Linkage clustering file formats

This document describes the formats of two types of mandatory files for record linkage clustering techniques and corresponding ground truth data. A third optional file format is described for specifying constraints between record pairs (as implied by factors such as time).

It is assumed each record contains a unique record identifier, denoted as: **rec_id_x**

Similarity file format

Such a file contains the pair-wise similarities (links) calculated between records, where each pair is made of two record identifiers, a calculated overall similarity between them, as well as optional information for each record pair / link.

Optional information is shown in italics.

File format:

```
num_rec
num_links
rec_id_1, optional attributes about record with rec_id_1
rec_id_2, optional attributes about record with rec_id_2
…
rec_id_[num_rec], optional attributes about record with rec_id_[num_rec]
rec_id_a,rec_id_b,sim_val, optional attributes about link rec_a to rec_b
…
rec_id_x,rec_id_y,sim_val, optional attributes about link rec_x to rec_y
```

Cluster file format

Such a file contains information about clusters (groups of records). These can either be ground truth clusters or predicted clusters (records grouped by linkage or clustering algorithm). The same file format can be used for both. It is assumed that such a file contains each record (i.e. record identifier) only once (in one row / one cluster), which means the clusters are not overlapping; and the union of all clusters (all rows) corresponds to all record identifiers available in a data set.

Optional information is shown in italics. Note that if no optional information is provided then the last character in a line must be a trailing comma. This is because the number of records in a cluster can vary, and therefore the assumption is that all comma separated values are expected to be record identifiers, with the exception of the one after the last comma (so it can either empty or contain additional information about the cluster, such as the average similarity between its records). Note that the optional attribute about a cluster must not contain any commas.

File format:

```
num_cluster
rec_id_a,rec_id_b,rec_id_c, optional attribute about this cluster
…
rec_id_x,rec_id_y, optional attribute about this cluster
…
rec_id_z, optional attribute about this cluster
```

Note that the last line above is a cluster of size one (rec_id_z), also known as a singleton.

Constraint plausibility file format

Such a file contains information about which which record pairs are plausible, and which are not, with respect to constraints implied by factors such as time (known as temporal plausibility). The plausibility is specified in a matrix format, where a cell value of 1 indicates a pair that is temporally plausible, and an empty string (‘’) indicates that a record pair is not plausible.

File format:

```
           <rec_id_z, …, rec_id_c, rec_id_b, rec_id_a>
rec_id_a          1,  …,          ,         1
rec_id_b           ,  …,        1
rec_id_c          1,  …
…                 …
rec_id_z
```

Note that the columns are in reverse order of the row ID values in this file.