LaNe Plot: A Visual Fingerprinting Technique for Sequential Data

Harith Rathish, Ginés Carreto Picón, and Hans-Jörg Schulz

Aarhus University, Denmark

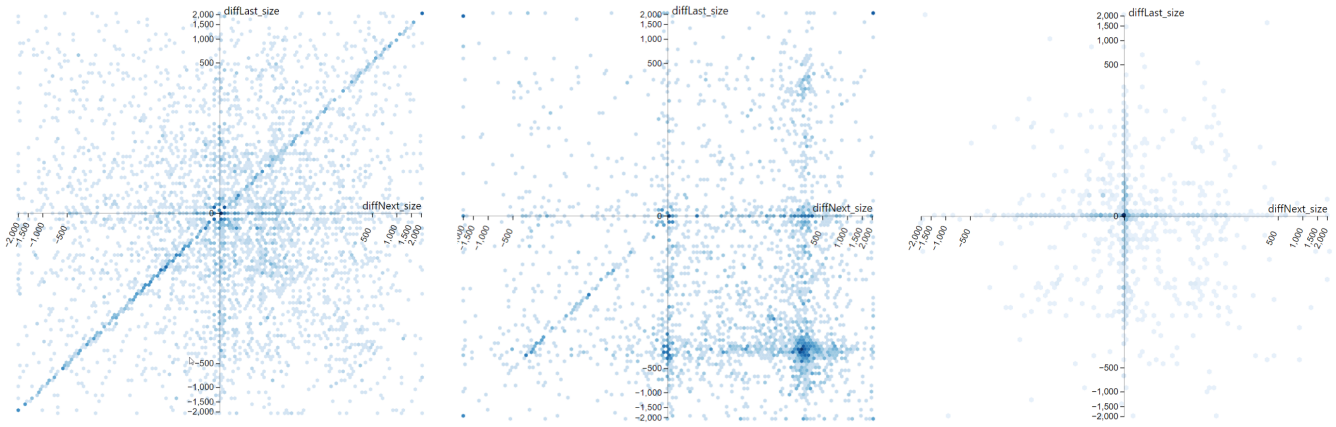

Figure 1: *LaNe plots of last 5000 revisions of select Wikipedia articles. Note that the axes are log-scaled, and the number of points within each bin is encoded by its color. Article (a) is highly subject to vandalism and debate compared to (b) and (c), as indicated by the intensity of the $y = x$ diagonal line. Most revisions in article (b) added items to a list of endorsements with around 300 bytes for each entry, as indicated by the cross-shaped pattern centered at around (300,-300). In article (c), most revisions updated existing tennis records without changing the article size, as indicated by the cross pattern centered at (0,0)*

Abstract

Visual summaries of sequential data are often used to identify common trends at a glance. In this poster, we propose a visualization technique to fingerprint sequential data by showing the difference between contiguous data points. For each data point in the sequence, we visualize the difference between itself and the last data point as well as the next data point. As an application, we visualized the revision histories of Wikipedia articles to demonstrate the exploratory value of this technique.

CCS Concepts

• ***Human-centered computing*** → ***Visualization techniques;***

1. Introduction

Sequential data is common in many data domains ranging from event sequences to time series to graph traversals and beyond. When such datasets become large, it is useful to have a dense overview of significant patterns within a fixed screen space. Such representations also allow the analyst to compare multiple datasets *at a glance*. In this poster, we contribute to this class of visualizations by presenting the **LaNe plot** (**La**st and **Ne**xt) - a technique for visually fingerprinting sequential data based on the *last and next data points* in the sequence.

2. Design

Consider a sequential dataset with quantitative values $d_1, d_2, d_3, ... d_n$. For any arbitrary d_i, where $1 < i < n$ we calculate its difference between its adjacent values as follows:

$$diffLast_i = d_{i-1} - d_i \qquad diffNext_i = d_{i+1} - d_i$$

We now map each d_i onto a point in a scatterplot, encoding $diffLast_i$ and $diffNext_i$ onto position, as shown in Figure 2. This

DOI: 10.2312/evp.20241087

mapping places the points into one of four quadrants, based on the following criteria:

- If $diffLast_i, diffNext_i > 0$, then d_i is on a *local minima*.
- If $diffLast_i, diffNext_i < 0$, then d_i is on a *local maxima*.
- If $diffLast_i < 0$ and $diffNext_i > 0$, then d_i is in between an *increasing* trend.
- If $diffLast_i > 0$ and $diffNext_i < 0$, then d_i is in between a *decreasing* trend.

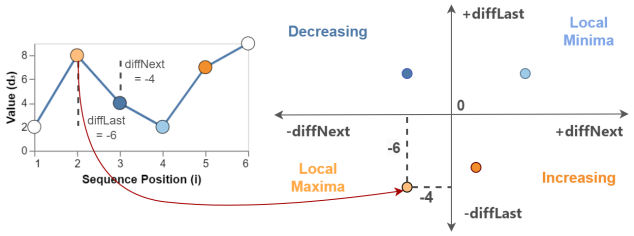

Figure 2: *Mapping data points in a sequence onto LaNe plot using the difference between adjacent data points*

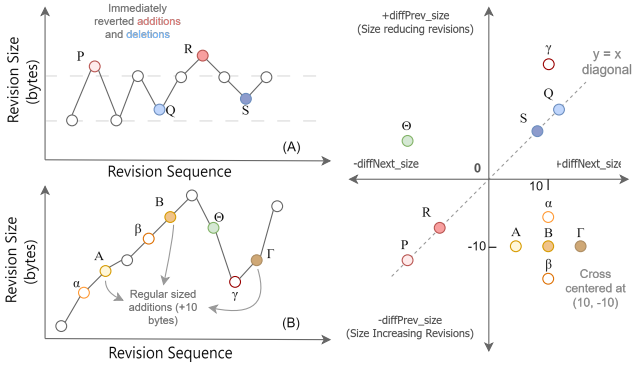

Figure 3: *Patterns in the LaNe plot corresponding to the type of Wikipedia revisions. In sequence (A), the labelled revisions, which were then immediately reverted, lie on the $diffLast = diffNext$ diagonal on the LaNe plot. In sequence (B), revisions A, B, and Γ are all 10 bytes larger than their previous revisions, and lie on the line $diffLast = -10$ on the LaNe plot. Meanwhile, the revisions α, β and γ, which precede them, lie on the line $diffNext = 10$. Both combined forms the cross centered at $(10, -10)$*

3. Application: Wikipedia Revisions

We applied this technique to the sequence of the sizes of the last 5000 revisions of select Wikipedia articles (see Figure 1). The articles were curated from the list of the most edited Wikipedia articles [Wik], along with a few articles about 2016 US Election. For the sake of visibility, we used a 2D density plot instead of a scatterplot. The articles exhibited different visual features depending on their content, and we discuss the meaning of these features below. This use case is similar to that of history flow visualizations by Viégas et al. [VWD04]. However, our work spans an even larger number of revisions per article.

3.1. Diagonal pattern ($y = x$)

The diagonal line defined by $y = x$ is comprised of all points d_i for which $diffLast_i = diffNext_i$. In the context of Wikipedia revisions, we observed that most revisions (78 %) on this diagonal were *immediately reverted* revisions, and thus the sizes of the revisions before and after revision i were equal (see Figure 3 (A)). These revisions were often a form of vandalism, especially on potentially controversial topics, such as 'Jesus' in Figure 1(a). On the contrary, this feature was near absent in Figure 1(c) on "ATP Tour records", whose content tended to be more objective with fewer schools of thought.

Interestingly, in Figure 1(b) for "Endorsements in the 2016 Rep. Party presidential primaries", the diagonal is more prevalent in the bottom-left quadrant for local maxima. These were edits which caused a peak (maxima) in the revision size, and were then later reverted. This is because most of the immediate reversions on this article were on edits that *added* incorrect endorsements, compared to those that removed them.

3.2. Cross-shaped pattern

All d_i for which $diffLast_i = \gamma$ will lie on a horizontal line that intersects the y-axis at γ, and vice-versa for $diffNext_i = \gamma$. Both lines combined will form a cross-shaped pattern that has its center at (γ, γ) (see Figure 3 (B)).

For Wikipedia articles, we hypothesize that the cross-shaped pattern indicates the presence of *a heavily edited table or list*. For example, in Figure 1(b) we see the cross at around (-300, 300). This is because in this article, the addition of each endorsement increased the article's size by roughly 300 bytes and thus $\gamma = 300$. Note that the cross is present only on the bottom-right quadrant because there were more revisions that *added* endorsements than those which removed.

Similarly, in Figure 1(c), the cross is along the axes, with the center at (0,0). This is because most edits on the "ATP Tour Records" article updated existing records, which rarely changed the size of the article, thus $\gamma = 0$. This pattern is near absent in Figure 1(a) since the 'Jesus' article did not have any heavily edited lists or tables.

4. Conclusions and future work

In this work, we have presented a dense overview visualization for large sequential datasets. We applied this technique in a real-world context with Wikipedia data and saw distinct visual features based on the content types of articles. Moving forward, we aim to apply the LaNe plot to other sequential datasets and test its effectiveness with a user study, broadening its potential impact.

5. Acknowledgment

We acknowledge the support of Aarhus University Research Foundation, whose generous funding made this research possible.

References

[VWD04] VIÉGAS F. B., WATTENBERG M., DAVE K.: Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2004), CHI '04, Association for Computing Machinery, p. 575–582. URL: https://doi.org/10.1145/985692.985765, doi:10.1145/985692.985765. 2

[Wik] WIKIPEDIA: Pages with the most revisions. Accessed on 14th Feb 2024. URL: https://en.wikipedia.org/wiki/Special:MostRevisions. 2