Supplemental Material: Designs from Earlier Iterations
Understanding User Behaviour in Action Sequences: from the Usual to the Unusual

Phong H. Nguyen, Cagatay Turkay, Gennady Andrienko, Natalia Andrienko, Olivier Thonnard, and Jihane Zouaoui

1 Overview of Sessions

Before the current design, we have made several attempts. Note that the goal is to provide an overview of sessions through three attributes:

- anomaly score
- one numerical attribute
- one categorical attribute

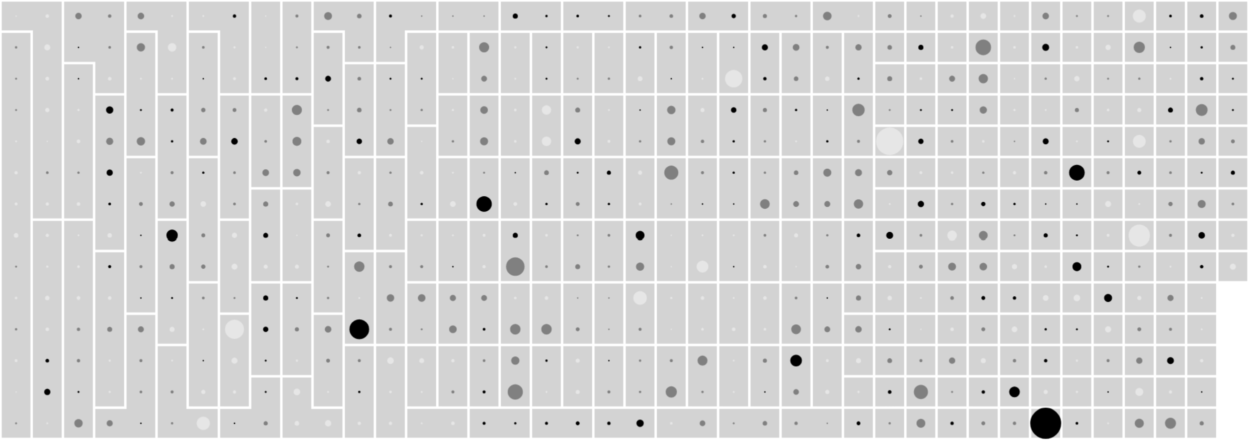


Figure 1: Score is encoded with color lightness. A numerical attribute is encoded with area size. Items are followed a sequential, alternative up-down layout. Groups are separated by space and colored background.

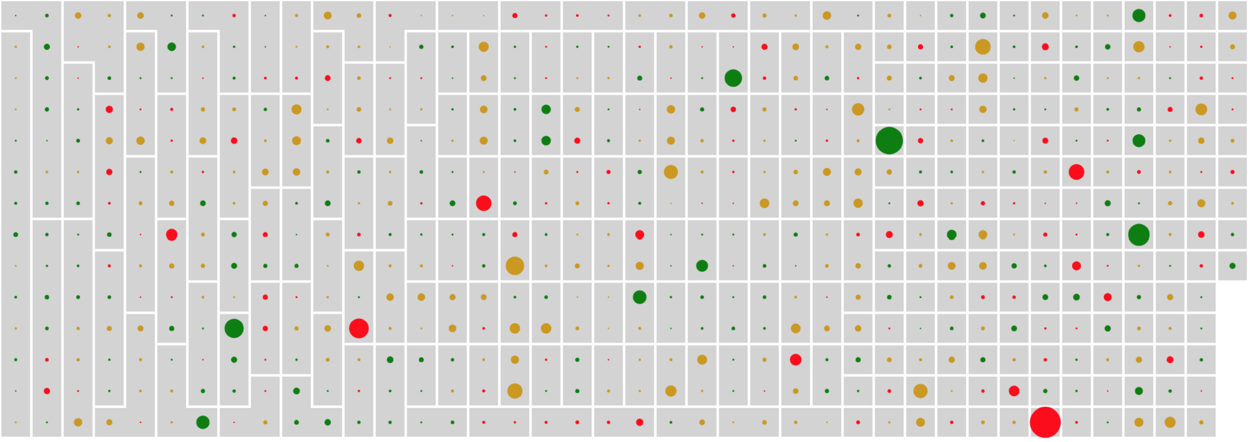


Figure 2: Colors are shown with three different hues: green for low scores, yellow or medium scores, and red for high scores. This encoding reflects a conventional use by security analysts and strongly highlights the high-scoring ones. However, it is sensitive to an arbitrary selection of threshold values.

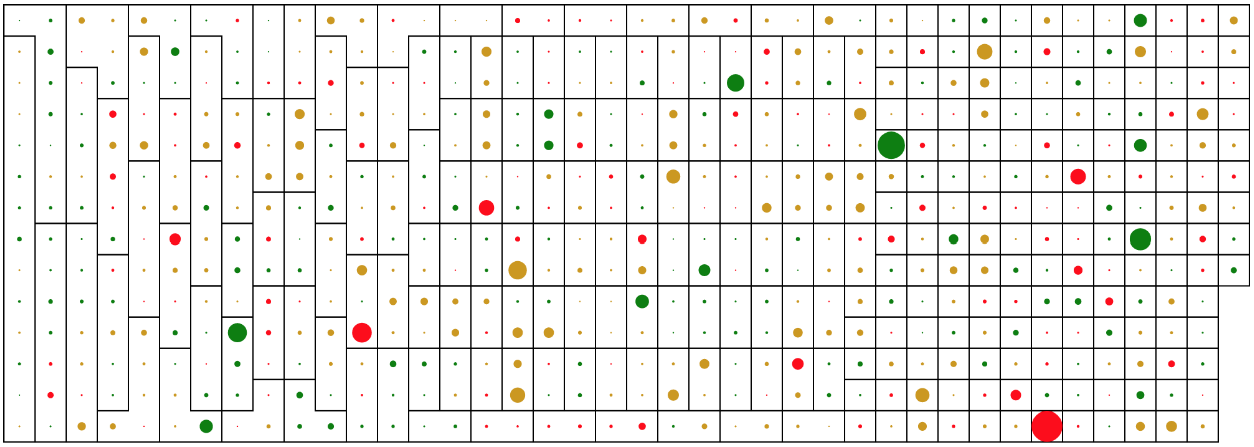


Figure 3: Groups are contained by colored border. It could be weaker than using colored background, but the visualization is cleaner.

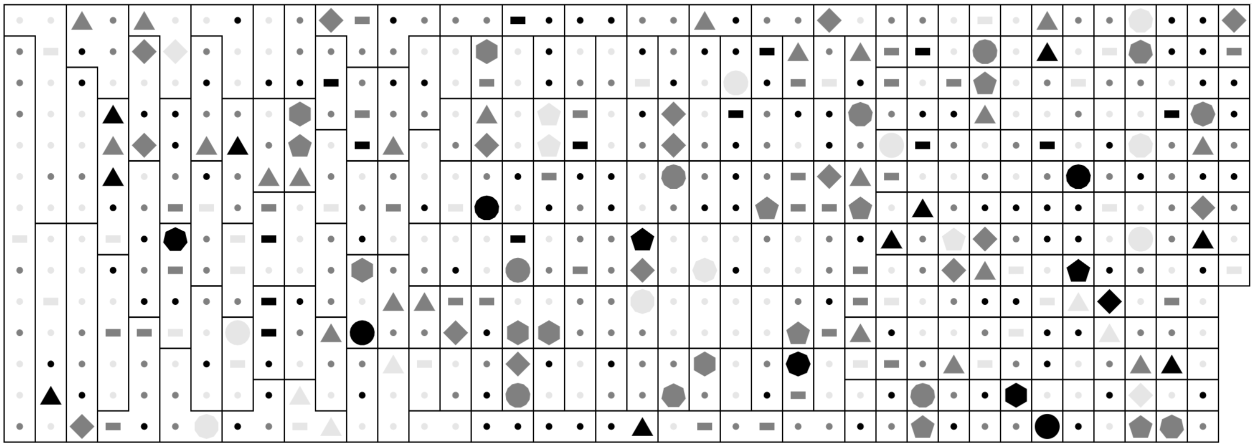


Figure 4: Shapes are used to encode a numerical attribute. Using size, such as circle radius, suffers from extreme range values. As absolute comparison of session lengths might not be essential, this design splits the attribute values into bins and maps to the sides of a shape. For example, a dot ≈ 10, a dash ≈ 20, a triangle ≈ 30, etc.

2 Overview of Action Sequences

Before applying sequential pattern mining to extract patterns, we have designed a compact view to reveal the *following* or *trailing* relationship between actions within a session. For instance, how likely sequence ABC will be followed by D (becoming ABCD) or E (becoming ABCE)? A clear visual evidence of repeating patterns supports us in algorithmically mining them.

Figure 5: A sub-sequence is represented by a rectangle with multiple equal-sized columns, each representing an action in the sub-sequence, from left to right. The rectangle height indicates its number of occurrences in the entire dataset; i.e., the total number of times that sub-sequence is found in all sessions.

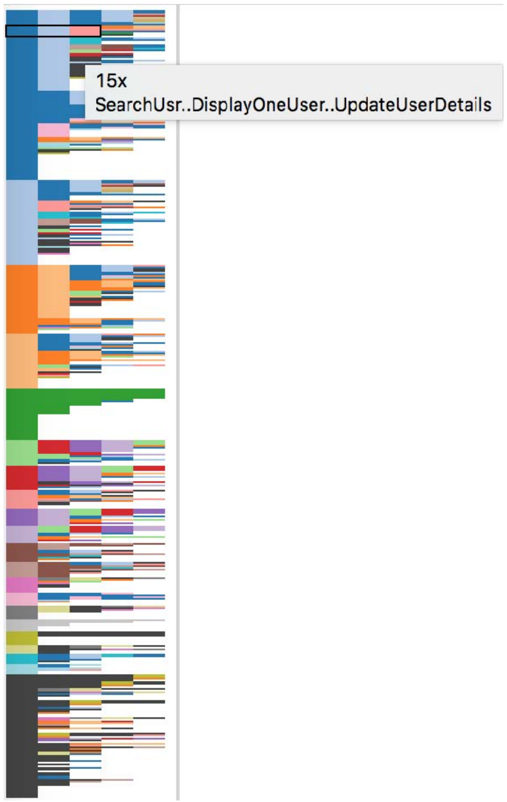


Figure 6: Sub-sequences are then aggregated in a tree metaphor representation so that the ones sharing the same ancestor are placed next together. For instance, ABCD is placed right above or below ABCE so that ABC can be seen as their parent. As a result, the sequence tree consists of multiple equal-sized columns (5 in this figure), each representing a color-coded action. The first column shows sub-sequences with one action, which actually are just the actions themselves. The combination of the first and the second columns shows sub-sequences with two actions, and so on. We can see that **Search User** is the most common action, but what do users commonly do after a search? The actions following are **Display One User** or **Search User** again, i.e., after displaying a user, the users usually search again or update details.

3 Overview of Mined Activities

To gain understanding into the mining results, we have started with a scatter plot, showing different pattern attributes using two axes and circle size. This is to explore patterns with multiple dimensions, namely:

- the number of occurrences
- the number of sessions having the pattern
- the number of users performing the pattern
- the median score of all sessions having the pattern (the action frequency score from the model)
- the median length (number of actions) of all sessions having the pattern
- the median duration (time) of all sessions having the pattern

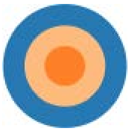


Figure 7: A "ring" representation of a pattern of actions. Each ring represents an action type, ordering outward.

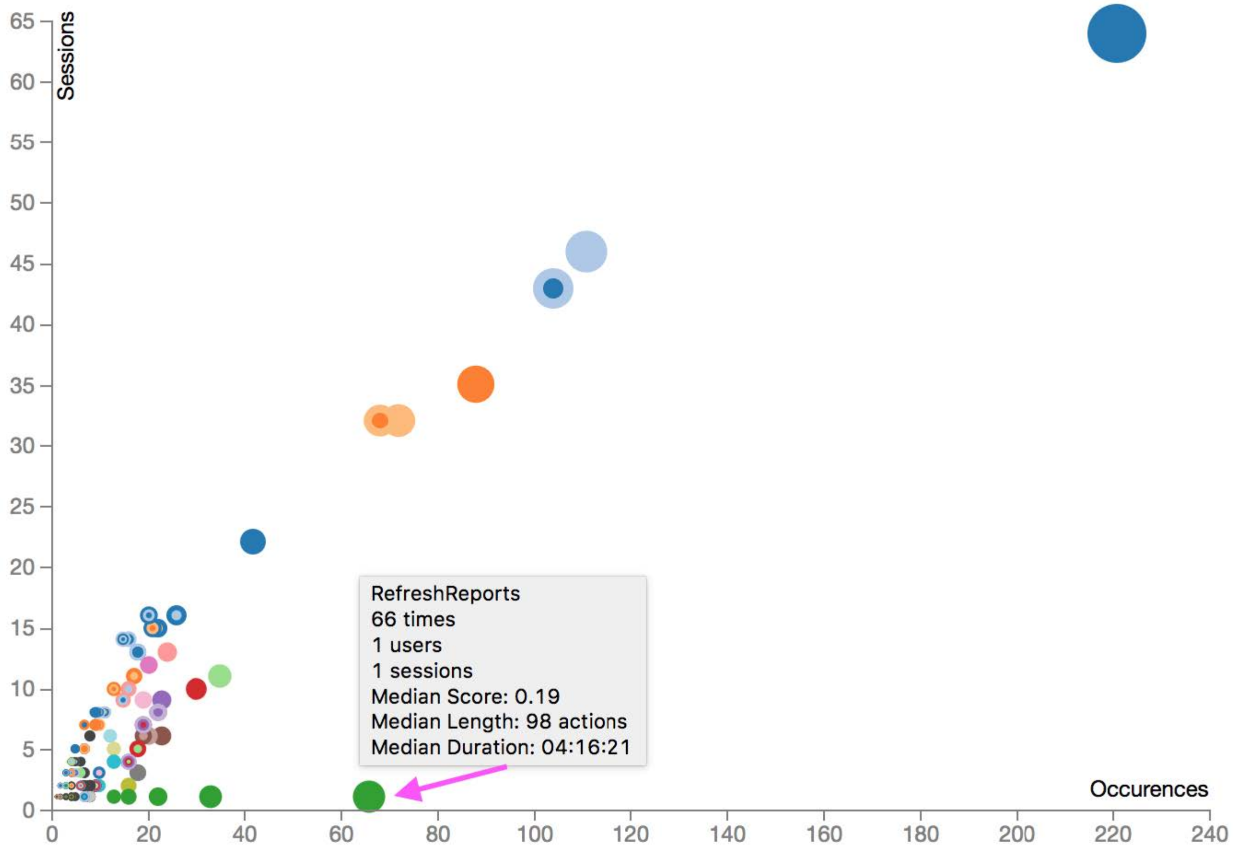


Figure 8: Changing x-axis to *Occurrences* and y-axis to *Sessions*, we expect these two dimensions to be positively correlated and circles placed along a diagonal line. Most of patterns follow this but the green ones.

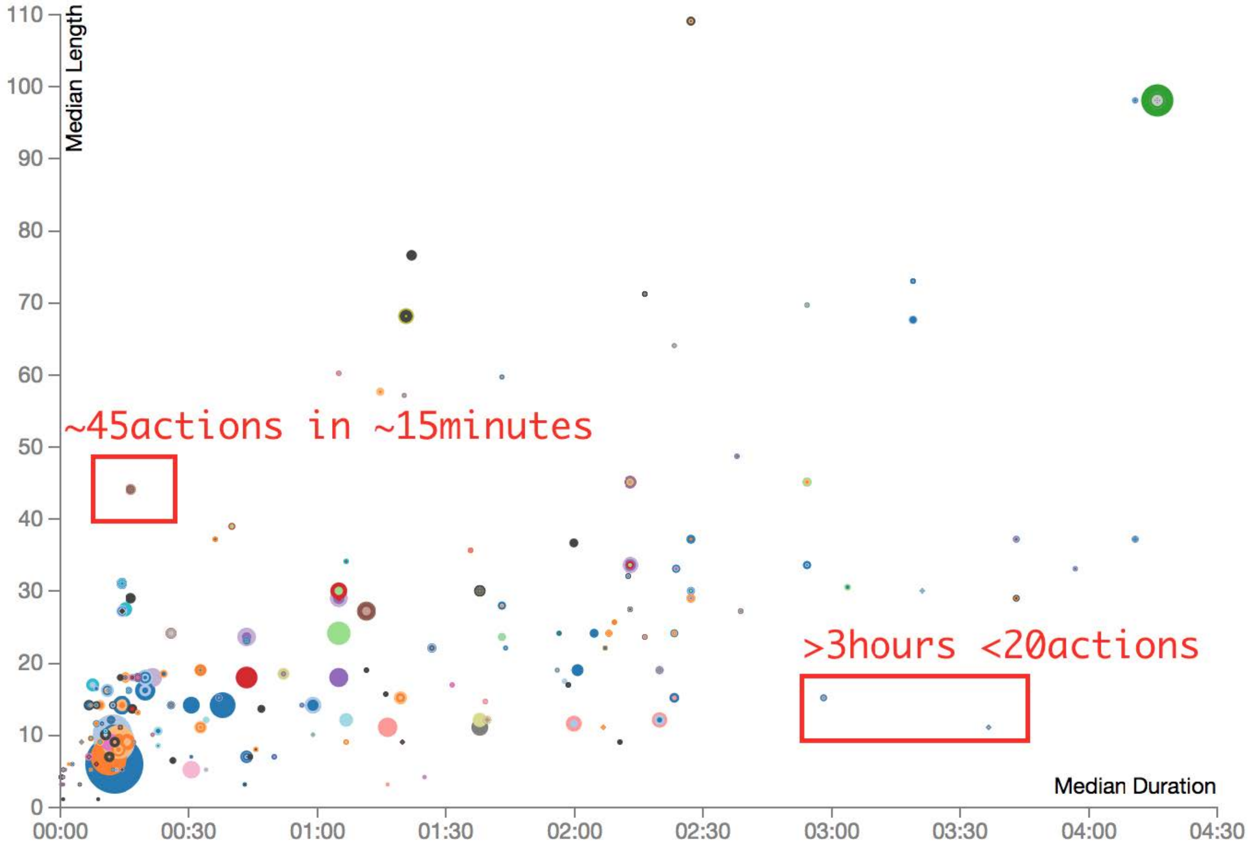
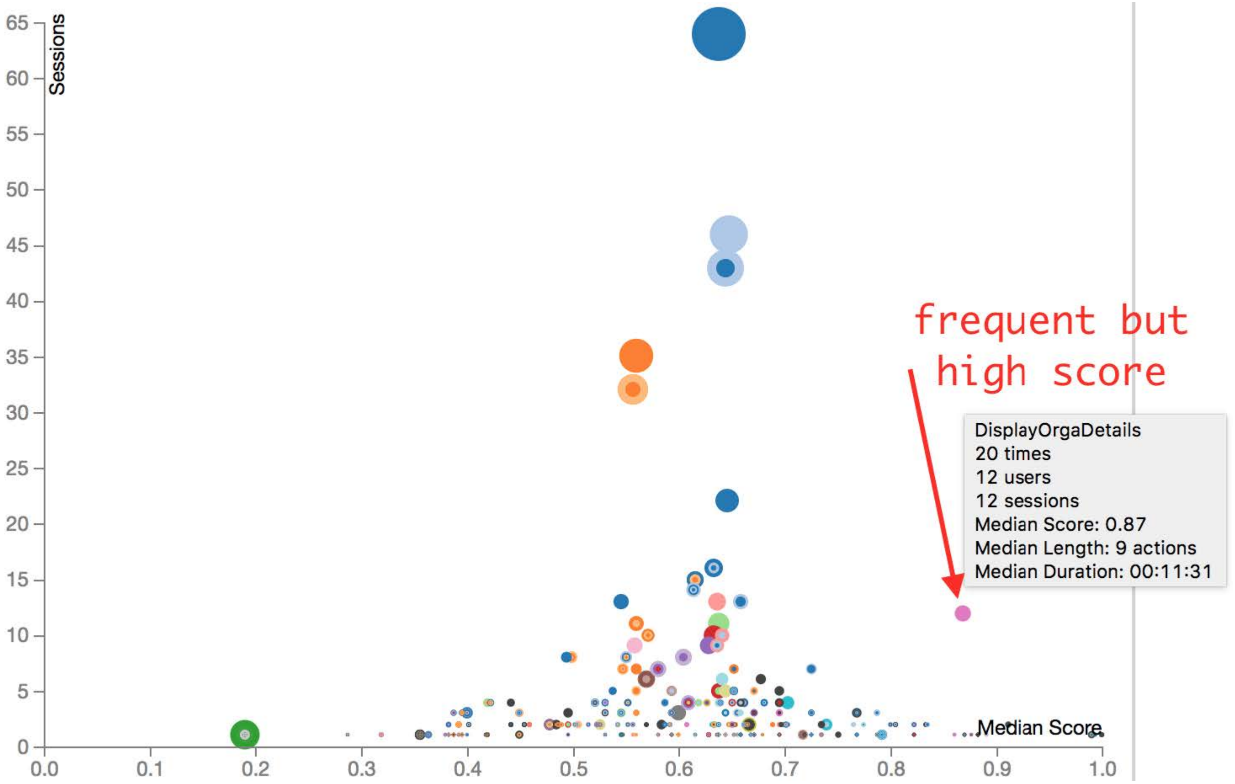


Figure 9: Changing x-axis to *Median duration* and y-axis to *Median length*, we still expect a positive correlation, nonetheless, we can spot some exceptions as highlighted.



frequent but high score

DisplayOrgaDetails
20 times
12 users
12 sessions
Median Score: 0.87
Median Length: 9 actions
Median Duration: 00:11:31

Figure 10: Changing x-axis to *Median score* and y-axis to *Sessions*, we can see a quite normal distribution of score. Frequent patterns (big circles) seem to have average score and some less frequent patterns have higher score. An exception is a pink circle (DisplayOrgaDetails) that is occurred 20 times in total of 12 sessions (all by different user) and received a median of 0.87.