Summary of my papers

Among the various trends analyzed was one of co-authorship and collaboration. [8] suggests that a large number of collaborators have a first degree connection amongst them. The empirical proof was given by analyzing 4 databases from different fields and comparing the results. It showed that, irrespective of size of database, 75% of the collaborators have a first degree connection. [7] supports this inference by providing a different view for the same problem. It clusters the collaborators by harmonic distance to find the degree of supportiveness, which is the metric of how much a co-author of a particular paper supports the research.

Hive stuff

(Can Ansa please put in how the files were flattened for the HIVE inputs)

For certain metrics, we thought of using HIVE because of its easy support for data retrieval. As only certain number of fields were required from the entire data, we flattened the data into a tab-delimited format containing the key, year, conference and all the authors of the paper. The resulting record was as shown

conf/icod/AusielloBM80 1980 icod Giorgio Ausiello|Carlo Batini|Marina Moscarini

The various Tier files were created by taking the names of conferences in each tier and putting them in a file. Then each conference name in a particular tier file was compared with the conference name in the original file. If they were a match, it meant that the paper was presented in a conference belonging to that tier, and was entered as such. This was done using a JAVA program and new files were made containing records for each tier. The analytics were run against the entire original file and each tier files as well. These records were loaded into various different HIVE tables. The statistics obtained from then were

* Number of paper published each year
* Average number of authors per paper in each year