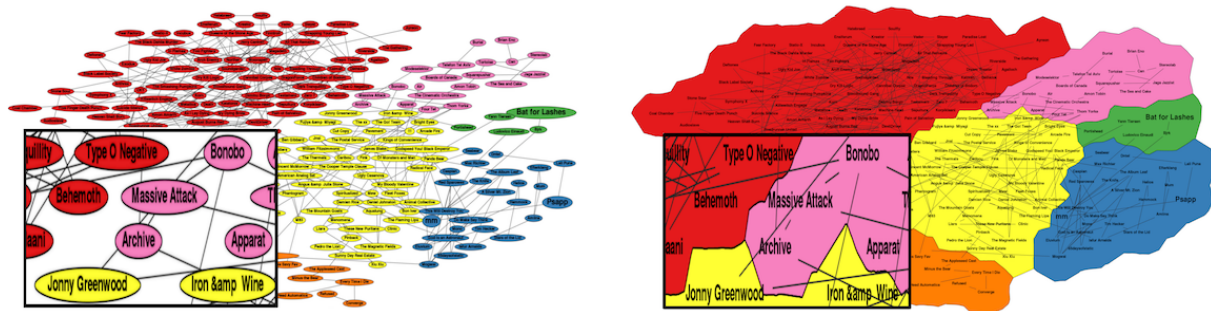


# Map-based Visualizations Increase Recall Accuracy of Data

Bahador Saket,<sup>1</sup> Carlos Scheidegger,<sup>1</sup> Stephen G. Kobourov,<sup>1</sup> and Katy Börner<sup>2</sup>

<sup>1</sup>Department of Computer Science, University of Arizona, Tucson, AZ, USA

<sup>2</sup>Department of Information and Library Science, Indiana University, Bloomington, IN, USA



**Figure 1:** We investigate the memorability of relational data represented with *node-link* (left-side) and *map-based* (right-side) visualizations; shown are a node-link and a map-based visualization with 200 nodes and 500 links from the LastFM dataset.

## Abstract

We investigate the memorability of data represented in two different visualization designs. In contrast to recent studies that examine which types of visual information make visualizations memorable, we examine the effect of different visualizations on time and accuracy of recall of the displayed data, minutes and days after interaction with the visualizations. In particular, we describe the results of an evaluation comparing the memorability of two different visualizations of the same relational data: node-link diagrams and map-based visualization. We find significant differences in the accuracy of the tasks performed, and these differences persist days after the original exposure to the visualizations. Specifically, participants in the study recalled the data better when exposed to map-based visualizations as opposed to node-link diagrams. We discuss the scope of the study and its limitations, possible implications, and future directions.

## 1. Introduction

Researchers have long recognized that the visual display of information can be more effective than tables and numeric summaries [Ans73]. We also know that different visual designs offer significantly different reading precision [CM84]. In contrast, we do not understand nearly as well the *memorability* of the data that underlies the visualization. *Is the design of a visualization responsible for how well users will remember its content?*

In this paper, we present evidence that different visual designs can impact the recall accuracy of the data being visualized. Several recent studies have tested the memorability of different types of visualizations [BMG\*10, BARM\*12, MPWG12, VMTW\*12, IXTO11, BVB\*13]. These seminal

studies focused on which types of visual information are memorable [BVB\*13]. To the best of our knowledge, no study has yet been performed to assess long-term memorability of the underlying data represented in these visualizations. In this paper, we focus on two alternative visualizations for relational data. Specifically, we compare *node-link* visualizations to *map-based* visualizations.

*Node-link* visualizations date back to 1735 and are a standard way of depicting relational datasets. In node-link diagrams, entities are depicted as points (typically dots or circles) in low-dimensional space, and two related entities are connected with a curve (typically a straight-line segment). Cluster membership is typically indicated by filling each circle with a color that is unique for each cluster.

	Visualization	Number of tasks
Phase 1	shown	0
Phase 2	shown	6
Phase 3	not shown	9

**Figure 2:** Our study has three phases. In phases 1 and 2 the participants are involved in reading of the visualizations. In phase 3, they are required to recall the visualization contents.

Map-based visualizations are also old [BCB03, Bör10]. More recently several fully automated tools were developed to generate map-based visualizations for non-spatial data. In map-based diagrams, the visualization is enriched by enclosing in a contiguous region elements that belong to the same set; see Figure 1-Right. This is the output of several recent InfoVis techniques which visualize sets, groups, and clusters: BubbleSets [CPC09], LineSets [ARRC11], GMap [GHK10], KelpFusion [MRS\*13], and MapSets [EHKP14].

In previous work, Saket et al. [SSKB14] investigated the effectiveness of three different relational data visualizations in terms of task accuracy and time, *point clouds*, *node-link diagrams*, and *maps*. We borrow their terminology and refer to these respectively as N (node), NL (node-link), and NLG (node-link-group) diagrams. Here, we report on experiments to measure the extent to which people remember the *data* depicted in such visualizations by having participants in the study perform tasks long after being exposed to visual stimuli (four days between exposure and assessment).

Our study can be summarized in three phases; see Figure 2. In phase 1, subjects examine the visualizations with no preset time limits, and are not asked to answer any tasks. In phase 2, subjects still have access to the visualization, and are now asked to perform a set of six tasks. In phase 3 (which happens after a predetermined period of time depending on whether the subject receives the immediate or long-term treatment), subjects are asked to perform the same six tasks, together with three new additional tasks. As we will discuss in more detail later, we find that *subjects recall data in map diagrams more accurately, but not any faster*; we find, in addition, that *in phase 1, subjects interact longer with map data*.

## 2. Related Work

**Comprehension and Recall Experiments** There have been a number of studies to investigate the effect of embellishments on visualization memorability and comprehension [BMG\*10, BARM\*12, MPWG12, VMTW\*12]. Bateman et al. [BMG\*10] conducted a study to test the comprehension and recall of charts using an embellished version and a plain version. In the recall phase of their experiment, the participants were first asked to recall as many of the charts as possible. The experimenter then asked the participants to describe the charts as completely as they could. If the answer was incomplete, the experimenter went through a series of

increasingly specific prompts until either the participant sufficiently recalled the information or the list of prompts was exhausted. In contrast, we test for *unprompted recall of the data* in the visualizations. Bateman’s study has been somewhat controversial, and Li et al. [LM14] recently reported a replication, limiting their selection to those charts that consisted of data sets with 10 or more observations. They found that the presence of a time limit affected comprehension and short-term recall performance, while the type of chart significantly affected short-term recall. Borgo et al. [BARM\*12] showed that visual embellishment improves information retention in terms of both accuracy of and time required for memory recall. Since their focus was on “visual perception and cognitive speed-focused tasks” that leverage cognitive abilities, they used analytical tasks, where they enforced attention to switch from one task to another. Another study by Vande Moere et al. [VMTW\*12] showed that visual metaphors do not have a significant impact on perception and comprehension.

Ghani and Elmqvist [GE11] further studied the effect of visual landmarking in node-link diagrams and found that landmarking is generally promising for *graph revisitation*, i.e., the “task of remembering where nodes in the graph are and how they can be reached”. Marriott et al. [MPWG12] investigated the cognitive impact of various layout features, such as symmetry and alignment, on the recall of graphs. They asked participants to look at drawings and redraw them. Perceptual characteristics and memorability in dynamic graphs have also been studied [AP12, AP13, FQ11, GEY12]. Our study focuses on data recall; we defer a discussion of its relationship to graph features to Section 6.1. As a part of an experiment measuring the effectiveness of four visualizations (BubbleSets, Node-link, LineSets, GMap) Jianu et al. [JRHT14] asked participants to perform 10 different tasks, including one task related to the memorability of the data. Their results suggest that GMap might be more memorable than the other three visualizations, but the effect disappeared when labels (which were only present in GMap) were removed.

Isola et al. [IPTO11, IXTO11] measured the memorability of annotated natural images and found that “people and human-scale objects in the images contribute positively to memorability”. Borkin et al. [BVB\*13] applied the same methodology to measure the memorability of “real world visualizations” which they extracted from a variety of websites covering different areas of visualization publications. These results indicate that attributes such as color and the inclusion of human-recognizable objects enhance memorability. However, these studies aimed in identifying “which type of visual information is memorable or forgettable”, in contrast with the content of the visualization stimuli, the object of our study.

**Recalling Map-based Visualizations** Geographers have studied knowledge about the world in general, and about foreign countries in particular. Jahoda used verbal interviews to assess the knowledge of Scottish children [Jah62]. Wiegand asked participants to write down the names of all the foreign

countries they knew, and to circle the places they had visited [Wie91]. In a subsequent study, the participants were asked to draw a map of the world [Wie95], which was meant to elicit whether young children can remember the general shape of the map of the world, various countries, and their neighbors. In this study, Wiegand points out that individual differences in drawing skills made the drawing requirement challenging for the study.

**Perception of Visualizations** One of the main theories which is relevant to this study is Norman's level of processing [Nor04]. This theory divides the process of perception into three levels: visceral, behavioral and reflective. The visceral level includes basic perceptual operations of distinguishing objects and makes quick judgments. The behavioral level uses the output of the visceral level and acts on it, but typically as a result of inherent skills built up with practice. "This is the level where the understandability of a stimulus is most important" [BRSG07]. The highest level, reflective thought, reflects on what is happening at the behavioral level, tries to find meaning in it, and attempts to influence it. "This high level is strongly affected by context, including the culture and experience of the perceiver, and the viewing circumstances" [BRSG07].

Several of the visualization studies mentioned earlier focused on the visceral level, and on fast, immediate perceptions [IPTO11,IXTO11,BVB\*13,MPWG12]. Since their aim was not to study the memorability or comprehensibility of the underlying data presented in the stimulus, each stimulus was shown to participants for just a few seconds. In contrast, our focus is on the behavioral level, where our main purpose is to understand the memorability and comprehension of the *underlying data*, rather than to identify features that make a visualization memorable. Thus, in our study we are interested in the effect of different visualizations (NL or NLG diagrams) of the same data, on the immediate and long-term memorability about the data shown in the visualization (and not just the visualization itself).

### 3. Preliminaries

We designed a three-phase experiment to measure differences in recall between *Node-link* (NL) and *Map-based* (NLG) visualizations. In the first phase, participants simply look at the NL and NLG visualizations for as long as they want to, with no direction or predetermined tasks. In the second phase, participants were asked to perform various tasks using the NL or NLG visualizations with different *Size*, *Density* and *Dataset*. In the last phase, the visualization is removed and the participants are asked to perform the same tasks. Three additional questions are asked to measure how well participants remember the overall shape, cluster colors, and cluster-adjacency. Since the visualization is not shown in the third phase, in order to perform the required tasks, participants need to recall the information presented in the visualizations from the

first phases. Participants were placed in one of two recall conditions (immediate recall or long-term recall). For the immediate recall condition, phase three followed phase two after a 2-minute visual diversion. For the long term recall condition, phase three occurred four days after phase two. We chose the length of our long-term recall condition as in previous long-term evaluations, e.g., Li et al. [LM14].

#### 3.1. Size and Density

In a recent evaluation of N, NL, and NLG diagrams, Saket et al. conducted a preliminary study to choose suitable minimum and maximum number of nodes. The average response time for a single task was in the range from 5 to 30 seconds, and found  $N = 50$  nodes as minimum and  $N = 200$  nodes as maximum [SSKB14]. Determining a good range for *Density* is a difficult problem. We use  $L = 1.5N$  links for the sparse setting and  $L = 2.5N$  links for the dense setting. In order to have a reasonable experiment length and complexity we have two settings: the *small visualization* has *Size* 50 and *Density* 1.5 and the *large visualization* has *Size* 200 and *Density* 2.5. We discuss these decisions in greater depth in Section 6.

#### 3.2. Datasets and Clusters

When deciding the number of different datasets to show users, two goals conflict with one another. On the one hand, we would clearly like to limit the chance that the observed effects are due to an unfortunate choice of dataset: this would suggest that the more graphs we have, the better. At the same time, we would like to be able to assess the extent to which features of each particular graphs are responsible for the results: this suggests we want to *control* these features as levels in the design, and increasing the number of datasets would mean an increase in the necessary sample size to achieve any kind of power. Because our experiment includes a long-term condition which requires a repeated visit (precluding popular venues for high-powered studies such as Amazon's Mechanical Turk), for logistical reasons we limited the size of the study to 40 participants. A power analysis (included in the supplemental material) suggests that more than two different datasets would not yield sufficient statistical power: we would not be able to tell the presence of an effect, *even if it existed*.

As a result, our compromise is to use two real-world datasets for our evaluation (we go back to this point in Section 6). The *Book* dataset contains 3,204 popular books. The links are obtained with a breadth-first traversal following Amazon's "Customers Who Bought This Item Also Bought" links [GHK10]. The *LastFM* dataset contains 2,588 popular bands and musicians crawled from the Last.fm online radio station. The links correspond to similarities between musicians as determined by the radio station [GHKV09]. The nodes in the datasets are labeled with familiar words: books, bands and musicians. We selected 50 and 200 nodes from the

Name	# Nodes	Size	# Links	Density	Dataset	# Clusters
Small Visualization of Book Dataset	50	N	75	1.5	Book	3
Large Visualization of LastFM Dataset	200	4N	500	2.5	LastFM	6

**Table 1:** Characteristics of the two datasets used in this study.

*Book* and *LastFM* datasets and a subset of the links between them to match the desired densities. We also have two different settings of 3 and 6 for the number of clusters in the datasets. The number of clusters we use is determined by our preliminary study (see Section 3.3) showing that participants could not identify or remember more than six unique colors.

The graphs are embedded in the plane with a multi-dimensional scaling (MDS) [KW78] algorithm and clustered using modularity clustering [New06], with link weights treated as similarities. For both algorithms we used the implementations provided in GRAPHVIZ [EGK\*01]. To generate instances of NLG diagrams we use GMap. Since the original GMap implementation [GHK10] generates fragmented countries, which can be confusing [JRHT14], we use a new and improved version of GMap, which is guaranteed to generate contiguous regions [KPS14]. From the map-based visualizations, we obtain the node-link visualizations by removing the group regions. Thus the positions of the nodes and links in the two settings (NL and NLG) are identical. We created two visualizations (books and music) for each technique (node-link and map-based); see Table 1.

### 3.3. Cluster Colors

Since the experiment requires colors to be distinguishable and memorable, we ran a preliminary study to verify that the colors we use can be quickly and unambiguously named and distinguished. This is particularly important in our case since the participants are expected to remember the clusters and their relative positions.

We chose colors using ColorBrewer [Bre14], selecting a map-friendly, qualitative color scheme with eight different colors: red, green, yellow, blue, orange, pink, purple, brown. We presented the colors to six participants and asked them to look at the colors for two minutes. We then removed the colors and asked them to write down the names of as many colors as they remembered. We ranked these colors based on the number of times that participants could remember them correctly and their place in the lists of colors that participants could remember. The best remembered and consistently named colors were red, yellow, green, blue, orange and pink; these are the colors we use for the visualizations with six clusters. For the visualizations with three clusters we used red, yellow and green.

### 3.4. Node and Link Colors

The color of the nodes and links is another important factor in our study since participants need to perform several tasks

that assume the readability of the nodes and the links. Clearly, we cannot use any of the colors selected for clusters. Similarly, white is not a good option since links become invisible in the Node-link setting. The standard choice for node and link color is a dark gray or black. We generated four sets of visualizations (each set has two node-link and map-based visualizations) using a dataset consisting of 200 nodes, 500 links, and six clusters. We used four colors on the gray-to-black spectrum to color nodes, starting with black RGB (0, 0, 0) and lightening it by increment of 45 until RGB (135, 135, 135). We then presented the four different sets of visualizations to six participants and asked them to decide which set has the most readable nodes. All participants selected the set RGB (0, 0, 0) set as the most readable one. We then generated another four sets of visualizations. This time the color of nodes was fixed to RGB (0, 0, 0). We now varied the color of the links in the same way. Five of the participants chose the gray color links RGB (90, 90, 90) as the most readable set. Thus, the color of the nodes is RGB (0, 0, 0) and the color of the links is RGB (90, 90, 90).

### 3.5. To draw or not to draw

In another preliminary, informal study, we asked six participants to redraw the shapes of the two visualizations to the best of their ability, using the immediate condition as described above. This was a similar approach to that of Wiegand [Wie91]. Two of the participants were unhappy with their drawings, because of their (self-reported on a subsequent interview) weak drawing skills. Three of the remaining four participants gave up on this task altogether, because they found this part of the study very difficult and frustrating. Thus we decided not to ask the participants to draw, but rather to answer three questions that capture some of the information that we were hoping to capture with the drawing. Tasks 7 through 9 in Table 3 deal with the recall of the shapes of the clusters, the colors used in the visualization, and the cluster-color match.

## 4. Experiment

The experiment had two parts: a visualization reading part (phases 1 and 2), and a recall part (phase 3). To prevent intentional learning, participants were not told about the recall part.

### 4.1. Participants and Setting

In order to maximize the power of the tests with the different factors we wanted to control, we decided on a strat-

<p><b>Node-Based Task</b></p> <p><b>T1.</b> How many nodes are there in this visualization?  <b>Why.</b> The purpose of the task is to count the nodes depicted in the visualization. The targets are nodes in the visualization and the location is entire visualization. Thus, both targets and location are known. Participants need to identify the nodes and count them.  <i>(DISCOVER + LOOK UP + COUNT)</i>  <b>What.</b> The input is the entire visualization and the output is the total number of nodes.  <b>How.</b> The participants need to count number of nodes in the visualization.  <i>(COUNT)</i></p>	<p><b>Node-Based Task</b></p> <p><b>T2.</b> Identify three nodes with labels in large font size.  <b>Why.</b> The purpose of the task is to find a subset of nodes given a specific characteristic. The targets (nodes with large font size) and the location (visualization) are both unknown. The participants compare sizes and select three nodes with large font.  <i>(DISCOVER + EXPLORE + SUMMARIZE)</i>  <b>What.</b> The input is the entire visualization and the output is list of nodes with large font sizes.  <b>How.</b> The participants need to compare the font size of different nodes and identify three particular nodes.  <i>(COMPARE + SELECT)</i></p>	<p><b>Network-Based Task</b></p> <p><b>T3.</b> Identify two nodes connected to node X.  <b>Why.</b> The purpose of this task is to find neighbors of a given node in the visualization. The search target is given (node X) but the location of the target is not. The participants need to find two different neighbors of the given node.  <i>(DISCOVER + LOCATE + SUMMARIZE)</i>  <b>What.</b> The input is a specific node. The output is list of nodes adjacent to the given node.  <b>How.</b> The participants need to find nodes that have links to the given node.  <i>(SELECT)</i></p>
<p><b>Network-Based Task</b></p> <p><b>T4.</b> Find a path that connects the node “X” to the node “Y” and passes through node “Z”.  <b>Why.</b> The purpose of the task is to find paths between two given nodes (e.g., X and Y) and select one that goes through a specified intermediate node (e.g., Z). The targets are given (nodes X, Y, Z) but their location is not given.  <i>(DISCOVER + LOCATE)</i>  <b>What.</b> The input for the task are three X, Y and Z nodes and the output is a path which passes through node Z.  <b>How.</b> The participants need to find paths from X to Y and identify one which passes through Z.  <i>(DERIVE + SELECT)</i></p>	<p><b>Group-Based Task</b></p> <p><b>T5.</b> How many clusters are there in this visualization?  <b>Why.</b> The purpose of the task is to count the number of clusters in the visualization. The search target is known since the participants need to count every cluster in the visualization. The location is the whole visualization. The participants need to identify clusters and count them.  <i>(DISCOVER + COUNT)</i>  <b>What.</b> The input is the entire visualization and the output is a number.  <b>How.</b> The participants need to look at the visualization and count the number of different clusters in the visualization.  <i>(COUNT)</i></p>	<p><b>Group-Based Task</b></p> <p><b>T6.</b> Specify the clusters colors of nodes X, Y, Z.  <b>Why.</b> The purpose of the task is to determine whether three nodes belongs to the same clusters. Participants know the target since they worked with the nodes in earlier tasks. The location might be known (if the participants remember the location), or unknown. After finding the three nodes, the participants need to identify the background color of each of the nodes.  <i>(DISCOVER + LOOKUP/LOCATE + IDENTIFY)</i>  <b>What.</b> The input for the task are three nodes and the output is a number.  <b>How.</b> The participants need to distinguish three nodes and their background colors.  <i>(SELECT)</i></p>

**Table 2:** List of tasks used in the second phase of our evaluation.

ified sample, balanced design, and between-subjects experiment. Purchase [Pur12] encourages balancing participants in between-subjects experiments, so we recruited 40 participants (20 male and 20 female) aged 21-30 years with normal vision (not color blind). We divided the participants into two groups: 20 participants (10 male, 10 female) performed tasks using NL diagrams and the other 20 participants (10 male, 10 female) performed tasks using the NLG diagrams. In each group half of the participants were science and engineering students and the other half were from other majors/background (e.g., music, art). In other words, we stratified both on gender and background. Tasks were presented using a software application on a computer with i7 CPU 860 @ 2.80GHz processor and 24 inch screen with 1600x900 pixel resolution. The participants interacted with the mouse to complete the tasks.

#### 4.2. First Phase

**Procedure** In this phase, participants were simply instructed to look at the NL and NLG visualizations for as long as they wanted.

#### 4.3. Second Phase

**Procedure** We used a between-subjects design: for each technique (NL or NLG diagram), we had two different visu-

alizations (small and large). In this phase, each participant performed 12 tasks: 2 visualizations  $\times$  6 tasks.

Before the controlled experiment, participants were briefed about the purpose of the study, data, and technique used. We also explained the technical definitions (e.g., node, link, group, path). We then asked participants to complete six training tasks as quickly and accurately as possible. Participants were highly encouraged to ask questions during this stage and we did not record time and accuracy in this stage.

Phase two consisted of 12 tasks for a specific technique (NL or NLG diagram). First, the participants were shown two visualizations that they could examine for as long as needed. Then they were asked to perform tasks while the visualization corresponding to each task was provided below the description of the task. The tasks were presented in reduced latin square to counterbalance learning. The participants were able to zoom and pan the visualization on the screen (if needed) and were required to select one of the provided multiple choices. The software recorded time and accuracy for each task.

**Tasks** We selected tasks based on several considerations. First, the task should represent standard problems, commonly encountered when analyzing relational data. Second, the tasks should be present in existing graph/network task taxonomies and utilized in prior user studies. With these two main considerations in mind, we selected six different tasks, grouped into

three categories based on the information required to perform them:

- **Node-Based Tasks:** Tasks in this category can be performed by considering only nodes, so that no other information is required. **For example:** Given node "X", what is its background color?
- **Network-Based Tasks:** Tasks in this category can be performed by considering only nodes and links. **For example:** Find a node with the highest degree.
- **Group-based Tasks:** Tasks in this category can be performed by considering nodes, links, and groups. **For example:** Given a group X, find the group neighbors of group X.

Most of the tasks in the first two categories are listed under "Attribute-Based Tasks" and "Topology-Based Tasks" in the taxonomy of Lee et al. [LPP\*06]. The tasks in the third category are "Group-Based Tasks" in the taxonomy of Saket et al. [SSK14]. Task descriptions, along with a Brehmer and Munzner [BM13] discussion of the why/what/how questions about the selected six representative tasks, T1 to T6, are provided in Table 2.

**Overview of results** In phase 2, we asked the participants to perform tasks in order to familiarize themselves with the data. Phase 3 asked the participants to recall what they saw. Even though our main goal of measuring memorability and long-term recall is in the third phase, we also analyzed data from phases 1 and 2.

Assessing the effect of each visualization technique on phase 2 performance revealed that map-based visualizations on our sample are about 2.5 seconds ( $p = 0.02$ , 95% confidence interval for mean = [0.33, 5.06]) and 7% more accurate ( $p = 0.056$ , 95% confidence interval for mean of [-11.8%, 0.14%]) than node-link visualizations across all tasks. However, this accuracy improvement was not noticeably higher, in agreement with earlier results by Saket et al. [SSKB14]. However, unlike the results in [SSKB14] we did not find significant improvement in accuracy and time for group-based tasks for NLG over NL diagrams. There are two possible explanations. First, in our experiment we did not ask the participants to perform the tasks as fast as possible. Second, the difference between NLG and NL diagrams for group-based tasks seemed to be due to just one group-based task (*Given a group X, find the group neighbours to group X*) [SSKB14]; we do not use this particular task in our study.

We did observe a substantial difference in performance (about 12%) between people with background in science/engineering and those without, but this is not statistically significant. We discuss this in more detail in Section 6.

#### 4.4. Third Phase

**Procedure** After the second phase, participants in each group (NL or NLG) were divided into two subgroups. Each

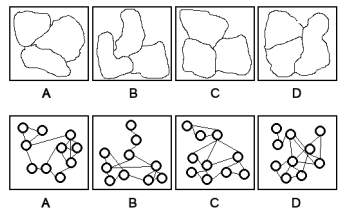
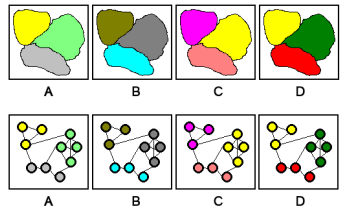
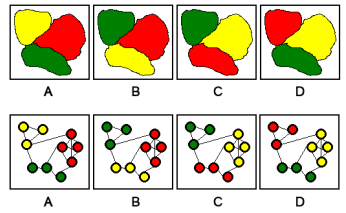
subgroup corresponds to one of two recall conditions (immediate recall or long-term recall), with ten participants in each. Stevanov et al. [SSAK12] advocate using motion illusions for clearing the visual memory of participants. Participants in the immediate recall condition watched five visual illusions for about 2 minutes, followed by phase three of the experiment. Participants in the long-term condition performed the three phase of the study four days later.

At the beginning of the recall part (both for immediate and long-term recall conditions) the participants were verbally reminded of the two visualizations that they worked with (e.g., "Previously you worked with visualizations of datasets about books and music. We would like you to answer a few questions based on these two visualizations."). We then asked the participants to perform nine tasks (T1 to T9). The first six tasks (T1 to T6) were exactly the same tasks from the second phase of the experiment, but this time we removed the visualizations and we expected the participants to perform the tasks using their recollection of the visualizations that they worked with before. The order of tasks T1 through T6 and their answers (multiple choices) was changed from phase two to phase three, in order to prevent participants from extrapolating new judgements from previous ones. Tasks T7 to T9 are new and used to evaluate how well the participants remembered the shapes of the clusters, the colors used in the visualizations, and the cluster neighbors. We showed the map-like visualizations (see Table 3-first row figures) to participants in the NLG group and node-link visualizations (see Table 3-second row figures) to participants in the NL group. We did not reorder T7 to T9 since each task builds on information from the previous ones. Detailed task descriptions, along with a discussion of the why/what/how questions [BM13] about the additional three tasks, T7 to T9, are provided in Table 3.

## 5. Data Analysis

An R script to reproduce the results presented in this section is available as supplemental material, along with the anonymized results dataset. We provide online <http://sites.google.com/site/eurovis2015/> all relevant materials for this study: datasets, software for running the experiment, anonymized results, and detailed statistical analysis.

**Exploratory Data Analysis** The collected dataset has 720 answers (9 tasks, 40 participants, 2 datasets), and six factors: *Gender, Background, Task, Visualization, Size* and *Condition*. Through a power analysis, we expected to be able to find effects of around 15% with group samples of  $n = 20$ . In order to briefly assess feature selection, we performed a logistic regression on all factors but *Task*. Somewhat surprisingly, a Wald test found little evidence for different performance with respect to *Gender* or *Background* (but plenty of evidence for an effect due to *Visualization* and *Condition*, and some evidence due to *Size*). Combined with the power analysis

Group-Based Tasks	Group-Based Tasks	Group-Based Tasks
<p><b>T7.</b> Which of the overall shapes below is same as the overall shapes of any of the visualizations that you worked with before.</p>	<p><b>T8.</b> Which visualization has <i>exactly</i> the same colors as any of the visualizations that you worked with before?</p>	<p><b>T9.</b> Which visualization has clusters whose colors match <i>exactly</i> any of the visualizations that you worked with before?</p>
		
<p><b>Why.</b> The purpose of the task is to discover whether the participants can remember the overall shape of the visualization that they worked with. The participants need to recall the two visualizations that they worked with before and compare them with overall shape of the provided visualizations and select one of the options. (DISCOVER + LOOKUP + COMPARE) <b>What.</b> The input consists of visualizations with different overall shapes and the output is one of the provided options. <b>How.</b> The participants compare the given shapes with a memory of the shape they saw earlier and select one of them. (COMPARE + SELECT)</p>	<p><b>Why.</b> The purpose of the task is to discover whether the participants can remember the colors of the clusters in a specific visualization that they worked with. The participants need to recall the specific visualization they worked with earlier and compare the colors of the clusters in the visualizations provided with those in their memory. (DISCOVER + COMPARE) <b>What.</b> The input consists of the same shapes of clusters colored differently and the output is one of the choices. <b>How.</b> The participants need to compare the colors in the given visualizations with the memory of the colors they saw earlier. (COMPARE + SELECT)</p>	<p><b>Why.</b> The participants need to remember the neighbors of the clusters in the visualizations that they worked with. They need to recall the specific visualization that they worked with before and compare the colors and neighbors of the clusters in the given visualizations with those in their memory. (DISCOVER + COMPARE) <b>What.</b> The input consists of the same shapes of clusters, each colored using the correct set of colors and the output is one of the choices (where the colors match the clusters). <b>How.</b> The participants need to compare cluster neighbors in the given visualizations with the memory of the cluster-color-neighbors they saw earlier. (COMPARE + SELECT)</p>

**Table 3:** List of additional tasks used in the third phase of our evaluation. Participants in the NLG group were shown map-like visualizations (see visualizations in the first row) and participants in the NL group were shown node-link visualizations (see visualizations in the second row)

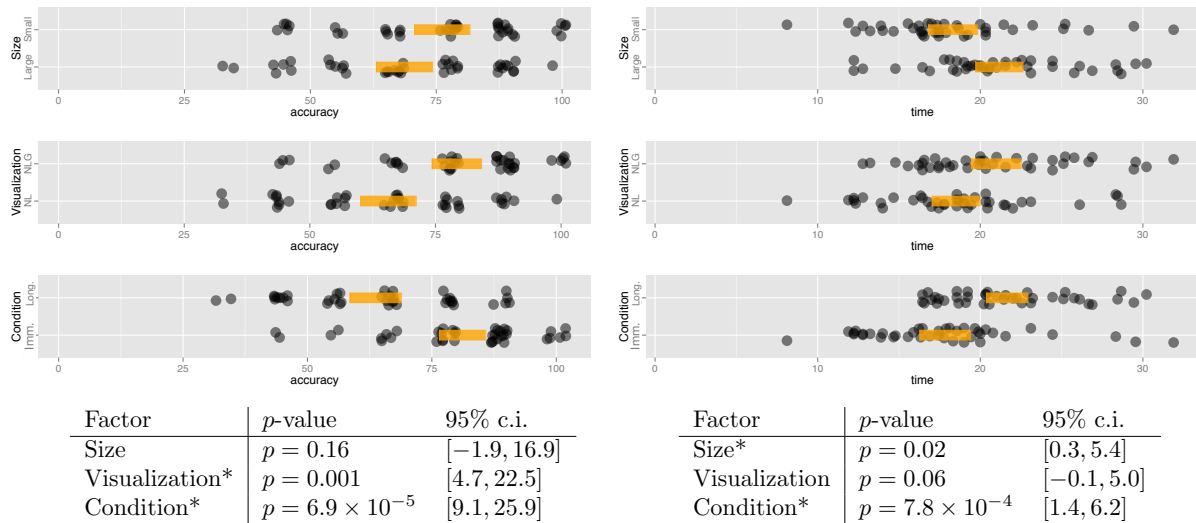
above, we decided that the effect, if it existed, of *Gender* and *Background* was too small to further analyze. Please refer to the supplemental material for details.

**Hypotheses** Based on prior work and our intuition we consider the following hypotheses:

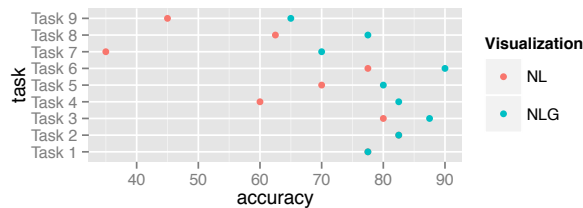
- We expected recall accuracy to be better for participants using NLG diagrams compared to those using NL diagrams, but we did not expect response times to change. Our expectations came from, respectively, the intuition of the designers of GMap [GHK10], and prior work comparing N, NL, and NLG diagrams [SSKB14]. Specifically, the null hypotheses are that (H1a) there will be no difference between recall accuracy from NL to NLG diagrams, and that (H1b) there will be no difference between response times from NL to NLG diagrams.
- In addition, we expected the *effect size* of going from NL to NLG diagrams to be smaller in the immediate recall condition than in the long-term recall condition (i.e., the effect of improved accuracy of NLG diagrams would become stronger, the longer the period between stimulus and recall). Specifically, the null hypothesis is that (H2) the ratio of recall accuracy for NLG and NL diagrams in the immediate recall condition will be the same as that in the long-term recall condition.

- Finally, we expected that recall accuracy for visualization of small graphs would be higher than that of a visualization of large graphs across all settings. Specifically, the null hypothesis is that (H3) the recall accuracy for visualizations of small graphs is the same as that of a visualization of a large graph.

**Summary** We performed one Welch two-sample test (the Welch test is a generalization of *t*-test that does not assume equal sample variances) for each of the *Visualization*, *Condition*, and *Size* factors, and found evidence to reject the null hypothesis for the first two factors. This gives statistical support for hypothesis H1, but for rejecting hypothesis H3. The Bonferroni-corrected *p*-values are, respectively,  $p_V = 0.001$ ,  $p_C = 6.9 \times 10^{-5}$ , and  $p_S = 0.16$ . The Bonferroni-corrected 95% confidence intervals for the *Visualization* and *Condition* factors are [4.63%, 22.59%] and [9.05, 25.94] (loosely speaking, the true difference means lies inside the interval with 95% chance). A Wald test on the interaction between *Visualization* and *Condition* terms in a logistic regression failed to reject the null ( $p = 0.411$ ): this suggests that the loss of accuracy due to the NL diagram might happen *independently* of the loss of accuracy due to the long-term condition. That, together with the rejection of the null of the *Condition* test leads us to reject H2. Figure 3 illustrates the tests performed. At the  $p < 0.05$  level, the statistical evidence we found was:



**Figure 3:** A summary of the data analysis results. The left column shows the results for accuracy, while the right column shows the results for time to completion. All tests are Bonferroni-corrected Welch 2-sample *t*-tests. The plots show a jittered dotplot of the mean accuracy, and the orange bar indicates the range of the 95% confidence interval of the true means. The confidence intervals of the true mean differences (computed from the *t*-tests) are shown in the row below. Factors highlighted by an asterisk indicate statistically significant rejections of the null.



**Figure 4:** Per-task breakdown of outcomes. Note that the gap in performance for Task 7 seems larger than other tasks in the study.

- sufficient to reject H1a, matching our expectations
- insufficient to reject H1b, matching our expectations
- insufficient to reject H2, against our expectations
- insufficient to reject H3, against our expectations

## 6. Discussion

Participants who used NLG diagrams had significantly more accurate recall ( $p = 0.001$ , sample mean difference of 13.6%, 95% confidence interval of  $[9.1, 25.9]$ ; a comparable effect exists when splitting on immediate vs. long-term condition) than those who used NL diagrams. While there is some evidence for a difference between NL and NLG diagrams in the time to perform the tasks, it is not statistically significant:  $p = 0.06$ , with a confidence interval of  $[-0.1, 5.0]$ .

When comparing immediate to long-term recall conditions, as expected, the participants who had to recall the data after two minutes had significantly more accurate recall ( $p = 6.9 \times 10^{-6}$ , confidence interval of  $[9.1\%, 25.9\%]$ , sample mean difference of 17.5%) than those who were tested after four days. With respect to time, we also observed a significant difference in time to complete the recall tasks ( $p = 7.8 \times 10^{-4}$ , confidence interval of  $[1.4s, 6.2s]$ , sample mean difference of 3.8s). We expected the recall decay rate would be different for NL and NLG diagrams, with memories of NL diagrams decaying faster, but found no evidence for this effect. In other words, while it is the case that NLG diagrams had better recall four days later, they were not comparatively better. This suggests that, at least in the case of NL vs NLG diagrams, it might be possible to restrict one's attention to immediate recall experiments (notably opening the possibility for using Amazon's Mechanical Turk).

We analyzed information recall accuracy and time in the small and large settings. In the immediate recall condition, recall of information for the large visualizations is associated with lower accuracy and more time. However, these differences were not statistically significant. It is possible that a higher-powered experiment could find evidence for such an effect. In the long-term condition, recall of information for the large NLG diagrams is as good as for small NLG diagrams. However, large NL diagrams are associated with lower accuracy and greater time than small NL diagrams, but not significantly so after a Bonferroni correction. Still, the difference is intriguing, and so we decided to perform



some exploratory checking by breaking down the accuracy of individual tasks, and present the results in Figure 4.

While we performed no formal hypothesis tests, visual inspection suggests that tasks T7 to T9 appear associated with a decrease in performance for large NL diagrams compared to small NL diagrams, while the change between small and large NLG diagrams is negligible. Once again, the explicit presence of boundaries and the creation of clearly identifiable outlines for the graphs likely plays an important role.

There also appears to be an effect of the background of the participants on the results, although the effect is not statistically significant. Participants with science and engineering background performed NL diagram tasks with about 12% more accuracy than those with backgrounds other than science and engineering. We found no such difference between the two groups performing NLG diagram tasks.

Finally, Figure 4 suggests a difference in effect sizes between tasks 1–6 and tasks 7–9. This could imply a situation where the significance of our results stems entirely from this smaller set of tasks with larger performance differences. To check this scenario, we ran additional hypotheses tests for this specific task grouping (tasks 1–6 and 7–9 separately). Although the effect size for the first six tasks is in fact reduced to just under 10%, it remains significant at  $p < 0.05$ , even after multiple-comparisons corrections. We leave a more comprehensive study of performance vs. specific task for future work.

### 6.1. Limitations and threats to validity

In our experiment we attempted to control several variables that typically impact such studies. In particular, for a given dataset, we fixed the location of the nodes and links in the node-link and map-based visualizations. We used the same font size and the same colors to indicate groups in all visualizations. We ran preliminary experiments to determine colors for the nodes, links, and groups, as well as to determine how evaluate memorability (e.g., asking the subjects to draw was too difficult a task, which we replaced by three tasks aiming to capture what people remembered: shapes, colors, neighbors). Nevertheless, we want to point out limitations and possible threats to the validity of our study.

First, we use a between subjects experiment design. Since we are expecting the participants to work with two datasets in depth so that they can remember the underlying data four days later, a within-subjects design (where subjects see both NL and NLG diagrams) could have made an already difficult task too difficult to obtain meaningful results.

Second, T7 to T9 appear to have significantly worse accuracy when comparing NL to NLG diagrams. It is possible that, even though Figure 4 suggests that NLG plots perform better for both immediate and long-term recall conditions, the effect would have been too small to be significant in this

study. We need to perform a study with more power in order to say anything significant about these apparent differences.

Finally, we only consider four fixed examples across all our subjects (one small node-link diagram, one large node-link diagram, one small map, one large map). Due to the long-term condition, services such as Mechanical Turk could not be used since we cannot find the same participants in a multi-phase study, where the phases are days apart. Having a small number of graphs allowed us to directly measure the effect of size on the recall. Even though we could not reject the null hypothesis, the power analysis we performed suggests that if effects due to size existed, they are substantially smaller than those due to visualization type or condition. We chose a study of long-term recall knowing that the possibility remains that the particular datasets could have confounded our results. With the lack of evidence for interaction between Condition and Visualization factors, we feel more confident in a future high-powered study limited to the immediate recall condition.

## 7. Conclusion and Future Work

The main result of our study is that subjects recall data shown with NLG diagrams more accurately than data shown with NL diagrams. The per-task breakdown of Figure 4 shows a more consistent pattern of decreased accuracy from immediate to long-term recall conditions which suggests that the recall rate decay is *also* independent of the task performed. This would again suggest the applicability of a high-powered study through crowdsourcing as future work.

In the beginning of the first phase, the participants were allowed to work and explore the visualizations for as long as they needed. Our data indicates that the average time that participants spent to explore the NLG diagrams was about 20 seconds more than the NL diagrams, roughly a 25% increase. We noticed this difference and asked several of the participants about it. Some of the responses included “This is beautiful”, “I like it”, and “How did you draw this map?”. This could be an indication that NLG visualizations might be more memorable because they more effectively engage the viewers. We plan to study engagement in future work.

## References

- [Ans73] ANSCOMBE F. J.: Graphs in statistical analysis. *The American Statistician* 27, 1 (1973), 17–21. 1
- [AP12] ARCHAMBAULT D., PURCHASE H. C.: The mental map and memorability in dynamic graphs. In *Pacific Visualization Symposium (PacificVis)* (2012), pp. 89–96. 2
- [AP13] ARCHAMBAULT D., PURCHASE H. C.: Mental map preservation helps user orientation in dynamic graphs. In *Graph Drawing* (2013), pp. 475–486. 2
- [ARRC11] ALPER B., RICHE N. H., RAMOS G., CZERWINSKI M.: Design Study of Linesets, a Novel Set Visualization Technique. In *IEEE Trans. Visualization and Computer Graphics (TVCG)* (2011), pp. 2259–2267. 2

- [BARM\*12] BORGIO R., ADUL-RAHMAN A., MOHAMED F., GRANT W. P., REPPA I., FLORIDI L., CHEN M.: An empirical study on using visual embellishments in visualization. In *IEEE Transactions on Visualization and Computer Graphics (InfoVis '12)* (2012). 1, 2
- [BCB03] BÖRNER K., CHEN C., BOYACK K. W.: Visualizing Knowledge Domains. *Annual review of information science and technology* 37, 1 (2003), 179–255. 2
- [BM13] BREHMER M., MUNZNER T.: A Multi-level Typology of Abstract Visualization Tasks. In *Symp. Information Visualization (InfoVis '13)* (2013), pp. 2376–2385. 6
- [BMG\*10] BATEMAN S., MANDRYK R. L., GUTWIN C., GENEST A., MCDINE D., BROOKS C.: Useful junk? the effects of visual embellishment on comprehension and memorability of charts. In *CHI '10* (2010). 1, 2
- [Bör10] BÖRNER K.: *Atlas of science*. MIT Press, 2010. 2
- [Bre14] BREWER C.: ColorBrewer, <http://www.colorbrewer.org>, 2014. 4
- [BRSG07] BENNETT C., RYALL J., SPALTEHOLZ L., GOOCH A.: The aesthetics of graph visualization. In *3rd Eurographics Conf. on Computational Aesthetics in Graphics, Visualization and Imaging* (2007), Computational Aesthetics'07, pp. 57–64. 3
- [BVB\*13] BORKIN M. A., VO A. A., BYLINSKII Z., ISOLA P., SUNKAVALLI S., OLIVA A., PFISTER H.: What makes a visualization memorable? *IEEE Transactions on Visualization and Computer Graphics* 19, 12 (2013), 2306–2315. 1, 2, 3
- [CM84] CLEVELAND W. S., MCGILL R.: Graphical perception: Theory, experimentation, and application to the development of graphical methods. *J. American Statistical Association* 79, 387 (1984), 531–554. 1
- [CPC09] COLLINS C., PENN G., CARPENDALE S.: Bubble Sets: Revealing Set Relations with Isocontours Over Existing Visualizations. In *IEEE Trans. Visualization and Computer Graphics (TVCG)* (2009), pp. 1009–1016. 2
- [EGK\*01] ELLSON J., GANSNER E. R., KOUTSOFIOS E., NORTH S. C., WOODHULL G.: Graphviz - Open Source Graph Drawing Tools. In *International Symposium on Graph Drawing (GD'01)* (2001), pp. 483–484. 4
- [EHKP14] EFRAT A., HU Y., KOBOUROV S., PUPYREV S.: Mapsets: Visualizing embedded and clustered graphs. In *22nd Symposium on Graph Drawing (GD)* (2014), pp. 450–461. 2
- [FQ11] FARRUGIA M., QUIGLEY A.: Effective temporal graph layout: a comparative study of animation versus static display methods. *Information Visualization* 10, 1 (2011), 47–64. 2
- [GE11] GHANI S., ELMQVIST N.: Improving Revisitation in Graphs Through Static Spatial Features. In *Graphic Interface (GI '11)* (2011), pp. 737–743. 2
- [GEY12] GHANI S., ELMQVIST N., YI J. S.: Perception of Animated Node-link diagrams for dynamic graphs. *Computer Graphics Forum* 31, 1 (2012), 1205–1214. 2
- [GHK10] GANSNER E. R., HU Y., KOBOUROV S. G.: Visualizing Graphs and Clusters as Maps. In *IEEE Computer Graphics and Applications* (2010), pp. 2259–2267. 2, 3, 4, 7
- [GHKV09] GANSNER E., HU Y., KOBOUROV S., VOLINSKY C.: Putting recommendations on the map: Visualizing clusters and relations. In *Proceedings of the Third ACM Conference on Recommender Systems* (2009), RecSys '09, pp. 345–348. 3
- [IPTO11] ISOLA P., PARIKH D., TORRALBA A., OLIVA A.: Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems* (2011). 2, 3
- [IXTO11] ISOLA P., XIAO J., TORRALBA A., OLIVA A.: What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2011), pp. 145–152. 1, 2, 3
- [Jah62] JAHODA G.: Development of scottish children's ideas and attitudes about other countries. *The Journal of Social Psychology* 58, 1 (1962), 91–108. 2
- [JRHT14] JIANU R., RUSU A., HU Y., TAGGART D.: How to Display Group Information on Node–Link Diagrams: an Evaluation. *IEEE Trans. Visualization and Computer Graphics (TVCG)* 20 (2014), 1530–1541. 2, 4
- [KPS14] KOBOUROV S. G., PUPYREV S., SIMONETTO P.: Visualizing graphs as maps with contiguous regions. In *EuroVis14, Accepted to appear* (2014). 4
- [KW78] KRUSKAL J. B., WISH M.: *Multidimensional Scaling*. Sage Press, 1978. 4
- [LM14] LI H., MOACDIEH N.: Is "chart junk" useful? An extended examination of visual embellishment. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 58, 1 (2014), 1516–1520. 2, 3
- [LPP\*06] LEE B., PLAISANT C., PARR C., FEKETE J.-D., HENRY N.: Task Taxonomy for Graph Visualization. In *Workshop on Beyond Time and Errors: Novel Evaluation Methods For information Visualization (BELIV)* (2006), pp. 81–85. 6
- [MPWG12] MARRIOTT K., PURCHASE H., WYBROW M., GONCU C.: Memorability of visual features in network diagrams. *Visualization and Computer Graphics, IEEE Transactions on* 18, 12 (Dec 2012), 2477–2485. 1, 2, 3
- [MRS\*13] MEULEMANS W., RICHE N. H., SPECKMANN B., ALPER B., DWYER T.: KelpFusion: A Hybrid Set Visualization Technique. In *IEEE Trans. Visualization and Computer Graphics (TVCG)* (2013), pp. 1846–1858. 2
- [New06] NEWMAN M. E. J.: Modularity and Community Structure in Networks. *Proc. Natl. Acad. Sci. USA* 103 (2006), 8577–8582. 4
- [Nor04] NORMAN D.: *Emotional Design: why we love (or hate) everyday things*. Basic books, 2004. 3
- [Pur12] PURCHASE H. C.: *Experimental Human-Computer Interaction*. Cambridge Press, 2012. 5
- [SSAK12] STEVANOV J., SPEHAR B., ASHIDA H., KITAOKA A.: Anomalous motion illusion contributes to visual preference. *Frontiers in Psychology* 3, 528 (2012). 6
- [SSK14] SAKET B., SIMONETTO P., KOBOUROV S. G.: Group-level graph visualization taxonomy. In *EuroVis14, Accepted to appear* (2014). 6
- [SSKB14] SAKET B., SIMONETTO P., KOBOUROV S., BORNER K.: Node, node-link, and node-link-group diagrams: An evaluation. *Visualization and Computer Graphics, IEEE Transactions on* 20, 12 (Dec 2014), 2231–2240. 2, 3, 6, 7
- [VMTW\*12] VANDE MOERE A., TOMITSCH M., WIMMER C., CHRISTOPH B., GRECHENIG T.: Evaluating the effect of style in information visualization. *Visualization and Computer Graphics, IEEE Transactions on* 18, 12 (Dec 2012), 2739–2748. 1, 2
- [Wie91] WIEGAND P.: The 'known world' of primary school children. *Geography* 76, 2 (1991), pp. 143–150. 3, 4
- [Wie95] WIEGAND P.: Young children's freehand sketch maps of the world. *International Research in Geographical and Environmental Education* 4, 1 (1995), 19–28. 3