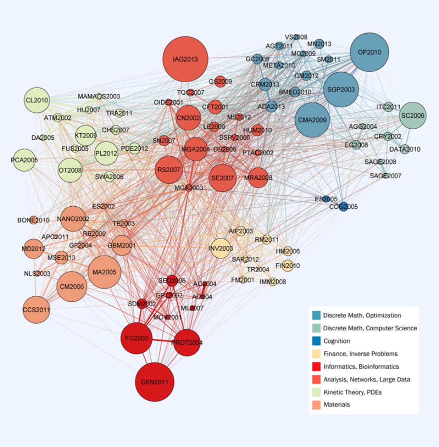
Both multi-label classification and social network community detection are phenomena present in our daily life. Thanks to classification we can find music by genre in Spotify or filter Youtube movies by topics. Multi-label classification is a task of teaching a classifier the relationship between objects and labels. Its purpose is to allow assigning multiple labels to an object, as opposed to the traditional task of classifying objects with just one category. Usually multi-label problems are reduced to single-label cases.

Communities in networks Detecting community structures is the problem of finding a group of people, things, books, movies, products, that are closer to each other because they are related by a larger number of relationships within the group than to outside objects. The simplest case of two products being close to each other can be based on the fact the two persons bought both of the products, read the same book, etc. A more detailed case can take into account that two persons rated a movie with the same grade. The relationship can also be weighted - the weight is a measure of importance of a relationship between two objects. An example of the weight could be the number of people who rated a move with the same score or number of people who both these two products. In the example Figure 1 community detection was used on a graph of relations between mathematics programmes/workshops. Each edge represents the fact that scientists went to both workshops/programmes. The weight - depicted by width - is the percentage of participants of the two, that took part in both of them[1].

Figure 1: Znalezione skupiska dzielą zbiór programów wsparcia nauki na dziedziny - umożliwiając dokładniejsza klasyfikację wewnątrz dziedzin



Skupiska a klasyfikacja Currently we have to perform multi-label classification on larger and larger data sets. The larger the data set, the higher the number of possible label combinations to assign, while it is not true that every combination has the same number of samples. In most cases it is the other way around - certain combinations are assigned to a majority of samples. One can easily imagine that when classifying youtube music covers, we will get more covers of pop stars than amateur performances of classical music.

The methods that are used for classification, are based on finding distictions between how objects differ to each other. By differntiating objects based on their features we can assign the music genre or instruments that are used in the performance. What is important is to make sure there is a balance between the accuracy with which the method can detect differences and the scale of differences. It is natural that in a situation when we classify into more homogenous groups we can perform better. Classifying music covers with the original performing artists is easier if we're doing it separately for pop covers and classical performances than when we mix all of them together and learn one classifier.

In order to be able to classify in homogenous subspaces we want to use community detection methods from social networks to divide the label space. Up until now methods that separate multi-label classification into subproblems were based on the similarity of objects input features and the division was performed into non-overlapping subgroups. We hypothesize that efficiency fo these methods can be improved by using exploring relationships between labels and using them to construct label space divisions. Thus in this project we plan to verify under what conditions and by what factors can we improve multi-label classification efficiency by dividing the labels space into overlapping subproblems - a in case of labels in multiple subproblems - how to select a relevant method of deciding whether the label should be assigned or not. We plan to evaluate both flat and hierarchical classification schemes as our preliminary results are promising[2].

[1] http://helper.ipam.ucla.edu/articles/ipamconnections.aspx

[2] https://arxiv.org/abs/1606.02346