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

Mi Feng, Evan Peck, Lane Harrison

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 identify needs for visualization behavior measurement, and develop corresponding candidate features that can be inferred from users' interaction data. 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 implications of these results for future studies, including the potential for applying these metrics in visualization interaction analysis, as well as emerging challenges in developing and selecting metrics depicting visualization explorations.

Index Terms—Interaction, Visualization, Quantitative Evaluation.

1 Introduction

As interactive visualizations migrate from standalone applications to the web, visualization 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? In this paper, we 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, etc. 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 interacted with*. 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 interact with?" or "how long did the average person explore the visualization?". Yet despite their diversity in goals, it's possible that users interact with 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 [23, 5, 6, 16]. Various analysis strategies have been used to answer these questions, including *statistical* and *visual* approaches (*e.g.*, [3, 5, 30]). However, these existing approaches have limitations with characterizing user explorations precisely. Many of the metrics used to summarize activity tend to overaggregate behavior, failing to identify differences between users, or by failing to capture detailed information such as *how long* has been spent on *which visual 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 measures, which take into account more information in peoples' interaction logs, and to better reveal facets of peoples' explorations. Related efforts 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 efforts influence our work of visualization interaction analysis, in that a user's open-ended exploration of a visualization containing visual elements can be considered analogous to the exploration of a website. 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.

The aims of this paper are three-fold:

1. Derive a requirements space to categorize existing and new metrics that quantify facets of users' exploration of data visualization and in doing so identify emerging analysis needs.
2. Derive two new metrics centered around user exploration diversity and pacing that 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 [1].

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.

- *Mi Feng, Worcester Polytechnic Institute, mfeng2@wpi.edu.*
- *Evan Peck, Bucknell University, evan.peck@bucknell.edu.*
- *Lane Harrison, Worcester Polytechnic Institute, lane@cs.wpi.edu.*

Manuscript received 31 March 2018; accepted 1 August 2018; posted online 13 October 2018; mailed on 4 October 2018.
For information on obtaining reprints of this article, please send e-mail to: tvcg@computer.org.

[1] The experiment data and analysis scripts are available on the Open Science Framework: https://osf.io/dx43q

2 Background

2.1 Characterizing Website Exploration

One closely related thread of research is clickstream analysis and visualization for websites or applications [22, 24, 25, 39, 46], under a broader research topic of event sequence analysis [15, 28, 43, 27, 45]. 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.* [46] created a visualization called MatrixWave to compare two clickstream datasets, and found it to scale better than commonly used Sankey diagrams.

2.2 Characterizing Visualization Explorations

Characterizing user behavior through interaction logs has been used for various purposes, such as learning user characteristics [7, 30], understanding system usage [32] and the reasoning process [12], and evaluating visualization design [5, 16, 13].

2.2.1 Visual Approaches

Visual approaches refer to strategies of showing users' interaction logs with visualizations [8, 12, 43, 37]. 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.* [30] 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. [27]

2.2.2 Statistical Approaches

Commonly used metrics to depict a user's exploration include *total exploration time* spent by a user, and *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, 16, 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.* [16] coded the raw interactions into semantic interactions, such as selecting, filtering and inspecting, according to Yi *et al.*'s [44] 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.* [16] further extracted the sub-sequences containing specific individual interactions, and then counted the subsequences for each user. Wall *et al.* [40] proposed six metrics to measure cognitive bias, including data coverage, data distribution and attribute coverage/distribution, etc.

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, which translate into two questions:

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 measuring needs for visualization behavior analysis, and 2) deriving low-level measurable 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 [29] 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, collaboration, ease of use, etc. O'Connell *et al.* then derive corresponding metrics for each of the heuristics by utilizing features from users' interaction data such as number of interactions performed by a user.

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, by 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 often develop metrics to capture various dimensions of user experience. While there is currently no widely used metrics framework targeting visualization activity, we might consider frameworks developed in human-computer interaction as a starting point. Consider the HEART metrics framework for web applications proposed by Rodden *et al.* [34]. In the context of evaluating web applications, Rodden *et al.* define five categories: *happiness*, *engagement*, *adoption*, *retention*, and *task success*. This framework, while not originally intended for visualization, can be used to some extent to categorize existing efforts in developing metrics for interactive visualizations. For example, some existing metrics aim to reveal users' *task success*. Time spent on exploration has been used as an indicator of how efficiently a task is performed by the user [31, 41]. Targeting bias, Wall *et al.*'s metrics aim to capture the quality of exploration in visual analytics contexts. Other metrics target proxies of *engagement* with visualizations. For example, interaction coverage metrics have been explored in several recent works, typically targeting numbers of specific types of interaction events (*e.g.*, [23, 5, 13, 38]).

The complexity of interactive visualization makes it difficult to develop metrics that fulfill all dimensions of measurement needs, especially by only adapting metrics from web applications. Challenges surrounding metrics development are especially salient given the question of measuring user "engagement" in visualizations, which indicates *users' willingness to invest effort to explore further and gain more information from the visualization* [26, 5, 17]. For example, longer exploration may indicate engagement, but may also be reflective of confusion and difficulties faced by a user learning a new interaction scheme. Similarly, an increase in interaction counts could either reflect users' interest or possibly their random clicks to orient themselves in a new environment. Given these ambiguities, one aim of this paper is to explore new formulations and perspectives on metrics that may serve as useful proxies for capturing part of the user experiences of an interactive visualization.

In this paper, we seek new metrics to reveal more aspects of user engagement with visualizations. The next section (3.2) deals with how to enumerate these metrics given users' interaction data.

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.* [40] listed two measurable fea-

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.
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.
$T(u_m, e_n)$	Time spent by 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.

tures, *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 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 [40, 4], visualization interaction frameworks [44], and 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 (see Table 1 for precise descriptions using mathematical notations):

- **type**: What type of interaction is it (*e.g.* selection, hovering, or the types from existing frameworks such as Yi *et al.* [44])?
- **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 exploration does each interaction take place?
- **exploration time**: How long does the user spend exploring facets of the visualization?

We list several metrics characterizing users' visualization explorations and identify the underused features. There are several basic metrics that are commonly used by researchers and practitioners. 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, 16]: equals to k, the number of certain type of interactions performed by the user during exploration.
- ***number-of-visited-elements*** [13]: equals to $|E(u)|$, the number of unique visualization elements visited by the user. A *visit* is defined as a meaningful (*i.e.* non-accidental) interaction with an element, which lasts for at least a short amount of time (*e.g.*, 500ms) [5, 13, 14].
- ***exploration-time*** [5, 23, 13, 14]: 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.* [40] propose metrics to measure bias in visualization exploration, by using multiple features from users' interaction data.

bias-data-point-coverage measures bias based on the user's coverage of the data points in the visualization. Since one data point is often mapped to one element in the visualization, in this paper, a data point is considered equivalent to a visual element.

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

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 elements 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***, measuring bias toward repeated interactions with individual data points or subsets of the data.

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

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

Among the existing metrics, we observe that the basic metrics (*e.g.*, *number-of-visited-elements*) reveal the *end results* of the exploration process, meaning that the *details in the process* (*e.g.*, elements and duration of each interactions) are not preserved in the metrics. Wall *et al.*'s bias metrics go beyond traditional methods by distinguishing the *element* of each interaction in the exploration process. However, we still observe that some other features revealing details are underused by existing metrics, *e.g.*, *duration* and *moment* of interaction. Examining the existing metrics such as *visits* and *exploration time* in light of these recent studies raise new questions about user exploration in visualizations. For example, instead of *how many* elements are visited in the visualization, what about the *diversity* of elements visited? And instead of *how much* time is spent in exploration, what about the *pacing* of peoples' exploration inside visualizations?

Given both the unfilled needs of measurement and underused features in users' interaction data, we propose two additional metrics that take into amount more features, and thus offer new perspectives user engagement, *i.e.*, the diversity and pacing of peoples' open-ended explorations with interactive visualizations.

4 Proposed Metrics

4.1 Exploration Uniqueness

The *exploration-uniqueness* (*EU*) metric aims to capture the diverse engagement of peoples' open-ended explorations, *i.e.*, to quantify how unique a user's exploration pattern is compared to patterns from others. The *EU* metric is defined as *aggregated visit duration over visual elements, weighted by the uniqueness in comparison to the crowd.* A low *EU* value suggests that a user's time distributed visiting the visual elements align with common patterns of exploration in the visualization. A high *EU* value suggests that a user's exploration strategy differentiates itself from most other users.

Measurable features: There are many ways to define a unique exploration pattern, *e.g.*, visiting a set of unique elements, or visiting elements in an unique order. This metric characterizes the uniqueness of a user's visited element sets, instead of visit orders. In open-ended explorations, visit orders may carry too much variance, obscuring the trends of unique explorations. We thus compute uniqueness based on 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 *elements* and *duration of interaction*, *i.e.*, the *order of interaction* and *exploration time* are discarded.

Modeling approach: As we develop the metric, we model the interaction behavior of a group of users as a *matrix* (Equation (2)), where each row represents a user, and each column represents an element in the visualization. An alternative approach is to model each user's interaction sequence as a Markov chain, which is used in Wall *et al.*'s bias metrics. Each interaction with a visual element is a state in a state space. A user performing the {element, interaction} combination has transitioned to the associated state in the Markov chain. We adopt *matrix* to model interactions for the *EU*metric mainly for two reasons. First, a user's behavior is represented as a vector, instead of a sequence. This representation aligns with the selection of features, *i.e.*, focus on the data distribution, rather than the visit order. Second, matrix enables comparisons across users, *i.e.*, each user's behavior can be compared to the crowd.

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

In order to depict the uniqueness of a users' exploration process, we need to know how unique each of her visit is, compared to other users' visits. From the field of information retrieval, we find one adaptable concept, *Term Frequency-Inverse Document Frequency (TF-IDF)*, describing how unique a word is in a document collection. The uniqueness of the appearance of a word in a document is the product of TF and IDF. *Term frequency (TF)* is the frequency of the word in a particular document. *Inverse document frequency (IDF)* is the inverse proportion of documents the word appears in. IDF acts as a weight to TF, rewarding the words appear in less documents, and penalizing those appear in more documents. For example, *a* and *the* tend to have lower IDF than *house*.

Given a document collection D, a word w, and an individual document $d \in D$, we calculate a TF-IDF value for each word in a document:

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

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 [36].

In several research initiatives, TF-IDF has been extended from characterizing words to modeling user behavior. For example, In the field of HCI, TF-IDF has been used to describe the uniqueness of peoples' visits to a website [9, 19, 20], and peoples' geospatial movement [42]. Herein we adapt TF-IDF to calculate the uniqueness of a user's visit to an element in a visualization. Specifically, we map a user's interaction log to a *document*, each visit of a visual element to a *word* in the document, and a collection of exploration sessions of multiple users to a *corpus*. A visual element visited by more users has a higher IDF value than an element visited by fewer users.

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 visual elements in a collection of interaction logs.

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

where each row represents a user u_m, each column represents a visual 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.

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

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 visual element e_m.

To calculate the *Term Frequency* of TF-IDF, we choose to use a user's aggregated time spent on a visual element $T(u_m,e_n)$, divided by the total time the user spent visiting all the visual elements. 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 noise, *i.e.*, during open-ended explorations, a user might accidentally interact with an element, and aggregating the time spent by the user on the element can better indicate the user's intentional visit to the element.

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

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

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

By aggregating the TF-IDF values of the visits from one user to all the visual elements, we get a metric depicting the overall uniqueness of the user's exploration. This 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. But the aggregation has at least two advantages. First, it enables the comparison among any users in a group. Before aggregation, each user's interaction behavior is modeled as a vector. A vector supports pairwise comparison between two users, which can be very useful under some circumstances (see Section 6.1). However, it does not support comparison among more than two users. The aggregated metric, instead, supports comparison between subgroups containing any number of users, which is useful for evaluating alternative visualization designs through user interaction. Second, a sum-based aggregation preserves all the unique visits during a user's exploration, *i.e.*, if a user has visited *any* rarely-visited element, it will be preserved in the final metric. Considering an alternative aggregation approach – averaging the uniqueness of a user's visits by the visited elements, if a user has visited lots of frequently-visited elements and one rarely-visited element, the latter will be averaged out after aggregation.

4.2 Exploration Pacing

The pacing metric aims to differentiate temporal strategies by users. Given that the temporal features, *e.g.*, *duration* and *moment* of interactions, are underused by existing metrics, we seek metrics that can utilize them and reveal peoples' diverse exploration patterns. The temporal information from a user's exploration process can be viewed as time-series signal, with the moments "visiting any element" marked as non-zero values, and the other moments marked as zeros.

In fields related to signal processing, *e.g.*, image and audio processing, people extract features not only from the temporal aspect, but also the frequency aspect of the signal, *e.g.*, the high-frequency parts of an audio piece. Similarly, the frequency-related information may also carry users' exploration characteristics. While users may visit the same set of visualization elements and spend same amount of time on exploration, the duration and frequency of those visits may be different, and thus may reflect different exploration strategies. By characterizing these differences with *exploration-pacing*, we may begin to quantify another aspect of user engagement with a visualization.

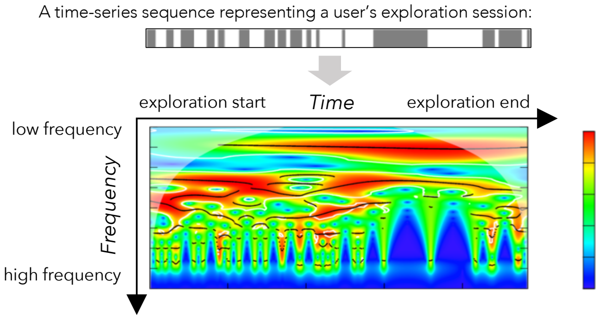


Fig. 1: A user's exploration interactions can be transformed into a time-series signal with $\{0,1\}$ representing her visiting status. (Time used to visit an element is marked as gray.) The signal sequence can further be transformed to a 2D wavelet power spectrum through continuous wavelet transform.

The *exploration-pacing* (*EP*) derived in this section is defined as *the density of a user's high-frequency visits to the visual elements during exploration.* A higher *EP* value suggests a user that rapidly moves from item to item. A lower value might reflect a user that explore individual elements for more time.

Measurable features: We compute the pacing metric based on the distribution of a users' interaction frequency. The features related to time from users' interaction logs (Section 3.2) to calculate the metric, *i.e.*, the *moments* and *duration of interaction*, and *exploration time*. We discard the other less relevant features, *e.g.*, *elements*.

Transform approach: One essential step to develop the metric is to select a function to transform a user's visit sequence over time to a visit-frequency sequence. One key intuition here is that merely averaging or binning time durations of a user's interaction with visual 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, *e.g.*, the wavelet transform, can readily transform duration data from the *time* domain to the *frequency* domain.

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):

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

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 called the wavelet power spectrum (Equation 6). This coefficient can be used to indicate the energy distribution of every moment in a user's exploration, *e.g.*, a moment with higher power values on high-frequency ranges indicate rapid visits to visual elements.

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

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.

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 an element, then it is marked as 1; otherwise if the user is not visiting any element, it is marked as 0. We sample the data every 0.1 second. If a user visited an element 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" [35], 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 $[f_{min}, f_{max}]$.

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

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

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, but can vary depending on visualization design. Precise modeling and parameters given visualization type and user behavior may be a valuable route for future work. For example, the range parameter can be changed to extract users' power density for other frequencies (*e.g.*, low frequency).

5 Metric Evaluation

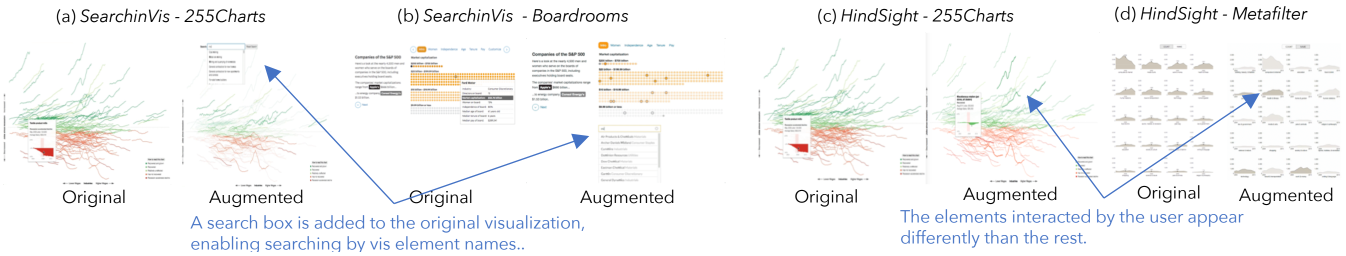
5.1 Interaction Data from Two Studies

We evaluate the two proposed metrics by applying them, together with other four metrics used in visualization literature, to interaction data collected from two previous studies from Feng *et al.* [14, 13]. ***SearchinVis*** [14] studied the effects of adding text-based search functionality to interactive visualizations on the web. Search enables a user to highlight elements in the visualization by typing keywords in a search box. ***HindSight*** [13] studied the effects of directly encoding people's personal interaction history in visualizations. During a user's exploration, visual elements the user interacts with are visually augmented to appear distinct from unexplored elements. Figure 2 shows visualizations from these studies whose data was adapted for the present work.

Participants and Conditions: Both studies include multiple experiments, one for each study-visualization pair, *e.g.*, *SearchinVis-255Charts*. All experiments had a between-subjects design, conducted on Amazon Mechanical Turk. Every participant was randomly assigned to interact with either the original or augmented version of a visualization. Table 2 shows the participant numbers of the control and experimental groups in each experiment.

Task and Procedure: Each participant was asked to analyze the visualization for as long as they liked before answering questions. The authors used an open-ended exploration task in order to simulate people's real-world explorations of web visualizations. Each participant went through several phases, including *introduction*, *exploration*, and *insight/strategy*. In the *introduction* phase, the participant was given instructions to interact with the visualization in any way they saw fit. Afterwards, the participant entered the *exploration* phase, where they could interact with the visualization without time limit. When the participant finished exploring, they entered the *insight/strategy* phase, where they answer questions about their findings aimed at highlighting possible differences in the control and experiment conditions.

Interaction Data: During the *exploration* phase, each participant's interactions with visual elements were recorded as $Ex(u) =$


(a) SearchinVis - 255Charts
(b) SearchinVis - Boardrooms
(c) HindSight - 255Charts
(d) HindSight - Metafilter
Original
Augmented
Original
Augmented
Original
Augmented
Original
Augmented

A search box is added to the original visualization, enabling searching by vis element names..

The elements interacted by the user appear differently than the rest.

Fig. 2: Four experiment datasets from two previous studies [13, 14] were used for the metric evaluation: *SearchinVis-255Charts*, *SearchinVis-Boardrooms*, *HindSight-255Charts* and *HindSifght-Metafilter*. Each dataset includes the interaction data of two groups of participants. Each group interacted with either the original or the augmented visualization.

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 [13, 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.

$(t_{start}, I_1, ..., I_k, t_{end})$, where t_{start} and t_{end} are the start and end time of the $exploration$ phase, and $I_i (0 \leq i \leq k)$ denotes every interaction (Table 1). Each interaction was recorded as $I = (t, a, E_i, d)$, where t is the start moment of the interaction, a is the interaction type (hovering), E_i is the set of interacted elements (there is one and only one element affected by each interaction), and d is the duration of the interaction.

Specifically, we select from these studies visualizations that were adapted from published visualizations on the web, including $255Charts$ and $Boardrooms$ from the $SearchinVis$ study, and $255Charts$ and $Metafilter$ from $HindSight$. We exclude the remaining visualizations in these studies in the present analysis, such as a bubble chart of exoplanet data (*i.e.*, condition 5 in [14]), because they were designed to test specific hypotheses in the prior studies, such as the effects of dataset familiarity and search behavior.

5.2 Applying Interaction Metrics: Case Studies

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

To evaluate the extent to which the proposed metric *exploration-uniqueness*(*EU*) shows different exploration patterns, we examine several individual exploration sessions in which users visited the same number of visual elements (*NVE*), but varied on the *EU* values. Then we visually examine whether an exploration with a higher metric value includes more elements rarely-visited by others.

In order to distinguish frequently- and rarely-visited elements, we calculate the percentage of visits for each element, and plot it as a baseline map (Figure 3), where each circle represents a visual element, 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 elements at the periphery of the visualization are frequently visited, while the elements in the middle are rarely visited.

As shown in Figure 3, user A, B and C from ***SearchinVis-255Charts*** visited the same number of elements (25) during their explorations. However, these explorations vary at *EU*, *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 elements on top of the baseline map, where each circle represents a visit, with its size corresponding to how long the user spent visiting this element. Visually comparing the visit maps from the user A, B and C, we find that the elements visited by A (lower *EU*) are mostly at the periphery of the visualization, which are frequently visited by other users, according to the baseline map. Instead, user C (higher *EU*) visited more elements located at the lower-middle part of the visualization, which are rarely visited by other users in the study. The elements visited by user B (middle *EU*) include some frequently-visited and some rarely-visited elements.

Similarly, user G, H and I from ***HindSight-255Charts*** visited the same number of elements (22) with varying *EU*, *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 G, H and I, we see that the elements visited by G (lower *EU*) are mostly at the periphery of the visualization (*i.e.*, frequently-visited elements). User I (higher *EU*) visited more elements at the middle part of the visualization, which are rarely visited by other users in the study. The elements visited by user H (middle *EU*) include both frequently-visited and rarely-visited elements.

We also find that *exploration-pacing* (*EP*) 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 metrics provide additional insight in experiment analyses?

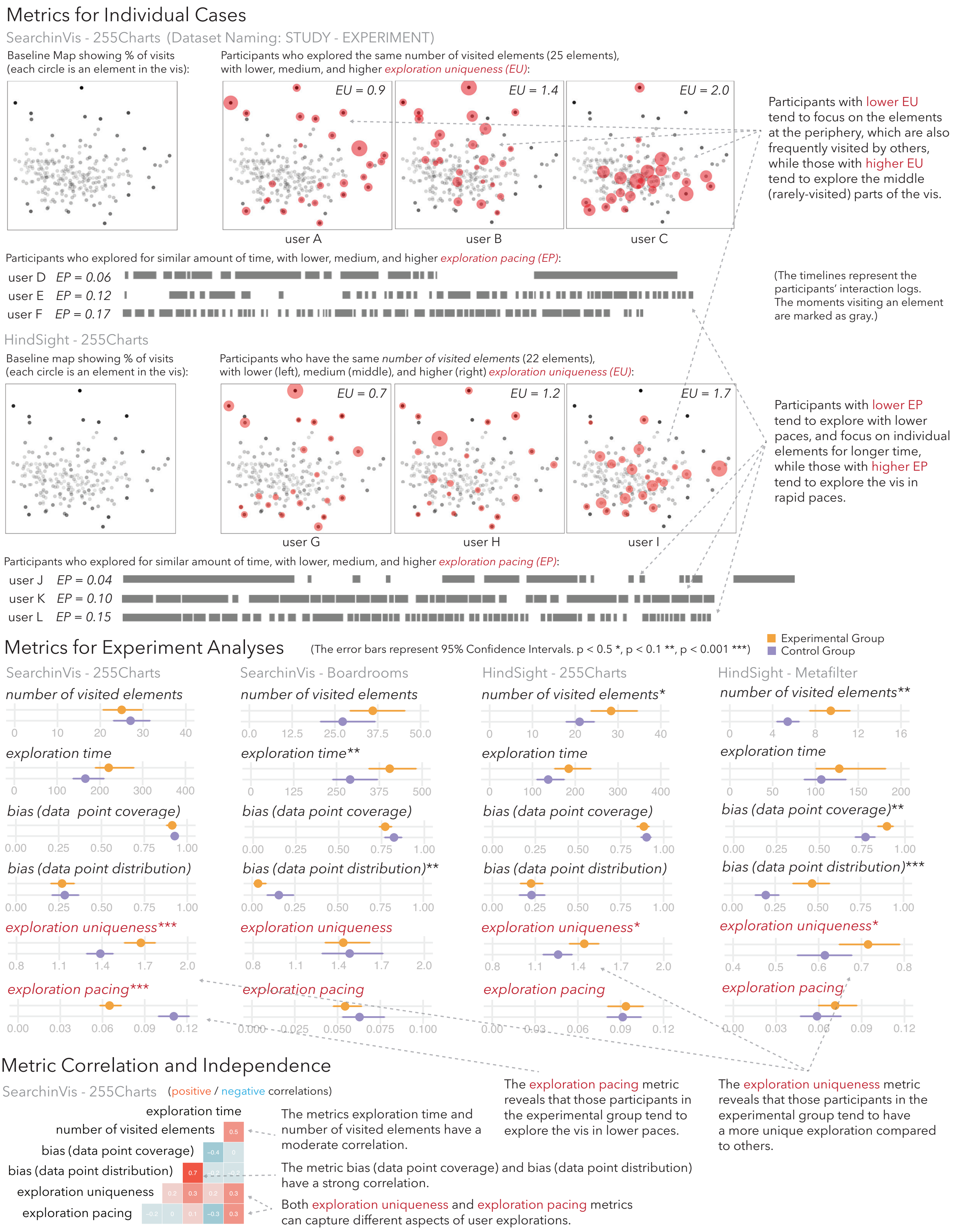
After examining individual cases to check the validity of the proposed metrics, we explore their effectiveness on one of the potential application scenarios, *i.e.*, to show the impact of visualization designs on user interaction behavior. Specifically, we compare the metric values between two user groups (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-elements*, *exploration-time*, *bias-data-point-coverage*, and *bias-data-point-distribution*.

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.

255Charts: We filtered out 5 (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 show significantly more unique explorations (***exploration-uniqueness***), and had fewer rapid-pace visits (***exploration-pacing***). The experimental group had a higher *exploration-uniqueness* 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 $W = 1165, p = 0.0002$,



Metrics for Individual Cases

SearchinVis - 255Charts (Dataset Naming: STUDY - EXPERIMENT)

Baseline Map showing % of visits (each circle is an element in the vis):

Participants who explored the same number of visited elements (25 elements), with lower, medium, and higher *exploration uniqueness (EU)*:

Participants with lower EU tend to focus on the elements at the periphery, which are also frequently visited by others, while those with higher EU tend to explore the middle (rarely-visited) parts of the vis.

Participants who explored for similar amount of time, with lower, medium, and higher *exploration pacing (EP)*:

(The timelines represent the participants' interaction logs. The moments visiting an element are marked as gray.)

HindSight - 255Charts

Baseline map showing % of visits (each circle is an element in the vis):

Participants who have the same *number of visited elements* (22 elements), with lower (left), medium (middle), and higher (right) *exploration uniqueness (EU)*:

Participants with lower EP tend to explore with lower paces, and focus on individual elements for longer time, while those with higher EP tend to explore the vis in rapid paces.

Participants who explored for similar amount of time, with lower, medium, and higher *exploration pacing (EP)*:

Metrics for Experiment Analyses

(The error bars represent 95% Confidence Intervals. p < 0.5 *, p < 0.1 **, p < 0.001 ***)

Metric Correlation and Independence

SearchinVis - 255Charts (positive / negative correlations)

The metrics exploration time and number of visited elements have a moderate correlation.

The metric bias (data point coverage) and bias (data point distribution) have a strong correlation.

Both exploration uniqueness and exploration pacing metrics can capture different aspects of user explorations.

The exploration pacing metric reveals that those participants in the experimental group tend to explore the vis in lower paces.

The exploration uniqueness metric reveals that those participants in the experimental group tend to have a more unique exploration compared to others.

Fig. 3: We applied the proposed *pacing* and *uniqueness* to two previous studies, with results suggesting that they capture different facets of user explorations of visualizations on the web.

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 visual elements.

Boardrooms: We filtered out 11 (from 96) users who did not interact with any visual element. 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 [14]. 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 255*Charts* and *Boardrooms* visualizations.

5.3.2 Study 2: HindSight

We apply existing and proposed metrics to the interaction data from the two experiments of the study *HindSight* [13], 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 (from 116) user who did not interact with any visual elements. We found significant differences on the metric ***number-of-visited-elements*** and ***exploration-uniqueness*** between the two user groups, while the other metrics are similar between groups. Specifically, the experimental group visited more visual 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 [13]. The experimental group has a higher *exploration-uniqueness* 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 significant differences on the metric ***number-of-visited-elements***, ***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-elements* 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 [13]. The experimental group also has a higher *exploration-uniqueness* 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.

5.4 Metric Correlation and Independence

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

We compute correlations between each pair of the metrics across all experimental datasets (Figure 3), *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-elements* and *exploration-time*, in *SearchinVis-255Charts* ($r = 0.57, p < 0.001$).

Both *exploration-uniqueness* and *exploration-uniqueness* have weak correlations or less with the other metrics 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 proposed uniqueness and pacing metrics aim to reveal new facets of how people interact with visualizations. To examine this claim, we applied the proposed metrics, together with metrics from prior work, to interaction data from prior information visualization studies. The results suggest that, first, the proposed metrics do reveal new characteristics of peoples' exploration behavior in visualizations. Second, the proposed metrics can be used as target metrics in comparative experiments, *i.e.*, quantitative analysis comparing control and treatment groups. Third, the proposed metrics are also 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 for exploratory visualization design and research.

In the analysis of the study *SearchinVis*, we found that the results differ 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 interaction. Larger interaction sets, then, may pose new challenges and opportunities for measures of interaction behavior.

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 visual 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. The proposed metrics thus lead to new questions, such as how much bias measures should weight revisits against the breadth of peoples' exploration?

6.1 Potential Applications of Interaction Metrics

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, 13, 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. Behavioral patterns of exploration can be used to infer users' characteristics (*e.g.*, locus of control [29, 30]), reasoning processes [12], and insight generation [16]. 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 visualization approaches for analyzing interaction data is that detailed information can be logged during user interaction, as shown in previous works that center on visualizing interaction logs *e.g.*, Blascheck *et al.* [3, 4]. 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 interaction log visualization, for example ordering or coloring 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 [30], 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 visual element (Equation 2), the vectors of TF-IDF values (Equation 3), or the highest level exploration uniqueness metric values. Similarly, both the two-dimensional wavelet power spectrum and the corresponding pacing metric can be used as features for machine learning algorithms.

6.2 Benefits and Tradeoffs of Interaction 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, 33, 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-elements*. 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-elements* 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.*, *exploration-uniqueness*, *exploration-pacing* and *bias-data-point-coverage*.

Knowledge coverage (cognition level): *How much additional knowledge does the metric cover given other metrics?* This dimension evaluates whether proposed metrics can uncover characteristics of users' exploration that other metrics do not capture. By examining correlations between metrics, we found that both proposed metrics, *exploration-uniqueness* and *exploration-pacing* may reveal different perspectives from other metrics.

7 Future Work & Conclusion

Despite their advantages, there are also limitations with the proposed metrics. For example, the uniqueness metric focuses on an interaction set, ignoring ordering effects, which may be another source of exploration diversity. The visualizations tested thus far are also constrained in interaction scope, implying that more complex interaction schemas may require more sophisticated approaches for developing useful metrics. Further, there are currently no established guidelines in the visualization community for evaluating other characteristics of proposed metrics, such as consistency, discriminability, and reliability. Given the changing landscape of visualization in the world, addressing challenges such as these may be fruitful areas for future work.

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. 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.

Acknowledgments

We would like to acknowledge the reviewers, who gave excellent feedback and suggestions, Emmanuel Agu, who shared valuable insights on analyzing time-varying data, Hua Guo, who pointed out closely-relevant literature on user behavior measuring needs, and the multiple journalists who created the original works we replicate in these experiments.

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] 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.

[14] M. Feng, C. Deng, E. M. Peck, and L. Harrison. The effects of adding search functionality to interactive visualizations on the web. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 137. ACM, 2018.

[15] 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.

[16] 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.

[17] S. Haroz, R. Kosara, and S. L. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1191–1200. ACM, 2015.

[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):1189–1196, 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, 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] N. Mahyar, S.-H. Kim, and B. C. Kwon. Towards a taxonomy for evaluating user engagement in information visualization. In *Workshop on Personal Visualization: Exploring Everyday Life*, volume 3, page 2, 2015.

[27] S. Malik, F. Du, M. Monroe, E. Onukwugha, C. Plaisant, and B. Shneiderman. Cohort comparison of event sequences with balanced integration of visual analytics and statistics. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 38–49. ACM, 2015.

[28] 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.

[29] 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.

[30] 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.

[31] C. Plaisant. The challenge of information visualization evaluation. In *Proceedings of the working conference on Advanced visual interfaces*, pages 109–116. ACM, 2004.

[32] 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.

[33] 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.

[34] K. Rodden, H. Hutchinson, and X. Fu. Measuring the user experience on a large scale: user-centered metrics for web applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 2395–2398. ACM, 2010.

[35] A. Roesch and H. Schmidbauer. *WaveletComp: Computational Wavelet Analysis*, 2018. R package version 1.1: `https://CRAN.R-project.org/package=WaveletComp`.

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

[37] 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.

[38] A. Sarvghad, M. Tory, and N. Mahyar. Visualizing dimension coverage to support exploratory analysis. *IEEE transactions on visualization and computer graphics*, 23(1):21–30, 2017.

[39] 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.

[40] 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.

[41] J. Walny, S. Huron, C. Perin, T. Wun, R. Pusch, and S. Carpendale. Active reading of visualizations. *IEEE transactions on visualization and computer graphics*, 24(1):770–780, 2018.

[42] 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.

[43] 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.

[44] 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.

[45] E. Zgraggen, S. M. Drucker, D. Fisher, and R. DeLine. (s— qu) eries: Visual regular expressions for querying and exploring event sequences. 2015.

[46] 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.