Preparation of a Graph Database on Indian Legal Corpora

A Term Project Report

By

Saptarashmi Bandyopadhyay

Examination Roll:510514006

Fifth Semester 2016-2017

Dual Degree (B.Tech. - M.Tech.) in Computer Science and Engineering

under the Esteemed Guidance of

Dr. Saptarshi Ghosh

Department of Computer Science and Technology

Indian Institute of Engineering Science and Technology, Shibpur
Howrah-711103
West Bengal, India

Contents

1. Introduction
2. Scope of the Work
   - 2.1. Development of Graph Databases a Review - A Review
   - 2.2. SNAP (Stanford Network Analysis Project) a Review
   - 2.3. LIIofIndia a Review
   - 2.4. Legal Graph Database Development in World and India
   - 2.5. Research significance of the dataset
3. Features of the Dataset
   - 3.1. LIIOFINDIA Legal Database
   - 3.2. Dataset Statistics
4. Programs and tools used to build the dataset
5. Validation of the Dataset
6. Various Patterns indigenous to the Legal Graph Database
7. Future Work and Conclusion
8. References

INTRODUCTION

- Information Extraction from Unstructured Text databases
- Graph Modeling
- Graph Database of Legal Documents
- Citation Analysis for developing Citation Networks
- www.liiofindia.org is our source of legal documents in India

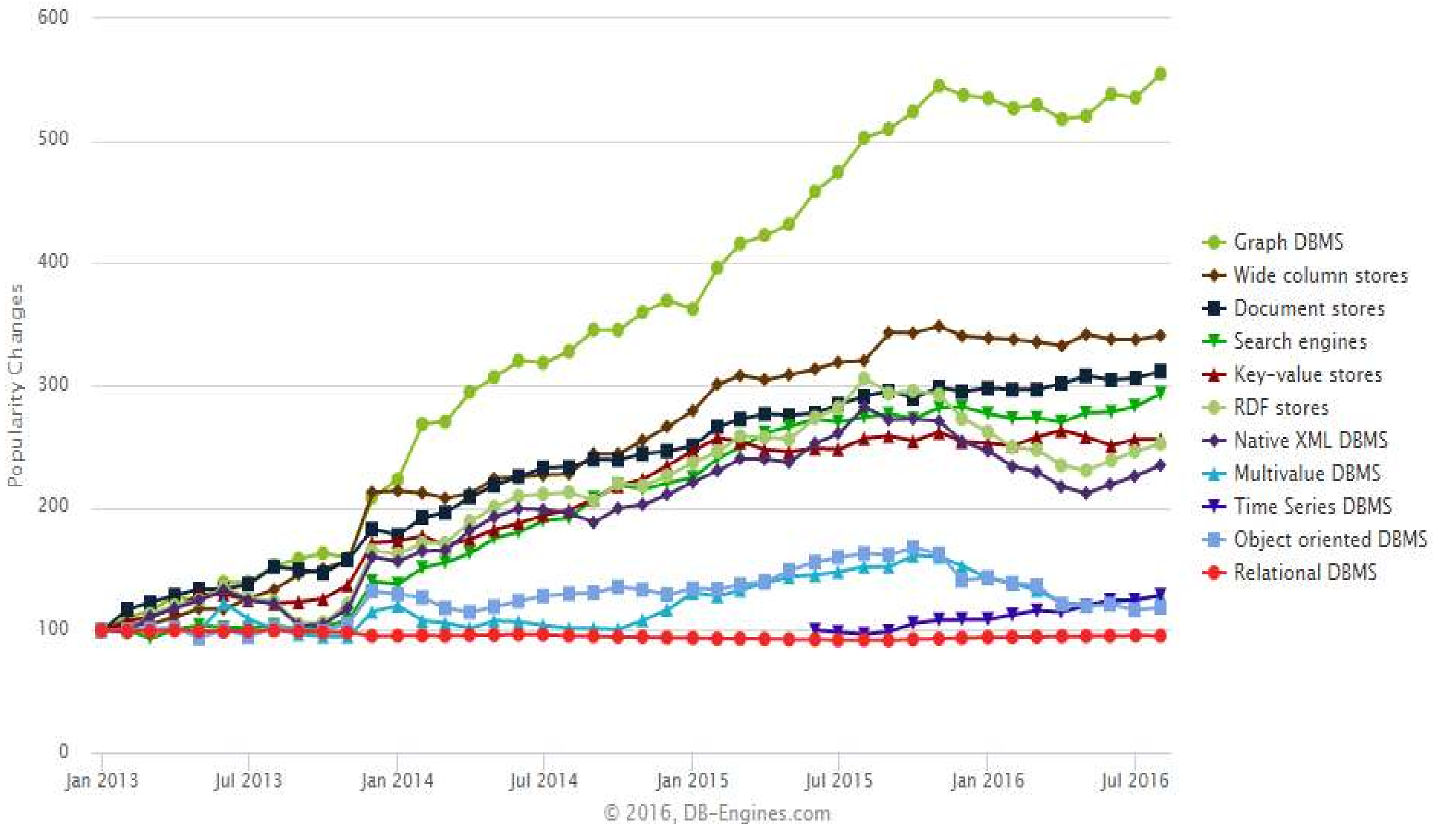
Scope of the Work – Graph Databases

- Graph Databases – Nodes represent the documents and Edges represent the relationship between the documents
- Graph Data stored in either table format or Key-value stores

The Law of Relational Database

If the only tool you have is a relational database,
everything looks like a table.

Popularity of graph database



Stanford Network Analysis Project (SNAP)

- general purpose network analysis and graph mining library
- Snap.py is a Python interface for SNAP in C++.
- A collection of more than 50 large network datasets from tens of thousands of nodes and edges to tens of millions of nodes and edges.
- Citation Networks – only research articles citation

Liiofindia

- Legal Information Institute of India
- NALSAR University of Law, Hyderabad; National Law School of India University, Bangalore; National Law University, Delhi, NUJS Kolkata, Rajiv Gandhi School of Intellectual Property Law, Indian Institute of Technology – Kharagpur.
- Open source consisting of 154 databases with more than 800,000 documents obtained from public sites of India

Present Legal Database

Scales of justice just got digital!

Legal Graph Database Development

- The Lillian Goldman Law Library of Yale Law School maintains a large open source collection of Foreign and International Law Resources.
- Manupatra, pioneer in online legal research in India since 2001, is India’s premier legal information resource.
  - The Service is available for a fee

Only Search is not “Research”!

Research Significance of the Dataset

- Most legal research involves relationship analysis
- Citation relationships among opinions create a cluster of opinions that are likely to address the same topic
- A major issue with graph analysis of legal content is how to filter the relationships to so that one can do a useful analysis.
- Relationship among content types

Features of the Dataset

The Indian Legal corpora in the LII legal database can be categorized into :

- The Constitution
- The Central/State/UT acts.
- The State/UT schemes
- The State/UT regulations
- **The Court cases (72.37% of the records)**
- The amendment bills
- The treaties

Graph Database

- **Node_list_for_metadata**
- 118 2011.INWBKOHCA.6402
- 119 1967.INSC.172
- 120 2010.INWBKOHCA.437
- 121 2011.INWBKOHCA.6409
- 122 2011.INWBKOHCA.6408
- 123 1968.Criminal.Law.Journal.India.231
- 124 2010.INWBKOHC.17218
- 125 AIR.1982.SC.626
- 126 2001.5.SCC.546
- 127 2001.5.SCC.540
- 128 2010.INWBKOHC.17219
- 129 32.ITR.737
- 130 2008.INWBKOHC.13899

Graph Database

- Connection_List
- 5 19076
- 6
- 7
- 8
- 9 710593 500003 34500 788645 400808 145647 117749 158712
- 10
- 11
- 12
- 13
- 14
- 15 189657
- 16
- 17 499008 493457 368231

Dataset Statistics

```
•   GraphInfo. build: 15:51:02, Nov 14 2016. Time: 17:08:48 [Jul 23 2016]
•   ================================================================================
•   Input graph (one edge per line, tab/space separated) (-i:)=connectionlist.txt
•   Directed graph (-d:)=Yes
•   Output file prefix (-o:)=graph
•   Title (description) (-t:)=
•   What statistics to plot string:
•           c: cumulative degree distribution
•           d: degree distribution
•           h: hop plot (diameter)
•           w: distribution of weakly connected components
•           s: distribution of strongly connected components
•           C: clustering coefficient
•           v: singular values
•           V: left and right singular vector
•            (-p:)=cdhwsCvV
•   ================================================================
```

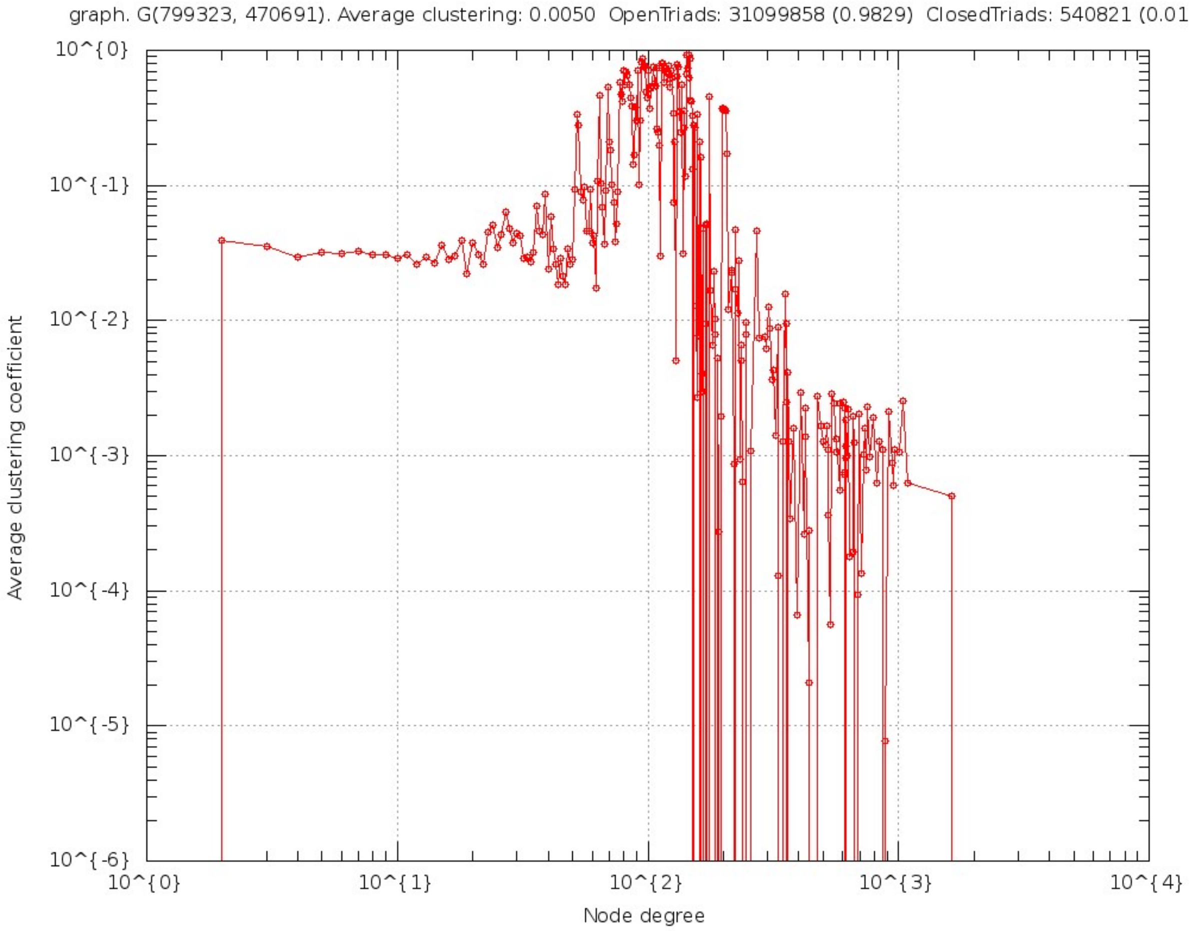
Dataset Statistics

- Loading...directed graph (TXT format)
- connectionlist.txt: Directed
-   Nodes: 799323
-   Edges: 470691
-   Zero Deg Nodes: 604335
-   Zero InDeg Nodes: 678467
-   Zero OutDeg Nodes: 715467
-   NonZero In-Out Deg Nodes: 9724
- Creating plots...
- size 19983076
- Calculating 400 eigen-values of 799323 x 799323 matrix
- 400
- Diameter (longest shortest path)   19
- 90-percentile effective diameter   7.6
-
- run time: 42.46s (17:09:31)

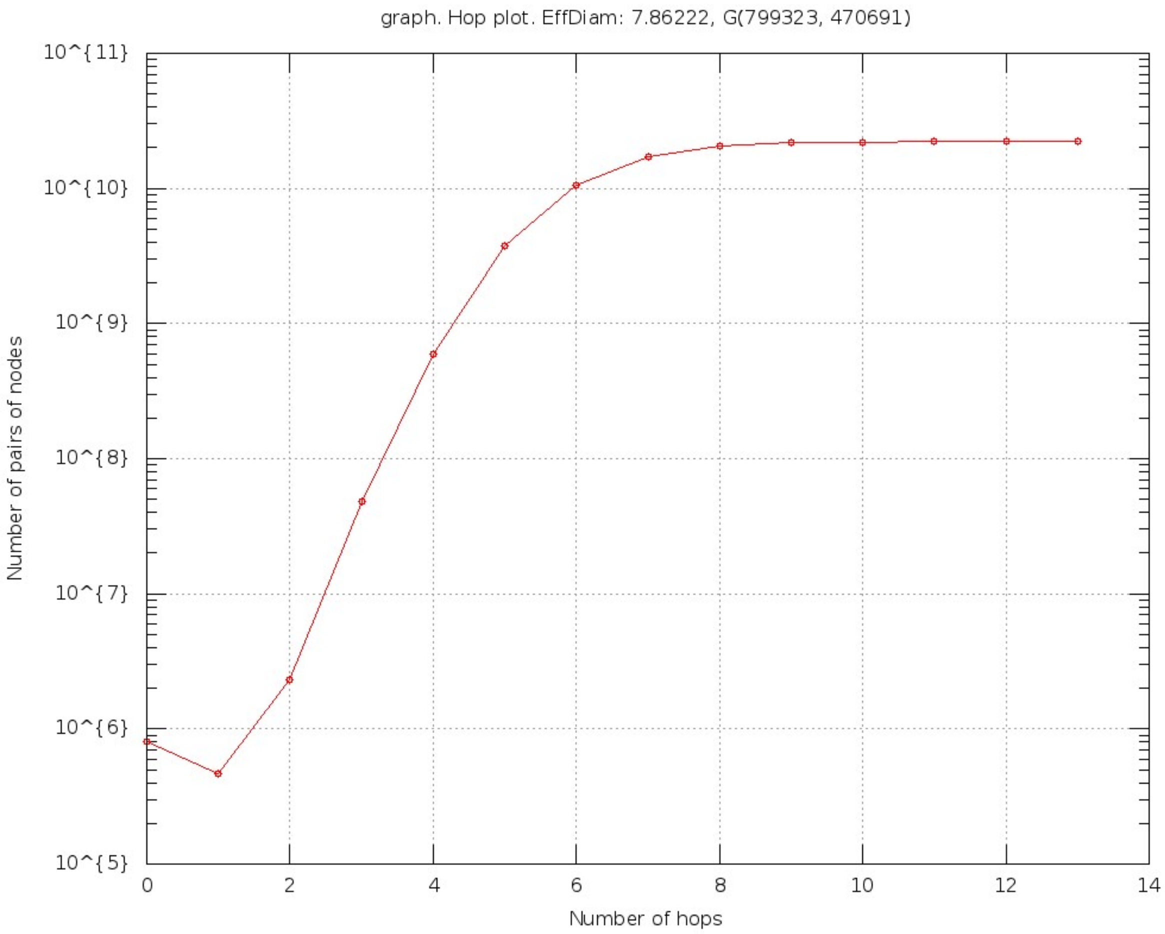
Dataset Statistics

- 33548 case of Supreme Court of India,
- 524800 cases from 25 High Courts across India,
- 2662 cases from 8 District courts,
- 17463 cases from 4 tribunals
- 103372 cases from Central Information Commission of India,
- 15 cases from an apellate board
- 1742 cases from 17 law journals, law research series, law reviews, treaty series, legislative summaries and law commission of India.

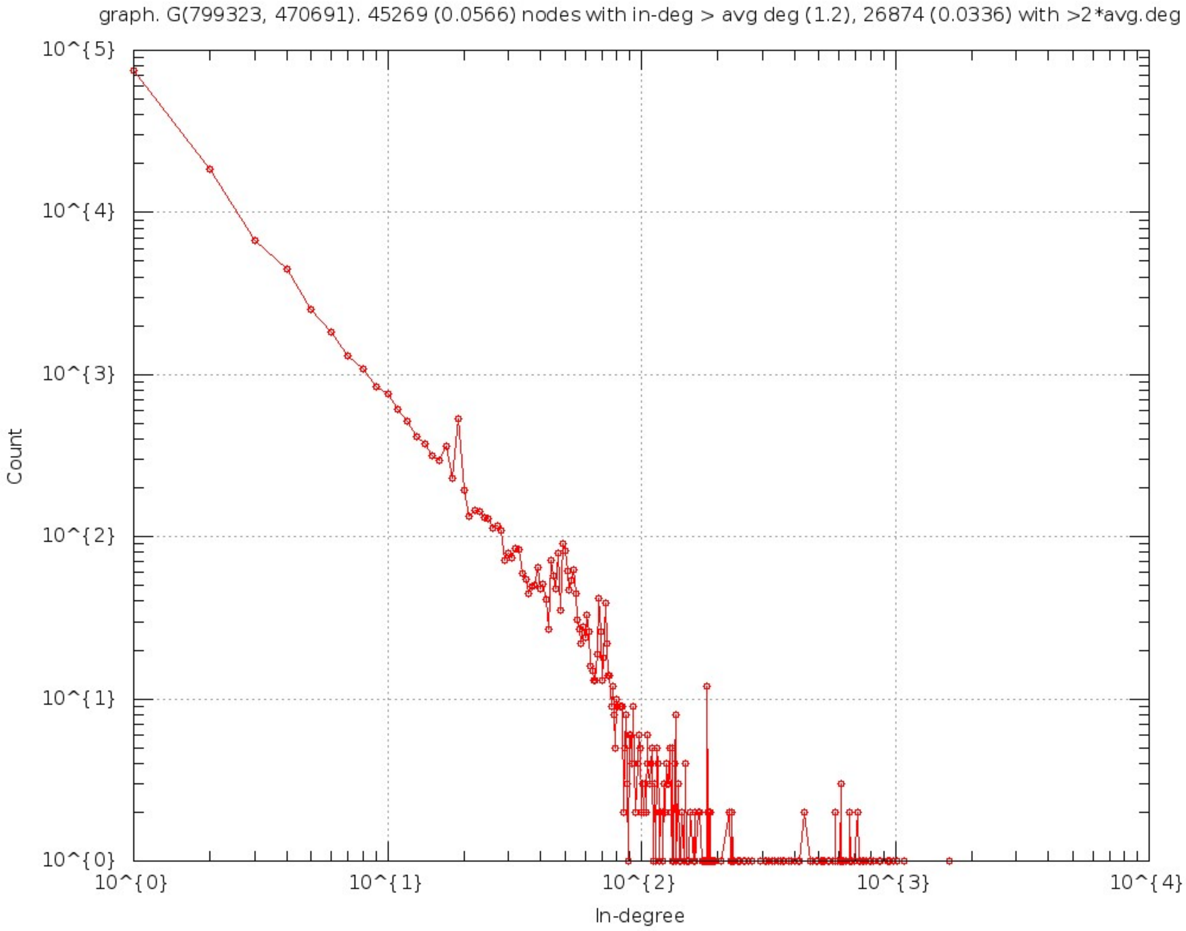
Graph on Average Clustering Coefficient vs. Node Degree



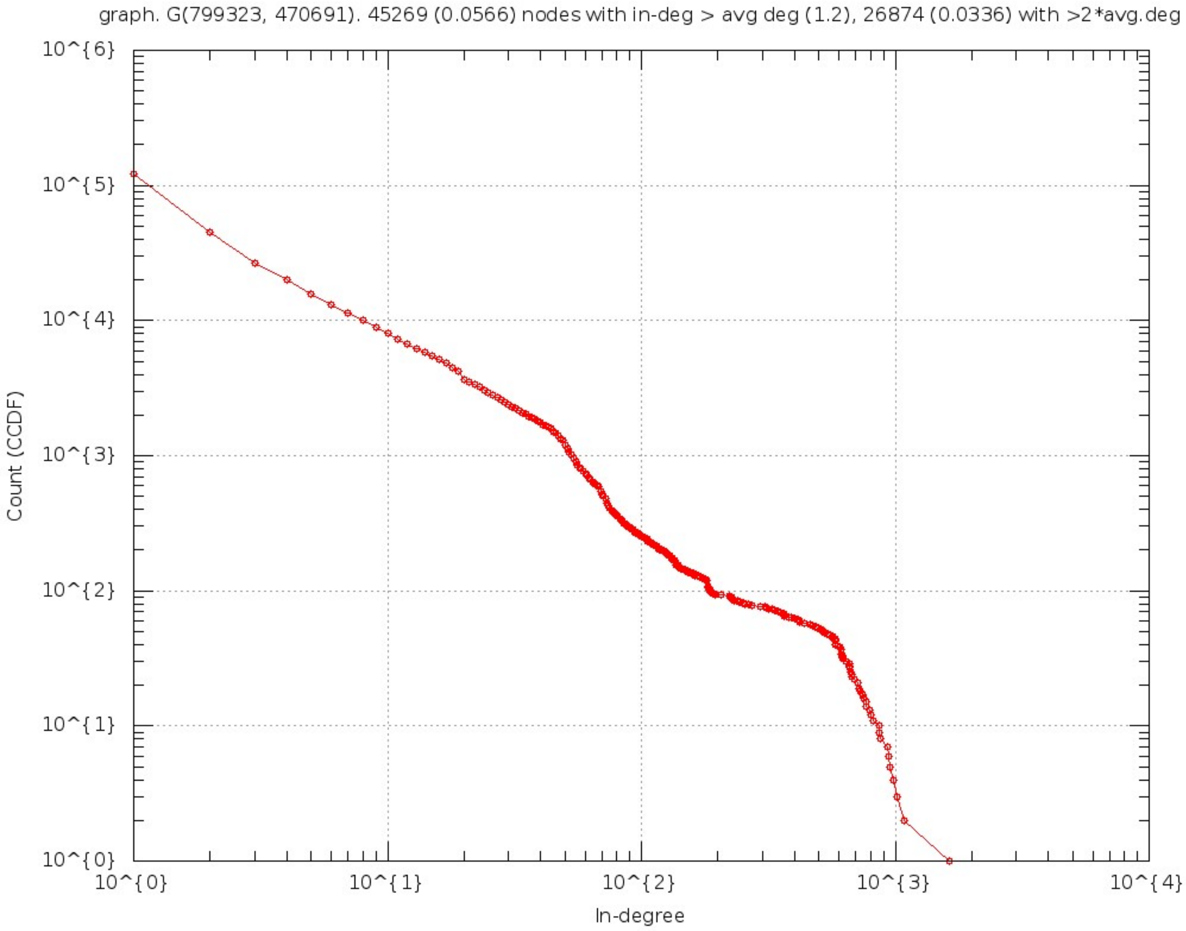
Graph on Number of pairs of nodes vs. Number of hops



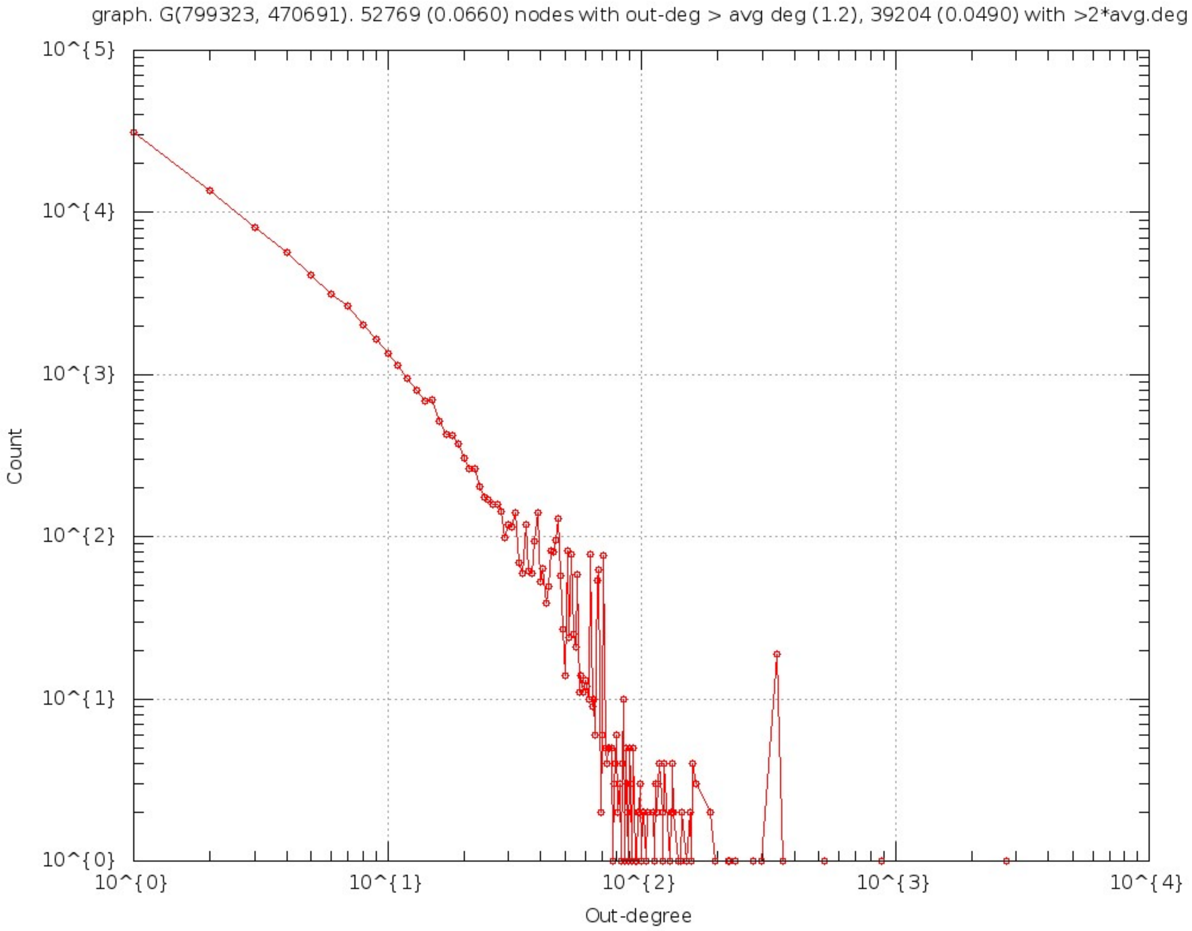
Graph on Count vs. In-degree



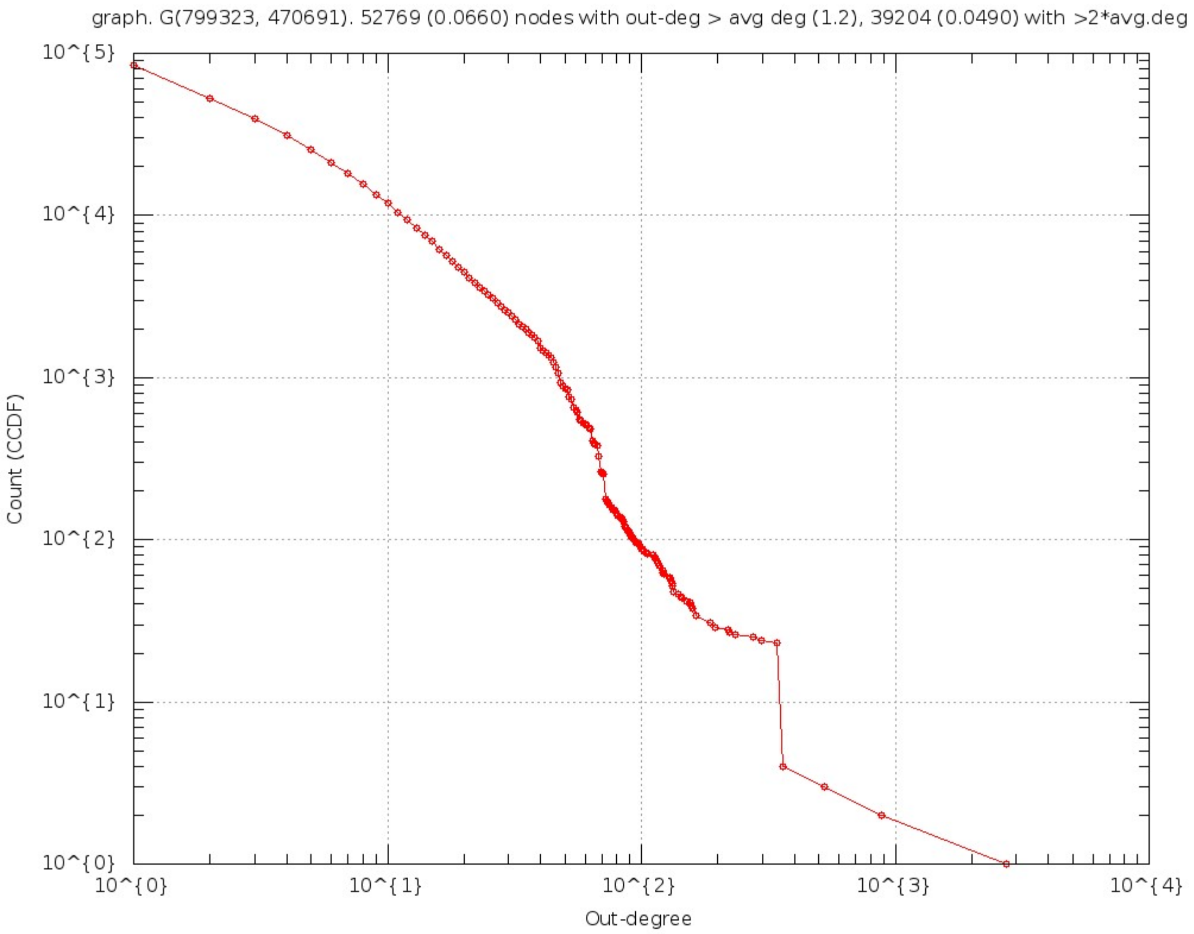
Graph on Count (CCDF) vs. In-degree



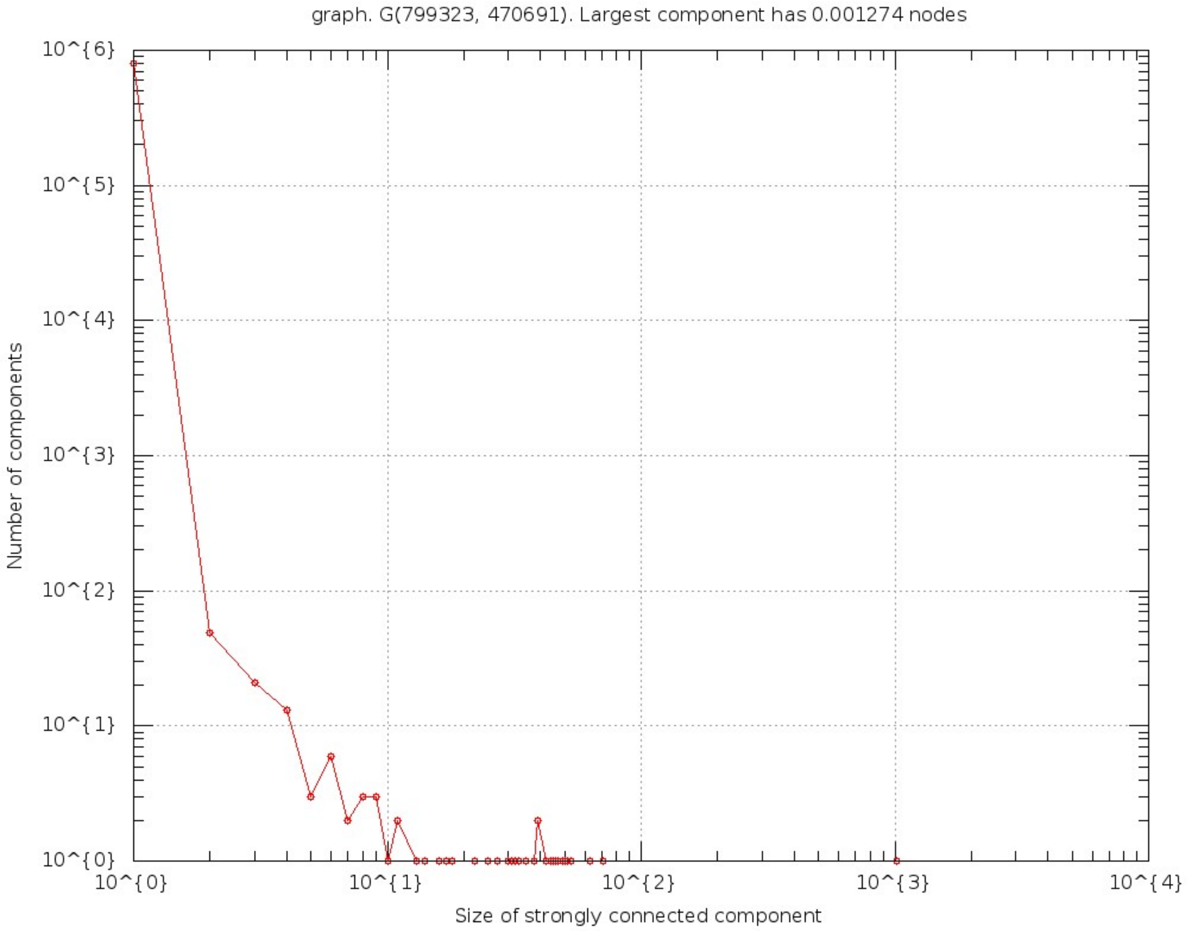
Graph on Count vs. Out-degree



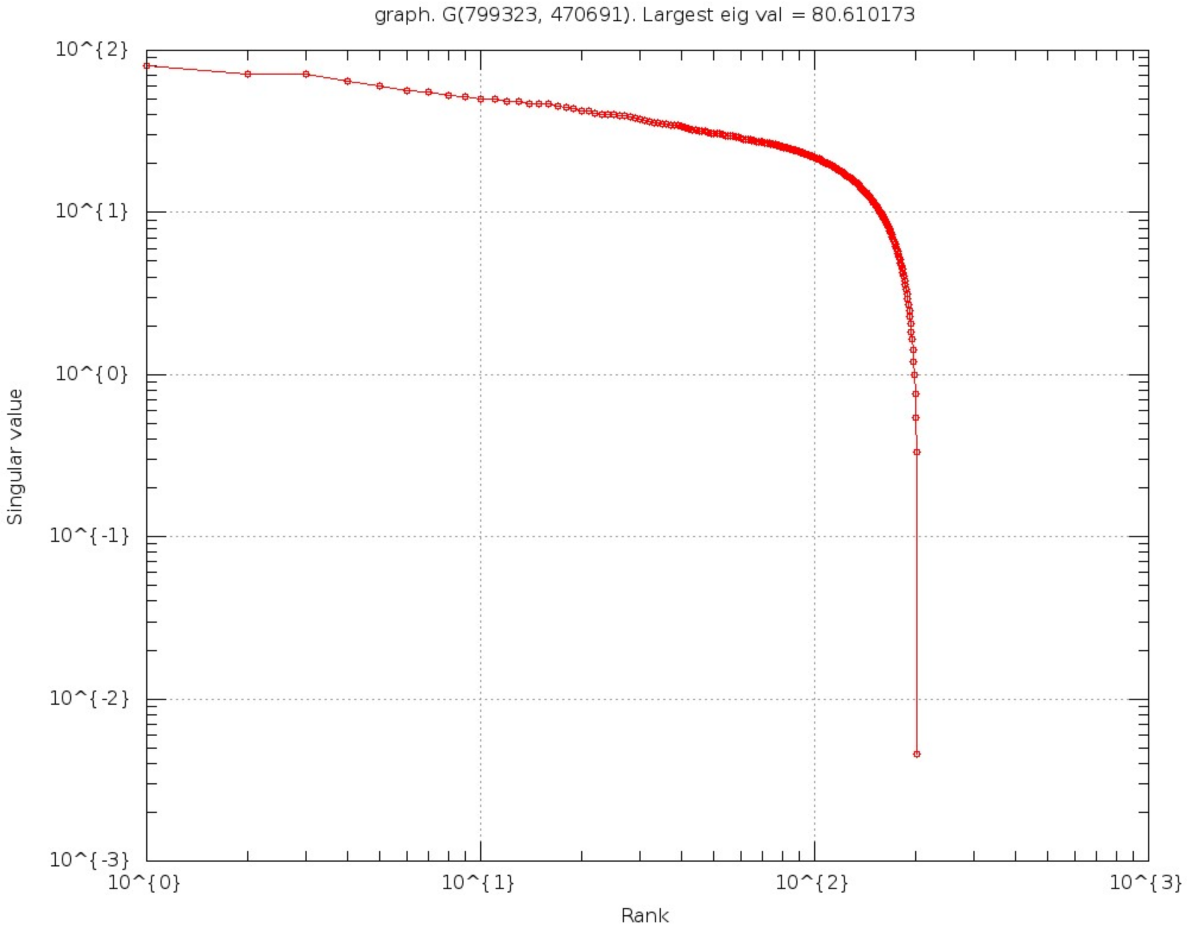
Graph on Count (CCDF) vs. Out-degree



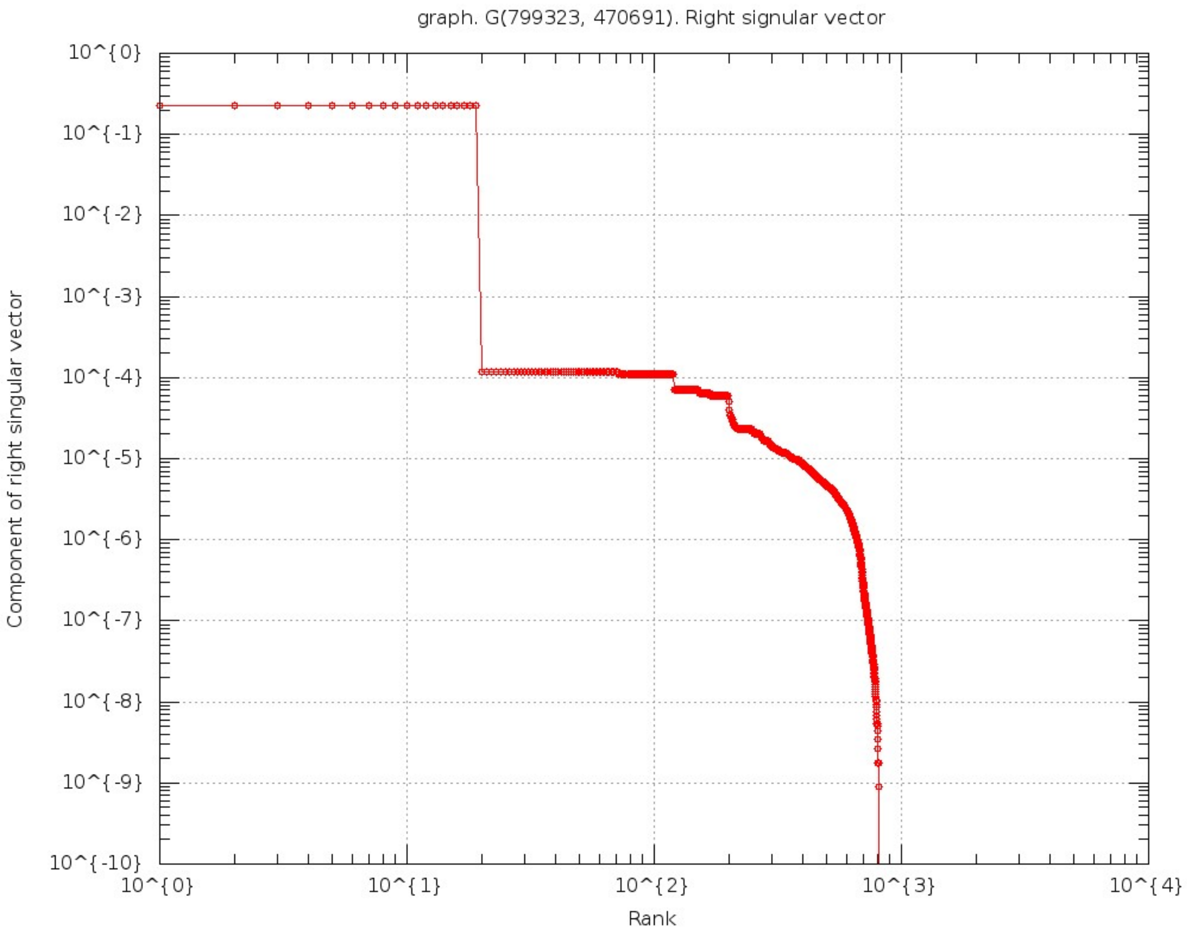
Graph on Number of components vs. Size of strongly connected component



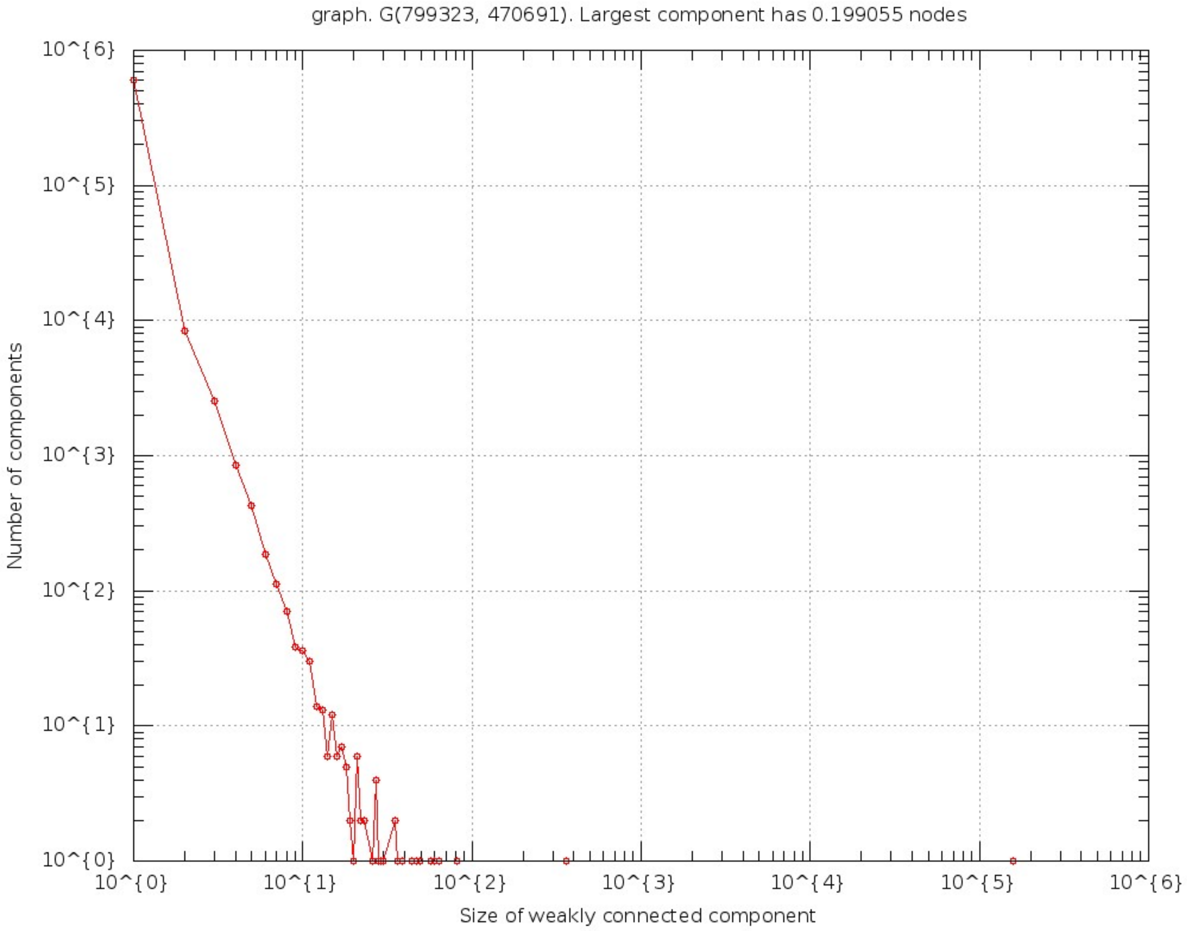
Graph on Singular value vs. Rank



Graph on Component of right singular vector vs. Rank



Graph on Number of components vs. Size of weakly connected component



PROGRAMS AND TOOLS USED FOR DATASET DEVELOPMENT

- The records were downloaded from the domain www.liiofindia.org
  - The records obtained were in HTML and PDF format which were converted to text (using script/online tool)
- The citation records were retrieved using a python script from the domain www.lawcite.com.
- The graph has been formed according to the citation data retrieved.
- Text parsing has been carried out to retrieve more citation links from the documents and complete the graph.

PROGRAMS AND TOOLS USED FOR DATASET DEVELOPMENT

- The program consists of mainly five functions:

- 1. store_current_progress()
- 2. resume_current_progress()
- 3. parse_1()
- 4. parse_2()
- 5. parse_3()

VALIDATION OF THE DATASET

- Sub-databases 1 to 41 include judgements delivered by the Supreme Court of India, various High Courts and the District Courts.
  - Citation details validated
- Sub-databases 42 to 134 contains acts, regulations etc. of the various states of India.
  - No citation information and not considered
- Sub-databases 135 to 154 include articles published in the various law journals
  - Citation details validated

Validation Summary Report

- All the cases in the adjacency list do not have a specific order.
- The database of lliofindia.org has several inconsistencies in the data records.
  - District Court of Allahabad records are occurring multiple times at 1 and 34. 1 is empty while there are records in 34.

Validation Summary Report

- Format of Foreign cases can be varying so interpretation of the metadata can be difficult. 2007.INAPHC.3 of High Court of Andhra Pradesh cites 2004.5.ALD.632 where the format changes from year.month.caseid to a 4 tuple format.

Validation Summary Report

- Citations to Constitution Articles not specific
- Citations are not hyperlinked in some cases
- Citation Text and Citation Hyperlink are mismatching in some cases
- Citations present in the judgements not hyperlinked
- Citations present in the lawcite with hyperlinks but not in the judgement

Citation text patterns not in lawcite

- The following representative text patterns have been identified that refer to citations in the judgements or journal articles which have not been recorded in the lawcite database.
-
- Article 227 of the Constitution of India
- Section 166 of the Motor Vehicles Act, 1988 (in Hindi)
- Section 3 of Explosive Substances Act, 1908
- Section 374 (2) of the Cr.P.C.
- Section 307 of the IPC
- u/s.8(1) of E.A. Act
- Section 4 of the Land Acquisition Act, 1894 (for short 'the ACt')
- Section 6 of the Act
- section 8 paragraph 8 of the Anti-discrimination Act of Slovak Government.
- Article 12 pragraph 2 of the Constitution both of Slovak Government.
- The Orissa (Alteration of Name) Bill, 2010

FUTURE WORK & CONCLUSION

- The various acts, regulations of the different states of India will be considered for their inclusion
- The validation tests have identified several issues which are to be addressed
- Lawcite citation has missed citations to the Constitution articles as well as to various acts
- Various text patterns have been identified
- Text analytic programs developed
- Topic based linking of the various graph nodes will be identified

REFERENCES

- 1. Network Analysis in the Legal Domain - A Complex Model for European Union Legal Sources, Marios Koniaris, Ioannis Anagnostopoulos, Yannis Vassiliou, arxiv.org/pdf/1501.05237.pdf
- 2. Graph Database, https://en.wikipedia.org/wiki/Graph_database
- 3. Stanford Network Analysis Project, Jure Leskovec, snap.stanford.edu
- 4. Legal Information Institute of India, www.liiofindia.org
- 5. Datasets for Empirical legal Research, http://library.law.yale.edu/datasets-empirical-legal-research
- 6. Indian Law Legal Database, www.manupatra.co.in