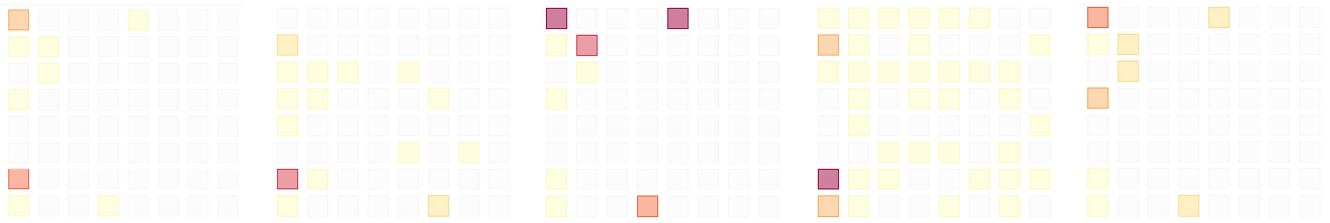


(a) current activities of an entity (b) historical activities of this entity (c) risk against the historical activities of this entity (d) historical activities of the peers (e) risk against the historical activities of the peers

Figure 3: The topic grids. The self risk in (c) is derived from comparing the current activities (a) and the historical activities (b) of a specific entity. The peer risk in (e) is derived from comparing the current activities (a) and the peers' activities (d) of a specific entity.

(b) and (f), topics are represented by the most relevant keywords, encrypted. Topics close to each other may share the same representative keyword. The SD algorithm follows to generate the topic grids and visualize different metrics about the behavior of a user (Figure 3).

When not directly displaying the detail keywords about a topic, the topic grids requires less space. At the same time, the human expert still can easily keep track of the topics based on their indexes over all dimensions and compare the difference between different sets of topic grids. Human interaction, which is the ultimate goal of the uniform placement of the data points, can be done more easily on the topic grids than on the raw dimension reduction output as in Figure 1 (a)-(d). For example, the mouse over event on a grid pops up the topical summary, and the click event to overlay the detailed topical activities.

It is also useful to monitor the behavior change over time. In such cases, we reserve a dimension in $\mathcal{L}$ as the time axis. For a 2D space $\mathcal{L}$, a 1D version of SD algorithm is applied to maintain the point-wise topology. The cumulative activities have a shape of curtain. Meanwhile, we can pile up the 2D topic grids on the time axis over the 3D $\mathcal{L}$, as shown in Figure 4. With normal or usual behavior, it is expected to see the consistent hot grids at the same locations over time.

4. FUTURE WORK

In addition to the cyber security domain, the topic grids can be applied to other domains having free-form text logs to analyze the behavior described by the logs. Some possible use cases include e-commerce, credit card transaction, customer service, or others with large volume of behavioral data to be analyzed.

It is also possible to apply the topic grids to the structured data, on which an arbitrary clustering algorithm can generate cluster centers. The data points are then organized into these cluster centers, the same way we use the topic to represent the log entries related to it.

5. REFERENCES

[1] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

[2] P. Geurts, D. Ernst, and L. Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

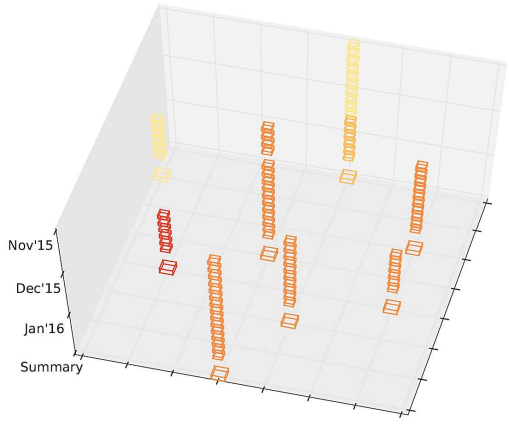
[3] K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

[4] J. B. Tenenbaum, V. D. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[5] W. S. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17(4):401–419, 1952.

[6] L. Van der Maaten and G. E. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

(a) Topic curtain



(b) Topic shower

Figure 4: Other formats of the topic grids