets. Realistic datasets have the advantage of supporting ecological validity. On the other hand, the analysis of the perception of similarity of graphs has to be based on a systematic variation of elements of the graph. Realistic datasets often do not contain or allow for all these variations. Therefore, in many cases a synthetic dataset is better suited to support such research. Such a dataset has to be developed according to the research questions of the study. Another important issue is also the content (e.g., the node labels) of the dataset. Previous knowledge of participants of a study about the dataset might vary considerably and produce unintended effects.

2. *Development of appropriate visualizations (stimuli):* When studying the utility and usability of a visualization the design of the study material is of huge importance. Researchers have to explain their assumptions about the influence of the features of the visualizations on the behavior of the users. In the context of graphs and especially graph comparison, there are three main influencing factors:

- *Graph layout and structure:* The influence of the graph layout and structure is obvious. On the one hand, the graph layout should reflect the nature of the tasks potential users of these graphs would want to solve. On the other hand, the appearance of the graph should enable the researchers to clearly identify influencing elements. Researchers have to consider the trade-off between these two. Sometimes, a less realistic visualization is a better way to identify which factors influence the users' behavior.
- *Complexity of the graphs:* Graphs which are too complex might be difficult to investigate because not all influencing factors can be isolated. Graphs can have different substructures. Even in simple graphs, there are central and peripheral nodes. Changes in these nodes have different consequences on the perception of graph similarity. More complex graphs can contain subgroups which are clustered together. Adding meaningful labels to graphs increases their complexity and also influences the perception of differences in certain ways.
- *Size of the graphs:* The size of the graphs to be compared is highly relevant. Basic research in psychology indicates that simple graphs should be used to be able to easily identify influencing factors and avoid confounding effects. On the other hand, larger graphs are more realistic and help to understand how graph comparison works in real life. We decided to use small graphs for the initial investigation presented in this paper to avoid confounding variables. The research process in psychology, and more generally, in social sciences is based on the idea that clear causal relationships between independent and dependent variables can only be identified when all additional influencing factors can be determined by the researcher and kept constant (Bortz and Döring, 2007; Zimbardo and Gerrig, 2008). When investigating large graphs, a considerable number of additional influencing variables has to be taken into consideration, because of the amount of information available and also because occlusion makes perception of nodes, links, and labels difficult (Huang et al., 2006b; McGee and Dingliana, 2012). We think that it is necessary to start with very simple graphs to identify basic mechanisms of perception. When this is achieved it is possible to add additional complexity in the future.

3. *Systematic variation of elements of the visualization:* An important issue is the question which elements of a visualization should be considered for the investigation. It is often not possible to consider all kinds of features of a visualization in a study, therefore, researchers have to restrict their research to specific elements. In graphs, such elements might be the structure or layout of the graph, the design of edges,

edge crossings, usage of color, design of nodes, content (e.g., labels) of nodes etc. If a researcher, for example, investigates the influence of edge crossings, the number of edges between nodes, or the node labels on the perception of graph similarity, these elements have to be varied according to a systematic plan. If these features are varied in combination, the number of possible combinations will quickly become huge. Therefore, not all variations can usually be taken into account because participants of experiments will get tired after a fairly limited number of trials. Thus, researchers have to choose which variations to include in their experiment. This decision will depend on the research question.

4. *Interaction possibilities:* Interaction possibilities are also elements of a visualization which influence the users' behavior, but they have a specific character. Sometimes, it makes sense to provide only a few interaction possibilities to be able to identify the influence of these possibilities on the sensemaking activities of the users more clearly. If there are too many interaction possibilities in a user test, then it becomes difficult to analyze the individual influence of each of these possibilities because they might influence each other. Interaction possibilities which are especially relevant for graph comparison are, for example, highlighting specific areas of the graph, the possibility for users to draw on top of the graph to annotate certain areas, or moving parts of the graph around.

5. *Measurement of graph similarity:* Beside the issues discussed above, the question on how to measure the responses from study participants deserves specific mention as it is an often underestimated yet essential factor in study design. Quantitative methods may be more economical when dealing with large sample sizes but may miss contextual detail such as why humans perceive graphs as more or less similar. However, the *why* may be especially important in such kind of investigations. Qualitative methods may shed light on this issue but are more time-consuming and may thus limit the number of participants. Thus, a mixed-methods approach, merging quantitative measurements with qualitative methods (e.g., thinking aloud, video capture, or annotations) seems promising. In terms of quantitative measurements, the type of response scale (e.g., dichotomous, nominal, ordinal, continuous) should also be selected with care as it affects how nuanced the characterization of similarity will be. Lastly, it also remains largely unclear how many scales are appropriate for measuring graph similarity, that is, should a single scale or multiple scales focusing on different aspects of similarity (e.g., structure, content) be used.

6. *Task:* Lastly, the types of tasks users should perform need to be considered when designing and evaluating visualizations (for a general discussion see, e.g., Schulz et al. (2013) while Lee et al. (2006) provide an overview of common tasks related to graph data analysis). Consequently and ideally, the types of tasks should thus also be taken into account when studying the perception of graph similarities. For example, if users are performing exploratory tasks their notion of similarity may be influenced by other factors than when users engage in specific goal directed tasks such as, for instance, assessing differences regarding the number of adjacent nodes.

3 RELATED WORK

In the following we briefly review visual network comparison techniques and work concerned with perception and cognition factors in graph readability.

3.1 Graph Comparison Visualization

Recent papers survey network visualization and visual comparison (Beck et al., 2014; Gleicher et al., 2011; Hadlak et al., 2015; Vehlow et al., 2015; Von Landesberger et al., 2011). Visual network comparison techniques deal with two or with many networks (Bremm et al., 2011; Graham and Kennedy, 2010). Gleicher et al. (2011) present three main categories of visual comparison techniques, including juxtaposition which is easy to create and understand. Juxtaposed views are sometimes extended by linking the corresponding nodes between the graphs (Collins and Carpendale, 2007; Holten and Van Wijk, 2008). However, this can be misleading if links do not represent the inner graph structure well (Bremm et al., 2011). Highlighting of similar parts shows commonalities (Bach et al., 2014; Bremm et al., 2011; Holten and Van Wijk, 2008), and collapsing the identical parts emphasizes differences (Archambault, 2009). Both, however, need a graph matching or similarity function. While juxtaposition relies on user's memory, superposition makes use of perception for comparison but can, in turn, lead to clutter and does not scale well. Explicit encoding shows the computer-determined comparison directly, but encounters problems with decontextualization and encoding understandability. In contrast, juxtaposition gives the user full freedom to determine his/her comparison interpretation – which is essential for our study. Thus, we aim to clarify, which factors influence human's notion of graph similarity eventually leading to design guidelines.

3.2 Perception and Cognition Aspects in Graph Visualization

Several works focus on the *readability* of single graphs. Research has identified the following graph readability factors: graph design, graph aesthetics and layout, graph size, graph semantics as well as users' background. Huang et al. (2006a) advise how to design node-link diagrams. They suggest, for example, to highlight important nodes, to put the most important nodes on the top or in the center, or to cluster nodes which belong to the same group. Holten and van Wijk (2009) recommend designs for directed edges and Tennekes and de Jonge (2014) suggest node coloring in hierarchies. McGrath et al. (1997) state that layout can influence the interpretation of graphs. Graph aesthetics, such as edge crossing minimization, edge length homogeneity, edge angle specifics, or edge bends play a role for creating readable layouts (see Kobourov et al., 2014; Purchase, 2002; Purchase et al., 2007). Several studies analyzing human views on graph layouts showed that humans prefer force-directed-like layouts for small undirected graphs, (e.g., Dwyer et al., 2009; Kieffer et al., 2016; McGee and Dingliana, 2012).

It is generally accepted that larger networks are more difficult to process because of the amount of information available and also because occlusions make the perception of nodes, links, and labels more difficult (Huang et al., 2006b; McGee and Dingliana, 2012). Semantic aspects also influence graph perception (Purchase et al., 2001). Novick (2006) indicates that performance in the interpretation of node-link diagrams depends on the layout of the graph as well as on the content knowledge of the users. Körner (2005) also argues that graph comprehension is a process integrating visual and conceptual knowledge. The layout of the graph has to reflect the specific nature of the content, otherwise comprehension may fail.

Only a limited number of studies on perception and cognitive aspects in comparison of node-link diagrams exists. Some interesting insights can be gained from user studies dealing with dynamic graph visualization (e.g., Archambault et al., 2011). Dynamic graphs are often presented as several time slices one after the other (Bach et al., 2014; Beck et al., 2014) and users have to compare these time slices. Archambault and Purchase (2012) argue that memorability plays an important role. For example, Diehl and Görg (2002) proposed a layout for dynamic graphs preserving the users' mental model. Memorability also plays an important role for juxtaposed graph comparison, especially for larger graphs (Tominski et al., 2012).

4 STUDY DESIGN

The goal of this preliminary study is to identify factors which possibly influence the perceived similarity of small star-shaped networks (see Section 4.1). To streamline our study we focused on a subset of factors inspired both by literature and our experience with visual network comparisons. This initial subset is, of course, by no means complete but it covers a variety of frequently occurring changes in graphs.

$\mathbf{F}_1$ – The number of visually apparent edge changes.

$\mathbf{F}_2$ – Changes in edge crossings as edge crossings are a factor in graph aesthetics known to influence graph readability (Purchase, 2002).

$\mathbf{F}_3$ – Altering the label of the central node.

$\mathbf{F}_4$ – Exchanging labels between peripheral nodes while keeping the graph structure the same (cf. *GP1* in Table 1).

$\mathbf{F}_5$ – Familiarity with the theme of the graph as familiarity may influence graph interpretability (cf. Körner, 2005; Novick, 2006).

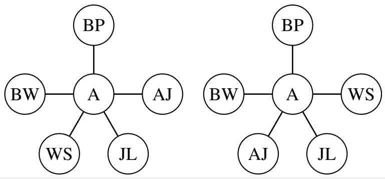
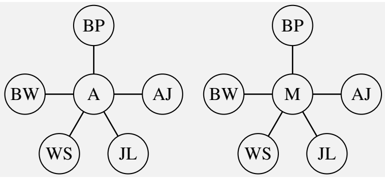
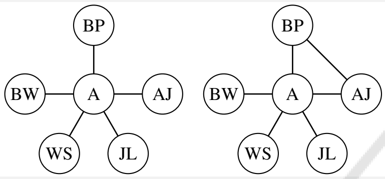
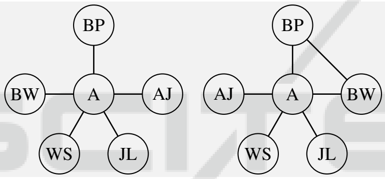
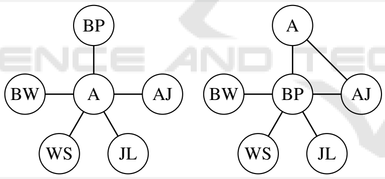
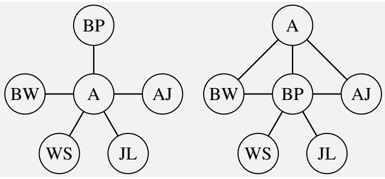
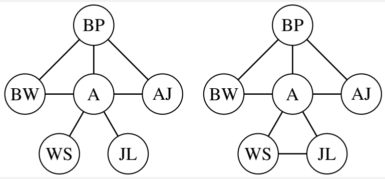
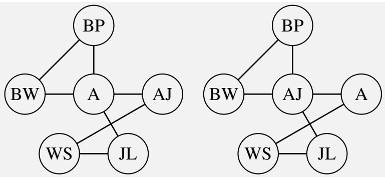
$\mathbf{F}_6$ – Professional background since people with graph theory knowledge may focus more on structural changes rather than on visual differences (McGrath and Blythe, 2004).

4.1 Data Set

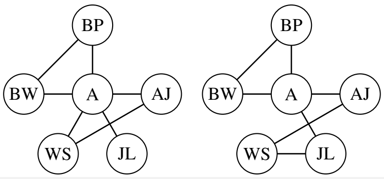
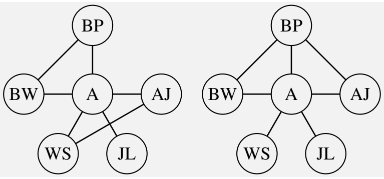
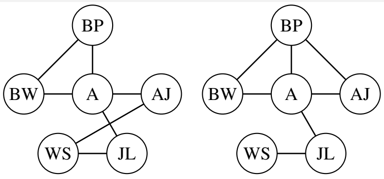
As pointed out above we focused on differences between pairs of juxtaposed labeled node-link diagrams for our first preliminary study. Because of the multitude of factors which may affect the perceived similarity of graphs we confined the experiment to small graphs. While this admittedly may limit the generalizability of our results to graphs of varying sizes this setting allowed for more control over confounding factors. Specifically, we used a star-shape-like graph structure with one central node and five nodes on the periphery which can also have edges between them. In these pairs we encoded various differences which – according to our hypotheses stated above – may influence the perceived similarity of two graphs. The dataset was composed of 11 graph pairs (GPs), each consisting of a reference graph and an altered graph (see Table 1).

Changes in Edge Structure. We included visual edge additions (see *GP3*, *GP4*, *GP7*) and edge deletions. Edge deletions were combined with additions (i.e., edge moves, see *GP9*, *GP10*, *GP11*). These changes occurred at various locations in the graph (compare, e.g., *GP3* and *GP7*).

Table 1: Overview of the experiment graphs (Hollywood theme) with their encoded change factors ($F1$ = number of edge changes, $F2$ = edge crossing change, $F3$ = label change altering the meaning, $F4$ = label change without changing the meaning). Node labels were shortened for illustration purposes. A = Actor, M = Musician, BP = Brad Pitt, AJ = Angelina Jolie, JL = Jennifer Lopez, WS = Will Smith, and BW = Bruce Willis.

	Graph Pair	F_1	F_2	F_3	F_4	Description
GP1	 BP BW A AJ WS JL BP BW A WS AJ JL	0	NO	NO	YES	Label switch between AJ and WS.
GP2	 BP BW A AJ WS JL BP BW M AJ WS JL	0	NO	YES	NO	Label of central node changed.
GP3	 BP BW A AJ WS JL BP BW A AJ WS JL	1	NO	NO	NO	Added edge between AJ and BP. Note, AJ and BP are a couple.
GP4	 BP BW A AJ WS JL BP AJ A BW WS JL	1	NO	NO	YES	Added edge between BP and BW, AJ and BW switched labels.
GP5	 BP BW A AJ WS JL A BW BP AJ WS JL	1	NO	YES	NO	A and BP switched labels and thus central node has changed; added edge between A and AJ.
GP6	 BP BW A AJ WS JL A BW BP AJ WS JL	2	NO	YES	NO	Same changes as in GP5, but with a second added edge (between A and BW).
GP7	 BP BW A AJ WS JL BP BW A AJ WS JL	1	NO	NO	NO	Added edge between WS and JL.
GP8	 BP BW A AJ WS JL BP BW AJ A WS JL	0	NO	YES	NO	A and AJ switched labels thus altering the central node.

... continued on next page

Graph Pair		F_1	F_2	F_3	F_4	Description
GP9	 BP BW A AJ WS JL BP BW A AJ WS JL	2	NO	NO	NO	Removed edge between A and WS; added edge between WS and JL.
GP10	 BP BW A AJ WS JL BP BW A AJ WS JL	2	YES	NO	NO	Added edge between AJ and BP; removed edge between AJ and WS. This also removes the edge crossing.
GP11	 BP BW A AJ WS JL BP BW A AJ WS JL	2	YES	NO	NO	Same changes as in GP10 (add and remove edge – removing edge crossing) but the reference graph is different.

Changes in Edge Crossings. Edge crossings were present in four reference graphs (*GP8* – *GP11*) of which two pairs also encoded a change in edge crossing (*GP10* and *GP11*).

Change of Center Node. The central node of the star-like graph has a label which attributes meaning to the graph. Peripheral nodes support this meaning. Consequently, this allowed us to test if changing the label of the central node has semantic implications.

Label Exchange between Peripheral Nodes. Switching labels on the periphery does not influence the graph structure but can be considered as a kind of node reordering. This allowed us to test how much influence the order of nodes has on the graph similarity.

Graph pairs *GP4*, *GP5*, and *GP6* encoded multiple of the above described differences. Each graph pair existed twice, once labeled with names of Hollywood actors and once with psychology personalities, resulting in a total of 22 graph pairs (2×11). This allowed us to assess the factor of familiarity (F_5).

4.2 Procedure

The study was conducted in a controlled environment (room, table, etc.) on a laptop with touchscreen and a touch-enabled pen. At the beginning the participants were informed about the goal and the procedure of the study. Next, they were instructed on how to use the tool after which a training phase with a set of three to five graph pairs followed. Participants were asked to consider content aspects as well as structural aspects. Once the participants felt confident with the tool the main study – consisting of two parts in random order – was conducted: comparing graph pairs with Hollywood actors and with the psychology theme. For that purpose the reference graph and the altered graph were shown side-by-side. Within each part all studied graph pairs were shown in random order. For each pair participants had to rate its similarity on an 8-point scale (0 = very different to 7 = identical). During the study thinking aloud protocols were recorded which were then transcribed and analyzed for reoccurring themes. Participants could also make drawings using the user interface. The study lasted between 30 and 60 minutes.

4.3 Participants

We recruited 24 volunteers – master students and research assistants – from the [anonymous] university. 12 subjects had a major in computer science and 12 subjects had a background in psychology (F_6). Subjects were between 18 and 35 years of age with equal proportion of females and males. Psychology subjects, in contrast to computer science subjects, had no background in graph theory or in network visualization. All participants but one stated to be familiar with Hollywood actors and only two participants stated to be familiar with psychological personalities. Thus, we assumed that participants were familiar with the Hollywood topic and the opposite was the case for the psychology topic (F_5).

5 RESULTS

Table 2 lists the distribution of the ratings for the individual graphs of the Hollywood set. Ratings of the psychology graph pairs did not differ substantially from the Hollywood graph pairs (as also confirmed by statistical analysis which showed no influence of the graph topic on the ratings, see below) and are thus not listed due to space restrictions. Informal inspection of these distributions revealed three patterns which will be described in detail in the following. We will also amend the discussion with insights gained from the analysis of the thinking aloud protocols (see Table 3).

+ Dominant similarity ratings correspond to graph pairs with small structural differences which do not alter the meaning of the graph (*GP1*, *GP3*, *GP4*, and *GP7*). Analysis of the think aloud protocols revealed that participants realized that structure and content of *GP1* did not change despite switching the labels of two nodes. Graph pairs *GP3*, *GP4*, and *GP7* have structural differences in the sense that one edge has been added. Contrary to *GP1* they have no label changes. As expected, the participants focused on edge related aspects (19, 16, and 20 statements, respectively). An interesting observation could be made for *GP3* where participants specifically stressed the newly added connection between *Angelina Jolie* and *Brad Pitt* which is also meaningful in a real-life context where both are in a relationship with each other.

− Graph pairs where the label of the central node (i.e., context node) changed (*GP2*, *GP5*, *GP6*, and *GP8*) received predominant dissimilar ratings with the majority of participants noting a change in context. More specifically, with respect to *GP2*, 22 statements of the thinking aloud protocols were concerned with a context change (actor → musician) and 16 statements were related to different content. The statements regarding *GP5*, *GP6*, and *GP8* were very similar. In all three instances, the respondents mentioned that the location of the context node changed (that is, it has been moved to the periphery) (19x, 14x, 21x), highlighted the change in the focus of the relationships (13x, 14x, 11x), and noted the deleted edge which formerly connected the persons to their profession (10x, 11x, 7x).

± Bipolar similar/dissimilar ratings correspond to graph pairs with both structural and content changes (*GP9*, *GP10*, and *GP11*). Interestingly, all these graph pairs together with *GP8* (also having a partly bipolar rating) have a reference graph with an edge crossing. The bipolar rating distribution is also reflected in the participants' statements. About half of the respondents mentioned edge changes (13x, 12x, 15x). Others, in turn, primarily focused on content aspects, noting, for example, the formation of groups (5x), the unclear meaning of edges (5x), or the focus on relationships (11x). It seems that several different factors influenced the participants' ratings.

Table 2: Rating distribution for the Hollywood graph pairs and rating modes(s). Ratings: *very different* (0) to *identical* (7). Background color reflects the number of ratings from white (0) to blue (max), D = distribution type, M = median.

GP	**D**	**Ratings**								**M**
		0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	
GP1	+	0	0	0	1	1	3	8	11	6
GP2	-	3	5	8	1	1	3	2	1	2
GP3	+	0	0	4	2	2	7	8	1	5
GP4	+	2	0	4	3	2	9	3	0	5
GP5	-	5	5	8	2	0	2	2	0	2
GP6	-	7	4	7	1	1	4	0	0	2
GP7	+	0	0	1	3	1	9	8	2	5
GP8	±	2	6	7	2	0	6	1	0	2
GP9	±	0	1	8	5	0	6	2	2	3
GP10	±	3	1	7	0	3	7	2	1	4
GP11	±	0	2	7	2	4	7	2	0	4

Next, we tested the influence of the above described factors ($F1$-$F6$) on the perceived graph similarity rating using a regression. Due to the ordinal response scale and the within-subject design we used general estimating equations (GEE) with an ordinal logistic regression model and the factors $F1 - F6$ as predictors (for an introduction to GEE models see, e.g., Hardin and Hilbe, 2012). Missing responses (3 out of 24 × 22) were excluded from the analysis. All factors relating to changes were compared by selecting the *no change* condition as reference category. GEE model estimates are shown in Table 4, significant factors are written in italics. To complement the evaluation, we also assessed differences in the ratings between specific graph pairs in order to develop a better understanding of the significant predictors. For that purpose we used pairwise Wilcoxon signed-rank tests with a significance level of .05. For reasons of comparability, however, we will restrict pair-wise comparisons to graph pairs which only have one difference in graph changes (e.g., one vs. two added edges) or have the same changes in the reference graphs. In the following, we discuss the GEE results in more detail.

F_1 - Number of Visually Apparent Edge Changes: The number of added and removed edges showed to be a significant factor lowering the similarity rating both for one edge ($B = -1.510$, $OR = 0.22$) and two edge differences ($B = -1.897$, $OR = 0.15$). This result was expected, as it seems natural that larger changes lead to lower similarity. This was

Table 3: Three most frequent statements for each graph pair.

#	Most frequent statements		
GP1	▬ **20** arrangement different	▬ **12** arrangement not important	▬ **7** content identical
GP2	▬ **22** context changed	▬ **16** content different	▬ **7** graph does not make sense
GP3	▬ **19** extra edge	▬ **15** relationship BP & AJ	▬ **6** graph does not make sense
GP4	▬ **16** extra edge	▬ **10** arrangement different	▬ **10** relationship BP & BW
GP5	▬ **19** location of category changed	▬ **14** focus is on relationships	▬ **10** deleted edge to profession
GP6	▬ **14** location of category changed	▬ **13** focus is on relationships	▬ **11** deleted edge to profession
GP7	▬ **12** extra edge	▬ **8** edge changed	▬ **6** meaning of edges unclear
GP8	▬ **21** location of category changed	▬ **11** focus is on relationships	▬ **7** deleted edge to profession
GP9	▬ **13** edge changed	▬ **13** deleted edge to profession	▬ **5** content identical
GP10	▬ **12** edge changed	▬ **6** edge crossing	▬ **5** formation of groups
GP11	▬ **15** edge changed	▬ **7** edge crossing	▬ **5** meaning of edges unclear

also confirmed by a pairwise Wilcoxon signed-rank test comparing *GP*5 with *GP*6 which differ only in one edge (changing one vs. two edges, $Z = -3.471, p = .001$).

F$_2$ - Changes in Edge Crossings lower the rating significantly ($B = -0.771$, $OR = 0.46$). The significance of this factor could be confirmed by comparing ratings of *GP*9 and *GP*10, both having two edge changes, but only one graph pair has an edge crossing change ($Z = -3.801$, $p < .001$).

F$_3$ - Change of Center Node is the factor which lowers the rating by the largest extent ($B = -3.221$, $OR = 0.04$). This is also supported by a Wilcoxon pairwise test of *GP*1 and *GP*2 which both have a label change – once on the periphery and once in the center ($Z = -4.120$, $p < .001$). The central node change seems to dominate over the size of visual edge changes. As an example, graph pairs *GP*5 and *GP*6 have the same central node change. Both graph pairs have also a visual edge change, but of a different size (one or two). Interestingly, the subjects rated both graph pairs similarly ($Z = -0.403$, $p = .687$).

F$_4$ - Label Switch Leading to Visual Rearrangement does not have a significant impact on the rating ($B = -0.121$, $OR = 0.89$). Analysis of the think aloud protocols showed that most users recognized the irrelevance of these label changes on the graph structure. Thus, they did not consider them in their ratings.

F$_5$ - Familiarity with the Graph Topic does not influence the rating significantly in our case ($B = 0.102$, $OR = 1.11$). Thinking aloud indicates that the participants paid attention to the same visual and structural factors in both cases. The users could infer meaning changes even for the psychology graphs despite their unfamiliarity. This result should, however, be interpreted with caution as not all graphs may be easily interpreted when users are unfamiliar with the topic. In our case, both conditions included names of personalities and consisted of a very small number of nodes. This could have eased their interpretation.

F$_6$ - Users' Professional Background was not a significant factor ($B = 0.092$, $OR = 1.1$). This suggests that the readability of changes in small labeled graphs with a star-shaped structure is irrespective from graph theory knowledge and visualization background.

6 DISCUSSION – LESSONS LEARNED

Apart from the results gained by the study described in this paper we could also identify methodological challenges for the investigation of the perception of graph similarity.

In general, it may seem preferable to use realistic datasets for the investigation of visualizations. For the investigation of the perception of similarity of graphs, however, in most cases only synthetic datasets are advisable because a significant subset of variations of features has to be analyzed. These variations are usually not available in a realistic dataset and have to be constructed according to the research questions. In addition, the previous knowledge of the participants of the study, which has some relationship to the content of a dataset, has also to be taken into account. In our investigation, this knowledge did not have much influence on the results. Nevertheless, there is some

Table 4: GEE predicting the effect of the tested factors on the perceived graph similarity rating. Odds ratios (OR) are calculated by exponentiation of the B coefficient. Reference categories for the individual predictors are given in brackets and significant predictors are highlighted in italics.

Predictor (reference)	**B**	**OR**	**95% CI**
F_1 – *Nr. of edge changes (0)*			
1 edge change	-1.510	0.22**	-2.08 to -0.94
2 edge changes	-1.897	0.15**	-2.59 to -1.20
F_2 – *Edge crossing change (no)*	-0.771	0.46*	-1.41 to -0.13
F_3 – *Label Change with meaning (no)*	-3.221	0.04**	-3.90 to -2.54
F_4 – Label change without meaning (no)	-0.121	0.89	-0.56 to 0.32
F_5 – Graph topic (Hollywood)	0.102	1.11	-0.18 to 0.38
F_6 – Profession (Computer Science)	0.092	1.10	-0.64 to 0.83

* $p < .05$; ** $p < .001$

indication that in other studies this influence was significant (e.g., Shah and Hoeffner, 2002). Further research has to identify the reasons for the presence or absence of this influence.

The different features of the graphs have to be investigated systematically. There is a large amount of features which can influence the perception of similarities of graphs. In our investigation we only used the number of edges, edge crossings, and content (specifically, node labels) as features influencing the behavior of the participants. Our first results suggest that content plays an important role and changing the label of the central node is more relevant than changing the label of peripheral nodes. Changes in the number of edges and edge crossings also have some influence. It depends on the preference of the participants which aspects are more important to them. If both types of changes (content and structure) are present, a bipolar distribution of the similarity ratings can occur, indicating that there are two distinct groups of participants, one taking content more into account, and the other for whom structure is more important. However, there are still many open issues in this context (e.g., are semantic changes distinguished from visual-only changes, that is, changes where the data stays the same but the layout changes), and many other features of graphs have to be investigated. The systematic variation is essential for the investigation, especially the combination of various features. In our case it was combining number of edges, number of edge crossings, and labels of graphs. Even developing an investigation plan for the variation of such a limited amount of features was challenging. We had especially to take into account that the cognitive load on the participants was manageable. Experience shows that participants are only able to rate a limited number of graph pairs in one experiment. Therefore, we think that a number of investigations following each other has to be conducted to study a relevant part of all possible variations.

We did not provide the participants with interaction possibilities to avoid confounding effects. It can be argued that, for example, the number of edges or the change of node labels can be detected more easily when the possibility of highlighting is provided. Therefore, the influence which edges or edge crossings play have to be studied in isolation without taking the confounding factors of interaction into account. This also indicates that a number of consecutive studies with an increasing amount of complexity and interactivity should be conducted.

7 CONCLUSIONS

We think that the challenge of developing an appropriate dataset and appropriate stimuli (i.e., variations of visualizations) for evaluation purposes is sometimes underestimated. For systematic basic research in this area which can uncover the influence of individual features of a visualization, this activity is nevertheless necessary. In this paper, we try to demonstrate issues related to this question using the example of graph comparison. We also provided a first overview of contributing factors which are relevant in this area. This is by no means a comprehensive list of such factors. We intend to investigate this issue in more detail in future work. Based on this work, we want to conduct several studies with graphs of increasing complexity and interactivity to identify influencing factors for graph comparison. We already presented first tentative results of such a study. The goal of this work is to develop guiding principles for the design of graphs which can support graph comparison efficiently.

REFERENCES

Archambault, D. (2009). Structural differences between two graphs through hierarchies. In *GI*, pages 87–94.

Archambault, D. and Purchase, H. (2012). The mental map and memorability in dynamic graphs. In *IEEE PacificVis*, pages 89–96.

Archambault, D., Purchase, H. C., and Pinaud, B. (2011). Animation, small multiples, and the effect of mental map preservation in dynamic graphs. *IEEE TVCG*, 17(4):539–552.

Bach, B., Pietriga, E., and Fekete, J.-D. (2014). Graphdiaries: animated transitions and temporal navigation for dynamic networks. *IEEE TVCG*, 20(5):740–754.

Beck, F., Burch, M., Diehl, S., and Weiskopf, D. (2014). The state of the art in visualizing dynamic graphs. *EuroVis STAR*.

Bortz, J. and Döring, N. (2007). *Forschungsmethoden und Evaluation für Human-und Sozialwissenschaftler*. Springer.

Bremm, S., Von Landesberger, T., Heß, M., Schreck, T., Weil, P., and Hamacher, K. (2011). Interactive visual comparison of multiple trees. In *IEEE VAST*, pages 31–40. IEEE.

Collins, C. M. and Carpendale, S. (2007). VisLink: Revealing relationships amongst visualizations. *IEEE TVCG*, 13(6):1192–1199.

Diehl, S. and Görg, C. (2002). Graphs, they are changing. In *Graph drawing*, pages 23–31. Springer.

Dwyer, T., Lee, B., Fisher, D., Quinn, K. I., Isenberg, P., Robertson, G., and North, C. (2009). A comparison of user-generated and automatic graph layouts. *IEEE TVCG*, 15(6):961–968.

Gao, X., Xiao, B., Tao, D., and Li, X. (2010). A survey of graph edit distance. *Pattern Analysis and applications*, 13(1):113–129.

Ghoniem, M., Fekete, J.-D., and Castagliola, P. (2005). On the readability of graphs using node-link and matrix-based representations: A controlled experiment and statistical analysis. *Inf Vis*, 4(2):114–135.

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

Graham, M. and Kennedy, J. (2010). A survey of multiple tree visualisation. *Inf Vis*, 9(4):235–252.

Hadlak, S., Schumann, H., and Schulz, H.-J. (2015). A survey of multi-faceted graph visualization. *EuroVis STAR*.

Hardin, J. W. and Hilbe, J. M. (2012). *Generalized Estimating Equations*. Chapman and Hall/CRC, 2 edition.

Holten, D. and Van Wijk, J. J. (2008). Visual comparison of hierarchically organized data. *CGF*, 27(3):759–766.

Holten, D. and van Wijk, J. J. (2009). A user study on visualizing directed edges in graphs. In *CHI '09*, CHI '09, pages 2299–2308.

Huang, W., Hong, S.-H., and Eades, P. (2006a). How people read sociograms: a questionnaire study. In *IEEE PacificVis*, pages 199–206.

Huang, W., Hong, S.-H., and Eades, P. (2006b). Predicting graph reading performance: a cognitive approach. In *IEEE PacificVis*, pages 207–216.

Kieffer, S., Dwyer, T., Marriott, K., and Wybrow, M. (2016). Hola: Human-like orthogonal network layout. *IEEE TVCG*, 22(1):349–358.

Kobourov, S. G., Pupyrev, S., and Saket, B. (2014). Are crossings important for drawing large graphs? In *Graph Drawing*, pages 234–245. Springer.

Körner, C. (2005). Concepts and misconceptions in comprehension of hierarchical graphs. *Learning and Instruction*, 15(4):281–296.

Lee, B., Plaisant, C., Parr, C. S., Fekete, J.-D., and Henry, N. (2006). Task taxonomy for graph visualization. In *BELIV*, pages 1–5, New York, NY, USA. ACM.

McGee, F. and Dingliana, J. (2012). An empirical study on the impact of edge bundling on user comprehension of graphs. In *AVI*, pages 620–627.

McGrath, C. and Blythe, J. (2004). Do you see what I want you to see? the effects of motion and spatial layout on viewers' perceptions of graph structure. *J. Soc. Structure*, 5(2):2.

McGrath, C., Blythe, J., and Krackhardt, D. (1997). The effect of spatial arrangement on judgments and errors in interpreting graphs. *Social Networks*, 19(3):223–242.

Novick, L. R. (2006). The importance of both diagrammatic conventions and domain-specific knowledge for diagram literacy in science: The hierarchy as an illustrative case. In *Diagrammatic representation and inference*, pages 1–11. Springer.

Purchase, H. C. (2002). Metrics for graph drawing aesthetics. *J. Vis. Languages & Computing*, 13(5):501–516.

Purchase, H. C., Hoggan, E., and Görg, C. (2007). How important is the mental map: an empirical investigation of a dynamic graph layout algorithm. In *Graph drawing*, pages 184–195. Springer.

Purchase, H. C., McGill, M., Colpoys, L., and Carrington, D. (2001). Graph drawing aesthetics and the comprehension of uml class diagrams: an empirical study. In *IEEE PacificVis*, volume 9, pages 129–137.

Schulz, H.-J., Nocke, T., Heitzler, M., and Schumann, H. (2013). A design space of visualization tasks. *IEEE TVCG*, 19(12):2366–2375.

Shah, P. and Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, 14(1):47–69.

Tennekes, M. and de Jonge, E. (2014). Tree colors: color schemes for tree-structured data. *IEEE TVCG*, 20(12):2072–2081.

Tominski, C., Forsell, C., and Johansson, J. (2012). Interaction Support for Visual Comparison Inspired by Natural Behavior. *IEEE TVCG*, 18(12):2719–2728.

Vehlow, C., Beck, F., and Weiskopf, D. (2015). The state of the art in visualizing group structures in graphs. In *EuroVis STAR*.

von Landesberger, T., Gorner, M., and Schreck, T. (2009). Visual analysis of graphs with multiple connected components. In *IEEE VAST*, pages 155–162.

Von Landesberger, T., Kuijper, A., Schreck, T., Kohlhammer, J., van Wijk, J. J., Fekete, J.-D., and Fellner, D. W. (2011). Visual analysis of large graphs: state-of-the-art and future research challenges. In *CGF*, volume 30, pages 1719–1749. Wiley.

Zimbardo, P. G. and Gerrig, R. J. (2008). *Psychologie*. Pearson Studium.