

A Comparison of User-Generated and Automatic Graph Layouts

Tim Dwyer, Bongshin Lee, Danyel Fisher, Kori Inkpen Quinn,
Petra Isenberg, George Robertson, and Chris North

Abstract—The research presented in this paper compares user-generated and automatic graph layouts. Following the methods suggested by van Ham et al. (2008), a group of users generated graph layouts using both multi-touch interaction on a tabletop display and mouse interaction on a desktop computer. Users were asked to optimize their layout for aesthetics and analytical tasks with a social network. We discuss characteristics of the user-generated layouts and interaction methods employed by users in this process. We then report on a web-based study to compare these layouts with the output of popular automatic layout algorithms. Our results demonstrate that the best of the user-generated layouts performed as well as or better than the physics-based layout. Orthogonal and circular automatic layouts were found to be considerably less effective than either the physics-based layout or the best of the user-generated layouts. We highlight several attributes of the various layouts that led to high accuracy and improved task completion time, as well as aspects in which traditional automatic layout methods were unsuccessful for our tasks.

Index Terms—Graph layout, network layout, automatic layout algorithms, user-generated layout, graph-drawing aesthetics.

1 INTRODUCTION

A great deal of research has been done on the algorithmic problems of automatically arranging the nodes and edges of a graph to make a pleasing picture. While algorithm designers have developed a number of “rules-of-thumb” for what makes an effective layout—such as minimizing edge crossings or maximizing the angles of incidence of edges where they connect to nodes—the problems of optimizing such criteria are computationally very challenging and so heuristics have to be adopted to achieve approximate solutions. Further, improving the drawing with respect to one of these criteria may require a trade-off with respect to one or more of the others. For example, algorithms exist to lay out a planar graph with no crossings but they do so at the expense of very poor angular resolution. For these reasons, algorithm design must be informed by studies of the human factors in readable layout to decide which heuristics are most important to optimize. As described in Section 2, such studies have relatively recently begun, but the work is far from conclusive.

In this paper we describe two studies: the first involving gathering user-generated layouts in a controlled setting; the second, evaluating these layouts and comparing them to the results of automatic layout algorithms. Our first study compares user interaction in a layout task using either a mouse interface or a multi-touch table-top interface. Compared to a simple sorting task (see [14] for details) we find that users use multi-touch less in the more cognitively demanding layout task, yet still, most users did use significant multi-touch interaction in manipulating groups of nodes. In our second study we find that the best user generated layouts were preferred by more people than either orthogonal or circular automatic layout styles. However, an automatic force-directed layout was the most popular layout style overall. In supporting task accuracy and speed the force-directed layout and the best user layouts were comparable while the automatic orthogonal layout was particularly

difficult for users, even though it had the fewest edge crossings of all the layouts.

2 RELATED WORK

Automatic graph layout is usually posed as an optimization problem where a “good” layout is found by searching for a configuration of nodes and edges that is optimal with respect to various aesthetic criteria. For example, edge crossings, bends, angle of incidence and length are routinely cited (e.g., [2][7]) as important in algorithm design. However, compared to the body of literature on graph-layout algorithms, the number of studies evaluating the aesthetic and perceptual criteria on which those algorithms are based is relatively small. Various studies (e.g., Purchase et al. [15], Huang et al. [11]) have found that unnecessary crossings do have an adverse effect on human understanding of graph drawings but crossing minimization alone is an intractable task [9]. Further, optimization of one aspect of layout may compromise other aspects of drawing quality. For example, some algorithms that seek to reduce edge crossings may decrease the symmetry of a graph. These and other studies (especially by McGrath et al. [12]) have also demonstrated that fairly well understood Gestalt principals also apply to graph layout. For example, spatial proximity between nodes implies logical groupings, whether or not such groupings are reflected by the connected structure of the graph. More recent studies by Purchase et al. [16][18] have focused on the problem of mental-map preserving dynamic graph layout.

Recent work by van Ham and Rogowitz [10] departed from the earlier “evaluation only” studies by allowing on-line users to generate their own layouts of a small graph. They examined their participants’ manually-arranged graphs for evidence of the commonly accepted criteria for good graph drawing, but also looked for new, or at least previously untested, principals and aesthetics. The authors did claim one such aesthetic, which was a tendency to arrange clusters inside convex-hulls. They attributed this result to the Gestalt principal of “closure” in which a closed boundary implies a grouping. This paper takes inspiration from the user-generated layout study of van Ham and Rogowitz (HR) [10], but the work presented here has a number of key differences.

The first and most significant difference is that our study is in two parts: generating the layouts and then evaluating the layouts. HR drew conclusions about what makes an effective layout based on how users tried to arrange their graphs. We contend that what users want or need from layout may be different from what they are able to easily achieve manually and therefore a separate evaluation phase is

-
- Tim Dwyer, Bongshin Lee, Danyel Fisher, Kori Inkpen Quinn and George Robertson are with Microsoft Research, E-Mail: {timdwyer, bongshin, danyelf, kori, ggr}@microsoft.com.
 - Petra Isenberg is with the University of Calgary, E-Mail: petra.isenberg@ucalgary.ca
 - Chris North is with Virginia Tech, E-Mail: north@cs.vt.edu

Manuscript received 31 March 2009; accepted 27 July 2009; posted online 11 October 2009; mailed on 5 October 2009.

For information on obtaining reprints of this article, please send email to: tvcg@computer.org.

required to tease these apart. By contrast, evaluation in HR was entirely by the researchers not by the general users who interpret the graphs. Since aesthetics are subjective and an important criterion in assessing the effectiveness of layout is how well it supports analysis tasks, we felt that a follow-up user study was a more impartial way to evaluate the user-generated layouts.

Another significant difference is that the layout generation part of our study was completed in a controlled lab environment with 32 participants under different conditions, resulting in 48 user-generated layouts. By contrast HR was an anonymous on-line task with 73 submitted layouts. Although we had fewer user-generated layouts, they still exhibited wide variety. We felt it was more important to have a large number of participants in the evaluation phase to find statistically significant differences in user preference.

We evaluated the graph layouts with a larger set: 194 participants in an online study to rate the graphs. The participants compared a selection of user-generated layouts with three layouts generated by popular automatic-layout algorithms. This allowed us to correct for aspects that are particularly difficult for users to handle.

Last, our selection of graphs attempted to illustrate a variety of different topological challenges, and directed users to address specific goals in both visualizing and evaluating the graphs. Further differences are addressed in more detail in Section 3.1.3 below.

3 COLLECTING USER-GENERATED LAYOUT

The first part of our study involved observing participants in a usability-lab as they completed a manual graph-layout task. This was part of a larger study that sought to examine how users manipulate many nodes to perform two types of tasks (sorting and graph-layout) when using three different types of user interface paradigms (a multi-touch surface, a mouse, and physical objects). The results for the sorting task are reported in [14]. Here we present the results for the graph-layout task. We chose to investigate multi-touch in the hopes that users would find multi-touch to be a more natural way to manipulate graphs, and so alleviate some of the difficulties in traditional graph layout. In particular, it might facilitate manipulation of large parts of the graph with a single gesture.

3.1 Study Design

The goal of the collection phase of the study was to examine what strategies users apply to graph layout, how they manipulate the graphs given different interactive paradigms, and what kind of layouts they produce.

3.1.1 Participants

We recruited 32 participants (25 males and seven females) and two pilot testers via email from our institution. They were primarily researchers and software developers who were frequent computer users. The average age of participants was 34, ranging from 21 to 61. Participants each received a \$10 lunch coupon for their participation. None of the participants were experts in graph drawing.

3.1.2 Conditions and Groups

We compared three interface conditions: Surface, Physical and Mouse. However, the graph-layout task had no corollary in the Physical setup, so only the Surface and Mouse conditions are relevant here.

The Surface condition used a Microsoft Surface system, with 1024×768 resolution. The multi-touch implementation enabled users to manipulate graph nodes by drag-n-drop in two ways: (1) Users can select and drag multiple individual nodes with multiple fingers. (2) Users can select spatial groups of nodes by touching the surface with at least three fingers at the same time (or a palm). All nodes inside the convex hull defined by the fingers become selected and can be manipulated with affine transformations based on the users' drag motions, until all fingers in the group are released. For example, users can condense or spread out a group by moving their fingers or hands together or apart.

The Mouse condition used a standard desktop system. This application supported standard Windows multi-selection techniques: clicking to select single nodes, and control- or shift-clicking or marquee selection (by drawing a rectangle) to select multiple nodes.

Participants were randomly divided into four groups that determined which two of the systems they used and the order (because of the goals of the broader experiment in [14]): Surface then Mouse (SM), Mouse then Surface (MS), Surface then Physical (SP), Physical then Surface (PS). All 32 participants generated a Surface graph layout; in addition, 16 also generated a layout with the Mouse.

3.1.3 Task

The graph-layout task immediately followed the sorting task, which asked users to pile and sort 200 small, colored tokens into piles. In the graph-layout task, participants were asked to lay out a social network consisting of 50 nodes and about 75 links. In a pilot study, we found that an initial random layout was so messy that users felt overwhelmed by the task and that force-directed layout already looked too "finished." As such, the initial arrangement was generated using the very basic circular layout in the yEd graph editor (V3.2 from www.yfiles.com). Feedback from pilot participants also suggested that they were less intimidated by the layout task if we told them that it was a social-network or a map of "friendship" relations. Van Ham and Rogowitz also posed the problem in this way but used random names for node labels to help people interpret the abstract graph as a social network. However, they found that these labels caused participants to associate too much meaning with the labels. For example, some participants focused on grouping male and female names or creating male/female pairings. We wanted users to focus on the graphs' connected structure so we replaced node labels with uniform anthropoid glyphs (smiley faces, about 1.5 cm in diameter, see Fig. 4). This was sufficient to help users understand that the nodes represented people.

We generated two graphs (for participants who used both Surface and Mouse interface, with graph order varied independently of interface order) with similar structure: each with 50 nodes, and similar edge density (74 and 77 edges). Each graph featured similar topological structures: a small complete sub-graph (a clique); a nearly complete sub-graph; a large cycle; a chain of degree 2 vertices; a cut vertex; and a several leaves (degree one nodes). In order to ensure that participants focused on graph utility, they were asked in written instructions to "arrange the nodes in a way that you think would best enable you to analyze four things: (1) tightly interconnected groups of friends, (2) persons that tie two different groups together, (3) long chains of friends (friend of a friend of a friend...), and (4) peripheral people that have few friends." The participants also asked to generate only one layout to analyze four things at the same time. Thus, in comparison to HR, our study used larger graphs (their graphs had only 17 nodes) with a more diverse topology (their topology consisted of two "clusters"), and provided a more specific task goal.

3.1.4 Procedure

Each participant completed one graph-layout task for each interface condition assigned to them (except for the Physical condition). The task was not timed, and participants could take as long as they desired until satisfied with their layout. Participants were also encouraged to 'think aloud'. On each interface they were given a brief tutorial about the interaction features prior to the graph-layout task. At the end of each condition participants answered a questionnaire about their experience. At the end of the session participants answered a final questionnaire comparing the systems. We recorded video, audio, and event logs of the sessions.

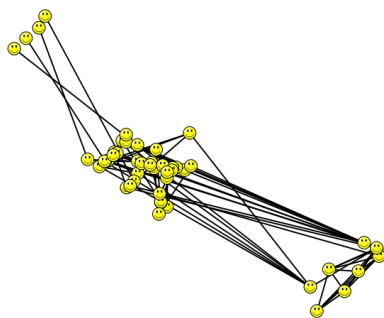


Fig. 1. This user-generated layout is a very literal interpretation of the task by one user who attempted to sort the nodes into three piles: leaves in the top-left, clique on the bottom-right and everything else in the middle.

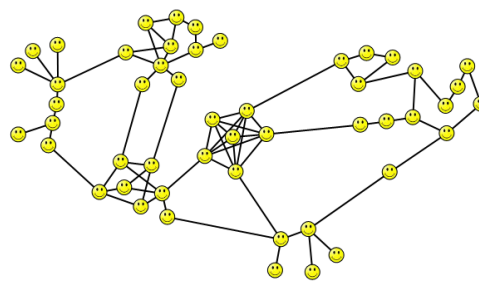


Fig. 2. A user-generated layout with an interesting grouping of nodes. Several users made such groupings of 5-7 nodes. The groups did not necessarily follow the graph connectivity, rather it seemed users were less overwhelmed when they could work on isolated parts of the graph.

3.2 Results

While users found the graph-layout task challenging, they generated a wide variety of different graph layouts. Two interesting sample graph layouts are shown in Figures 1 and 2; G3-G11 of Figure 9 show even more layouts.

Layout Metrics: As mentioned above, graph layout quality is highly subjective since it relies on both personal aesthetic preference and on the particular application domain. The task given to the users deliberately avoided mentioning any of the traditional aesthetic criteria, such as number of edge crossings or edge length variance specifically. We were interested to see which, if any, of these the users found important in completing a general graph exploration/layout task.

Generally we found that users focused on removing edge crossings in both conditions (Surface or Mouse). When we asked participants to describe their process, they used phrases like “untangle,” “separate the lines,” and “move overlapping lines” to describe their planarization process. After each condition, participants rated their satisfaction with the clarity of their final graph layout. Of the participants that used both the Surface and the Mouse, they were somewhat more satisfied with their final Surface layout than their Mouse layout (Surface 5.5, Mouse 4.9, $p=.055$)

Multi-Touch Surface Interaction: We wanted to understand to what degree users used multi-touch interaction to manipulate the graph. In our log files for the Surface condition we collected detailed finger contact information. We devised a multi-touch metric:

$$MT = (TC - IC) / TC \quad (1)$$

where TC is the *total* number of contacts made during the task and IC is the number of *isolated* contacts, i.e. single finger contacts that did not overlap (in time) with other contacts. Thus, a user who completed the entire task using only one finger contact at a time would have $MT = 0$ and a user who never lifted a finger off the table without first making at least a second contact would have $MT = 1$.

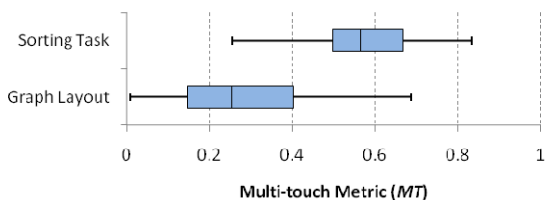


Fig. 3. Degree of multi-touch interaction employed by users in graph-layout was less than in the sorting task.

We found that users used multi-touch interaction in the graph-layout task (mean $MT = 0.27$) significantly less than they did in the far simpler sorting task (mean $MT = 0.56$) ($p=0.00$). Further, the degree of multi-touch interaction employed in the graph-layout task varied more from user to user than in the sorting task, see Fig. 3. We suggest that this variance points to the difficulty of the task: users would try to limit the number of changes happening at once by moving to simpler interactions. Nonetheless, while they used *less* multi-touch, it was still a significant part of their work. Our coding of the video observes that all but eight of the 32 tabletop users used multi-touch operations repeatedly throughout the task.

Graph Layout Operations: We observed a common sequence of operations: organizing a highly-connected sub-graph using single node operations, then using a group operation to move and/or shrink that sub-graph into position, then proceeding to another portion of the graph. Another common operation was to layout in terms of edges rather than nodes, by moving the two nodes on either end of a single edge. In interviews afterward, several users reported wanting to generalize this technique to paths, manipulating the shortest path between two nodes as a group operation.

The Surface appears to have encouraged users to “think with their hands,” as suggested by the principles of embodied interaction [4]. Of the 16 participants who used both Surface and Mouse, on the Surface they used on average 277 touches to move nodes about 240 times. With the Mouse, they used on average of 103 mouse clicks to move nodes 117 times. Yet they took about the same amount of time on average, about 5.5 minutes. These additional touches often came in the form of users ‘poking’ or wiggling nodes to make very minor adjustments. In some cases, it seemed that users did this in order to make very minor optimizations to the layout, or to better see the edges associated with the poked node (edges of a selected node were highlighted). This behavior was not observed in the Mouse condition. (It is also possible that the Surface may suffer from a lack of precision of the touch-based interface.)

The Surface affords efficient translation, rotation and scaling of group selections in a single combined gesture. Of the 16 participants who used both Surface and Mouse, 13 used grouping on the Surface (using the multi-finger convex hull selection technique) to translate nodes, and did this 18 times on average. About half of them used large rotations (more than 45 degrees) on one or more occasion and 19/32 used significant scaling (more than 50% increase or decrease). With the Mouse, 15 of them used grouping (using the standard multi-select features) to translate nodes an average of eight times. Rotation and scaling were not available on the Mouse interface.

Users on the Surface were observed to manipulate long chains of nodes and small highly connected cliques of nodes with these group operations. For example, the orientation of a chain of nodes, arranged in a linear way, can be quickly rotated (e.g., from vertical to horizontal) if doing so fits better with the arrangement of the

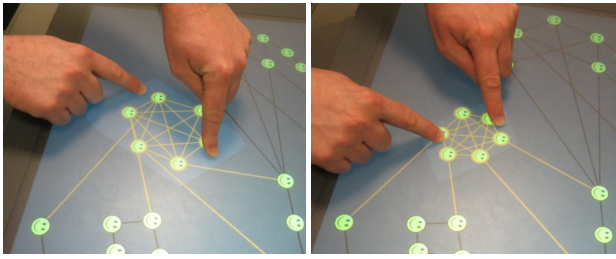


Fig. 4. A participant using the surface to scale and rotate a sub-graph so that the local topology is unchanged while improving the global topology. The group was initially selected using 4 fingers, but then 2 fingers were lifted while rotating the group.

surrounding graph. Further, the chain can be stretched along its principal axis in the same gesture. For cliques with several connections to the surrounding graph such a gesture is also useful to collapse the clique (since its internal structure is simply a complete graph and therefore possibly equally well represented as a tight “bundle”), and several users made such bundles after organizing the clique (Fig. 4). In general terms, a component of the graph can be scaled or rotated such that its local topology (in terms of crossings, relative edge lengths, and so on) can be preserved, while at the same time potentially improving the topology with respect to the surrounding graph.

4 EVALUATING USER-GENERATED LAYOUT

The first study generated 48 hand-made layouts. While some were obviously poor (some users gave up on the task), most seemed to largely succeed at the task. A second study was conducted to evaluate the user-generated layouts and to compare their effectiveness and aesthetics with that of popular automatic layout algorithms.

4.1 Study Design

To gather evidence for comparative effectiveness of user-generated versus automatically generated graph layouts, we designed a web-based study for a large set of participants.

4.1.1 Data Graphs

The nine best user-generated layouts were selected from the first study. We selected the 9 of the 32 tabletop layouts, prioritizing graphs with the fewest edge-crossings (since most studies in this area agree that edge-crossings are a significant factor in layout effectiveness, see Section 2).

Three layouts were also generated automatically with widely used types of layout algorithms for drawing general undirected graphs: force-directed, orthogonal and circular layout. The choices we made correspond to the state of the art in terms of practical layout algorithms for general undirected graphs. Other layout methods occasionally discussed in graph drawing papers, are mainly of theoretical interest since they can only be used with a very restricted set of graphs (e.g., planar, 3-connected, symmetric, etc.).

We used the yEd tool to generate each of the graphs. The yFiles documentation does not specify exactly which algorithms are used; however, the “smart organic layout” seems to follow standard force-directed approaches, the orthogonal layout (“classic orthogonal”) is a *topology-shape-metrics* approach [6], and the circular layout is similar to [3]. Each algorithm has many parameters. We decided to use the defaults because we found that they produced reasonable results. Furthermore, assuming that the yEd tool is implemented to have reasonable default behavior, we believe that it would be more objective and reproducible to use the default values rather than fiddling with the parameters ourselves.

4.1.2 Participants

We sent an invitation to a “Graph Layout Challenge” to approximately 6500 people with diverse backgrounds using email distribution lists within our company. The invitation described the goal of our study and indicated that the study was web-based and would take approximately ten minutes to complete. Two \$100 Visa gift cards were offered to motivate participation, one to be randomly drawn from the names of all participants and the other to be awarded to the person who completed all tasks fastest without any errors. Of those 6500 invitations, 194 people (179 male, 15 female) completed the online tasks and survey. Age range was between 21 and 60; most people were aged 26-30.

4.1.3 Tasks and Equipment

Since we wanted to evaluate the user-generated layouts from the first study, we used the same four tasks described in 3.1.3. To help participants understand the tasks better, we rephrased the task descriptions so that they could be easily mapped to the basic concepts given to the participants before the timed trials.

Our study platform was implemented as a web browser plugin using Microsoft Silverlight. It recorded task times, answers, as well as preferences and comments into the database once participants completed the entire session. The study materials were presented within a 1280×1024 window and the tasks were completed with a 1024×768 view of the graph.

Users were required to complete four different tasks for each graph they selected:

Task 1: Select a clique of six people. A clique is a group of more than two friends who are all directly connected to each other. The study graph had one clique of six.

Task 2: Select a chain of four people. A chain is a series of neighboring nodes, each with only two connections. The study graph had one chain of four.

Task 3: Select a cut node. A cut node is a person who, if removed, would leave two separate components. The study graph had one cut node.

Task 4: Select leaf nodes. That is, people with only one connection. The study graph had eight leaf nodes.

4.1.4 Procedure

The web-based study began by asking the participant’s gender, age group, and occupation. Since we could make no assumptions about knowledge of graph understanding tasks, the study then stepped through descriptions of basic concepts necessary to complete four tasks: the idea of a “friendship network,” what a clique is (Fig. 5), what a cut node is, what a chain is, and what leaf nodes are.

Next, the participant was asked to select the graph layout that they believed would be best for doing the four tasks (Fig. 6). The

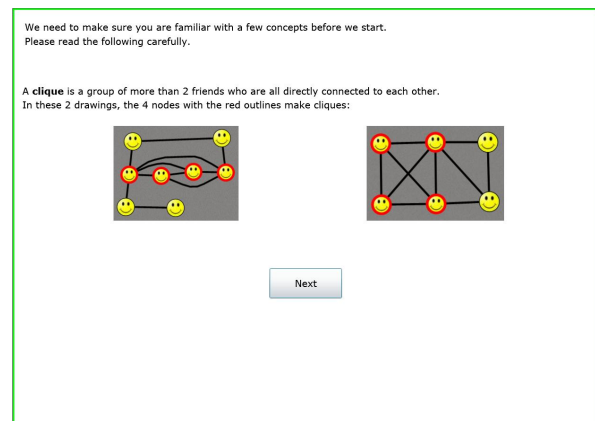


Fig. 5. Instruction page describing a clique.

nine user-generated layouts and three automatic layouts were intermingled, and the order of all twelve layouts was randomized for each participant.

Participants were then asked to select a second best layout from the other subset depending on their first choice. For example, if a user-generated layout was the first choice, the second choice had to be from the automatic layouts. Fig. 7 shows the remaining automatic layouts after the participant selected one of the user-generated layouts as his/her first choice. Notice that the position of those is the same as in Fig. 6.

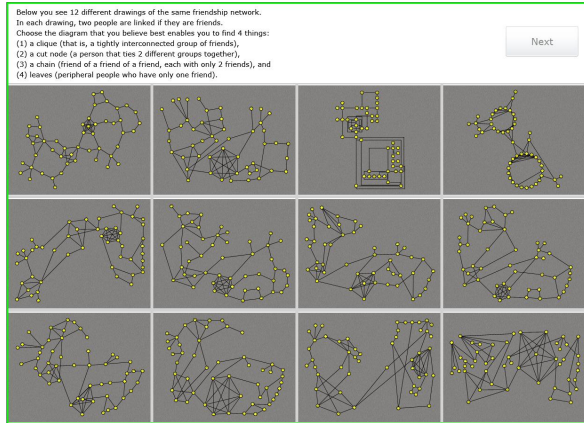


Fig. 6. Participant selects graph layout believed to be most effective for the tasks presented.

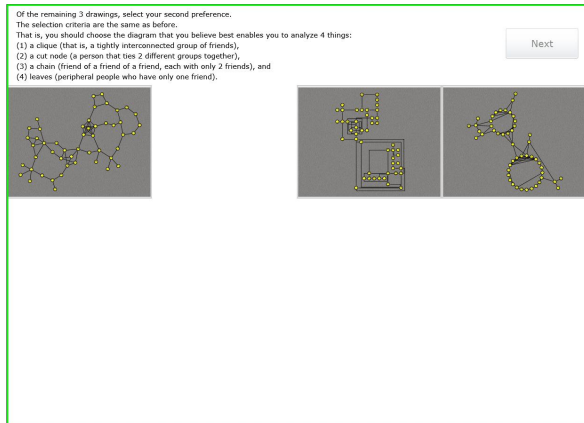


Fig. 7. Participant makes second choice selection amongst the automatic layouts when he/she selected one of the user-generated layouts as his/her first choice. Image cropped slightly for detail.

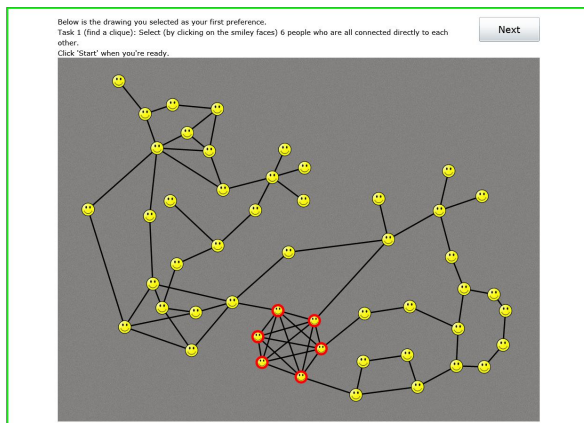


Fig. 8. The clique task after the participant has selected the six nodes.

Participants were then asked to perform the four tasks (fixed order), first with one of their choices, and then with the other choice (randomly ordered). In each case, the task was presented before showing the graph. The participant clicked on a “Start” button when ready, and the graph was shown. The participant then clicked on one or more nodes to select the answer, and then on a “Next” button when done. Fig. 8 shows the clique task after the participant selected the six nodes and before clicking on the “Next” button.

After performing the four tasks on both graph layouts, the participant was asked some simple survey questions, basically asking which graph layout between from the first and second graphs they selected before they performed the tasks - was more helpful for all tasks, which was more helpful for each task, and an open-ended question about what aspects of the layout influenced their selections.

As a final step, the participant was given the same set of twelve graph layouts, in the same order as was shown when the first choice was made (same as in Fig. 6), and asked to confirm their choice of the best layout. We asked this last question to see if performing the tasks would cause them to change their first choice.

4.2 Results

The web-based survey was answered by 194 distinct users. After eliminating six outliers (users who were more than three standard deviations from mean, getting zero or one tasks correct, or taking more than 280 seconds for the eight tasks), we have 188 distinct users, each of whom chose a first- and second-choice graph, finished four timed tasks with each graph, and chose a final choice afterward.

4.2.1 Layout Preference

Fig. 9 shows the twelve layouts labeled G0 through G11, and annotated with the number of first choice selections. The most popular first choice graph was G2, the automatic physics-based layout, with 48 choices. G11 was the second-most popular, with 41. The circular layout and orthogonal layout were far less popular.

Of the 117 users whose first choice was a user-generated layout (G3-G11), 103 of them chose the physics-based layout G2 as their second choice. In the final choice, after users had been through the study once, 80 of the full 188 selected G2. Although ten people picked G0, the orthogonal layout, as either their first or second choice, it was no-one’s final choice.

Chi-square analyses revealed that significantly more people selected one of their initial selections as their final choice than a different graph: $\chi^2=30.72$, $p<.001$; however, there was no significant trend observed as to whether they selected their first or second preference, $\chi^2=2.46$, $p=.117$.

4.2.2 Task Accuracy and Time

Table 1 shows average completion time per task only for the successful trials. G0, G4, G5, G6, G8, and G10 were chosen less than 5% of the time and were therefore not included in the following statistical analyses.

	Task 1 (Clique)		Task 2 (Chain)		Task 3 (Cut)		Task 4 (Leaf)	
	n	Avg. Time	n	Avg. Time	n	Avg. Time	n	Avg. Time
G1	1	38.00	16	14.00	25	2.76	26	14.23
G2	146	8.12	135	7.10	139	3.81	145	9.18
G3	26	5.92	25	6.40	26	3.62	25	12.52
G7	28	9.25	28	6.71	25	5.64	27	17.22
G9	25	16.12	24	6.63	25	5.12	27	10.37
G11	60	7.07	50	13.00	39	4.00	62	12.35

Table 1: Number of successful trials, and average completion time, per task. The time cell is colored from scheme: the best in green, to the worst in red.

Accuracy: A repeated-measures ANOVA revealed a significant main effect of graph ($F_{1,321}=18.15$, $p<.001$) and task ($F_{1,321}=17.66$, $p<.001$) for users’ accuracy. Post-hoc pairwise comparisons using a

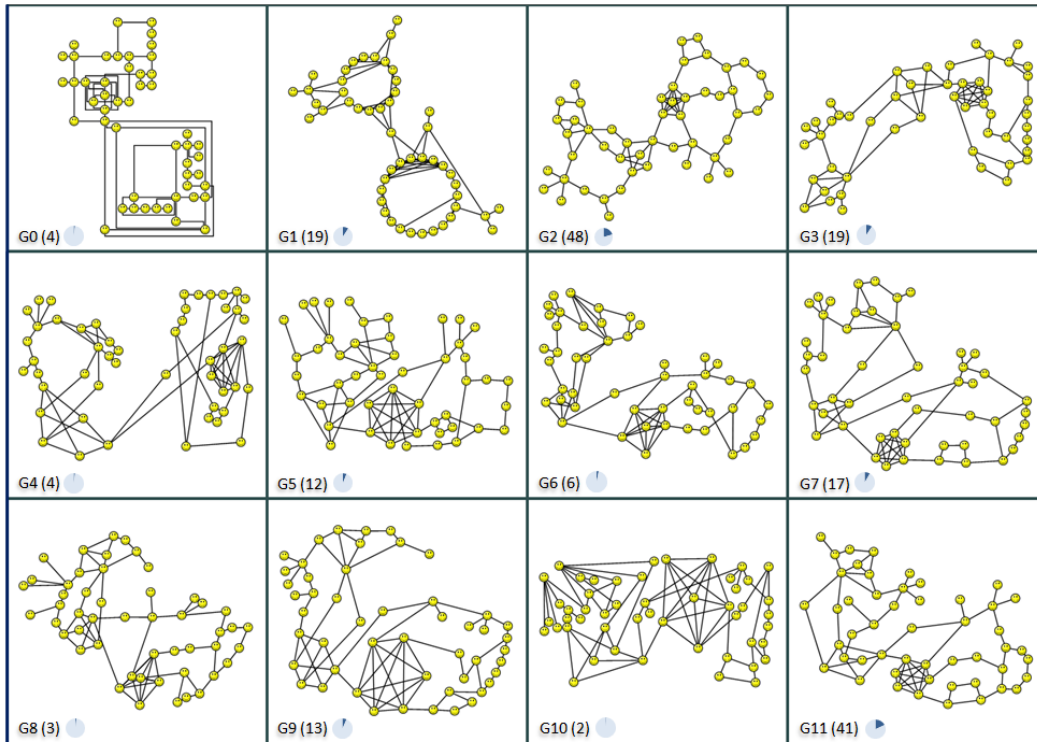


Fig. 9. The 12 layout alternatives presented to the users (labeled G0-11), annotated with the number of first-choice selections (out of 188 distinct users). G0, G1, and G2 are automatic layouts; the rest were user generated layouts.

Bonferroni correction indicated that users accuracy was significantly lower for G1 than any of the other graphs ($p < .001$).

A significant interaction effect was also found between task and graph in terms of accuracy ($F_{15,321}=47.30$ $p < .001$). Analyzing each task separately using ANOVAs revealed significant main effects of accuracy for Task 1 and Task 2 ($F_{5,321}=88.71$, $p < .001$ and $F_{5,321}=4.54$, $p = .001$). For Task 1 (the clique task), post-hoc pairwise comparisons using a Bonferroni correction revealed that G1 was significantly less accurate than all the other graphs ($p < .01$). For Task 2 (the chain task), G1 was found to be significantly less accurate than G2, G3, and G7 ($p < .01$). No significant differences were found in terms of accuracy for Task 3 (the cut point, $F_{5,321}=.73$, $p = .604$) and Task 4 (leaf nodes, $F_{5,321}=.61$, $p = .696$).

Task Time: For task time, only successful trials were analyzed and each Task was analyzed separately using ANOVAs. For Task 1, G1 only had one user complete the task correctly so this graph was omitted from the analysis. Significant main effects of graph selection were found for Task 1, 2, and 4, but not Task 3 ($F_{4,280}=8.08$, $p < .001$, $F_{5,272}=4.00$, $p = .02$, $F_{5,306}=10.23$, $p < .001$, $F_{5,293}=0.93$, $p = .459$, respectively).

For Task 1, G9 was significantly slower than the other graphs considered ($p < .01$). For Task 2, G2 was significantly faster than G11 ($p < .01$).

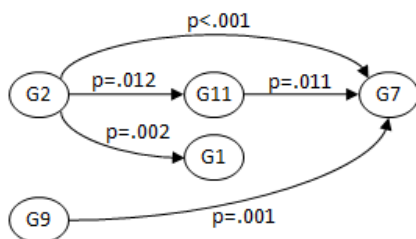


Fig. 10. Statistically significant time differences on Task 4. Each arrow indicates that the graph on the left side of the arrow is faster than the graph on the right side.

Task 4 is more complex. Fig. 10 illustrates the significant differences between the graphs. In G2 and G9, the fastest choices, the leaf nodes were easily accessible: in G2, leaves are all on the outside; on G9, most were clustered in two groups. In contrast, the lesser-performing graphs have leaves that are placed less consistently.

In order to attempt to understand the factors driving accuracy and speed, we examined the number of crossings, the stress [8], and the variance in edge length (see Table 2 and Section 5).

4.2.3 Graph Choice Criteria

Our final questionnaire asked for participants to report which aspects of the graphs most influenced their choices. These answers were coded according to ten main categories and several subcategories. The ten main categories included: node spacing, edge crossings, edge length, graph layout, nature of clique, nature of chains, nature of cut node, nature of leaf nodes, ability to show patterns, and others.

The most common answers are closely related to our tasks. Out of the top five criteria (see Fig. 11), four refer to the criteria we asked participants to focus on when picking the graphs.

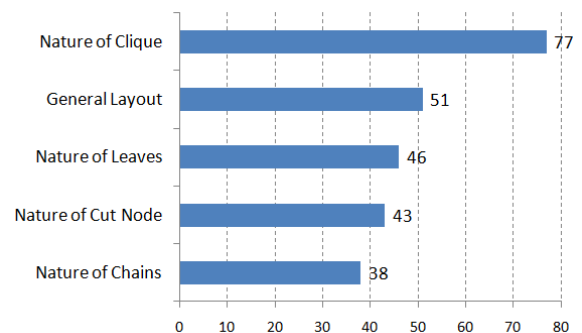


Fig. 11. Frequency of coded responses in the final questionnaire.

	# edge Crossings	Edge Length		User Choice		Fraction Correct by Task				Avg. Correct Time per Task			
		Std Dev	Stress	First	Final	1. Clique	2. Chain	3. Cut	4. Leaf	1. Clique	2. Chain	3. Cut	4. Leaf
G0	7	143.54	335.03	4	0	0.80	0.70	0.90	0.60	26.88	16.71	8.78	35.00
G1	27	57.75	176.11	19	11	0.04	0.59	0.93	0.96	38.00	14.00	2.76	14.23
G2	13	19.18	44.95	48	80	0.97	0.89	0.92	0.96	8.12	7.10	3.81	9.18
G3	18	59.49	136.72	19	20	0.96	0.93	0.96	0.93	5.92	6.40	3.62	12.52
G4	31	132.08	320.64	4	0	1.00	1.00	0.86	0.86	9.14	6.57	4.67	14.17
G5	27	68.47	259.18	12	9	0.94	0.78	0.89	0.94	10.24	6.21	21.69	9.59
G6	14	59.81	180.48	6	7	0.86	0.71	0.86	1.00	16.17	14.00	14.33	15.14
G7	19	63.02	158.36	17	17	0.93	0.93	0.83	0.90	9.25	6.71	5.64	17.22
G8	16	54.12	168.96	3	0	0.75	0.50	0.75	1.00	29.67	9.50	10.00	8.50
G9	23	86.43	250.14	13	12	0.89	0.86	0.89	0.96	16.12	6.63	5.12	10.37
G10	30	110.66	396.99	2	0	1.00	0.00	0.67	0.67	10.00	n/a	4.00	22.50
G11	17	48.04	169.67	41	32	0.94	0.78	0.92	0.97	7.07	13.00	4.00	12.35

Table 2: Graph layout properties and task performance data. The graph is colored along each axis from scheme: the best in green, to the worst in red.

Cliques: Participants reported to have paid most attention to the nature of cliques (77 of the 188 picked this), including tight clustering (19) and shape and symmetry of the clique (7) as most common layout criteria.

General Layout: The second most commonly named criteria related to the overall graph layout or its general aesthetics (51). This was somewhat surprising as this did not relate to our task instructions. Many participants preferred graphs with a clean, organized, or symmetric look (16) while compactness (4), shape (4), and perceived simplicity (4) were also named by a few participants.

Leaves: The nature of leaf nodes was reported as a criterion by 46 participants with the placement of leaf nodes towards the outside (edge) of the graphs being the most commonly reported layout criteria (27).

Cut Node: Visibility of the cut node was reported as a main criterion by 43 participants. Layout criteria were space around cut node (3) or the placement of the cut node in the center of the graph (7).

Chains: The nature of chains was the least commonly reported criteria relating to our task question (38). Layout criteria included: chains in geometric shapes such as lines or circles (8), and node spacing within the chains (6).

Other Criteria: Some common graph drawing aesthetics were also named by participants including node distribution (21), edge crossings (17), graph symmetry (11), edge length (5), and overall “graph shape” (4).

4.2.4 Graph Layout Metrics and Evaluation Results

Interpreting our observed differences in user preferences and also task time and accuracy requires looking more closely at the layout differences amongst automatic layouts (G0-G2), and user generated ones (G3-G11).

Some statistics for each graph, together with the average time and accuracy data, are given in Table 2. It is noteworthy that G0 (the orthogonal layout) had only 7 crossings while the next best layout in terms of crossings (the force-directed layout G2) had almost twice as many (13). Yet, G0’s performance in terms of user choice was dismal. The ten users who chose to complete the tasks using G0 were also slow and inaccurate in finding the clique and leaf-nodes, although the sample size is too small to be conclusive. In general we did not find any strong correlation between crossing count and either user preference or task time and accuracy.

A number of studies have used edge-length variance (or standard deviation) as a measure of drawing quality. However, this measure does not in any way consider the spacing between nodes that are not immediately connected. Statisticians measure the error of a 2-dimensional plot of dissimilarity data in terms of *stress* [8]. In graph drawing, stress measures how close the graph is to a layout that perfectly represents the distances between nodes in the graph:

$$\sum_{i,j \in V} w_{ij}(D_{ij} - d_{ij})^2 \quad (2)$$

where for all pairs of data points i, j we have D_{ij} is the dissimilarity measure, d_{ij} is the actual Euclidean distance in the plot, and w_{ij} is a weighting (e.g., confidence in the measurement). We used stress as a measure of layout quality by taking $D_{ij} = S_{ij}\mu$ where S_{ij} is the shortest path between nodes i and j and μ is the mean edge length; we balanced the stress of shorter paths against longer ones by setting $w_{ij} = D_{ij}^{-2}$. A graph layout with low stress would tend to have nodes that are far from each other on the graph be also far from each other in space; a graph with high stress would have remote nodes drawn near each other. Fig. 12 shows the relationship between stress and the number of users who selected the graph as their first and final choices; users appear to be sensitive to stress in choosing layouts.

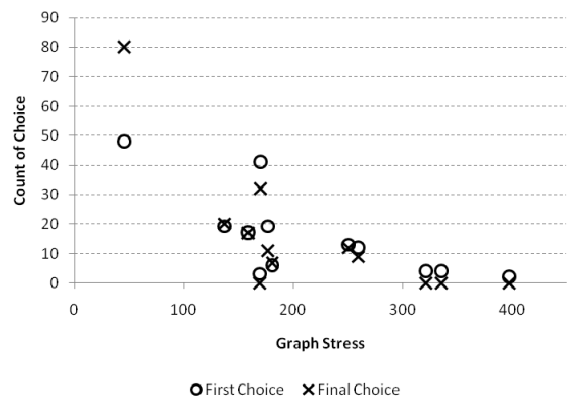


Fig. 12. The relationships between stress and choice. Users seem to prefer low-stress graphs.

5 DISCUSSION

Graph drawing is a skill, and graph-drawing algorithms embody aspects of that skill. G2, a physics-based layout, displays the nodes

on this graph very elegantly, and so is a favorite. None of the users in our first study were able to create graphs as desirable as G2. Users found graph layout to be a difficult task: many gave up on it part-way, while others spent a long time making small adjustments, trying to find a good layout. Several users found that their initial strategies for arranging the graph were so poor that they asked to restart the task.

During the graph-layout process, users adopted a variety of strategies. Some tried to find the highest-degree nodes and bring them to the centre, a strategy that would backfire if they selected a member of the clique. Users that started moving a cut-point fared much better. During their work, several users asked for automated assistance, such as a tool to bring nearby nodes with them.

Non-physics algorithms failed badly on this graph. The circular layout hid important features, including cliques and lines. The orthogonal layout (like the similarly-unpopular G10) buried leaves in its interior. And the most popular graph layouts tended to have easily-visible cut points.

Among the criteria reported by participants for their graph choices, the most common one related to the nature of cliques. Users seemed to have particular trouble with the circular and orthogonal layouts in this regard. In the orthogonal layout there were a large number of right angle bends in edges inside the clique. In the circular layout the nodes in the clique were arranged almost linearly along one side of a larger circle of nodes. The angular resolution between clique edges was very poor.

While many participants fared better with the force-directed arrangement of the clique some reported to have been misled by the placement of a node in the center of the clique, making it hard for edges to be distinguished. By contrast, user-generated layouts such as G3, G7 and G11 show the clique nodes distinctly in a tight, symmetric group. Unfortunately, clique layout may be a difficult task for automatic algorithms, since maximal clique detection is NP-hard. However, heuristics may work well enough in some cases.

Interestingly many participants reported to have chosen graphs based on general aesthetics, most commonly a symmetric, ordered, or clean look. This may be the reason for many people choosing the computer-generated layouts over the hand-drawn ones. We were particularly surprised by the relatively high number of first choices for the circular layout, which based on the task description should not have been a good option. The attraction of symmetric or ordered layouts may have led participants to making more emotional choices (as theorized by Norman [13]) than considering the tasks at hand.

The apparent relation between layout *stress* and participants' first and final preferences supports the choice of stress minimization [8] as a good general automatic layout heuristic. In particular the lowest stress drawings had reasonably well separated leaves, clear cut-nodes and compact clusters. There were, however, higher *stress* layouts that seemed to have a clearer delineation of cut-nodes and chains.

6 FURTHER WORK

The design of our evaluation study meant that we had less task accuracy and time data for the less popular graphs. It might be useful to get more users to complete the tasks using the unpopular graphs to get larger samples for these and see if *stress* is a useful predictor of task performance in addition to preference.

Our user-generated layouts were created using a very basic drag-and-drop interface. In future we would like to use a similar study model to see what layouts users create using semi-supervised layout strategies such as [5]. Semi-supervised layout provides automatic stress minimization while still allowing users a fine degree of control over the layout through a constraint interface. When some of the more labor-intensive aspects of manual layout are automated, users' layout strategies may be different.

In this study, although the graphs tested were nominally "friendship networks", some effort was made to keep the tasks as application agnostic as possible. Yet the balance of different layout

criteria may vary greatly depending on the application domain and on the specific semantics of the network data. Although some experimental studies on graph aesthetics have considered particular applications (e.g., UML [17] and social networks [11]), layout criteria for applications such as biological networks remain largely anecdotal.

ACKNOWLEDGMENTS

The authors wish to thank all the participants in our studies and other members of the Microsoft Research VIBE team for their support.

REFERENCES

- [1] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling: Theory and Applications*. 2nd Ed. Springer 2005.
- [2] G. Di Battista, P. Eades, R. Tamassia, and I. Tollis. *Graph Drawing: Algorithms for the Visualization of Graphs*. Pearson 1998.
- [3] U. Doğrusöz, B. Madden and P. Madden. Circular layout in the Graph Layout Toolkit. *Proc. Graph Drawing (GD '96)*, LNCS 1190, pp. 92–100. Springer 1996.
- [4] P. Dourish. *Where The Action Is: The Foundations of Embodied Interaction*. MIT Press 2001.
- [5] T. Dwyer, K. Marriott and M. Wybrow. *Dunnart: A Constraint-based Network Diagram Authoring Tool*. *Proc. Graph Drawing (GD '08)*, LNCS 5417, pp. 420-431. Springer 2009.
- [6] M. Eiglsperger, S. Fekete and G. W. Klau. Orthogonal Graph Drawing. In *Drawing Graphs: Methods and Models*, Kaufmann and Wagner (Eds.) LNCS 2025, pp. 121-171. Springer 2001.
- [7] R. Fleischer and C. Hirsch. Graph Drawing and Its Applications. In *Drawing Graphs: Methods and Models*, Kaufmann and Wagner (Eds.) LNCS 2025, pp. 1-22. Springer 2001.
- [8] E. Gansner, Y. Koren and S. North. Graph Drawing by Stress Majorization. *Proc. Graph Drawing (GD '04)*, LNCS 3383, pp. 239–250, Springer 2004.
- [9] M. R. Garey and D. S. Johnson. Crossing number is NP-complete. *SIAM Journal on Algebraic and Discrete Methods*, 4:312-316, 1983.
- [10] F. van Ham and B. Rogowitz. Perceptual Organization in User-Generated Graph Layouts. *IEEE Trans. on Visualization and Computer Graphics*, 14(6): 1333-1339, 2008.
- [11] W. Huang, S. Hong and P. Eades. Layout Effects on Sociogram Perception. *Proc. Graph Drawing (GD '05)*, LNCS 3843, pp. 262-273. Springer 2006.
- [12] C. McGrath, J. Blythe and D. Krackhardt. The effect of spatial arrangement on judgements and errors in interpreting graphs. *Social Networks*, 19:223-242. Elsevier 1997.
- [13] D. A. Norman, *Emotional Design: Why We Love (Or Hate) Everyday Things*. Basic Books 2004.
- [14] C. North, T. Dwyer, B. Lee, D. Fisher, P. Isenberg, K. Inkpen Quinn, and G. Robertson. Understanding Multitouch Manipulation for Surface Computing. *Interact 2009*. To appear: Springer 2009.
- [15] H. C. Purchase, D. Carrington, and H. Alder. Empirical Evaluation of Aesthetics-based Graph Layout. *Empirical Software Engineering*, 7(3):233-255, Kluwer 2002.
- [16] H. C. Purchase, E. Hoggan, and C. Görg. How Important is the "Mental Map"? – An Empirical investigation of a Dynamic Graph Layout Algorithm. *Proc. Graph Drawing (GD '06)*, LNCS 4372, pp. 262-273. Springer 2007.
- [17] H. C. Purchase, M. McGill, L. Colpoys and D. Carrington. Graph Drawing Aesthetics and the Comprehension of UML Class Diagrams: An Empirical Study. *Proc. Asia-Pacific Symp. on Information Visualization*, CRPIT 9, pp. 129-137. ACS 2001.
- [18] H. C. Purchase and A. Samra. Extremes Are Better: Investigating Mental Map Preservation in Dynamic Graphs. *Proc. Diagrammatic Representation and Inference (Diagrams '08)*, LNAI 5223, pp. 60-73. Springer 2008.
- [19] P. Saffrey and H. Purchase. The "Mental Map" versus "Static Aesthetic" Compromise in Dynamic Graphs: A User Study. In *Proc. 9th Australasian User Interface Conference (AUIC2008)*, CRPIT 76, pp. 85-93. ACS 2008.