Preprocessing functions
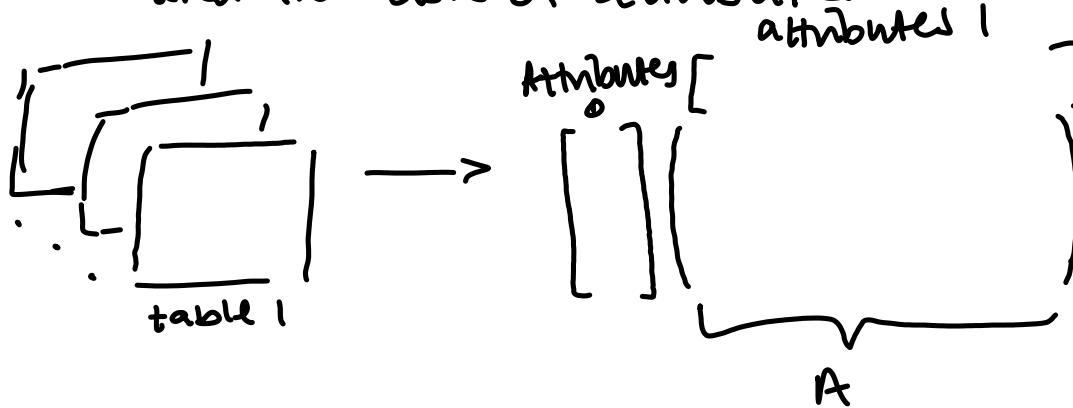
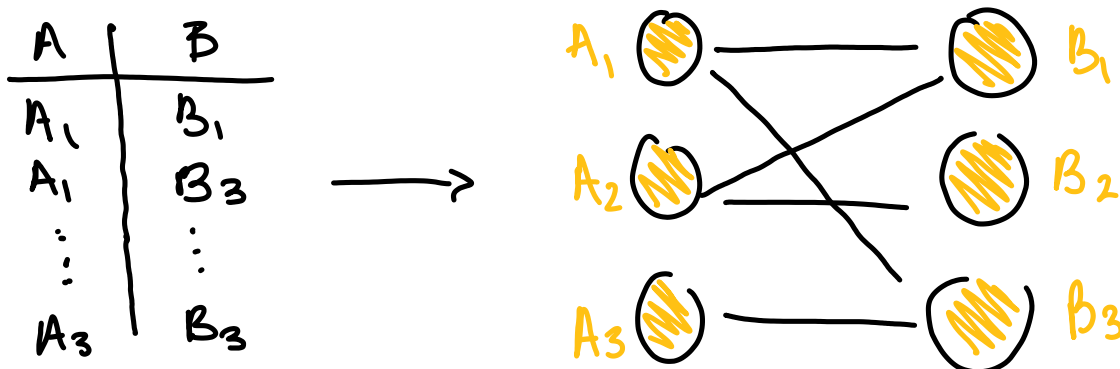15 November 2023   14:16

The idea of the preprocessing functions are t take the data and turn it into a matrix and two lists of attributes

---

|Inputs|

(1) <mark>tables</mark> : list of dataframes / tables with relational information
Each column is about an entity. The entries in the cells are items of that entity with an ID.
We specify which entries we want relationships between. eg

A	B
A_1	B_1
A_1	B_3
$\vdots$	$\vdots$
A_3	B_3

Note! if two columns are being connected their <mark>IDs must be unique</mark>

are if the intersection of two columns is not zero

Note: If the intersection of two columns is ___ ___
they will be considered part of the same entity

(2) Relationships: list pairs of columns to connect
                    This can take a couple of forms
                    forms:

(i) list of pairs: $[[\cdot,\cdot], \ldots [\cdot,\cdot]]$

  ↳ in this case: for each pair of entities, it will
    search all tables and link the columns
    with these names.

(ii) pairs in lists of lists: $[[[\cdot,\cdot],[\cdot,\cdot]], \ldots, [[\cdot,\cdot]]]$

  Here the list of pairs are linked to a particular
  table. → specify pairs of entities for each
  table. e.g.

  [table 1, table 2]   and $[[[\cdot,\cdot],[\cdot,\cdot]], [[\cdot,\cdot]]]$
      ①   ②   ①   ②

(3) dynamic col : if the data is dynamic then
                  it indicates which column is
                  the time column.
                  can be one str. if all tables have
                  the same name for time column
                  or list of strings if they're different

(4) join token : when creating the globally unique
                 names this character connects
                 the identity name to the ID

⌐Outputs⌐

(1) A : matrix of the graph.

(2) attributes : two lists of dictionaries. The first is for the nodes that don't change over time (rows of A). The second is the nodes that do change over time (columns of A)

Note: the information contained about the nodes are:
   name: partition_name + join token
                + ID

   partition: partition_name
   time: time code (nan if not dynamic).