Additional Information for paper Interactive Time Series Exploration Powered by the Marriage of Similarity Distances

Table 1: **List of notations**

Symbol	Definition
D	Data set
X	Time series
$(X_p)_j^i$	Subsequence of X of length i starting at position j
G_k^i	k^{th} similarity group of length i
R_k^i	Representative k of a similarity group of length i
ST	Similarity Threshold
$\mathcal{L}$	Length of Subsequence

This document contains additional information to supplement the material presented in the paper.

Figure 1 shows the way warping path is calculated using dynamic programming for dynamic time warping distance as described in Section 2 of paper.

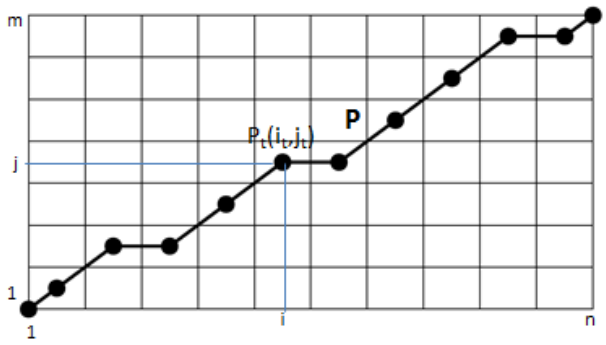

Figure 1: **Warping Path**

Table 1 describes the list of notations used through out the paper.

Figure 2 describes the Representative Space as described in Section 3 of the paper. It "represent" each group constructed over a data set D by only *one single* sequence, namely, the group's representative. We collect the representatives for all groups over G into a collection, called the **Representative Space** ($\mathcal{R}$-Space).

Discussion: DTW Clustering in ONEX framework. We now show that using DTW for ONEX similarity group formation would require us to guarantee the triangle inequality for DTW, which is still an open research problem [1], [2] due to the non-metric nature of DTW. This, along with the efficiency of ED for clustering compared to DTW, confirms our ED-DTW design choice of the framework.

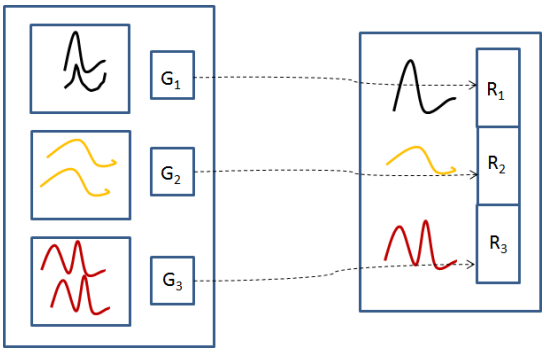

Figure 2: **ONEX Base Intuition**

Let's assume that some clustering method were to exist that would create clusters as in our Def. 8 of the paper in Section 3, i.e. it is guaranteed to place all subsequences in similarity groups for which the following conditions hold:
(1) $\overline{DTW}(X,Y) \leq ST$, where X and Y are any sequences in the same group and ST is the given similarity threshold.
(2) $\overline{DTW}(X,R) \leq ST/2$, where R is the representative of a group and X is any sequence in that group.
In other words, our ONEX framework would have to work now with an "adjusted" Lemma 2 (Section 3 of paper), where ED would be replaced by DTW. For this important foundation of our ONEX methodology to still be applicable, we would have to prove the following:
If $\overline{DTW}(Y,Y') \leq ST/2$ *and* $\overline{DTW}(X,Y) \leq ST/2$ *then we have* $\overline{DTW}(X,Y') \leq ST$, which unfortunately corresponds to the triangle inequality for DTW. □
In conclusion, using a clustering methodology based on DTW would impede our framework's functionality, rendering its core formal foundation unproven.

1. DISCUSSION OF ONEX BASE

1.1 ONEX Base Construction

While in the paper we have introduced a viable and robust solution for ONEX base construction as described in Section 4, the question arises if alternate clustering methods based on ED could equally be utilized in the context of our framework. The key observation here is that any clustering solution we employ must observe our core group requirements,

namely, they must produce clusters with a maximum diameter equal to ST and have centers (our representatives) whose ED to any sequence in the group is less than $ST/2$. While the well-known kmeans algorithm appears to be a contender on rst sight, we note that there are key dierences: (1) K-means must know the xed number of clusters in advance, while we give a lower bound to the number of groups, but we expect it to grow. (2) K-means is batch oriented, while Algorithm 1, after a small setup batch mode, is online. In a sense, Algorithm 1 is similar with online nearest neighbor clustering or the on-line k-center problem [1].

1.2 ONEX Base Maintenance under Updates

Next, we sketch strategies for accommodating insertions and deletions of time series. If a time series is deleted, the groups can be updated to remove any sequences with that specific time series id. These groups don't need to be reconstructed. However, the representatives (and their envelopes [3]) have to be re-computed according to the sequences that remain in the group. We keep the locations of all objects in an index, thus if we delete an object, we can nd it in constant time and set it to NULL. In short, deletions, so long as they are relatively rare (say less than 10effect on speed or accuracy. If a time series is inserted into the dataset, some groups have to be updated. When possible we use strategies that avoid re-construction of the groups. One possibility is to use the original ONEX ED based methodology to place the subsequences of the new time series into groups. The groups into which the new subsequences are placed are updated to include the new sequence, and their representatives are recomputed. Another option is to use each subsequence of the inserted time series as a query sample and to find its best match representative using the ONEX approach (See Section 5.2). Then we insert the subsequence into this group and recompute the representative.

Table 2 describes the general ONEX query syntax clauses.

Table 2: **List of clauses**

Clause	Description
$MATCH$	Exact($\mathcal{L}$) refers to a specific length $\mathcal{L}$. Any refers to any length.
X_p	Subsequence of the time series p
seq	NULL means no sample sequence is given. X_p, q – samples sequences provided by user
Sim	Similarity Distance
ST	Similarity Threshold
$\mathcal{L}$	Length

Table 3 display the statistics of the datasets used in the experiments (No. is the number of time series and L is the length).

[1] *http : //cseweb.ucsd.edu/ dasgupta/291geom/streaming.pdf*

Table 3: **Datasets Statistics**

	Italy Power	*ECG*	*Face*	*Wafer*	*Symbols*	*Two Patterns*	*Star Light Curves*
No.	67	200	560	1000	995	4000	9236
L	24	96	131	152	398	129	1024

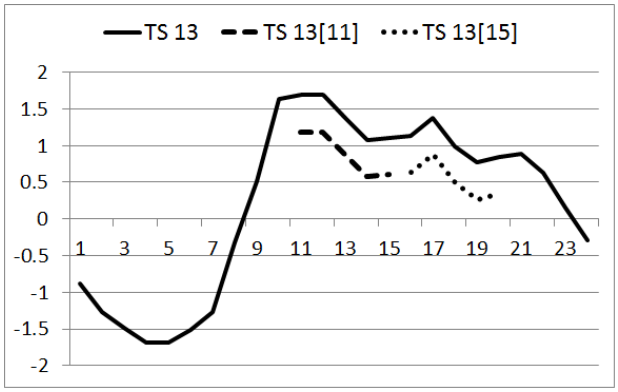

Figure 3: **Seasonal similarity for one time series**

For a better intuition of seasonal similarity queries as described in Section 6.2.2 of paper, we also provide a visual result using one sample time series in Fig. 3. This shows that for the 13th time series sample in the ItalyPower dataset, there are two similar subsequences of length 5, one starting at position 11 and the other one starting at position 15. In this figure we intentionally "shifted" the sample time series for better display, so they don't overlap with the other two subsequences.

2. REFERENCES

[1] L. Chen and R. Ng. On the marriage of lp-norms and edit distance. In *VLDB*. VLDB Endowment, 2004.

[2] P. Marteau. Time warp edit distance with stiffness adjustment for time series matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(2):306–318, 2009.

[3] T. Rakthanmanon, B. Campana, A. Mueen, G. Batista, B. Westover, Q. Zhu, J. Zakaria, and E. Keogh. Searching and mining trillions of time series subsequences under dynamic time warping. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 262–270. ACM, 2012.