Realtime and Big Data Analytics Project Proposal

Part 1. General Information

Team Members: Aditya Garg, Aashutosh Trivedi, Ansa Mary Ephraim
Project Title: Analysis of Citation Network
Project Description:
We access citation database of conferences over a period of time. Some analytics we hope to derive are

1) Potential Hot Topic : Paper with the least citation distance from all the published papers.
2) Find citation distance and collaboration distance between authors
3) Help suggest should two authors work together in future.
4) Suggest/predict conferences to submit entries to for authors

Data Sources - Use the table below to list and describe potential data sources.

Part 2. General Data Source Information

Data Sources	**Data Source Description**	**Data Size**
http://snap.stanford.edu/data/cit-HepTh.html	Paper citation network of Arxiv High Energy Physics Theory category	36.3
http://www.cs.cornell.edu/projects/kddcup/datasets.html	29,000 hep-th papers with 1.7 gigs of data	36.37
http://people.cs.umass.edu/~mccallum/data.html	Simulated/Real/Aviation/Auto UseNet data	1.02
http://konect.uni-koblenz.de/networks/subelj_cora	Cora citation network datase	2.05
Self made editor database	While we have author, journal or metadata information, we need to create a database of the editors of the conference	-

Part 3. Detailed Data Source Information

Data Sources	**Data Characteristics**	**Data Frequency**
http://snap.stanford.edu/data/cit-HepTh.html	This is a static historical data of which papers have cited which others in high energy theoretical physics	

http://people.cs.umass.edu/~mccallum/data.html	Historical data of about 73000 articles from 4 discussion groups for simulated and real auto and aviation.	
http://konect.uni-koblenz.de/networks/subelj_cora	Historic dataset of over 23000 cora citation papers network. The network is connected.	
Self made editors database	Historic. we need to visit all journals and conferences and find out the editors or panel members of the same.	

Part 4. Technologies

MapReduce and HDFS are currently clear choices. If required, other can be incorporated.

Part 5. References

Analysis of SIGMOD's Co-Authorship Graph
http://sigmod.acm.org/publications/sigmod-record/0309/3.coauthorship.pdf

Some Analyses of Erdos˝ Collaboration Graph
http://vlado.fmf.uni-lj.si/pub/networks/doc/erdos/erdos.pdf

The Chasms of CSCW: A Citation Graph Analysis of the CSCW Conference
http://www.tau.ac.il/~gailgf/pubs/Chasms.pdf

Understanding Importance of Collaborations in Co-authorship Networks: A Supportiveness Analysis Approach
http://www.cs.sfu.ca/~jpei/publications/SupportivenessAnalysis-sdm'09.pdf

The structure of scientific collaboration networks
http://www.pnas.org/content/98/2/404.full

Analysis of Papers from Twenty-Five Years of SIGIR Conferences: What Have We Been Doing for the Last Quarter of a Century ?
http://sigir.org/files/forum/F2002/smeaton.pdf