

Article

Design and evaluation of line symbolizations for origin–destination flow maps

Information Visualization
2017, Vol. 16(4) 309–331
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1473871616681375
journals.sagepub.com/home/ivi


Caglar Koylu¹ and Diansheng Guo²

Abstract

We present the results of a user study comparing variants of commonly used line symbolizations for directed origin–destination flow maps. Our design and evaluation consisted of five line symbolizations that employ a combination of following visual variables: arrowheads, origin–destination coloring (color hue, and value), line shortening, line width, tapered edges (varying width from wide to narrow, and narrow to wide), and curvature asymmetry and strength. To guide our evaluation, we used a task-by-type typology and chose four representative tasks that are commonly used in flow map reading: identifying dominant direction of flows, flows with the highest magnitude (volume), spatial focusing of long flows toward a destination, and clusters of high net-exports (net-outflow). We systematically analyzed user responses and task performance which we measured by task completion time and accuracy. We designed a web-based flow mapping and testing framework and recruited the participants from Amazon Mechanical Turk. To demonstrate the application and user experiment, we used 16 commodity flow data sets in the United States from 2007 and systematically rotated the layouts to evaluate the effect of layout orientation. From this study, we can conclude that there is potential usefulness for all of the five symbolizations we tested; however, the influence of the design on performance and perception depends on the type of the task. Also, we found that data and layout orientation have significant effects on performance and perception of patterns in flow maps which we attribute to the change in visual saliency of node and flow patterns in relation to the way users scan the map. We recommend that the choice of line symbolization should be guided by a task taxonomy which end users are expected to perform. We discuss various design trade-offs and recommendations and potential future work for designing and evaluating line symbolizations for flow mapping.

Keywords

Flow mapping, user study, evaluation, perception, line symbolization, cartography

Introduction

Both physical and intangible phenomena such as people, commodities, and information constantly move in the geographic space and create location-to-location networks (graphs) that are often referred to as spatial interactions. In a location-to-location network, each node represents a geographic location (or area), and a link (edge) represents an interaction between a pair of locations. For example, domestic freight shipments within the United States form a network of state-to-state commodity flows in which there are 50 nodes

(states) and thousands of links (commodity imports/exports between states).

Flow maps are commonly used to visualize spatial interactions and facilitate the understanding of patterns of spatial flows and the corresponding spatial

¹The University of Iowa, Iowa City, IA, USA

²University of South Carolina, Columbia, SC, USA

Corresponding author:

Caglar Koylu, The University of Iowa, Iowa City, IA 52242, USA.
Email: caglar-koylu@uiowa.edu

context. Compared to other cartographic approaches, flow mapping is unique, in that each data item (i.e. flow) involves two locations and connections across space. A flow map presents flows with straight or curved lines connecting origin and destination locations.^{1–3} The nodes (origins and destinations) and lines (flows) can vary in symbol, color, and size to represent attribute information pertaining to nodes (e.g. population and degree) and flows (e.g. flow volume).

Although an empirical evaluation of flow mapping has not been explicitly addressed in cartography, studies in the graph visualization community^{4–12} evaluated readability and aesthetics of graph layouts and suggested various principles such as minimizing edge crossings, maintaining large crossing angles, and obtaining symmetrical layouts. The heuristics learned for general graph drawing can only give a very limited guidance for flow map design. Flow maps are dramatically different from general graph drawing approaches because flow map layout is constrained by the geographic coordinates of nodes, while in non-spatial graph drawings, nodes can be moved freely to enhance visual clarity.

In addition to deriving heuristics for aesthetics and edge geometry in graph drawing, alternative line symbolizations (edge representations) that use arrowheads,¹³ biased curvature (Fekete, 2003),²⁶ tapered flow lines,¹⁴ gradual change of color hue, value and transparency,^{15,16} and shortened lines¹⁷ have been introduced and evaluated using tasks such as identifying shortest paths (Huang and Huang, 2011),⁶³ longest link,¹⁸ node connectivity,¹³ common neighbors,¹⁹ and number of edges connected to a node.²⁰ The findings of these studies are useful for flow map design but provide limited information for answering geographic questions, and thus, the tasks that flow maps are used for the following:

- Where are the major flows?
- Which direction do the majority of flows go to?
- Where are the clusters of import or export?
- Where are the flows most local or dispersed?
- Where are the clusters of uneven flow where total export is greater than total import?

The goal of this article is to evaluate the user perception and performance of line symbolizations for directed origin–destination flow maps. We design and evaluate five symbolizations based on quadratic Bézier curves as variants of existing techniques found to be effective in graph drawing and cartography literature. We use the following visual variables in our designs: arrowheads, origin–destination coloring (color hue and value), line shortening, line width, tapered edges (varying width from wide to narrow and narrow to

wide), and curvature asymmetry and strength. Using a task-by-type typology, we employ four representative tasks that are commonly used in the literature of spatial interaction analysis: identifying dominant direction of flows, flows with the highest magnitude (volume), spatial focusing of long flows toward a destination, and clusters of high net-exports (net-outflow). We systematically analyzed user responses and task performance which we measured by task completion time and accuracy. We designed a web-based flow mapping and testing framework and recruited participants from Amazon Mechanical Turk (AMT) which is an online crowdsourcing platform. To demonstrate the application and user experiment, we used 16 commodity flow data sets in the United States from 2007. We systematically rotated the layouts to evaluate the effect of layout orientation and account for learning effects.

We organized the remainder of the article as follows. First, we describe the related work on the design and evaluation of line symbolizations in flow mapping and graph visualization. We then introduce our task taxonomy and flow line symbolizations. Next, we introduce the details of our experiment including our hypotheses, experiment design, tasks, data, procedure, and participants. We then report the results of our evaluation and discuss our recommendations and limitations. Finally, we conclude by a summary of our evaluation and future work.

Related work

In this section, we provide a review of the studies on the design and evaluation of alternative flow line symbolizations and discuss the trade-offs when choosing visual variables for depicting direction, magnitude, length, and clustering. Our study and review focus on static line symbolizations; however, one can find a thorough discussion of animated representations in Ware and Bobrow²¹ and Holten et al.¹³

Slocum et al.²² identify five kinds of flow maps: distributive, network, radial, continuous, and telecommunications flow maps. Tobler's³ origin–destination flow maps of state-to-state migration are examples of network flow maps that include abstract links between origins and destinations. Depicting one-way flows with a measure such as net flow or total flow is commonly used to create flow maps. In this article, we focus on directed two-way origin–destination flow maps with abstract connections.

Directed two-way flow lines between a pair of locations are often drawn by clockwise (i.e. left-hand traffic rule)²³ and counter-clockwise (i.e. right-hand traffic rule)³ line orientation. The right-hand traffic rule draws a flow line on the right side, and the left-hand

traffic rule draws on the left side of an imaginary straight line from an origin to a destination. Both straight and curved lines have been commonly used to depict two-way directions among a pair of locations.^{1–3}

Tobler³ generated one of the first origin–destination flow maps, in which each flow is depicted as a straight line with an arrowhead at the destination end to indicate direction. Ware et al.²⁴ suggested that some form of asymmetry is needed to encode direction along the edge of a line. Arrowheads provide asymmetry by provoking stronger response at the head than the tail, but they are found to perform poorly on directional tasks in graph reading.^{14,24}

Quadratic Bézier curves provide curvature asymmetry (bias) by depicting lines as curvy at the origin and straight at the destination end.^{25,26} Purchase et al.¹⁹ discuss that curved lines would lead to improved interpretation of the connections since curvature produces wider angles between edges and the connections become more visible.¹⁹ In contrast, Ware²⁷ argues that visual processing of line curvature is weaker than factors such as color, orientation, and size. Holten et al.¹⁴ and Netzel et al.¹⁸ evaluated several design alternatives to visualize edges in node–link diagrams and found that tapered edges outperform curvature-based representations. Tapered edges employ varying line width from narrow to wide or from wide to narrow to indicate direction along a flow line. Xu et al.²⁰ studied the impact of edge curvature on graph readability and found that uniform (symmetrical) edge curvature had a detrimental impact on graph readability as increased curvature results in more visual clutter, and this negative effect increased with curvature level. Xu et al.²⁰ also found that users prefer straight over curved lines despite the contrasting findings by Bar and Neta.²⁸

Alongside with tapered edges and asymmetrical curved lines, gradual change of color value (i.e. from dark to light or light to dark) and color hue (i.e. between two divergent colors such as red and blue) along the length of a flow line have been commonly used as a redundant variable to indicate directionality.^{15,16} However, little is known on users' performance and perception of direction using color value and hue as a visual variable along the length of a flow line.

Line width is commonly used to depict magnitude on flow maps and weighted node–link diagrams. In addition to line width, color value, hue, and transparency have commonly been adopted as redundant variables to represent magnitude of flows. Gill²⁹ demonstrated that the most significant contribution to the prominence or visual weight of line symbols is made by line weight (or width); however, both line width and color value have effectively been used together to express flow magnitude.

In addition to direction and magnitude, flow length has been symbolized by visual variables, especially color hue and value. To aid the perception of edge lengths for performing length-related tasks, Holten et al.¹⁴ used color hue to classify the edges by length. Holten et al.'s¹⁴ evaluation concluded that users have increased difficulty in determining edge lengths when edges are longer, and the effect is amplified when the graph is denser. Alternatively, depth sorting, which sorts lines by their length so that longer lines are displayed on top, has been found effective in identifying the longest edges.¹⁸

In order to enhance the perception of clusters in graphs, a variety of node–link group diagrams have been introduced. Node–link group diagrams employ a variety of techniques such as node coloring,^{30,31} map-like areal representations,³² isocontours,³³ and connecting set of lines³⁴ to display group or cluster information overlaid on node–link diagrams.

A flow map can easily become cluttered when it depicts a large number of flows. To overcome this problem, interaction operators, such as linking,³⁵ brushing,^{36,37} filtering and zooming,³⁸ and computational methods, such as edge bundling, edge ordering, minimizing overlap with arrows, adjusting vertex positioning to optimize angular resolution and edge crossings,^{39–41} graph partitioning and regionalization,²⁵ and flow smoothing and clustering,⁴² have been successfully employed. Computational methods are often necessary to first reduce visual cluttering and then generate visually enhanced flow maps with decreased number of flows and regions (or nodes). In our experiment, data sets we use do not require further simplification as they do not have a severe cluttering problem. However, our findings can also provide design guidelines for the data reduction and flow map generalization techniques to reduce cluttering.

Aside from the interaction techniques and computational methods, alternative edge representations have been used to reduce visual cluttering. Becker et al.⁴³ used half-lines from an origin to a destination with a straight line in which only the first half of the line is drawn. This strategy reduces visual cluttering by reducing the number of edge crossings. However, it becomes difficult to distinguish origin–destination pairs. Borrowing from the Gestalt principle of closure, Rusu et al.⁴⁴ employed line shortening to reduce cluttering by drawing only the start and end segments of flow lines. Burch et al.¹⁷ further evaluated user performance and responses of the line shortening method and found that shortened lines help decrease response times; however, they increase error rates because of directional ambiguity.

Task selection

Flow map reading involves visual judgment and cognition of the properties of flows such as magnitude, orientation, direction, and distribution of connections.²⁷ Given the large number of possible flow map reading tasks, it is challenging to select tasks for the evaluation of flow maps. Andrienko et al.⁴⁵ introduced a task taxonomy for the analysis of temporal-spatial interaction data based on the search level (the number of map elements under consideration). While elementary tasks such as lookup, comparison, and relation seeking refer to searching for a characteristic of only one feature such as a flow or a node, synoptic tasks focus on several-to-all map features. Synoptic tasks refer to Bertin's⁴⁶ intermediate and overall levels and allow the exploration of geographic and holistic patterns which is the main focus of our analysis of tasks.

In order to guide the selection of appropriate visual tasks, we employ a task-by-type taxonomy^{47,48} which includes two dimensions which visual tasks vary. (1) Bertin's⁴⁶ three levels for map reading: individual, group (intermediate), and network (overall). The individual level refers to an elementary task to search for the characteristics of a single element (e.g. a flow or a node). While group level describes search tasks for a group of elements (e.g. a group of flows), and the network level describes search tasks that include all of the elements. (2) The second dimension of the taxonomy is type-centric operands that categorize spatial interaction data (i.e. flows, origins, and destinations) by the characteristics of flows and nodes. Our task-by-type taxonomy (Table 1) consists of four visual tasks. The first two tasks require a search task on a single attribute (direction, or magnitude) and commonly used in throughout the literature in spatial interaction analysis.

On the other hand, Task 3 and Task 4 require a search task using a combination of two attributes. We provide the use cases of the four tasks below:

Task 1 (dominant flow direction). The first task is to compare the direction of all flows and select the most common direction summarizing all flows.

Task 2 (flows with the highest magnitude). The second task requires participants to compare the magnitude of all flows and identify the top 3 flows with the highest magnitude.

Task 3 (spatial focusing of long flows toward a destination): The third task requires participants to compare and identify the top 10 flows in length, compare their directions, and identify the location that receives the largest number of these flows. The emphasis of this task is the convergence of long flows toward a destination and typically analyzed in migration research for identifying channelized or spatially focused flows.^{49,50}

Task 4 (cluster of net-exporters). The fourth task requires participants to compare the similarity of nodes in terms of their net-flows represented by color (blue) and identify the nodes with the high-net-outflows that are near one another based on their spatial proximity. Therefore, participants do not need to compare flows to complete the fourth task. The emphasis of this task is to identify a spatial cluster of net-exporters (or locations) with uneven flows where export is greater than import.⁵¹ McGrath et al.⁵² found that both the structural pattern of edges and the spatial arrangement of nodes affect users' perception of groups in graphs. As different line symbolizations could impact the perception of the pattern of edges between nodes, our purpose is to assess the influence of alternative line symbolizations on the perception of node clusters.

Table 1. Flow map tasks.

Task	Element	Search level	Dist.	Dir.	Mag.	Clust.
T1: Select the dominant direction that most flows are going to	Flow	Network		X		
T2: Select three flows that have the highest volume	Flow	Group			X	
T3: Select the circle that receives the highest number of flows in the top 10 in length	Node, flow	Individual, group	X	X		
T4: Blue circles illustrate net-exporters which have greater total volume of exports than imports. Select a cluster of blue circles that are near one another	Node	Group	X			X

Dist.: distance; Dir.: direction; Mag.: magnitude; Clust.: clustering.

Table 2. Flow line symbolizations used in the evaluation.

Design	Name	Direction	Magnitude
	Monotone Arrowhead (MA)	Biased curvature, arrowhead, counter-clockwise orientation (right-hand traffic rule)	Line width, color value
	Divergent Arrowhead (DA)	Biased curvature, arrowhead, counter-clockwise orientation, varying thickness, gradual change of color hue from blue, to gray mid-break, to red	Line width
	Fading Arrowhead (FA)	Biased curvature, arrowhead, counter-clockwise orientation, gradual change of color hue and transparency from blue to, transparent white mid-break, to red	Line width
	Tapered (TA)	Biased curvature, counter-clockwise orientation, varying line width from wide to narrow	Line width, color value
	Teardrop (TD)	Biased curvature, counter-clockwise orientation, varying line width from narrow to wide, gradual change of color value	Line width, color value

Flow line symbolization

In this section, we introduce our five design alternatives compared in our user study (Table 2). Our first design is Monotone Arrowhead (MA) with high curvature at the start and low curvature at the end (Fekete, 2003)²⁶ and a partial arrowhead to reduce the cluttering caused by the full arrow. Edge representations with arrowheads have been found to hinder observation of flow direction as a result of increased visual cluttering and lead to poor performance in unweighted and directed node-link diagrams.¹⁴

Our second and the third designs use the same Bézier curve with partial arrowhead and a variation of gradual change of color hue and transparency^{15,16} to indicate directionality. The second design is Divergent Arrowhead (DA) which depicts the flows with a gradual change of color hue from origin to destination using a divergent color scheme (blue at the origin, light gray at the middle break, and red at the destination end).

The third design is Fading Arrowhead (FA), which depicts the flows only at their start and ending points with the same divergent color scheme used by DA; however, FA employs line-shortening⁴³ by setting the middle of the flow line as transparent. The length of the transparency is set in proportion to the length of the flow, and line shortening is only applied to flow lines that are larger than 200 pixels. As a result of selective shortening by distance, flow lines drawn by FA and DA differ significantly for long-distance flows, whereas they tend to be similar or the same for shorter flows.

The fourth design is Tapered (TA) flow line, which is an altered version of Holten et al.'s¹⁴ tapered edge design which performed the best among many other edge representations for directed and unweighted node-link diagrams. Different from the original tapered design, we employed counter-clockwise biased curvature (from high to low curvature toward the destination) to be able to depict bi-directional flows and proportional line width and color value to depict magnitude. While its curvature is identical to that of MA, DA, and FA, TA's line width is variable and gradually decreases to become non-existent where it touches the destination end.

We adopted Teardrop (TD) as the fifth design from Ware et al.²⁴ which demonstrated the effectiveness of the TD design for depicting the direction of a vector field such as wind direction and ocean current. In contrast to the tapered edge used by Holten et al.,¹⁴ Ware et al.'s²⁴ TD employs a biased curvature from low to high curvature toward the destination; varying width and an additional visual variable of color value along the length of an edge from narrow to wide to depict directionality. As compared to other designs which employ the same curvature, TD has a reduced curvature to ease the perception of direction toward the destination end.

There are two alternatives for node symbolization in origin–destination flow maps: flow lines start and end inside an area or at point symbols. We use point symbols as the start and end point of flows and employ proportional circles to illustrate total flow or net flow depending on the visual task. For tasks that do not require assessing net flow per location, we use total flow (in and out) to determine the size of the point

symbol which improves the design of the layout as it provides a larger area (circle) for connecting to locations with a large number of flows. For the clustering task, we employed divergent node coloring to distinguish net-exporters and net-importers and used size proportional to the net flow.

Experiment

This section describes our research questions, hypotheses, tasks, data sets, procedure, and further details of our user study. We first introduce our general research questions and specific hypotheses based on previous research in cartography and graph visualization and our own experience obtained from working with flow mapping and user testing.⁵³

Hypotheses

In this section, we introduce our research questions and specific hypotheses:

H1. Which flow line symbolization(s) facilitate(s) a consistent performance and perception of patterns across different types of task, data, and layout settings? Overall, we expect MA to be the most successful design overall because of its simplicity and fewer design trade-offs.

H2. How do flow data influence task performance and pattern perception? Are there any differences in performance across different data sets? In light of our previous work,⁵³ we hypothesize that data would have a strong effect on task performance and pattern perception in flow maps, which can often be stronger than the design decisions.

H3. Does the layout orientation affect viewing strategies and thus the performance of designs? We expect that rotation of a layout will result in differences in performance and perception by changing the position of nodes, and flows, and thus, altering the saliency of patterns.

H4. Following the top-down perspective, we argue that arrowhead is a symbol in culture that indicates direction, and therefore, may be perceived easier than other visual variables that encode direction. Therefore, we hypothesize that designs with arrowhead (MA) will outperform the tapered design without arrows (TA) for identifying dominant direction (Task 1) and spatial focusing of long flows toward a destination (Task 3).

H5. We hypothesize that the use of the divergent color scheme as an additional visual variable to encode direction between origins and destinations, FA and DA will produce higher correctness and lower response time for the dominant direction task.

H6. Using varying thickness from narrow to wide (instead of wide to narrow) with less curvature and a gradual change of brightness, TD will outperform TA on perceiving dominant direction.

H7. MA and TA will outperform other designs in perceiving magnitude as they both use redundant visual variables of line width and color value, which has been found to be effective in perceiving magnitudes.²⁹ DA and FA will increase the difficulty in reading magnitude information because of using a divergent color scheme. FA will produce increased difficulty as a result of line shortening, and TD will likely produce decreased correctness because of its double use of color value for depicting both direction and magnitude.

H8. As previous work suggests, tapered flow line symbolizations (TA and TD) have been found to be ineffective for long flows.¹⁸ However, we expect TD to perform well and outperform TA on Task 3 because TD's design emphasizes the destination end of flows which Task 3 is looking for. Also, we expect FA to be ineffective for perceiving patterns that involve long flows (Task 3) as line shortening results in directional ambiguity especially in displaying long flows.¹⁷

H9. As the structural pattern of edges affect users' perception of groups in graphs,⁵² we hypothesize that participants' perception of clusters will be affected by line symbolization. We hypothesize that designs that emphasize the origin of flows will make net-exporters easier to detect. Specifically, we expect the design FA and DA to outperform other designs as the origin of the flow lines are symbolized with blue hue, which increase the emphasis on the point symbol for distinguishing net-exporters.

Experiment design

We used a mixed design with one between-subjects variable: line symbolization (five levels) and two within-subjects variables: data (four levels) and 180° rotation (two levels). The test included 27 questions, in which only 16 questions require participants to complete tasks using flow maps, while the rest are background and a final feedback question for the test and given flow maps.

We evaluated the complexity of the type of tasks based on the number of attributes (i.e. direction, magnitude, distance, and clustering) that participants are required to assess and user interactions needed to complete the task. We then ordered the task from simple to more complex. We kept the first two types of tasks (i.e. direction and magnitude) in a fixed order. We selected the direction task to be the first in order because of its simplicity: users are asked to evaluate the direction of all flows and select a choice among four alternatives (e.g. south, north, west, and east).

We selected magnitude to be the second task, as it requires users not only to evaluate the magnitude of all flows but also to interact with the map to select (identify) the top 3 flows with the highest magnitude. The complexity of the task increases for the last two tasks as users are asked to evaluate a combination of characteristics (flow distance and direction, node type, and distance and clustering) and interact with the map. Because there is no distinct difference between the complexities of the last two types of tasks, we randomized their order of appearance (Tasks 3 and 4).

Participants were given four distinct data sets for the four types of tasks, and we specifically kept the questions of the same task together in order to alleviate the confusion that may result from switching back and forth between the different task types. Random ordering for the two tasks and four data sets per task with 180° rotation was applied to all five designs to generate 480 (5 design × 4! data set × 2 rotation × 2 task) unique combinations which we randomly assigned to participants.

Tasks

We chose our tasks by a task-by-type taxonomy (Table 1) as follows:

- *Task 1.* Select the dominant direction that most flows are going to.
- *Task 2.* Select three flows that have the highest volume.
- *Task 3.* Select the circle that receives the highest number of flows in the top 10 in length.
- *Task 4.* Blue circles illustrate net-exporters which have greater total volume of exports than imports. Select a cluster of blue circles that are near one another.

To reduce the cognitive load induced by instructional materials, we kept the instructions as short as possible. Wording in task questions greatly influence the quality of response. In order to reduce the complexity of tasks, especially for the non-expert online users, we presented abstract tasks to the participants and avoided domain-specific words such as “node,” “edge,” and “location.” We directed users’ attention to the visual variables using abstract phrasing such as “Select the circle” instead of using the true verbal description of phenomenon that those visual variables present (e.g. “Select the state or location”).

Flow data

As a result of being constrained by geographic coordinates of nodes, flow map layout suffers from the visual

complexity induced by the number of flows, flow lengths, and crossings. However, geographies of flows have particular characteristics, and such characteristics are crucial for understanding holistic and geographic patterns in flow maps. To account for the characteristics of geographic flows and any bias that would be introduced by a particular layout, we designed the experiment with 16 real-world data sets on commodity flows in the United States which is collected and generated by the Commodity Flow Survey (CFS) in 2007. Each data set exhibits a particular set of patterns such as a dominant direction of flows, high-magnitude flows with varying length and position, convergence of long flows toward a destination, and clustering of net-exporters. Each data set was assigned to a task and used only once. We did not control any properties such as number of nodes, edges, and edge crossings. Our purpose was to observe whether data would have a significant effect on task performance and pattern perception in flow maps using four different data sets for each task.

Layout orientation (rotation)

We hypothesize that visual saliency of flows potentially impact the perception of node characteristics (prominence), and visual saliency is greatly influenced by the position of the flows and nodes, and orientation of the layout. In order to understand the relationship between the particular layout orientation and typical display viewing strategies, one can use layout rotation or creating a mirror image of the flow map. We employed a 180° rotation to consecutive flow maps so that the participants are given a 180° rotated layout for every other flow map in the test. In addition to evaluating the effect of layout orientation on viewing, rotation help take into account the learning effect on the position of nodes.

Procedure

The test is made available to the public using the following link: <http://tinyurl.com/flowmaptest>. Participants are first prompted with an instruction window that briefly describes interactive flow mapping and the online system that would be used by the participant. The test included 27 questions, in which only 16 questions require participants to complete tasks using flow maps, while the rest are background questions and a final open-ended question to receive feedback on the test and given flow maps.

There was no time limit to answer any of the questions; however, the participants were encouraged to spend 30–45 s to complete each task. The whole session took about 12 min on average. The test interface

detects screen resolution and the browser window's size to adjust the size of the map. If the height of the browser window is narrower than 768 pixels, the participant is not eligible to take the test. Zooming and panning were disabled in the test interface; however, users could still interact with the flow map to select (click on) nodes and flows to complete the tasks. The first task of each task type was given as a practice, and the expected answer after the practice was highlighted. We did not include the practice task in evaluating the results. We allowed participants to participate in the study only once.

Participants

Increasing number of studies has proven the usefulness of online crowdsourcing services for conducting usability experiments.^{54–56} Following this trend, we used AMT crowdsourcing service (<https://www.mturk.com>) to recruit participants. We paid each participant 50 cents to conduct the test that took 12 min on average. To ensure motivation (1) we required the participants to have greater than 1000 approved hits with a 90% hit approval rate and (2) we paid a bonus of 5 cents for each correct answer which added up to a total of 60 cents as bonus.

A total of 551 subjects participated in the test. We used a threshold of 2 s of average response time in order to eliminate “the spammers,” participants who quickly respond without thoughtfully considering the prompt.⁵⁷ There were 37 participants whose average response time was below 2 s. We also excluded six participants who self-reported to have impaired vision or English level as “Do not know English.” We recorded maximum time for unchanged cursor position to identify the participants who were idle during the test. We omitted 11 responses which were idle between approximately 3 and 60 min. We omitted a total of 54 participant responses which corresponded to approximately 10% of all responses.

We analyzed the responses of 496 participants (59% male, 41% female) after elimination. The ages of the participants were between 18 and 68 years, and median age was 33 years. The majority of the participants declared to have a college (41%) and graduate degree (39%), whereas there were participants with a high school degree (19%) and a degree with less than high school (1%). Most participants were from the United States (67%) and India (27%), and the rest 6% were from 18 different countries. A total of 47% of the participants stated that they had never seen a flow map before. After seeing a flow map, 91% of the participants stated (i.e. agree and strongly agree) that they understand what a flow map represents. The majority (95%) of the participants use computers

more than 3 h a day. Most of the participants (94%) use maps regularly (e.g. Google Maps) and feel comfortable about using online mapping services. To account for the performance variation due to screen size, we recorded screen resolution and used it as a factor in our statistical analysis of the results. Screen resolutions varied from 1024×768 to 2560×1440 . While 48% of the participants had 1366×768 and 28% of the participants had screen height equal or greater than 1024. In order to account for its effect on performances, we included screen resolution as a categorical variable derived from screen height (small screen: < 1024 , large screen: ≥ 1024). A total of 72% of participants had smaller screen height (< 1024), whereas 28% had larger (≥ 1024).

Results

We provide an application to view participants' responses: <http://tinyurl.com/flowmaptestresponses>. We use the following abbreviations for four unique data sets in each task: D1, D2, D3, and D4 and two rotations: U for Un-rotated and R for Rotated 180°.

Task 1

To complete the first task, participants selected one of the four choices that summarizes the direction of most flows in a given flow map. We illustrate the usability metrics by design, data set, and rotation: percentage of correct responses in Figure 1(a) and response time distributions using bean plots in Figure 1(b). We converted each participant's answer to a correctness score of 1 and 0. We selected design MA, no rotation, and data set D1 as the reference group for performing a series of logistic regressions on correctness score with independent variables: design, data set, rotation, and screen resolution. To meet the assumptions on the normality and homogeneity of residuals, we log transformed the response times and performed an analysis of variance (ANOVA) using the same independent variables. Screen size was found to have no significant effect on percent correctness; however, unsurprisingly, smaller screen resolution (height: < 1024) produced higher response time ($p < .001$).

Dropping screen resolution from the models, we performed logistic regression on correctness and ANOVA on log response time using design, data set, and rotation as independent variables (Table 3). As none of the interactions were significant, we ran the models without the interaction terms. All main effects were significant which highlight the significant differences among the levels of design, data set, and rotation (Table 3). MA outperformed all other designs. Data sets D2 and D4 resulted in higher accuracy than D1

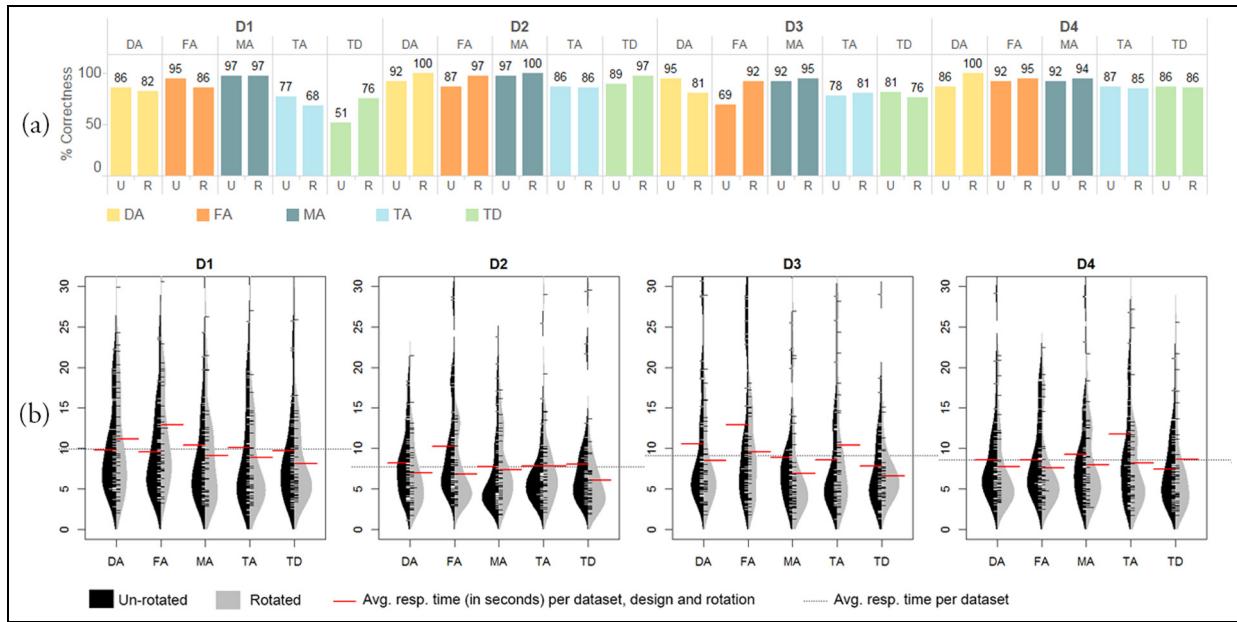


Figure 1. Usability metrics for Task 1 by data set, design, and rotation: (a) percent correctness and (b) distributions of response time (in seconds). Increased response time is often a result of stretched distributions and correlates with lower accuracy.

Table 3. Significant results on correctness and response time for Task 1.

Logistic regression on correctness (Reference group: MA, D1, U)			ANOVA on log response time		
Coefficients	Z	p	Coefficients	F	p
DA	-2.540	.011	Design (MA, DA, FA, TA, TD)	5.784	<.001
FA	-2.964	.003	Data set (D1, D2, D3, D4)	8.388	<.001
TA	-5.144	<.001	Rotation (U, R)	11.944	<.001
TD	-5.244	<.001	Tukey's HSD on log response time		
D2	4.635	<.001	Pairs		
D4	3.461	<.001	FA-MA	.1680	.008
R	1.694	.090	FA-TD	.2254	<.001
MA: ○	DA: ○	FA: ○	DA-TD	.1510	.026
TA: ○	TD: ○	U: Un-rotated; R: Rotated	D1-D2	.2211	<.001
			D1-D3	.1258	.029
			D1-D4	.1608	.002
			U-R	.1113	<.001

MA: Monotone Arrowhead; DA: Divergent Arrowhead; FA: Fading Arrowhead; TA: Tapered; TD: Teardrop; ANOVA: analysis of variance.

and D3. Although the effect of rotation depended on the orientation of the data set, rotation led to higher accuracy.

Similar to the analysis of correctness, the results of ANOVA on log response time did not produce any significant interaction effect; however, all three main effects were found to be significant (Table 3). Post-hoc pairwise comparisons showed that FA produced higher average log response time than MA; FA and DA produced higher response times than TD. Great variation in response times between FA-TD and DA-TD could

clearly be observed from the bean plots (Figure 1(b)). TD produced relatively more compact distributions than DA and FA, which indicates less variation among users. D1 resulted in higher average response time than other data sets: D2, D3, and D4. In addition, rotation resulted in a significant decrease in time to respond. Logistic regression on accuracy with response time as an independent variable revealed a significant negative correlation ($p < .001$) between response time and accuracy which could also be observed by comparing Figure 1(a)–(b).

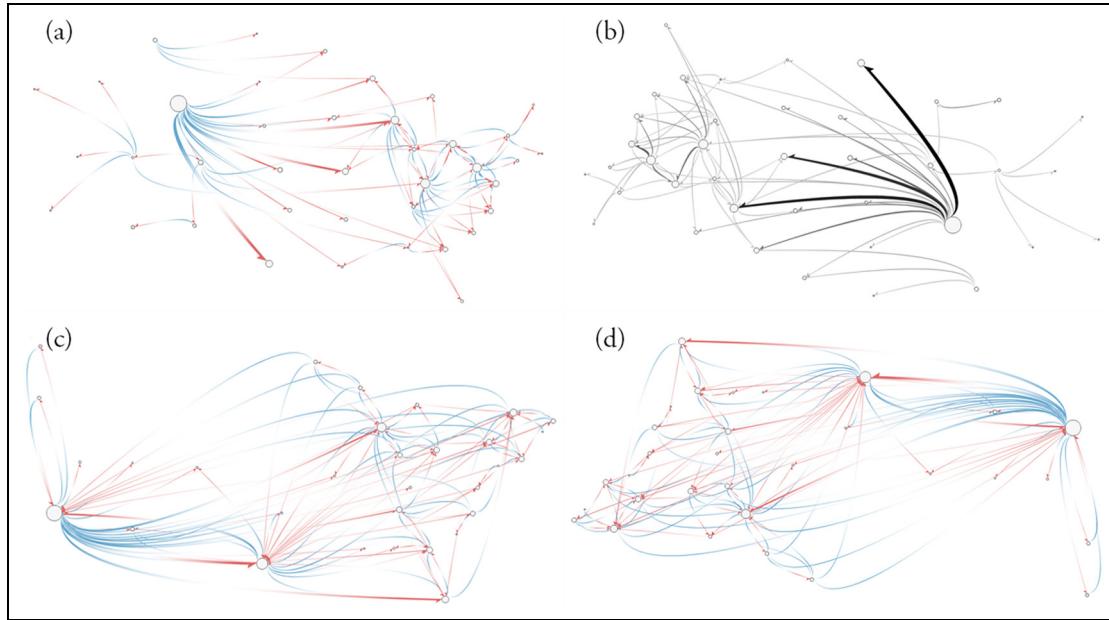


Figure 2. Example flow maps used in Task 1: (a) FA on un-rotated D2 (correctness [C]: 87%, avg. resp. time [T]: 10.9 s); (b) MA on rotated D2 (C: 100%, T: 7.4); (c) FA on un-rotated D3 (C: 69%, T: 13.1); and (d) FA on rotated D3 (C: 92%, T: 9.5). Rotation on D2 and D3 resulted in significant differences in correctness and in response times and affected FA the most.

To illustrate the observed significant differences, Figure 2 depicts the flow maps created using the following: (1) FA on un-rotated D2, (2) MA on rotated D2, (3) FA on un-rotated D3, and (4) FA on rotated D3. Participants were quicker and more accurate in picking the most common direction for flows when D2 was rotated (Figure 2(b)). Rotation resulted in 23% increase in correctness and 4-s decrease in average response time for FA on D3 (Figure 2(c) and (d)).

Significant effects of data set on correctness and response time support our hypothesis H2 that data have a strong effect on the performance of users. This is not surprising as the responses are most likely to be influenced by the availability of alternative answers, the visual complexity, and layout configuration of each flow map. For example, while the most dominant direction of flows in D3 is west-to-east, there are also a large number of flows from east-to-west which some participants likely to choose as the answer (Figure 2(c)). On the other hand, the alternative answer in D2, south-east to north-west, has much less number of flows than the correct choice north-west to south-east (Figure 2(a)). Visual complexity also influences user responses and is determined by various properties of data sets such as number of nodes, edges, edge crossings, clustering of nodes and flows, and differences in magnitude, direction, and length among flows.

Significant effect of rotation on correctness and response time support our hypothesis H3 that rotation (or layout orientation) has an impact on user

perception and responses. Although our findings could still happen by chance, we argue that higher accuracy on rotation may be a result of re-positioning of flows relative to the direction the participants scan the map from top-left.⁵⁸ When the destination ends of flows, which users attention is focused when looking for directions, are positioned closer to top-left, participants were more accurate and faster (i.e. compare Figure 2(a) and (b) and Figure 2(c) and (d)).

MA outperformed all other designs. In order to evaluate whether designs with arrowheads produce higher accuracy, we ran logistic regressions using DA and FA as reference level. Both DA and FA were found to produce significantly higher accuracy than TA and TD. These findings support our hypothesis H4 that arrowheads produce higher accuracy on identifying dominant direction. On the other hand, DA and FA produced higher average response time than MA, and TD and lower correctness than MA which contradicts our hypothesis H5 that divergent color scheme would result in higher accuracy and lower response times. We attribute this finding to increased cognitive load caused by divided attention among multiple visual variables (e.g. divergent color hues and arrowhead) to represent direction.⁵⁹ We expected TD to outperform TA (H6); however, we did not find any significant differences in correctness among the two designs. However, TD resulted in decreased average response time than FA and DA, which could be attributed to its additional visual variable of varying color value for depicting direction.

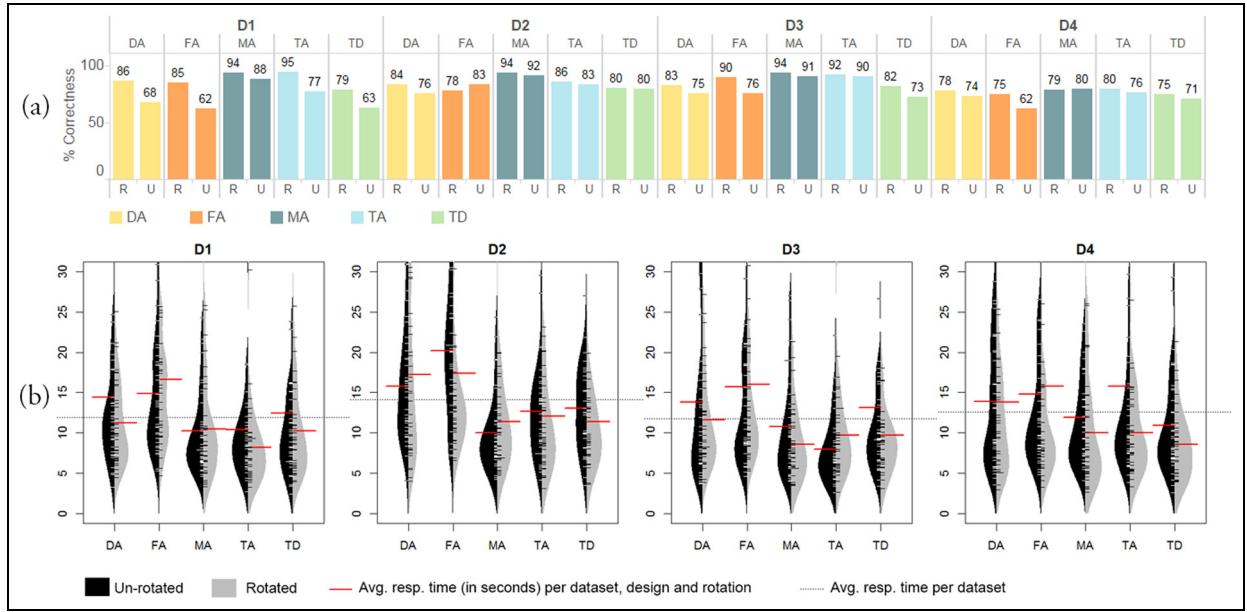


Figure 3. Usability metrics for Task 2 by data set, design, and rotation: (a) percent correctness and (b) distributions of response time (in seconds). FA, DA, and TD resulted in lower accuracy and higher response times which was largely due to increased performance variation between participants.

Task 2

To complete the second task, participants were instructed to compare the volume of the flows and select three flows which appeared to have the highest volume. For each flow map (i.e. a combination of design, data set, and rotation), we calculated a percent correctness score by dividing the number of times the top 3 flows are selected by the total number of selections ($3 \times \#$ participants). ANOVA on percent correctness revealed significant main effects of design, data set, and rotation and a two-way significant interaction effect on data set and rotation (Table 4). Screen size was found to have no significant effect on correctness and excluded from the model. Figure 3 illustrates (1) percent correctness and (2) distributions of response time by data set, design, and rotation.

Pairwise comparisons revealed 11 significant rotation–data set interactions (Table 4). Among these pairs, D1:R–D1:U compares the rotated and unrotated versions of D1 and highlighted a significant increase in correctness when D1 was rotated (Figure 3(a) and (b)). On the other hand, comparisons between the levels of design revealed that MA and TA outperformed all other designs, which confirms our hypothesis H7 on magnitude. We also found significant differences among the data sets which support our hypothesis H2 that flow data impact performances. D4 resulted in lower correctness than other data sets, and D3 was found to produce higher correctness than D1. Rotation also produced higher

correctness which supports our hypothesis H3 that layout orientation impacts performance.

On the other hand, FA and DA showed relatively more stretched distributions that highlight greater variation of response time among participants (Figure 3(b)). The results of ANOVA on log response time revealed significant main effects of design, data set, and rotation and the interaction between design and data set. Screen resolution was not found to have a significant effect on correctness or response time. Pairwise comparisons between the levels of design and data set using Tukey's honest significant difference (HSD) showed that FA led to higher log response time than all other designs. DA resulted in higher average log response time than MA, TA, and TD.

In order to discuss possible reasons behind the performance variation, we compare four flow maps: MA on rotated D1 (Figure 4(a)) with FA on un-rotated D1 (Figure 4(b)) and MA on rotated D3 (Figure 4(c)) with FA on un-rotated D3 (Figure 4(d)). Due to its use of only line width to depict magnitude and origin-destination coloring acting as distractors, FA's lower performance can be observed in both data sets D1 and D3. Findings of eye movements on various media sources commonly agree that most users start scanning an image from top-left region which attracts the most and earliest attention.^{58,60} Rotation alters the relative position of the flow lines and thus the salient areas in a flow map. As a result, visually salient flows that are placed along the direction of users' gazing behavior (from top-left) possibly stay in their short-term memory,

Table 4. Significant results on correctness and response time for Task 2.

ANOVA on percent correctness			ANOVA on log response time		
Coefficients	F	p	Coefficients	F	p
Design (MA, DA, FA, TA, TD)	16.991	<.001	Design	34.389	<.001
Data set (D1, D2, D3, D4)	11.930	<.001	Data set	13.049	<.001
Rotation (U, R)	33.469	<.001	Rotation	12.362	<.001
Rotation × data set	6.093	.003	Design × data set	1.669	.06

Tukey's HSD on percent correctness						Tukey's HSD on log response time					
Pairs	Diff.	p	Interaction pairs	Diff.	p	Pairs	Diff.	p	Interaction pairs	Diff.	p
MA-DA	.1082	<.001	D1:R-D4:R	.1047	.006	DA-MA	.2550	<.001	FA:D4-TD:D4	.3772	.005
TA-DA	.0703	.012	D1:R-D1:U	.1614	<.001	DA-TA	.2390	<.001	FA:D3-MA:D3	.4263	<.001
MA-FA	.1214	<.001	D1:R-D4:U	.1511	<.001	DA-TD	.2182	<.001	FA:D3-TA:D3	.4937	<.001
TA-FA	.0835	.002	D2:R-D1:U	.1279	<.001	FA-DA	.1675	.002	FA:D3-TD:D3	.3271	.04
MA-TD	.1350	<.001	D2:R-D4:U	.1176	.002	FA-MA	.4225	<.001	FA:D2-MA:D2	.5585	<.001
TA-TD	.0972	<.001	D3:R-D4:R	.1085	.004	FA-TA	.4065	<.001	FA:D2-TA:D2	.4414	<.001
D1-D4	.0472	.062	D3:R-D1:U	.1652	<.001	FA-TD	.3857	<.001	FA:D2-TD:D2	.4447	<.001
D2-D4	.0865	<.001	D3:R-D4:U	.1549	<.001	D2-D1	.1945	<.001	FA:D2-DA:D2	.4438	<.001
D3-D4	.0965	<.001	D2:U-D1:U	.1122	.003	D3-D2	.2321	<.001	FA:D1-MA:D1	.4223	<.001
D3-D1	.0493	.048	D3:U-D1:U	.0947	.017	D2-D4	.1756	<.001	FA:D1-TD:D1	.5310	<.001
R-U	.0735	<.001	D2:U-D4:U	.1018	.008	U-R	.1003	<.001	FA:D1-TD:D1	.3881	.003

MA: DA: FA: TA: TD:
U: Un-rotated; R: Rotated

MA: Monotone Arrowhead; DA: Divergent Arrowhead; FA: Fading Arrowhead; TA: Tapered; TD: Teardrop; ANOVA: analysis of variance; HSD: honest significant difference.

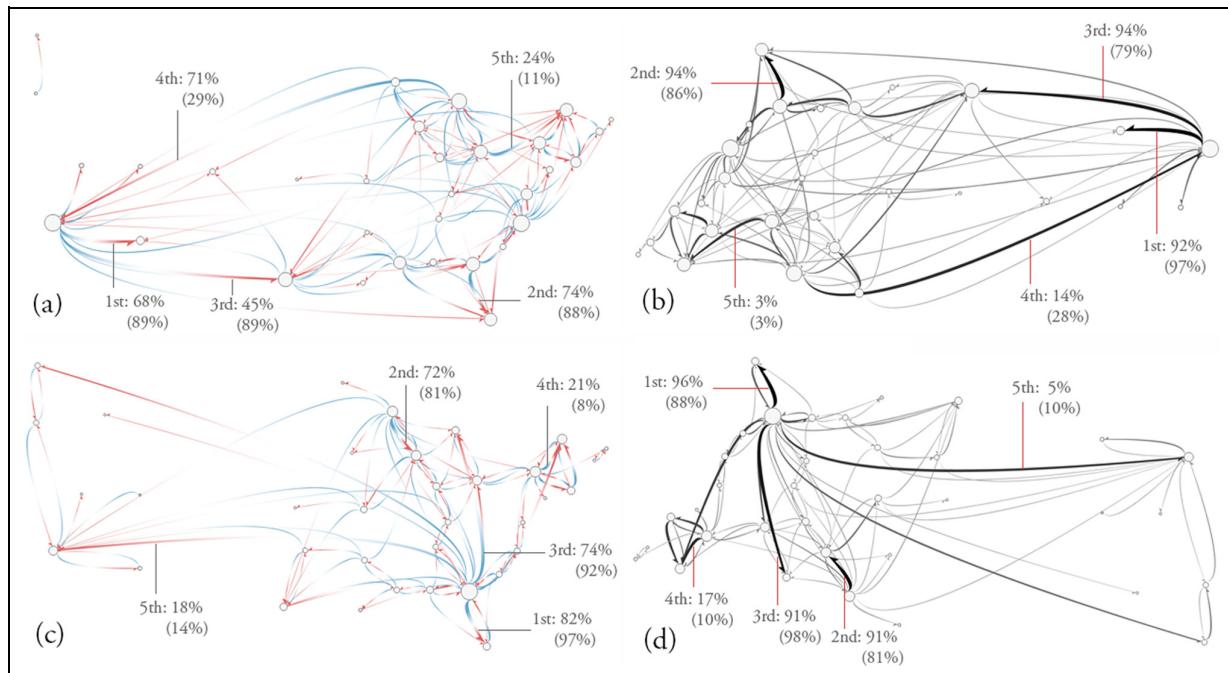


Figure 4. Example flow maps used in Task 2: (a) FA on rotated D1, (b) MA on un-rotated D1, (c) FA on un-rotated D3, and (d) MA on rotated D3. The percentage of participants that selected the top 5 flows are shown for each map, and the percentage values for the 180° rotated version of each map is given in parenthesis.

as they compare the magnitude of all flows in the map. We observe the effect of rotation in flow maps shown in Figure 4. For example, the top 3 flows were at the bottom of the layout in un-rotated D1 (Figure 4(a)), and the percentage of participants that selected the top 3 flows was 68%, 74%, and 89%. When D1 was rotated, the top 3 flows were placed at the top of the layout, and the percentages increased to 89%, 88%, and 89%. We provide a further discussion on the influence of gazing behavior, saliency of flows, and task-dependent factors in the “Discussion” section.

Overall, the results confirm our hypothesis H7 on the success of MA and TA as a result of their use of color value as an additional visual variable to depict the volume of a flow. We attribute lower percent correctness for FA and DA to their use of only line width for depicting magnitude and the use of gradual change of color hue and transparency (i.e. for FA) for depicting directionality which act as distractors when assessing the magnitude of flows. On the other hand, we attribute TD’s lower percent correctness to its double use of color value to encode magnitude and direction between origin and destination.

Task 3

For the third task, participants were asked to select a location (circle) that received the highest number of flows that are top 10 in length. Figure 5 illustrates (1) the most common participant responses with their percentages and (2) distribution of response times. We

converted each response to an accuracy score of 1 and 0 and ran a series of logistic regressions on accuracy with independent variables design, data set, and rotation and reference levels: MA, no rotation, and D1 (Table 5). Significant interaction effects exist between rotation and data set and design and data set. Pairwise comparisons revealed that the odds of correctly performing the task increased when TA and TD were used on D3 and TD was used on D4, whereas the odds for correct response were decreased when D3 and D4 were rotated (Table 5). Besides the interaction effects, the main effects of data set and design were found to be significant, while rotation was not. Overall, TA performed significantly worse than MA, whereas there were no significant differences between MA and other designs. The results of logistic regressions using TD, FA, and DA as the reference level also showed that each of the designs significantly increased the odds of correct response as compared to TA ($p < .001$). We attribute TD’s increased correctness to its emphasis on the destination end of flows which Task 3 is asking. We also expected FA to be ineffective for perceiving patterns that involve long flows as line shortening results in directional ambiguity especially in displaying long flows. However, the results did not show significant differences between FA and other designs: MA, TD, and DA. On the other hand, the odds of accurately performing the task increased on D2 and D4 regardless of design and rotation.

ANOVA on log response times revealed none of the interaction effects as significant, whereas main effects

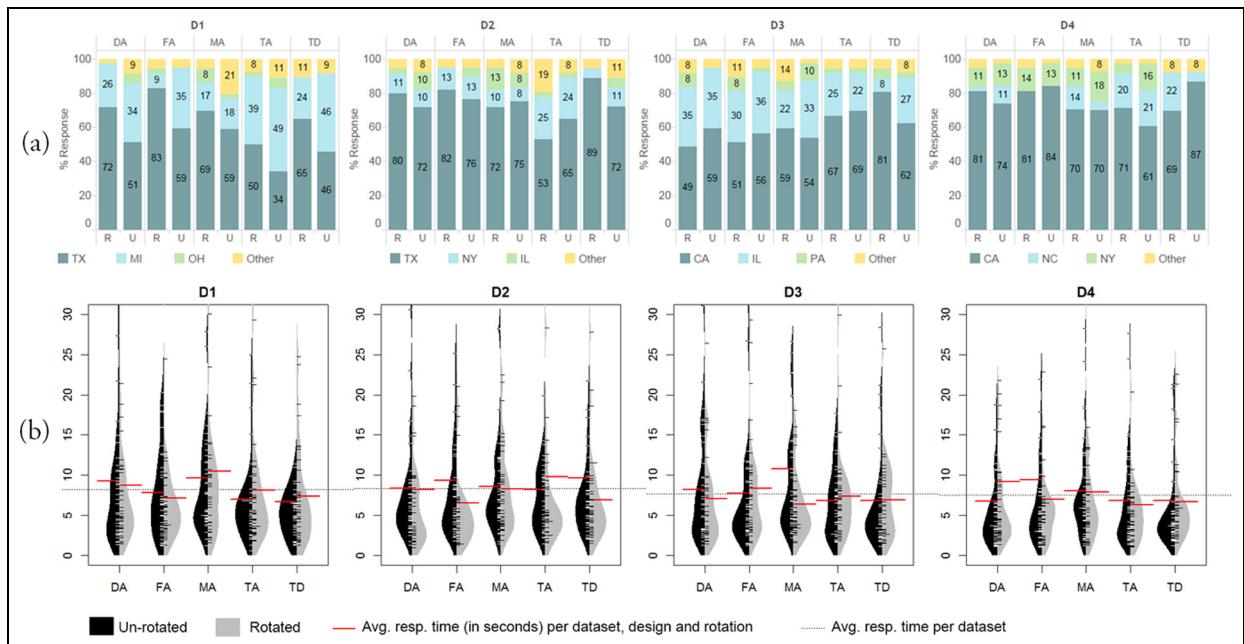
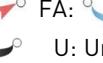


Figure 5. Usability metrics for Task 3: (a) responses and percent correctness and (b) distributions of response time (in seconds).

Table 5. Significant results for correctness and response time for Task 3.

Log regression on correctness			ANOVA on log response time		
Coefficients	Z	p	Coefficients	F	p
R:D3	-2.252	.024	Design	3.438	.008
R:D4	-2.278	.022	Data set	4.022	.007
TD:D4	1.684	.092	Tukey's HSD on log response time		
D2	4.162	<.001	Pairs	Diff.	p
D4	4.841	<.001	MA-TD	.1680	.008
TA	-2.101	.035	MA-TA	.2254	<.001
MA: 	DA: 	FA: 	D2-D4	.1768	.009
TA: 	TD: 	U: Un-rotated; R: Rotated	D1-D4	.1610	.02

MA: Monotone Arrowhead; DA: Divergent Arrowhead; FA: Fading Arrowhead; TA: Tapered; TD: Teardrop; ANOVA: analysis of variance; HSD: honest significant difference.

of design and data set resulted in significant differences (Table 5). Although the main effect of rotation on correctness was not found to be significant, rotation resulted in observable differences in response time (Figure 5(b)): FA on D2, MA on D3, and DA and FA on D4). Pairwise comparisons showed that TA and TD resulted in shorter response times than MA. This finding suggests increased difficulty in performing the task with MA, which also correlates with MA's decreased accuracy in comparison with TA and TD on D3. Participants responded significantly faster on D4 than D2 and D1. D4's faster response also correlates with its higher accuracy.

The presence of the significant interaction between data and rotation supports our hypotheses that rotation of the layout may impact the performance (H3). We attribute the significant interaction between data and rotation to the relative positioning of the nodes and flows in relation to the way users scan the map from top-left. Correct nodes were selected more often when they were positioned at the top-left and top of the layout. This finding is also consistent with the findings of the first two tasks that indicated increased visual saliency and thus easier perception of the flows that were placed at the top-left and top of the interface. As Figure 5(a) suggests, rotation seemed to impact responses in all designs, whereas MA was the least affected. TD benefits from a less cluttered layout, thanks to their tapered design and the absence of arrowheads, and convergence of long flow lines toward the destination regardless of the difference between the two curves (Figure 6(c)).

We discuss that participants' responses are likely to be influenced by a variety of factors including the number of competing nodes (answers) and the saliency of nodes which is determined by the size and position of the circle and length, width, color value,

and position of the flows that converge at each competing node. In order to shed light into our discussion, we illustrate the best performing designs per data set: FA on rotated D1 (Figure 6(a)), TD on rotated D2 (Figure 6(b)), TD on rotated D3 (Figure 6(c)), and FA on un-rotated D4 (Figure 6(d)). The top 2 most common responses are labeled with the percentage of users that selected them both for the map in the figure and its 180° rotated version in parenthesis. Figure 6(a)–(c) highlights that nodes were selected more often when they were positioned closer to the top of the layout. Because of the complexity of the task, we hypothesize that some participants took shortcuts and used their intuition in performing the task instead of thoroughly scanning the map to compare direction and length of flows around the candidate nodes and determine whether the node receives most of the top 10 flows. As a result, nodes that are visually more salient (e.g. larger in size, placed closer to top-left, and top) were selected more often as shortcuts.

Task 4

To complete the fourth task, participants were asked to select a cluster of net-exporters that are near one another. As the experiment does not control for the factors of perceptual grouping (e.g. proximity, similarity), our goal for the analysis is exploratory, and we do not aim to predict perceived groups, we rather focus on what those groups are based on the combination of design, data, and rotation. Each participant selected a minimum of three circles to define a cluster.

ANOVA on log response times revealed significant main effects of design, data set, and rotation (Table 6). None of the interactions were found to be significant and screen resolution was dropped from the model as

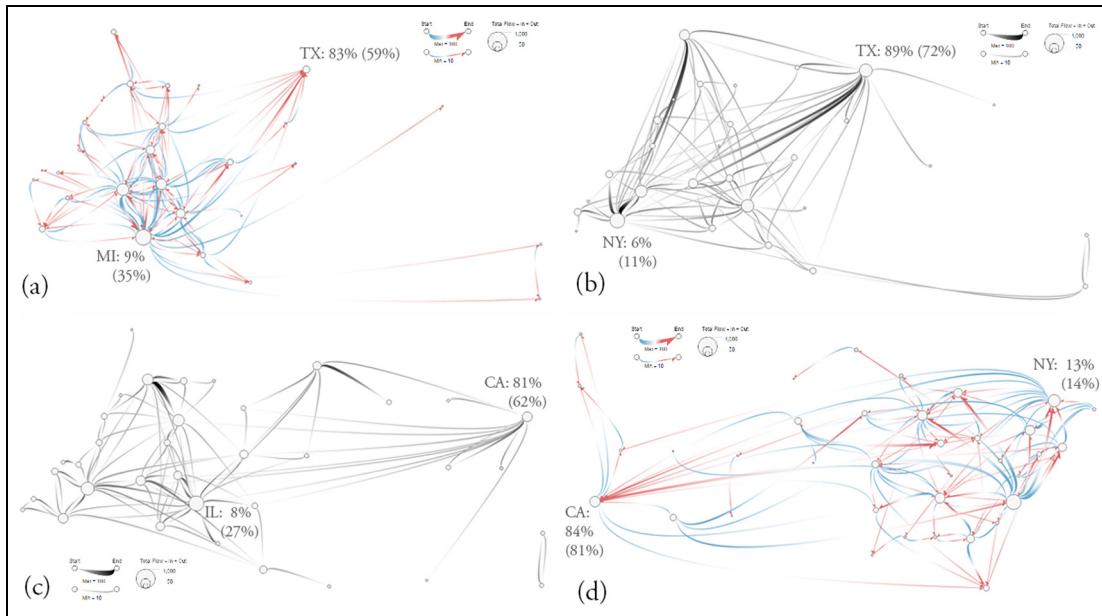


Figure 6. Example flow maps used in Task 3: (a) FA on rotated D1, (b) TD on rotated D2, (c) TD on rotated D3, and (d) FA on un-rotated D4.

Table 6. The effects of design and data set on response agreement and time for Task 4.

ANOVA on response agreement			ANOVA on log response time		
Coefficients	F	p	Coefficients	F	p
Design	4.586	.018	Design	3.381	.009
Data	17.479	<.001	Data set	13.955	<.001
Tukey's HSD on response agreement			Tukey's HSD on log response time		
Pairs	Diff.	p	Pairs	Diff.	p
FA-TA	.0890	.027	MA-TD	.1756	.009
TD-TA	.0852	.034	MA-FA	.1613	.019
D3-D1	.1485	<.001	D2-D4	.2993	<.001
D3-D2	.1265	<.001	D2-D3	.2067	<.001
D3-D4	.1114	.002	D2-D1	.2043	.002
			U-R	.0659	.050

MA: Monotone Arrowhead; FA: Fading Arrowhead; TA: Tapered; TD: Teardrop; ANOVA: analysis of variance; HSD: honest significant difference.

it did not have a significant effect on response time. Tukey's HSD showed that D2 resulted in higher response times than other data sets. This was because participants' answers varied greatly as D2 included two alternative clusters, while the other data sets included only one alternative. MA resulted in higher response time than FA and TD.

We evaluated the differences among line symbolizations by comparing the variation in participants' delineation of clusters. First, we calculated the percentage of nodes selected as a part of a cluster for each

combination of design, data set, and rotation. Using the percentage value of the nodes that were selected at least once, we performed an ANOVA for each data set with independent variables of rotation and design. Neither design nor rotation resulted in a significant difference. This suggests that unlike the first three tasks, user responses in Task 4 were not influenced by rotation or design.

Second, we calculated the percentage of the most commonly selected clusters by participants for each data set and design (Figure 7(a)). In Figure 7(a), light

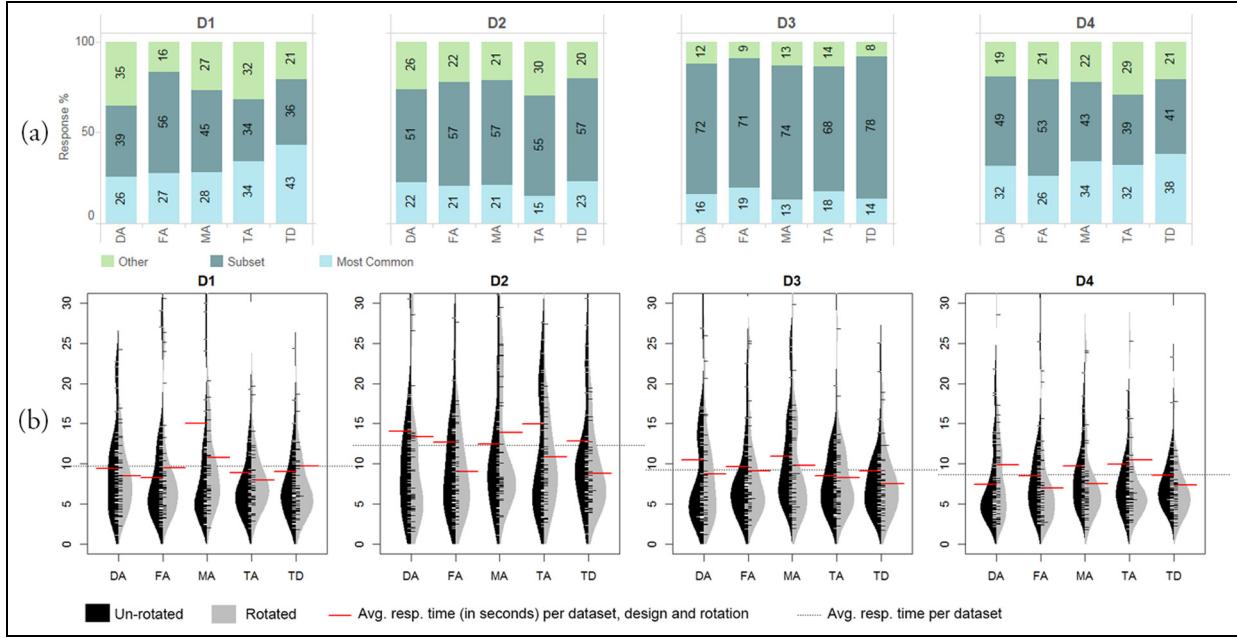


Figure 7. Participant responses and response times for Task 4: (a) the most common answers by their percentage. Categories “Most Common” and “Subset” illustrate the percentage of responses that participants agreed upon per data set and design and (b) distributions of response time (in seconds).

blue indicates the percentage of the most common response per data set and design, and darker blue indicates the percentage of the second most common and the subset of the first two most common responses. The subset includes user responses with nodes in which each node exists either in the first or the second most common answer. We then combined the two most common and subset categories to derive a measure of agreement on responses. We performed an ANOVA on the agreement measure with independent variables data set and design (Table 6). Pairwise comparisons revealed that D3 resulted in less variation in cluster descriptions (higher percentage of agreement) than D1, D2, and D4; and FA and TA resulted in less variation than TD, while differences between other designs and data sets were not found to be significant.

Third, in order to analyze which common pairs of nodes were selected as a part of a cluster for each data set and design, we summed the number of times each pair of node was selected and normalized it by the number of participants to derive percent co-occurrence of node pairs in all responses. We then constructed an undirected graph in which the width and color value of an edge represent the percent occurrence of a node pair, and the size of the node represents percent occurrence of a node in participants’ responses. Figure 8 illustrates (1) FA on un-rotated D1 and (2) percent occurrence of nodes and node pairs in cluster definitions. The most common response for FA on D1 was 27% and included the

following nodes: CO, KS, MO, NE, IA, SD, and WI. A total of 56% of the responses were a subset of the most common response, and 22% was a subset of the second most common response. Undirected graph of node pairs (Figure 8(b)) reveals that nodes IA, MO, and WI appeared in participant responses slightly more than other node pairs. We attribute this finding to all three nodes’ visual saliency and similarity in size and proximity.

As another example, Figure 9 illustrates (1) MA on un-rotated D4 and (2) percent occurrence of nodes and node pairs in cluster definitions. The most common response for FA on D1 was 34% and included the following nodes: IL, IN, KY, MA, OH, PA, and WV. A total of 43% of the responses were a subset of the most common response, and 16% was a subset of the second most common response and included IL, IN, and OH. Undirected graph of node pairs (Figure 9(b)) reveals that pairs of nodes IL, IN, OH, and PA appeared in participant responses more often than the other node pairs. While proximity and similarity of circle sizes are potential factors, we hypothesize that higher magnitude flows between the nodes may contribute to the perception of these nodes to appear as a cluster.

Rotation or orientation of the flow map significantly affected all tasks that involved comparison of direction, magnitude, and distance, whereas clustering was not affected by rotation. We attribute this finding to the nature of clustering task, as participants

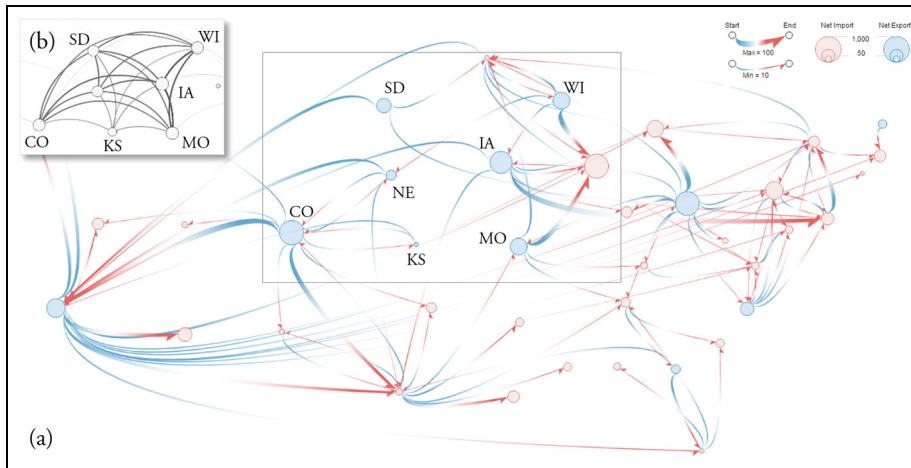


Figure 8. Example flow map for Task 4: (a) FA on un-rotated D1 and (b) undirected graph of percent co-occurrence of nodes that commonly appeared in participant responses.

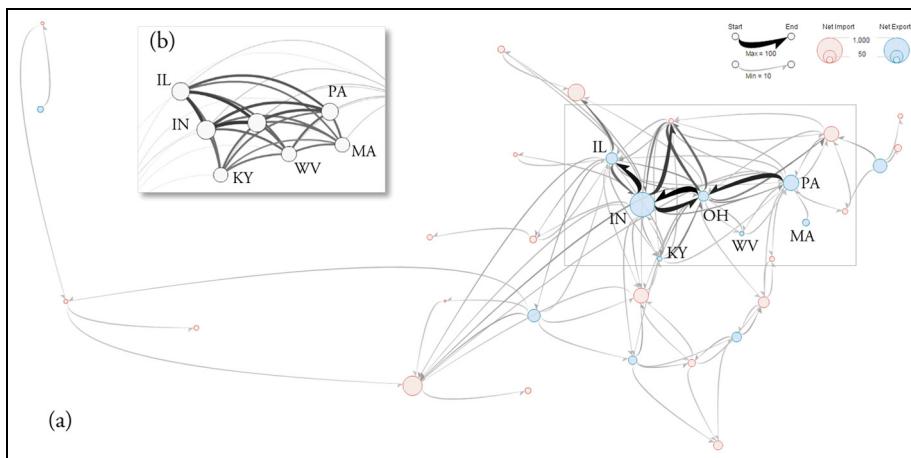


Figure 9. Example flow map for Task 4: (a) MA on un-rotated D4 and (b) percent occurrence of nodes that commonly appeared in participant responses. Labeled nodes represent the most common answers.

predominantly used spatial attention to focus on a region of the map to find out similar nodes positioned near one another.

We hypothesized (H9) that designs DA and FA would ease the perception of export clusters as color scheme of the flows would enhance the discriminability of node clusters. The results showed that FA and TD led to more accurate and faster responses, whereas DA was not found to have a significant on correctness or response time. This could be attributed to the ability of FA to produce a less cluttered display which eventually enhances the perception of clusters. Decreased difficulty for TD can be attributed to its asymmetric tapered design where the flow lines are thin and start from a point at the origin and thus result in less cluttering for detecting export patterns.

Discussion

Our findings indicated that there is potential usefulness for all of the five symbolizations we tested; however, the influence of the design on performance and perception depends on the type of the task. We recommend that the choice of line symbolization should be guided by a task taxonomy which end users are expected to perform. We guide our discussion on our results by comparing the efficiency (response time) and effectiveness (correctness) of the five line symbolizations across tasks (Table 7).

Hypotheses

H1. The results confirmed our hypothesis that MA was significantly faster to read and more

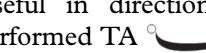
Table 7. Performance differences among the five designs by task.

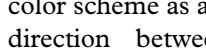
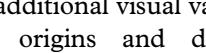
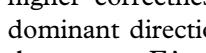
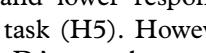
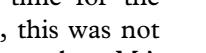
Task	MA 		DA 		FA 		TA 		TD 	
	C	T	C	T	C	T	C	T	C	T
T1: Dominant direction	✓	✓								✓
T2: Flows with the top 3 magnitude	✓	✓					✓	✓		✓
T3: Spatial focusing of long flows toward a destination	✓		✓		✓			✓	✓	✓
T4: Clustering of net-exporters					✓	✓			✓	✓

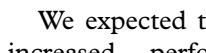
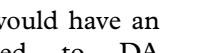
C: correctness; T: response time; MA: Monotone Arrowhead; DA: Divergent Arrowhead; FA: Fading Arrowhead; TA: Tapered; TD: Teardrop. Check marks illustrate the most accurate (C) and fastest response (T) for each task.

accurate in judging relative magnitude (Task 2) and direction tasks (Task 1), and its performance was found to be no different than other designs in spatial focusing of long flows (Task 3) and clustering of net-exporters (Task 4). Because of its consistent performance across tasks, data and rotation, and its simple design, we recommend MA for designing flow maps for exploratory visualization that involve multipurpose tasks.

H2 and 3. The results confirmed our hypotheses that data (H2) and layout orientation (H3) have strong effect on performance and perception of patterns in flow maps. However, the effect was dependent on the type of the task. While Tasks 1, 2, and 3 were significantly affected by rotation, Task 4 was not. Earliest attention of users and fixations are influenced by three main factors: (1) salience of areas in the image, (2) memory and expectations about where to find information, and (3) task and information at hand.^{58,60} We attribute the unchanged response of users in Task 4 to the nature of the clustering task where participants predominantly use spatial attention to focus on a region of the map to find out similar nodes positioned near one another (i.e. visually salient blobs), and this task potentially overrides the effect of eye gazing behavior. On the other hand, in Tasks 1, 2, and 3, users compare individual elements (i.e. nodes, flows) rather than blobs, and rotation alters the patterns by changing the relative position of the flow lines and thus the salient areas in the image. As a result, visually salient flows that are placed along the direction of users' gazing behavior (i.e. from top-left, top-center, and left-center) possibly stay in their short-term memory, as they compare properties of all other flows in the map.

H4. The results confirmed our hypothesis H4 that arrowheads are useful in direction tasks as MA  outperformed TA  in accuracy for tasks to identify dominant direction (Task 1) and spatial focusing of long flows toward a destination (Task 3).

H5. We hypothesized that the use of the divergent color scheme as an additional visual variable to encode direction between origins and destinations, FA  and DA  would produce higher correctness and lower response time for the dominant direction task (H5). However, this was not the case as FA and DA were less accurate than MA , but more accurate than TA  and TD  in Task 1. Moreover, FA and DA resulted in relatively more stretched response time distributions with higher average response times, which was caused by greater variation in response times among participants. Despite its use as a redundant visual variable, we argue that increased difficulty in performing direction tasks with the divergent color scheme was likely because participants' attention to compare directions was divided among multiple attributes (e.g. two divergent color hues, arrowhead, and curvature) which made perceptual judgments more difficult than MA when participants could attend to only arrowhead and curvature to evaluate directions.⁵⁹

We expected that FA  would have an increased performance as compared to DA  and would perform as well as MA , thanks to producing a less cluttered flow map by line shortening. However, the results

showed that FA's performance on identifying the dominant direction of flows was lower than that of MA, and there were no significant differences between FA and DA. We argue that this could be due to the use of line shortening, which makes flows with smaller magnitudes become more salient as compared to other designs. As a result, participants' attention is divided among a larger set of flows and making the comparisons more difficult and time-consuming, this results in decreasing correctness and increasing response time:

H6. Because of its decreased curvature, additional use of color value from light to dark and use of line width from narrow to wide instead of wide to narrow, we hypothesized (H6) that TD would outperform TA; however, we did not find any significant differences in correctness among the two designs. However, TD was the most efficient in response time and was significantly quicker than FA and DA. We recommend TD for direction tasks when response time is important.

H7. The results confirmed our hypothesis (H7) that the redundant visual variables of color value and line width would produce higher correctness and lower response times for MA and TA in the magnitude task (Task 2). Grayscale coloring scheme used in MA and TA helped users see the contrast between the darker, more dominant flows, versus the lighter and less salient smaller flows which significantly increased the performance in magnitude task. The test results also confirmed our hypothesis on TD's failure on magnitude task due to its mixed use of varying color value and thickness across the length of a flow line. While TD could potentially be useful in direction tasks, it must be used in caution when displaying an edge attribute (e.g. magnitude or length). Both FA and DA resulted in lower correctness and increased response time in judging the magnitude of flows because of increased difficulty of the search task caused by the divergent color scheme and use of just one variable (line width) to encode magnitude while other designs employed color darkness as a redundant variable.

H8. We expected MA, TD, and DA to be successful on Task 3 (H8) which was confirmed by our results. As it emphasizes the destination end of flows which Task 3 is looking for, TD was effective in perceiving convergence patterns toward a destination, whereas TA resulted in lower correctness as it has strong emphasis on the origin of flows. We also expected FA to be ineffective for perceiving patterns that involve long flows as line shortening results in

directional ambiguity especially in displaying long flows;¹⁷ however, FA was as effective as MA, TD, and DA.

H9. We hypothesized (H9) that design DA and FA would ease the perception of net-export clusters as color scheme of the flows would enhance the discriminability of net-exporters. The results confirm this hypothesis with FA, whereas DA was not found to have a significant. In addition to FA, TD produced more accurate and faster responses. The success of both FA and TD indicates that reduced cluttering help the perception of node clusters. Decreased difficulty for TD can be attributed to its asymmetric tapered design where the flow lines are thin and start from a point at the origin and thus result in less cluttering for detecting net-exporters which potentially include a large number of out-flows. Line symbolization could override node proximity and similarity effects when perceiving groups (clusters). Therefore, for tasks that require the perception of node clusters could use line designs that emphasize the nodes.

Limitations

We would like to acknowledge that our findings are limited by the experimental parameters, and some of the conclusions may not apply to the general comparison of flow map designs. We speculated on user strategies in order to explain our results; however, further studies that employ eye-tracking must be conducted to study cognitive processes and behaviors linked to flow map reading.

Since the participants were from AMT with a certain level of computer skills, the findings of the study are not necessarily representative of a broader population with diverse backgrounds. Because the test is not administered by an experimental facilitator, confounding factors such as multi-tasking and factors related to the test environment such as display type and size, lighting, and subjects' viewing distance and angle.⁵⁷

User performance and perception are greatly influenced by not only the complexity of the flow map and task but also user motivation and inattention. It is difficult to know how much time is due to users' inattention and response submission. While the participants' ability to identify and click on a circle is irrelevant of their judgment and perception, it affects the process of pattern search and thus the test performance. Divided attention becomes more difficult when tasks are harder. For Task 3, participants evaluated both node and flow elements and searched for patterns of direction and length. We hypothesize that some participants took shortcuts to select nodes that are visually more salient (e.g. larger in size, placed closer to the top).

Administered tests coupled with eye-tracking studies could reveal the pattern search process and help understand the actual perception of patterns.

Conclusion

We introduced a user evaluation study that compared line symbolizations for directed origin–destination flow maps. Our study is the first that tested the usefulness of five commonly used line symbolizations in origin–destination flow mapping: MA , DA , FA , TA , and TD . Participants in our study were asked to perform four tasks: identifying the dominant direction of flows, top flows with the highest magnitude, spatial focusing of long flows toward a destination, and clustering of locations with net-outflows. We evaluated the performance (i.e. correctness and response time) and perception of the five line symbolizations using real-world flow data sets and systematic rotation to take into account the layout orientation and potential learning effects. We recruited our participants from AMT which is an online crowdsourcing platform.

The results supported our hypothesis that data and orientation (rotation) both have significant effect on performance and perception of patterns in flow maps. Building upon the previous literature in image viewing and gazing behavior, detailed results on user responses suggest that the effect of rotation on performances could be due to the change in visual saliency of node and flow patterns in relation to the way users scan the map.

According to the results of our study, MA  performed well with minimized subject variations across data sets and rotation, higher performance, and efficiency in judging relative magnitude (Task 2) and direction tasks (Task 1). Our results also highlighted FA  as a potentially useful design, thanks to its use of line shortening and origin–destination coloring with a divergent color scheme. While line shortening improved FA's performance in direction tasks, it also created increased variability in responses as less cluttered flow maps increased the saliency of obscured and small flows and thus the cognitive load and number of flows for comparison. From this study, we can conclude that there is potential usefulness for all of the five symbolizations we tested; however, the influence of the design on performance and perception depends on the type of the task. We recommend that the choice of line symbolization should be guided by a task taxonomy which end users are expected to perform.

In order to share the designs and data sets, we developed an interactive flow mapping application (<http://tinyurl.com/commodityflows>) that allows users to dynamically adjust a variety of flow map symbolization (including the Bézier curve designs evaluated in this study) and visualize data sets of commodity flows (Figure 10).

Future work

In this experiment, we evaluated five prominent flow line symbolizations. However, one can combine different visual variables to obtain designs that could perform better or worse. Thanks to its divergent color scheme and line shortening to reduce crossings of flows, FA  is potentially useful for tasks that emphasizes origins or destinations such as Task 3 which emphasizes destinations and Task 4 which emphasizes origins. Line shortening is useful in reducing visual clutter, and many different styling of partially drawn lines are possible which may perform better than FA . For graphs with relatively short and local flows, FA  resembles DA  as line shortening is applied in proportion to flow length. We recommend future studies that analyze divided attention and compare varying lengths of line shortening to evaluate its effect on a series of flow map tasks. While TD  produced faster response times, its correctness was found to be significantly lower due to the double use of color value to depict direction and magnitude. To address directional ambiguity and benefit from line shortening, one can combine TD and TA with FA to derive the following designs: Fading Tapered (FATA)  and Fading Teardrop (FATD) .

Given the large number of possible flow map reading tasks, it is challenging to select tasks to evaluate the effectiveness and efficiency of flow maps. For a comprehensive evaluation of flow map reading, there is a need to construct a typology of patterns and visual tasks. Similar work has been done in movement pattern analysis,⁶¹ group level comprehension in graphs,^{30,31} and a multi-level typology of abstract visualization tasks.⁶²

While we focused on the influence of flow map symbolization, design decisions, and the effect of data on understanding a set of patterns in flow maps, future studies are needed to focus on higher level processes such as derivation of meaning and decision-making using flow maps. We believe that an insight-based approach that analyzes how users generate insights into flow data and visualization would be valuable.

For future work, there is also a need to assess other alternative designs which could incorporate cartographic interaction techniques such as highlighting,

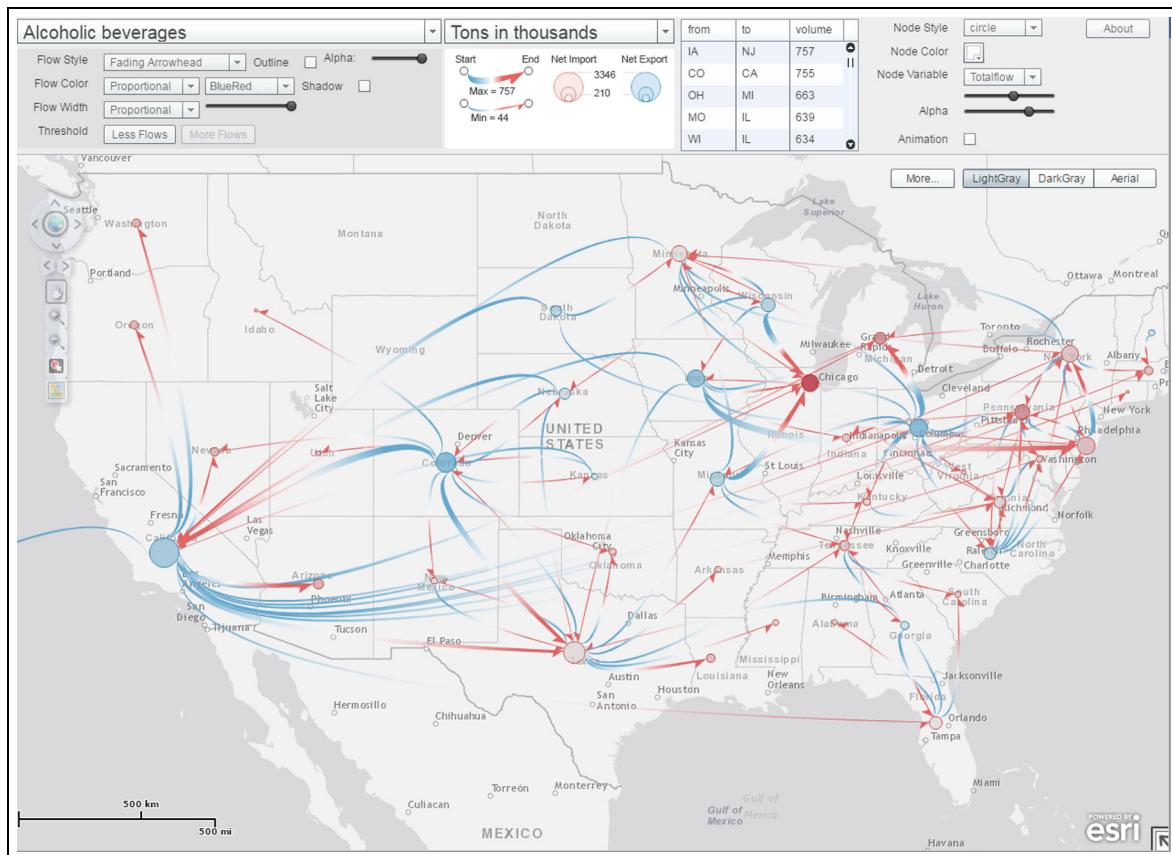


Figure 10. Commodity Flow Mapper—an interactive flow mapping application to visualize commodity flows: <http://tinyurl.com/commodityflows>.

isolations, and flow animations that could help reduce the cognitive load associated with effects such as edge tunneling, edge crossings, and crossing angles. We also recommend future studies to control and compare properties such as number of nodes, edges, and edge crossings using flow map tasks that are designed to explore holistic and geographic patterns.

Acknowledgements

The experiment in the study was approved by the Institutional Review Board of the University of Iowa with an ID number 201510743. The authors thank Mary Windsor and Sarah Battersby for their valuable feedback on various stages of our study. They also thank all the reviewers for their constructive comments.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: An earlier version of this work was in part supported by The National Science Foundation under grant no. 0748813, and The

Institute of Museum and Library Services grant no. LG-00-14-0030-14.

References

1. Koblin A. Flight patterns. *Paper presented at the ACM SIGGRAPH 2006 computer animation festival*, Boston, MA, 30 July–3 August 2006.
2. Phan D, Xiao L, Yeh R, et al. Flow map layout. *Paper presented at the IEEE symposium on information visualization*, Minneapolis, MN, 23–25 October 2005.
3. Tobler WR. Spatial interaction patterns. *J Environ Syst* 1976; 6: 271–301.
4. Alper B, Bach B, Henry Riche N, et al. Weighted graph comparison techniques for brain connectivity analysis. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Paris, 27 April–2 May 2013.
5. Battista GD, Eades P, Tamassia R, et al. *Graph drawing: algorithms for the visualization of graphs*. Upper Saddle River, NJ: Prentice Hall, 1999.
6. Dwyer T, Lee B, Fisher D, et al. A comparison of user-generated and automatic graph layouts. *IEEE T Vis Comput Gr* 2009; 15(6): 961–968.
7. Ghoniem M, Fekete J-D and Castagliola P. On the readability of graphs using node-link and matrix-based

- representations: a controlled experiment and statistical analysis. *Inf Vis* 2005; 4(2): 114–135.
8. Huang W. Using eye tracking to investigate graph layout effects. *Paper presented at 6th international Asia-Pacific symposium on the visualization, APVIS'07*, Sydney, NSW, 5–7 February 2007.
 9. Körner C. Eye movements reveal distinct search and reasoning processes in comprehension of complex graphs. *Appl Cognitive Psych* 2011; 25(6): 893–905.
 10. McIntire JP, Osesina OI, Bartley C, et al. Visualizing weighted networks: a performance comparison of adjacency matrices versus node-link diagrams. *Paper presented at the SPIE defense, security, and sensing*, Baltimore, MD, 23–27 April 2012.
 11. Purchase HC, Carrington D and Allder J-A. Empirical evaluation of aesthetics-based graph layout. *Empir Softw Eng* 2002; 7(3): 233–255.
 12. Purchase HC, Cohen RF and James MI. An experimental study of the basis for graph drawing algorithms. *J Exp Algorithm (JEA)* 1997; 2: 4.
 13. Holten D, Isenberg P, Van Wijk JJ, et al. An extended evaluation of the readability of tapered, animated, and textured directed-edge representations in node-link graphs. *Paper presented at the Pacific Visualization Symposium (PacificVis)*, Hong Kong, China, 1–4 March 2011.
 14. Holten D, Isenberg P, Fekete J-D, et al. Performance evaluation of tapered, curved, and animated directed-edge representations in node-link graphs, 2010, <https://hal.inria.fr/hal-00696823>
 15. Boyandin I, Bertini E and Lalanne D. Using flow maps to explore migrations over time. *Paper presented at the geospatial visual analytics workshop in conjunction with the 13th AGILE international conference on geographic information science*, Guimaraes, 11–14 May 2010.
 16. Fowler D and Ware C. Strokes for representing univariate vector field maps. *Paper presented at the proceedings of graphics interface*, London, ON, Canada, 19–23 June 1989.
 17. Burch M, Vehlow C, Konevtsova N, et al. Evaluating partially drawn links for directed graph edges, 2011, http://www.visus.uni-stuttgart.de/uploads/tx_vispublications/GD2011_Burch.pdf
 18. Netzel R, Burch M and Weiskopf D. Comparative eye tracking study on node-link visualizations of trajectories. *IEEE T Vis Comput Gr* 2014; 20(12): 2221–2230.
 19. Purchase HC, Hamer J, Nöllenburg M, et al. On the usability of Lombardi graph drawings. *Paper presented at the graph drawing*, Bordeaux, 23–25 September 2013.
 20. Xu K, Rooney C, Passmore P, et al. A user study on curved edges in graph visualization. *IEEE T Vis Comput Gr* 2012; 18(12): 2449–2456.
 21. Ware C and Bobrow R. Motion to support rapid interactive queries on node-link diagrams. *ACM Trans Appl Percep (TAP)* 2004; 1(1): 3–18.
 22. Slocum TA, McMaster RB, Kessler FC, et al. *Thematic cartography and geovisualization*. Upper Saddle River, NJ: Prentice Hall, 2009.
 23. Wood J, Slingsby A and Dykes J. Visualizing the dynamics of London's bicycle-hire scheme. *Cartographica* 2011; 46(4): 239–251.
 24. Ware C, Kelley JG and Pilar D. Improving the display of wind patterns and ocean currents. *B Am Meteorol Soc* 2014; 95(10): 1573–1581.
 25. Guo D. Flow mapping and multivariate visualization of large spatial interaction data. *IEEE T Vis Comput Gr* 2009; 15: 1041–1048.
 26. Fekete JD, Wang D, Niem D, et al. Overlaying graph links on treemaps. In: *Proceedings of IEEE symposium on information visualization (compendium)*, Seattle, WA, 21 October 2003.
 27. Ware C. *Information visualization: perception for design*. Amsterdam: Elsevier, 2013.
 28. Bar M and Neta M. Humans prefer curved visual objects. *Psychol Sci* 2006; 17(8): 645–648.
 29. Gill GA. Experiments in the ordered perception of coloured cartographic line symbols. *Cartographica* 1988; 25(4): 36–49.
 30. Saket B, Simonetto P and Kobourov S. Group-level graph visualization taxonomy (arXiv preprint arXiv:1403.7421), 2014, https://www.cs.arizona.edu/~kobourov/group_tax.pdf
 31. Saket B, Simonetto P, Kobourov S, et al. Node, node-link, and node-link-group diagrams: an evaluation. *IEEE T Vis Comput Gr* 2014; 20(12): 2231–2240.
 32. Gansner ER, Hu YF and Kobourov SG. Visualizing graphs and clusters as maps. *IEEE Comput Graph* 2010; 30(6): 54–66.
 33. Collins C, Penn G and Carpendale S. Bubble sets: revealing set relations with isocontours over existing visualizations. *IEEE T Vis Comput Gr* 2009; 15(6): 1009–1016.
 34. Alper B, Riche NH, Ramos G, et al. Design study of line sets, a novel set visualization technique. *IEEE T Vis Comput Gr* 2011; 17(12): 2259–2267.
 35. Buja A, Cook D and Swayne DF. Interactive high-dimensional data visualization. *J Comput Graph Stat* 1996; 5(1): 78–99.
 36. MacEachren AM, Wachowicz M, Edsall R, et al. Constructing knowledge from multivariate spatiotemporal data: integrating geographical visualization with knowledge discovery in database methods. *Int J Geogr Inf Sci* 1999; 13(4): 311–334.
 37. Shepherd I. Putting time on the map: dynamic displays in data visualization and GIS. *Innovat GIS* 1995; 2(2): 169–187.
 38. Shneiderman B. The eyes have it: a task by data type taxonomy for information visualizations. In: *Proceedings of IEEE Symposium on visual languages*, Boulder, CO, 3–6 September 1996.
 39. Holten D and Van Wijk JJ. Force-directed edge bundling for graph visualization. *Comput Graph Forum* 2009; 28(3): 983–990.
 40. Lambert A, Bourqui R and Auber D. Winding roads: routing edges into bundles. *Comput Graph Forum* 2009; 28(3): 853–862.
 41. Van de Ven B. *Algorithms for flow maps*. Eindhoven: Technische Universiteit Eindhoven, 2007.
 42. Guo D and Zhu X. Origin-destination flow data smoothing and mapping. *IEEE T Vis Comput Gr* 2014; 20: 2043–2052.

43. Becker R, Eick SG and Wilks AR. Visualizing network data. *IEEE T Vis Comput Gr* 1995; 1(1): 16–28.
44. Rusu A, Fabian AJ, Jianu R, et al. Using the gestalt principle of closure to alleviate the edge crossing problem in graph drawings. In: *Proceedings of 15th International Conference on Information Visualisation (IV)*, London, 13–15 July 2011.
45. Andrienko G, Andrienko N, Bak P, et al. A conceptual framework and taxonomy of techniques for analyzing movement. *J Visual Lang Comput* 2011; 22(3): 213–232.
46. Bertin J. *Graphics and graphic information processing*. Berlin: Walter de Gruyter, 1981.
47. Roth RE. Cartographic interaction primitives: framework and synthesis. *Cartographic J* 2012; 49(4): 376–395.
48. Wehrend S. Appendix B: taxonomy of visualization goals. Los Alamitos, CA: IEEE Computer Society Press, 1993, pp. 187–199.
49. Long L, Tucker CJ and Urton WL. Migration distances: an international comparison. *Demography* 1988; 25(4): 633–640.
50. Rogers A and Sweeney S. Measuring the spatial focus of migration patterns. *Prof Geogr* 1998; 50(2): 232–242.
51. Pandit K. Differentiating between subsystems and typologies in the analysis of migration regions: a U.S. example. *Prof Geogr* 1994; 46(3): 331–345.
52. McGrath C, Blythe J and Krackhardt D. The effect of spatial arrangement on judgments and errors in interpreting graphs. *Soc Networks* 1997; 19(3): 223–242.
53. Koylu C. *Understanding geo-social network patterns: computation, visualization, and usability*. PhD Thesis, University of South Carolina, Columbia, SC, 2014.
54. Kinkeldey C, Mason J, Klippel A, et al. Assessing the impact of design decisions on the usability of uncertainty visualization: noise annotation lines for the visual representation of attribute uncertainty. In: *Proceedings of the 26th international cartographic conference*, Dresden, 25–30 August 2013.
55. Mason W and Suri S. Conducting behavioral research on Amazon's Mechanical Turk. *Behav Res Methods* 2012; 44(1): 1–23.
56. Paolacci G, Chandler J and Ipeirotis PG. Running experiments on Amazon Mechanical Turk. *Judgm Decis Mak* 2010; 5(5): 411–419.
57. Willett W, Heer J and Agrawala M. Strategies for crowdsourcing social data analysis. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Austin, TX, 5–10 May 2012.
58. Buscher G, Cutrell E and Morris MR. What do you see when you're surfing? Using eye tracking to predict salient regions of web pages. In: *Proceedings of the SIGCHI conference on human factors in computing systems*, Boston, MA, 4–9 April 2009.
59. Corbetta M, Miezin FM, Dobmeyer S, et al. Selective and divided attention during visual discriminations of shape, color, and speed: functional anatomy by positron emission tomography. *J Neurosci* 1991; 11(8): 2383–2402.
60. Nielsen J and Loranger H. *Prioritizing web usability*. Upper Saddle River, NJ: Pearson Education, 2006.
61. Dodge S, Weibel R and Lautenschutz A-K. Towards a taxonomy of movement patterns. *Inf Vis* 2008; 7(3–4): 240–252.
62. Brehmer M and Munzner T. A multi-level typology of abstract visualization tasks. *IEEE T Vis Comput Gr* 2013; 19(12): 2376–2385.
63. Huang W and Huang M. Exploring the relative importance of number of edge crossings and size of crossing angles: a quantitative perspective. *Intl J of Advanced Intelligence* 2011; 3: 25–42.