

# Metrics for criminal records

Walter A. Kusters, Jeroen F. J. Laros  
Leiden University



Universiteit Leiden

## Ten criminals, four crimes

	1	2	3	4	5	6	7	8	9	10
$\mathcal{A}$	0	2	10	0	0	0	0	2	0	2
$\mathcal{B}$	0	0	0	2	0	0	2	4	0	2
$\mathcal{C}$	0	0	0	0	1	0	2	0	3	2
$\mathcal{D}$	0	0	0	0	0	1	1	0	5	2

Here we see a dataset with crimes  $\mathcal{A}$ ,  $\mathcal{B}$ ,  $\mathcal{C}$  and  $\mathcal{D}$  of increasing severity, and criminals ranging from 1 to 10. For each criminal the number of crimes in each category is given. For instance, 1 is innocent, 2 is an incidental small criminal, 6 is a one-time offender of a serious crime, number 10 is an all-rounder and 9 is a severe criminal.

The generic **distance measure** is

$$d_f(X, Y) = \frac{\sum_{i=1}^n f(x_i, y_i)}{|S(X) \cup S(Y)|}$$

The parameter  $f$  can be tuned. It specifies the difference between the number of occurrences of a particular element in two multisets. Constructing such a function is natural and can easily be done by domain experts.

Q: What is the difference between 2 and 10 crimes of the same category?

A:  $f(2, 10)$ .

The well-known **Jaccard** distance measure is a function defined on sets. This measure is used in the market basket analysis.

$$d(X, Y) = \frac{|X \setminus Y| + |Y \setminus X|}{|X \cup Y|}$$

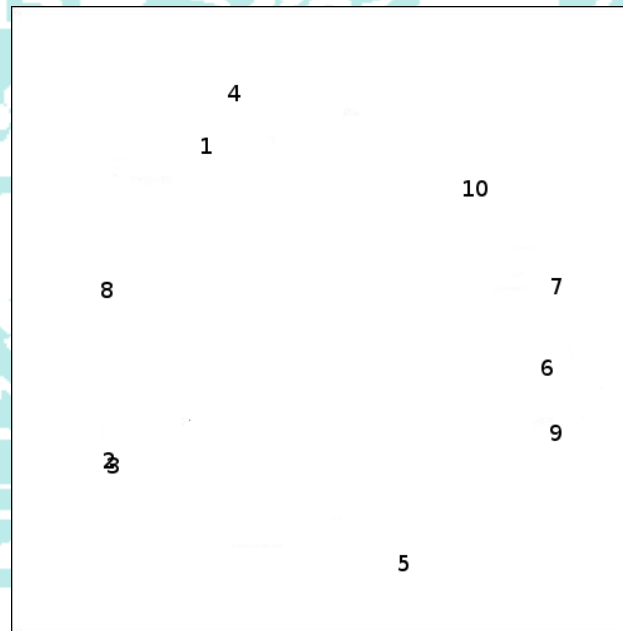
By choosing the right  $f$ , we get this formula.

The seal of Universiteit Leiden is a circular emblem. It features a central figure of a woman in a blue and white dress, holding a book and a quill. The figure is flanked by two shields: the left one shows a lion, and the right one shows a cross. Above the figure are the numbers '15' and '75'. The outer ring of the seal contains the Latin text 'ACADEMIA • LUGDUNO • BATAVA • LIBERTATIS • PRAESIDIUM'.

We use a **push and pull** algorithm to cluster the data once all pairwise distances are determined.

This algorithm iterates over all pairs of (randomly initialized) points and corrects the distances until no changes are visible.

## Visualization of the data with the Jaccard distance



Notice that criminal 2 and 3 are considered the same. They both are light criminals, but the number of crimes differ greatly.

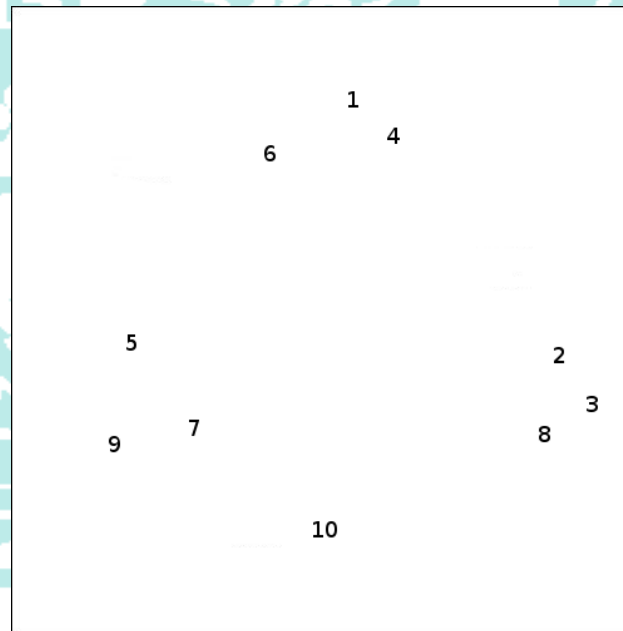
To distinguish between the number of crimes, we need a more complicated formula.

$$f_0(x, y) = \frac{|x - y|}{(x + 1)(y + 1)}$$

Notice that the difference between 0 and 1 is large, but the difference between 9 and 10 is not.

This comes in handy when comparing two criminals, the difference between an innocent man and someone who had committed a crime is larger than two criminals that differ in the number of crimes they have committed.

## Visualization of the data with the multiset distance



Here we see a clustering of the ten criminals, but now with our  $f_0$  function.

Notice that criminal number 2 and 3 still are close together, but do have a considerable distance.

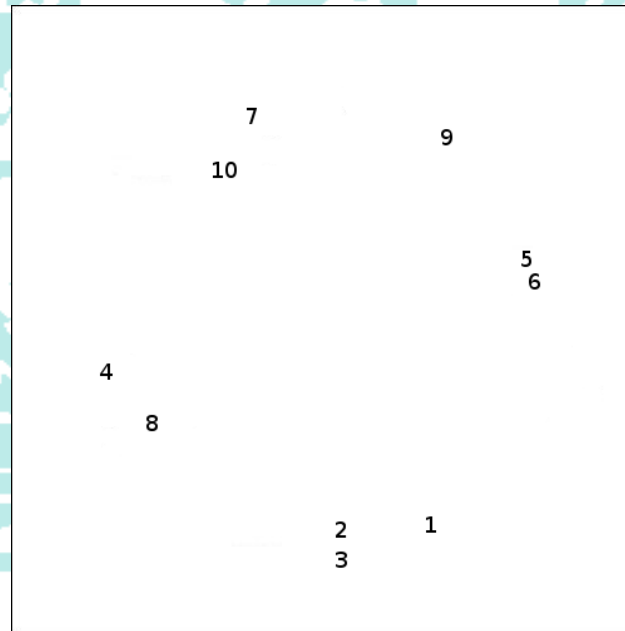


The seal of Universiteit Leiden is a circular emblem. It features a central figure of a woman in a blue and white gown, holding a book and a quill. She is flanked by two shields: the left one is red with a white lion, and the right one is blue with a white cross and a red lion. Above the shields are the numbers '15' and '75'. The entire seal is surrounded by a blue border with the Latin text 'ACADEMIA • LUGDUNO • BATAVA • LIBERTATIS • PRAESIDIUM'.

Now we add weights (1, 10, 100, 1000) to the categories of crimes.  
This indicates the severity of the crime.

Universiteit Leiden

## Visualization of the data with weighted multiset distance



Notice that criminal 2 and 3 now are close to number 1 (the innocent one). Also note that criminal 7 and 10 are now very close together, these are convicted of heavy crimes.

## Other applications

Other applications include, but are not limited to:

- Genetic research, we mention:
  - Comparison of DNA using small fragments of a fixed length.
  - Comparison on the occurrence of SNPs.

Both techniques can be used for phylogenetic tree construction and crime investigation.

- Comparison of documents: plagiarism, textmining, ...

## Conclusions and further research

We have devised an interesting class of distance measures for multisets.

For the clustering of criminal behaviour through time, we want to make a string of multisets. Then we want to use this distance measure to align these strings. Doing so will hopefully give more insight in criminal behaviour through time and might be used to make predictions for starting criminals.

The seal of Universiteit Leiden is a large, circular emblem in the background. It features a central figure, likely a personification of Wisdom or Justice, holding a book and a scale. The figure is flanked by two shields, one on each side. The shield on the left contains a lion, and the shield on the right contains a cross. The entire seal is surrounded by a Latin inscription: "ACADEMIA • LUGDUNO • BATAVA • LIBERTATIS • PRAESIDIUM • 1575".

This research is part of the DALE (Data Assistance for Law Enforcement) project as financed in the ToKeN program from the Netherlands Organization for Scientific Research (NWO) under grant number 634.000.430.

The logo for the Netherlands Organization for Scientific Research (NWO) consists of the letters "NWO" in a stylized font. The "N" and "O" are red, while the "W" is black. A red swoosh is positioned above the "N", and a black swoosh is positioned above the "O".

NWO

The logo for ToKeN2000 features the text "ToKeN2000" in a black, sans-serif font. A large, black, curved swoosh is positioned above the text, starting from the right and curving towards the left.

ToKeN2000

Universiteit Leiden