Computational Aesthetics in Graphics, Visualization, and Imaging (2009)
O. Deussen and P. Hall (Editors)

Comparing the Readability of Graph Layouts using Eyetracking and Task-oriented Analysis

Mathias Pohl, Markus Schmitt, and Stephan Diehl

[1] Computer Science, University of Trier, Germany

Abstract

In this paper we present the results of a user study comparing the readability of force-directed, orthogonal, and hierarchical graph layouts. To this end we identified prototypical tasks which are solved using visual representations of graphs. Based on the correctness of answers and the related response time we evaluated for each task which layout is better suited. In addition, we found possible explanations for these results by analyzing the eye-tracking data. Finally, we discuss some implications of our findings for algorithm designers and application developers.

Categories and Subject Descriptors (according to ACM CCS): I.3.6 [Computer Graphics]: Methodology and Techniques—Ergonomics

1. Introduction

Over the years many different graph drawing methods and variants thereof have been developed. Usually, these approaches are designed to produce layouts that are optimal with respect to certain aesthetic criteria. As a result, many algorithms focus primarily on minimizing the number of edge crossings, the number of edge bends, or the size of the resulting drawing. According to previously conducted user studies these criteria certainly have a strong influence on the readability of node-link visualizations and their understanding. However, the relation between the visualization's usefulness and the used layout style has not been examined so far.

To answer this question, we evaluated the effect of three different layout styles on the readability of graphs – a force-directed, an hierarchical, and an orthogonal layout style. We conducted a study that is supposed to give answers to two general issues. The first was to have an idea about what properties should be optimized. Primarily, this addresses the domain of algorithm designers. The second question was which algorithm to choose when designing an application.

To come to a result we conducted a user study with 36 subjects. Several drawings of graphs were shown to the participants who were asked to solve five different prototypical tasks for each drawing. We not only evaluated answering time and correctness statistically but also analyzed the subjects' answer strategies using an eye-tracking system. That way it is possible to give explanations why a certain layout is better suited for a specific task than another.

2. Related Work

For more than a decade, Purchase has performed empirical studies related to the aesthetics of graph layouts. Previous studies conducted by her [PMCC01] revealed that graph layout aesthetics can have a significant impact on the usability of drawings. However, she also found out that useful layouts for certain application domains obey different aesthetic criteria [PC02]. Whereas in earlier work she found that reducing the number of edge crossings was the most important aesthetic consideration [PCJ96, Pur97], in recent work continuity turned out to be an important factor as well [WPCM02]. Here, continuity means the sum of the angular deviations of the incoming and outgoing edges for each node on a path.

Bennett et al. gave a comprehensive summary on which aesthetic heuristic has which effect on readability and understanding of graph drawings [BRSG06]. However, in the same paper they also stated that the perceptual basis of these heuristics is not fully understood.

To the best of our knowledge in this paper we present the most extensive eye-tracking study of the readability of graph layouts. While an earlier study of Huang and Eades [HE05] was performed with 13 participants, and the subsequent

DOI: 10.2312/COMPAESTH/COMPAESTH09/049-056

study by Huang [Hua07] with 16 participants, we had 5 participants in the pre-study and 36 in the main study. In the studies by Huang and Eades subjects had to find shortest paths and most connected nodes. While Huang and Eades investigated the effect of the number and angles of crossings and the distance of the shortest path from the geometric path, we are evaluating the effect of the layout method on the readability of the graph.

3. Experimental Setup

For the study we identified five prototypical tasks that are discussed below. We decided to use random graphs with 10, 15, and 20 nodes with an average degree of 3.

These graphs then were layouted using three different layout algorithms:

- The force-directed approach according to Fruchterman and Reingold [FR91].
- The layer-based approach by Sugiyama et al. [STT81] also often referred as hierarchical layout.
- The orthogonal layout by Fößmeier and Kaufmann [FK95].

To obtain the nine final drawings the DGD-system [PRB08], a system primarily designed for dynamic graph drawing, was used. The final drawings are shown in Figure 1.

The actual experiment was performed with an eye tracking system (Tobii x50) that uses corneal reflection of infrared light to locate the position and movement of the eye. The questions and graph drawings were shown on a computer screen (1280x1024) and two cameras mounted on the screen recorded the eye movements at a frequency of 50Hz, i.e. an image is taken every 20ms. Prior to each task a small cross in the center of the screen was shown. That way all subjects started at the same position on the screen and hence, the obtained results are more comparable.

For the analysis of the recorded eye-tracking data we used heatmap visualizations. To create the heatmaps the points of fixation are aggregated over all subjects and over time. The higher the aggregated fixation count of a pixel the more red is the color of that pixel in the heatmap. The resulting heatmap is visualized on top of the original drawing of the graph.

In a pre-study we tested the experimental setup with 5 subjects to avoid erroneous results because of the multitude of parameters that we had to take into account. As a consequence of the pre-study we replaced the fixed order of the questions by a random order of question blocks to avoid a learning effect in the main study. We also relabeled the nodes in each graph and used only letters from a set of phonetically discriminable letters to reduce possible misunderstandings when recording the subjects' answers to questions. The participants (22 female, 14 male) in the main study were students from various fields including law, geography, computer science and psychology with an average age of 23.5 years with the youngest 20 and the oldest 29 years of age.

4. Task 1

As a warm-up question we asked the subjects to answer with "yes" or "no", if the displayed graph contained a node with a given label.

Results: As expected all subjects could answer this question correctly. It took them on average less than 3 seconds to decide this question.

Eye-tracking: The information from the eye-tracking system reveals an interesting result. The subject's strategies to detect the node is independent from the layout. We visualized the search strategy by separating the fixations into consecutive heatmaps – one for each second of the search time. Figure 3 shows the consecutive heatmaps for the orthogonal layout. The target node in this example was "Z". In the heatmap representing the first second of the search time the main focal point is in the center because the subjects were asked to focus the center of the screen before the graph is displayed. The consecutive heatmap shows that some subjects used a spiral search pattern starting in the center and already found the target node whereas other participants move their focal point to the left upper corner of the screen and started their search from there. The next two heatmaps show that the subjects then moved their focal points downwards and to the right and finally to the target node. Nodes below this node have received none or only little attention. Note that due to the small cross that is displayed before the actual graph all subjects start in the center of the screen. Hence, this is not an effect of the drawing or the task itself but an effect of the experimental setup.

5. Task 2

The second task was to identify whether there is a path between two given nodes. The participants were asked to name the labels of the nodes along the path they found. We did not ask explicitly for the shortest path.

Results: Here, for the hierarchical layout only 58% of the answers were answered correctly, whereas the force-directed and the orthogonal layout had 93% and 91% correct answers respectively. The poor performance of the hierarchical layout compared to the force-directed and the orthogonal layout was statistically significant for graphs of size 15 and 20 as well as when combining the results for all three graph sizes.

Furthermore, in the case of the force-directed layout the path found was in 68% of the cases the shortest path – compared to 2% for the orthogonal and 40% for the hierarchical layout.

Eye-tracking: The heatmaps for the force-directed and the orthogonal layouts show that there were almost only fixations on the nodes, whereas the heatmaps for the hierarchical layout show many fixations on edge crossings indicating

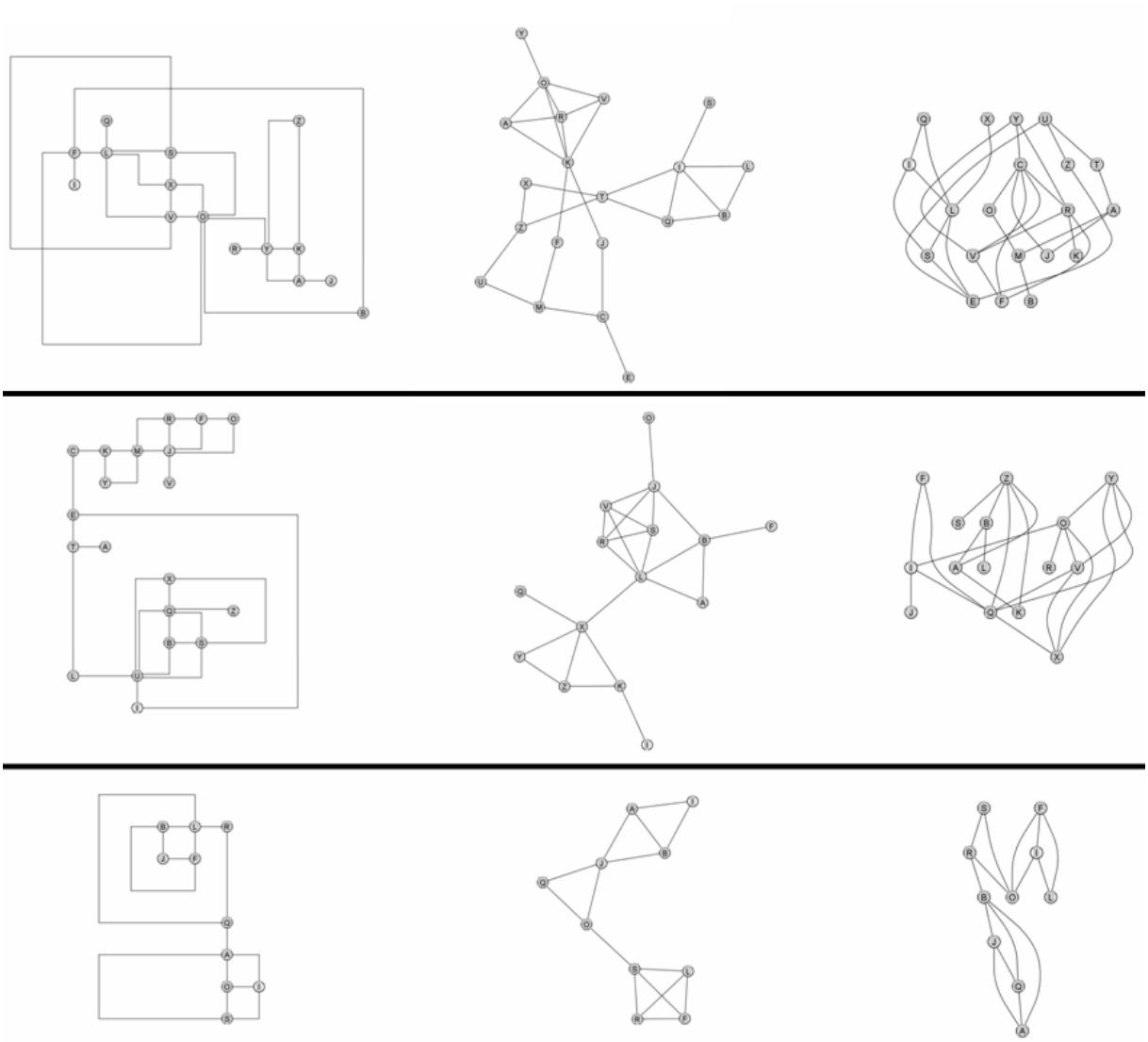

Figure 1: *The orthogonal, force-directed and hierarchical layouts of three graphs used for the study.*

that tracking edges in the hierarchical layout took more effort. Figure 4 displays the resulting heatmaps for the graph with 20 nodes.

The heatmap also contains an explanation for why many subjects did not find the shortest path in orthogonal layout. Long edges seem to be ignored during the search. This is less important in force-directed layout since nodes are always connected directly by a straight line but it is crucial for orthogonal layout.

6. Task 3

This task aimed at the problem to find specific patterns in a graph, i.e. isomorphic subgraphs. All subjects were asked to say whether the graph shown contains a given graph as a subgraph and to mention all labels belonging to that subgraph. Since an abstract description of the requested pattern appeared too complex all participants received a visual description of the requested subgraph.

Results: Most correct answers (more than 81%) were given for the force-directed layout. The orthogonal (52,8%) and the hierarchical (58,3%) layout could not reach this degree of precision. Furthermore, the average time spent for correctly

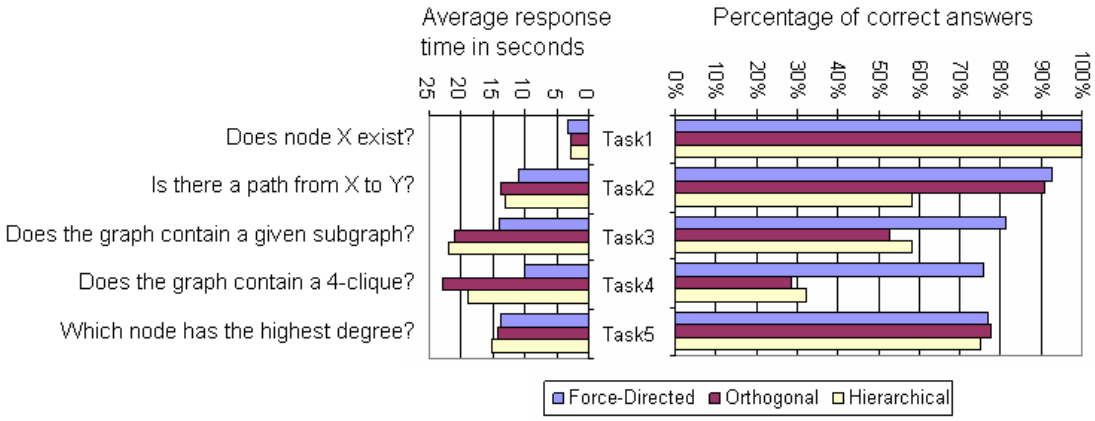

Figure 2: *Correct answers and average response time.*

finding the subgraph was the lowest for the force-directed layout (14s) compared to the orthogonal (22s) and the hierarchical (22s) layouts. It seems that finding this specific pattern is much easier in the force-directed layout. Both results, that the force-directed layout outperformed the other layouts with respect to correctness and response time, were statistically significant for graphs of size 15 and 20 as well as when combining the results for of all three graph sizes.

Eye-tracking: Figure 5 shows the heatmap for solving Task 3 using an orthogonal layout. We can see that there are three red areas in the heatmap, each around a number of nodes that are placed close to each other. These clusters contain up to nine nodes. The subjects seem to first identify these and then inspect the connection among the nodes of each cluster. For the force-directed layout we saw the least number of clusters that the subjects focused on.

Although the subgraph was not given in an abstract description but in a graphical representation the participants obviously do not try to find a matching shape. Instead they were really looking for a set of nodes with the matching connectivity.

7. Task 4

Similar to Task 3 all participants had to find a 4-clique in all of the drawings. However, in contrast to the previous task only an abstract description of the requested pattern was given.

Results: The results of this task are similar to those of Task 3. Again, the force-directed layout produced most correct answers (75,9%) compared to the hierarchical (28,7%) and the orthogonal (30,6%) layout. The average response time of correct answers was the shortest for the force-directed approach (10s), while for the hierarchical (19s) and orthogonal (23s) layouts the participants took much longer to answer correctly.

Both findings with respect to correctness and response time were statistically significant for graphs of size 10, 15, 20, as well as when taking the results for all graphs into account.

Eye-tracking: Here we observed basically the same search behavior as in Task 3. Subjects identified clusters and then inspected the connection among the nodes of these clusters. Again, for the force-directed layout we saw the least number of clusters that the subjects focused on. Since this task was given in an abstract description instead of a visual one this result also shows that the success of force-directed layout in Task 3 was independent from the task description. The subjects used a similar search strategy than that in the previous task.

8. Task 5

The final task of the study addressed the degree of nodes. The subjects had to find the node with the highest degree.

Results: The results for this question do not show any significant difference between the three layouts. To answer the question correctly each person took about 15 seconds on average. Furthermore the correctness of the answers was between 75,0% (hierarchical) and 77,8% (orthogonal). According to this results, it seems that inspecting nodes is not affected by the used method for node placement and edge routing.

Eye-tracking: As depicted in Figure 6 the subjects only focus three to four nodes. Only for these they count the number of outgoing edges to find the node with the highest degree. This observation emphasizes the result that inspection of nodes is independent from the used layout method.

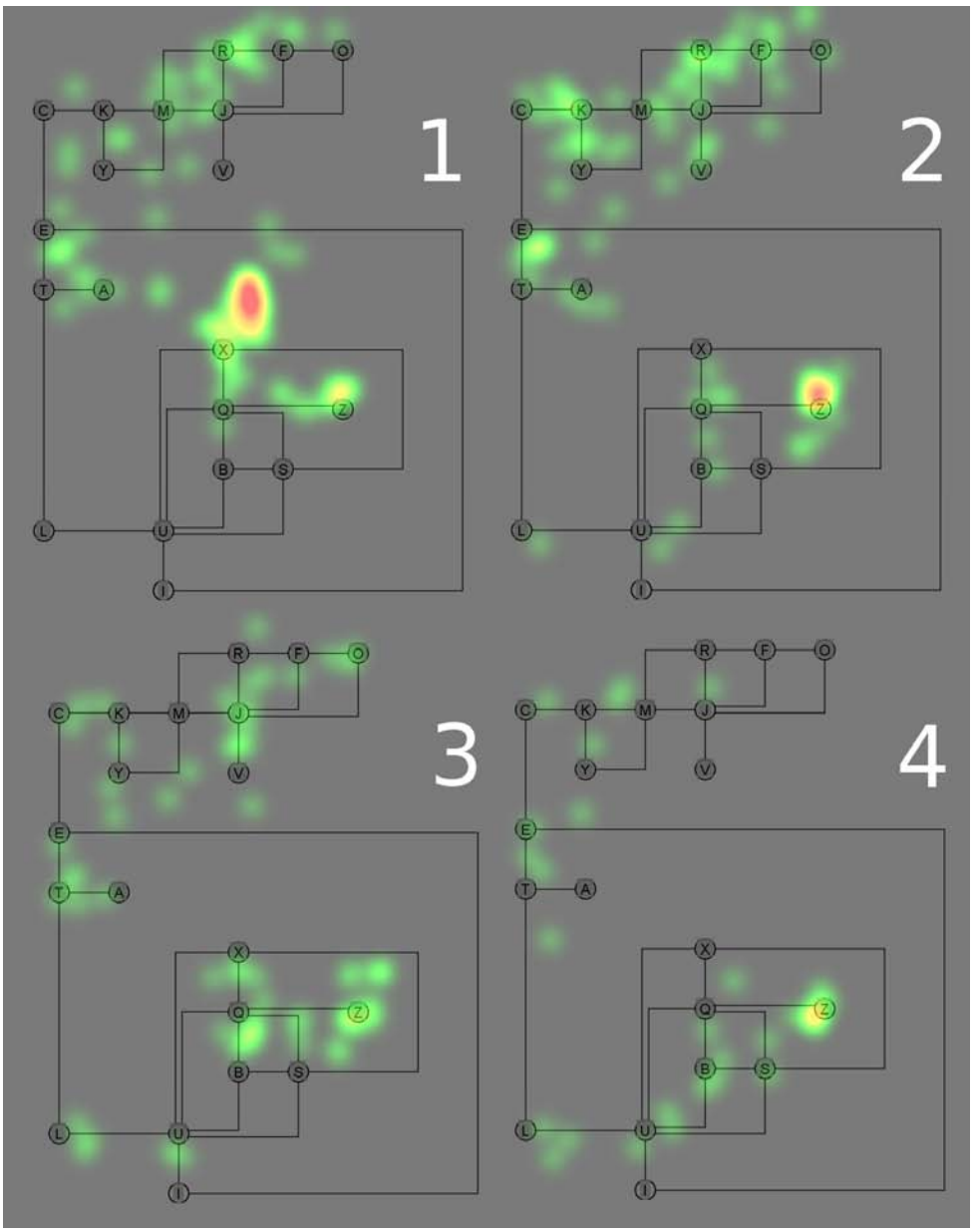

Figure 3: *Task 1: The four heatmaps for the first four seconds of search time (orthogonal layout).*

9. Statistics

The statistical results are presented in Tables 1 and 2. For the evaluation we used T-Tests. Here, $p(f,o)$ is the error probability that the means of the force-directed f and the orthogonal o test results or response times are different in the samples although there is no difference in the population. If $p(f,o)$ is smaller then 5% then we say that the difference of the means is statistically significant. Analogously, $p(o,h)$ is the error probability for comparing the orthogonal and the hierarchical layout, and $p(f,h)$ for comparing the force-directed and the hierarchical layout. In the tables, we set all statistically relevant results in bold face.

The force-directed layout outperformed the other layout methods for all tasks considered in this study. We don't want to imply that there is no need for the other layout methods. There are many tasks that we did not cover in this study. For example, we would expect that the hierarchical layout would perform better for finding parent nodes or following

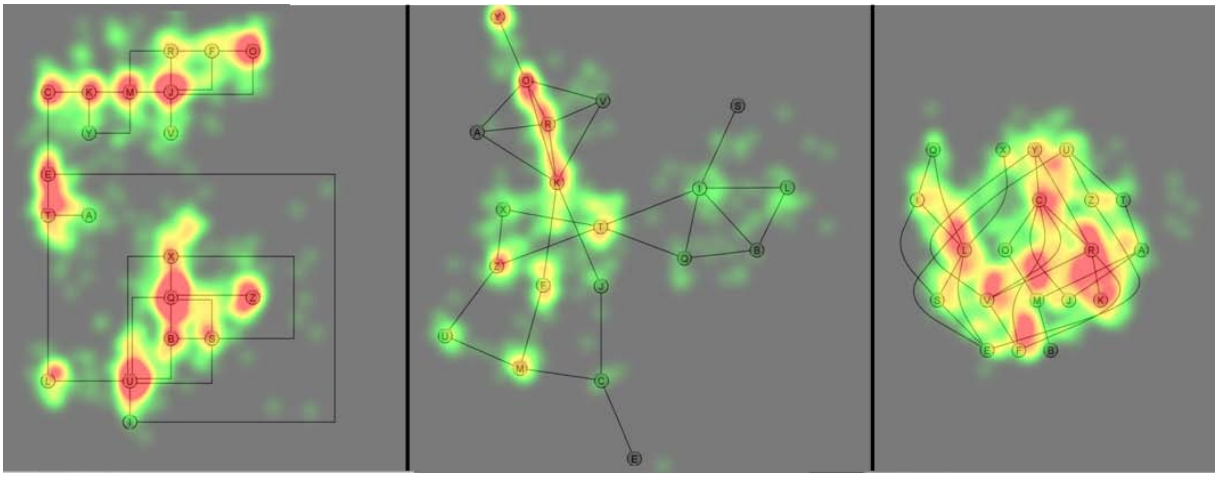

Figure 4: *Task 2: Heatmaps for the orthogonal, force-directed and hierarchical layouts of the graph with 20 nodes.*

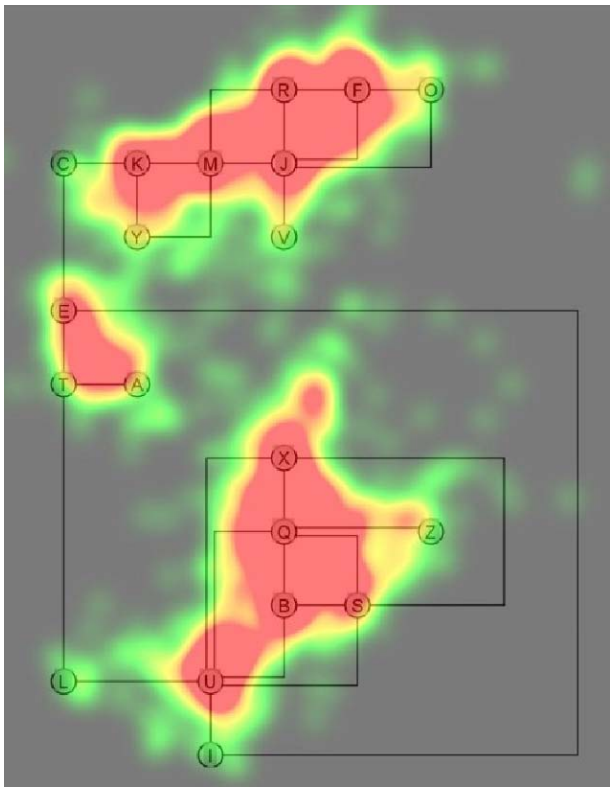

Figure 5: *Task 3: Heatmap for orthogonal layout with 20 nodes.*

flows in graphs with a clear hierarchical structure. We also did not evaluate the augmentation of graphs with additional information, e.g. complex nodes or color coding.

When designing the experiment we discussed the pros and cons of most decisions made among the team members. We strived to reduce/avoid a selection bias (choice of sample graphs, tasks, subjects, and tools and parameters for actually drawing the graphs).

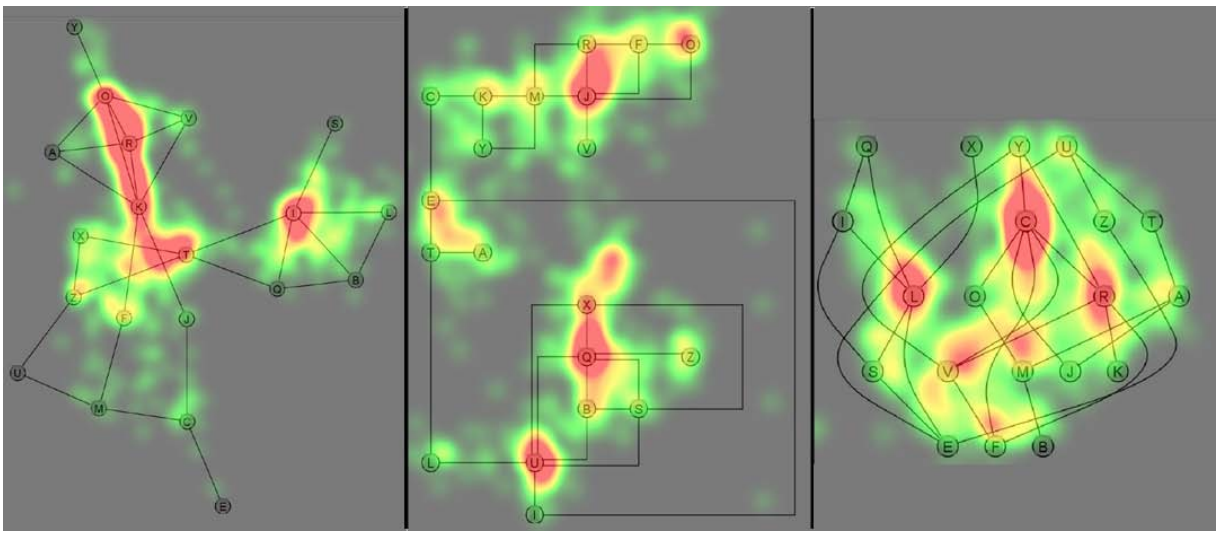

Figure 6: *Task 5: Heatmaps for the three different layouts of the graph containing 20 nodes.*

Table 1: *T-test statistics for number of correct answers.*

		Size of Graph							
		10		15		20		10,15, or 20	
		Mean	T-Test	Mean	T-Test	Mean	T-Test	Mean	T-Test
Task 1	force-directed	1,000	p(f,o) = 0,3208	1,000	p(f,o) = 1,0000	0,972	p(f,o) = 0,3208	0,991	p(f,o) = 1,0000
	orthogonal	0,972	p(o,h) = 0,3208	1,000	p(o,h) = 1,0000	1,000	p(o,h) = 0,3208	0,991	p(o,h) = 1,0000
	hierarchical	1,000	p(f,h) = 1,0000	1,000	p(f,h) = 1,0000	0,972	p(f,h) = 1	0,991	p(f,h) = 1,0000
Task 2	force-directed	1,000	p(f,o) = 0,0788	0,972	p(f,o) = 0,3102	0,806	p(f,o) = 0,3327	0,926	p(f,o) = 0,6244
	orthogonal	0,917	p(o,h) = 0,0788	0,917	p(o,h) = **0,0041**	0,889	p(o,h) = **9E-16**	0,907	p(o,h) = **2E-08**
	hierarchical	1,000	p(f,h) = 1,0000	0,639	p(f,h) = **0,0002**	0,111	p(f,h) = **1E-11**	0,583	p(f,h) = **1E-09**
Task 3	force-directed	0,750	p(f,o) = 0,6048	0,972	p(f,o) = **0,0000**	0,722	p(f,o) = **0,0040**	0,815	p(f,o) = **5E-06**
	orthogonal	0,694	p(o,h) = 0,2828	0,500	p(o,h) = **0,0138**	0,389	p(o,h) = **0,0356**	0,528	p(o,h) = 0,4137
	hierarchical	0,806	p(f,h) = 0,5771	0,778	p(f,h) = **0,0122**	0,167	p(f,h) = **3E-07**	0,583	p(f,h) = **0,0002**
Task 4	force-directed	1,000	p(f,o) = **6E-09**	0,861	p(f,o) = **1E-09**	0,417	p(f,o) = **0,0412**	0,759	p(f,o) = **2E-13**
	orthogonal	0,444	p(o,h) = **0,0324**	0,222	p(o,h) = 0,3651	0,194	p(o,h) = 0,5338	0,287	p(o,h) = 0,5568
	hierarchical	0,694	p(f,h) = **0,0002**	0,139	p(f,h) = **8E-13**	0,139	p(f,h) = **0,0080**	0,324	p(f,h) = **2E-11**
Task 5	force-directed	0,917	p(f,o) = 0,4603	0,667	p(f,o) = 0,4438	0,722	p(f,o) = 1,0000	0,769	p(f,o) = 0,8717
	orthogonal	0,861	p(o,h) = 0,4603	0,750	p(o,h) = 0,3129	0,722	p(o,h) = 0,7989	0,778	p(o,h) = 0,6327
	hierarchical	0,917	p(f,h) = 1,0000	0,639	p(f,h) = 0,8078	0,694	p(f,h) = 0,7989	0,750	p(f,h) = 0,7516

10. Threats to validity

At this point it should be mentioned that the presented results might be influenced by the actual setup. All layout algorithms can be configured by many parameters. The coefficients for attractive and repulsive force in Fruchterman and Reingold's approach certainly have an impact on the final drawings. The same holds for the space between two consecutive layers in Sugiyama's hierarchical layout. However, based on the results from the pre-test we don't believe that different configurations of the layout algorithms lead to completely opposing results.

11. Conclusion

In our task-oriented analysis we found that force-directed layout outperformed the other layouts for Task 2, 3 and 4, while for Task 1 and 5 all three layouts performed equally well. By analyzing the heatmaps produced from the recorded eye-tracking data, we tried to explain these results. For Task 2 we found edge crossings to be confusing in the hierarchical layout, and for Task 3 and 4 the number and size of groups of nodes inspected was much lower for force-directed layout. For Tasks 1 and 5 the subjects basically read the graph from left to right, top down pretty much independent of the edge routing.

Coming back to the two questions raised in the introduc-

Table 2: *T-test statistics for response time of correct answers.*

		Size of Graph							
		10		15		20		10,15, or 20	
		Mean	T-Test	Mean	T-Test	Mean	T-Test	Mean	T-Test
Task 1	force-directed	2	p(f,o) = 0,4421	3	p(f,o) = 0,9382	4	p(f,o) = **0,0212**	3	p(f,o) = 0,1599
	orthogonal	2	p(o,h) = **0,0379**	3	p(o,h) = 0,6073	3	p(o,h) = 0,9428	3	p(o,h) = 0,2603
	hierarchical	2	p(f,h) = 0,0867	3	p(f,h) = 0,4828	3	p(f,h) = **0,0031**	3	p(f,h) = **0,0064**
Task 2	force-directed	8	p(f,o) = 0,0920	8	p(f,o) = **2E-11**	19	p(f,o) = 0,1891	11	p(f,o) = **0,0009**
	orthogonal	9	p(o,h) = 0,4613	15	p(o,h) = 0,1592	18	p(o,h) = **3E-06**	14	p(o,h) = 0,5696
	hierarchical	9	p(f,h) = 0,3018	18	p(f,h) = **9E-09**	32	p(f,h) = **4E-06**	13	p(f,h) = 0,0651
Task 3	force-directed	17	p(f,o) = 0,2341	11	p(f,o) = **3E-09**	16	p(f,o) = **0,0054**	14	p(f,o) = **3E-08**
	orthogonal	20	p(o,h) = 0,1955	24	p(o,h) = 0,3762	23	p(o,h) = 0,7792	22	p(o,h) = 0,7829
	hierarchical	23	p(f,h) = **0,0215**	22	p(f,h) = **1E-08**	21	p(f,h) = **0,0342**	22	p(f,h) = **2E-09**
Task 4	force-directed	6	p(f,o) = **2E-09**	11	p(f,o) = **1E-08**	17	p(f,o) = **0,0056**	10	p(f,o) = **1E-13**
	orthogonal	18	p(o,h) = 0,1087	29	p(o,h) = 0,6060	28	p(o,h) = 0,2292	23	p(o,h) = 0,0654
	hierarchical	14	p(f,h) = **0,0000**	26	p(f,h) = **5E-05**	33	p(f,h) = **0,0023**	19	p(f,h) = **9E-08**
Task 5	force-directed	12	p(f,o) = 0,0553	15	p(f,o) = 0,3791	14	p(f,o) = 0,2755	14	p(f,o) = 0,1590
	orthogonal	14	p(o,h) = 0,5258	14	p(o,h) = 0,4267	16	p(o,h) = 0,9744	15	p(o,h) = 0,3899
	hierarchical	15	p(f,h) = **0,0393**	15	p(f,h) = 0,9546	16	p(f,h) = 0,3732	15	p(f,h) = **0,0436**

tion. We think that for algorithm designers our most important finding is that edge crossings pose little problems in orthogonal graph drawings, but that the length of edges and the number of bends make finding subgraphs difficult. In many heatmaps we found indications of a tunnel effect, i.e. subjects did rarely focus nodes on the periphery.

Finally, one goal of the task-oriented analysis was to come up with a table were application developers could identify the layout method best suited for the tasks relevant for their application. As it turned out, for the tasks considered in this study we need no such table, because force-directed worked best for all tasks.

As part of our future work we want to perform a similar study with a larger variety of tasks. In particular, we want to identify tasks that are supposed to be better solved using hierarchical or orthogonal layout methods.

Acknowledgements

Felix Bott helped during the eye-tracking study, Carsten Görg and Peter Birke provided helpful comments on earlier revisions of this paper.

References

[BRSG06] BENNETT C., RYALL J., SPALTEHOLZ L., GOOCH A.: The aesthetics of graph visualization. In *Proceedings of the International Symposium on Computational Aesthetics in Graphics, Visualization, and Imaging* (Banff, Alberta, Canada, 2006), pp. 57–64.

[FK95] FÖSSMEIER U., KAUFMANN M.: Drawing high degree graphs with low bend numbers. In *Proceedings of Symposium on Graphdrawing, GD* (1995), vol. 1027 of *LNCS*, pp. 254–266.

[FR91] FRUCHTERMAN T. M. J., REINGOLD E. M.: Graph Drawing by Force-directed Placement. *Software, Practice, and Experience. 21*, 11 (1991), 1129–1164.

[HE05] HUANG W., EADES P.: How people read graphs. In *Proceedings of Asia Pacific Symposium on Information Visualisation 2005 (APVIS 2005)* (2005), vol. 45 of *Conferences in Research and Practice in Information Technology*, Australian Computer Society Inc.

[Hua07] HUANG W.: Using eye tracking to investigate graph layout effects. In *Proceedings of Asia Pacific Symposium on Information Visualisation 2007 (APVIS 2007)* (2007), IEEE Computer Society.

[PC02] PURCHASE H. C., CARRINGTON D. A.: Empirical evaluation of aesthetics-based graph layout. *Empirical Software Engineering 7*, 3 (2002), 233–255.

[PCJ96] PURCHASE H. C., COHEN R. F., JAMES M.: Validating graph drawing aesthetics. In *Graph Drawing (Proceedings of GD '95)* (Berling, Heidelberg, New York, 1996), Brandenburg F. J., (Ed.), vol. 1027 of *Lecture Notes Computer Science*, Springer, pp. 435–446.

[PMCC01] PURCHASE H. C., MCGILL M., COLPOYS L., CARRINGTON D. A.: Graph drawing aesthetics and the comprehension of UML class diagrams: An empirical study. In *Australasian Symposium on Informatoin Visualization* (2001).

[PRB08] POHL M., REITZ F., BIRKE P.: As time goes by - integrated visualization and analysis of dynamic networks. In *Proc. of 9th Int. Working Conference on Advanced Visual Interfaces, AVI* (2008).

[Pur97] PURCHASE H. C.: Which aesthetic has the greatest effect on human understanding? In *GD '97: Proceedings of the 5th International Symposium on Graph Drawing* (Berlin, Heidelberg, New York, 1997), Springer, pp. 248–261.

[STT81] SUGIYAMA K., TAGAWA S., TODA M.: Methods for Visual Understanding of Hierarchical Systems. *IEEE Transactions on System, Man and Cybernetics SMC 11*, 2 (1981), 109–125.

[WPCM02] WARE C., PURCHASE H., COLPOYS L., MCGILL M.: Cognitive measurements of graph aesthetics. *Information Visualization 1*, 2 (2002), 103–110.