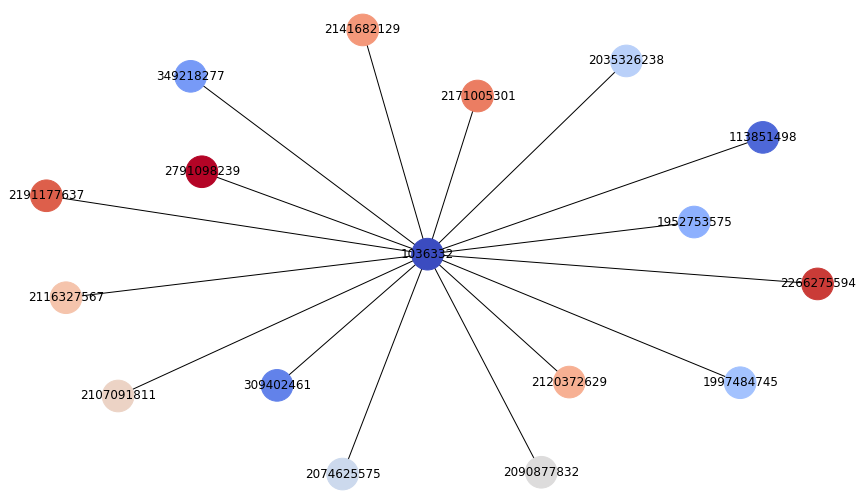
Data Analysis :

1. Graph Data:

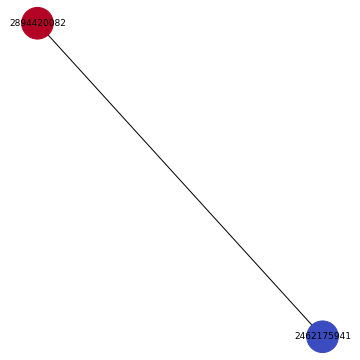
We are dealing with unweighted and undirected graph that models the co-authorship network, such that vertices correspond to authors where two authors (vertices) are linked by an edge if they have co-authored at least one paper, we have 231 239 nodes and 1 777 338 edges. Let’s see an example of what our graph looks like:



In the example below, we notice that the author with has co-authored with 16 authors in writing at least 16 papers.

We observe also that there are no isolated nodes in the graph, however when we examine the *author\_papers.txt* document, we realize that even if there are two connected nodes by an edge which means that they co-authored in at least one paper, we didn’t find any single common paper between them.

To illustrate what is stated before, we take two nodes with degree 1, which means we should find at least one co-authored paper between them in the *author\_papers.txt.*



Let’s have a look on the papers set for each author:



One can notice easily that there is no common paper between these 2 nodes.

1. Text Data

For the text data, we have at our disposal a set of papers with their abstracts, there is a total of papers that are present in *abstract.txt* file, however we have only papers that have abstracts, so papers have missing abstracts, we noticed also that there is more than one language in the abstracts set.

1. Train and test data

The training data contained labeled authors, each sample have the author ID with the corresponding *h\_index,* each author has his top-cited papers in *author\_papers.txt.* For the test data, we have unlabeled authors, the goal is to use the feature extracted from the graph and textual information to predict their corresponding h\_index.