APPENDIX A: GRAPH GENERATION

This appendix provides details of graph generation, step (a) in our process, as described in Section 3.1.

GUIDANE constructs a graph from tabular data using any common characteristic shared by records (such as an identification number like *familyID*, spatial facts such as the suburb a person lives in, and temporal facts such as the week of year number, 1 to 52). For any given dataset/problem, the user can identify one such data characteristic to form groups of records, where each record becomes a node in the graph we create, and edges connect nodes with the same data characteristic (such as the same *familyID*). We provide two illustrative examples to describe our approach below.

While existing approaches for anomaly detection in graphs either aim to find anomalous (individuals or groups of) nodes or edges, considering either structural or attribute information, the novelty of our approach is its ability to consider both structural and attribute information to identify anomalous nodes and edges using an unsupervised approach, given a particular context/problem. Hence, the characteristic to form the edges from tabular data is based on this problem the user wishes to solve.

We will describe our approach using the following two examples, where Example 1 considers a tabular dataset of COVID-19 patients, and Example 2 considers a tabular dataset of birth certificates.

Example 1

The dataset named *QLD*[1] is a synthetic dataset with details of individuals in Queensland, Australia, who have contracted COVID-19. Each record in this dataset represents one patient, covering the time period January (week 5) to August (week 33) 2020. While the dataset contains 10,000 records, we aggregated the records of each week and each postcode to form the nodes of a graph. Hence, a node represents the details of all patients who have contracted COVID-19 in a particular week and living in a particular postcode. Edges can connect nodes in different ways, as we now discuss. Using these nodes, we can address three different problems/contexts as follows.

(1) Problem to solve: Identify anomalies in the change of the number of patients in a particular postcode over time.

Edge criterion/characteristic: Form edges between consecutive weeks of the same postcode in chronological order. In this scenario, we consider the temporal aspect to form the edges.

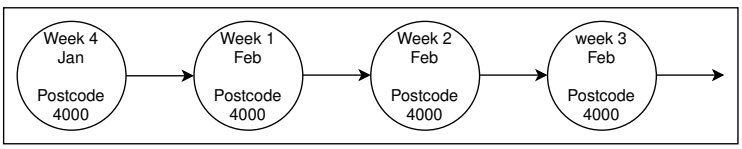

Fig. 1. A temporal graph formed between consecutive weeks of the same postcode in chronological order.

(2) Problem to solve: Identify anomalies in the spread of the COVID-19 between a particular postcode and its neighboring postcodes within a particular week of consideration.

Edge criterion/characteristic: Form edges between neighboring postcodes of the same week. In this scenario, we consider the spatial aspect to form the edges.

[1]https://github.com/anusii/RelM/blob/master/examples/20200811_QLD_dummy_dataset_individual_v2.xlsx

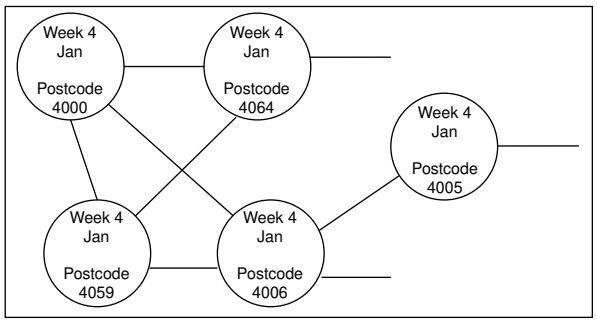

Fig. 2. A spatial graph formed between neighboring postcodes of the same week.

(3) Problem to solve: Identify anomalies in the spread of COVID-19 from one week in a particular postcode to its neighboring postcodes in the week after.

Edge criterion/characteristic: Form edges between one week in a particular postcode to its neighboring postcodes in the week after. In this scenario, we consider both temporal and spatial aspects to form the edges.

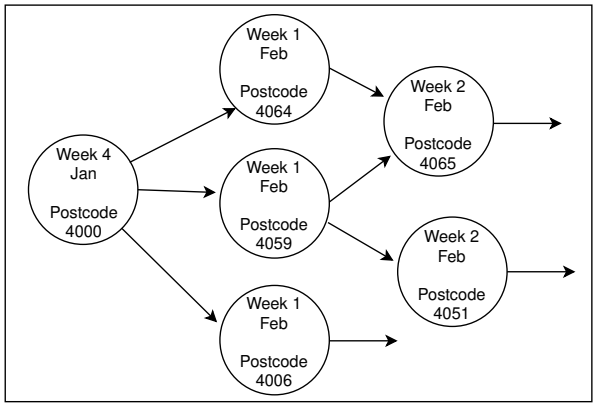

Fig. 3. A spatial and temporal graph formed from one week in a particular postcode to its neighboring postcodes in the week after.

As per Example 1, using one tabular dataset we can construct different graphs based on different edge criterion/characteristic to address different problems. The selection of the edge criteria/characteristic is problem dependent.

Example 2

Consider the following tabular dataset containing personal details of people's birth certificates.

Table 1. Tabular dataset of peoples' birth certificate information.

BirthID	SiblingID	Baby First Name	Baby Family Name	Date of Birth	Father's Name	Mother's Name
B100	F101	John	Smith	10/04/1847	Mike Smith	Nancy Smith
B101	F101	Lauren	Smith	12/12/1855	Mike Smith	Nancy Smith
B102	F101	Katie	Smith	20/04/1872	Mike Smith	Nancy Smith
B103	F102	Andrew	Brown	13/10/1860	John Brown	Nancy Smith
B104	F103	Kirby	Mcleod	13/02/1887	Josh Mcleod	Maleen Mcleod
B105	F103	Listy	Mcleod	18/07/1893	Josh Mcleod	Maleen Mcleod

In the domain of family reconstruction, where we assume the "SiblingID" is obtained via census data collection or a record linkage program based on domain knowledge, we can use graph-based anomaly detection to identify abnormal edges among siblings. This highlights potentially incorrect attribution of the "SiblingID", or families with unexpected birth patterns of their children. In this example, a node is the person (baby) represented by the birth certificate in the dataset of Table 1. Using these nodes, we can construct the following graphs as follows.

(1) Problem to solve: Identify abnormal edges among siblings.

Edge criterion/characteristic: Form edges using the attribute *SiblingID*. Draw an edge if two babies have the same *SiblingID*.

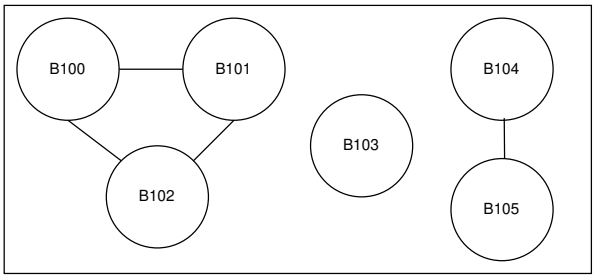

Fig. 4. The graph with *SiblingID* forming the edges.

(2) Problem to solve: Identify abnormal edges among siblings.

Edge criterion/characteristic: Form edges using the attribute *SiblingID*. Draw a directed edge if two babies have the same *SiblingID*, where the arrowhead demonstrates the chronological order of the date of births of any two siblings.

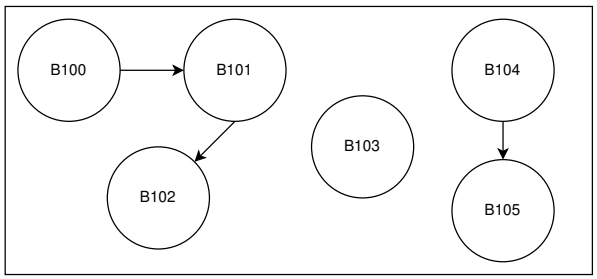

Fig. 5. A directed graph that is sequential in the order of the births with *SiblingID* forming the edges.

As per Example 2, using one tabular dataset we can construct different graphs based on the same edge criterion/characteristic to address one problem. Alternative to the *SiblingID*, we can use the attribute *Mother's Name* to establish links among nodes. This will create edges among the nodes *B100*, *B101*, *B102*, and *B103* as they share the same Mother's name. However, this selection of the edge criterion is meaningless as it is highly likely to have many people identified by the same name.

While the selection of the edge criterion is based on the domain expertise, as future directions of our research, with the use of data profiling we aim to create an approach to automatically identify the edge criterion, and perform grouping on the tabular data based on thus identified edge criterion.