Patterns and Pace: Quantifying Diverse Exploration Behavior with Visualizations on the Web

Category: Research

Abstract— The diverse and vibrant ecosystem of interactive visualizations on the web presents an opportunity for researchers and practitioners to observe and analyze how everyday people interact with data visualizations. However, existing metrics of visualization interaction behavior used in research do not fully reveal the breadth of peoples' open-ended explorations with visualizations. One possible way to address this challenge is to determine high-level goals for visualization interaction metrics, and infer corresponding features from user interaction data that characterize different aspects of peoples' explorations of visualizations. In this paper, we address this challenge by identifying needs for visualization behavior analysis, and developing candidate features that can be inferred from users' interaction data with visualization. We then propose metrics that capture novel aspects of peoples' open-ended explorations, including exploration uniqueness and exploration pacing. We evaluate these metrics along with four other metrics recently proposed in visualization literature by applying them to interaction data from prior visualization studies. The results of these evaluations suggest that these new metrics 1) reveal new characteristics of peoples' use of visualizations, 2) can be used to evaluate statistical differences between visualization designs, and 3) are statistically independent of prior metrics used in visualization research. We discuss the implications of these results for future studies, including the potential for applying these metrics in visualization interaction analysis, as well as the emerging challenges in developing a design space of metrics for visualization engagement.

Index Terms—Interaction, Visualization, Quantitative Evaluation.

1 Introduction

As interactive visualizations migrate from standalone applications to web pages, their users have expanded from domain experts to the general population. Alongside this expansion of both visualization creators *and* consumers comes an expansion in the goals of both - from casual exploration to focused analysis. But do the metrics we use to assess visualizations capture this diversity in objectives and goals? In this paper, explore how the rapid development of expressive and interactive forms on the web has demanded an extension of the metric toolbox in which we equip content creators, and how we can better align assessment with the goals of the designers.

Consider an example where someone explores an interactive scatterplot visualization showing a company's profit and income. Each point represents a company, and upon mousing over a point the user will uncover the company's income over several years, the employees' age distribution, *etcetera*. A person's goals can be diverse here, ranging from specific (gathering information on a possible stock purchase) to broad (getting to know more companies). Two likely metrics to describe their behavior include *time spent on exploration* and *points visited*. These metrics could be used to answer basic questions about how an audience uses a published visualization, for example "how many points did the average person visit?" or "how long did the average person explore the visualization?". Yet despite their diversity in goals, it's possible that users visit a similar number of points and engage with the visualization for a similar amount of time. While simple metrics might not reveal differences between users, in reality, their behavior may not align with what the creator of the visualization had in mind for their audience.

Although research has made strides in designing and evaluating interaction in visualization, we lack *low-barrier, expressive* metrics that capture the breadth of user interaction with visualizations [23, 5, 6, 17]. Various approaches have been used to analyze the logs so as to answer these questions, including *statistical* and *visual* analysis approaches. Statistical approaches stand for summarizing users' interactions statistically. For example, Boy *et al.* [5] evaluated the impact of storytelling on users' interaction behavior by comparing the time of exploration and the number of actions performed by users interacting with a visualization with or without storytelling. Visual approaches stand for visualizing users' interaction logs. For example, Blascheck *et al.* [3] developed an interactive visualization system to display each user's interaction history with pattern search function. Some other works aggregated the interaction traces of different user groups, and make visual comparisons [28, 14].

However, the existing approaches, including both the statistical and visual approaches, have limitations with characterizing user explorations automatically and precisely. Many of the metrics used to summarize activity tend to over-aggregate behavior, failing to identify differences between users, or captured detailed information such as *how long* has been spent on *which data elements*. On the other hand, the visual approaches usually keep the details of users' interaction logs, but visual inspections can hardly lead to reliable inferences.

One possible way to bridge this gap is to develop metrics, *i.e.*, statistical summaries, which take into account more information in users' interaction logs, and to automatically reveal the characteristics of user explorations. Related examples can be found in the field of HCI. Chi *et al.* [9] quantified the saliency of a user's visit to a website when modeling users' information needs and actions on the web. Heer *et al.* [20] further used this measure to cluster web users. These works have inspired our work of visualization interaction analysis, in that a user's open-ended exploration of a visualization containing visual elements can be analogized to the exploration of a website containing webpages. However, it is impractical to directly adapt these methods developed to analyze website explorations, due to the differences between the website clickstream analysis and visualization interaction analysis, such as different scales (*i.e.*, usually millions of users versus tens to thousands of users) and different complexity of interaction types.

1.1 Contributions

The aims of this paper are three-fold:

1. Derive a requirements space to categorize existing and new metrics that quantify aspects of users' exploration of data visualization and to identify emerging needs.
2. Derive two new metrics centered around user exploration diversity and pacing that aim to provide new perspectives into users' open-ended exploration.
3. Evaluate both these new metrics and metrics recently proposed in visualization literature across hundreds of interaction traces from previously published visualization experiments.

We further discuss how these metrics can help both statistical and visual approaches to analyze interaction logs, such as 1) quantifying the impact of visualization designs on user behavior; 2) organizing the visual representation of interaction logs; and 3) serving as features to machine learning models.

2 BACKGROUND

2.1 Characterizing Website Explorations

One thread of research that is closely related to this paper is the website or application clickstream analysis and visualization [22, 24, 25, 37, 42], under a broader research topic of event sequence analysis [16, 26, 40]. Clickstream research includes the data processing, analysis and visualization methods to analyze users' website visit logs. For example, Liu *et al.* [25] developed algorithms to extract sequence patterns from clickstreams. Zhao *et al.* [42] created a visualization called MatrixWave to compare two clickstream datasets, and found it to scale better than commonly used sankey diagrams.

The exploration of a website is *similar* to the exploration of a web visualization in two ways: 1) a visit to a webpage can be analogized to a visit to a visual element, and 2) most explorations are open-ended, *i.e.*, with unspecified tasks. While the methods of clickstream analysis inspire the research of visualization interaction analysis, they are *not fully applicable* because of two reasons: 1) the webpages of a website are structured, which makes the website visits constrained, *i.e.*, one page is only accessible through certain pages, while the visual elements of a visualization are less structured, which makes the visits independent, *i.e.*, all elements are accessible. While most clickstream analysis methods focus on sequential features, this difference makes the sequential features less weighted in vis interaction analysis than the clickstream analysis. 2) the clickstream analysis approaches, such as pattern mining, are typically applicable to a much larger scale (millions of users) than the visualization interaction analysis (tens to thousands of users).

2.2 Characterizing Visualization Explorations

Characaterizing user behavior through interaction logs have been used for various *purposes*, such as learning user characteristics [7, 28], understanding the system usage [29] and the reasoning process [12], and evaluating visualizaton design [5, 17, 14]. Various *approaches* have also been used for these interaction analyses, including visual and statistical approaches.

2.2.1 Visual Approaches

Visual approaches refer to strategies of showing users' interaction logs with visualizations [8, 12, 40, 34]. The interaction logs can be shown in an aggregated way in order to reveal the behavioral differences of user groups in experiment analyses. For example, Ottley *et al.* [28] used aggregated maps to show different exploration patterns of tree visualizations. Users' interaction logs can also be shown individually. Blascheck *et al.* [3] introduced a visual analytic approach to study users' interactions with visual analytics. These visual approaches have the advantage of preserving the details of the interaction logs. However, visual examination alone cannot provide robust analyses of user behavior, as they are often better paired statistical approaches.

2.2.2 Statistical Approaches

Commonly used metrics to depict a user's exploration include *total exploration time*, the time spent by a user exploring a visualization, and *number of raw interactions*, the number of raw interactions performed by a user during exploration, such as hovering and clicking. Boy *et al.* [5] evaluated the effectiveness of storytelling by comparing users' exploration time and raw interaction counts (hovers and clicks) between the experimental and control groups. Liu *et al.* [23] measured the effects of latency on users' exploration behavior of visual analytics by using raw interaction counts (drag, brushing and linking, etc). There are many other works using the basic metrics to characterize users' interaction with visualizations [7, 17, 21].

However, these raw counts have limitations with delivering semantic meanings of user explorations. Interaction coding was thus used to describe interaction behavior. Boy *et al.* [6] and Guo *et al.* [17] coded the raw interactions into semantic interactions, such as selecting, filtering and inspecting, according to Yi *et al.*'s [41] visualization interaction framework. They counted the coded interactions afterwards.

Some work went beyond counting individual interactions (including raw and semantic ones), in order to reveal more characteristics of user explorations. Guo *et al.* [17] further extracted the interaction patterns (defined as a sequence of individual interactions), and then counted the number of each extracted pattern for each user, and the numbers of every pattern were used as metrics. Wall *et al.* [38] proposed six metrics to measure cognitive bias, including data coverage, data distribution and attribute coverage/distribution, etc. However, Guo *et al.* [17] also pointed out that the temporal aspect of interaction history - a dimension largely absent from their work - might contain important information.

In this work we create a feature space to categorize the existing metrics, and develop new metrics by filling the gaps in the framework, by fully utilizing the information in interaction sequences, in order to reveal more characteristics of users' visualization explorations.

3 A REQUIREMENTS SPACE FOR METRIC DEVELOPMENT

The aim of this work is to explore and evaluate metrics that characterize the diversity of peoples' explorations with interactive visualizations on the web. We therefore situate our metric development activities by deriving a set of requirements (top-down) and examining the possible common data sources (bottom-up) from which new metrics can be derived. Two questions need to be answered to achieve these goals:

1. What do we *need* to measure for behavior analysis?
2. What *can* we measure given users' interaction logs?

These questions drive two dimensions in this requirements space: 1) identifying unfilled needs for visualization behavior analysis, and 2) deriving low-level features from visualization interaction logs, (Section 3.1 and 3.2).

We form the structure of this requirements space based on O'Connell *et al.*'s [27] framework for deriving metrics measuring human interaction with interactive visualizations. In their framework, high-level needs include human-interaction heuristics, *i.e.*, measures that assess how well visualizations empower analysis, improving resulting analytic products, collaboration, ease of use, *etcetera*. O'Connell *et al.* then derive corresponding metrics for each of the heuristics. For example, to measure how well a given visualization system empowers analysis, O'Connell describes a metric based on the number of entities analysts draw into clusters.

One aim of this work is to move beyond system-specific visualization interaction metrics towards metrics that can be applied across a range of visualizations and for a range of creator goals, whether they be visualization practitioners or researchers. Our requirements space therefore expands and generalizes O'Connell *et al.*'s metric framework in two ways. First, centered on needs of visualization creators, we identify desirable avenues for visualization interaction metric development. Second, examining commonly available interaction data in web-based visualizations we derive novel features which are intended to enable new means for comparison and reasoning about how audiences interact with visualizations.

3.1 Visualization Interaction Analysis: Identifying Needs

What are unfilled needs in visualization interaction analysis?

Researchers and practitioners may choose to record and analyze user interactions with visualizations for multiple reasons, as noted by Eltayeby and Dou [13]. One common use case for collecting visualization interaction data, for example, is to use interaction data as part of visualization system evaluation, for example as was done with the visual analytics system WireVis [8]. Another goal of capturing user interaction data with visualizations is to infer aspects of users' cognitive states during exploration. For example, Wall *et al.* [38] developed metrics to measure users' cognitive bias while interacting with a visualization. Similarly, Ottley *et al.* [28] used interaction data in a user study, finding that personality factors such as the person's *locus of control* were related to underlying interaction patterns. In any case, for each of these goals, we can follow O'Connell *et al.*'s observations and develop a corresponding set of *measurement needs*, in other words, we can determine what any derived low-level metrics should reveal.

Notation	Description
$E = \{e_1, ..., e_N\}$	The set of N interactive elements in the visualization.
$U = \{u_1, ..., u_M\}$	The set of M users who explore the visualization.
$\|E(u)\|$	The number of visualization element a user u interacts with.
$A = \{click, hover...\}$	The set of available action types of interaction in the visualization.
t	The moment when an event occurs.
$I = (t, a, E_i, d)$	An interaction event, including the moment t when it occurs, the type of action $a \in A$, the set of interacted visualization elements $E_i \subseteq E$, and the duration of the interaction d.
$Ex(u) = (t_{start}, I_1, ..., I_k, t_{end})$	The interaction log of a user u exploring the visualization, including a time-series ordered sequence of events, the moment when the exploration starts t_{start}, followed by k ordered interaction events $I_1, ..., I_k$ and the end moment t_{end}.
$C(u_m, e_n)$	The count of interactions with the visualization element $e_n \in E$ by the user u_m.
$T(u_m, e_n)$	The time spent by the user u_m interacting with the vis object $e_n \in E$.

Table 1: Notations used to describe user interactions with visualizations, and to describe the metrics in this paper.

In our case, we observe an unfilled need in that existing studies of how people interact with visualizations in web-based settings tend to focus on measures such as number of items visited and time spent in exploration. To give a few examples, studies from Boy *et al.* [5, 6], Feng *et al.* [14, 15], and Wall *et al.* [38] center on these metrics. (Walls *et al.* [38] is an exception, however, in that the aim of their work was similarly to move beyond traditional metrics towards metrics that measure aspects of bias in exploration.) One other purpose is to examine users' exploration strategies. Examining *visits* and *time* in light of these recent studies raise new questions about user exploration in visualizations. For example, instead of *how many* items are visited in the visualization, what about the *diversity* of items visited? And instead of *how much* time is spent in exploration, what about the *pacing* of peoples' exploration inside visualizations?

3.2 Deriving Features from Visualization Interaction Data

What can we measure from peoples' visualization interaction data?

Working from the bottom-up, we observe that multiple candidate low-level features can be extracted from peoples' visualization interaction data. For example, Wall *et al.* [38] listed two measurable features, *types of interaction* (*e.g.* clicking and hovering) and *objects of interaction* (*i.e.* the visual elements interacted with), that can be extracted from users' interaction data and used in interaction modeling and metric development. Similarly, Blascheck *et al.* [4] point out that *the time spent for inspecting particular data items* is a widely available and useful feature to be considered in evaluating user behavior.

Based on the existing literature of visualization interaction analysis [38, 4], visualization interaction frameworks [41], as well as the related topic of website clickstream analysis [24, 18], we list several low-level interaction features that can commonly be obtained from a user's exploration session of visualizations encountered on the web:

- **type**: What type of interaction is it (*e.g.* selection, hovering, or using existing frameworks such as Yi *et al.* [41])?
- **element(s)**: Which element(s) in the visualization are users interacting with?
- **duration**: How long does the user interact with the visualization element(s)?
- **order**: In what order do the interactions take place?
- **moment**: At what moment in time does each interaction take place?
- **exploration time**: How long does the user spend exploring facets of the visualization?

To illustrate these features in practice, consider the situation where a user explores an interactive scatterplot, by hovering over each point the user can see the details of the point through a pop-up chart. When the user starts the exploration, she immediately hovers over a point A, the *type of interaction* is hovering, and the point A she interacts with is considered the *element of interaction*. She spends 10 seconds (*duration of interaction*) reading the pop-up chart showing the details of this point. After closing the details of point A, she pauses for 5 seconds, and then hovers over point B. The *order of interaction* is {A, B}, and the *moment of interaction* with point B happens at the 15th second from the start (after 10 seconds on point A and 5 seconds pause). In the end, she has spent 10 minutes in total exploring this visualization (*exploration time*). The user's exploration process of the visualization can be described using more exact notation (Table 1).

3.3 Deriving Existing Metrics from Visualization Literature

There are several basic metrics that are commonly used by researchers and practitioners to measure users' interaction with visualizations. We simplify the description of the metrics by assuming that there is only one type of interaction in the visualization:

- ***number-of-actions***[5, 23, 17]: equals to k, the number of certain type of interactions performed by the user during exploration.
- ***number-of-visited-charts*** [14]: equals to $|E(u)|$, the number of unique visualization elements interacted by the user.
- ***exploration-time*** [5, 23, 14, 15]: equals to $t_{end} - t_{start}$, the time spent by the user on the exploration.

A recent study in the visual analytics area from Wall *et al.* [38] propose metrics to measure bias in visualization exploration, by using multiple features from users' interactions with visual analytics.

bias-data-point-coverage [38] is measuring bias based on the user's coverage of the data points in the visualization. In this paper, a data point is considered equivalent to a visualization element, in that we consider a common scenario in which one data point is mapped to one element in the visualization. The metric is calculated as:

$$b_{DPc} = 1 - min(\frac{|E(u)|}{\hat{\kappa}(E(u))}, 1)$$

where $|E(u)|$ denotes the number of unique visualization elements interacted by the user, and $\hat{\kappa}(E(u))$ denotes the expected value of the number of unique data points visited in k interactions.

$$\hat{\kappa}(E(u)) = \frac{N^k - (N-1)^k}{N^{k-1}}$$

Another metric Wall *et al.* propose is ***bias-data-point-distribution*** [38], measuring bias toward repeated interactions with individual data points or subsets of the data.

$$b_{DPd} = 1 - p$$

where p is the p-value obtained from the χ^2 distribution with N - 1 degrees of freedom

$$\chi^2 = \sum_{n=1}^{N} \frac{(C(u, e_n) - \hat{C}(u, e_n))^2}{\hat{C}(u, e_n)}$$

where $\hat{C}(u, e_n) = [k/N]$.

In the next section, we propose two additional metrics that offer new perspectives on the diversity and pacing of peoples' open-ended explorations with interactive visualizations.

4 Proposed Metrics

4.1 Exploration Uniqueness Metric

The uniqueness metric is developed to quantify how unique a user's exploration pattern is compared to those from others. Low exploration-uniqueness suggests that a user's visits align with common patterns of exploration in the visualization. High exploration-uniqueness suggests that a user's exploration strategy differentiates itself from most other users.

What measurable features do we use for calculating uniqueness? We compute uniqueness based on the data distribution of a user's visits, *i.e.*, the distribution of time spent visiting the data items in the visualization. The features from users' interaction logs (Section 3.2) taken into account are the *objects* and *duration of interaction*, *i.e.*, the *order of interaction* and *exploration time* are discarded.

4.1.1 Adapting the Concept of Term Frequency-Inverse Document Frequency (TF-IDF)

We adapt *Term Frequency-Inverse Document Frequency (TF-IDF)* for the metric calculation. TF-IDF is most commonly used in Information Retrieval to reveal the **uniqueness** of words in a document collection, by calculating the value for each word in a document through an inverse proportion of the frequency of the word in a particular document to the percentage of documents the word appears in [30].

Given a document collection D, a word w, and an individual document $d \in D$, we calculate a TF-IDF (or Uniqueness) value $S_{w,d}$ for each word in a document:

$$TFIDF(w,d) = TF \times IDF = \frac{f(w,d)}{|d|} \times log(\frac{|D|}{f(w,D)}) \qquad (1)$$

where $f(w,d)$, equals the number of times w appears in d, $|d|$ is the number of words in d, $|D|$ is the size of the corpus, and $f(w,D)$, equals the number of documents in which w appears in D [33].

TF-IDF has also been used in areas outside Information Retrieval [36]. In the field of Human-Computer Interaction, it has been used to model user behavior, such as to describe the uniqueness of a user's visit to a website [9, 19, 20], and a person's visit to a geolocation [39].

In this paper, we adapt the TF-IDF concept to calculate the uniqueness metric from users' visualization interaction logs, specifically, by mapping a document to a user's interaction log exploration session, and a word to a chart in the visualization. Then a TF-IDF value reveals the uniqueness of a user's visit to a visualization element in the whole collection of users' interaction logs. The following section show the detailed steps of how this metric is calculated by adapting TF-IDF.

4.1.2 Metric Calculation Steps

The *exploration-uniqueness* metric is computed in three steps.

Step 1: Form a matrix $V_{N\times M}$ representing the distribution of visits from the M users to the N visualization elements in a collection of interaction logs. (Equation (2))

$$V_{M\times N} = \begin{pmatrix} T(u_1,e_1) & \dots & T(u_1,e_N) \\ \vdots & \ddots & \vdots \\ T(u_M,e_1) & \dots & T(u_M,e_N) \end{pmatrix} \qquad (2)$$

where each row represents a user u_m, each column represents a visualization element e_n, and each element $T(u_m,e_n)$ is the aggregated time (measured in *ms*) the user u_m spent visiting the e_n.

Step 2: For each element in the matrix $V_{M\times N}$, calculate a TFIDF value. We adapt Equation (1) to calculate the TF-IDF values, which represent how unique the visit is from each user u_m to each visualization element e_n.

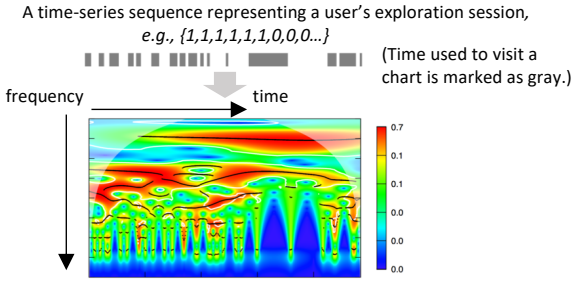

Fig. 1: A user's exploration interactions can be transformed into a time-series signal with $\{0,1\}$ representing her visiting status. The signal sequence can further be transformed to a 2D wavelet power spectrum through continuous wavelet transform.

$$TFIDF(u_m,e_n) = \frac{T(u_m,e_n)}{\sum_{i=1}^{N} T(u_m,e_i)} \times log(\frac{M}{f(e_n,Ex)}) \qquad (3)$$

where $T(u_m,e_n)$ is the aggregated time the user u_m spent visiting the element e_n, M is the total number of the users, and $f(e_n,Ex)$ denotes the number of users in the exploration collection Ex who spent time on the visualization element e_m.

To calculate the *Term Frequency* of TF-IDF, we choose to use a user's aggregated time spent on a visualization element $T(u_m,e_n)$, divided by the total time the user spent visiting all the charts. There are two alternative options, 1) to use the count of visits from a user to an element $C(u_m,e_n)$, divided by the total number of visits from the user to all the elements, and 2) to use the binary value $\{1,0\}$ to mark a user's visit to an element (1 if visited, 0 if not visited), and then divide it by the total number of elements visited by the user. We choose the time option over the other two to minimize the noise, *i.e.*, during open-ended explorations, a user might accidentally interact with a chart element, and aggregating the time spent by the user on the chart can better indicate the user's intentional visit to the chart element.

Step 3: Aggregate the uniqueness scores $Uniq$ for each user u_m.

$$Uniq(u_m) = \sum_{n=1}^{N} TFIDF(u_m,e_n) \qquad (4)$$

where $TFIDF(u_m,e_n)$ is the TF-IDF value calculated for each visit from a user u_m to a visualization element e_n, using Equation (3).

By aggregating the TF-IDF values of the visits from one user to all the chart elements, we get a metric depicting the overall uniqueness of the user's exploration. The aggregation process can omit the variation of the TF-IDF distribution, *i.e.*, a user having only one visit with extremely high TF-IDF value may have the same uniqueness metric value as another user having many visits with low TF-IDF values. However, the goal of this metric is to upweight the users who may have less but unique visits, and to downweight those who have more visits to the charts frequently visited by others.

4.2 Exploration Pacing

The pacing metric is developed to differentiate temporal strategies by users - integrating the frequency, duration, and timing of interaction with data. Higher exploration-pacing suggests a user that rapidly moves from item to item. Lower exploration-pacing might reflect a user that explore individual elements for more time.

While users may visit the same data during interaction, the duration and frequency of those visits may reflect different exploration strategies. By characterizing these differences with exploration-pacing, we may begin to quantify the depth of engagement with a visualization.

What measurable features do we use for calculating users' exploration pace? One key intuition here is that merely averaging or binning time durations of peoples' interaction with chart elements is insufficient for developing a single metric, as in-depth interactions will

be dominated by multiple shorter-duration interactions. Instead, we observe that common mathematical techniques can readily transform duration data from the *time* domain to the *frequency* domain.

We compute the pacing metric based on the distribution of a users' interaction frequency. The features from users' interaction logs (Section 3.2) taken into account are the *moments* and *duration of interaction*, and *exploration time*.

4.2.1 Adapting the Concept of Continuous Wavelet Transform

The *Continuous Wavelet Transform* (CWT) is often used for extracting the frequency information from a time-series signal by conducting a convolution of the signal with a wavelet function. [1] The CWT of a function $x(t)$ at a scale s ($s > 0$, $s \in \mathbb{R}^{+*}$) and translational value $\tau \in \mathbb{R}$ is expressed by Equation (5):

$$Wave_w(s,\tau) = \frac{1}{\sqrt{s}} \int_{-\infty}^{\infty} x(t)\, \overline{\psi}(\frac{t-\tau}{s}) dt \quad (5)$$

where $\overline{\psi}(t)$ is the *mother wavelet*, a continuous function in both the time domain and the frequency domain, and the over-line represents operation of complex conjugate.

The power of CWT has an interpretation as time-frequency wavelet energy density, and is called the wavelet power spectrum.

$$Power(s,\tau) = \frac{1}{s} |Wave_w(s,\tau)|^2 \quad (6)$$

Compared to the traditional *short-time Fourier transform* (STFT), CWT can also construct a time-frequency representation of a signal that offers reliable time and frequency localization. This property makes it better at extracting frequency features from the non-periodic time-series user interaction sequences. Thus we adapt CWT to automatically detect the frequency distribution of a user's visualization exploration over time.

4.2.2 Metric Calculation Steps

The *exploration-pacing* metric is computed in three steps. We show the computation steps as well as the key parameters we use as follows (also shown in Figure 1).

Step 1: For each user, form a time-series sequence $S(t)$, representing the user's visiting status over time, from a user's exploration interaction sequence $Ex(u)$. $S(t)$ contains a sequence of values of $\{0,1\}$ over time, sampled from a user's exploration process. If at moment t, the user is visiting a chart element, then it is marked as 1; otherwise if the user is not visiting any chart, it is marked as 0. We sample the data every 0.1 second. If a user visited a chart for one second at the beginning of her exploration, then the sequence $S = (1,1,1,1,1,1,1,1,1,1,0,0,0,...)$.

Step 2: Apply *continuous wavelet transform* to the sequence $S(t)$ to obtain a 2D time-frequency wavelet power spectrum. We use the R package "WaveletComp" [32], which computes CWT and obtain a wavelet power spectrum (Figure 1) according to Equation (5) and (6). The Morlet wavelet is used as the mother wavelet for the convolution.

Step 3: Obtain the metric value $Pacing_{HF}$ by computing the average power over time and a high-frequency range $[f_{min}, f_{max}]$.

$$Pacing_{HF} = \frac{1}{(f_{max}-f_{min})(t_{end}-t_{start})} \sum_{f=f_{min}}^{f_{max}} \sum_{t=t_{start}}^{t_{end}} Power(t,f) \quad (7)$$

where t_{start} and t_{end} denote the start and end moments of a user's exploration.

We use $1/32$ Hz and $1/8$ Hz as the minimum and maximum bounds for the high-frequency range to compute the metric. Given that the sampling rate we use is once per 0.1 second, the high-frequency range corresponds to a period range of 0.8 and 3.2 seconds. This range aims to generally align with high-frequency (*i.e.* rapid) visit behavior, and to mitigate possible accidental interactions. Precise modeling and parameters given visualization type and user behavior may be a valuable route for future work. The range parameter can be changed to extract users' power density for other frequencies (*e.g.*, low-frequency visits).

Studies	**SearchinVis**		**HindSight**	
Experiments	*Colleges*	*255Charts*	*255Charts*	*Metafilter*
control	67	57	57	44
experimental	72	72	59	48
Total	*139*	*129*	*116*	*92*

Table 2: Number of participants in 4 experiments in the previous studies, *SearchinVis* [15] and *HindSight* [14]. The control user group in each experiment includes the users randomly assigned to explore original visualizations. The experimental group includes those exploring augmented visualizations, *i.e.*, those who were presented and used the search functionality in *SearchinVis*, and those who interacted with the *HindSight*-applied visualizations.

5 Metric Evaluation

5.1 Four Datasets from Two Studies

We evaluate the two proposed metrics by applying them, together with other four metrics used in visualization literature, to four datasets of interaction logs collected from two previously published studies from Feng *et al.*, *SearchinVis* [15] and *HindSight* [14]. The goal of each study was to examine the impact of an interaction technique on user behavior, specifically, the effects of text-based search in *SearchinVis* and the effects of direct encoding of personal interaction history in *HindSight*. We extract two visualizations from each study for our metric evaluation. Specifically, select the visualizations that were adapted from published visualizations on the web, including *255Charts* and *Boardrooms* from the *SearchinVis* study, and *255Charts* and *Metafilter* from *HindSight*. Other conditions in these studies are less reflective of real world visualizations, *e.g.* an interactive bubble chart of colleges.

In each of these crowdsourced studies, participants were given instructions to interact with the visualization for as long as they liked before answering questions about their exploration. During this time, their interactions with chart elements was recorded. Each of these studies was a between-subjects design, *i.e.*, every participant was in either experimental or control group. Table 2 shows the participant numbers in each experiment dataset. One experiment dataset stands for an study-visualization pair, *e.g.*, *SearchinVis-255Charts*, which contains the interaction logs of all the participants in two groups, the experimental group and the control group.

We now describe three evaluation activities to assess how well these metrics reveal characteristics of user explorations, from both qualitative and quantitative perspectives. First, we inspect several individual cases, to determine whether corresponding metric values aligned with noticeable patterns in peoples' visualization interaction data. Positive results here may suggest that the metrics are appropriate for visualization practitioners in that they reveal some aspect of user behavior that is difficult to obtain with baseline metrics such as number of items visited or time spent in exploration. Second, we compare different user groups using the metrics to make statistical inferences, in the same manner as the original studies. Positive results with statistical inference suggest that the metrics could have been used in controlled user experiments with visualization to examine the effects of the interaction techniques in question from new perspectives. Finally, we examine possible relationships between the proposed metrics and other metrics, to see if these metric values are correlated or independent when applied to participant data.

5.2 Applying Interaction Metrics: Individual Cases

Can the proposed metrics reveal new characteristics of peoples' interaction with visualizations?

To evaluate the extent to which the proposed metric *exploration-uniqueness* shows different exploration patterns, we examine several individual interaction logs in which users visited the same number of charts of the visualization, but varied on the *exploration-uniqueness*

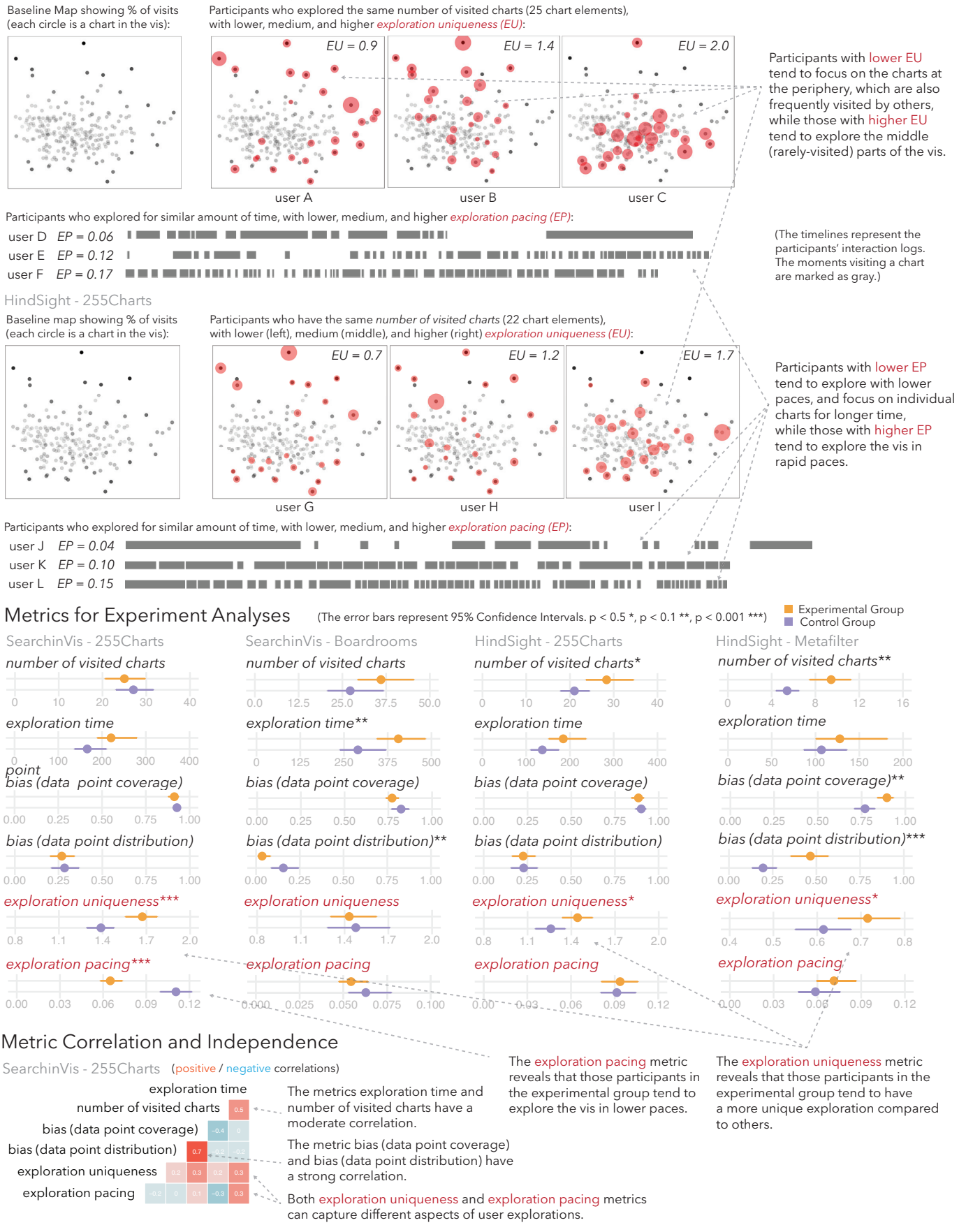

Fig. 2: We applied *pacing* and *uniqueness* to two previous studies, showing that they capture different aspects of user explorations.

values. Then we visually examine whether an exploration with a higher metric value includes more charts rarely-visited by others.

In order to distinguish frequently- and rarely-visited charts, we calculate the percentage of visits for each chart, and plot it as a baseline map (Figure 2), where each circle represents a chart, and its opacity mapped to the percentage of users who visited it. From the baseline maps of both *SearchinVis-255Charts* and *HindSight-255Charts*, visual inspection suggests that the charts at the periphery of the visualization are frequently visited, while the charts in the middle are rarely visited.

As shown in Figure 2, user A, B and C from ***SearchinVis-255Charts*** visited the same number of charts (*i.e.*, *number-of-visited-charts* = 25) during their explorations. However, these explorations vary at *exploration-uniqueness*, *i.e.*, A's is the lower (0.9), B's is the medium (1.4), and C's is the higher (2.0). For each of them, we plot all visited charts on top of the baseline map, where each circle represents a visit, with its size corresponding to how long the user spent visiting this chart. Visually comparing the visit maps from the user A, B and C, we find that the charts visited by A (lower *exploration-uniqueness*) are mostly at the periphery of the visualization, which are frequently visited by other users, according to the baseline map. Instead, user C (higher *exploration-uniqueness*) visited more charts located at the lower-middle part of the visualization, which are rarely visited by other users in the study. The charts visited by user B (middle *exploration-uniqueness*) include some frequently-visited and some rarely-visited charts.

Similarly, user G, H and I from ***HindSight-255Charts*** visited the same number of charts (*i.e.*, *number-of-visited-charts* = 22) during their explorations, while these explorations vary at *exploration-uniqueness*, *i.e.*, G's is the lower (0.7), H's is the medium (1.2), and I's is the higher (1.7). By visually comparing the visit maps from the user G, H and I, we see that the charts visited by G (lower *exploration-uniqueness* value) are mostly at the periphery of the visualization (*i.e.*, frequently-visited charts). Instead, user I (higher *exploration-uniqueness*) visited more charts at the middle part of the visualization, which are rarely visited by other users in the study. The charts visited by user H (middle *exploration-uniqueness*) include both frequently-visited and rarely-visited charts.

We also find that the metric *exploration-pacing* can reveal differences in the pacing of user explorations. Specifically, we examine the individual cases from *SearchinVis* (user D, E, F) and *HindSight* (user J, K, L). Each of these users explored the visualization for similar amount of time, but with lower, medium or higher paces. User D, for example, appears to intersperse rapidly-paced interactions with longer interactions. User F, in contrast, spends nearly all of their time performing rapid exploration. These cases illustrate that the pacing metric can aid in distinguishing between the temporal behavior of users, essentially by transforming the temporal observations to the frequency space.

5.3 Metrics for Experiment Analyses

Can the proposed metrics provide additional insight in experiment analyses?

After examining individual cases to check the validity of the proposed metrics, we further explore their effectiveness on one of the potential application scenarios, *i.e.*, to show the impact of visualization designs on user interaction behavior. Specifically, we compare the metric values between two user groups (experimental and control) in each study by applying the same statistical tests as in the original studies. We also evaluate four other metrics proposed or used in visualization literature, *number-of-visited-charts*, *exploration-time*, *bias-data-point-coverage*, and *bias-data-point-distribution*. We report the results of the two studies, *SearchinVis* and *HindSight*.

Following the statistical methods used in the previous studies, we compute 95% confidence intervals using the bootstrap method, and effect sizes using Cohen's d - which is the difference in means of the conditions divided by the pooled standard deviation. We also use the non-parametric Mann-Whitney test to compare different user groups.

5.3.1 Study 1: SearchinVis

We apply the existing and proposed metrics to the interaction logs from the two experiments of the study *SearchinVis*, *i.e.*, *255Charts Boardrooms* visualization stimuli, to examine the behavioral impact of the text-based search functionality by comparing the two user groups in each experiment.

255Charts: We filtered out 5 users (from 129 users) who did not interact with any elements of the visualization. We found that, compared to the users in the control group, the users from the experimental group spent similar amount of time on exploration, visited similar number of charts, have similar levels of bias, while they also show significantly more unique explorations (***exploration-uniqueness***), and had fewer rapid-pace visits (***exploration-pacing***). The experimental group had a higher *exploration-uniqueness* value on average (M=1.7 95% CI [1.6, 1.8]) than the control group (M=1.4 95% CI [1.3, 1.5]). The Mann-Whitney test shows that $W = 1165, p = 0.0002$, and the effect size is $d = 0.7$ $[0.3, 1]$. The experimental group has a lower *exploration-pacing* value on average (M=0.07 95% CI [0.06, 0.07]) than the control group (M=0.11 95% CI [0.1, 0.12]). The Mann-Whitney test shows that $W = 3024, p = 2.3^{\times}10^{-8}$, and the effect size is $d = -1.2$ $[-1.57, -0.8]$.

Importantly, these metrics align with and quantify the findings and intuitions of that study. The addition of search to *255Charts* encouraged more diverse spatial patterns of exploration, while also nudging users to look more in-depth at specific data elements.

Boardrooms: We filtered out 11 users (from 96 users) who did not interact with any elements of the visualization. We found that, compared to users in the control group, users in the experimental group visited slightly more chart elements, spent significantly more time, had lower bias measures in data distribution, had slightly lower bias measures in data coverage and rapid-pacing measures, and finally had similar uniqueness measures. We found significant differences on the ***exploration-time*** and ***bias-data-point-distribution*** between the two user groups. Specifically, the experimental group spent longer time on average (M=405 95% CI [337, 480]) in seconds than the control group (M=290 95% CI [234, 368]). The Mann-Whitney test shows that $W = 598, p = 0.007$, and the effect size is $d = 0.51$ $[0.05, 0.96]$. This difference has been reported in the previous study [15]. The experimental group has a lower *bias-data-point-distribution* value on average (M=0.03 95% CI [0.01, 0.09]) than the control group (M=0.16 95% CI [0.09, 0.25]). The Mann-Whitney test shows that $W = 1109.5, p = 0.01$, and the effect size is $d = -0.6$ $[-0.95, -0.14]$.

In summary, by applying the existing and proposed metrics to two experiments in the study, we found that our proposed metrics appear to provide additional insight in experiment analyses that could have appeared in previous published work, *i.e.*, by quantifying the impact of text-based search functionality on the uniqueness or "diversity" of user explorations. At the same time, we find that in Feng *et al.*'s text-based search study, the presence of the search functionality appears to have a different impact on user behavior when added to the *255Charts* and *Boardrooms* visualizations. We discuss possible explanations for this difference in Section 6.

5.3.2 Study 2: HindSight

We apply existing and proposed metrics to the interaction data from the two experiments of the study *HindSight* [14], which includes *255Charts*, a visualization from The New York Times, and a comparatively simpler *Metafilter* visualization stimuli. The aim here is to examine the behavioral impact of the direct encoding of personal interaction history, by comparing the two user groups in each experiment.

255Charts: We filtered out one user (from 116 users) who did not interact with any elements in the visualization. We found that, compared to the users in the control group, the users in the experimental group spent a similar amount of time on exploration, visited more chart elements, had similar level of bias, while explore significantly more of the rarely-visited charts in the visualization, compared to those in the control group. We found significant differences on the metric ***number-of-visited-charts*** and ***exploration-uniqueness*** between the two user groups, while the other metrics are similar between groups. Specif-

ically, the experimental group visited more chart elements on average (M=28 95% CI [24, 35]) than the control group (M=21 95% CI [18, 24]). The Mann-Whitney test shows that $W = 1392.5, p = 0.11$, and the effect size is $d = 0.41$ $[0.07, 0.73]$. This difference has been reported in the previous study [14]. The experimental group has a higher *exploration-uniqueness* value on average (M=1.44 95% CI [1.34, 1.54]) than the control group (M=1.26 95% CI [1.17, 1.35]). The Mann-Whitney test shows that $W = 1165, p = 0.0002$, and the effect size is $d = 0.48$ $[0.11, 0.84]$.

Metafilter: We found that compared to the users in the control group, the users from the experimental group visited more chart elements, spent slightly more time on exploration, had significantly more unique explorations, and have higher level of bias. We found significant differences on the metric ***number-of-visited-charts***, ***bias-data-point-coverage***, ***bias-data-point-distribution***, ***exploration-uniqueness*** between the two user groups, while the other metrics are similar between groups. Specifically, the experimental group has a higher *number-of-visited-charts* value on average (M=9.4 95% CI [7.7, 11.3]) than the control group (M=5.4 95% CI [4.3, 6.5]). ($W = 686, p = 0.004, d = 0.75$ $[0.37, 1.12]$) This difference has been reported in the previous study [14]. The experimental group also has a higher *exploration-uniqueness* value on average (M=0.7 95% CI [0.6, 0.8]) than the control group (M=0.6 95% CI [0.6, 0.7]). The Mann-Whitney test shows that $W = 801.5, p = 0.047$, and the effect size is $d = 0.44$ $[0.02, 0.86]$. In addition, we found that the experimental group has a higher *bias-data-point-coverage* value on average (M=0.9 95% CI [0.8, 0.9]) than the control group (M=0.8 95% CI [0.7, 0.8]). The Mann-Whitney test shows that $W = 646.5, p = 0.001$, and the effect size is $d = 0.71$ $[0.26, 1.13]$. The experimental group also has a higher *bias-data-point-distribution* value on average (M=0.5 95% CI [0.4, 0.6]) than the control group (M=0.2 95% CI [0.1, 0.3]). The Mann-Whitney test shows that $W = 616.5, p = 0.0006$, and the effect size is $d = 0.82$ $[0.41, 1.24]$.

In summary, by applying the existing and proposed metrics to two experiments in the study *HindSight*, we found that our proposed metrics can provide additional insight in experiment analyses, *i.e.*, uncover the impact of direct encoding of personal interaction history on the uniqueness of user explorations. We also find that the HindSight technique has a different impact on user behavior, specifically, users' bias levels, when added to the real-world and less complex visualizations. We further discuss the possible causes and implications of this difference in Section 6.

5.4 Metric Correlation and Independence

Are the proposed metrics correlated or independent when applied to real data?

We compute correlations between each pair of the metrics across all experimental datasets (Figure 2), *SearchinVis-255Charts*, *SearchinVis-Boardrooms*, *HindSight-255Charts* and *HindSight-Metafilter*. We expect that the metrics measuring different high-level aspects of user explorations are independent from each other.

We found strong and moderate correlations, $r = [0.5, 1)$, between the metric *bias-data-point-coverage* and *bias-data-point-distribution*. Specifically, they are strongly correlated in *SearchinVis-255Charts* ($r = 0.72, p < 0.001$) and *HindSight-Metafilter* ($r = 0.76, p < 0.001$), and moderately correlated in *HindSight-255Charts* ($r = 0.67, p < 0.001$) and *SearchinVis-Boardrooms* ($r = 0.61, p < 0.001$). This indicates that for these cases, the two metrics play similar roles characterizing exploration behavior.

We also found a moderate correlation, $r = [0.5, 0.7)$, between *number-of-visited-charts* and *exploration-time*, in *SearchinVis-255Charts* ($r = 0.57, p < 0.001$).

The metric *exploration-uniqueness* has a weak correlation or less with any other metric across all the datasets. Similarly, the metric *exploration-pacing* also has a weak correlation or less with any other metric across all the datasets. These results suggest that the proposed metrics carry different information than the others when applied to user exploration data. However, we note that linear correlation is just one of many possible measures of dependence, and further analyses with larger datasets may be necessary to make definitive claims.

6 Discussion

The results of these evaluation activities suggest that the proposed uniqueness and pacing metrics reveal new facets of how people interact with data visualizations. We discuss how these metrics can be used by visualization creators and researchers, and explore emerging challenges and opportunities for future work in this space.

Specifically, we applied the proposed metrics, together with metrics explored in prior work, to interaction data from prior information visualization studies. These results suggest that, first, the proposed metrics can reveal new characteristics of peoples' exploration behavior in visualizations. Second, the results suggest that the proposed metrics can be used as target metrics in comparative experiments, *i.e.* quantitative analysis comparing control and treatment groups. Finally, the proposed metrics are generally independent of prior metrics used in visualization research, as indicated by the correlation analysis, implying that they may be a source of *new information* in exploratory visualization design and research.

In the analysis of the study *SearchinVis*, we found that the results are different for 255*Charts* and *Boardrooms*. The metric values of *exploration-uniqueness* and *exploration-pacing* are significantly different between groups in 255*Charts* while in *Boardrooms* they are similar. One possible explanation is that the *Boardroom* visualization is in a storytelling form with multiple types of interactions enabled. Besides the hovering enabled for each dot element, the user can also scroll up and down, or click tabs to change views. The extra available interactions create challenges for the measure of user interaction behavior. (further discussed in Section 6.3)

We also found that in the experiment *HindSight-Metafilter*, the experimental user group has higher values in the bias metrics on average than the control group. By further examining the interaction logs, we found that a higher value in the bias metric is contributed by the revisits to previously visited chart elements. On the other hand, the users in the experimental group visited more charts on average than those in the control group, in the whole exploration process. More questions have thus been raised. For example, is "revisiting a chart element" always related to bias? Should a user's level of bias be indicated through revisit chance, or the breadth of the whole open-ended process?

6.1 Potential Generalized Applications

New and newly evaluated metrics for visualization interaction analysis may open several opportunities in visualization research.

As *metrics* to quantify the impact of visualization designs. One goal of researchers and practitioners is to examine the comparative impact of competing techniques on user behavior. However, users' open-ended explorations of visualizations can be complex, as described in Section 3, and cannot be adequately summarized into *number of actions taken* or *total time spent on exploration*, which are the basic metrics commonly used in previous evaluations, *e.g.* [5, 14, 23]. Instead, we have shown that from these low-level user interaction components we can develop metrics that provide new perspectives into users' open-ended explorations, which may allow us to better assess the impact of a given visualization or interaction technique on user behavior.

As *proxies* to infer user characteristics and reasoning processes. Certain behavioral patterns of visualization exploration can be used to infer users' characteristics (*e.g.*, locus of control [27, 28]), reasoning processes [12], and insight generation [17]. By providing new ways to characterize users' exploration behavior, we could possibly explore new avenues of individual differences in visualization use and preference.

As *attributes* to visualize users' interaction logs. One advantage of existing visual analytics approaches for analyzing users' interaction data is that detailed information can be logged during user interaction, as shown in the previous works that center on visualizing interaction logs *e.g.* Blascheck *et al.* [3, 4]. However, one limitation of this approach is that it relies primarily on expert analysts to visually identify trends across user interaction traces. Fortunately, recent work has begun to explore automatic approaches to assist in navigating interaction

traces, such as sequence search and extraction [3]. We contend that new metrics can also aid in this direction of research, by serving as relatively low-barrier features that could be encoded in the interaction log visualization, for example ordering or coloring logs by uniqueness, bias, or pacing [2].

As *features* to support machine learning algorithms. Machine learning techniques have also been used to analyze users' interactions with visualizations, *e.g.*, to classify user characteristics [28], to extract interaction sequences [7], and to cluster users by behavioral patterns [3]. The proposed metrics in this paper, as well as intermediate variables generated from the computation process of the metrics, may be useful as features for these machine learning approaches. For example, users' explorations can be clustered using the feature vectors containing the time spent on each visualization element (Equation 2), the TF-IDF feature vectors (Equation 3), or the highest level exploration uniqueness metric values. Similarly, both the wavelet power spectrum and the corresponding pacing metric can be used as features for machine learning algorithms.

6.2 Benefits and Tradeoffs of the Metrics

All of the metrics in our evaluations may prove beneficial to user behavior analysis in visualizations, due to their ability to uncover different facets of peoples' explorations. However, potential adopters need to be aware of certain properties of these metrics in order to apply them correctly. We now compare the metrics in our evaluations according to a list of criteria focusing on barriers such as interpretability, and derive initial guidelines on when and how to use these metrics.

The criteria used for metric comparison are adapted from the works in relevant fields evaluating metrics [11, 31, 10], and are listed from lower to higher perspectives (*i.e.*, from metric computation to human perception and cognition):

Computational cost (computational level): *How much does it cost for the metric computation?* The computation of some metrics is trivial, such as *number-of-visited-charts*. However, there is a certain level of complexity required to compute other metrics *e.g.*, *exploration-uniqueness*. The computation of *exploration-pacing* requires more resources and its complexity depends on the choice of convergence parameters. The computation of *bias-data-point-coverage* includes power operations on the number of interactions performed by the user, *i.e.*, N^k where N is the number of all the interactive elements in the visualization, and k is the number of interactions performed by the user. This suggests that extra steps may be needed to avoid the overflow caused by large numbers when dealing with an exploration session where a user interacts with the visualization a lot, *e.g.*, $k > 100$.

Computational context (computational level): *Does the metric computation require extra context,* i.e.*, out of the single user scope?* Among all the metrics in our evaluation, the computation of *exploration-uniqueness* depends on the interaction logs not only of the current user being considered, but also those of the other users within the same group, while the computations of other metrics, *e.g.*, *exploration-pacing* are only based on the current user, *i.e.*, no extra context needed. This property influences the practical usage of a metric, *e.g.*, a reasonable number of users should be selected when computing the *exploration-uniqueness* metric.

Comparability (application level): *(How) can the metric values be compared?* All of the evaluated metrics are comparative because they are quantitative measures of scale. The comparability of *exploration-uniqueness* is constrained because the metric values are only comparative within a TF-IDF computation group, *i.e.*, it is not feasible to compare the metric values of two users in different computation groups. The values of the *exploration-pacing* metric can be compared across user groups if the same set of parameters are used for the computation.

Interpretability (cognition level): *How easily can the metric be understood or interpreted by human?* The proposed metrics have different levels of interpretability. Metrics such as *exploration-time* and *number-of-visited-charts* could be considered readily interpretable, since people can easily understand the meaning of the values (*e.g.*, 10 elements, 15 seconds). The values of some other metrics may require more cognitive effort to interpret, *e.g.*, the *exploration-uniqueness* metric, *exploration-pacing* and the *bias-data-point-coverage*.

Knowledge coverage (cognition level): *How much additional knowledge does the metric cover given other metrics?* In our study, this means whether a metric or a group of metrics, can uncover the characteristics of users' exploration that other metrics can hardly capture. By examining correlations between metrics, we found that both proposed metrics, *exploration-uniqueness* and *exploration-pacing* may reveal different perspectives from other metrics.

6.3 Challenges in Designing and Evaluating Visualization Interaction Metrics

As part of this work, we develop a guiding set of requirements for metrics that quantify people's open-ended exploration using visualizations. The requirements space includes several dimensions, including high-level facets such as inferring designer needs for audience behavior analysis, and lower-level facets such as determining what features can be mathematically derived from visualization interaction logs. Within these requirements dimensions, existing visualization interaction metrics can be categorized, and new metrics can be derived and placed.

There at least two ongoing challenges regarding the development of this space, specifically regarding the measuring needs and measurable features:

1. Gaps remain in measurement needs.
2. There is a not currently a systematic way to define the set of measurable features from users' interaction logs.

For each of the measurement goals for behavior analysis, there is a set of measurement needs. For example, the time and breadth of a user's exploration (measuring needs) are used to indicate how a user is engaged in the visualization (purpose). However, it remains unclear whether a longer exploration time or a broad range of exploration is always related to an "engaged" exploration.

The current measurable features from our requirements space (including exploration time, elements and duration of interaction, etc.) can only capture and describe users' interactions with a subset of interactive visualizations. In more general cases, there are many different kinds of visualization representations and interaction schemas, for some of which the features that can be measured might differ. For example, the interactions enabled in the visualizations can be more complicated (*i.e.*, brushing and linking), and one interaction may be dependent on another. Given this reality, we note that there is much more information that remains to be captured from users' interaction logs with visualizations. Thus, a more comprehensive framework of visualization parameters and interaction grammar, such as Satyanarayan *et al.*'s recent work on Vega [35], could make it much easier to scope features can be captured and how to capture them as part of the visualization design.

7 Conclusion

Each day, thousands of people interact with thousands of interactive visualizations across the web's vibrant and growing visualization ecosystem. However, our metrics for quantifying facets of peoples' open-ended explorations with these visualizations are lacking, as they are primarily based on low-level metrics such as *elements visited* or *time spent exploring*. The aim of this work is to characterize, develop, and evaluate metrics for visualization interaction that can be used in a variety of settings. We introduce two new metrics, *uniqueness* and *pacing*, and evaluate these metrics alongside those proposed in earlier and more recent research in visualization. The results of these evaluations suggest that, indeed, new metrics may provide new perspectives on how people interact with the visualizations they come across. Several metrics are also shown to be suitable candidates in prior visualization studies. We discuss the broad potential applications of new metrics for visualization interaction analysis, and enumerate some of the challenges future work in interaction metrics may face in the future.

REFERENCES

[1] P. S. Addison. *The illustrated wavelet transform handbook: introductory theory and applications in science, engineering, medicine and finance*. CRC press, 2017.

[2] E. Bertini, A. Tatu, and D. Keim. Quality metrics in high-dimensional data visualization: An overview and systematization. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2203–2212, 2011.

[3] T. Blascheck, M. John, K. Kurzhals, S. Koch, and T. Ertl. Va 2: a visual analytics approach for evaluating visual analytics applications. *IEEE transactions on visualization and computer graphics*, 22(1):61–70, 2016.

[4] T. Blascheck, S. Koch, and T. Ertl. Logging interactions to learn about visual data coverage. *LIVVIL: Logging Interactive Visualizations & Visualizing Interaction Logs*, 2016.

[5] J. Boy, F. Detienne, and J.-D. Fekete. Storytelling in information visualizations: Does it engage users to explore data? In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1449–1458. ACM, 2015.

[6] J. Boy, L. Eveillard, F. Detienne, and J.-D. Fekete. Suggested interactivity: Seeking perceived affordances for information visualization. *IEEE transactions on visualization and computer graphics*, 22(1):639–648, 2016.

[7] E. T. Brown, A. Ottley, H. Zhao, Q. Lin, R. Souvenir, A. Endert, and R. Chang. Finding waldo: Learning about users from their interactions. *IEEE Transactions on visualization and computer graphics*, 20(12):1663–1672, 2014.

[8] R. Chang, M. Ghoniem, R. Kosara, W. Ribarsky, J. Yang, E. Suma, C. Ziemkiewicz, D. Kern, and A. Sudjianto. Wirevis: Visualization of categorical, time-varying data from financial transactions. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 155–162. IEEE, 2007.

[9] E. H. Chi, P. Pirolli, K. Chen, and J. Pitkow. Using information scent to model user information needs and actions and the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 490–497. ACM, 2001.

[10] A. Croll and B. Yoskovitz. *Lean analytics: Use data to build a better startup faster*. " O'Reilly Media, Inc.", 2013.

[11] B. Donmez, P. E. Pina, and M. Cummings. Evaluation criteria for human-automation performance metrics. In *Performance evaluation and benchmarking of intelligent systems*, pages 21–40. Springer, 2009.

[12] W. Dou, D. H. Jeong, F. Stukes, W. Ribarsky, H. R. Lipford, and R. Chang. Recovering reasoning processes from user interactions. *IEEE Computer Graphics and Applications*, 29(3), 2009.

[13] O. ElTayeby and W. Dou. A survey on interaction log analysis for evaluating exploratory visualizations. In *Proceedings of the Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 62–69. ACM, 2016.

[14] M. Feng, C. Deng, E. M. Peck, and L. Harrison. Hindsight: encouraging exploration through direct encoding of personal interaction history. *IEEE transactions on visualization and computer graphics*, 23(1):351–360, 2017.

[15] M. Feng, C. Deng, E. M. Peck, and L. Harrison. The effects of adding search functionality to interactive visualizations on the web. 2018.

[16] D. Gotz. Soft patterns: Moving beyond explicit sequential patterns during visual analysis of longitudinal event datasets. In *Proceedings of the IEEE VIS 2016 Workshop on Temporal & Sequential Event Analysis*, 2016.

[17] H. Guo, S. R. Gomez, C. Ziemkiewicz, and D. H. Laidlaw. A case study using visualization interaction logs and insight metrics to understand how analysts arrive at insights. *IEEE transactions on visualization and computer graphics*, 22(1):51–60, 2016.

[18] J. Heer. Capturing and analyzing the web experience. In *Proc. CHI 2002*, 2002.

[19] J. Heer and E. H. Chi. Identification of web user traffic composition using multi-modal clustering and information scent. In *Proc. of the Workshop on Web Mining, SIAM Conference on Data Mining*, pages 51–58, 2001.

[20] J. Heer and E. H. Chi. Separating the swarm: categorization methods for user sessions on the web. In *Proceedings of the SIGCHI Conference on Human factors in Computing Systems*, pages 243–250. ACM, 2002.

[21] J. Heer, J. Mackinlay, C. Stolte, and M. Agrawala. Graphical histories for visualization: Supporting analysis, communication, and evaluation. *IEEE transactions on visualization and computer graphics*, 14(6), 2008.

[22] H. Lam, D. Russell, D. Tang, and T. Munzner. Session viewer: Visual exploratory analysis of web session logs. In *Visual Analytics Science and Technology, 2007. VAST 2007. IEEE Symposium on*, pages 147–154. IEEE, 2007.

[23] Z. Liu and J. Heer. The effects of interactive latency on exploratory visual analysis. *IEEE transactions on visualization and computer graphics*, 20(12):2122–2131, 2014.

[24] Z. Liu, B. Kerr, M. Dontcheva, J. Grover, M. Hoffman, and A. Wilson. Coreflow: Extracting and visualizing branching patterns from event sequences. In *Computer Graphics Forum*, volume 36, pages 527–538. Wiley Online Library, 2017.

[25] Z. Liu, Y. Wang, M. Dontcheva, M. Hoffman, S. Walker, and A. Wilson. Patterns and sequences: Interactive exploration of clickstreams to understand common visitor paths. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):321–330, 2017.

[26] M. Monroe, R. Lan, H. Lee, C. Plaisant, and B. Shneiderman. Temporal event sequence simplification. *IEEE transactions on visualization and computer graphics*, 19(12):2227–2236, 2013.

[27] T. A. O'Connell and Y.-Y. Choong. Metrics for measuring human interaction with interactive visualizations for information analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1493–1496. ACM, 2008.

[28] A. Ottley, H. Yang, and R. Chang. Personality as a predictor of user strategy: How locus of control affects search strategies on tree visualizations. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 3251–3254. ACM, 2015.

[29] M. Pohl, S. Wiltner, and S. Miksch. Exploring information visualization: describing different interaction patterns. In *Proceedings of the 3rd BELIV'10 Workshop: Beyond time and errors: novel evaluation methods for information visualization*, pages 16–23. ACM, 2010.

[30] J. Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 133–142, 2003.

[31] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit. Saliency and human fixations: state-of-the-art and study of comparison metrics. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1153–1160. IEEE, 2013.

[32] A. Roesch and H. Schmidbauer. *WaveletComp: Computational Wavelet Analysis*, 2018. R package version 1.1.

[33] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5):513–523, 1988.

[34] P. Saraiya, C. North, and K. Duca. An insight-based methodology for evaluating bioinformatics visualizations. *IEEE transactions on visualization and computer graphics*, 11(4):443–456, 2005.

[35] A. Satyanarayan, D. Moritz, K. Wongsuphasawat, and J. Heer. Vega-lite: A grammar of interactive graphics. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):341–350, 2017.

[36] M. J. Schuemie, M. Weeber, B. J. Schijvenaars, E. M. van Mulligen, C. C. van der Eijk, R. Jelier, B. Mons, and J. A. Kors. Distribution of information in biomedical abstracts and full-text publications. *Bioinformatics*, 20(16):2597–2604, 2004.

[37] Z. Shen, J. Wei, N. Sundaresan, and K.-L. Ma. Visual analysis of massive web session data. In *Large Data Analysis and Visualization (LDAV), 2012 IEEE Symposium on*, pages 65–72. IEEE, 2012.

[38] E. Wall, L. M. Blaha, L. Franklin, and A. Endert. Warning, bias may occur: A proposed approach to detecting cognitive bias in interactive visual analytics. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*, 2017.

[39] D. Wang, D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. Human mobility, social ties, and link prediction. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1100–1108. Acm, 2011.

[40] K. Wongsuphasawat, J. A. Guerra Gómez, C. Plaisant, T. D. Wang, M. Taieb-Maimon, and B. Shneiderman. Lifeflow: visualizing an overview of event sequences. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1747–1756. ACM, 2011.

[41] J. S. Yi, Y. ah Kang, and J. Stasko. Toward a deeper understanding of the role of interaction in information visualization. *IEEE transactions on visualization and computer graphics*, 13(6):1224–1231, 2007.

[42] J. Zhao, Z. Liu, M. Dontcheva, A. Hertzmann, and A. Wilson. Matrixwave: Visual comparison of event sequence data. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 259–268. ACM, 2015.