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

Department of Computer Science and Technology

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

ACKNOWLEDGEMENT

I must take this opportunity to place on record my deep sense of respect and gratitude to **Dr. Saptarshi Ghosh and Dr. Asit Kumar Das,** our laboratory teachers for the term project for their valuable advice, resourceful guidance, active supervision and constant encouragement without which it would not have been possible to complete this project report.

I also want to express my heartfelt gratitude towards Prof. Biplab Kumar Sikdar, Head, **Department of Computer Science and Technology, IIEST Shibpur**, and to all the teachers in the department for their guidance and active support.

Date 15/11/2016 SAPTARASHMI BANDYOPADHYAY

Examination Roll:510514006
Fifth Semester 2016-2017
Dual Degree (B.Tech. - M.Tech.) in Computer Science and Engineering

CONTENTS

CHAPTER 1: INTRODUCTION

A great part of the world's knowledge is stored using text in natural language, but using it in an effective way is still a major challenge. **Natural Language Processing** (NLP) techniques provide the basis for harnessing this huge amount of data and converting it into a useful source of knowledge for further processing. NLP is used in a wide variety of disciplines to solve many different types of problems. Analysis is performed on text from different sources, such as blogs, tweets, and various social media, with size ranging from a few words to multiple documents. Machine learning and text analysis are frequently used to enhance already existing services or to create completely new functionality. Some of the application areas could be: Search, Sentiment Analysis, Summarization, Named Entity Recognition and Question Answering.

Text is often referred to as unstructured data. However, in reality, free text has a lot of structure - it's just that most of it isn't explicit, making it difficult to search for or analyze the information within the text. NLP uses computer science, artificial intelligence and formal linguistics concepts to analyze natural language, aiming at deriving meaningful and useful information from text.

In particular, **Information Extraction** (IE) is the first step of this process. It attempts to make the text's semantic structure explicit so that it can be more useful. More precisely, IE is the process of analysing text and identifying mentions of semantically defined entities and relationships within it. These relationships can then be recorded in a database to search for a particular relationship or to infer additional information from the explicitly stated facts. In order to build a useful database, IE must do much more than to find a sentence in the text: it must identify the event's participants and properties, resolve pronouns and compute dates and times.

Due to the highly connected nature of the data produced, a **suitable model** for representing them is in the form of a **graph**. It not only stores the main data and relationships extracted during IE process but allows further extension by adding new information computed in a post-processing phase, such as similarity, sentiment extraction, and so on.

Legislation is a large collection of different normative documents, which keeps growing and changing with time. As legislation increases in size and complexity, finding a relevant norm may be a challenging task even for experts. Furthermore, the process of drawing up a consistent and coherent legislation framework becomes a more and more challenging task. Drafting of new or amending existing legislation are very complicated processes. As a result authorities at national and state level, often consider proposed regulations for months or years before they finally become effective. Thus, it is critical to firstly quantify the legal complexity and then work

towards the provision of a model that will assist us to reveal the emergent dependencies among the legislation corpus.

Typically, legal documents refer to authoritative documents and sources (e.g. most commonly regulations, treaties, court decisions, and statutes). Computer scientists and legal experts have used citation analysis methods, in order to construct case law citation networks, as well as to further model and quantify the complexity of the legislation corpus.

Citation analysis has been used in the field of law to construct case law citation networks. The American legal system has been the one that has undergone the widest series of studies in this direction. In all of the above studies, the law graph is treated as a citation network, thus showing the effectiveness of network analysis in the legal domain. In one hand, it was proven that case law citation networks contain valuable information, capable of measuring legal authority, identifying authoritative precedent, evaluating the relevance of court decisions, or even predicting the cases that will receive more citations in the future. Yet, on the other hand, citation network analysis over the legislation corpus, provides us information over a single dimension view. Edges on the graph are of the same type and just simple references between documents.

However, in the real-life paradigm of legal domain, there are multiple and heterogeneous networks, each representing a particular kind of relationship, and each kind of relationship plays a distinct role in a particular legal norm. Thus, in order to construct a network model that simulates legislation in a quite robust way, we have to take into account the multi-scale structure of law. Distinct features of the law as the hierarchy between the sources of law, or different types of relations between legal documents should be properly carved and incorporated into a model.

During the course of study, it has been noticed that the only text database for legal documents (articles, judgements, acts, regulations, journals etc.) in the Indian context are available in the freely available website www.liiofindia.org. The web site includes citation analysis carried out manually and the task is not yet complete. The site has only limited search capabilities which are not sufficient for the information need of the legal communities, common people and the law makers. The present situation has prompted us to design and develop the citation graph model of the Indian legal database as present in the website www.liiofindia.org.

CHAPTER 2: SCOPE OF THE WORK

2.1. Graph Databases - A Review

In computing, a **graph database** is a database that uses graph structures for semantic queries with nodes, edges and properties to represent and store data. A key concept of the system is the *graph* (or *edge* or *relationship*), which directly relates data items in the store. The relationships allow data in the store to be linked together directly, and in many cases retrieved with a single operation.

This contrasts with conventional relational databases, where links between data are stored in the data itself, and queries search for this data within the store and use the JOIN concept to collect the related data. Graph databases, by design, allow simple and rapid retrieval of complex hierarchical structures that are difficult to model in relational systems. Graph databases are similar to 1970s network-model databases in that both represent general graphs, but network-model databases operate at a lower level of abstraction[1] and lack easy traversal over a chain of edges.

The underlying storage mechanism of graph database products varies. Some depend on a relational engine and store the graph data in a table while others use a key-value store or document-oriented database for storage, making them inherently NoSQL structures. Most graph databases based on non-relational storage engines also add the concept of *tags* or *properties*, which are essentially relationships lacking a pointer to another document. This allows data elements to be categorized for easy retrieval *en masse*.

Retrieving data from a graph database requires a query language other than SQL, which was designed for relational databases and does not elegantly handle traversing a graph. As of 2016, no single graph query language has been universally adopted in the same fashion as SQL was for relational databases, and there are a wide variety of systems - most often tightly tied to a particular product. Some standardization efforts have taken place, leading to multi-vendor query languages like Gremlin, SPARQL, and Cypher. In addition to having query language interfaces, some graph databases are accessed through APIs.

In the present work, each legal document has been assigned an index number. The index number is unique throughout the system. Each legal document is coded in a dot notation that includes the year, the code used by the website www.liiofindia.org referring to the document and the document id within the liiofindia database. In addition to this metadata information, another citation information table is maintained. The citation details for each legal document

are maintained in a linked list format in which the index number for each document is the header and the index numbers of the cited documents are the nodes in the linked list. In future, such citation information may be stored in edge list format in order to facilitate query processing.

2.2. SNAP (Stanford Network Analysis Project) - A Review

Stanford **N**etwork **A**nalysis **P**latform (**SNAP**) is a general purpose network analysis and graph mining library. It is written in C++ and easily scales to massive networks with hundreds of millions of nodes, and billions of edges. It efficiently manipulates large graphs, calculates structural properties, generates regular and random graphs, and supports attributes on nodes and edges. SNAP is also available through the NodeXL which is a graphical front-end that integrates network analysis into Microsoft Office and Excel.

Snap.py is a Python interface for SNAP. It provides performance benefits of SNAP, combined with flexibility of Python. Most of the SNAP C++ functionality is available via Snap.py in Python

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. In includes social networks, web graphs, road networks, internet networks, citation networks, collaboration networks, and communication networks. The Stanford Large Network Dataset collection includes citation networks among others but that includes citations of research papers. It is planned in the present work, to design and develop the citation network of Indian legal database so that the final citation network follows the Snap model.

2.3. LiiofIndia - A Review

An attempt has been made to document the Indian legal corpora in a graph database for the first time. The Indian Legal corpora is one of the largest in the world, due to its sheer number of litigants.

The Legal Information Institute of India (LII of India) www.liiofindia.org is an international standard, free-access and non-profit, comprehensive online ollection of Indian legal information. The prototype is open for public use on 25 November 2010. Four leading Indian Law Schools are the initial Indian project partners: three National Law Schools (NALSAR University of Law, Hyderabad; National Law School of India University, Bangalore; and National Law University, Delhi), plus Rajiv Gandhi School of Intellectual Property Law, Indian Institute of Technology – Kharagpur. The technical hub of the project is at NALSAR.

In 2005 - 08 AustLII developed a substantial number of databases of Indian law under its Asian Legal Information Institute (AsianLII) project, funded by AusAID , and its Commonwealth Legal Information Institute project, funded by the Australian Research Council (ARC). NLU Delhi, NALSAR and NLUI in Bangalore offered their support. In 2010 the new LII of India website was

given a ‘soft launch’ in Australia. In 2010 and 2011 five more National Law Universities joined in supporting the LII of India project, bringing the total to 8 Partner Institutions.

In 2010 two of AustLII’s Co-Directors attended meetings in India to plan the development of LII of India, and in 2011 returned for further meetings of all Law Schools involved, and for a major launch of LII of India in Delhi (hosted and funded by NLU Delhi), and satellite launches in Hyderabad, Bangalore and Kolkata (hosted respectively by NALSAR in Hyderabad, NLSIU in Bangalore and NUJS in Kolkata). Up to this point, AustLII developed the LII of India to include over 150 databases, and to include over 800,000 court decisions, obtained from public web sites in India.

AustLII is keeping some of the LII of India databases updated by automated means, and will continue to do so as best it can, including maintenance of the Supreme Court of India database and databases of journal articles from NALSAR, NLU-Delhi and NUJS. At the Indian end, no funds to sustain LII of India have been made available, nor has any entity been established in India for the purposes of its long-term management.

2.4. Legal Graph Database Development in the World and in India

The Lillian Goldman Law Library of Yale Law School maintains a large collection of Foreign and International Law Resources.

Manupatra, pioneer in online legal research in India since 2001, is India’s premier legal information resource. It is the largest content aggregator of Indian and International material, linking primary information, secondary material and proprietary analytical content.

The Company started operations in the year 2000 and launched its flagship product, the online database www.manupatra.com in August 2001. Today, it has delivery capabilities in Online, Print and Mobile media.

Over the years, Manupatra has reinvented legal research by including intuitive and smarter legal research tools with database access, thus strengthening lawyer’s practice. Research on Manupatra ensures that users spend their time analyzing information, and in context and not gathering it. Sophisticated legal research tools, which were the province of the privileged, has been brought to the masses by Manupatra.

India’s premier legal information resource, is designed to be used by a wide variety of users across Legal, Educational, Finance, Tax, Accounting, Corporate, Risk Management, Banks, Consulting, Government, Law Enforcement, Intellectual Property, Media markets and others. The important supports provided by the system are:

- authoritative and editorially enhanced content accepted by Indian Courts;
- Enterprise Search Platform with advance search options makes research easy, quick and accurate;
- quality legal research on the move, through our Web and Mobile App;

- features such as Analytics, Visualisation Tools, Integrated Citation, Apps , News Alerts & more;
- documents have backward and forward cross referencing and extensive hyper linking saving time;
- a database of over 20 Lakh Case Laws with Citation Search on 300+ equivalent citations in addition to other content.

The service is available for a fee.

2.5. Research Significance of the Dataset Developed

Most legal research involves relationship analysis. One searches or uses analytical material to find a starting point then examines relationships to do the research. If one finds an opinion that is on point to a legal problem, it is likely that opinions that cite or are cited by that opinion are also on point. Citation relationships among opinions create a cluster of opinions that are likely to address the same topic. In fact, a legal search engine should incorporate relationship analysis. Because relationship analysis is the main part of legal research, a graph database is a natural way to represent the relationships among legal content.

A major issue with graph analysis of legal content is how to filter the relationships to so that one can do a useful analysis. An important court opinion can be cited by thousands of other opinions. A subgraph, starting from such an opinion and having just two levels of citation, could have so many nodes that it would be unmanageable for analysis. The most useful way to filter a court opinion citation subgraph for legal research purposes is by legal topic. That kind of filtering requires taking a different approach to organizing legal topics.

Citation relationships are the best-known category in legal research. One type of relationship that current legal research systems tend not to exploit is relationships among content types. For example, if we are doing statutory research, we will examine relationships like these that usually do not exist in legal research system. A direction of research recognizes the graph relationships among content types and allows the customer to start from a document of a specific type and return all the documents of different types that are related to the original document.

Current legal research systems merely are search engines on top of legal content. However, search is not research. A legal research system that provided the ability to easily analyze the graph structures inherent in legal information would be a great improvement over the systems that exist now.

CHAPTER THREE : FEATURES OF THE DATASET

3.1. LIIOFINDIA Legal Database

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
- The amendment bills
- The treaties

In addition to the above, a court case can refer to another famous case outside of Indian judiciary. Also, a case may refer to Law journal articles.

The data being downloaded includes all the cases cited in www.lawcite.com from the above classification except the foreign cases.

Also, a graph dataset is being prepared based on the dataset downloaded. The graph is a directed graph in which the nodes are the documents(it may be court cases, an article, an act, a regulation a scheme, a part of the constitution etc.) and an edge is defined from A to B when A has cited B.

The graph has been represented in both edge list format and connection list format. The graph has been represented in edge list format as a variant of adjacency-lists. Unlike adjacency-lists which contains only one list for each node we have considered five lists for each node, Each containing the list of:

- Cases referring to that case.
- Law Reforms referring to this case.
- Law Journal Articles referring to this case.
- Legislation cited.
- Cases and Articles.

NOTE:- only the fourth and fifth types gives the outwardly adjacent nodes and the first three gives the inwardly adjacent nodes.

The integration of both inwardly and outwardly adjacent nodes apparently seem to be redundant but needs to be considered for the following reasons:

- We had not considered every single citation document as nodes in the legal graph dataset. For example, the case [1951] INSC 61 cites an Australian legislation named the oaths act for which there is no citation record available under the "liiofindia.org". However presently, we are adding the node name to these cases as well, even though they do not have any record of the judgement and they will remain there as isolated nodes.
- Later, it may prove to be efficient for use in algorithms and query processing.

The graph has also been prepared in a connection list format in accordance with SNAP dataset. Two files have been used, a node_list_for_metadata which contains the list of the name of the nodes of the individual law documents in the lawcite record and their respective indices while the connection list shows the connectivity of one node with other nodes in a directed manner.

The task to obtain the data was tedious given the sheer size of the data. The initial idea was to get the whole database of cases using variations of the wget command. But, then we discovered that not only cases but other acts, or parts of the constitution were also citable by a case of court. So, we needed to download the acts as well as the constitution and the acts and all that were citable. In due course we understood that it literally meant that we needed that the whole data present on the website "www.liiofindia.org". So, we tried to get the whole data in a structured way. But, various attempts remained failed including proprietary windows software such as internet download manager which has an in-built site grabber. But, they were all unable to grab the whole information. Finally, we had to code a crawler in python without using any crawling framework to retrieve the data.

But now there was a new problem. We needed to determine edges to visualise the data as a graph. But to do so we needed to correctly identify the citations from the text of the data. This approach needed massive text processing which would both be very inaccurate and time consuming. Also, different courts of law had different formats of documenting their cases. So, now that the data was available the problem was to generate edges.

Next an important observation was made that made the edge retrieval possible and accurate as well. Another domain named LawCite under the original liiofindia.org was maintained just for the purpose of maintaining citation records of different cases. Later we found that it held citation records for not only court cases but also journal articles, International treaties, Amendment bills as well. Also these citation records were found to be exceptionally well formatted in tabular fashion. And were complete and non-redundant. The next step was simple to derive. It was to just get the citation links directly from the Lawcite site. These citation links are the edges of the representational large graph. These citation records of different court cases, journal articles or treaties were stored in the form of html files and were named according to their node names for ease of later use.

The technique used to get the data and hence form the graph can be stepwise described briefly as:

- For each documented case downloaded, it corresponding LawCite record was visited to fetch its citation record and corresponding entries were added in the adjacency list and stored it in a html file for further use.
- And recursively fetch lawcite records in a breadth-first manner of those cases whose judiciary falls under India and has no documentation available.

There were certain primary observations made on the citation records:

- If there exists a connected component within the Lawcite site none of whose nodes are documented then we would miss that. Although it is highly unlikely.
- we have not considered the nodes which has neither any documentation nor any citation. And apparently they are undetectable. So, they need not be considered.
- the records included court cases from courts of non- Indian judiciary too, i.e. court cases from outside India.
- There were even certain Indian cases for which there were citation records but no online documentation was available.
- Every documented case had complete citation record including citations to and from undocumented nodes.
- The data is verified to be non-redundant. Although there can be several names of the same case but they only had one documentation under the domain.

For example, the cases [1961]INSC 128, [1962] 1 SCR 694 and AIR 1961 SC 1471 are all the same cases. Although the significance, requirement or cause of these different names were not clear, it was hypothesized that these undocumented names are either alternate conventions for naming or are the names when this cases were in lower courts.

3.2. DATASET STATISTICS

The two tables node_list_for_metadata and connection_list have been developed in the SNAP format. Detailed statistics about the developed graph dataset has been obtained using SNAP tools and are mentioned below:

```
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: cummulative 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
=======================================================
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)
```

The following statistics identifies the total number of judgements delivered by the Supreme Court of India, the various High Courts and the district courts as well as the numerous articles published in the various law journals. The codes used by the www.liiofindia.org database to identify the judgements or the articles have also been mentioned in the detailed statistics.

```
Supreme Court of India INSC 33548
High Court of Himachal Pradesh INHPHC  10
High Court of Karnataka INKAHC 6
High Court of Allahabad, Lucknow Bench INUPLUHC 172
High Court of Jharkhand INJHHC 24
High Court of Uttarakhand INUTHC 13282
High Court of Bombay at Goa INGAHC 2292
High Court of Punjab and Haryana INPBHC 63741
High Court of Calcutta Port Blair Bench INWBKOHCPB 1522
High Court of Kerala INKLHC 21603
High Court of Orissa INORHC 7371
High Court of Calcutta INWBKOHC 139160
High Court of Chattisgarh INCTHC 108
High Court of Judicature at Allahabad INUPHC 54834
High Court of Rajasthan INRJHC 11504
High Court of Madhya Pradesh INMPHC 931
High Court of Gujarat INGJHC 11
High Court of Jammu and Kashmir INJKHC 44
High Court of Gauhati INASHC 63
High Court of Judicature at Patna INBRHC 38
High Court of Judicature at Bombay INMHHC 975
High Court of Andhra Pradesh INAPHC 3984
High Court of Judicature at Allahabad INAHHC 54834
High Court of Delhi INDLHC 39833
```

High Court of Calcutta (Appellate Side) INWBKOHCA 102358
High Court of Madras INTNHC 6100
District Court of Chandigarh INCHCHDC 22
District Court of Nainital INUTNADC 14
District Court of Jodhpur INRJJODC 17
District Court of Kamrup INASKPDC 31
District Court of Bhopal INMPBPDC 14
District Court of Allahabad INUPAHDC 4
District Court of Delhi INDLDLDC 2519
District Court of Ranchi INJHRADC 41
Indian Central Administrative Tribunal INCAT 17426
Indian Cyber Appellate Tribunal INCyberAT 7
Central Information Commission of India INCIComm 103372
Indian Intellectual Property Appellate Board INIPAB 15
Indian Railway Claims Tribunal INRCT 15
Indian Appellate Tribunal for Electricity INATEl 15
NALSAR Law Review NALSARLawRw 75
Indian Treaty Series INTSer 748
Indian Journal of Intellectual Property Law INJlIPLaw 46
NALSAR Student Law Review NALSARStuLawRw 61
NUJS Law Review NUJSLawRw 72
NALSAR Law Research Series NALSARLRS 10
Indian Journal of Constitutional Law INJlConLaw 44
Indian Journal of Law and Economics INJlLawEcon 9
GNLU Journal of Law, Development and Politics GNLUJlLDP 18
Journal of Intellectual Property Rights INJlIPR 324
NALSAR Media Law Review NALSARMLawRw 29
Indian Journals of Law and Technology INJlLawTech 17
NLUD Student Law Journal NLUDStuLawJl 13
NLUD Law Research Series NLUDLRS 109
Law Commission of India INLC 80
NALSAR Environmental Law and Practice Review NALSAREnvLawPRw 10
Indian Parliamentary Research Service Legislative Summaries INPRSLS 139

The following graphs have been obtained using the SNAP tools. The graphs represent the following statistics:

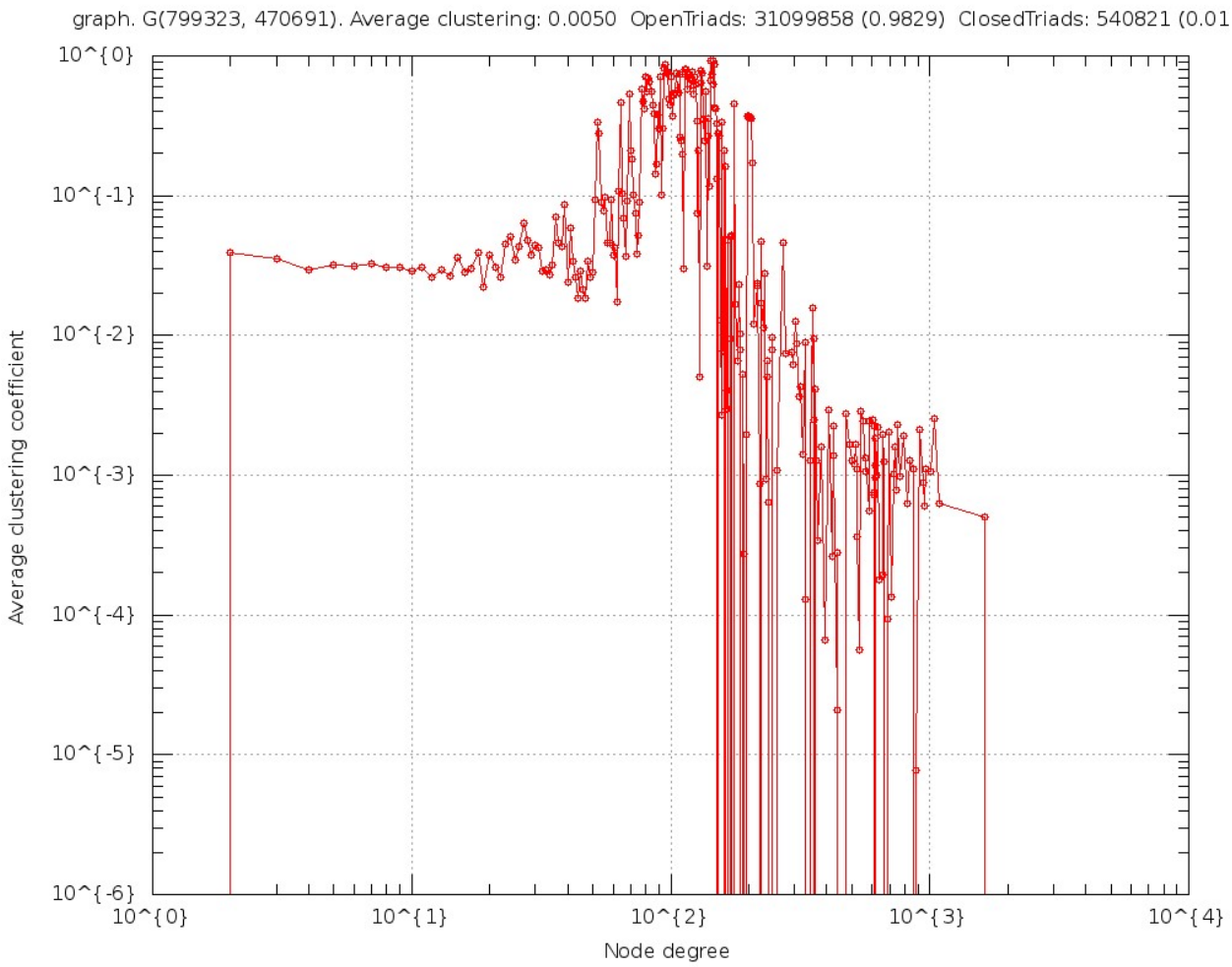

Figure 1. Graph on Average Clustering Coefficient vs. Node Degree

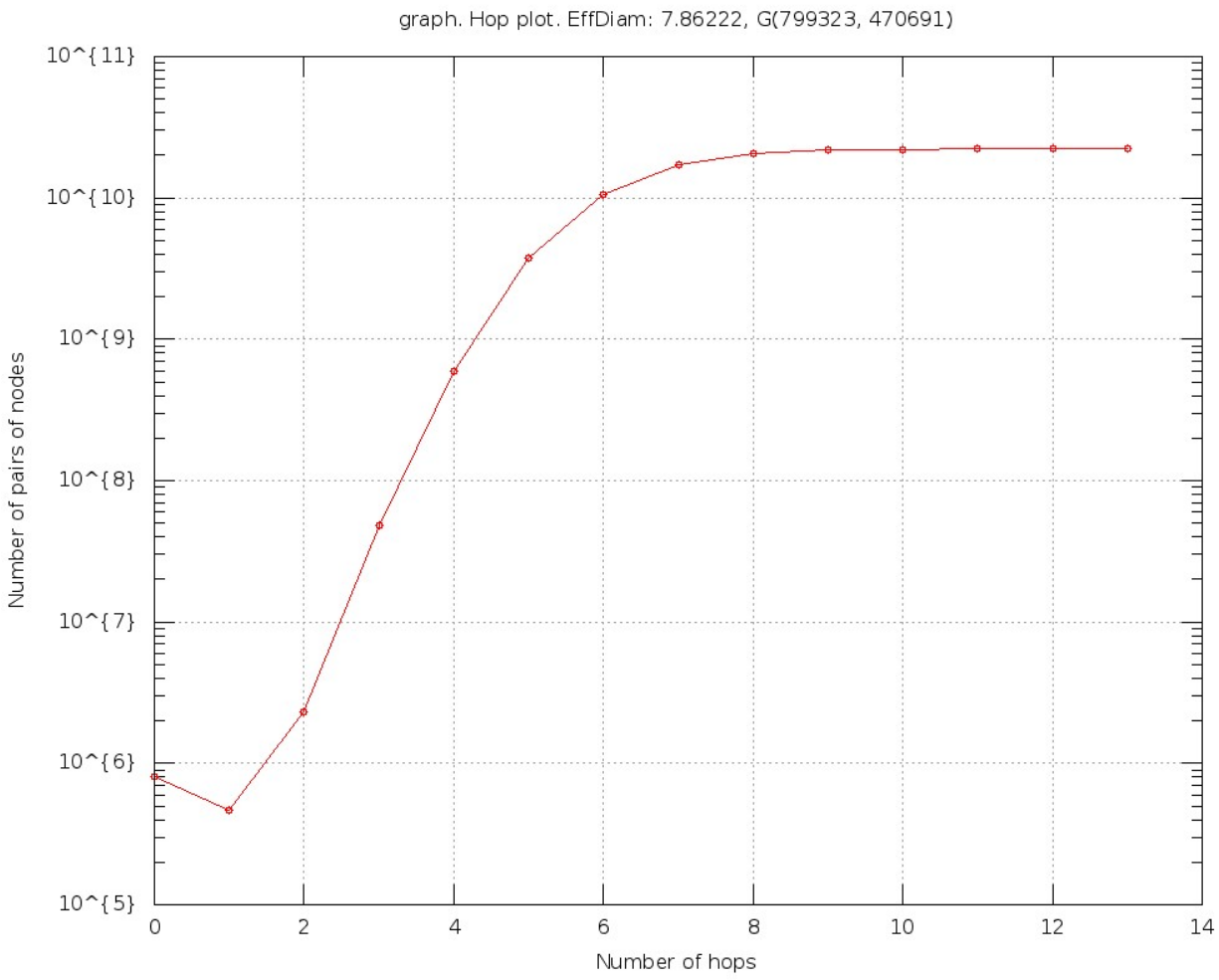

Figure 2. Graph on Number of pairs of nodes vs. Number of hops

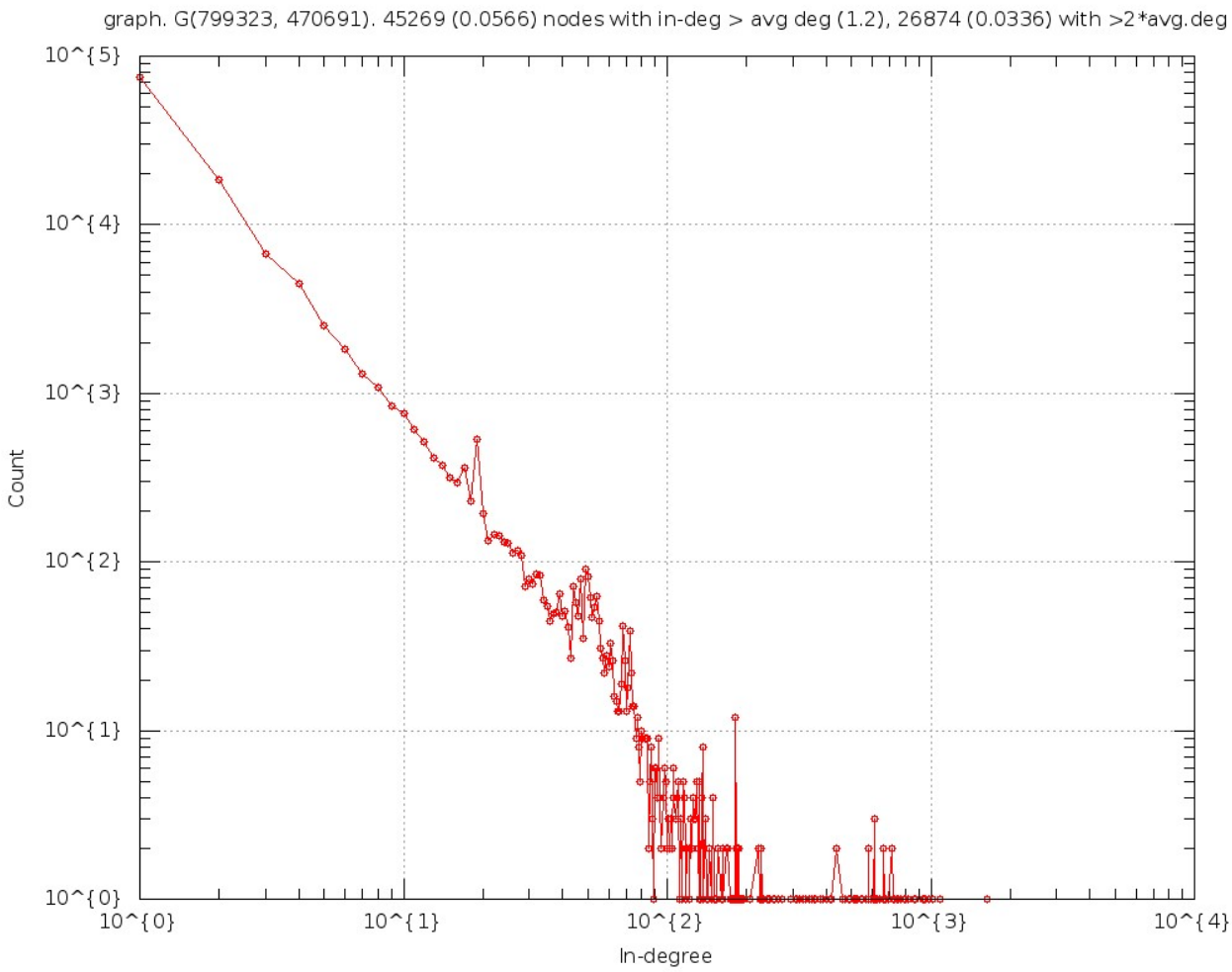

Figure 3. Graph on Count vs. In-degree

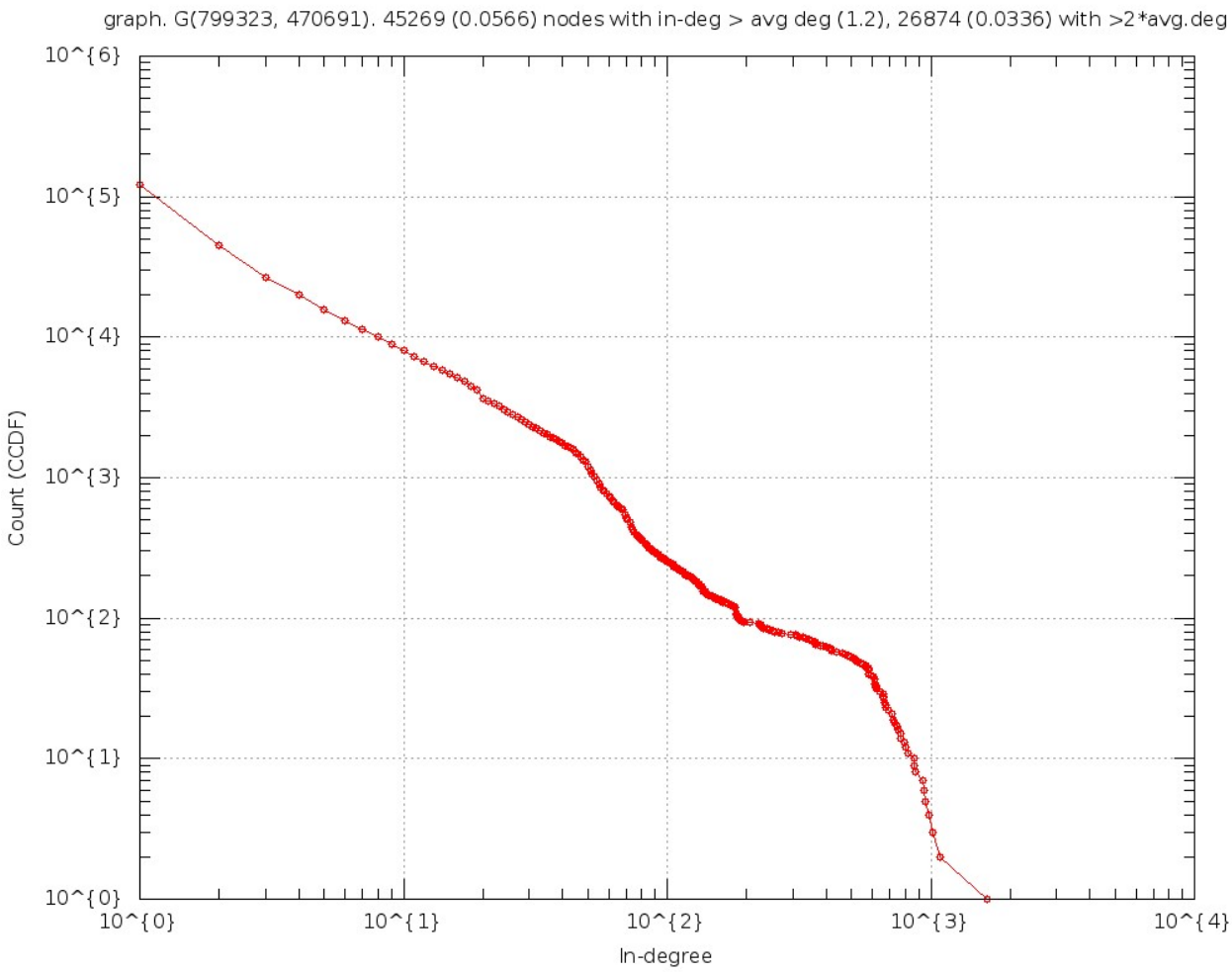

Figure 4. Graph on Count (CCDF) vs. In-degree

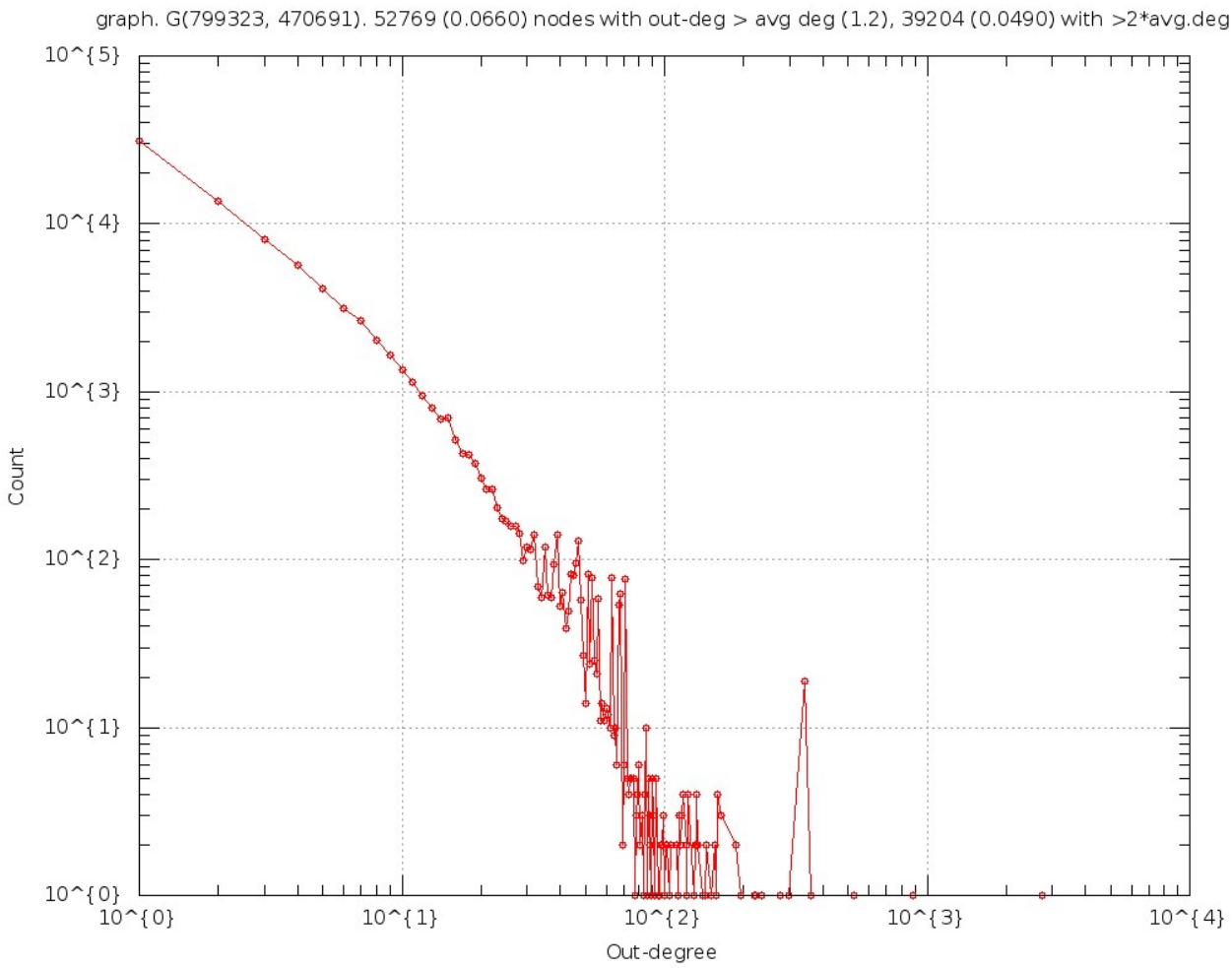

Figure 5. Graph on Count vs. Out-degree

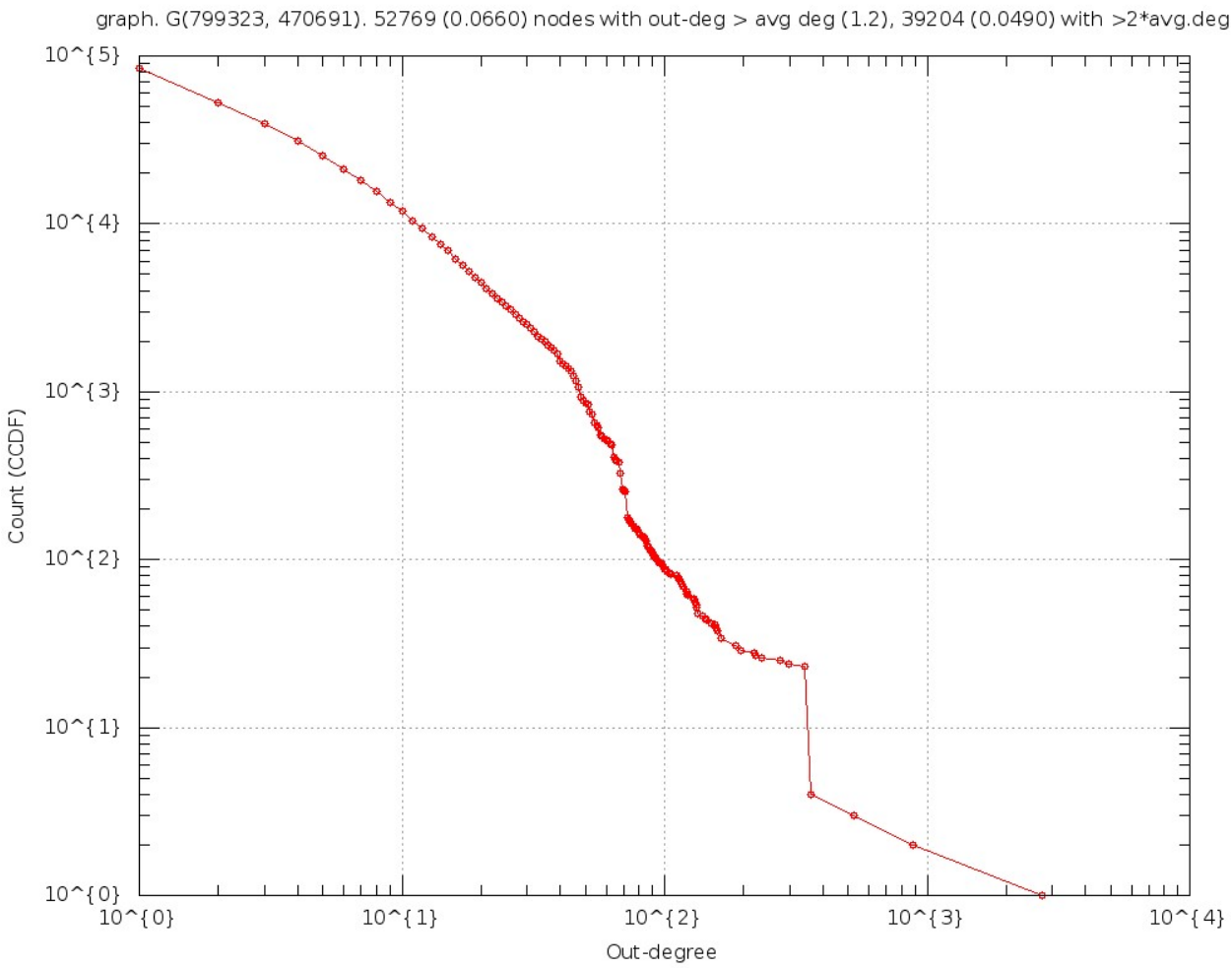

Figure 6. Graph on Count (CCDF) vs. Out-degree

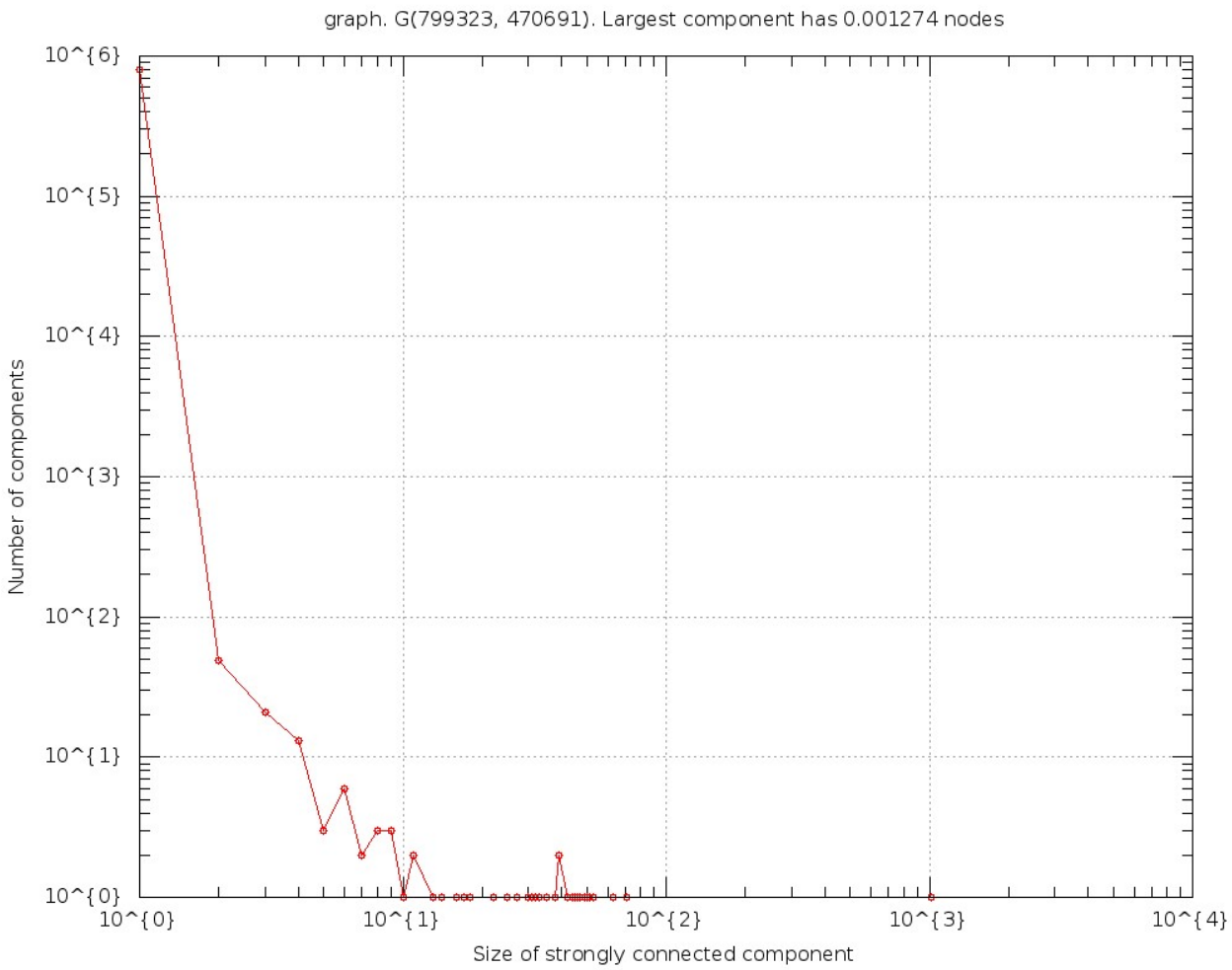

Figure 7. Graph on Number of components vs. Size of strongly connected component

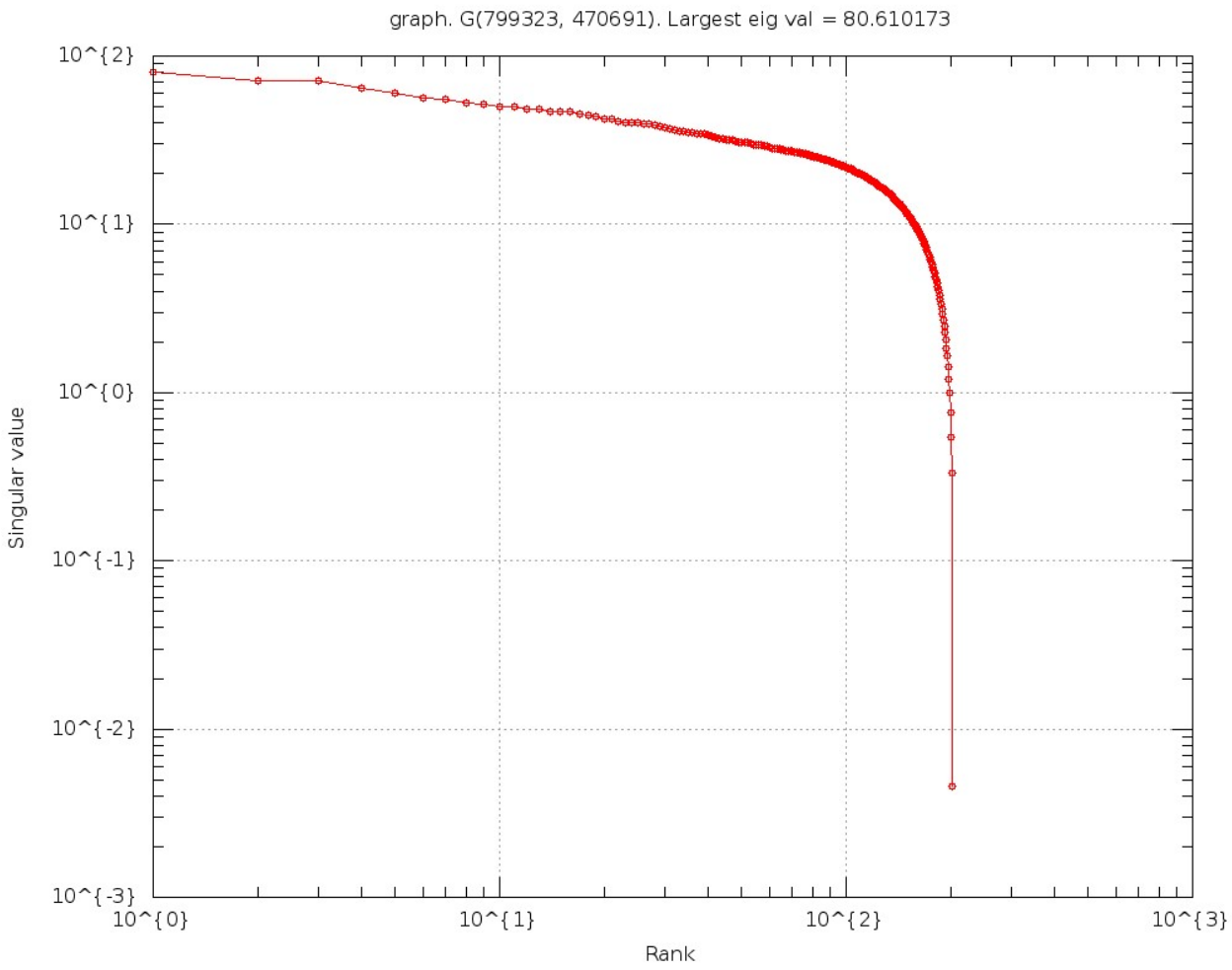

Figure 8. Graph on Singular value vs. Rank

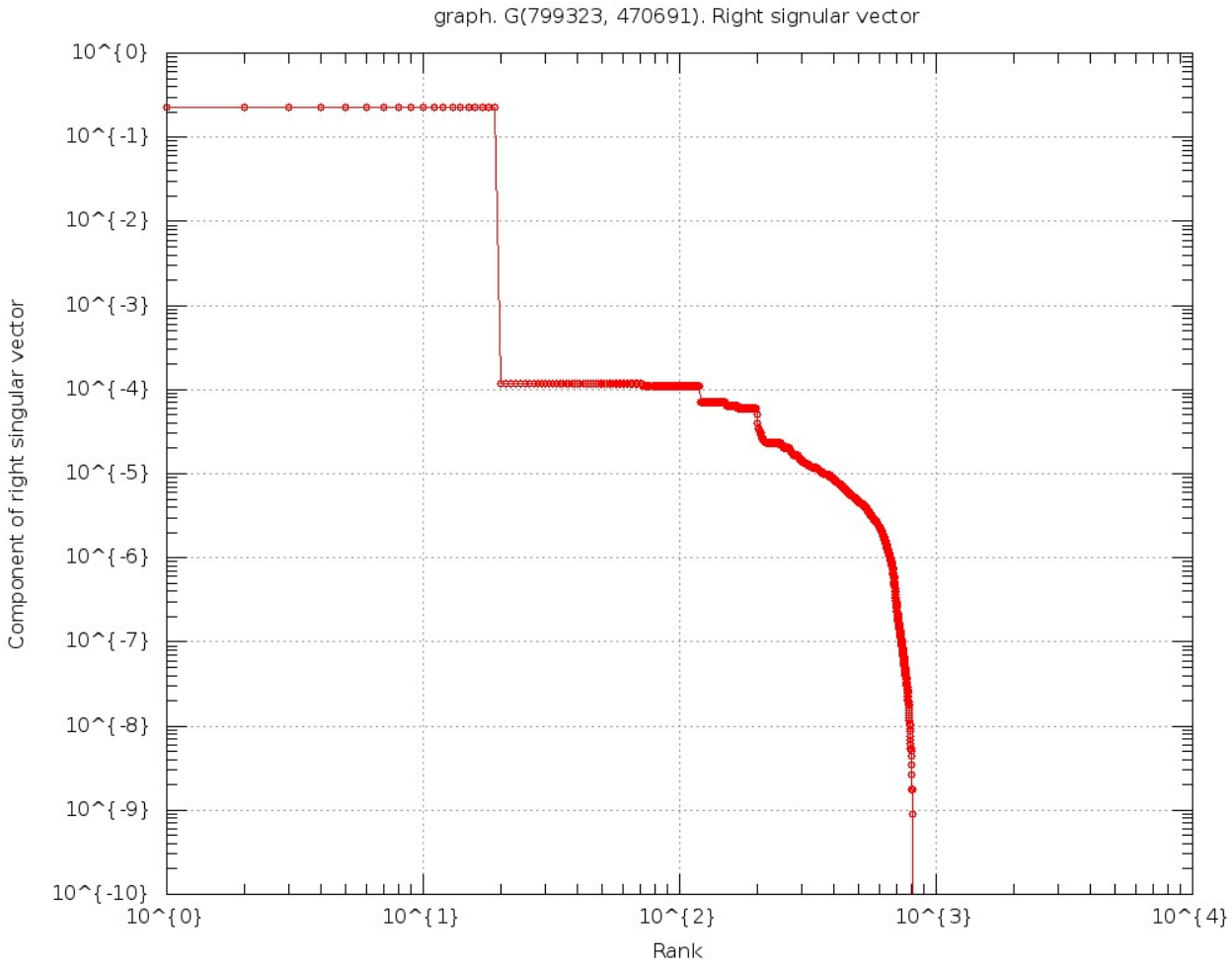

Figure 9. Graph on Component of right singular vector vs. Rank

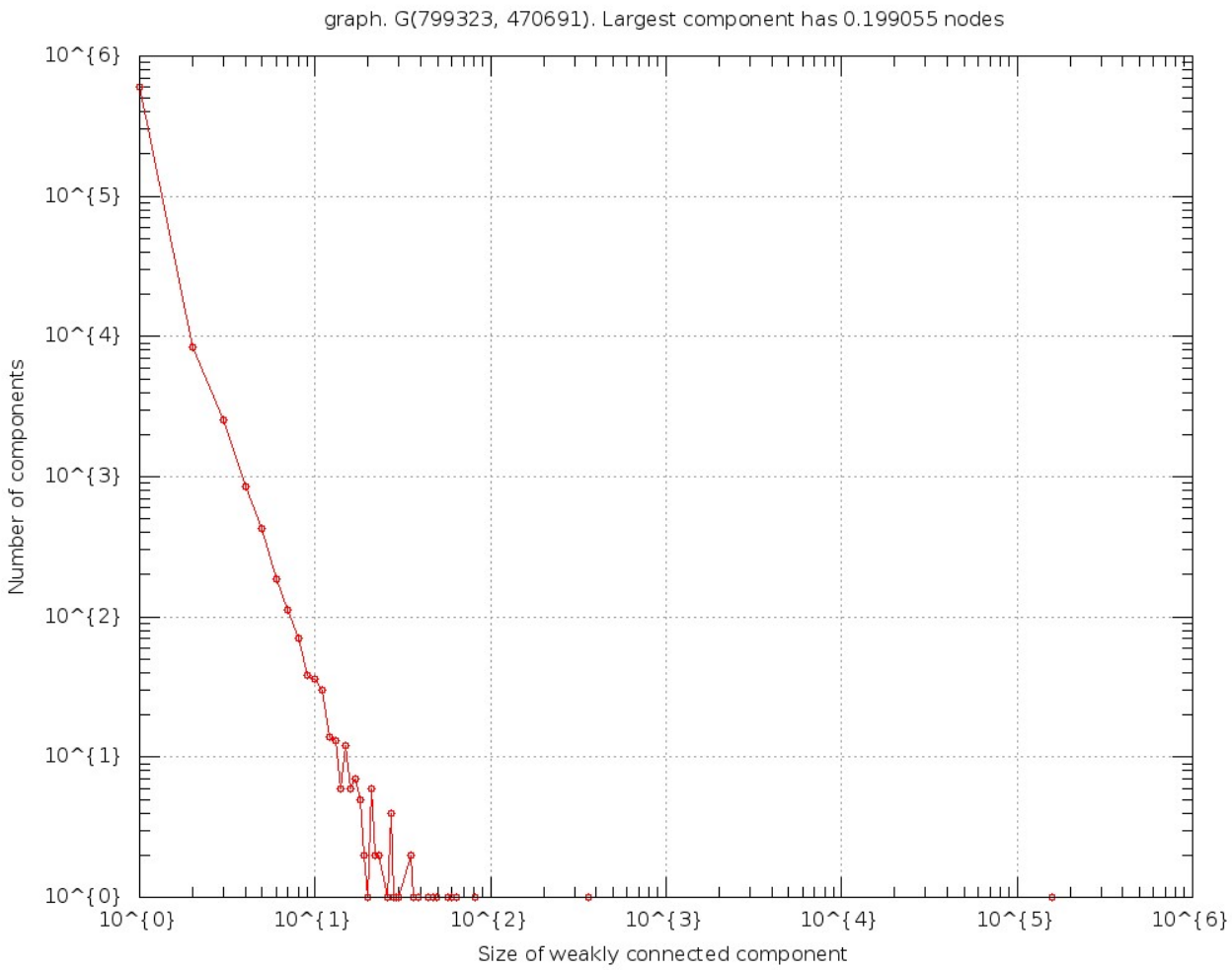

Figure 10. Graph on Number of components vs. Size of weakly connected component

CHAPTER FOUR : PROGRAMS AND TOOLS USED FOR DATASET DEVELOPMENT

The process of data creation was carried out in the following parts:

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

The data has been collected the legal database available online at www.liiofindia.org.

The program consists of mainly five functions:

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

resume_current progress()

this is the main controller function which resumes the state of the download. this enables the download to be continued if and when if there's any kind of abnormal termination of the program before the download has completed.

store_current_progress()

this is the function that stores the current state of download. The state of download is simply described by three integers written in a specially named text file.

parse_1()

this parses the required html links from the main page of www.liiofindia.org/databases/ and the obtained pages are passed on to the parse_2()

parse_2()

the pages sent to this function contains the index of all cases or articles of that particular court or journal.

Here there are 27 useful links which are passed on to the parse_3() function. Each of this links contains the list of cases/articles whose name start with a particular alphabet. The 27th link is the list of cases or articles starting with digits.

parse_3()

This is the function where the actual download occurs. After the download of each final html file we check if there is any external link to another pdf file if so, then we download the pdf file and store it instead of the html file.

All, the three parse functions call the store_current_progress() function with appropriate arguments.

CHAPTER FIVE : VALIDATION OF THE DATASET

There are 154 sub-databases in the Liiofindia legal database of which sub-databases 1 to 41 include judgements delivered by the Supreme Court of India, various High Courts and the District Courts. One judgement was randomly selected from each of these 41 sub-databases and validated for the citation details included in the metadata and the connection list data structures. Sub-databases 42 to 134 contains acts, regulations etc. of the various states of India. Since no citation information was present in the lawcite database, these documents have not been considered for graph modeling in the first phase of the work. Sub-databases 135 to 154 include articles published in the various law journals and these have been considered in the graph database preparation. One article was randomly selected from each of these journals and then validated for the citation details in the metadata and the connection list data structures.

```
1) District Court of Allahabad
Number of decisions: 0

2) High Court of Judicature at Allahabad
1987.INAHHC.1

3) High Court of Andhra Pradesh
2007.INAPHC.3
In Australian cases, the format changes to 2004.5.ALD.632. This
metadata information is different from the metadata used for the
Indian cases.
K Adivi Naidu v E Duruvasulu Naidu for this case,
1995 5 SCALE 455; (1995) 9 JT 593; (1995) 6 SCC 150 have been skipped.
Likewise for Venkata Reddi v Pothi Reddi
1963 2 SCR Supl 616; AIR 1963 SC 992 have been skipped.

4) High Court of Gauhati
2005.INASHC.1
SR Bommai v Union of India
(1994) 3 SCC 1; [1994] 2 SCR 644; 1994 2 SCALE 37; 1994 2 JT 215; AIR
1994 SC 1918  have been skipped.
Gujarat State Cooperative Land Development Bank Ltd v P R Manded
1979 3 SCC 123; AIR 1979 SC 1203 have been skipped.
T Cajee v U Jormanik Siem & Anr
[1961] 1 SCR 750; AIR 1961 SC 276 have been skipped.

5) District Court of Kamrup
2006.INASKPDC.1

6) High Court of Judicature at Patna
2003.INBRHC.2

7) Indian Appellate Tribunal for Electricity
2010.INATEl.5
```

8) Indian Central Administrative Tribunal
9) Central Information Commission of India
2014.INCIComm.1799

10) Indian Cyber Appellate Tribunal
2010.INCyberAT.5

How are we attending the Legislation Cited case when it is only mentioning Constitution?
Taylor v Taylor
45 LJCh 373 why missing?
Rao Shiv Rahadur Singh & Anr v Vindhya Pradesh
[1954] SCR 1038; AIR 1954 SC 322 " " ?
Deep Chand v Rajasthan
AIR 1961 SC 1527 " " ?
Air 1964 SC 358
[1964] 1 SCWR 57 " " ?
Raja Soap Factory v S P Shantharaj
[1965] 2 SCR 800; AIR 1965 SC 1449 " "?
Hussainara Khatoon v Home Secretary, Bihar, Govt of Bihar, Patna
(1980) 1 SCC 81; [1980] SCC (Cri) 23; AIR 1979 SC 1360; (1979) 3 SCR 169 " "?

Assistant Collector of Central Excise Chandan Nagar, West b v Dunlop India Ltd
(1985) 1 SCC 260; (1985) 2 SCR 190; 1984 2 SCALE 819; AIR 1985 SC 330; [1985] SCC (Tax) 75 " "?

11) Indian Intellectual Property Appellate Board
2009.INIPAB.5

12) Indian Railway Claims Tribunal
2009.INRCT.3

13) Supreme Court of India
2016.INSC.605 case missing.

14) District Court of Chandigarh
2010.INCHCHDC.7
[2009] ACJ 1298 Missing?

15) High Court of Chattisgarh
2005.INCTHC.6

16) District Court of Delhi
2003.INDLDLDC.17

17) High Court of Delhi
2012.INDLHC.1925
Adivekka v Hanamavva Kom Venkatesh,(2007)
7 SCC 91 is missing.

Kalyan Singh v SMT Chhoti
(1990) 1 SCC 266; 1989 2 SCALE 1238; 1989 4 JT 439; 1989 2 SCR Supl 356; AIR 1990 SC 396 " "?

Seth Beni Chand v SMT Kamla Kunwar
1976 4 SCC 554; 1977 1 SCR 578; AIR 1977 SC 63 " "?

18) High Court of Bombay at Goa
2005.INGAHC.179
Madhu v Bihar(AIR [1995] SC 1467 is missing.

19) High Court of Gujarat
2008.INGJHC.4
Legislation cited in constitution are not recorded.
Jai Dev v Punjab
[1962] INSC 219; 1962 3 SCR 489;
Of them even the first one is hyperlinked but is missing.

20) High Court of Himachal Pradesh
2007.INHPHC.4

21) High Court of Jharkhand
[2009] INJHHC 7

The non hyperlinked cases 7 SCC 507 and 4 SCC 578 are not mentioned although they have and 2 and 11 citation indices respectively.

22) District Court of Ranchi
[2007] INJHRADC 6
General Observation: the cases in the list are not in any specific order.

Air 1981 Pat 102; (2002) 7 SCC 764 (case name) AIR 1981 Pat 102 (Citation name) are different.

2002.7.SCC.764 Please take note similar case is cited in duplicates what should be our approach in this case?

Veeraswami v Ramanna AIR 1935 Mad 365 India - Tamil Nadu Missing no hyperlink

23) High Court of Jammu and Kashmir
[2006] INJKHC 15
Constitution has been referred to but not been mined for the article.
789412 1980.INSC.170
646437 1981.INSC.3
568735 1988.INSC.339
These cases have been labeled which are not even present. Also non hyperlinked cases are not being labeled.
[1988] INSC 347
[1981] INSC 12

[1961] INSC 172

IMPORTANT:
2014.INSC.164

24) High Court of Karnataka
[2009] INKAHC 1
There are two records in 2009 and 4 records in 2010. The corresponding filed in pdf are either blank or scrambled. All the six records have been noted in the database and all the six are isolated reports.

25) High Court of Kerala
[2007] INKLHC 5152
The Law Cite is blank but the there is a reference to 'Section 3 of Explosive Substances Act, 1908' which is not recorded.

26) High Court of Judicature at Bombay
[2001] INMHHC 10
The following five cases have been recorded in the database
349418 1969.INSC.174 ok
378075 1969.INSC.81 present in the law cite. How it has been identified?
377118 1970.INSC.44 ok
349254 1980.INSC.58 ok [1980] 3 SCR 224; 1980 3 SCC 162; AIR 1980 SC 1201 all these 3 have been additionally cited but without hyperlink. These 3 have not been included in the database.

458678 AIR.1980.Bom.484 ok
Constitution 'Article 227 of the Constitution of India' is present in the law cite but that has not been recorded in the database.

27) District Court of Bhopal
[2008] INMPBPDC 10

The judgement is in Hindi and in pdf format.
The Hindi pdf file contains reference to the Section 166 of the Motor Vehicles Act, 1988 (in Hini) but this has not been recorded in the database.

28) High Court of Madhya Pradesh
[2006] INMPHC 37
There are three citations 216528 604846 436551
There are references to the following and not recorded in the database:
Section 374 (2) of the Cr.P.C., Section 307 of the IPC

29) High Court of Orissa
[2007] INORHC 54
Only one citation in the judgement 327924 -- 1961.INSC.186
correctly mentioned in the database but associated citations 1962 2 SCR 333; AIR 1961 SC 1655 not retrieved and stored.
The following act is mentioned in the text.

u/s.8(1) of E.A. Act

30)High Court of Punjab and Haryana
[2009] INPBHC 31
Citations 774688 --> 2006.1.LACC.416
714751 --> 2004.1.LACC.164
209127 --> 2005.2.LACC.537

All these judgements relate to Australia - New South Wales. The graph nodes representing these judgements have indegree > 0 but outdegree = 0

There are act citations of the following type:
Section 4 of the Land Acquisition Act, 1894 (for short 'the ACt')
Section 6 of the Act

31) High Court of Rajasthan
[2004] INRJHC 25
There are no citations in the judgement. Correctly recorded.

32) District Court of Jodhpur
[2010] INRJJODC 9
There is one citation 9.F.1 which when clicked does not go to any judgement but goes to a law cite where 2010.INRJJODC.9 is referred.
The judgement is in Hindi pdf and includes a reference to 2003(1) Western Law Cases (SC) 501 but is not identified in the database.

33) High Court of Madras
[2003] INTNHC 101
There is no citation in the judgement. There is a reference to the constitution Art.226 of the Constitution of India not included in the database. The Law Cite cites the legislation.

34) District Court of Allahabad
[2007] INUPAHDC 2
It is correct with no references. But District Court of Allahabad records are occurring multiple times at 1 and 34.

35) High Court of Judicature at Allahabad
High Court of Judicature at Allahabad records are occuring multiple times at 1 and 34. It is perplexing that in 2 the label is INAHHC while in 35 the label is INUPHC. The same cases essentially will have different record names and treated as different entities in our graph. Why not make them equivalent as this is basically an LII database inconsistency.

[2002] INUPHC 17
1990.INSC.173 HAS THE FOLLOWING missing non hyperlinked references: 1990 2 SCC 715; [1990] SCC (L & S) 339; 1990 1 SCALE 839; 1990 2 JT 264; 1990 2 SCR 900; AIR 1990 SC 1607

1959.INSC.19 HAS THE FOLLOWING missing non hyperlinked references:
1959 2 SCR Supl 316; AIR 1959 SC 725.
[1996] INSC 107 HAS THE FOLLOWING missing non hyperlinked references:
1996 1 SCALE 636; 1996 1 JT 643
(208942) 1995.INSC.843 IS NOT PRESENT IN THE HYPERLINKS AT ALL.
(351788) 1990.INSC.1 IS NOT PRESENT IN THE HYPERLINKS AT ALL.
1965.INSC.86 HAS THE FOLLOWING missing non hyperlinked references:
[1965] 3 SCR 536; AIR 1966 SC 81.
[1961] INSC 172; HAS THE FOLLOWING missing non hyperlinked references:
[1962] 2 SCR 169; AIR 1961 SC 1731
[1962] INSC 290; HAS THE FOLLOWING missing non hyperlinked references:
1963 1 SCR Supl 676; AIR 1963 SC 786
[1954] INSC 57; HAS THE FOLLOWING missing non hyperlinked references:
[1955] 1 SCR 250; AIR 1954 SC 440
(337745) 1990.INSC.185 IS NOT PRESENT IN THE HYPERLINKS AT ALL.
[1985] INSC 244; HAS THE FOLLOWING missing non hyperlinked references:
(1986) 1 SCC 100; 1985 2 SCALE 1123; 1985 3 SCR Supl 766; AIR 1986 SC 391
(547670) 1954.INSC.32 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list.
(505244) 1996.INSC.8 IS NOT PRESENT IN THE HYPERLINKS AT ALL.
634813 1962.INSC.109 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1962.INSC.290 which is in the hyperlink and also in the connection list.

36) High Court of Allahabad, Lucknow Bench
2007.INUPLUHC.3
(797762) 1985.INSC.1 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list.
[1990] INSC 32; HAS THE FOLLOWING missing non hyperlinked references:
1990 1 SCALE 156; 1990 2 SCC 48
[1985] INSC 112; HAS THE FOLLOWING missing non hyperlinked references:
1985 1 SCALE 1091; [1985] SCC Supl 94; 1985 1 SCR Supl 101; AIR 1985 SC 1124
[1990] INSC 173; HAS THE FOLLOWING missing non hyperlinked references:
1990 2 SCC 715; [1990] SCC (L & S) 339; 1990 1 SCALE 839; 1990 2 JT 264; 1990 2 SCR 900; AIR 1990 SC 1607
[1953] INSC 21; HAS THE FOLLOWING missing non hyperlinked references:
[1953] SCR 655; AIR 1953 SC 250
(351788) 1990.INSC.1 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list.
(62054) 1985.INSC.116 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list. Is there a confusion with 1985.INSC.12?
[1983] INSC 204; HAS THE FOLLOWING missing non hyperlinked references: (1984) 2 SCC 141; [1984] 2 SCR 200; 1983 2 SCALE 1060; AIR 1984 SC 541

(337745) 1990.INSC.185 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list.
[1982] INSC 24; HAS THE FOLLOWING missing non hyperlinked references:
[1982] 3 SCR 298; HAS THE FOLLOWING missing non hyperlinked references: 1982 1 SCC 618; 1982 1 SCALE 110; AIR 1982 SC 879; [1982] SCC (L & S) 11
(292791) 1991.INSC.22 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list.
(760825) 1981.INSC.191 is in the connection list but not in the hyperlinks. Has some mismatch occurred with 1954.INSC.57 which is in the hyperlink and also in the connection list.
[1988] INSC 306; HAS THE FOLLOWING missing non hyperlinked references: 1988 2 SCALE 827; 1988 4 JT 53; 1988 3 SCR Supl 288; AIR 1989 SC 19

Prem Chand Somchand Shah v Union of India,(1991) 2 SCC 48 has no hyperlink and no mention. Still it is an entry.

[1995] INSC 846; HAS THE FOLLOWING missing non hyperlinked references: (1995) 7 SCALE 224; (1995) 9 JT 142; (1996) 7 SCC 256

37) High Court of Uttarakhand
[2009] INUTHC 223 correct with no reference. Randomly [2009] INUTHC 185 and [2009] INUTHC 97 were picked which were also correct with no references.

38) District Court of Nainital
[2007] INUTNADC 11 is correct with all 3 references pointed correctly.

39) High Court of Calcutta
[2006] INWBKOHC 90 correct with no reference. Randomly [2011] INWBKOHC 15 was picked which were also correct with no references.

40) High Court of Calcutta (Appellate Side)
Randomly [2004] INWBKOHCA 2, [2011] INWBKOHCA 69, [2011] INWBKOHCA 78 are all correct with no citations. Basically in the appeal side, citations are few.
But the consistency of LII Database is inconsistent with no records for many non trivial cases.

41) High Court of Calcutta Port Blair Bench
[2011] INWBKOHCPB 79 [2011] INWBKOHCPB 40 and [2011] INWBKOHCPB 33 were randomly picked but have no citations. In the LII Database, it seems there will be fewer edges for a set of nodes than expected.

136) GNLU Journal of Law, Development and Politics
 [2009] GNLUJlLDP 13
 contains no citations and correctly recorded in the database

137) Indian Journal of Constitutional Law
 [2007] INJlConLaw 9

References to section 8 paragraph 8 of the Anti-discrimination Act and Article 12 paragraph 2 of the Constitution both of Slovak Government.
The database does not contain these references

138) Indian Journal of Intellectual Property Law
[2012] INJlIPLaw 10
It contains no citations and this has been correctly noted in the database

139) Journal of Intellectual Property Rights
[2010] INJlIPR 5
contains references to Public Funded Intellectual Property Bill, 2008 but not in the database.

140) Indian Journal of Law and Economics
[2010] INJlLawEcon 7

The site contains 9 articles. All have been included.

141) Indian Journal of Law and Technology
[2007] INJlLawTech 4

All included.

Note: the Journal articles contains references to other articles.

142) ISIL Year Book Of International Humanitarian and Refugee Law
The liiofindia site says 404 file not fund while the metadata database shows 2 entries.

143) NALSAR Environmental Law and Practice Review
[2011] NALSAREnvLawPRw 7
Contains references to Marine and Coastal Access Act
Correctly recorded

144) NALSAR Law Research Series
[2012] NALSARLRS 10

145) NALSAR Law Review
[2013] NALSARLawRw 4

146) NALSAR Media Law Review
[2010] NALSARMLawRw 7

147) NALSAR Student Law Review
[2011] NALSARStuLawRw 10

148) NLUD Law Research Series
[2012] NLUDLRS 19

149) NLUD Student Law Journal
[2012] NLUDStuLawJl 3

150) NUJS Law Review
[2009] NUJSLawRw 1

151) Indian Parliamentary Research Service Legislative Summaries
[2010] INPRSLS 5
mentions The Orissa (Alteration of Name) Bill, 2010 but not in the database

152) Law Commission of India
[2009] INLC 17
The database links 4 citations but the law cite link mentions only one.

153) Indian Treaties
The hyperlink points to the main database meu.

154) Indian Treaty Series
[2009] INTSer 1
The following two legislations are cited Database Search and Name Search But they link to the search programs. There is no citation in the database.

CHAPTER SIX: VARIOUS TEXT PATTERNS INDIGENOUS TO THE LEGAL GRAPH DATABASE

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
```

CHAPTER SEVEN: FUTURE WORK & CONCLUSION

The first phase of the design and development of the Graph Modeling of Indian Legal database based on the legal documents and articles in the website www.liiofindia.org of the Legal Information Institute of India has been described in the present report. At present, only judgements delivered by the Supreme Court of India, various High Courts and the District Courts of India have been modeled with emphasis on citation analysis. In addition, journal articles published in the various law journals have been included in the graph database with emphasis on citation analysis. In the next phase, the various acts, regulations of the different states of India as mentioned in the liiofindia database will be considered for their inclusion in the graph database. Moreover, the validation tests have identified several issues which are to be addressed. It has been specifically identified that in many judgements there are citations to the Constitution articles as well as to various acts. The citation records in the lawcite database have missed such citation details. Various text patterns have been identified to identify the occurrences of the citations to Constitution articles or the various acts. Text analytic programs have already been developed for these purpose and these are now being integrated in the main system. In subsequent phases, topic based linking of the various graph nodes will be identified.

CHAPTER EIGHT: 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