A graphical tool for the interpretation of medical data

Marie Ely Piceno*
mpiceno@cs.upc.edu
Universitat Politècnica de Catalunya

Laura Rodríguez-Navas
laura.rodriguez@cttc.cat
Centre Tecnològic de Telecomunicacions de Catalunya

ABSTRACT

The analysis of highly frequent co-occurrences of data is often apply and commonly yield interesting but large amount of results. Graphical tools are useful in the interpretation and better understanding of data. This research is based on specific variants and implementations of the so-called Gaifman graphs.

KEYWORDS

Exploratory data analysis, Gaifman graph, modular decomposition.

1 INTRODUCTION

In [2] we show that Gaifman graphs, a mathematical structure introduced in Logic, can be used in an advantageous way for exploratory data analysis. The method we used there is a modular decomposition method, in terms of the so-called prime clans [1].

The construction of the Gaifman graph is based on co-occurrence, or lack of it, of items in the data. From it we can obtain a complete graph with an equivalence relation among its edges (so-called 2-structure[1]). It is known [1] that prime clans allow us to decompose a Gaifman graph into a tree-like form.

There are some variants of Gaifman graph. In [3] we studied the exponential variant. Here we extend our exploration to the shortest path version of the decomposition of that Gaifman graph.

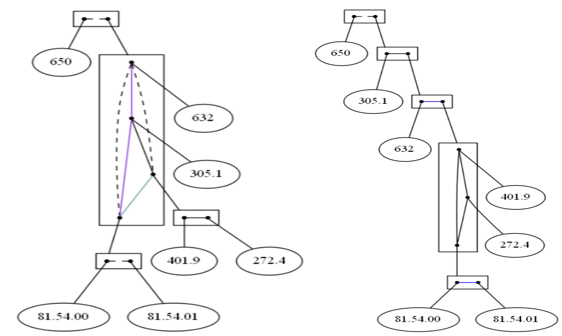
2 ANALYSIS OF DIAGNOSTIC DATA

The dataset we focus on was provided to us by *Hospital de la Santa Creu i Sant Pau*. The dataset contains information of all hospitalizations for the years 2015-2016. We have already studied this dataset in [3]. We analyze a part of the dataset, specifically we use the information about the diagnostics and procedures. The dataset consists of 7741 values. Therefore, it has been necessary to consider a minimum frequency threshold, since we cannot display all values graphically. In the following, by way of example, we consider 100 as a threshold; this leads to only 7 elements. This means that diagnostics and procedures that appear less often than 100 times were not taken into account for the visualization.

2.1 Exponential Gaifman graph

We examined the exponential Gaifman graph in [3]. We observed the decomposition in the figure (left). At the top of the figure we notice that the item 650 *"Normal delivery"* is not related with any other procedure or diagnostic item because it is connected by a dotted line with the clan containing the remaining items, represented by the larger box.

We can find diagnostics 272.4 *"Other and unspecified hyperlipidemia"* and 401.9 *"Unspecified essential hypertension"* are linked by a high-frequency joint occurrence, representing over 11000 cases. However, in another clan we found procedures with similar codes, 81.54... *"Total knee replacement"* left and right, are connected by a



dotted line that indicates they are incompatible since is almost never the case both knees are replaced at once. Inside the larger box, we see no co-occurrences from 632 *"Missed abortion"* to knee replacement clan and to hypertension diagnostic clan. While, diagnostic 305.1 *"Tobacco use disorder"* appears to have a few co-occurrences with the missed abortion node and knee replacement clan, whereas it is significantly more related to the hypertension diagnostic clan, about 2000 occurrences.

2.2 Shortest path

Here we propose instead to explore a shortest path version of the decomposition of the Gaifman graph. In comparison, this is an improvement of the previous analysis, as it allows for a better understanding of the behavior of the items within the larger box in the figure (right). In this analysis we can verify that the node 650 *"Normal delivery"* is not connected with any other node. The node 305.1 *"Tobacco use disorder"* is directly related, with different co-occurrences values, within all the remaining nodes. While 632 *"Missed abortion"* is not directly related with hypertension diagnostic clan items and knee replacement clan, there could be cases where the patient medical conditions consist on 632 *"Missed abortion"*, 305.1 *"Tobacco use disorder"* and some of the items into the large box clan. Again, with different multiplicities of co-occurrences, the members of the larger box clan are directly correlated. This shows how the hypertension diagnostic clan items are connected to each other and with both cases of knee replacement.

REFERENCES

[1] T. Harju A. Ehrenfeucht and G. Rozenberg. 1999. *The Theory of 2-Structures - A Framework for Decomposition and Transformation of Graphs.* World Scientific. http://www.worldscibooks.com/mathematics/4197.html

[2] M.E. Piceno J.L. Balcázar and L. Rodríguez-Navas. 2018. Decomposition of Quantitative Gaifman Graphs as a Data Analysis Tool. In *Advances in Intelligent Data Analysis XVII - 17th International Symposium, IDA 2018, 's-Hertogenbosch, The Netherlands, October 24-26, 2018, Proceedings (Lecture Notes in Computer Science)*, Wouter Duivesteijn, Arno Siebes, and Antti Ukkonen (Eds.), Vol. 11191. Springer, 238–250. https://doi.org/10.1007/978-3-030-01768-2_20

[3] M.E. Piceno J.L. Balcázar and L. Rodríguez-Navas. 2019. Hierarchical Visualization of Co-Occurrence Patterns on Diagnostic Data. (2019). Accepted for presentation at 32th IEEE CBMS International Symposium on Computer-Based Medical Systems, Córdoba, Spain, June 5-7, 2019.