* Calculate the similarity pairwise by 1-year/ 3 year  
  Expect the value of the similarity will jump when there is higher probability of frauds.
* Outlier Detection   
  Detect the change in the similarity values.
* Record the parameters
* Fill 0 as similarity value for the missing files
* Further step: using the WDM to calculate the similarity value