1. **Qué ofrezco y en qué contexto**

I propose to employ decomposition techniques on Gaifman graphs as a graphical tool for exploratory data analysis on medical data. The Gaifman graphs record the co-occurrences of data items in datasets and, then, graph decompositions provide valuable information that is not directly observable on the data, since the graph decompositions displays a hierarchical visualization of the co-occurrences.

Highly frequent co-occurrences of data items have been a target for several types of data mining frameworks for decades, in all types of data, including medical data. Most commonly studied notions for this sort of analysis are frequent sets and variants thereof, such as frequent closed sets or association rules. In most cases, frequent set mining actually returns a large textual list; then, a visualization of this result can allow us to achieve a better understanding of the data, as well to discover implicit and potentially useful information. But these notions are not really reaching out to end users, mainly due to the difficulty of finding explanatory descriptions. In fact, the results of these notions of data analysis are found in spaces of huge dimensionality, and their reductions to 2D or 3D plots almost never offer enough interpretability.

Gaifman graphs are a notion originated in mathematical studies of logical structures that, actually, can support a relational database model. Gaifman graphs can be advantageously employed for exploratory data analysis via their decomposition in terms of so-called clans, either in their original form or in one of its variants.

The decompositions of Gaifman graphs have the potential to reveal “co-occurrence” patterns or, alternatively, “incompatibility” patterns, since they relate together data items that, pairwise, appear together somewhere in the whole dataset. Generalizations of the notion allow us to adjust the co-occurrence thresholds so as to account for particularly frequent joint occurrences, or for different intervals of co-occurrence counts.

In the presentation, I will focus on specific variant of Gaifman graph, the exponential Gaifman Graph, showing how this combinatorial technique is applicable to medical data corresponding to joint diagnostics of patients. I believe that the resultant visualization, like the others that we created, can act in a useful way complementing existing statistical approaches; for example, by pointing out specific pairs of elements, possibly conditioned to other elements, whose correlation studies could be candidates for priority analysis.

2. **Porqué es importante para mi interlocutor/para otros**

The construction of data sets like the one we work on is not so simple. Initially, many diagnoses are expressed as natural language expressions, and ICD coding of information is a separate and subsequent process with often specialized people or even companies. We think that a graphical tool trained in frequent concurrent diagnoses can help accelerate this type of process by offering common options for automatic completion and/or by checking for double verification the rare ones, which could be either correct or the result of coding errors (like prostate surgery along with normal delivery as an extreme example).

3. **Como lo he hecho/hago/haré**

Our research is based on the Gaifman graphs, a mathematical structure introduced in Logic, that can be used in an advantageous way for exploratory data analysis. The method that we used is a modular decomposition method, in terms of the so-called strong clans.

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

The dataset that we used for the study was provided by the Hospital de la Santa Creu i Sant Pau. The dataset contains information of all hospitalizations for the years 2015-2016. We analyse a part of this 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 couldn't display all values graphically. For example, in the presentation that I will present, it was needed to consider 100 as a threshold. This means that diagnostics and procedures that appear less often than 100 times were not taken into account for the visualization.

With our research we can demonstrate a novel application of Gaifman graphs and their decomposition, providing a general visualization of the data behaviour that could be used as a tool to complement statistical approaches. And through this work we have illustrated the process and some of the possible results to apply the data analysis approach based on the tree decomposition of Gaifman graphs on the medical dataset.

Indeed, in order to obtain the previous results, we had to carefully observe the general behaviour of the data. In general, at the moment, the human brain us essential during the exploration of interesting parameter settings.

That said, we believe that our visualizations can act in a useful way complementing the statistical approaches, as an example among many others, it could point to the user specific pairs of elements possibly conditioned to other elements, whose correlation studies could be candidates for priority analysis. By itself, on the other hand, our approach did not provide any interesting results when we directly applied standard concepts of quantitative pattern mining as support or confidence thresholds.

In future contributions, we would like to offer self-descriptive, more informative, perhaps even animated, visualizations that trained medical staff can immediately capture.

4. **Qué pido y a quién**

I am a recent PhD who has just created a Spin-off with my thesis co-directors, and I am looking for a collaboration contract for conducting research studies with the Hospital de la

Vall d'Hebron in Barcelona. And I will ask to the director of this hospital, who will be represented by my teacher.