Eurographics Conference on Visualization (EuroVis) 2017
J. Heer, T. Ropinski and J. van Wijk
(Guest Editors)

Volume 36 (2017), Number 3

Finding a Clear Path: Structuring Strategies for Visualization Sequences

Jessica Hullman,[1] Robert Kosara,[2] and Heidi Lam[2]

[1]University of Washington, Seattle WA, USA [2]Tableau Software, Seattle WA & Palo Alto CA, USA

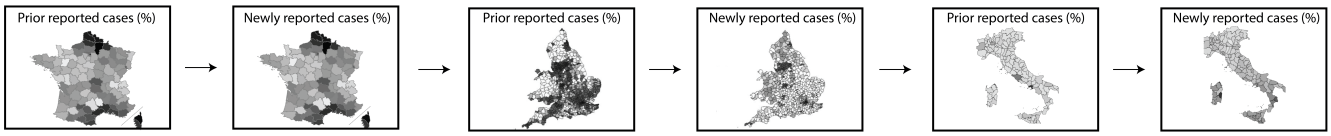

Figure 1: *Views depicting reported cases of a disease across space and time. Even for this small set of views, many sequences are possible.*

Abstract

Little is known about how people structure sets of visualizations to support sequential viewing. We contribute findings from several studies examining visualization sequencing and reception. In our first study, people made decisions between various possible structures as they ordered a set of related visualizations (consisting of either bar charts or thematic maps) into what they considered the clearest sequence for showing the data. We find that most people structure visualization sequences hierarchically: they create high level groupings based on shared data properties like time period, measure, level of aggregation, and spatial region, then order the views within these groupings. We also observe a tendency for certain types of similarities between views, like a common spatial region or aggregation level, to be seen as more appropriate categories for organizing views in a sequence than others, like a common time period or measure. In a second study, we find that viewers' perceptions of the quality and intention of different sequences are largely consistent with the perceptions of the users who created them. The understanding of sequence preferences and perceptions that emerges from our studies has implications for the development of visualization authoring tools and sequence recommendations for guided analysis.

Categories and Subject Descriptors (according to ACM CCS): H.5.2 [Information Interfaces and Presentation]: User Interfaces—Graphical User Interfaces (GUI)

1. Introduction

Visualizations are often presented in multiples. For example, data-based reports, dashboards, and many narrative visualizations consist of multiple static views which are sequentially presented to illustrate comparisons relevant to a topic or story. Fig. 1 shows a visualization consisting of six views portraying the reported cases of a preventable disease in three countries over time. These visualizations may have been created by a data analyst as part of a routine analysis for inclusion in a report, or by a visualization recommender system that automatically generates views of relational data [WMA*16]. Even for this small data set, there are multiple ways the views could be organized. For example, the countries could be grouped as in the sequence shown, and the prior cases always reported before the new cases. Or, the views depicting prior cases could be shown together in a group, followed by the views depicting the newly reported cases. Whether or not view order should be kept consistent across groups is another decision: without such consistency, it may be possible to compare adjacent views that would otherwise be separated (e.g., comparing newly reported cases in France with those in Germany if the third and fourth views in Fig. 1 were swapped). Currently, little guidance exists on what makes some sequences of related visualizations preferable to others, limiting researchers' ability to create tools that can recommend preferable sequences for designers or analysts.

In particular, how designers and viewers perceive higher-level organizations for sequences of visualization views (e.g., grouping views by time period versus the measure, etc.) remains largely unknown. Prior work in cognitive psychology shows that people gravitate toward certain visual structures (e.g., tables with time as an axis) for organizing information about agents in space and time [KT11]. Similarly, prior work in visualization sequencing has indicated that users find some types of view-to-view transitions

easier to comprehend than others [HDR*13]. However, how such preferences impact longer sequences of visualizations remains unknown. If users of visualization sets show systematic preferences for some sequential structures over others based on properties of the data that are shown (such as the measure or dependent variable, aggregation and filter level applied to data, time period, etc.), this knowledge could pave the way for sequence recommendations.

We present the results of two controlled studies to examine how people perceive sequential visualization structures. Our first study asks users to play the role of designers by sequencing sets of related visualizations (either bar charts or thematic maps) to clearly convey multivariate data. Our results indicate that *hierarchical structure* characterizes most preferred visualization sequences: participants create sequences by grouping subsets of visualizations with shared data properties, such as a common measure, time period, spatial region, or level of aggregation and filtering (Fig. 2a,b,c,d). Parallel structure – repeating a pattern of transitions two or more times in a sequence – and other hierarchical patterns that group views based on a single data factor are strongly preferred over schemes like minimizing an approximation of the *cognitive cost* (or perceived amount of difference) defined across adjacent views. Moreover, we find that groupings based on certain data properties, such as a shared spatial region or aggregation level, are much more likely to be created than others, like grouping based on shared time period or measure. In a second study focused on interpretation of visualization sequences, we investigate whether differences in preferences for sequence types among users playing the role of designers are also evident in viewers' interpretations. We find that viewers show the same preferences as authors when asked to rate the clarity of sequences. Viewers also perceive certain comparisons between views–for example, comparing different measures or time periods across the same spatial region or aggregation level–as more intentional than others–like comparing the same measure across different spatial regions or aggregation levels–independent of sequence. Our results have implications for predicting effective sequences for visualization applications like visualization recommendation tools, where more comprehensible sequences can facilitate analysis, or narrative visualization, where a novice designer might be helped by sequence suggestions during the design process.

2. Related Work

2.1. Narrative Cognition

Our studies of perceived structure in sets of visualization are inspired by the structures observed in the plots of many fictional and non-fiction stories [BB79,Gle78,MJ77,Tho77]. Mandler and Johnson [MJ77] propose that the structure of any story can be represented hierarchically, as a tree containing basic nodes: non-terminal nodes that represent groups of events. According to this model, connections between nodes on either level are causal or temporal. An alternative model posits that events in a story can be mentally indexed by various dimensions, like time, space, protagonist, causality, and intentionality. When a reader encounters discontinuities in these dimensions, more intensive cognitive processing occurs as the reader updates the appropriate indices in her mental representation [ATT97,ZLG95]. We are interested in how common data dimensions, like spatial region, time period, level of aggregation and filtering (e.g., what unit data is reported at, such as county versus state level, and how data is filtered, e.g, to only the U.S.), or measure (e.g. the dependent variable), may perform a similar "indexing" role in how people perceive structure in visualization sets.

2.2. Sequence in Visualization

Initial work on narrative visualization identified expository structures like drill-down and martini glass [SH10] and described how rhetorical devices can operate procedurally as a user navigates a visualization [HD11]. More recently, Amini et al. [ARL*15] characterized data videos using the cinematic categories from Cohn's visual narrative grammar for comics [Coh13], which analyzes how narrative tension changes over a data story. While related to sequence, we are primarily interested in models that can be defined a priori from the state space of views of a relational data set, so as to enable sequence recommendations in visualization recommender systems of narrative visualization authoring systems.

The closest work to our framing of the problem of finding good visualization sequences is Hullman et al. [HDR*13]. This scenario assumes that an analyst or designer starts with a relational data table. Possible sequences are paths through the state space of data and view combinations generated from the table. The set of possible views V represent nodes in an undirected graph G, where edges between each pair of nodes $v1$, $v2$ in G are weighted by a transition cost measure based on the number of transformations required to convert $v1$ to $v2$. Using controlled studies of peoples' preferences for different types of single visualization-to-visualization transitions, Hullman et al. [HDR*13] proposed a transition cost model that approximates the cognitive cost of moving from one visualization to the next in a sequence of static views. Transition cost is defined as the number of transformations required to convert the data shown in the first view to the second. A single transformation is defined as a change to one of the data fields shown from the first view to the second. For example, in Fig. 1, a transition from a view depicting newly reported cases of a disease in France to a view depicting prior reported cases of the disease in England would have a transition cost of 2, representing a change to the time factor and the spatial filter. The prior work also finds that when costs are equal, people are more likely to prefer certain types of transitions. When asked to choose which of two possible "next steps" representing changes to the time period, measure, dimension, or level of aggregation and filtering of an initial visualization was more understandable, subjects were most likely to change the time period, followed by changing either the dimension or measure. Subjects were least likely to change the aggregation level.

Parallel structure, also addressed by Hullman et al. [HDR*13], is a particular structure in which a visualization sequence is comprised of multiple parallel groupings in a set of views, which are presented one after another. For example, in the parallel structure sequence shown in Fig. 2a, the four groupings are defined by different spatial regions, and the measure, or dependent variable, differs between the two views in each spatial region grouping. For the parallel structure sequences shown in Fig. 2b and c, the two groupings are defined by different measures, and the spatial region differs between the views within each measure grouping. In such sequences where a single data factor is used to create groupings, we refer to

the data factor used to create chunks at a high level as the "between factor" and the data factor that varies within the chunks as the "within factor." We extend the prior work on visualization sequence by examining whether the preferences for certain transition types observed at the local (visualization-to-visualization) level can predict the higher level structures that people prefer in visualization sequences.

More recently, Kim et al. [KWHH17] presented a cognitive cost model that accounts for changes in both the data and encoding in a pair of visualizations. Motivated by parallel structure [HDR*13], their cost model incorporates hierarchical structure by prioritizing sequences of views that use parallel transitions. However, their model is agnostic as to whether certain types of transitions are more likely to be perceived by users as appropriate "between" versus "within" factors.

2.3. Sequence in Exploratory Analysis

View sequencing is implicit in some visualization recommender systems, which aim to present a manageable number of views to analysts exploring multidimensional data. In a grand tour, an analyst is shown a dynamic, seemingly continuous sequence of views (planes) of low dimensional projections [Asi85]. Projection pursuit presents a set of static low dimensional projections believed to depict interesting features of the data [FS81]. Recent work explores automated suggestion of sets of 2D views of relational data more broadly [WMA*16]. Suggested views vary data combinations (data variation) but also visual encodings (design variation). Sequencing is provided only to the extent that the top ranked visual encodings for various combinations and transformations on a set of variables selected by the user are presented at the top of the set of views. Suggesting effective paths through the recommended views could ease the analyst's interpretation process, increasing their ability to explore the search space in breadth and/or depth.

Other visualization tools allow an analyst to capture the sequence of steps they use [CFS*06, HMSA08]. However, the direct discovery sequence is not necessarily appropriate for communicating the important points in the data outside of collaborative analysis. Sequence is also implicit in interactive visualization analysis (e.g., [HS12]), but is always determined by the user's interactions. Sequence is related to, but often not made explicit, in spatial organizations of views. Andrews et al. [AEN10] found that analysts using large screens tended to group views based that contained similar items given two dimensional continuous spaces (e.g., large screens). We investigate whether similar preferences exist given sequential presentations.

3. Study 1: Structuring A Set of Views

Our first research question asks, **RQ1: To what extent are hierarchical structuring strategies preferred over non-hierarchical strategies?** We are interested in the prevalence of hierarchical structures like parallel structure and reverse parallel structure. Other "non-parallel" hierarchical structures are also possible, in which a property like shared time period or measure is used to group subsets of related views, but transitions are not consistent across groups. We hypothesize:

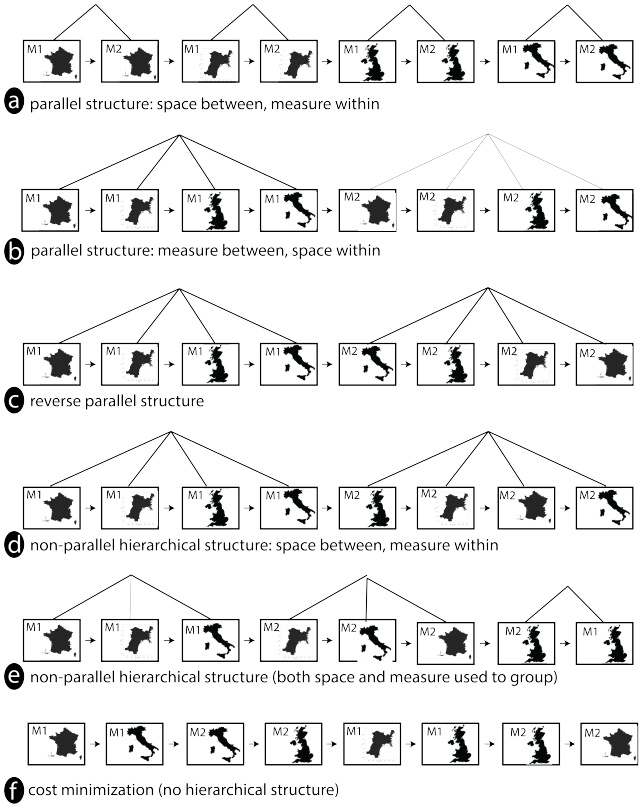

Figure 2: *Different types of sequences possible for a set of 8 views including different measures (e.g., Gross Domenstic Product (GDP) and birth rate) for four different spatial regions. A viewer could group spatial regions together (a) or measures together (b); she could keep the order of transitions the same across groupings (a, b), reverse them (c), or use inconsistent ordering within groups (d, e). Or, she can simply arrange views such that a minimum of transformations occur from one view to the next (f).*

- **H1 (Hierarchy)**: Strategies that use hierarchical structure will be preferred over other strategies.
  - **H1a**. Simple hierarchical structures, defined as sequences that use a single shared data factor (e.g., time period, spatial region, measure, or aggregation and filter level, Fig. 2a,b,c,d) to create high level groupings will be preferred over non-hierarchical or more complex hierarchical strategies (Fig. 2e,f).
  - **H1b**. Parallel structure and reverse parallel structure (Fig. 2a,b,c) will be preferred over other simple hierarchical structures (Fig. 2d).
  - **H1c**. Perfect parallel structure (Fig. 2a,b) will be preferred over reverse parallel structure (Fig. 2c).

Our second research question concerns how people choose between multiple possible global structures. If we assume that people prefer parsimony, and tend toward hierarchical structures where variation for a single data aspect defines groupings (e.g., Fig. 2a,b,c,d), then a question arises as to whether certain groupings are more likely than others. Prior work suggests preferences exist

for individual transitions (see summary of results from [HDR*13] above). Our second research questions asks **RQ2: Can grouping structures in preferred sequences be predicted from local (visualization-to-visualization) transition preferences?** Based on the transition preferences identified in prior work [HDR*13], we hypothesize:

- **H2**: People will systematically prefer to group views based on certain shared properties over others.
  - **H2a**. In visualization sets that vary aggregation and filter levels (the unit at which data is reported and how much of the data are shown), people will be more likely to group by aggregation and filter level than by time period, measure, or spatial region.
  - **H2b**. In visualization sets that vary spatial regions or measures, people will be more likely to group by spatial region or measure than by time period.
  - **H2c**. In visualization sets that vary spatial regions and measures, people will be equally likely to group based on region and measure.

Due to the importance of visual perception in visualization interpretation, one might expect that the degree of visual difference between views would impact structures such as what factors are preferred to create groupings. Using spatial encodings to present data (e.g., showing unemployment rate as color value in a choropleth map or radius in a graduated symbols map) for several different spatial regions (e.g., South American versus European) countries results in views with different looking marks (e.g., polygons) compared to a more abstract encoding such as bar length in a bar chart. Similarly, the length of a sequence may impact the sequences that are perceived as more understandable: a larger set of views may be more overwhelming for a user to understand than a shorter one, requiring more careful or different sequencing decisions. Our third research question asks **RQ3: Are preferred groupings affected by visual encoding or sequence length?**

3.1. Stimuli

Our goal was to present participants with sets of visualizations that can be organized using high level strategies like simple hierarchical structure, parallel structure, or reverse parallel structure. Structuring the visualization sets we presented should also require participants to choose between multiple groupings that differ only in the nature of the shared data property (e.g., time, spatial region, etc.). To achieve these goals, we created sets of visualizations that are analogous to "crosstabs", a common analysis for two dimensional data where the levels of two variables are crossed with one another. Each visualization set included views that varied along two of the data factors representing transition types from Hullman et al. [HDR*13]:

- **Time period**: year or range (2015 vs 2016, 1960's vs 1980's)
- **Spatial region** (Maine vs Georgia, Germany vs France)
- **Measure** (population vs unemployment, number of incidents vs cost of damage)
- **Aggregation and filter level** (data for several adjacent states aggregated by state vs data for a specific state aggregated by county)

For each pairing of data factors, we kept the other data factors in the set constant. This helped ensure that a participant who chose to group by a single data factor could also keep transitions consistent across the groupings. For example, if M1 and M2 in the views in Fig. 2 represent two measures like mean unemployment and annual net income, the two data factors are measure and spatial region. Time period and aggregation and filter level are kept consistent between views. If the participant chooses to group by spatial region (Fig.2a), they can keep the measure direction consistent if they desire, for example, transitioning from annual net income to unemployment in each spatial region grouping.

The four data factors time period, spatial region, measure, and aggregation and filter level yield six combinations of two. To investigate the impact of encoding, for each combination, we created sets that varied the visual encoding used across the set (2 levels, choropleth or graduated symbols maps or bar charts). We also varied the length of the sets (2 levels, 4 or 8 visualizations). To further ensure that our findings were not overfit to a particular data set used to generate the views, we also varied the data set (creating two sets of views using two data sets for each combination of data factor pairing, length, and encoding combination). Data sets were taken from the U.S. Census data [Bur12a], the American Community Survey [Bur12b], the World Bank Development Indicators [Ban16], and the Federal Aviation Administration's (FAA) Wildlife Strike Database [Adm16]. This produced 48 stimuli sets.

Each data factor in a pairing (e.g., time and measure) had two or more levels. For example, in stimuli sets of length 4, two time periods and two measures were available. For sets of length 8, one data factor was varied over two levels, the other over four levels. The number of levels of each data factor was randomized (not counterbalanced). The specific increment of time period, measure, spatial region, or aggregation and filter level from one view to another across a set of views was not held constant across all stimuli sets: for example, changes to the time period could include a single year change in some sets, or a change of 4 years, or a change of 10 years. However, within a stimuli set, the increment was constant. For example, in a set of 8 measure-time views where 4 views showed one measure and the other 4 views showed a second measure, the time steps would not include 1990, 2000, 2010, 2014, since the increment differs across these views (10 vs 4 years).

We generated all views using Tableau Desktop version. To reduce errors on how participants interpreted the data shown in each view, views had titles that clearly stated the measure, the spatial region, and the time period (e.g., "Unemployment in 2011 for Washington") using an identical format. To control for the influence of visual characteristics on sequencing choices, we kept all visual encodings (color scales, etc.) constant across views in a set, with the exception of our intentional manipulation of the encoding of measure values using bar length or a choropleth map. Bar charts were always sorted from greatest to least values reading left to right. A single hue was applied within and across views in the each stimuli set, but we varied the hues across stimuli sets. All quantitative color scales used in choropleth map stimuli were single hue sequential scales, with the exception of two stimuli sets where the measure was a rate that sometimes had a negative value and a two-hue di-

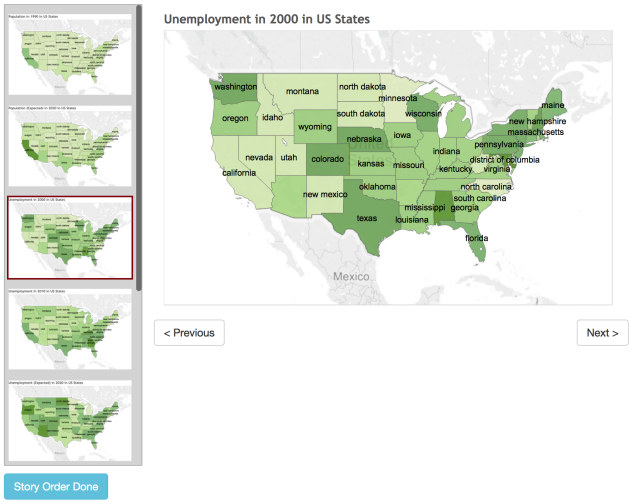

Figure 3: *The user interface used in the study. Users were able to reorder frames in the list on the left and flip through the set with the Previous and Next buttons.*

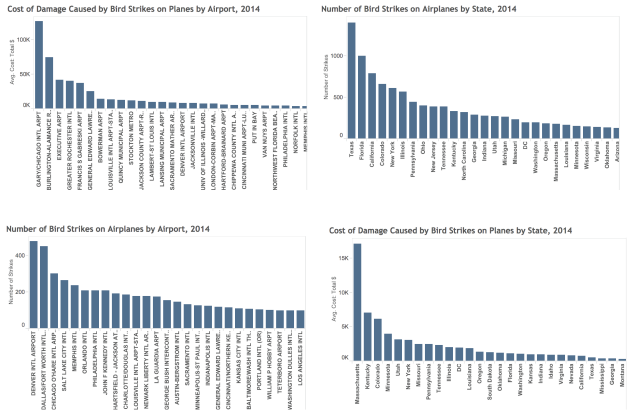

Figure 4: *Example 4-frame stimulus set for measure and hierarchy.*

verging color scale was used instead. All stimuli are available in supplemental material.

3.2. Procedure

We conducted the study as a repeated measures between-subjects experiment. Each participant saw twelve trials, including each combination of the six data factors twice, once with four and once with eight views. For each combination, each participant saw one set of stimuli that used a map encoding and one set that used a bar encoding. Assignment of stimuli sets to participants was otherwise random. Trials were presented in random order.

Each trial consisted of one set of visualization stimuli, which were vertically presented in random order on one side of the interface (Figure 3). For each trial, the participant was asked to arrange the views so that "they give a clear idea of the data." We avoided the word "story" so as to not evoke connotations commonly associated with stories (such as protagonists or setting) that might confuse participants. Our study prompt intentionally left the interpretation of a clear sequence to the participants in order to see whether people display systematic preferences for how to sequence certain data sets in the absence of a specific narrative or analysis goal.

Participants reordered the views by dragging and dropping into an order, clicking a button to indicate when they were satisfied with their ordering. After each trial, participants were asked to give a brief explanation how they picked the ordering. At the end of the study, participants filled out a demographic survey of ten-year age group, gender, highest level of education, and how often they used data visualizations. For each trial, we recorded the initial and final orders, the time it took the participant to reorder the views, the number of reorderings and the explanation.

We advertised the study through email and on Mechanical Turk. Emails were sent out to students and faculty via an HCI-related email list at a large university and to professionals working at a visualization software company. We expected these populations to be familiar with visualization presentations while not necessarily representing expert designers themselves. Participants were entered into a raffle for $100 gift cards for every 50 people who took part in the study. Workers on Mechanical Turk with an approval rating of 95%+ were paid $5 to participate in the study.

4. Results

124 participants took part in the study. 80 participants had been recruited via email, and 44 via Mechanical Turk. 70 participants (56%) were female, 49 (40%) male, and 5 (4%) answered other.

36 email participants (45%) had a bachelor's degree, 37 (46%) a master's degree, 6 (8%) a Ph.D., and only 1 (1%) only high school education. 25 Mechanical Turk participants (57%) had completed high school, and 19 (43%) had a bachelor's degree. We therefore conducted all core results analyses separately by sample to confirm that the patterns we identify are consistent for the separate populations.

We collected 1488 total sequences (12 per participant). The mean and median number of participants per stimuli set across the 48 sets was 31 (minimum 19 participants, maximum 43 participants). However, we omitted 36 sequences from analysis that were created for two stimuli sets of length 4 that combined spatial region and aggregation and filter level, due to a labeling issue with the views that prevented recording the sequences that participants had intended. Participants completed the study in just under 30 minutes (median: 19), reordering an average of 6.5 times before submitting (4 view sequences: 3.3. reorders, 8 view sequences: 9.8). The minimum number of reorders was 1, indicating that all participants reordered the views at least once before submitting.

4.1. Hierarchy

H1a posits that using high level groupings based on a single data factor will be preferred over non-hierarchical strategies or more complex hierarchical strategies that group based on multiple data

factors. Overall, we find that 1297 of 1452 (89%) of total sequences use a single factor to group views (8 views: 588, 4 views: 681). We can compare these observed frequencies to the expected frequencies of such sequences assuming that sequencing choices are random. Due to the limited ordering possibilities with only 4 views, the majority of sequences are expected to be grouped even if we assume random sequencing: we would expect two thirds (66%, or 16 out of 4!) to have a clear grouping. However, we observe a much higher rate than expected: 96% of the 4 view sequences created used a single between factor to group views.We also see a relatively high frequency of 8 view sequences that use a simple hierarchical structure (79%, or 588 out of 744 total). Assuming random sequencing, we would expect a much smaller 4% of the 8 view sequences to show grouping by a single factor (including either of the two possible factors).

H1b states that parallel structure (e.g., 2a,b) and reverse parallel structure (e.g., 2c) will be preferred over non parallel hierarchical structure (e.g., 2d). We observed 1044 total sequences (72%) that used either form of parallel structure. Of sequences of length 4, 641 used perfect parallel structure, while only 57 used reverse parallel structure (698 or 99% of 4 view sequences). Of sequences of length 8, 343 used perfect parallel structure while only 3 used reverse parallel structure (346 or 47% of 8 view sequences). These results, (presented in Table 1), strongly support H1c, which states that parallel structure will be preferred over reverse parallel structure.

The predicted rate at which sequences of length 4 would use a form of parallel structure is 67%. This equates to an expected total number of perfect and reverse parallel results in our results set of 472, but we saw 698 (99%), for a factor of about 1.5x. If we consider only the predicted (236) and observed (641) perfect parallel sequences for 4 view sequences, we would expect 33%, but we see 91%, 2.7x more than expected. However, if we consider only the predicted (236) and observed (57) reverse parallel sequences for 4 view sequences, we would again expect 33%, but we see only 8.1%. Similarly, the predicted rate at which sequences of length 8 would use a form of parallel structure if sequencing choices were random was predicted to be 0.2% (192 out of $8! = 40320$). If participants randomly sequenced views, we would predict seeing only 0.4% (approximately 3 sequences of 744) to be perfect or reverse parallel sequences across all 8-view results, but we actually observed 346 (47%), a proportion that is larger by a factor of about 94x. If we consider only the predicted perfect parallel sequences expected for 8 view sequences (0.2% or 1.5) versus the observed (46% or 343), we see 190x more than expected. Again, however if we consider only the predicted reverse parallel sequences expected for 8 view sequences (0.2% or 1.5) versus the observed (0.004% or 3), we see only 2x as many as expected. These results strongly support H1c, which states that parallel structure will be preferred over reverse parallel structure.

As a final comparison, we also manually generated all possible lowest cost sequences for each stimuli set (for example, all possible reverse parallel sequences as well as all possible sequences like that in Fig. 2f). This resulted in 437 cost minimizing sequences (64 4-view sequences, 373 8-view sequences). We calculate the number of participants who used a lowest cost sequence to be 5% (73/1452) overall, 9% (68/744) for 4 view sets, and 0.6% (4/708) for 8 view sets. These results indicate the prevalence of hierarchical strategies when sequencing visualization sets.

Structure	Obs-4	Exp-4	Obs-8	Exp-8
Hierarchical, Single Between-Factor	$\frac{681}{708}$ (96%)	$\frac{16}{24}$ (67%)	$\frac{588}{744}$ (79%)	$\frac{1536}{40320}$ (3.8%)
Perfect-Parallel	$\frac{641}{708}$ (91%)	$\frac{8}{24}$ (33%)	$\frac{343}{744}$ (46%)	$\frac{96}{40320}$ (0.2%)
Reverse-Parallel	$\frac{57}{708}$ (8.1%)	$\frac{8}{24}$ (33%)	$\frac{3}{744}$ (0.4%)	$\frac{96}{40320}$ (0.2%)

Table 1: *Observed proportions of sequences for 4- and 8-view sequences with various structures, compared to expected proportion given assumption of random sequencing. Highlighted cells indicate that we see more of that structure than would be expected if structures were equally likely.*

4 Views		**Between**		
Within	**Agg./Filter**	**Measure**	**Space**	**Time**
Agg./Filter		41 (29%)	66 (76%)	46 (36%)
Measure	100 (70%)		75 (68%)	66 (59%)
Space	19 (22%)	36 (32%)		29 (22%)
Time	79 (62%)	42 (38%)	99 (77%)	

8 Views		**Between**		
Within	**Agg./Filter**	**Measure**	**Space**	**Time**
Agg./Filter		23 (22%)	53 (58%)	45 (48%)
Measure	46 (44%)		76 (55%)	34 (33%)
Space	25 (27%)	15 (11%)		23 (19%)
Time	59 (63%)	68 (66%)	84 (70%)	

Table 2: *Structuring factors by total number and percentage (of total sequences created for those stimuli sets) for visualization sets with 4 (top) and 8 (bottom) views. Highlighted cells indicate that we see more sequences that use that between-factor (column header) than would be expected if both structures were equally likely.*

4.2. Grouping Views by Shared Data Properties

H2 states that the groupings people use will be predictable from the preferences observed by Hullman et al. [HDR*13] for certain types of visualization-to-visualization transitions. Table 2 presents the raw counts and rates of sequences that used different between and within factors. Rather than seeing proportions close to 50% for most combinations of factors, Table 2 instead indicates preferences toward grouping by one of the two factors over the other.

To analyze our results while accounting for the repeated measures design and nature of the task, we used multinomial logit models as implemented in R's mlogit [Cro11]. The multinomial logit is commonly used to analyze forced choice experiment data where subjects complete multiple trials with varying sets of alternatives. Applied to our results, the multinomial logit model accounts for the repeated measures design as well as how often a given pair of factors appeared in a trial. For each stimuli set, we included in the analysis all sequences that use a hierarchical structure based on a

Reference	Agg./Filter	Measure	Space
Agg./Filter	Reference	-0.51 (0.10)***	0.72 (0.11)***
Measure	0.51 (0.10)***	Reference	1.23 (0.11)***
Space	-0.72 (0.11)***	-1.23 (0.11)***	Reference
Time	0.42 (0.10)***	-0.08 (0.10)	1.14 (0.11)***
L.R. Test	($\tilde{\chi}^2$ = 1873.1)***		
McFadden R^2	0.54		

Table 3: *Multinomial Logit Regressing "Between Factor Choice" on factor types. Coefficients and standard errors (in parentheses) are reported. *** denotes significance at α=0.001. L.R. test describes likelihood that at least one of the predictors' regression coefficient is not equal to zero.*

single grouping factor (89% of total sequences). We regressed "between factor choice" (a binary variable indicating whether a factor was chosen to chunk views at the higher level) on the factors "measure", "space", "time", and "aggregation/filter" to distinguish how the probability of choosing to group by that factor is influenced by the alternative factor for that stimuli set. We also included a dummy variable called "present" in the model included to account for the constrained set of alternatives available in a trial (omitted from Table 3 as the coefficient in not interpretable). The reported models in Table 3, which present estimated intercepts and standard error for each factor, differ only in which possible factor is set to the omitted reference category.

Each coefficient is relative to a baseline of 0 representing the reference category. Hence, a negative value means that relative to the reference category, participants will be less likely to group by that factor, while a positive value means that relative to the reference category, participants will be more likely to group by that factor. Our results do not clearly support H2a, that aggregation and filter level groupings will be preferred in all visualization sets that pair aggregation and filter level with either time, measure, or spatial region. The logit results indicate that the spatial groupings are much more likely to be chosen than aggregation and filter based groupings. However, aggregation and filter based groupings are more likely to be created than measure groupings, and more likely to be created than time groupings.

Our results provide partial support for H2b, that measure or spatial groupings will be preferred over time-based grouping in sets that combine time with one of these factors. The logit results indicate that spatial groupings are more commonly used than time groupings. However, we do not see evidence of a preference either way when stimuli included time and measure as factors.

The results do not support H2c, that visualization sets that include different spatial regions and measures will show no dominant pattern in which is the preferred grouping factor. Instead, we find that grouping by spatial regions is much more commonly preferred than grouping by measure.

Participants in the study viewed a combination of map and non-map stimuli. It is possible that the prevalence of spatial groupings was in part due to participants being primed to think about space based on the map stimuli. To ensure that the map basis of half of our stimuli did not prime respondents to pay particular attention to spatial groupings, we reran our multinomial logit model on only bar encoding visualization sets that were structured by participants who had not yet encountered any map encoding visualization sets (79 total trials completed by 42 workers). We observe the same pattern of results for all comparisons of grouping factors, and continue to see a significantly stronger tendency ($p < 0.01$) among participants to group based on space compared to measure or time. Full results for this analysis are available in supplemental material.

To evaluate whether peoples' tendencies to use certain types of groupings are affected by visual encoding (RQ3), we ran a separate multinomial logit identical to those in Table 3, but added encoding as a dummy variable. When encoding is included, we see that a map encoding results in an increased tendency to group by space over measure, as well as to group by space over time.

To evaluate whether peoples' tendencies to use certain types of groupings are affected by the length of the sequence, we similarly ran a separate multinomial logit where we added the sequence length as a dummy variable. Overall, we find that the preferences for certain types of groupings that are significant in Table 3 all still hold, with one exception. When we account for the sequence length, we observe that the coefficient for grouping by time as compared to measure is positive and significant. However, people are more likely to group by measure when 8 views are presented. Full results for both analyses are included in supplemental material.

4.3. Preferred Transition Orders

It is reasonable to expect that people may prefer certain orders within and across sets of views that vary along a single factor. For example, forward (chronological) time is likely to be more common than backward. We calculated the total number of sequences that used various directions in time, space (e.g., East to West), and aggregation and filter levels to order views, either across or within groups. We did not include sequences that did not use consistent directions, such as reverse parallel structure sequences (Fig. 2c), in any of our counts for prevalence of directions (approximately 30% sequences did not use consistent directions).

For four view sequences that included time as a factor, 80% used forward time progression, versus 19% who used backward progression of time. Only 1% of four view sequences that included time as a factor did not use a consistent direction (e.g., instead used reverse parallel structure). For eight view sequences, where more orderings are possible, a total of 81% used a consistent forward or backward progression, with 75% of these using forward progression versus 6% using backward progression.

For four view sequences that included aggregation and filter as a factor, 72% progressed from more aggregated and less filtered to less aggregated and more filtered, versus 23% who used the reverse progression. For eight view sequences however, 77% used either direction, with 62% of these using progression from more aggregation and less filtering to less aggregation and more filtering, versus 15% for progression from less to more aggregation.

We did not observe any clear majority directions for ordering spatial regions, such as East to West or North to South.

4.4. Participants' Strategies

To gain insight into how participants made decisions in structuring the views, we qualitatively analyzed explanations they provided for their chosen sequences. Two of the authors first examined a random sample of 50 explanations and developed a coding scheme to characterize participants' reasons for their between-factor and within-factor choices. One of the authors then coded 50% of the explanations using the developed scheme. The second coder confirmed agreement with a random sample of the coded explanations.

For between-factor choices, we noted two distinct strategies in sequence construction: top-down and bottom-up. Approximately two-thirds of of the coded explanations described a top-down strategy where the participants determined the between factor directly after examining the views, then applied it to all views. The remaining one-third of participants described a bottom-up strategy where perceived relationships between individual views drove their choice of larger groupings.

We illustrate the two strategies with examples from a stimuli set where the participants were asked to sequence a series of bar charts showing either life expectancy or mobile phone usage for one year in a number of countries. An example of a top-down strategy is "First of all, align them according to the year. Then, always make the mobile phone usage first, and the life expectancy next," while an example of a bottom-up strategy is "I grouped this by showing the growth of mobile phone ownership by year chronologically (earliest to latest), and then by life expectancy by year to maybe show a correlation between phone ownership and increased life expectancy.". Even though in both cases, the time factor appears to drive the participants' decision, the sequence that resulted from the top-down strategy uses time as the between factor while the sequence that resulted from the bottom-up strategy uses time as the within factor.

For both between-factor and within-factor choices, we observed a variety of considerations in the coded explanations, both subjective and objective. Reasons cited for a chosen an ordering or grouping of views included perceived importance (e.g., "I went with the states over the planes because I think that is more relevant than specific plane companies."), perceptual similarity (e.g., "I determined the order of measures such that it kept similar visual patterns together"), convention (e.g., ordering regions alphabetically or by cardinal directions, presenting what is perceived as an independent variable before a dependent variable), data inference (e.g., "I wanted to show the delays by state. Then I focused on what caused those delays, which were the bird strikes."), and even personal reasons (e.g., "west before east since I am from the west.").

5. Study 2: Interpreting A Set of Views

A primary finding of Study 1 is that preferences for groupings based on certain data similarities (e.g., spatial region or aggregation and filter level) emerge when people are asked to sequence views with different combinations of data factors. If these tendencies are in fact evidence that certain sequences are more comprehensible to viewers, than the findings could inform the development of sequence recommendations in authoring tools and visualization recommender systems. To check whether the perceptions we observed among sequence "authors" also hold for viewers, we conducted a follow-up study. We presented a new set of viewers with sequences from Study 1 and asked them to judge how effectively each sequence portrayed the data, and what comparisons between views they thought the author intended to convey.

5.1. Stimuli

We limited our stimuli sets to those with four views. Sets with four views afforded a more controlled comparison because the number of views in each between-factor grouping in a four view sequence is always consistent (two) but can vary in eight view sequences (e.g., two groups of four or four groups of two). For each combination of data factors that we varied in Study 1, we selected two sequences where the between- and within-factor were swapped. For example, for the combination of measure and hierarchy, we tested two sequences of 4 maps depicting population growth and unemployment with different aggregation and filter levels applied to data on the northeastern U.S. One sequence grouped more aggregated, less filtered views for both variables, followed by less aggregated, more filtered views for the variables. The second sequence grouped views with the measure (population growth or unemployment) as the between factor, then ordered the views from more aggregated to less aggregated within each measure-based group. Both sequences used perfect parallel structure, as this structure was most prevalent among sequences created in Study 1. The two sequences for each combination of data factors also always used the most common ordering of measures observed in Study 1 for that visualization set (e.g., chronological time, or more aggregated to less aggregated). All stimuli are available in supplemental material.

5.2. Procedure

We used a repeated measures between-subjects design. Each participant again saw twelve trials, which we found to be a manageable number for a single Mechanical Turk HIT. Each trial consisted of one visualization sequence, which appeared in an interface identical to that shown in Fig. 3. The four views appeared vertically along the left side of the screen, but could be paged through to examine one by one. On a first screen, participants were told to examine the views and think about the data. We required participants to page through all views before continuing. On the following screen, the participant was asked to rate "how clear and effective the ordering of the visualizations is for conveying the data" using a 5pt Likert scale where 1 was "Very Ineffective" and 5 was "Very Effective."

Because we were also interested in how viewers perceive the author's intention in visualization sequences, participants were asked to describe what they thought to be the author's goal in the form of intended comparisons. We presented a set of checkboxes that included all pairwise combinations of views (e.g., "View 1 and View 2", "View 1 and View 3", etc.). Participants were encouraged to select the boxes for all comparisons that they believed thought were intended. Our goal in including this question was to enable examination of how viewers' perceptions of the intended comparisons

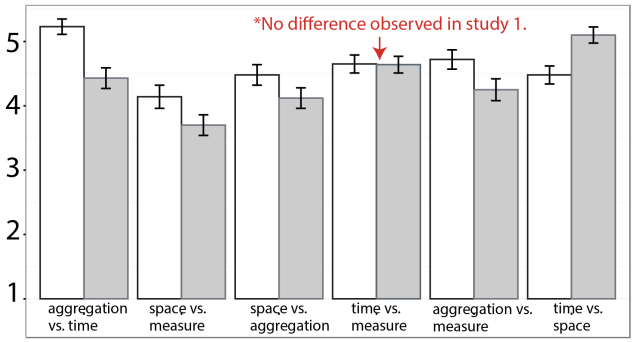

Figure 5: *Mean ratings of sequence effectiveness from viewers with 95% confidence intervals. For each combination of factors, the left-most bar represents the sequence where the first factor in the label is the between-factor. We see the same pattern of preferences observed in Study 1 (as summarized in Table 3), with observable differences between all pairs of sequences except when using time versus measure as the between-factor.*

differed by sequence. We also asked the participant to describe what he or she perceived to be the author's goal using a text box.

We advertised the study to Mechanical Turk workers with an approval rating of 95%+ and reward of $3.00.

5.3. Results

64 participants completed the study. Participants completed the study in just under 7.5 minutes on average (37s per sequence).

Figure 5 shows mean ratings with 95% confidence intervals for each combination of data factors. These results are highly consistent with Study 1, despite the fact that study 2 asks viewers to examine sequences as end-users only, rather than create them. We observe the same direction of preferences for certain groupings among viewers that we observed in Study 1. Spatial region is the preferred between-factor in all sequences that involve varying spatial regions. Aggregation and filter level is preferred when sequences vary aggregation and filter level and either time period or measure. Participants again showed no preference when presented with sequences with views that varied in time period and measure, suggesting that neither of these factors is perceived as more dominant as a means of forming groupings in a sequence.

Participants' perceptions of authors' intentions provide unique insight into how the properties of data that is shown in a set of visualizations can dominate over the impact of a particular sequence. Participants' specified 2 comparisons per sequence on average. In contrast to participants' clearly preferring comparisons between consecutive views in a sequence, as might be expected if sequence has a strong impact on what connections viewers draw between presented data, the most commonly selected comparisons were the same for both sequences for a stimuli set. Additionally, those comparisons that were chosen for a given stimuli set matched the preferences for between-factors observed in the ratings data and in Study 1's results. That is, for sequences that involve varying spatial regions, the two most prevalent comparisons among responses compared the views that used the same spatial regions to one another. For sequences involving varying aggregation and filter levels and either measures or time periods, the two most prevalent comparisons compared the views that used the same aggregation and filter level to one another. Only in the case of time and measure stimuli sets did we observe a difference in the two most prevalent comparisons based on the sequence that was shown: for this stimuli set, participants who saw a sequence where time was the between factor were slightly more likely to choose comparisons across the same time period over other comparisons, and participants who saw a sequence where measure was the between factor were more likely to choose comparisons across the same measure over other comparisons. Results are available in supplemental material.

Participants' free text explanations tended to align with their responses for the intended comparison question. The most common explanation type provided described what types of comparisons they felt the author intended, such as changes over time and increases or decreases in a measure between regions.

6. Discussion

The results of Study 1 indicate that people tend toward using certain patterns when asked to find a clear way to present data depicted in a set of visualizations. Like narratives and sequences of actions and events, visualization sequences are often structured hierarchically into groups of views with shared characteristics. Structures that are simpler (i.e., contain few, more homogeneous groupings) and more consistent (e.g., parallel transitions within groupings) dominated the sequences that people created.

We found that when visual encodings were held constant, shared spatial regions and aggregation and filter levels were more likely to be used to group views than shared measures or time periods. This suggests that people perceive variations along some data dimensions as more separable or "discrete" than variations along other dimensions. These results held when we changed the visual encoding used in the visualizations and the number of views. Our finding that time is typically used as a within-factor, rather than to construct groupings at a high level, aligns with results from studies that ask people to organize or interpret entities in time and space. Given such tasks, time is often interpreted as a continuum or "path" relative to these other factors [KT11, ZT99]. Visualizations, however, bring additional types of categories that are unique to data compared to traditional psychology studies. For example, sets of visualizations can be classified in relation to dependent variables (or measures), independent variables, the complexity or appropriateness of various statistical aggregations or transformations, visual encoding relationships, and others. We find that variation in the measure presented across a set of views behaves like time, in that it is rarely used to construct groupings at a high level. The results of Study 2 on interpretation confirm that some data factors are perceived as more appropriate for grouping views among viewers of visualization sequences as well. The results of Study 2's intended comparison question provide additional insight into the extent to which people see some types of groupings as more intended than others: participants tended to name the same several intended comparisons independent of how a set of stimuli were ordered.

These results can inform the development of algorithms for iden-

tifying effective sequences, such as in visualization recommenders for exploratory analysis. One implication of our results for the graph-based representation of the state space of views proposed by Hullman et al. [HDR*13] is the need for globally-based constraints in searching the graph. This is suggested by study participants' observed tendencies to structure views hierarchically and to maintain global consistency through patterns like parallel structure, How to implement such constraints effectively for sets of views of varying size remains an open problem. The cost model introduced by Kim et al. [KWHH17] provides an example of a generic constraint used to prioritize parallel structure, but does not differentiate between different between-factors. The preferences we observe for certain types of data-derived categories (space, aggregation and filter level) to be used more often than others (time, measure) for grouping could be implemented in an path prediction algorithm that prioritizes sequences with certain clustering profiles (e.g., as found using graph partitioning/community finding, e.g., [For10] techniques). For example, possible clusterings (i.e., sequences) for a given set of views could be ranked in terms of a cluster cohesiveness measure that accounts for the nature of the data differences between the clusters (e.g., clusters based on shared spatial regions would have higher cohesion than clusters based on shared time periods), as well as other definitions of cohesiveness of a set of views like the similarity of encodings, transformations, and other factors.

Variations along some data dimensions are subject to ordering expectations. Time steps are much more likely to be presented in forward progression. Levels of aggregation or amounts of filtering are more likely to proceed from more aggregated or less filtered to less aggregated or more filtered. The implication of these results for modeling sequence using a graph-based representation of the state space is that some edges, specifically between time periods and levels of aggregation or filtering, are best characterized as directed.

Though top-down structuring was most often used to find a clear order, this does not mean that low level differences between pairs of views are not considered. For example, choropleth maps increased the likelihood that participants would group views using a single factor, and in particular a shared spatial region, over bar charts. This suggests that bottom-up influences are also at work, as we would expect such groupings if the visual similarity between pairs of views influences the grouping structure. Additionally, some participants' explanations indicated that the sequences that they created were influenced by the specific data, such as ordering the views to show those with the most prominent difference between time periods first. An important step for future work is to conduct more controlled studies of the impacts of different marks, encodings, distributional properties, and other forms of view similarities on perceived structure in sets of visualizations.

6.1. Limitations

Our study provided a controlled setting in which participants did not receive information on the intended message or "story" that the views should convey. This assumption is realistic in the two settings we target: visualization recommenders for exploratory data analysis, and systems that provide design suggestions as an author transitions from exploratory analysis to visualization presentation authoring. However, we expect that a priori context will impact which structures are most effective in some narrative scenarios, for example, for supporting desired comparisons to make a point.

Our study focused on how people create sequences they believe are clear, and how viewers perceive the clarity and effectiveness of sequences, but did not evaluate whether performance differed for those sequences in a specific application. Studying how sequence impacts data interpretation is important for future work.

We did not systematically vary the number of views that could be grouped by the two possible factors in each stimuli set. For example, each set of 8 views allowed two groupings of four by one of the data factors or four groupings of two by the other factor. It is possible that the size of a subset defined by a data factor (e.g., how many different spatial region groups could be created) impacts peoples' choices. Similarly, our studies constrained views to depicting limited subsets of variables at a time (e.g., unemployment rate by county) using either spatial or abstract length encodings. We intentionally controlled other aspects of the views, like transformations applied to variables (e.g., log scaling), color encodings, and the formatting of the titles (e.g., always listing the year last). The visualization sets that our study participants reordered represented relatively balanced and controlled summaries of a relational data set. Future work should systematically investigate how people respond given more disparate sets of visualizations.

Our stimuli varied aggregation and filter level together (e.g., changing the reporting unit, such as from state to county, while increasing the filter level, e.g., from the midwestern U.S. to the state of Michigan). We made this choice based on the frequency with which these two properties of views tend to covary in actual visualizations. However, future work should differentiate the impact of these and other design factors on the sequences people prefer.

7. Conclusion

Identifying structuring principles for visualization sequences is valuable as InfoVis systems move toward recommending sets of views for analysis and presentation. We present two studies aimed at understanding what high-level patterns dominate what people perceive as the clearest sequences for presenting sets of related visualizations. In a first study, users played the role of designers, ordering sets of visualizations to give a clear depiction of the data. Most sequences that participants produced used hierarchical structure of some type, typically by using a single data factor to group views into two more chunks. Shared spatial region was preferred for grouping, though levels of aggregation were also common. Shared measures or time periods were not commonly used to chunk views, suggesting that people perceive differences along these factors more as continua along which comparisons can be made between specific views than as categories. These preferences were also prevalent among viewers who were asked to evaluate the clarity and effectiveness of sequences in a second study. Our results suggest representing transitions in the space of possible sequences for a set of views as directed edges, and accounting for preferred groupings within sequences and recurrent edge motifs (e.g. parallel structure in transitions) to enable visualization systems to predict paths that mimic those people construct to present data clearly.

References

[Adm16] ADMINISTRATION F. A.: F.a.a. wildlife strike database, accessed March 2016. URL: http://wildlife.faa.gov/default.aspx.

[AEN10] ANDREWS C., ENDERT A., NORTH C.: Space to think: large high-resolution displays for sensemaking. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (2010), ACM, pp. 55–64.

[ARL*15] AMINI F., RICHE N. H., LEE B., HURTER C., IRANI P.: Understanding data videos: Looking at narrative visualization through the cinematography lens. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (2015), ACM, pp. 1459–1468.

[Asi85] ASIMOV D.: The grand tour: a tool for viewing multidimensional data. *SIAM journal on scientific and statistical computing 6*, 1 (1985), 128–143.

[ATT97] A. TAYLOR H., TVERSKY B.: Indexing events in memory: Evidence for index dominance. *Memory 5*, 4 (1997), 509–542.

[Ban16] BANK W.: Development indicators. , accessed March 2016.

[BB79] BLACK J. B., BOWER G. H.: Episodes as chunks in narrative memory. *Journal of Verbal Learning and Verbal Behavior 18*, 3 (1979), 309–318.

[Bur12a] BUREAU U. C.: Census bureau data, 1990 - 2012. URL: http://factfinder2.census.gov.

[Bur12b] BUREAU U. C.: American community survey, 2012. URL: http://factfinder2.census.gov.

[CFS*06] CALLAHAN S. P., FREIRE J., SANTOS E., SCHEIDEGGER C. E., SILVA C. T., VO H. T.: Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD international conference on Management of data* (2006), ACM, pp. 745–747.

[Coh13] COHN N.: Visual narrative structure. *Cognitive science 37*, 3 (2013), 413–452.

[Cro11] CROISSANT Y.: Estimation of multinomial logit models in r: The mlogit package. URL: http://cran.r-project.org/web/packages/mlogit/vignettes/mlogit.pdf.

[For10] FORTUNATO S.: Community detection in graphs. *Physics reports 486*, 3 (2010), 75–174.

[FS81] FRIEDMAN J. H., STUETZLE W.: Projection pursuit regression. *Journal of the American statistical Association 76*, 376 (1981), 817–823.

[Gle78] GLENN C. G.: The role of episodic structure and of story length in children's recall of simple stories. *Journal of verbal learning and verbal Behavior 17*, 2 (1978), 229–247.

[HD11] HULLMAN J., DIAKOPOULOS N.: Visualization rhetoric: Framing effects in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on 17*, 12 (2011), 2231–2240.

[HDR*13] HULLMAN J., DRUCKER S., RICHE N. H., LEE B., FISHER D., ADAR E.: A deeper understanding of sequence in narrative visualization. *Visualization and Computer Graphics, IEEE Transactions on 19*, 12 (2013), 2406–2415.

[HMSA08] HEER J., MACKINLAY J. D., STOLTE C., AGRAWALA M.: Graphical histories for visualization: Supporting analysis, communication, and evaluation. *Visualization and Computer Graphics, IEEE Transactions on 14*, 6 (2008), 1189–1196.

[HS12] HEER J., SHNEIDERMAN B.: Interactive dynamics for visual analysis. *Queue 10*, 2 (2012), 30.

[KT11] KESSELL A., TVERSKY B.: Visualizing space, time, and agents: production, performance, and preference. *Cognitive Processing 12*, 1 (2011), 43–52.

[KWHH17] KIM Y., WONGSUPHASAWAT K., HULLMAN J., HEER J.: Graphscape: A model for automated reasoning about visualization similarity and sequencing. In *Proceedings of the ACM Conference on Computer-Human Interaction (CHI 2017)* (2017), ACM.

[MJ77] MANDLER J. M., JOHNSON N. S.: Remembrance of things parsed: Story structure and recall. *Cognitive psychology 9*, 1 (1977), 111–151.

[SH10] SEGEL E., HEER J.: Narrative visualization: Telling stories with data. *Visualization and Computer Graphics, IEEE Transactions on 16*, 6 (2010), 1139–1148.

[Tho77] THORNDYKE P. W.: Cognitive structures in comprehension and memory of narrative discourse. *Cognitive psychology 9*, 1 (1977), 77–110.

[WMA*16] WONGSUPHASAWAT K., MORITZ D., ANAND A., MACKINLAY J., HOWE B., HEER J.: Voyager: Exploratory analysis via faceted browsing of visualization recommendations. *Visualization and Computer Graphics, IEEE Transactions on 22*, 1 (2016), 649–658.

[ZLG95] ZWAAN R. A., LANGSTON M. C., GRAESSER A. C.: The construction of situation models in narrative comprehension: An event-indexing model. *Psychological science* (1995), 292–297.

[ZT99] ZACKS J., TVERSKY B.: Bars and lines: A study of graphic communication. *Memory & Cognition 27*, 6 (1999), 1073–1079.