APPENDIX A
PROPERTIES OF THE METRIC

TABLE IV
OVERVIEW OF THE SYMBOLS USED TO DESCRIBE THE DIFFERENT PARTS OF OUR METRIC

S	set of events
e	event — combination of an object and an identifier
o	object (e.g. a course)
i	identifier which can be ordered (e.g. semester, can be ordered in time)
S^O	set of objects from set of events
dist(S_A, S_B)	metric distance function
δ	positional relation (before, concurrent, after)
M_S	matrix of positional relations in set S
w	contribution factor
pos_dist(S^*_{AB}, S^*_{BA})	distance between the sets of identical events of the input

Given a set of objects O and an ordered set of identifiers I, we define an event e as $e \in O \times I$ and a set of events S as

$$S = \{e : e \in O \times I\}, \tag{3}$$

as well as its projection onto just the object space S^O as

$$S^O = \{o : \exists e = (o, i) \in S\}, \tag{4}$$

A set of events S is the basic representation of data used in our metric, its projection onto the object space is used to determine all objects that appear in both compared sets.

Given two sets S_A and S_B, the contribution factor w is the number of objects that occur in both S_A and S_B divided by the total number of objects in both sets.

$$w = \frac{|S^O_A| + |S^O_B| - (|S^O_A \setminus S^O_B| + |S^O_B \setminus S^O_A|)}{|S^O_A| + |S^O_B|} \tag{5}$$

Each object contributes equally, therefore the overlap of the two sets $|S_A \cap S_B|$ is counted twice, because each object appears twice in the overlap.

$$w = \frac{2(|S^O_A \cap S^O_B|)}{|S^O_A| + |S^O_B|} \tag{6}$$

The distance function of our metric consists of two parts, $w \times$ pos_dist(S^*_A, S^*_B) is the positional distance, which is used for the part of the sets, where the objects are identical. $(1 - w) \times 1$ describes the contribution of objects that only occur in one of the two sets, since the maximum distance possible in our function is one, these objects each contribute one to the distance.

$$\text{dist}(S_A, S_B) = w \times \text{pos\_dist}(S^*_{AB}, S^*_{BA}) + (1 - w) \times 1 \tag{7}$$

S^*_{AB} and S^*_{BA} are the subsets of S_A and S_B, where the objects are identical.

$$S^*_{AB} = \{s = (o, i) \in S_A \text{ and } \exists i_B, \text{ s. t. } (o, i_B) \in S_B\} \tag{8}$$

We define the matrix M_S as the positional relations δ between all events in S. This matrix is computed for each set of events individually.

$$M_\text{S} = (\, \delta(e_i, e_j)\,) \quad : \quad \forall e_i, e_j \in S \tag{9}$$

The positional relation $\delta(e_1, e_2)$ indicates the relation of event e_1 to event e_2 using the order of the identifiers I. The function returns -1 if e_1 happens before e_2, 0 if both are concurrent and 1 if e_1 happens after e_2. It is defined as

$$\delta(e_1, e_2) := \begin{cases} -1 & i_1 < i_2 \\ 0 & i_1 = i_2 \\ 1 & i_1 > i_2 \end{cases} \tag{10}$$

where i_1 and i_2 are the positions given by the identifiers of the events e_1 and e_2 respectively.

Since the relation between two identical events is always zero, the diagonal entries of the matrix M_S are zero. Furthermore M_S is skew symmetric ($m_{i,j} = -m_{j,i}$), because if e_1 happens before e_2, symmetry implies that e_2 happens after e_1 and vice versa.

The positional distance function sums up the differences in the matrices of positional relation of the subsets with identical objects of both input sets S_A and S_B.

Let M_A and M_B be the matrices of positional relations of the sets of events S^*_{AB} and S^*_{BA}, and $n = |S^*_{AB}| = |S^*_{BA}|$ be the number of events with common objects in one of the two sets, then the metric distance between S^*_{AB} and S^*_{BA} is defined by

$$\text{pos\_dist}(S^*_{AB}, S^*_{BA}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{sgn}(\, |a_{i,j} - b_{i,j}|\,) \tag{11}$$

where $a_{i,j} \in M_A$ and $b_{i,j} \in M_B$.

The sum of differences is averaged over all non-diagonal entries — of which there are $n(n-1)$ many — to ensure the result is between one and zero.

The signum function ensures that all differences contribute equally to the sum of differences. Further, this distance metric only makes sense if the position i of event e_i containing object o is identical in both matrices A and B.

A. Metric Properties

This subsection proves that our distance measure satisfies all metric properties.

Lemma 1. *Our distance measure $dist(S_A, S_B)$ is non-negative.*

Proof: $|a - b| \geq 0$ for any values of a and b, the number of rows/columns n in the matrix M_S is also always positive, and therefore the positional distance function is always greater or equal to zero.

$$\frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{n} \text{sgn}(\, |a_{i,j} - b_{i,j}|\,) \geq 0. \tag{12}$$

$$w \times \text{pos\_dist}(S^*_{AB}, S^*_{BA}) + (1 - w) \times 1 \geq 0 \qquad (13)$$

The contribution factor w (see eq. 6) describes a percentage of the input, thus it can only be between 0 and 1, therefore our metric is always greater or equal than zero. ■

Lemma 2. *If two sets of events are indiscernible, the result of our distance measure is zero.*

Proof: Two sets of events X and Y are indiscernible, if they contain the same objects, and those are in the same order. If this is the case their matrices of positional relations are identical ($M_X = M_Y$) and the sum of differences between two identical matrices is zero.

A distance of zero is only the case if X and Y are identical, if they differ in only one event the sum of differences is not zero anymore, which thus concludes that if $\text{dist}(X, Y) = 0$, then $X = Y$. ■

Lemma 3. *Our distance measure is symmetric.*

Proof: The positional distance function (equation 11) is the sum of differences between the positional relations of the events. One is added to the sum if the positional relations are different, zero otherwise. Changing the order is equivalent to changing $\text{sgn}(\,|x_{i,j} - y_{i,j}|\,)$ to $\text{sgn}(\,|y_{i,j} - x_{i,j}|\,)$. Since the sum is formed on the absolute value of differences, the sign is cancelled and the signum function guarantees that no other values than one and zero can be added. Therefore one is added to the sum whenever $x_{i,j}$ and $y_{i,j}$ are different independent of their order, which concludes $\text{dist}(X, Y) = \text{dist}(Y, X)$. ■

Lemma 4. *Our distance measure fulfills the triangle inequality, which means*
$dist(X, Z) \leq dist(X, Y) + dist(Y, Z)$ for all X, Y and Z.

Proof: We are proving this inequality component-wise. The distance $\text{dist}(X, Z)$ is calculated by summing up the differences between all positional relations (entries) in M_X and M_Z. $\text{sgn}(\,|x_{i,j} - z_{i,j}|\,)$ can either take the value zero if $x_{i,j}$ and $z_{i,j}$ have the same positional relation, or one if their positional relation is different.

- **Case 1**: $\text{sgn}(\,|x_{i,j} - z_{i,j}|\,) = 0$
  Zero is the lowest value that can appear, therefore $0 \leq \text{sgn}(\,|x_{i,j} - y_{i,j}|\,) + \text{sgn}(\,|y_{i,j} - z_{i,j}|\,)$.
- **Case 2**: $\text{sgn}(\,|x_{i,j} - z_{i,j}|\,) = 1$
  Our proof is by contradiction. Let us assume that $\text{sgn}(\,|x_{i,j} - y_{i,j}|\,) + \text{sgn}(\,|y_{i,j} - z_{i,j}|\,) = 0$. The only case where this happens is, if $x_{i,j}$ and $y_{i,j}$, as well as $y_{i,j}$ and $z_{i,j}$, are concurrent events. Hence, by transitivity, $x_{i,j}$ and $z_{i,j}$ are concurrent as well. This means $\text{sgn}(\,|x_{i,j} - z_{i,j}|\,) = 0$ which is a contradiction.

By proving that the difference between individual entries of the positional relations fulfill the triangle inequality and since our distance measure is the sum of all individual differences, all of which are non-negative, it follows that it fulfills the triangle inequality. ■

Appendix B
Instructions for Experiment Replication

All data, code as well as the instructions and answers to our user study can be found here: https://github.com/VDA-univie/set-of-events-distance-metric. The instructions are written for the bash shell, since it is available on all operating systems. To run the experiments described in this work, the following steps need to be done first to set up the environment.

1) download the 'code' directory to your machine
2) navigate to the 'code' directory in a terminal
3) create a virtual environment using python3
   `python3 -m venv env`
4) activate the virtual environment
   `source env/bin/activate`
5) install all required packages
   `pip install -r pip-requirements`

Once completed the experiments can be run with the commands shown in table V.

TABLE V
Commands used to replicate the different experiments from section IV.

Clustering generated paths
`python path_clustering_generic.py`
Clustering of real paths
`python path_clustering.py`
Predicting path lengths
`python path_predict.py`

All parameters are set such that the results shown in this work are replicated. The parameters can be changed by editing the source code, where each parameter is named accordingly.

Appendix C
Detailed Clustering Results

Tables VI, VII and VIII show more detailed results from clustering the generated study paths. The probabilities used for generating the paths have been varied, all other parameters stayed the same. DBSCAN was run with epsilon of 0.8 and the minimum number of points was 5.

TABLE VI
DETAILED RESULTS FROM CLUSTERING 100 GENERATED STUDY PATHS, USING A CHANGE PROBABILITY OF 70-20-10

metric	clustering	homogeneity score	completeness score	v_measure score	adjusted rand score	adjusted mutual info score	normalized mutual info score	silhouette score	clusters	noise
our metric	kMeans	1	1	1	1	1	1	0.740380813	4	0
Energy	kMeans	0.75	0.892601039	0.815110624	0.652883569	0.740611948	0.818199718	0.694392994	4	0
EM	kMeans	0.75	0.892601039	0.815110624	0.652883569	0.740611948	0.818199718	0.66375268	4	0
Dlev	kMeans	1	1	1	1	1	1	0.470027398	4	0
our metric	DBSCAN	1	1	1	1	1	1	0.740380813	4	0
Energy	DBSCAN	0.792802453	0.679549349	0.731820194	0.601756278	0.658452905	0.733994817	0.633068367	6	1
EM	DBSCAN	0.77935697	0.575018842	0.661773399	0.520272708	0.535539742	0.669436287	0.457225376	9	4
DLev	DBSCAN	0.498300818	0.431093361	0.462267097	0.18400579	0.360822106	0.463480501	-0.142133249	10	53

TABLE VII
DETAILED RESULTS FROM CLUSTERING 100 GENERATED STUDY PATHS, USING A CHANGE PROBABILITY OF 50-30-20

metric	clustering	homogeneity score	completeness score	v_measure score	adjusted rand score	adjusted mutual info score	normalized mutual info score	silhouette score	clusters	noise
our metric	kMeans	1	1	1	1	1	1	0.66528544	4	0
Energy	kMeans	0.75	0.869013399	0.805132372	0.632383295	0.740779426	0.807316573	0.619289976	4	0
EM	kMeans	0.75	0.869013399	0.805132372	0.632383295	0.740779426	0.807316573	0.572138792	4	0
Dlev	kMeans	1	1	1	1	1	1	0.213137912	4	0
our metric	DBSCAN	1	1	1	1	1	1	0.66528544	4	0
Energy	DBSCAN	0.806200476	0.649006172	0.719113104	0.609812209	0.623622465	0.723345758	0.593211605	7	1
EM	DBSCAN	0.731543477	0.441038545	0.550304993	0.349707838	0.374413587	0.568013091	0.197628526	14	14
DLev	DBSCAN	0.020447872	0.289137359	0.038194612	0.000830441	0.007903456	0.076891117	0.092768369	1	98

TABLE VIII
DETAILED RESULTS FROM CLUSTERING 100 GENERATED STUDY PATHS, USING A CHANGE PROBABILITY OF 30-40-30

metric	clustering	homogeneity score	completeness score	v_measure score	adjusted rand score	adjusted mutual info score	normalized mutual info score	silhouette score	clusters	noise
our metric	kMeans	1	1	1	1	1	1	0.613383675	4	0
Energy	kMeans	0.75	0.858419754	0.800555718	0.621994795	0.740884491	0.802380717	0.563085631	4	0
EM	kMeans	0.809563453	0.825070928	0.817243632	0.753666841	0.802826335	0.817280411	0.459402372	4	0
Dlev	kMeans	1	1	1	1	1	1	0.109391893	4	0
our metric	DBSCAN	1	1	1	1	1	1	0.613383675	4	0
Energy	DBSCAN	0.764070217	0.64863478	0.701636248	0.563013699	0.629685354	0.703990424	0.490243353	5	3
EM	DBSCAN	0.714315308	0.406440431	0.518090805	0.247930083	0.338350653	0.538819656	0.148717497	14	19
DLev	DBSCAN	3.20E-16	1	6.41E-16	0	9.61E-16	4.44E-06	0	0	100